MAPPING AND ANALYSIS OF HUMAN DISEASE NETWORK MAP (*DISEASOME*) ON MOUSE GENOTYPE & PHENOTYPE NETWORK

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

SULTAN NİLAY CAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE
IN
BIOINFORMATICS

AUGUST 2017

MAPPING AND ANALYSIS OF HUMAN DISEASE NETWORK MAP (*DISEASOME*) ON MOUSE GENOTYPE & PHENOTYPE NETWORK

Submitted by **SULTAN NILAY CAN** in partial fulfillment of the requirements for the degree of **Master of Science in Bioinformatics, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin _____
Director, Graduate School of **Informatics**

Assoc. Prof. Dr. Yeşim Aydın Son _____
Director, **Health Informatics**

Prof. Dr. Rengül Çetin-Atalay _____
Supervisor, **Health Informatics**

Dr. Tunca Doğan _____
Co-Supervisor, **Health Informatics**

Examining Committee Members:

Prof. Dr. Tolga Can _____
Computer Engineering, Middle East Technical University

Prof. Dr. Rengül Çetin-Atalay _____
Health Informatics, Middle East Technical University

Assist. Prof. Dr. Aybar Can Acar _____
Health Informatics, Middle East Technical University

Assoc. Prof. Dr. Özlen Konu _____
Molecular Biology and Genetics, Bilkent University

Dr. Nurcan Tunçbağ _____
Health Informatics, Middle East Technical University

Date: 11.08.2017

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: SULTAN NİLAY CAN

Signature: _____

ABSTRACT

MAPPING AND ANALYSIS OF HUMAN DISEASE NETWORK MAP (*DISEASOME*) ON MOUSE GENOTYPE & PHENOTYPE NETWORK

Sultan Nilay CAN

MSc, Bioinformatics

Supervisor: Prof. Dr. Rengül ÇETİN-ATALAY

Co-Supervisor: Dr. Tunca DOĞAN

June 2017, 146 Pages

Mouse is the primary model organism to study mammalian genetics. The genome of mouse is incisively and specifically modified and controlled to study the mutations in the human genome, to discover the molecular mechanisms of various complex human diseases such as cancers, diabetes, hereditary and neurological disorders. Various ontology systems have been constructed to express metabolic functions and diseases as controlled vocabulary terms. This way, abstract definitions such as gene functions, diseases or phenotypes become machine readable and quantifiable data. Mammalian Phenotype Ontology (MPO) is one of these databases that generates standardized terms to define phenotyping textures in mammals by carrying out gene knock out experiments in mice, which was followed by the observation of abnormal phenotypes.

In a previous study, biological networks were designed to analyse the relationships between complex human diseases and the genes responsible for the occurrence of those diseases. Human disease network focused on 22 different disease classes and brought insight to the complex relations between different disease classes. This study aims to map the human disease network onto the mouse genotype/phenotype data by generating multi-partite networks of human diseases – human/mouse genes – phenotypic abnormalities observed in targeted knock-out-mouse models. The resulting networks are presented to the research community in an online interactive platform. The output of this work is expected to aid experimental researchers to select the appropriate targeted knock-out mouse models to study a specific human disease. Furthermore, the mappings between disease and phenotype terms is expected to enrich the ongoing efforts to curate specific symptoms and effects of diseases to improve medical diagnosis.

Keywords: Human diseases, abnormal phenotypes, mouse knock out genes, biological networks

ÖZ

FARE FENOTİPİ VE GENOTİPİ ÜZERİNDE, İNSAN HASTALIK AĞININ (*DISEASOME*) HARİTALAMA VE ANALİZİNİN YAPILMASI

Sultan Nilay CAN
Yüksek Lisans, Biyoenformatik
Tez Yöneticisi: Prof. Dr. Rengül ÇETİN-ATALAY
Ortak Tez Yöneticisi: Dr. Tunca DOĞAN

Haziran 2017, 146 sayfa

Fare, memeli genetiğini çalışmak için kullanılan temel model organizmadır. İnsan genomundaki mutasyonları çalışmak ve kanser, diyabet, kalıtsal ve sinirsel birçok kompleks insan hastalığının mekanizmasını anlamak için, memeli genetiğinde temel bir organizma olan fare genomu isabetli ve spesifik olarak değiştirilebilir ve kontrol edilebilir olarak kullanılmaktadır. Metabolik fonksiyonları ve hastalıkları, organizmalar üzerindeki fenotipik yansımalarını da hesaba katarak anlamak için birçok ontoloji sistemi yapılandırılmıştır. Memeli sistemleri için fenotipleme özelliğini tanımlamak amaçlı standartlaştırılmış birçok terimi barındıran Memeli Fenotipi Ontolojisi (MPO) bu özelleşmiş veri bankalarından biridir ve farede anormal fenotiplerle sonuçlanan nakavt çalışmalarını yürütmek, fenotipik terimleri tanımlamak

için standartlaştırılmış tanımlar üretir. Önceki bir çalışmada biyolojik haritalamalar, kompleks hastalıklar arasındaki ilişkileri ve bu hastalıklardan sorumlu olan genleri ve kendi aralarındaki ilişkileri çalışmak amaçlı dizayn edilmişlerdir. Bu tez, insan hastalıklarının ve nakavt fare çalışmalarından elde edilmiş fenotipik anormalliklerin çok parçalı ağlarını üreterek, insan hastalık ağını, fare fenotipi ve genotipi veri setinin üzerine haritalamayı amaçlamaktadır. Sonuç olarak elde edilecek olan haritalamalar, araştırma dünyasına çevrimiçi bir platform olarak sunulmuştur. Bu çalışmanın, insan hastalıkları üzerine gerçekleştirilmekte olan deneysel araştırmalarda uygun nakavt fare modellerinin seçilmesine yardımcı olması beklenmektedir. Ayrıca, hastalıklar ve fenotipik terimler arası yapılan bu haritalamanın, tıbbi teşhis ve tedavilerin geliştirilmesi amacıyla yapılan ontolojik çalışmalara katkıda bulunması beklenmektedir.

Anahtar Sözcükler: İnsan hastalıkları, hastalık fenotipleri, fare nakavt genleri, biyolojik ağlar.

To my father Yusuf CAN

and

My mother Salime CAN

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor Prof. Rengül Atalay for her support of my Master thesis and research, for her patience, motivation, enthusiasm, and knowledge. Her guidance helped me a lot in all the time of this thesis.

I would like to thank also my co-advisor Dr. Tunca Doğan for his remarkable support, precious comments, and hard questions, which led me to success.

I would also like to thank my committee members, professor Tolga Can, associated professor Aybar Can Acar, associated professor Özlen Konu and associated professor Nurcan Tunçbağ for serving as my committee members. I also want to thank them for taking time, letting my defense be a great moment, and for their brilliant comments and suggestions.

My thanks go to associated professor Aybar Can Acar because of his insightful comments and supports in analyzing and writing this thesis.

I thank my fellow group mates Kübra Narcı, Damla Gözen, Mona Shojaei and Alperen Dalkıran who support me, helped me in editing, for answering my never-ending questions and criticized me while writing this thesis.

I would like to show my gratitude to my friends Tuğrul Tuncer, Hakan Özkök, Gamze Tanık, Volkan Orhan, Güzin Erdem, Ecem Uzun, Onur Akyürek, Ece Beşiroğlu and who gave me their unconditional support, help and advices during these years.

I wish to present my huge thanks and great love to my dear brother Önay Can who helped me a lot in many ways during this thesis. Thank you for your encouragement,

and all your supports. I also would like to thank my spiritual sister Güneş Carmichael, nephews Frida Maya and Martin Toprak for their sincere support and love.

Special thanks to my family, words cannot express how grateful I am to my mother, father and brother for all the sacrifices that they made on my behalf. Your prayer for me was what sustained me thus far.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DAG	**Directed Acyclic Graph**
DGN	**Disease Gene Network**
DISEASOME	**The Human Disease Network**
DNA	**Deoxyribonucleic acid**
DO	**Disease Ontology**
GO	**Gene Ontology**
HDN	**Human Disease Network**
HPO	**Human Phenotype Ontology**
MGI	**Mouse Genome Informatics**
MP	**Mammalian Phenotype**
MPO	**Mammalian Phenotype Ontology**
OBO	**Open Biological and Biomedical Ontologies**
OMIM	**Online Mendelian Inheritance in Man**
XML FORMAT	**Extensible Markup Language Format**

CHAPTER 1

INTRODUCTION

1.1 MOTIVATION

Model organisms have long been experimented on to understand complex molecular mechanisms in the human body. When the Human Genome Project (HGP) was started in 1990, mouse was included as one of the five central model organisms with the purposes of understanding the gene functions, disease mechanisms and for discovering new drugs (Waterston *et al.*, 2002).

There are biological data resources to store and freely publish the finding obtained from the experimental studies on mouse. One example is Mouse Genome Informatics - MGI (Blake *et al.*, 2011), which is documenting the relations between mouse genotype and phenotype. MGI is discussed in section 2.2, in detail. There are also open-access resources that document the relations between the human genome and the genetic diseases, such as OMIM (Hamosh *et al.*, 2005) and *Diseasome* (Goh *et al.*, 2007), which is discussed under section 2.5. Both the human and mouse resources are extremely valuable for the research community and the underlying knowledge have significant overlaps due to the genetic similarities between human and mouse. However, the resources on mouse has a lot more to offer compared to resources on human, due to extensive systematic experimental research carried out on mouse. As a result, integrating the information found in mouse data resources to human datasets have the potential to extensively enlarge our understanding about the relation between human genome and phenome, especially in terms of genetic human diseases. This

understanding may in turn help researchers to develop novel treatments to stop these diseases.

As far as we are aware, the studies aiming to integrate the biomedical data on model organisms with the human data exists, though scarce. We believe that more focus is required for biological data integration and inferring biological insight from the results. Automated computational approaches should be used for this purpose, as the data volume is now beyond the capabilities of manual curation now. Open access tools and services that will be generated to house and to present the integrated data to the life science research community is the key to be able to analyze this huge amount of data and to obtain biological knowledge from it.

1.2 SCOPE AND GOAL

The main objective of this thesis is to generate a biological network composed of disease records, disease causing genes and observed abnormalities in the form of phenotypic terms. This is done by analyzing the *Diseasome* (the Human Disease Network) and mapping it onto the mouse genotypic vs. phenotypic relation data. This way, associations between abnormal mouse phenotypes and human diseases are provided by using mouse knock-out genes and their human orthologues as the key attribute between two data sources. The main output of this thesis is an open access online network that visualizes these relations interactively, in a map format.

As the first step, *Diseasome* database source published by Goh et. al. in 2007 was used as a data resource and the list of human disorders, disease genes, and associations between them were obtained from their datasets. Mouse Genome Informatics (MGI) and Mammalian Phenotype Ontology (MPO) databases were used to collect mouse affected systems (abnormal phenotypes) and the associated mouse knock-out genes. The data derived from MGI/MPO and *Diseasome* were integrated to generate the data tables.

In second part of the study, python scripts were written to produce gexf formatted files, to run on Gephi graph visualization tool, for constructing biological networks in two different approaches, which are Genes-Node and Genes-Edge versions. Genes-Node version treat mouse genes, affected systems (phenotypes) and human diseases as nodes and the edges represent the direct relations in-between. Genes-Edge version patterns genes as edges that connect human disease nodes with mouse affected system (i.e. abnormal phenotypes) nodes. Gephi tool was used to process these gexf files, to analyze them in terms of simple graph theory concepts and finally to visualize the undirected graphs on web-browsers via an exporter tool.

The main objective of this thesis is to provide a practical online tool for the use of experimental and computational researchers working on genetic diseases, and providing case studies on how the proposed tool can be utilized to infer biological insight. We seek to aid laboratory scientists to prepare their knockout mouse models by using our online tool. This study is also expected to aid the computational studies on the development and the annotation of ontological systems for medical diagnosis and treatment.

1.3 CONTRIBUTION

The main contribution of this study is to provide an open access tool that displays the associations between diseases, genes -that cause these diseases when they possess specific mutations- and the observed abnormalities when those genes do not function properly. The produced output is expected to help laboratory scientists to observe targeted knock-out mouse gene models to select relevant models for studying specific human diseases. Another contribution is encouraging researchers to investigate the novel human disease - phenotype associations, which may aid the development of ontological medical diagnosis systems. Newly discovered relations between phenotypes, diseases and genes can be utilized in the forthcoming studies in the field of biomedicine.

1.4 OUTLINE

This thesis comprises 5 chapters. These chapters are entitled as “Introduction”, “Background and Related Works”, “Materials and Methods”, “Results” and “Discussion”, respectively.

The first chapter gives a short introduction followed by the scope and objective of the study. The second chapter provides a short description of the basic concepts used in the proposed research together with the related work in the literature (i.e. information about the targeted knock-out mice studies, *Diseasome* with its analysis details, MPO and MGI databases and a brief information about the general phenotypic studies in the mouse and human genomes). In the third part (Material and methods), the details about data preparation and integration are provided and illustrated exclusively. Moreover, network visualization tool Gephi is exhibited under this section, together with the technical details on the proposed web-service Mouıse2HumanNet. The fourth chapter (Results) includes the information related to statistical and network analysis of the generated networks, together with case studies. Finally, the fifth chapter (discussion) sums up the work done, discusses the results and offers possible modifications, alterations and developments as potential future studies.

CHAPTER 2

BACKGROUND AND RELATED WORKS

2.1 MOUSE AS A MODEL ORGANISM

Mouse is one of the most preferred model organisms for the research on human physiology and pathophysiology (Rosenthal and Brown, 2007). Mouse models have been used comprehensively to understand the mechanisms of human diseases, to explore the effects of drugs and to predict patient provisions.

Genetic resemblance between mouse and human organisms is the reason behind using mouse as a model organism to study human diseases. More than 90% of the mouse and human genomes can be divided into related conserved synteny regions, which show the gene order in the genomes. These regions are highly conserved in both human and mouse genomes (Waterston *et al.*, 2002). It is also stated that, both species have similar number of protein coding genes (Guénet, 2005).

Another reason why mouse is a suitable model organism to model human diseases and deficiencies is that mice is easy to maintain and breed in the laboratory conditions. The typical "life span" of the mouse approximately ranges from 1.3 to 3 years for various strains. As a result, their lifetime can be studied in a reasonable period (Comfort, 1959). Furthermore, mice have been used in numerous experimental genetic studies up to this day; as a result, the generated collective practical experience is shared by researchers all over the world.

However, there also exist genetic differences between the mouse and human, which is reflected onto their distinct physiological and anatomical characteristics. These

differences are resulted from the accumulation of various types of mutations on the genomes of human and mice after their divergence from their common ancestor. Naturally, there also is a divergence between the human and mouse at the systemic level such as the regulatory factors, immune system gene activities, stress response and metabolic periods (Comparing the mouse and human Genomes, 2015).

Various studies in the literature have investigated the genetic differences between mouse and human with the purpose of modifying the mouse genome to study human physiology on mouse models. The way to achieve this lies in the field of genetic engineering, which is dealing with the direct manipulation of DNA to change an organism's genotype in a desired way. Gene targeting –one of the various genetic manipulation methods– allows researchers to introduce mutations at specific loci in the target organism. For example, targeted deletion of a specific gene in mouse is frequently used to determine the biological role of the in-activated/deleted gene.

2.1.1 THE KNOCK-OUT MOUSE

A targeted knock-out mouse is a laboratory animal where a specific gene was inactivated, in other words "knocked out" by researchers. The practical application is usually carried out by replacing the existing gene or damaging it with an artificial piece of DNA. During the 1980's, a Dr. Mario Capecchi invented a procedure to remove or change any single gene in the mouse genome (Capecchi, 2008). Mouse strains were constructed in such a way that the altered genes pass from parent to its offspring.

The discovery of mouse embryonic stem (ES) generating cell lines allowed for the generation of the efficiently targeted knock-out mouse (Limaye, Hal, & Kulkarni,2009). ES cells were reproduced from embryos at a developmental stage before implantation. Fertilization normally occurs in the oviduct, and throughout few days a series of cleavage divisions occur. The embryo rides down the oviduct and into the uterus. Embryo cells are undifferentiated in each cleavage-stage. Indeed, each of these cells has a potential to give rise to any cell for the body. The first fractionation in human organism occurs at about five days of development. Outer layer of cells self-

dedicate themselves to become a part of the placenta and separates from the inner cell mass (ICM). The ICM cells can generate any cell type of the body. If the ICM is removed from its environment and cultured, these cells can continue to proliferate and replicate themselves indefinitely. These cells can maintain the developmental potential to form any cell type of the body. These ICM-derived cells are ES cells. It is important to notice that ES cells do not exist in vivo; they should be considered as a tissue culture artifact (Winslow, 2017).

Gene targeting and homologous recombination are the preferred ways of building a targeted knock-out mutation in a mouse. Homologous recombination is a DNA repair mechanism and it has been made up by inserting a specific mutation into the homologous genetic locus (Majzoub and Muglia, 1996). During gene targeting or homologous recombination, manipulation of the gene is occurred in the nucleus of an ES cell. This is done by introducing an artificial piece of DNA that shares identical or homologous sequence to the gene. This homologous sequence flanks the existing gene's DNA sequence both upstream and downstream of the gene's location. The cell recognizes the identical stretches of sequence and wipes out the existing gene or portion of this gene with the artificial piece of DNA. Because the artificial DNA is inactive, the wipe eliminates, or "knocks out," the function of this gene. In the second strategy, called gene trapping, again a gene in an ES cell is manipulated. However, instead of directly targeting a gene of concern, a random process is preferred. A piece of artificial DNA containing a reporter gene is constructed to be inserted randomly into any gene. The inserted piece of artificial DNA prevents the cell's RNA "splicing" mechanism to work properly, thus gene’s function is knocked out.

When the gene loses its activity, various alterations can be observed in the mouse phenotype. These phenotypical alterations can be anatomical, behavioral, biochemical or physical (Austin *et al.*, 2004). The knockout mice specifically constructed to study human mutations are eminent sources to study pathophysiology and may serve to find novel therapies for genetic diseases (Majzoub and Muglia, 1996). These works have led to various discoveries about human diseases, from cancer to obesity.

2.1.2 PURPOSES BEHIND USING KNOCK-OUT MICE

Human organism shares various similar genes with the mice. Therefore, observing the main characteristics of knocked-out mice can give valuable information regarding the human genetic disorders. A study stated that mice have been used widely to enlighten the mechanism behind human diseases and increase the efficacy of drugs (Vandamme, 2014; Justice and Dhillon, 2016). There have been progress for understanding critical human diseases such as cancer, obesity, heart disease, diabetes, anxiety, aging and Parkinson disease thanks to the mouse studies, as the knockout mice serve critical information about how the knocked-out gene normally functions in the body.

IMPC (International Mouse Phenotyping Consortium) is a freely available and useful platform for human disease investigations (White *et al.*, 2013) and this consortium is creating targeted knock-out mutations for various protein coding orthologue genes in the mouse genome. Orthologous genes are defined as homologs in different species, which diverged from each other following a speciation event (Jensen, 2001). It is reported that usually the function is conserved between orthologous genes. The main aim of IMPC is to explore the machinery and functions of 20,000 common genes between mouse and human. It provides a platform to examine the mechanisms of human disorders.

It has been stated that the identification of the essential genes in mouse will help to analyze genetic human diseases. Essential genes can be defined as the genes required for the life of any human cell. In the perspective of the collaboration with The Exome Aggregation Consortium (ExAC) it was demonstrated that, these genes are valuable nominees for various undiagnosed human genetic conditions (Lek *et al.*, 2016). ExAC is created for harmonizing and clustering the exome sequencing data of large scale sequencing projects.

2.2 MOUSE GENOME INFORMATICS (MGI)

Mouse Genome Informatics (MGI) is an international database of scientific information obtained by experimenting with the genome of laboratory mouse. It is considered and acknowledged as the most comprehensive resource covering the genomic features of the mice. It also facilitates human health and disease studies.

There exist various projects contributed to MGI can be listed as:

- Mouse Genome Database (MGD) Project
- Gene Expression Database (GXD) Project
- Mouse Tumor Biology (MTB) Database Project
- Gene Ontology (GO) Project at MGI
- MouseMine Project

First project that contributed to MGI is MGD project (Blake *et al.*, 2011), which was carried out in Jackson's laboratory. MGD includes various types information such as GO, MPO and human diseases in OMIM. It provides a genetic map, a genome browser (Mouse Jbrowse), Single Nucleotide Polymorphisms (SNPs) information and mammalian orthology data.

Second project is the Gene Expression Database (GXD) and constructed to extract gene expression profiles for the laboratory mouse. There exists emphasis on endogenous gene expression during the development of mouse.

Another project is the Mouse Tumor Biology Database (MTB), established to mine experimental models, review specific cancers and detecting genes that are mutated in cancers.

Other one is Gene ontology project at MGI, which is a part of the Gene Ontology Consortium that provides vocabularies for describing the MF, BP, and CC of gene products. GO team members at MGI contribute to develop specific ontological terms for mouse and functional curation of mouse gene products.

Lastly, MouseMine is a very powerful online platform which serves a system using mouse data from MGI. It includes nomenclature, synonyms, database cross references, genome coordinates, the mouse allele catalog, spontaneous and engineered mutants, mutant cell lines, mouse strains and genotypes. Also, it consists mouse functional (GO) annotations, phenotype (MP) annotations, disease (OMIM) annotations, human genes and their genome coordinates (via EntrezGene); mouse/human orthologues and mouse/mouse paralogues, mouse/mouse and mouse/human protein-protein interaction data from Database of Protein, Chemical, and Genetic Interactions (BioGrid) and European Bioinformatics Institute (IntAct); plus, publications, notes, and external database references. MGI also provides an investigation tool called as "batch summary".

2.3 BASIC CONCEPTS IN GRAPH THEORY AND NETWORK ANALYSIS

The computational methodology to generate the proposed tool in this study is based on the graph theory and on network analysis. As a result, an introduction on the basic concepts in graph theory and network analysis is required.

A graph is a pair of sets (V, E) where V is defined as a finite set called the set of vertices and E is a set of 2-element subsets of V, called the set of edges. A network can be defined as a graph where nodes and/or edges have labels in other words attributes. In graph theory, various concepts are employed to analyze a network. One of the basic terms, a walk is defined as any route from vertex to vertex along edges and it can end on the same vertex where it began or on a different vertex. A path is a walk that does not include any vertex twice, except that its first vertex can be the same as its last. A trail is defined as a walk with no repeated edge. A cycle is defined as a closed path. Edges do not have an orientation in undirected graph and undirected graph is connected if there is a path between each pair of vertices and if it has no cycle, it is called as acyclic, it is defined as a tree if any two vertices are connected by exactly one path and it is named as acyclic - bipartite if V is partitioned into two independent sets.

Networks used in this thesis are constructed as undirected and modified from the bipartite *Diseasome* design mapping.

Furthermore, degree, average weighted degree, graph diameter, graph density, modularity and eigenvector centrality terms are frequently used to reveal various characteristics of networks.

For undirected networks, the node degree term is the number of edges linked to node n. A self-loop of edges is counted as two edges for the node degree (Seymour, Schrijver, and Diestel, 2005). In degree and out degree terms are used for directed graphs, not applicable for the undirected ones. In degree means incoming edges to a node n and out degree means outgoing edges from the node n. Weighted degree is the weight of each edge related to node n.

Graph diameter can be defined as the maximum of the shortest paths between any two of the vertices in graph, in other words, it is the maximum eccentricity of any vertex in the graph. The maximum eccentricity is the graph diameter. The eccentricity of any vertex denoted as v in a connected graph is the maximum graph distance between this vertex v and any other vertex u.

Graph density is a measure that shows how strongly network elements have connected each other. It is calculated as dividing the number of edges in network to the all possible connections. It takes a value between 0 and 1.

Modularity can be one of the most frequently used quality function for community detection in networks (Jin, Girvan, and Newman, 2001). It is a representation of sum of the number of edges in the communities minus the expected fraction of such edges if they are placed at random with the same distribution of vertex degree (Newman and Girvan, 2004). In other words, modularity compares the number of edges in a cluster with the expected number of edges that can be found in a cluster. It indicates the importance of a node while considering its connections in a network and it gives relative scores to each node. Modularity also measures the robustness of a network (Labs, 2012). It has been stated that the modularity issue suffers from resolution limit and therefore sometimes it is unable to detect small communities or cliques. If a

network is considered as large enough, the expected number of edges between two groups of nodes in a model with null modularity can be smaller than one. In that case, a single edge between two clusters would be interpreted by modularity. Therefore, even weakly interconnected complete graphs would be merged by modularity optimization if the network were sufficiently large (Fortunato and Barthélemy, 2007).

Another important term is directed acyclic graph (DAG) for understanding the logic of this study. In DAG structure, one node is named as a root node, and all the other nodes are constructed as leaf nodes. It is declared that DAG having established hierarchical parent-child relations between all neighbor nodes proceeding from the root node down to any leaf nodes. The difference between a tree and a DAG is the possibility of more paths between two nodes in the DAG structure. In other words, an undirected graph is named as a tree if there exist exactly one simple path between each pair of vertices.

A Connected component defines a subgraph where any two of its vertices are connected to each other by common paths (i.e. there is no non-connected vertices in a connected component), whereas a maximal clique defines a component whose all vertices are fully connected to each other.

2.4 MAMMALIAN PHENOTYPE ONTOLOGY (MPO)

Phenotype is a term that describes observable morphological, physiological and behavioral characteristics of an individual. Phenotypic characters can appear, disappear, increase or decrease in lifetime. Environmental facts can change the phenotypic characters. Phenotypic variation can be explained with the individual's genetic and environmental history. Various human diseases are associated with both environmental and genetic characters. Also, it is possible that some variantions in germline cells may lead to inherited syndromes that are passed to the offspring (Smith and Eppig, 2009).

A comprehensive database called the Mammalian Phenotype Ontology (MPO) has been constructed under the MGI resource to catalogue tens of thousands of mutations in the mouse genome and their related phenotypes. Phenotypic terms are stored in a specialized format to describe abnormal mammalian phenotypes in a hierarchical format. Root node is named as the "Mammalian Phenotype" in that hierarchy and it divaricated into 30 different terms called high-level phenotypes, which are related to the physiological systems, survival and behavioral conditions. Each term describes a unique phenotype and displayed with its unique MP ID. Besides this ID, it consists term name, a synonym (if any) and a detailed definition of the content. Every phenotypic term that is inherited from a term in a higher level in the hierarchy is called as "child" of the parent term. Their direct parent phenotypes called as "parent" of the child term. Any term should have at least one parent except the root term "Mammalian Phenotype".

All phenotypic information in MPO is kept in OBO (Open Biological and Biomedical Ontologies) format. The OBO is one of the machine-readable formats implemented for easy data query, mining, and manipulation. One of the properties of OBO is that it is constructed as easily human readable compared to the XML.

Mammalian phenotype browser serves the users with the stored phenotypic terms and their relations. Under phenotype search bar, the recorded phenotypic information can be viewed in a DAG format. Additional information is given under "Phenotype Term Detail" part with terms, synonyms, definitions, parent terms and IDs. According to the MGI statistics, as of 2017, 11,464 mammalian phenotype (MP) terms are generated and stored. MP terms can be searched by typing its name directly on the query column. For example, the term "abnormal brown fat cell morphology" (id: MP:0009116) was searched and the relationships are illustrated in Figure 1. In this example, "abnormal brown fat cell morphology" term has two parent phenotype terms namely: "abnormal brown adipose tissue morphology" and "abnormal fat cell morphology".

To sum up, MPO is a collection of controlled vocabulary terms to define abnormal phenotypes observed in mouse experiments. These phenotypes have been annotated to

mouse genes, which lead to the corresponding phenotypic traits due to certain mutations.

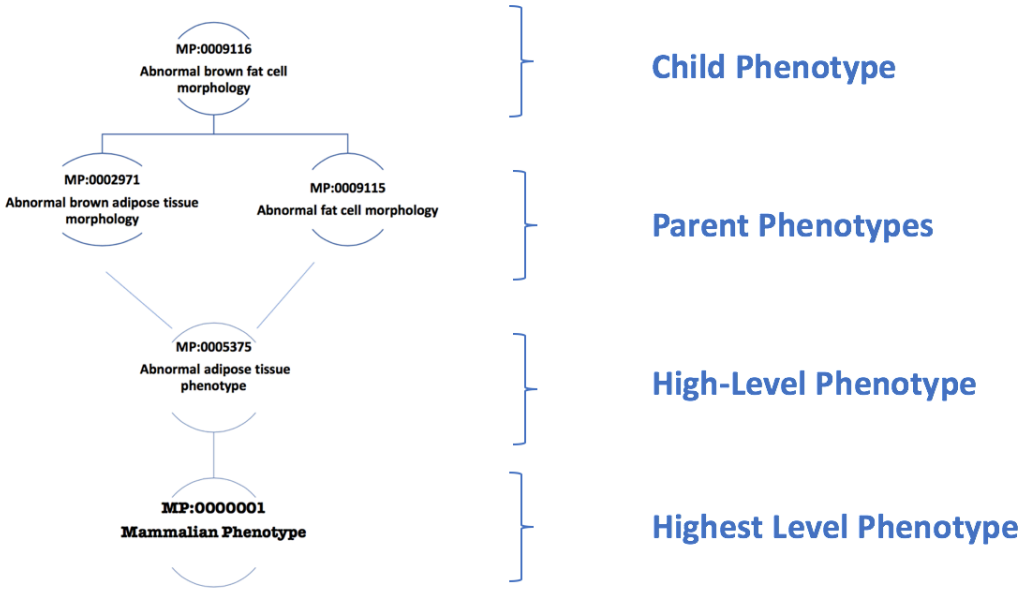


Figure 1: Phenotypic relationships of "abnormal brown fat cell morphology" in MPO

2.5 DISEASOME

Diseasome is a collection of networks that relates human diseases with the disease causing human genes (Goh *et al.*, 2007). It is proposed as a network based approach to study the relations between human genetic disorders and the genes. The Online Mendelian Inheritance in Man (OMIM) is used as the data source for disease-gene relations in *Diseasome*. The *Diseasome* mapping consists of multiple networks namely: the human disease network (HDN), the disease genes network (DGN) and the bi-partite human disease and gene network. More details about these two networks are given under section 2.5.1. In Figure 2, *Diseasome* design is illustrated. In their study, Goh *et al.* stated that disorders can be associated with each other using the shared

disease-causing genes. The main list of *Diseasome* contained 1,284 disorders and 1,777 disease genes and all diseases are categorized based on 22 distinct disease classes.

Diseasome particularly focuses on the molecular relationships between genetic variation and phenotypic information, and it is a seminal work in terms of discovering the mechanisms of complex diseases. It is important here to note that, revealing complex disease mechanisms is one of the most crucial problems in biomedical research, currently (Botstein and Risch, 2003, Kann, 2009). It had already been stated in the literature that many human diseases occur due to the factors related to genetic variations (Hirschhorn and Daly, 2005). Up to date, various databases are constructed for annotating the relations between genes and diseases of human such as OMIM (Hamosh *et al.*, 2005), CTDTM (Davis *et al.*, 2010) and NHGRI-EBI GWAS catalog (Welter *et al.*, 2013). Due to the nature of database curation process the associations are not complete, so the integration of multiple existing resources usually leads to more comprehensive view of the current biomedical knowledge. DisGeNET is one of these platforms and constructed for the integration of gene and disease information and associations from various resources (Piñero *et al.*, 2015). The source of disease-gene relation information is obtained from the OMIM database. The Online Mendelian Inheritance in Man (OMIM) was constructed by Dr. Victor A. McKusick in early 1960's to catalogue genetic diseases/traits and the corresponding disease causing genes (Hamosh *et al.*, 2005).

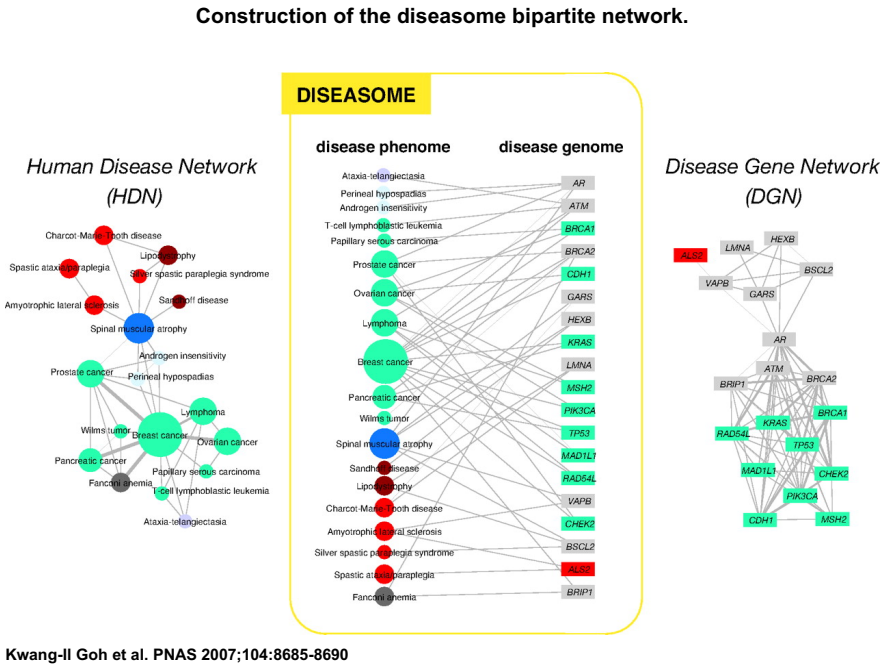


Figure 2: Illustration of *Diseasome* networks, Re-printed from: Physical Sciences - Applied Physical Sciences: Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert László Barabási, the human disease network PNAS 2007 104(21) 8685-8690; published ahead of print May 14,2007, doi:19.1073/pnas.0701361104).

2.5.1 HUMAN DISEASE NETWORK AND DISEASE GENE NETWORK

Human Disease Network (HDN) shows the relations between human disorders. A representative sub-network of HDN is shown on the left side of Figure 2. Every node in HDN shows a distinct disorder and two disorders have a link if they share at least one gene in common. Disorder classes inform the user regarding which physiological system is affected by that disease. Classifications were made for twenty primary disorder classes but additionally two categories were preferred to be added as “multiple” and “unclassified”. If the primary classification does not seem clear and this disorder belongs to more than one classes, then it was put into the multiple class.

If there is no sufficient and obvious information for classification, then the disorder put into the unclassified class. At the visualization level, 22 different disorder classes are differentially colored to investigate if the diseases belong to the same system share their genes as well. The edge between the diseases from the same disorder classes is colored according to the color of this class; otherwise they are shown in gray. The size of every disease node depends on the number of genes associated with that disorder. Also, the edge thickness between two disorders is proportional to the number of shared genes. Name of disorder is shown on the network if it has ten or more genes associated with it for practical reasons. 867 of 1,284 disorders have at least one link to other and 516 disorders constitute a giant component. This result suggests that the origins of most of the hereditary/genetic diseases are shared.

Disease Gene Network (DGN) displays the associations between genes according to their shared diseases. In DGN, each node represents a distinct gene. Two distinct genes are connected to each other if they are both associated with the same disorder. Therefore, the link thickness is proportional to the number of disorders commonly shared by two distinct genes. The size of each node is proportional to the number of diseases it is related to. Nodes are coloured as gray if they play a role in more than one disorder, otherwise they are colored according to the disorder class of related disease. The name of gene is indicated only if it is associated with more than five disorders, for practical reasons. It can be said that the link between two genes may indicate the phenotypic associations, protein-protein interactions (PPIs) (Rodriguez-Caso, Medina, and Sole, 2005) and the discovery of novel genetic interactions. 1,377 out of 1,777 of disease genes are connected to at least one other gene and 90 of them compose a giant component.

2.5.2 INVESTIGATION OF THE DISEASOME NETWORK

Morbid Map (MM) of the OMIM is one of the most comprehensive and highly curated disorder gene association database. The OMIM MM shows the cytogenetic map location of disease genes in OMIM. The data in *Diseasome* were downloaded from the 2005 version of MM and contains 2,929 entries of 4,043 with the “(3)” tag, which

shows at least one mutation exists in that gene causing the disorder. After this pre-processing, the authors have parsed 2,929 entries into 1,284 distinct disorders by gathering all same subtypes of the same diseases under one entry. For example, 11 distinct groups of Fanconi Anemia were merged. Each distinct disease was assigned with unique disease ID. Similarly, each gene was indicated with its distinct ENTREZ ID, which is a specific indicator of it for the organism of interest. Entrez Gene is a gene bank and maintained in the National Center for Biotechnology Information (NCBI) (Maglott *et al.*, 2010).

In the *Diseasome* mapping, circle shaped nodes represent diseases and rectangle shaped nodes show disease genes. There exists a link between two disorders if a mutation in the commonly shared gene lead to these disorders. Colors are attained according to these disease classes. Size of giant component of a randomized network was computed both for the HDN and the DGN. It had been shown that the giant component sizes of the randomized networks are larger than the actual ones. This result gave the indication that there is a pathophysiological clustering between the disorders and the disease genes. The researchers stated that actual disorders and diseases genes show tendency to link with the same classes (Goh *et al.*, 2007).

A specifically described term in the *Diseasome* study is the "locus heterogeneity", which was employed to reveal the hub diseases clusters in *Diseasome*. Locus heterogeneity term is specified according to the mutations in more than one genes which cause similar disorders. It has been found that cancer and neurological disorders show high locus heterogeneity and they are the most connected nodes. On the other hand, metabolic, skeletal, and multiple disorder classes are the less connected ones and shows low genetic heterogeneity (Goh *et al.*, 2007).

It was seen from the results of Disesome that several disorders arise from mutations in few genes. Therefore, it was thought that corresponding protein product of these genes tend to participate in the same cellular pathways, molecular complexes or functional modules. Disease genes associated with the same disorders share common cellular and functional characteristics in terms of their annotated "Gene Ontology (GO) Terms". GO is a controlled vocabulary ontology system to describe gene/protein functions and

it is highly used in functional genomics studies. GO is composed of 3 main categories: Molecular Function (MF), which shows the molecular activities of gene products, Cellular Component (CC), which shows location of activity for the gene products, and the Biological Process (BP), indicating the involvement of gene products in the systemic processes such as the metabolic pathways.

Finally, an investigation in *Diseasome* study worth mentioning is the prediction of the essential gene information for the human. If a targeted knock-out mouse gene ends with lethality at the end of the experiment, then the researchers called human orthologue of that mouse gene as an essential gene. They obtained human related data from MGI in 2006 (MGI-Mouse genome informatics-the international database resource for the laboratory mouse, 2014). Embryonic/prenatal lethality and postnatal lethality classes are considered as lethal and the rest as marked as non-lethal. 398 of 1,267 mouse lethal human orthologue genes were found to have known human disease associations (Goh *et al.*, 2007), which shows 22% of them are already known human disease genes. This result leads to a separation in two classes of human disease related genes: 1,267 essential disease genes and 1,379 nonessential disease genes.

2.6 THE HUMAN PHENOTYPE ONTOLOGY PROJECT (HPO)

HPO provides a controlled vocabulary set to define phenotypic traits in human diseases. These phenotype terms mostly cover symptoms and they are associated with human disease records by manual curation. Köhler *et al.*, reported in 2014 that the system contains 10,088 classes (terms) describing human phenotypic abnormalities. HPO also provides phenotype-gene relations using OMIM disease-gene associations. Combination of phenotype and genomic data serves the identification of complications of disease subtypes (Köhler *et al.*, 2014). The HPO project (Robinson *et al.*, 2008) has started in 2007 and it has enhanced the coverage, usage, complexity and cross connection with other projects, particularly from the OBO Foundry (Smith *et al.*, 2007). HPO covers a wide range of phenotypic abnormalities seen in human diseases. Each class is named starting with “HP” letters with a unique and a stable number. On

average, each disease entry has 15 HPO annotations and the mapping is enriched at every database release. There are various biomedical projects that link to HPO.

DECIPHER project interconnects with HPO and its aim is to find clusters of rare diseases that have phenotypes and structural rearrangement with strong correlation (Firth *et al.*, 2009). The Biomedical Research Centers/Units Inherited Diseases Genetic Evaluation consortium uses the HPO database for saving the phenotypes of patients with rare inherited disorders.

Another crosslinking project is European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations (ECARUCA), which is established in 2003 and collecting and providing clinical and molecular information related to rare unbalanced chromosome abnormalities (Feenstra *et al.*, 2006). Currently this database includes information for more than 4800 cases that are crosslinked to HPO (Vulto-van Silfhout *et al.*, 2013).

Yet another one is Nijmegen Genetics Phenotype Database (NGPD), aiming to use and collect phenotypic information of patients with unexplained intellectual disability and/or congenital anomalies using the HPO. The NGPD currently includes more than 8000 patients with 73,496 HPO associations (Moss *et al.*, 2014).

2.7 NOVEL DISEASE - GENE IDENTIFICATION USING PHENOTYPE DATA

There are studies in the literature aiming to discover novel disease-gene associations using phenotype data. A study was conducted in 2012 by Chen *et al.* to find the candidate disease genes by using mouse phenotypes. The authors developed a web application to compare the mouse organism with human. Data collection comprises the most comprehensive part of this study. Human Phenotype ontology (HPO) annotations of OMIM diseases, and the HPO itself, and MPO annotations of mouse models, MGI asserted disease models and OMIM human gene to MGI gene mappings were downloaded. At the end, a database was created that consists of HPO annotations

for almost all clinical OMIM entries for a large part of HPO and MPO terms. Investigation was extended that covers nearly all known Mendelian diseases and a new software called OWLSim was constructed. The database was started to be constructed in 2011 and contains mappings from HPO, MPO and OMIM databases. 5,035 OMIM diseases (1858 with known gene associations and 3,177 with no known gene) and 1,791 OMIM genes with HPO annotations, along with the MPO annotations of 24,904 mouse models and 8,124 mouse genes are stored in the database. Additionally, 2,624 associations between OMIM diseases and models from MGI of the literature are also published (Chen *et al.*, 2012).

The main reason of using OWLSim software is to compare each HPO related OMIM genes or diseases with all MPO related mouse genes or mutant lines. It uses merged OWL file of PATO, UBERON, MPO plus logical definitions, HPO plus logical definitions and a mapping of HPO and MPO lexical matches for pairwise comparisons. OWL is the acronym for Web Ontology Language and a standard produced by the W3C. GO terms in OWL are based on a translation from OBO to OWL. Uberon is an integrated cross-species ontology that covers anatomical structures in animals. PATO can be used along with other ontologies such as GO or anatomical ontologies.

Another resource called PhenomeNET was conducted in 2011 by Hoehndorf *et al.*, with the same annotations, ontologies and definitions used for comparing human and mouse phenotypes; however, this algorithm differs from OWLSim (Hoehndorf, Schofield and Gkoutos, 2011) in methodological manners. While calculating the least common ancestor, PhenomeNET uses the idea of subsuming between classes, while OWLSim prefers to use other ontology relations. PhenomeNET calculates the average of all pairs of phenotypes, however, OWLSim uses the average of best matches.

MouseFinder is a web tool, which provides users with the opportunity to investigate mouse phenotypes and their comparison to disease records (Chen *et al.*, 2012). Users can search for various types of features by entering OMIM disease, gene names or HPO terms. Also, MGI asserted mouse models can become visible if it is provided. Another aim of this web tool is to discover the novel genes for OMIM diseases with

unknown gene. 468 OMIM diseases were taken with a mapped locus with no known genes.

In 2007, a study authored by Chen *et al.* improved the novel gene prioritization by using mouse phenotype information. It was shown that genes that because diseases have functional relationships. ToppGene database was created for gene prioritization and claimed to have higher performance compared to resources such as SUSPECTS and ENDEVAOUR (Chen *et al.*, 2007). Since most of the diseases are genetically polygenic, intricate, multifactorial and present different clinical phenotypes, it is hard to identify the disease-causing genes. Therefore, a different approach was applied with the use of integrative genomics-transcriptomics-phenomics-bibliomics sources. These sources were compounded with human gene annotations, mouse phenotype data and literature co-citations of genes.

In the same study, ToppGene was compared to the other gene prioritization methods: SUSPECTS and ENDEVAOUR. SUSPECTS is a tool that matches within GO terms, InterPro domains and gene expression data built on top of the PROSPECTR. PROSPECTR uses sequence features to rank genes (Adie *et al.*, 2005). The user interface was written in JAVA script, JSP and servlets, and integrated with the Tomcat web server. GO, pathways, phenotype, protein domains, PubMed and protein interaction terms are displayed (Chen *et al.*, 2007). While comparing it with SUSPECTS and ENDEVAOUR it was observed that percentage of top 10% and 5% ranked target genes results were higher in ToppGene.

ToppGene Suite is a portal for gene enrichment and novel gene prioritization based on functional annotations and protein interactions. Moreover, literature identifiers were used such as PubMed, PMIDs. As an example, for simple interpretation, if two genes have the same cross-reference in PMID result, it means that they have either direct or indirect biological association.

CHAPTER 3

MATERIALS AND METHODS

3.1 MATERIALS

This section includes the processing steps of website and the required inputs for design and analysis. The inputs' preparation and related soft wares are illustrated in detail.

3.1.1 GEPHI FOR NETWORK ANALYSIS & VISUALIZATION

Network visualization of large graphs has been a challenging subject for various years but it also is crucial to examine and understand the biological mechanisms (Bastian, Heymann and Jacomy, 2009). Gephi is an open free source software written in Java on the NetBeans platform for analyzing and visualizing networks and graphs. It is claimed that Gephi can handle large and complex data and both dynamic and static networks can be displayed and manipulated with Gephi tool (Bastian, Heymann, and Jacomy, 2009). It is freely available for academic purposes under the public license agreements (gephi.org). Gephi provides a visual platform, which bridges the complex biological data and mechanisms onto a tangible virtual environment. Gephi has various modules for importing, visualizing, filtering all types of networks. Multiple networks can run at the same time in separate workspaces.

Any algorithm, tool or filter can easily be added to Gephi with moderate programming skills. Nodes and edges output files can be exported manually or using filtering system. It provides various network analysis tools and their results also can be exported in

various formats. Also, with the help of various plugins in it, both static and dynamic results can be gathered. It provides user to manipulate and anticipate the data during handling with the network (Bastian, Heymann and Jacomy, 2009).

There are various visualization tools as alternatives to Gephi, which can be listed as yEd Graph Editor, Graphviz, Cytoscape, and Neo4j. The reason behind using Gephi arised from the fact that it was used in the *Diseasome* project. yEd is more suitable in diagramming rather than network analysis. The Graphviz takes descriptions of graphs in a text language and can create diagrams in several formats. Cytoscape is an open source platform for visualizing complex networks. Neo4j is an online platform for graph visualization and for the generation of graph based databases.

It is possible to perform various types of analysis with Gephi and results can be exported in different formats. Here are the important graph properties that Gephi calculates: Connected components, modularity, node degree, graph diameter, centrality, graph density, average path length and clustering coefficient. These network statistics can be computed under statistics part belongs to the "Overview" menu. Users can found filter options and node/edge overviews under this menu. Possible formats to export a network can be given as:

- A CSV is a comma separated values file and it allows data to be saved in a table structured format.
- A GML, Geography Markup Language (GML) is the XML grammar defined for expressing geographical features. Image exporters makes user to export view of a graph to .png, w.svg of .pdf formats
- Portable Network Graphics is a raster graphics file format that supports lossless data compression.
- Scalable Vector Graphics (. svg) is an XML-based format that can be edited using either text editors or image editing software. SVG can also be used for the Web, as it looks well when zooming or panning a visualization.

- The Portable Document Format (PDF) is the output can be written in terms of the wide-spread cross-platform document format. Even though this is the most trivial way among other exporting variations, any possibility for interactive alterations is not fully accessible.
- Seadragon exporter is suitable for the dynamic networks.
- Sigma.Js creates web based network graphs using a template driven approach.
- Loxa web site export also uses a sigma. Js and it provides user an interactive filtering and zooming.
- Terminally, an HTML/JS project which is gexf. Js master makes user to drag and drop a GEXF file to create a web export.

3.2 METHODS

This section explains the data preparation, aggregation and integration work. Also, processing steps of Gephi tool, details of python scripts in pseudo format are presented. A short website tutorial is described under this part (for more details see Appendix A). Data processing steps are summarized in Figure 3 and detailed information is provided in the following sub-sections.

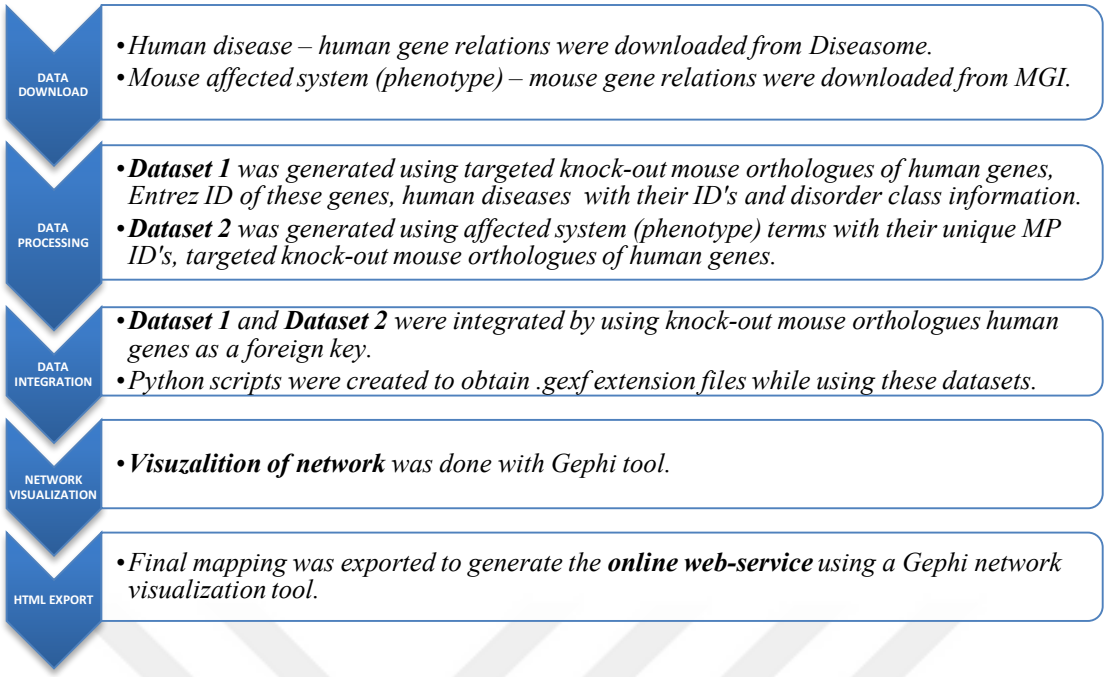


Figure 3: Working diagram of the study.

3.2.1 DATA DOWNLOAD AND PROCESSING

Under this part, datasets are illustrated according to their purpose of usage, content and modifications that were made on them. Two different datasets about human and mouse organisms were extracted. Dataset 1 contains the Human disease – human gene relation information and downloaded from *Diseasome* resource and the Dataset 2 contains Mouse affected system (phenotype) – mouse gene information derived from MGI. Mouse genes attribute was chosen as a foreign key, to relate these two sets. A more detailed information about the datasets is provided under sections 3.2.1.1 and 3.2.1.2 respectively.

3.2.1.1 DATASET DOWNLOAD FROM *DISEASOME* & DATA PROCESSING

Diseasome dataset shown in Table 1 was used as the source to constitute Dataset 1. It includes disease ID, disease name, disorder class, size (s) that show the number of associated genes, degree (k) shows number of disorder classes it connects to, class degree (K) is the number of distinct disorder classes it connects to and genes written as comma delimited at the last column.

Table 1: The *Diseasome* dataset

Supporting Information Table 2. **Network characteristics of diseases.**

Disease ID	Name	Disorder class	Size (s)	Degree (k)	Class-degree (κ)	Genes implicated (Entrez ID) [comma-deliminated]
1	17,20-lyase_deficiency	Endocrine	1	0	0	CYP17A1 (1586)
3	2-methyl-3-hydroxybutyryl-CoA_dehydrogenase_deficiency	Metabolic	1	0	0	HADH2 (3028)
4	2-methylbutyrylglycinuria	Metabolic	1	0	0	ACADSB (36)
5	3-beta-hydroxysteroid_dehydrogenase,_type_II,_deficiency	Metabolic	1	0	0	HSD3B2 (3284)
6	3-hydroxyacyl-CoA_dehydrogenase_deficiency	Metabolic	1	0	0	HADHSC (3033)
7	3-Methylcrotonyl-CoA_carboxylase_deficiency	Metabolic	2	0	0	MCCC1 (56922), MCCC2 (64087)
8	3-methylglutaconic_aciduria	Metabolic	1	0	0	AUH (549)
9	3-methylglutaconicaciduria	Metabolic	1	1	1	OPA3 (80207)
10	3-M_syndrome	multiple	1	0	0	CUL7 (9820)

Information about datasets are available under the supported information (SI) part. Curated Morbid Map file with disease ID, class assignment (SI Table1), network characteristic of diseases (SI Dataset 2) from and disease genes (SI Table 3) were examined and combined in Dataset 1. Mouse orthologues of human genes were converted and extracted with the online converter tool called as HCOP: Orthologue Predictions Search (European Bioinformatics Institute, HCOP: Orthologue Predictions Search. Retrieved [04.07.2016] from [http://www.genenames.org/cgi-bin/hcop]).

SI Table1 contains the Disease ID, Disorder name, Human Gene Symbols, OMIM ID, Chromosome Position of the related gene and Disorder Class information. Disorder names were aligned in an alphabetical order and distinct consecutive numbers are given in ascending order starting from 1. These numbers are called as Disease ID and assigned for analysis in Gephi. Disorder names are distinctly ordered with their related

human genes and in accordance OMIM Ids are retrieved. If a disorder has more than one genes related to it, these genes are separated with comma.

SI Table 2 covers the information based on disease network statistics. Columns are separated as Disease ID, Disorder name, class Size(s), Degree(k), Class-degree(κ), Genes implicated (Entrez ID) as comma delimited. Size(s) is the number of genes associated with that disorder, degree(k) is the total number of connectivity to disorder classes and class-degree(κ) is the number of distinct disorder classes.

SI Table 3 was constructed according to the disease gene information. This table contains Entrez ID, Symbol, Disorder class, Size (s), Degree (k), Number of classes associated, Implicated diseases (Disease ID) as comma delimited. Size(s) is the number of diseases associated with that gene, degree(k) is the total number of genes belonging to disorder(s) interact with this gene expect itself.

SI Table2 was used as a reference source to compose Dataset 1. As the final step, Dataset 1 was linked together with the dataset obtained from MGI, which is explained in the following section.

Dataset 1 shown in

Table 2 consists of targeted knock-out mouse orthologues of human genes, Entrez ID of these targeted knock-out mouse genes, human disease ID, human disease and disorder class information. The remaining information except mouse orthologues of human genes and their IDs are the same with the *Diseasome* dataset information. Human gene column was added to ease the understanding for orthologue idea between human/mouse organisms. This dataset is based on human data.

Table 2: Dataset 1

ENTREZ ID	MOUSE GENE	HUMAN GENE	DISEASE ID	HUMAN DISEASE	DISORDER CLASS
13074	Cyp17a1	CYP17A1	1	17 20 lyase deficiency	Endocrine
66515	Cul7	CUL7	6	3 M syndrome	multiple
403187	Opa3	OPA3	10	3 methylglutaconicaciduria	Metabolic
22017	Tpmt	TPMT	12	6 mercaptopurine sensitivity	Metabolic
13618	Ednrb	EDNRB	15	ABCD syndrome	multiple
238055	Apob	APOB	17	Abetalipoproteinemia	Metabolic
20682	Sox9	SOX9	18	Acampomelic campolelic dysplasia	Skeletal
107146	Cat	CAT	21	Acatalasemia	Hematological
17246	Mdm2	MDM2	22	Accelerated tumor formation	Cancer
223921	Aaas	AAAS	24	Achalasia addisonianism alacrimia syndrome	multiple
56873	Lmbr1	LMBR1	25	Acheiropody	Skeletal
12824	Col2a1	COL2A1	26	Achondrogenesis hypochondrogenesis type II	Bone
13521	Slc26a2	SLC26A2	27	Achondrogenesis Ib	Bone
14184	Fgfr3	FGFR3	28	Achondroplasia	Skeletal
30952	Cngb3	CNGB3	29	Achromatopsia	Ophthamological
12790	Cnga3	CNGA3	29	Achromatopsia	Ophthamological
14686	Gnat2	GNAT2	29	Achromatopsia	Ophthamological
16005	Igfals	IGFALS	30	Acid labile subunit deficiency of	Endocrine
192775	Kcnh2	KCNH2	31	Acquired long QT syndrome	Cardiovascular

3.2.1.2 DATASET DOWNLOAD FROM MGI & DATA PROCESSING

Mouse affected systems information (i.e. phenotypes) was collected from the MGI database. Collected mouse orthologue genes with HCOP were imported to the MGI batch summary tool for creating Dataset 2. Only the targeted null/knock-out mouse genes were taken into consideration during the generation of Dataset 2. MGI data shown in Table 3 were used as the source to constitute Dataset 2. It includes affected system information with unique “Mammalian phenotype ID” of all recorded mouse genes with marker symbols in that database. It also provides unique MGI IDs for these genes, allele type and allele attribute information.

Table 3: The MGI Dataset

MGI Allele Accession ID	Allele Type	Allele Attribute	Marker Symbol	Mammalian Phenotype ID (comma-delimited)
MGI:2180117	Targeted	Null/knockout	A2m	MP:0005379,MP:0005370,MP:0005370,MP:0005370,MP:0005370
MGI:5445373	Targeted	Null/knockout	A4galt	MP:0005387
MGI:3624807	Targeted	Null/knockout	A4galt	MP:0010768,MP:0005376
MGI:3621812	Targeted	Null/knockout	Aaas	MP:0005389,MP:0003631,MP:0005379
MGI:3577725	Targeted	Null/knockout	Abca1	MP:0005376
MGI:3525100	Targeted	Null/knockout	Abca1	MP:0005397,MP:0005376
MGI:1935192	Targeted	Null/knockout	Abca1	MP:0005388,MP:0003631

MPO website was used as a main source for the phenotypic information in both OBO and OWL format. In this resource, gene list either can be pasted directly or imported as a file into “ID/Symbols List” part in batch summary tool. “Mammalian Phenotype (MP)” option under additional information part was selected and search was initiated. From the resulting list, related data was imported with human readable file formats, such as .xlsx, .csv or .txt.

A Microsoft Office Excel tool function Vlookup was used for gathering the related affected systems of mouse genes in Dataset 1. Vlookup function finds common parts in a table or in each range, with respect to rows.

Except from the affected systems that are directly taken from MPO, four different versions of Dataset 2 were created by taking the different levels of phenotypes. Purpose of this idea is to observe the change in the network size. Detailed information about this method can be found under “3.1.2.3 Phenotype Levels” section.

Furthermore, it was noticed that some of these mouse orthologs of human genes were not annotated with any phenotype in the MGI database. According to MGI batch summary results, it was found that 1,375 of these genes have mammalian phenotype id and 170 of them do not have any recorded information. Detailed list is provided under Table 5.

Dataset 2 is shown in Table 4, and it consists of phenotype terms with their MP ID’s and targeted knock-out mouse orthologues of human genes. Human gene column again was added for the ease of understanding. This dataset is based on mouse data.

Table 4: List of mouse gene symbols, which do not have any MPO annotation

Aass	Baat	Eno3	Golga5	Kir3dl2	Ndufs8	Phkb	Serpina3b	Tpm2
Abat	Bckdha	Etfa	Gypc	Krt13	Ndufv1	Pkp1	Serpina3f	Trappc2
Acadsb	Bckdhb	Etfb	Hbb-bs	Lct	Oas1c	Pla2g2a	Serpina3i	Tspan7
Acsl6	Bpgm	Etfdh	Hbb-bt	Lrrc8a	Oas1e	Plekhg4	Serpina3j	Tspyl1
Adamts10	Btnl2	Fance	Hlcs	Maml2	Oas1f	Plod2	Serpina3k	Umps
Aggf1	C1s1	Fancf	Hmcn1	Mccc1	Oas1h	Pqbp1	Serpina3m	Upb1
Alad	C1s2	Fgb	Hmgcs2	Mccc2	Opcml	Prcc	Serpina7	Usp26
Aldh6a1	Cog7	Fgd1	Hnmt	Mcfd2	Pabpn1	Pus1	Slc22a18	Vps13b
Aldoa	Col9a3	Fmo3	Hs1bp3	Mllt10	Pdgfrl	Pygl	Slc25a15	Whsc1l1
Alg1	Creld1	Ftcd	Hsd17b3	Mllt11	Pdhb	Rap1gds1	Slc25a22	Xpnpep2
Alg12	Crybb1	Ftl1	Hsd3b1	Mmaa	Pdhx	Rfxank	Slc5a5	
Alg3	Ctdp1	Fuca1	Hsd3b2	Mmab	Pdlim4	Rfxap	Sncaip	
Alg6	Cyb5r3	Gale	Hsd3b3	Mmp1b	Pex12	Rnf139	Spg21	
Alg8	Cyp2r1	Gch1	Hsd3b4	Mpdu1	Pex19	Rnf6	Stox1	
Alg9	Cyp4v3	Gcsh	Hsd3b5	Mvk	Pex6	Rp9	Suox	
Arhgef10	Dguok	Gm10681	Hsd3b6	Myh2	Pgam2	Scgb3a2	Tas2r138	
Arl11	Dpys	Gm4450	Iqcb1	Myl3	Pgk1	Serpina1c	Tcn2	
Aspscr1	Dsg1a	Gm6904	Jrk	Ncf2	Phf11b	Serpina1d	Timm8a1	
Atic	Eif2b1	Gmps	Kif21a	Ndufs2	Phf11c	Serpina1e	Tnni2	
B4galt7	Eif2b4	Gns	Kir3dl1	Ndufs7	Phka2	Serpina3a	Top2a	

Table 5: Dataset 2.

MP ID	AFFECTED SYSTEM (PHENOTYPE)	MOUSE GENE	HUMAN GENE
MP:0005370	liver/biliary system phenotype	A2m	A2M
MP:0002006	neoplasm	A2m	A2M
MP:0002169	no abnormal phenotype detected	A2m	A2M
MP:0001869	pancreas inflammation	A2m	A2M
MP:0005388	respiratory system phenotype	A2m	A2M
MP:0008874	decreased physiological sensitivity to xenobiotic	A4galt	A4GALT
MP:0009767	decreased sensitivity to xenobiotic induced morbidity/mortality	A4galt	A4GALT
MP:0009747	impaired behavioral response to xenobiotic	A4galt	A4GALT
MP:0001516	abnormal motor coordination/ balance	Aaas	AAAS
MP:0005384	cellular phenotype	Aaas	AAAS
MP:0001262	decreased body weight	Aaas	AAAS
MP:0001417	decreased exploration in new environment	Aaas	AAAS
MP:0005381	digestive/alimentary phenotype	Aaas	AAAS
MP:0005379	endocrine/exocrine gland phenotype	Aaas	AAAS
MP:0001926	female infertility	Aaas	AAAS
MP:0005376	homeostasis/metabolism phenotype	Aaas	AAAS
MP:0001402	hypoactivity	Aaas	AAAS
MP:0003631	nervous system phenotype	Aaas	AAAS
MP:0011729	abnormal pineal gland melatonin secretion	Aanat	AANAT
MP:0011728	abnormal pineal gland physiology	Aanat	AANAT

3.2.2 HIERARCHICAL APPROACH TO PHENOTYPES

This section explains the different approaches used to generate the dataset 2. Batch summary result for affected systems of related genes in MGI contains concurrently the direct results of experiments. In other words, phenotypes provided by MGI tool are the directly recorded phenotypes (i.e. only the most specific phenotype terms in the MPO

DAG). We propagated the phenotypic term annotations through the root of MPO (i.e. mammalian phenotype term) and applied at cut-off only to provide the annotations at that certain level. Since the specific terms merge under the same parent terms at each level, the number of terms decrease going from specific to generic. This way, the total number of phenotype nodes decreases when we use higher levels of MPO instead of the most specific ones.

Figure 4 illustrates the MPO relations and how it can be possible to reach the Mammalian Phenotype (i.e. the root term) in varying number of steps, according to the actual level of the most specific annotated phenotype term.

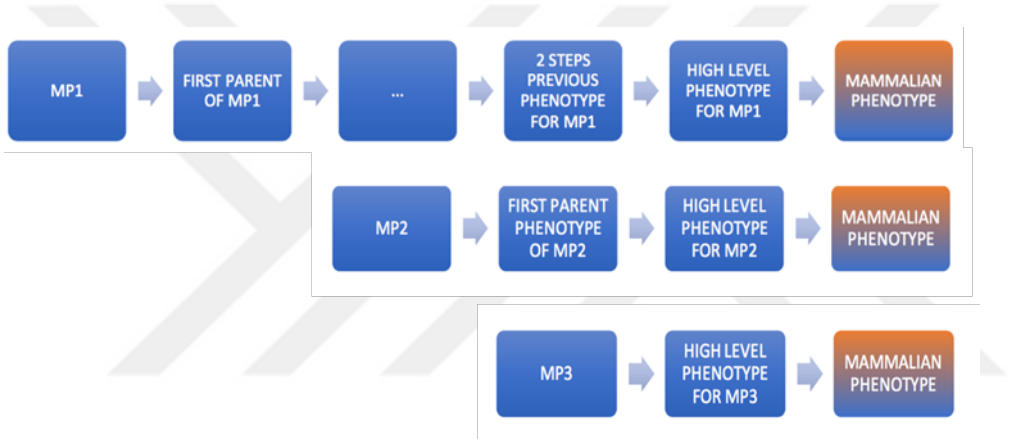


Figure 4: Three types of paths for affected systems to reach the root MP term.

In the toy example displayed in Figure 4, the annotated terms can connect the root of the MP tree in varying steps, according to their specificity. In this sense, we divded MP terms in 3 groups. First group is represented by MP1 phenotype, which can reach the root in multiple steps. The number of steps change from term to term, as some terms are more specific compared to the others. In the second group, MP2 connects to the root in 3 steps. Lastly for the third group, MP3 reaches the root node at 2 steps. To reduce the number of phenotype nodes in the generated networks, we generated 4 levels of phenotype annotations: using i) most specific phenotypes (i.e. direct annotations to genes), ii) first/direct parent phenotypes, iii) 2-steps before the root

affected systems, and iv) high-level affected systems. By using the main source from the bio portal, desired levels of affected systems were gathered.

In Figure 5, statistics for the networks generated using different phenotype levels are displayed. First level covers child affected systems directly taken from the batch summary result in MGI. Number of nodes on this version is 8,355 where 1,116 of them are diseases and the number of distinct affected systems is 5,696 and the number of edges is 111,207. This network is called as "child affected systems version". Second version represents the one step higher level (i.e. direct parent) of the asserted affected systems. Number of nodes is 3,675 and for distinct affected systems it is 2,558, where 1,116 of them are diseases and number of edges is 89,603. This second network is called as "parent affected systems version". The third one is generated with the affected systems that stands for two steps before mammalian phenotype. This one is called "two step before root version". In this version, number of nodes is 1,248, number of edges is 26,009 and there exist 131 distinct affected systems in this network. The last network is formed according to high-level affected systems. Here the total number of nodes is 1,146. Number of edges is 13,661 and the number of phenotypes was just 30. This version is called as “high-level affected systems”. It was observed that generalizing the affected systems decreases both edge and node numbers.

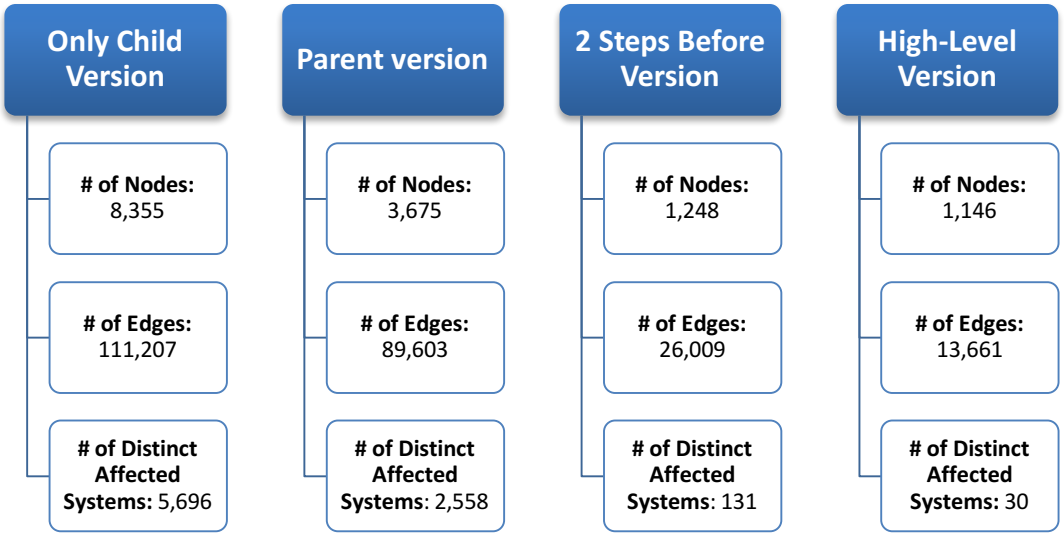


Figure 5: Node and Edge Statistics for Dataset 2 Versions.

Large datasets such as the network generated using the asserted phenotypes (i.e. child affected system version) can reveal diverse and extensive information; however, it has a drawback of very populated and dense visualization render, which is computationally intense. To avoid it, one can use the other versions with less nodal degrees of freedoms. In this sense, parent affected systems version did not provide a significant improvement as the number of edges is nearly the same as the child version. Moreover, both two steps previous version and the high-level version has very low number of distinct phenotypes, 131 and 30 respectively, to cause a loss in specificity. At the end, we decided to continue with the child affected version of the network as in the beginning.

3.2.3 INTEGRATION OF DATA & GENERATING THE NETWORKS

The data integration was based on connecting human diseases and mouse affected systems (i.e. phenotypes) by using mouse/human orthologous genes. Two strategies were followed to generate the networks: treating the genes i) as nodes, and ii) as edges. The idea behind this design is to generate a comprehensive network that display all

relations in-between genes-diseases-phenotypes. Human diseases are indirectly connected to the mouse phenotypes (i.e. affected systems) while using mouse/human orthologous genes as the mediator.

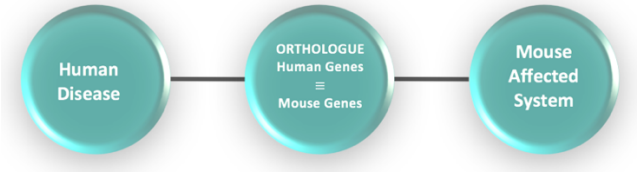


Figure 6: Genes-Node version of the Mouse2Human network.

In Figure 6, Genes-Node version design is illustrated and all terms are classified as nodes in that network.

Second version was constructed by treating mouse genes as edges. The idea behind this design is to decrease the number of nodes, to provide a less crowded network and visually perceivable network by only displaying relations between human diseases and mouse affected systems. Figure 7 displays the representation of Genes-edge network version, where the knock-out mouse genes / orthologues human genes were treated as edges.

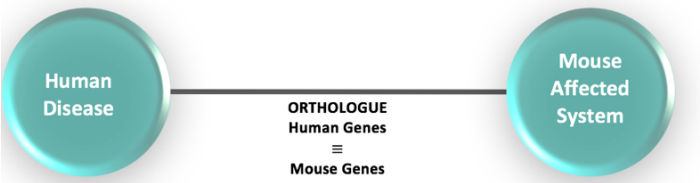


Figure 7: Genes-Edge version design of the Mouse2Human network

At this part of the study, Dataset 1 and Dataset 2 were merged by integrating the human and mouse data tables. A link was established between human and mouse data using the targeted knock-out mouse orthologues of human genes. If any gene is related to more than one disorder, the gene name is repeated multiple times in the merged dataset for each related disease.

Each human disease has a unique ID. Normally, OMIM IDs are used for diseases; however, it was indicated in the *Diseasome* study, the derivatives of the same diseases are gathered into a single category for the sake of simplicity. For example, all the derivatives of Alzheimer diseases were grouped into a one category and named as "Alzheimer disease". Therefore, another unique numbering system was developed by *Diseasome* and called as "Disease ID". In our study, *Diseasome* Disease IDs are used, as well. Disorder classes were attached to the diseases collaterally from the *Diseasome* dataset. Targeted knock-out mouse genes were indicated with their ENTREZ IDs. Knock-out mouse gene and human disease columns consist string values and rest of the table is composed of integers. Table 6 illustrates a portion of the combined dataset.

Table 6: The combined dataset

MP ID	AFFECTED SYSTEM (PHENOTYPE)	MOUSE GENE	HUMAN GENE	DISEASE ID	HUMAN DISEASE	DISORDER CLASS
MP:0005370	liver/biliary system phenotype	A2m	A2M	98	Alzheimer Disease	Neurological
MP:0002006	neoplasm	A2m	A2M	98	Alzheimer Disease	Neurological
MP:0002169	no abnormal phenotype detected	A2m	A2M	98	Alzheimer Disease	Neurological
MP:0001869	pancreas inflammation	A2m	A2M	98	Alzheimer Disease	Neurological
MP:0005388	respiratory system phenotype	A2m	A2M	98	Alzheimer Disease	Neurological
MP:0008874	decreased physiological sensitivity to xenobiotic	A4galt	A4GALT	212	Blood group	Hematological
MP:0009767	decreased sensitivity to xenobiotic induced morbidity/mortality	A4galt	A4GALT	212	Blood group	Hematological
MP:0009747	impaired behavioral response to xenobiotic	A4galt	A4GALT	212	Blood group	Hematological
MP:0001516	abnormal motor coordination/ balance	Aaas	AAAS	24	Achalasia addisonianism alacrimia syndrome	multiple
MP:0005384	cellular phenotype	Aaas	AAAS	24	Achalasia addisonianism alacrimia syndrome	multiple
MP:0001262	decreased body weight	Aaas	AAAS	24	Achalasia addisonianism alacrimia syndrome	multiple
MP:0001417	decreased exploration in new environment	Aaas	AAAS	24	Achalasia addisonianism alacrimia syndrome	multiple
MP:0005381	digestive/alimentary phenotype	Aaas	AAAS	24	Achalasia addisonianism alacrimia syndrome	multiple
MP:0005379	endocrine/exocrine gland phenotype	Aaas	AAAS	24	Achalasia addisonianism alacrimia syndrome	multiple
MP:0001926	female infertility	Aaas	AAAS	24	Achalasia addisonianism alacrimia syndrome	multiple
MP:0005376	homeostasis/metabolism phenotype	Aaas	AAAS	24	Achalasia addisonianism alacrimia syndrome	multiple
MP:0001402	hypoactivity	Aaas	AAAS	24	Achalasia addisonianism alacrimia syndrome	multiple
MP:0003631	nervous system phenotype	Aaas	AAAS	24	Achalasia addisonianism alacrimia syndrome	multiple
MP:0011729	abnormal pineal gland melatonin secretion	Aanat	AANAT	411	Delayed sleep phase syndrome	Psychiatric
MP:0011728	abnormal pineal gland physiology	Aanat	AANAT	411	Delayed sleep phase syndrome	Psychiatric

Python dictionary and list objects, sorting and string functions, parsing methods were used for the generation of the networks. The pseudo code of Genes-Node network and

the objects of the program can be found in Figure 8: Pseudo code for the network generation for the genes-node version (for more details see Appendix B.2).

```
"This part parses and allocates the Diseasome data"

function Open Workbook (Dataset1)
humandis2gene ← empty dictionary
disclass ← empty dictionary
gene2entrez ← empty dictionary
currentrow=1
currententrezid ← first row of Dataset1, keeps EntrezID information
currentgene ← second row of Dataset1, keeps gene information
currenthumandis ← third row of Dataset1,keeps disease information
currentdisclass ← fourth row of Dataset1, keeps disease class information
currenttuple ← currentgene,currentdisclass,currentrow, keeps whole information of Dataset1

"This part parses and allocates the MPO data"

function Open Workbook (Dataset2)
gene2affected ← empty dictionary
mp2id ← empty dictionary
currentrow=1
currentmp ← first row of Dataset2, keeps MP ID information
currentgene ← second row of Dataset2, keeps mouse gene information
currentaffected ← third row of Dataset2,keeps affected system information
affectedtuple ← currentmp,currentaffected

"This part attains source and target node information for
Gephi"
gene2geneid ← empty dictionary
currentrow=1
previousrow[max(mp2id.values()+1) ← taking maximum value prevents the duplicate Id numbers.
"Mp2id.values" here shows the current phenotype id.

"This part specifies the node size information"
"Disease Node Size Specification"
currentsize=0
for
human diseases in humandis2gene(keys)() ← if this disease exists in the
Dataset1, take it and attain a size to it
try
currentsize=currentsize+len(gene2affected[currentgene]) ← disease node
size is proportional to the number of genes it is related to

"Affected System Node Size"

import necessary modules
import counter
import chain
row to pairs ← attains the row that related to the affected system
sheet to pairs ← ranges the affected systems
count affected in sheet ← counts how many genes are related to them
currentcol ← give same colour to all affected systems
for
affected in gene2affected[gene] ← count the number of genes in
concerned affected system
    do
        currentsize=str(counter.get(currentaffected)) ← attain its final size

"This part is related to node colours"
currentcol=bgr[(len(disclasskeys()), disclass[currentdisclass]) ← give
colours to diseases according to their 22 distinct disease classes

label1 ← humandisease
attribute1 ← disease class
attribute2 ← number of knock out mouse genes

" This part is related to edge weights"
label1 ← mouse knock out gene
attribute1 ← ENTREZ ID
edgedic ← empty dictionary
edgecounter=1
for
human disease is in humandis2gene
affected in gene2affected[gene]
    do
        current gene target node id= gene2geneid[gene] ← it assigns
edge weight between gene and disease
for
affected is in gene2affected
    do
        currentmp,current affected=str(current mp),str(current affected)  ←
it assigns edge weight between gene and affected system
```

Figure 8: Pseudo code for the network generation for the genes-node version

The written scripts take these two datasets as input, process them and attain node and edge features. After executing the written scripts, output file was attained in. gexf format which is readable by the open viz platform, Gephi. Both versions of the networks were imported to Gephi as undirected.

Gephi provides a coloring option according to the node category. To differentiate different types of nodes in the network, mouse genes and affected systems are colorized as black and red, respectively in the Genes-Node version. Similarly, mouse affected systems were colored as red in the Genes-Edge version. Human diseases (i.e. disorders) were colored differently according to the disease classes. Node coloring in Gephi is set using the options under the overview menu, through node attributes and

type choices. The number of selected colors can be increased, as well as decreased using the palette widget option. Twenty-four and twenty-three distinct node coloring were generated respectively for Genes-Node and Genes-Edge versions.

Color code tables were constructed for both versions of network. Color code table for Genes-Node and Genes-Edge versions can be observed to gether with the number of nodes for each node type in Figure 9 and in Figure 10, respectively.

TYPES OF NODES	# OF NODES
Affected system	5696
Mouse knock-out Gene	1479
Disorder Class: Metabolic	172
Disorder Class: Multiple	140
Disorder Class: Cancer	120
Disorder Class: Neurological	119
Disorder Class: Hematological	87
Disorder Class: Immunological	65
Disorder Class: Skeletal	57
Disorder Class: Ophthamological	57
Disorder Class: Endocrine	54
Disorder Class: Dermatological	47
Disorder Class: Cardiovascular	44
Disorder Class: Developmental	34
Disorder Class: Renal	34
Disorder Class: Bone	30
Disorder Class: Muscular	26
Disorder Class: Unclassified	28
Disorder Class: Connective Tissue Disorder	21
Disorder Class: Gastrointestinal	18
Disorder Class: Psychiatric	16
Disorder Class: Respiratory	6
Disorder Class: Ear,Nose,Throat	5
Disorder Class: Nutritional	

Figure 9: Color code table with the number of nodes for Genes-Node version

TYPES OF NODES	# OF NODES
Affected system	5696
Disorder Class: Metabolic	172
Disorder Class: Multiple	140
Disorder Class: Cancer	120
Disorder Class: Neurological	119
Disorder Class: Hematological	87
Disorder Class: Immunological	65
Disorder Class: Skeletal	57
Disorder Class: Ophthamological	57
Disorder Class: Endocrine	54
Disorder Class: Dermatological	47
Disorder Class: Cardiovascular	44
Disorder Class: Developmental	34
Disorder Class: Renal	34
Disorder Class: Bone	30
Disorder Class: Muscular	28
Disorder Class: Unclassified	26
Disorder Class: Connective Tissue Disorder	21
Disorder Class: Gastrointestinal	18
Disorder Class: Psychiatric	16
Disorder Class: Respiratory	6
Disorder Class: Ear,Nose,Throat	5
Disorder Class: Nutritional	

Figure 10: Color code table with the number of nodes for Genes-Edge version

Different layouts can be applied to the networks in Gephi. In *Diseasome* network, "Force Atlas'' layout was employed. In this layout, repulsive forces between the distant nodes in the same cluster are approximated by a Barnes-Hut calculation and it stops after the range of convergence is achieved iteratively. Barnes-Hut calculation is an approximation algorithm to perform an n-body simulation.

It was also investigated to use the OpenOrd layout to emphasis divisions. This layout provides undirected weighted graphs and can divide clusters virtually in a tangible manner. It also stops automatically and this algorithm is also based on Fruchterman and Reingold and works with an upper limit for the number of iterations till convergence is achieved.

The name of the nodes and edges were automatically imported to the network by selecting show labels and show edges options in Gephi. The resulting can be exported in graph file format and it provides either a text or an xml of the trimmed gexf file. Also, it is possible to save the image of the network by selecting one from various other formats like pdf, jpeg or png.

Finally, a web exporter was used to provide the networks to the research community in a web-server, which was gexf Js master tool (Velt, 2011). This plug-in is used for undirected and static graphs and it provides a user firiendly interface (downloaded from: https://github.com/raphv/gexf-Js). Under the config.Js folder, output was be replaced by changing gexf extension name with the desired one. Output of network was kept under index.html part. The HTML file can be opened with any web browser and after a few seconds of loading time, the desired network becomes visible. It is possible to type node names in the search column, which provides a list of possible terms related to the searched word. One of the nodes can be selected from the list to display the related sub network (isolating the sub-network requires clicking on three dots sign at the lower left side of the screen). Connected nodes will become highlighted while approaching any node with the mouse cursor without clicking it.

CHAPTER 4

RESULTS

4.1 NETWORKS ANALYSES WITH GEPHI

Under statistics menu in Gephi, network analysis options are available for static and dynamic graphs for average degree, average weighted degree, network diameter, graph density, modularity and eigenvector centrality. We carried out the network analyses using Gephi's options and the results are given below.

Degree gives the number of edges linked to a node. Average degree is the information for all nodes in the network and it can be found by taking mean of all degrees. In Genes-Node and Genes-Edge versions respectively the average degrees are 12,725 and 26,620. In other words, it is the average number of links per node and naturally the value is larger in Genes-Edge version because making genes as nodes reduce the number of edges per node by the increasing the number of nodes. Degree distribution graphs are illustrated under Figure 11 and Figure 12 respectively for Genes-Node and Genes-Edge versions.

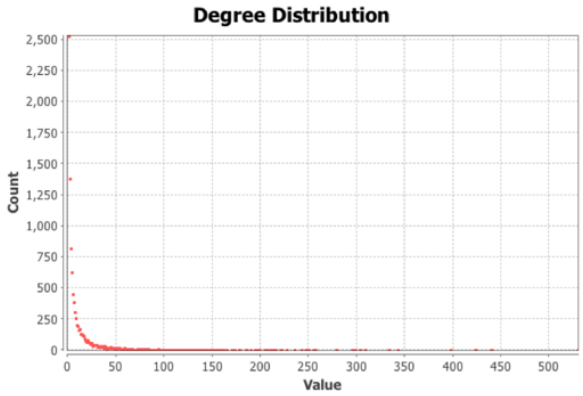


Figure 11: Degree distribution for Genes-Node version network

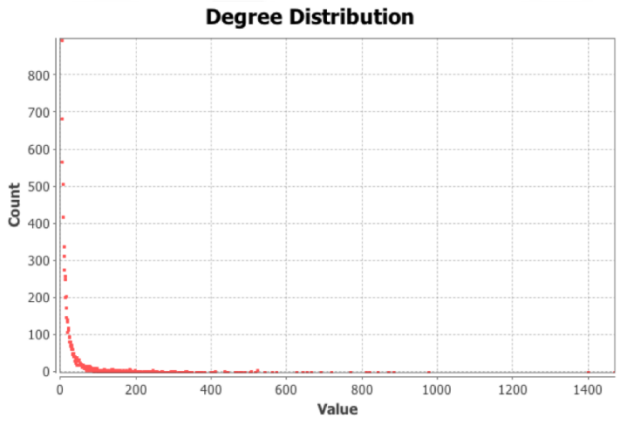


Figure 12: Degree distribution for Genes-Edge version network.

Average weighted degree results for Genes-Node and Genes-Edge versions are 26,203 and 30,095 respectively. It is expected to see a decrease in total node number from Genes-Node version towards to the Genes-Edge version. However, edge number has been increased up to 1.77 times in Genes-Edge version.

Network diameter can be defined as the maximum eccentricity of any vertex in the graph, in other words, it is the longest one of all shortest paths between any pair of vertices. Under the "Network diameter" section, betweennes centrality, closeness centrality and eccentricity analysis has also been performed. In Genes-Node version, diameter has been found as 8. In Genes-Node version, the shortest path was found

between any pair of vertices shows the diameter of a graph. Radius was calculated as 1 and average path length is 3,84. Number of shortest paths were calculated as 97,130,898 totally. In Genes-Edge version, diameter has been found as 7 and it is expected to be seen smaller in this version while comparing with the Genes-Node version. Radius again calculated as 1 and average path length is 3,45. Because genes were no more treated as nodes, it is expected to see the decline in this number also. Number of shortest paths had been found as 70,216,028 in total.

Another measure that we calculated was the Graph density analysis. Dense graph is where the number of edges is close to the maximum number of edges. The opposite term is a sparse graph that is a graph with only a few edges. Graph density had been found as 0,001 in Genes-Node version and 0,003 in Genes-Edge version. This value increase with the number of edges in the same direction. In this way, Genes-Node version network can be considered as a relatively sparse graph while comparing with the Genes-Edge version.

Modularity shows how well a network decomposes into its modular communities. It is directly proportional to the departmentalize issue in the network. Gephi looks for the nodes that are more densely connected in the network (Blondel *et al.*, 2008). A high modularity score indicates complicated internal construction. In other words, community structure shows how network is disaggregated into various sub-networks. Randomization provides a better disaggregation resulting in a higher modularity score, however randomizing procedure drastically increases the calculation time. In our analysis, “randomize” and “use weights” options were chosen to produce better disaggregation, and the edge weights were considered while computing the modularity. This these options, modularity was calculated for both Genes-Node and Genes-Edge versions.

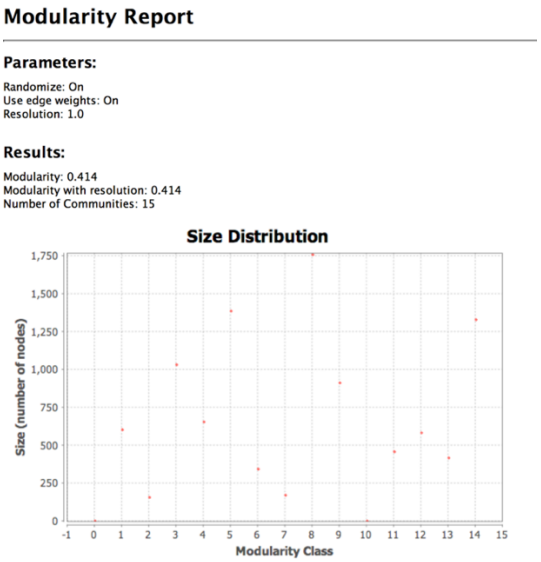


Figure 13: Modularity class sizes for the Genes-Node version network.

Modularity report in Figure 13 shows that modularity was calculated as 0,414 and 15 communities were found for Genes-Node version network. This positive modularity score indicates the presence of modularity structure and it is an average value for this network. This score is acceptable because 24 kind of nodes exist in network structure which are genes, affected systems and 22 disorder classes and they all show different patterns and edge properties. In Figure 14 OpenOrd layout was applied to reveal communities more clearly. Also, distinct colors and respective percentages for each community are visualized to distinguish clusters.

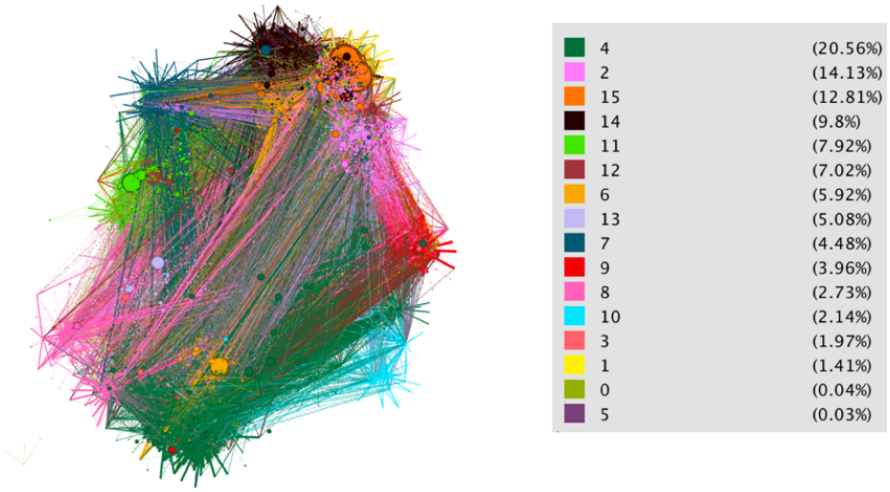


Figure 14: Modularity class colorization and percentage information in total for Gene Node version.

Some small sub-networks exist in the network and their connections are isolated and displayed for the selected examples of Aanat and Hal genes together with their connections.

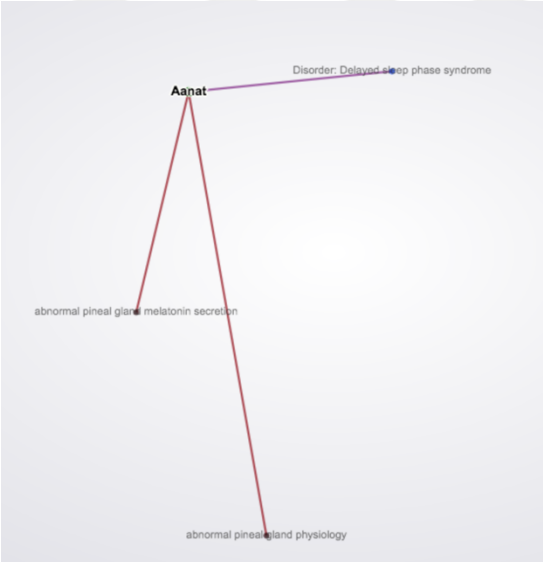


Figure 15: Aanat targeted knock-out mouse gene and its connections as a subnetwork.

In Figure 15, Aanat gene and its connections “Disorder: Delayed sleep phase syndrome”, and affected systems “abnormal pineal gland melatonin secretion” and “abnormal pineal gland physiology” show that this sub-network constituted a distinct community. These nodes are only connected to each other and is separate from rest of the network and its class number was “0”.

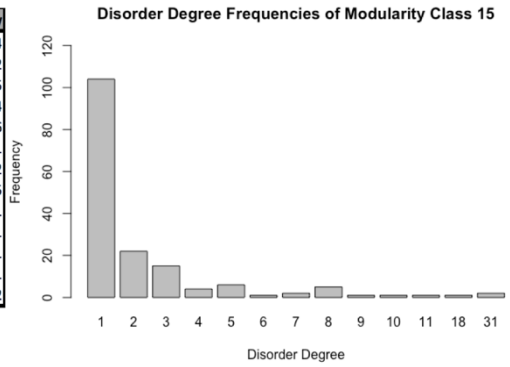
Figure 16: Hal targeted knock-out mouse gene and its connections as a subnetwork.

Also, in Figure 16, Hal gene and its connections “Disorder: Histidinemia”, and affected system “increased urine histidine level” show that this sub-network constituted a distinct community and its class number was “5”.

Modularity clusters kept in group number 4 and 15 were analyzed to observe the network characteristics. These are distinct hub clusters and show different features. Degree frequencies for disorders in modularity class 15 in Figure 17shows that mostly cancer, hematological and metabolic diseases belong to that class. Disorder class frequencies are illustrated at the right-side pane in Figure 17.

Also, the average disorder degree frequency for modularity class 15 is calculated as 2,4787 and they are separately illustrated at the left side pane in Figure 17. This table shows that 104 diseases in that modularity class have degree 1 and 22 diseases have degree of 2, etc. The diseases “Leukemia” and “Colon Cancer” that have 31 connections, stand together at the last line of this table.

degree	degree frequency
1	104
2	22
3	15
4	4
5	6
6	1
7	2
8	5
9	1
10	1
11	1
18	1
31	2



Modularity Class 15 Disorder Classes	Frequency
Disorder Class: Cancer	51
Disorder Class: Hematological	31
Disorder Class: Metabolic	19
Disorder Class: Multiple	17
Disorder Class: Cardiovascular	7
Disorder Class: Dermatological	5
Disorder Class: Endocrine	5
Disorder Class: Developmental	4
Disorder Class: Neurological	4
Disorder Class: Ophthamological	4
Disorder Class: Unclassified	3
Disorder Class: Bone	2
Disorder Class: Gastrointestinal	2
Disorder Class: Immunological	2
Disorder Class: Psychiatric	2
Disorder Class: Renal	2
Disorder Class: Skeletal	2
Disorder Class: Connective Tissue Disorder	1
Disorder Class: Respiratory	1

Figure 17: Degree analysis result for disorders in Modularity class 15

Degree indicates number of total connected mouse knock-out genes for an affected system. The average affected system degree frequency for modularity class 15 is calculated as 10,3482 and distinct degree frequencies are illustrated in Figure 17. Bar plot in Figure 18 also shows that this class keeps affected systems with degree 1 mostly but also some affected systems with high degrees exist, as well. Affected system with the degree value of 440 is the ‘‘premature death’’.

degree	degree frequency	degree	degree frequency	degree	degree frequency
1	216	35	2	124	1
2	132	36	3	130	1
3	78	37	3	135	1
4	60	39	1	152	1
5	39	42	3	159	1
6	43	43	1	440	1
7	35	44	2		
8	27	45	1		
9	28	46	1		
10	20	49	1		
11	15	50	1		
12	16	51	2		
13	12	53	3		
14	15	54	1		
15	13	55	1		
16	12	56	2		
17	11	58	2		
18	6	60	1		
19	8	61	1		
20	6	63	2		
21	5	64	1		
22	5	69	1		
23	6	75	2		
24	5	76	1		
25	3	78	1		
26	7	79	1		
27	1	84	1		
28	3	86	1		
29	1	88	1		
30	2	91	1		
31	1	99	1		
32	2	105	1		
33	1	110	1		
34	4	111	1		

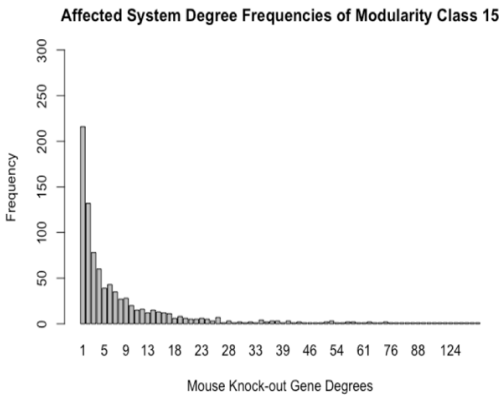


Figure 18: Degree analysis result for affected systems in Modularity class 15.

Degree shows the number of total connected diseases and affected systems to the corresponding mouse knock-out gene. The average targeted knock-out mouse gene degree frequency for modularity class 15 is calculated as 42,058 and distinct degree frequencies are illustrated in Figure 19. The mouse knock-out gene that has degree 308 is Trp53.

degree	degree frequency	degree	degree frequency	degree	degree frequency
2	2	42	1	100	1
3	4	43	5	108	1
4	6	44	6	110	1
5	8	46	1	112	1
6	5	47	3	115	1
7	6	48	2	116	1
8	9	50	1	120	1
9	4	51	1	123	1
10	3	52	2	134	1
11	3	54	1	142	1
12	5	58	1	143	1
13	3	60	1	149	1
14	3	61	1	153	1
15	5	65	1	156	1
16	5	66	1	170	1
17	1	67	2	173	1
18	4	69	1	186	1
19	4	71	2	215	1
20	3	72	3	308	1
21	3	73	2		
22	6	74	1		
23	2	75	1		
24	4	76	1		
25	3	77	2		
26	6	78	1		
27	2	80	3		
28	3	83	1		
31	2	85	1		
32	1	86	1		
35	3	89	1		
36	5	92	1		
37	1	93	1		
38	1	94	2		
39	5	97	1		

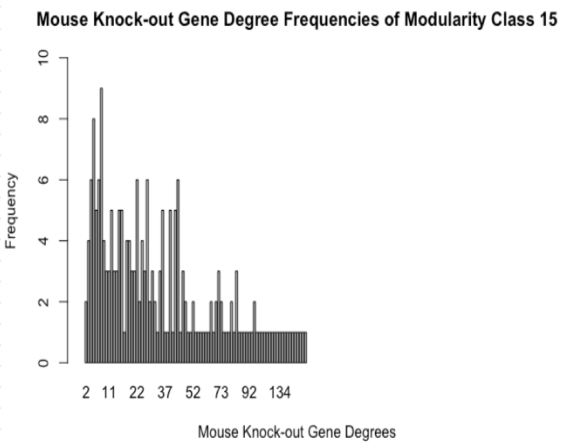
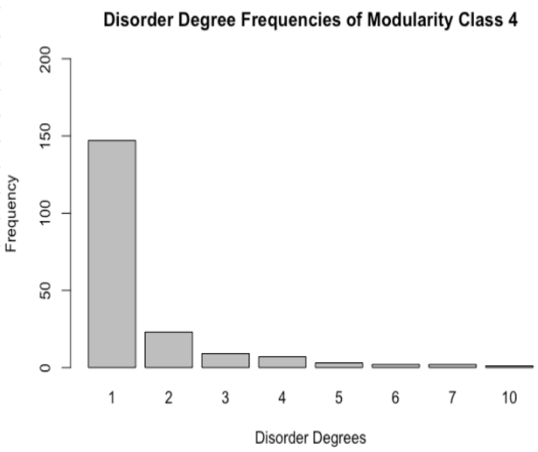


Figure 19: Degree analysis result for mouse knock-out genes in modularity class 15

The biggest modularity class in Genes-Node version was the 4th class. Figure 20 shows the degree frequencies for disorders in modularity class 4. As observed, mostly skeletal, multiple and dermatological disorder classes located in this group. Also, the average disorder degree frequency for modularity class 4 was calculated as 1.54123 and they are separately illustrated at the table stands on the left side in Figure 20. The disease which have 10 connections is Epidermolysis bullosa. Bar plot shows degree distribution for disorders in modularity class 4.

Degree	Degree Frequency
1	147
2	23
3	9
4	7
5	3
6	2
7	2
10	1



Modularity Class 4 Disorder Classes	Frequency
Disorder Class: Skeletal	45
Disorder Class: Multiple	29
Disorder Class: Dermatological	20
Disorder Class: Bone	14
Disorder Class: Developmental	13
Disorder Class: Cancer	11
Disorder Class: Connective Tissue Disorder	11
Disorder Class: Endocrine	11
Disorder Class: Metabolic	11
Disorder Class: Neurological	11
Disorder Class: Hematological	6
Disorder Class: Ophthamological	4
Disorder Class: Cardiovascular	2
Disorder Class: Gastrointestinal	2
Disorder Class: Immunological	2
Disorder Class: Nutritional	1

Figure 20: Degree analysis result for disorders in Modularity class 4.

The average affected system degree frequency for modularity class 4 is calculated as 6.9760 and distinct degree frequencies are illustrated in Figure 21. Bar plot in Figure 21 also displays that this class mostly contains affected systems with degree 1 but also some affected systems with higher degrees exist, as well. The affected system with the degree value of 423 is the decreased body weight.

Degree	Degree Frequency	Degree	Degree Frequency
1	462	38	2
2	288	39	1
3	184	40	2
4	137	41	2
5	100	43	1
6	69	44	2
7	45	45	3
8	38	47	1
9	29	48	2
10	26	49	1
11	25	50	1
12	18	51	2
13	23	52	2
14	14	53	1
15	15	56	1
16	16	57	1
17	11	58	1
18	3	66	1
19	9	67	1
20	7	73	1
21	11	78	1
22	6	79	1
23	9	83	1
24	5	87	1
27	7	91	1
28	4	152	1
30	7	199	1
31	4	250	1
32	3	298	1
33	6	423	1
34	4		
35	2		
36	4		
37	1		

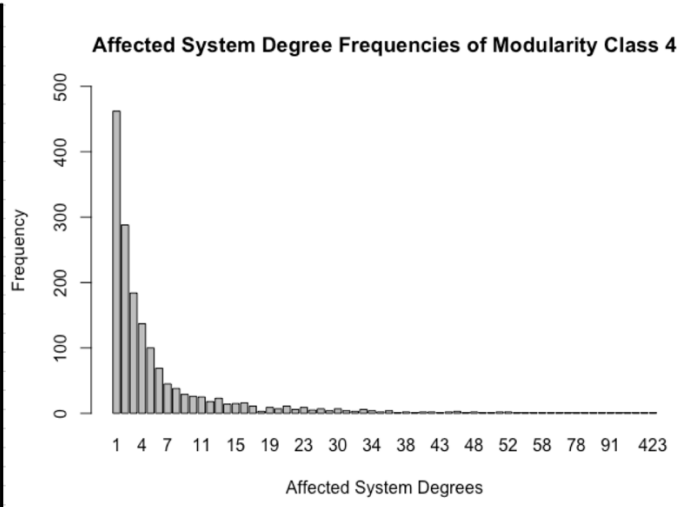


Figure 21: Degree analysis result for affected systems in Modularity class 4

The average targeted knock-out mouse gene degree frequency for modularity class 4 was calculated as 51,136 and distinct degree frequencies are illustrated in Figure 22. This modularity class have higher average degree frequency compared to the class 15, since most of the loosely interconnected disorders classes (such as bone, skeletal, multiple etc.) and nearly all connected mouse knock-out genes are kept in class 4. Therefore, it is usual to see higher average degree for mouse knock-out genes in modularity class 4. The mouse knock-out gene that has the degree value of 343 is Fgfr2.

Degree	Degree Frequency	Degree	Degree Frequency	Degree	Degree Frequency	Degree	Degree Frequency
2	3	42	2	82	2	227	1
4	3	44	1	84	1	303	1
5	3	45	2	85	1	333	1
6	4	47	2	86	1	343	1
8	2	48	1	93	1		
9	1	49	2	94	2		
10	3	50	1	99	1		
11	1	51	1	106	1		
12	3	52	2	109	2		
13	4	53	3	113	1		
14	6	54	2	115	2		
15	6	55	4	116	1		
16	2	57	1	117	1		
17	7	58	3	127	1		
18	2	59	3	135	1		
19	1	60	2	141	1		
20	2	61	3	142	1		
21	2	62	1	145	4		
22	3	63	1	148	1		
24	1	64	1	149	1		
25	5	65	1	156	1		
26	3	66	3	159	1		
27	2	67	4	162	1		
28	2	68	1	165	2		
29	6	69	1	170	1		
30	2	70	1	173	1		
32	3	71	1	186	1		
33	3	72	1	194	1		
36	2	73	1	196	1		
37	2	74	1	197	1		
38	5	76	2	204	1		
39	2	78	1	206	1		
40	1	79	1	211	1		
41	1	80	2	214	1		

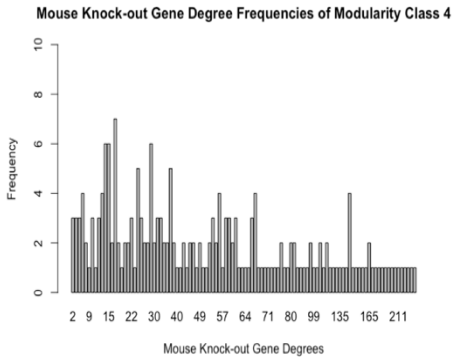


Figure 22: Degree analysis result for mouse knock-out genes in modularity class 4

Figure 23 displays the modularity class sizes for Genes-Edge version network. The report in Figure 23 shows that the modularity was calculated as 0.368 and 12 communities were formed in total. This positive modularity score indicates the presence of modularity structure and it is an average value for this network. The modularity score is reduced in Genes-Edge version because genes are not treated as nodes anymore and the diseases are directly connected to the affected systems. Therefore, it is possible that some nodes belong to different classes in the previous network may remained in same class in this network version.

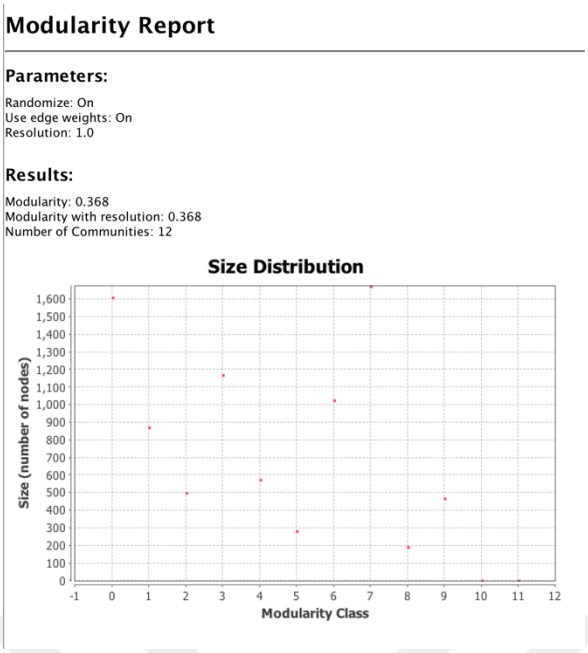


Figure 23: Modularity class sizes for the Genes-Edge version network

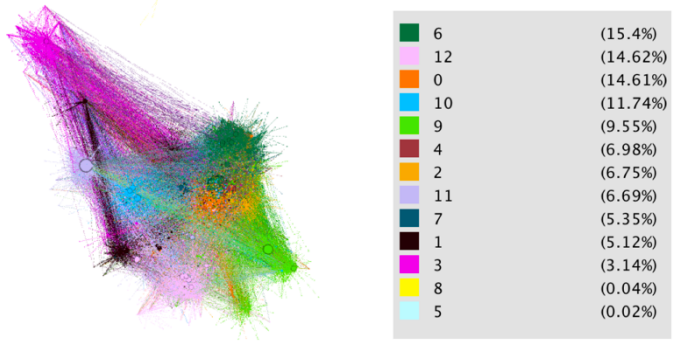


Figure 24: Modularity class colorization and percentage information in total

In Figure 24, OpenOrd layout was again applied to reveal communities (i.e. classes). Small sub-networks similar to the ones shown in Figure 15 and Figure 16 are isolated from the whole network and illustrated in Figure 25 and Figure 26. These are "Disorder: Delayed sleep phase syndrome", Disorder: Histidinemia", and their connections.

Figure 25: Delayed sleep phase syndrome disorder & its connections as a subnetwork.

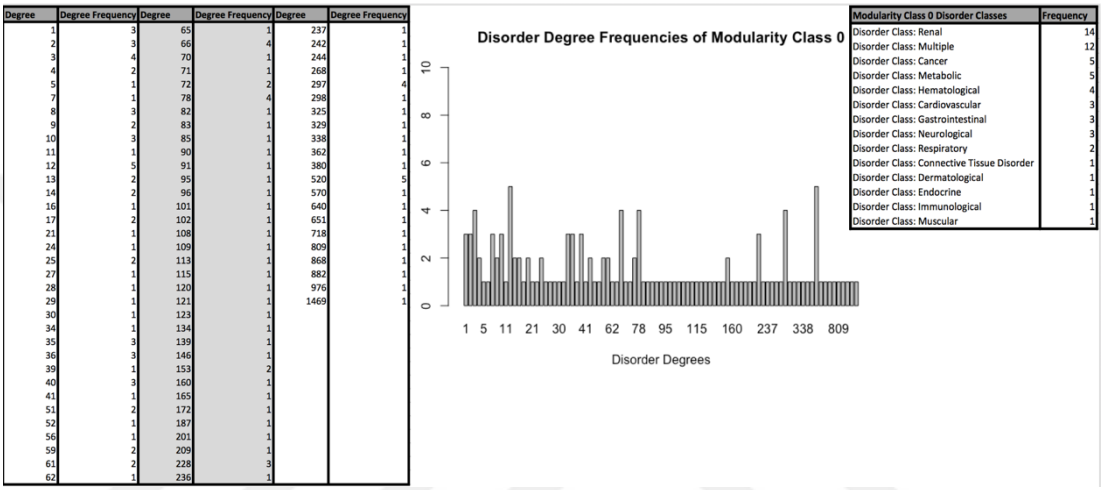
In Figure 25, “Disorder: Delayed sleep phase syndrome” and its connection “abnormal pineal gland melatonin secretion” and “abnormal pineal gland physiology” affected systems are shown and that this sub-network constituted a distinct community number “8”. It stays separated from rest of the network.

Figure 26: Hal targeted knock-out mouse gene and its connections as a subnetwork

In Figure 26, “Disorder: Histidinemia”, and affected system “increased urine histidine level” show that this sub-network constituted a distinct community with number “5”.

Modularity clusters 0 and 7 were analyzed to see network characteristics. These are distinct hub clusters and show varying features. Modularity class "0” mostly has cancer class diseases and some connected hub genes. Degree frequencies for disorders in modularity class 0 (Figure 27) reveals that mostly renal, multiple and cancer diseases

belong to this module. Also, the average disorder degree frequency for modularity class 0 is calculated as 155.5073 and they are separately illustrated at the table stands at left in Figure 27. This score is increased when it is compared with Genes-Node version because diseases directly connect to the affected systems in Genes-Edge version. The diseases “Colon Cancer” and “Breast Cancer” have 1469 and 976 connections, respectively.



Degree	Degree Frequency	Degree	Degree Frequency	Degree	Degree Frequency
1	3	65	1	237	1
2	3	66	4	242	1
3	4	70	1	244	1
4	2	71	1	268	1
5	1	72	2	297	4
7	1	78	4	298	1
8	3	82	1	325	1
9	2	83	1	329	1
10	3	85	1	338	1
11	1	90	1	362	1
12	5	91	1	380	1
13	2	95	1	520	5
14	2	96	1	570	1
16	1	101	1	640	1
17	2	102	1	651	1
21	1	108	1	718	1
24	1	109	1	809	1
25	2	113	1	868	1
27	1	115	1	882	1
28	1	120	1	976	1
29	1	121	1	1469	1
30	1	123	1		
34	1	134	1		
35	3	139	1		
36	3	146	1		
39	1	153	2		
40	3	160	1		
41	1	165	1		
51	2	172	1		
52	1	187	1		
56	1	201	1		
59	2	209	1		
61	2	228	3		
62	1	236	1		

Modularity Class 0 Disorder Classes	Frequency
Disorder Class: Renal	14
Disorder Class: Multiple	12
Disorder Class: Cancer	5
Disorder Class: Metabolic	5
Disorder Class: Hematological	4
Disorder Class: Cardiovascular	3
Disorder Class: Gastrointestinal	3
Disorder Class: Neurological	3
Disorder Class: Respiratory	2
Disorder Class: Connective Tissue Disorder	1
Disorder Class: Dermatological	1
Disorder Class: Endocrine	1
Disorder Class: Immunological	1
Disorder Class: Muscular	1

Figure 27: Degree analysis result for disorders in Modularity class 0.

The average affected system degree frequency for modularity class 0 is calculated as 19.3448 and distinct degree frequencies are illustrated in Figure 28. Bar plot in Figure 28 also shows that this class contains affected systems with degree 1 mostly, together with some affected systems with high degrees. Affected system with the degree value of 440 is mortality/aging.

Degree	Degree Frequency	Degree	Degree Frequency	Degree	Degree Frequency
1	103	35	10	72	1
2	38	36	7	75	2
3	44	37	5	80	1
4	63	38	8	83	1
5	36	39	5	84	3
6	37	40	7	89	1
7	45	41	5	91	1
8	36	42	4	93	1
9	57	43	2	97	1
10	42	44	4	100	1
11	35	45	2	103	1
12	32	46	3	105	1
13	29	47	1	107	3
14	25	48	7	118	1
15	24	49	3	123	1
16	23	50	2	124	2
17	22	51	2	150	1
18	22	52	5	158	1
19	33	53	5	170	1
20	23	54	1	177	1
21	21	55	2	180	1
22	16	56	2	191	1
23	18	57	2	193	1
24	13	58	1	219	1
25	12	59	1	223	1
26	10	60	7	226	1
27	12	61	1	247	1
28	14	62	1	251	1
29	14	64	1		
30	14	65	2		
31	7	67	2		
32	7	68	2		
33	8	69	2		
34	3	71	3		

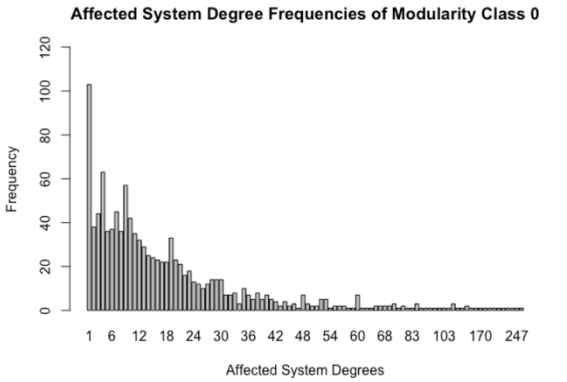
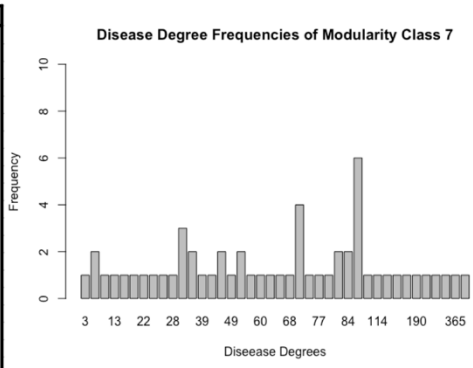


Figure 28: Degree analysis result for affected systems in Modularity class 0

Another modularity class in 7th group is analyzed as being one of the hub classes in Genes-Edge version. Degree frequencies for disorders in modularity class 7 (Figure 29) indicates that mostly renal, multiple and cancer disorder classes located in this group. Also, the average disorder degree frequency for modularity class 15 is calculated as 82.5357 and they are separately illustrated at the table stands at the left side in Figure 29. Bar plot shows degree distribution for disorders in modularity class 4 and disorders with degree 85 are the most frequent ones as the tallest bar. These are Yemenite deaf blind hypopigmentation syndrome, Frasier syndrome, WAGR syndrome, PCWH, Denys Drash syndrome and Mesangial sclerosis diseases.

Degree	Degree Frequency	Degree	Degree Frequency
3	1	190	1
8	2	208	1
12	1	222	1
13	1	272	1
14	1	365	1
17	1	367	1
22	1		
24	1		
26	1		
28	1		
30	3		
32	2		
39	1		
43	1		
47	2		
49	1		
51	2		
54	1		
60	1		
65	1		
66	1		
68	1		
69	4		
74	1		
77	1		
82	1		
83	2		
84	2		
85	6		
93	1		
114	1		
134	1		
162	1		
173	1		



Modularity Class 7 Disorder Classes	Frequency
Disorder Class: Renal	14
Disorder Class: Multiple	12
Disorder Class: Cancer	5
Disorder Class: Metabolic	5
Disorder Class: Hematological	4
Disorder Class: Cardiovascular	3
Disorder Class: Gastrointestinal	3
Disorder Class: Neurological	3
Disorder Class: Respiratory	2
Disorder Class: Connective Tissue Disorder	1
Disorder Class: Dermatological	1
Disorder Class: Endocrine	1
Disorder Class: Immunological	1
Disorder Class: Muscular	1

Figure 29: Degree analysis result for disorders in Modularity class 7

The average affected system degree frequency for modularity class 4 is calculated as 19.6465 and distinct degree frequencies are illustrated in Figure 30. Bar plot in Figure 30 also shows that this class contains the affected systems with degree 1. Affected system with the degree value of 350 is postnatal lethality, incomplete penetrance.

Degree	Degree Frequencies	Degree	Degree Frequencies
1	47	38	1
2	32	39	1
3	36	43	2
4	28	47	1
5	35	51	1
6	33	55	1
7	19	56	1
8	18	57	1
9	11	59	1
10	15	63	1
11	9	66	1
12	11	68	1
13	8	76	1
14	6	91	1
15	7	109	1
16	8	116	1
17	5	125	1
18	4	138	1
19	3	151	1
20	4	350	1
21	3		
22	4		
23	3		
24	4		
25	3		
26	3		
27	2		
28	3		
30	1		
31	2		
32	1		
33	1		
37	3		

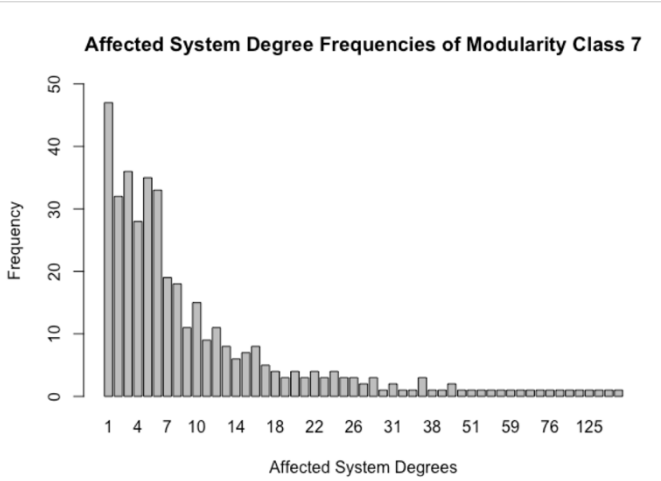


Figure 30: Degree analysis result for affected systems in Modularity class 7

There is another measure, which shows the importance of a node in a network based on a node's connections. Sum change was found as 0.061 and 0.076 in Genes-Node and Genes-Edge versions respectively. It can be said that the nodes connected to central nodes are considered central themselves. Eigenvector centrality distributions for Genes-Node and Genes-Edge networks are visualized under Figure 31 and Figure 32, respectively.

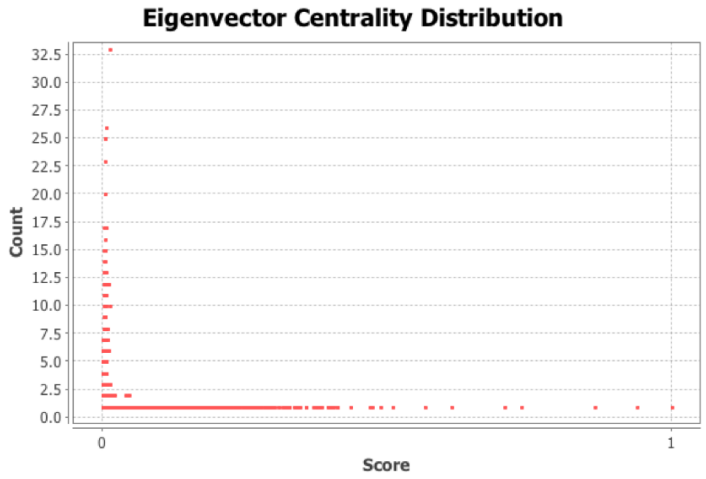


Figure 31: Eigenvector Centrality distribution for Genes-Node version network

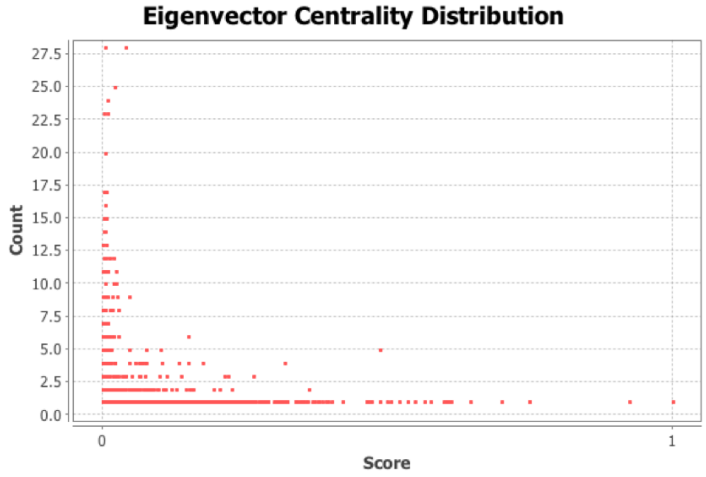


Figure 32: Eigenvector Centrality distribution for Genes-Edge version network

All the graphical statistical analyses are summarized for both versions of networks under the Table 7.

Table 7 : Gephi statistical analysis results for Genes-Node and Genes-Edge version networks.

GEPHI STATISTICAL ANALYSIS RESULTS	GENE-NODE VERSION	GENE-EDGE VERSION
AVERAGE CLUSTERING COEFFICIENT	0	0
AVERAGE DEGREE	12.725	26.62
AVERAGE PATH LENGTH	3.84	3.45
AVERAGE WEIGHTED DEGREE	26.203	30.095
DIAMETER	8	7
EIGENVECTOR CENTRALITY SUM CHANGE	0.0616058649	0.0760103896
GRAPH DENSITY	0.001	0.003
MODULARITY	0.427	0.384
MODULARITY with RESOLUTION	0.427	0.384
NUMBER OF THE SHORTEST PATHS	97130898	70216028
NUMBER OF TOTAL COMMUNITIES	15	13

4.2 STATISTICAL ANALYSIS OF THE NETWORKS

The following sections describes the statistical analyses done in R platform. R is a free software environment for making statistical computing and analyze graphics. It can compile and run on an extensive variety of UNIX platforms, Windows and MacOS (The R Project for Statistical Computing. 1993. R Core Team. [ONLINE] Available at: https://www.r-project.org/). In the following analyses, the diseases were ranked according to number of targeted knock-out genes they have and the genes are ranked regarding both the number of diseases and the number of affected systems they are related to.

4.2.1 DISEASE STATISTICS

In this analysis, the diseases are arranged from the most to least populated in terms of the connected genes. Histogram plot of diseases vs. genes are illustrated under Figure 33 and the top5 diseases were shown in Figure 34. These diseases are Deafness, Leukemia, Colon cancer, Retinitis Pigmentosa and Diabetes Mellitus, having connections with 38, 31, 31, 26 and 22 genes, respectively.

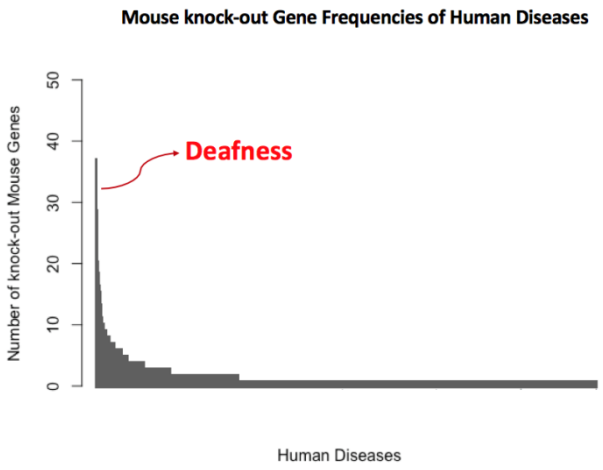


Figure 33:Frequency plot of all diseases in terms of their connected targeted knock-out mouse genes

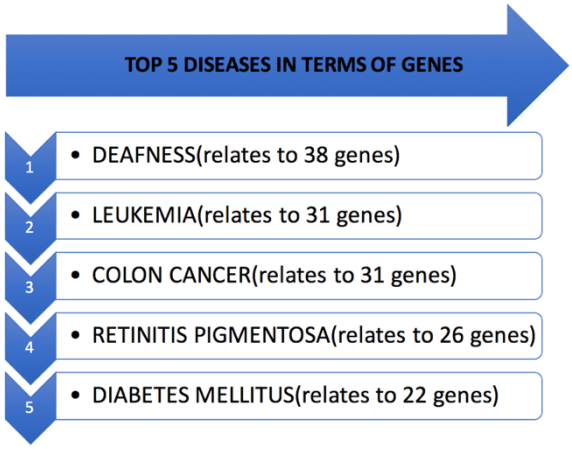


Figure 34:Top five diseases are listed according to their related total number of genes.

In Table 8, frequencies for all the diseases are shown. Total gene number column shows the number of targeted knock-out mouse genes and column total diseases number shows how various diseases have the corresponding number of mouse knock-out genes. For example, it was calculated that 881 diseases are correlated with just 1 mouse knock-out gene.

Table 8: Gene frequencies for diseases in total

Total Gene Number	Total disease Number
1	881
2	165
3	65
4	39
5	15
6	19
7	11
8	9
9	5
10	4
11	2
12	1
13	1
14	1
15	2
16	1
17	1
18	1
19	1
20	1
26	1
28	1
30	1
38	1

4.2.2 GENE STATISTICS AND A CASE STUDY

Statistical computing was done to see distributions of genes in terms of diseases by using R programming. In Figure 35, the histogram of genes vs. connected diseases was illustrated. The gene mostly distinguished in the histogram is Trp53 and has 11 disease connections in total. Also, in Figure 36, the top 5 genes are listed.

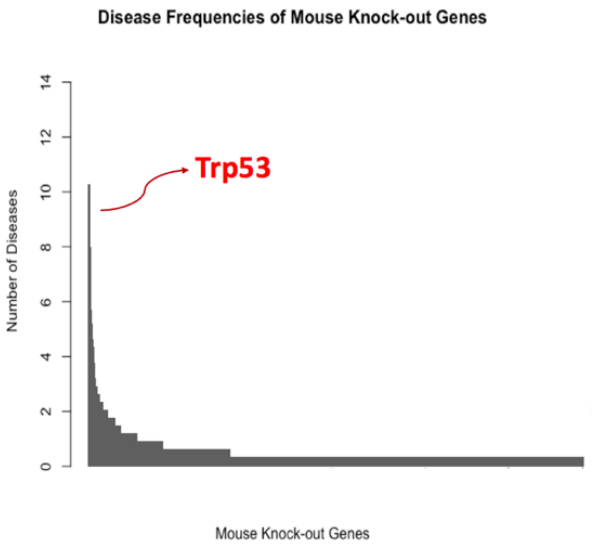


Figure 35: Frequency plot of all mouse knock-out genes in terms of their connected diseases.

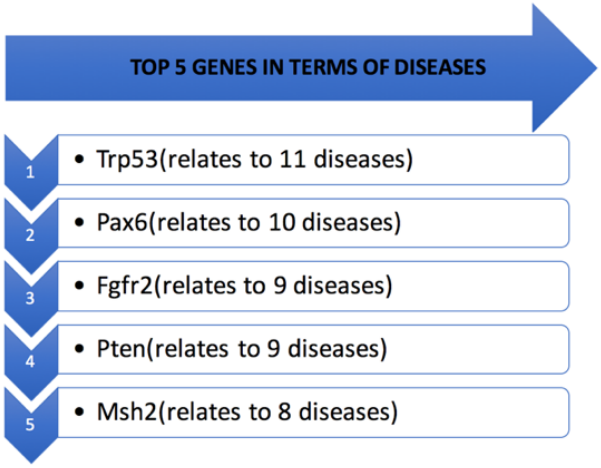


Figure 36: Top 5 genes specified according to the number of diseases they are related to.

In Table 9: Disease frequencies for genes in total, all the genes are grouped according to their total disease frequencies. Total disease number column shows the number of diseases and total gene number shows how various genes have the corresponding

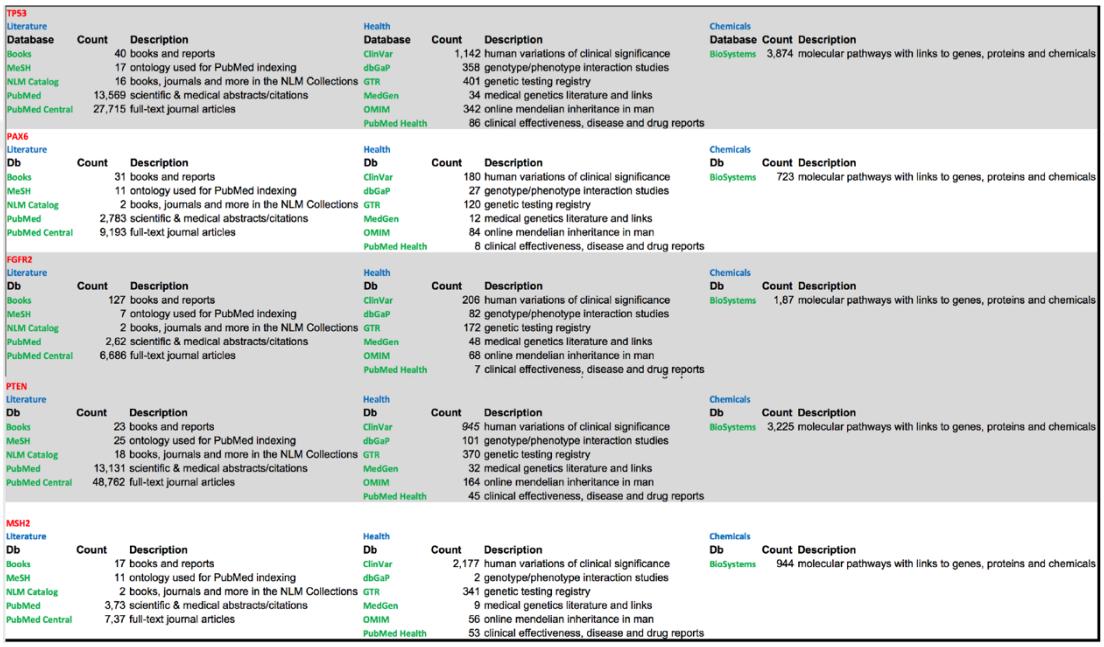
number of diseases. For example, it was calculated that 1031 genes are connected to just 1 disease. 257 genes are connected to 2 diseases.

Table 9: Disease frequencies for genes in total

Total Disease Number	Total Gene Number
1	1031
2	257
3	109
4	35
5	29
6	7
7	3
8	3
9	2
10	1
11	1

A literature and a database review was done for the top5 genes. In this review, the hub genes in terms of diseases are investigated to make inferences about Mouse2Human Network node statistics. Human orthologues of top5 mouse knock-out genes were searched from NCBI databases and from the relevant literature, various types of statistics about these genes are listed in Figure 37. The literature review indicated that these genes have been studied and investigated extensively. Chemical results signify the number of molecular pathways linked to these genes. For example, one of the top5 genes in Figure 36, the p53 tumor suppressor gene (TP53 in humans or Trp53 in mice) is crucial for inhibiting tumor growth (Blackburn & Jerry, 2002). Another gene, the PAX6 belongs to a PAX gene family and plays a critical role in the formation of tissues and organs during embryonic development (Thakurela *et al.*, 2016). PTEN is also highly studied and it belongs to a tumor suppressor gene family (LESLIE & DOWNES, 2004). Tumor suppressor genes are related to the cell growth control, and acting to block cell proliferation and tumor development, which is the reason why these genes are highly studied (Lee & Muller, 2010). It was stated that the FGFR2 abnormalities underlie a wide range of bone, skin and cancer pathologies because FGF

group of genes are involved in fetal morphogenesis, adult tissue homeostasis, and tumorigenesis (Dailey *et al.*, 2005; Eswarakumar *et al.*, 2005; Grose and Dickson, 2005; Wilkie, 2005; Chaffer *et al.*, 2007). The MSH2 gene codes a protein that plays an important role in DNA repair, as it aids fixing errors in DNA replication. As a result, it plays roles in various fatal human diseases (Pereira *et al.*, 2013). These findings correlate with these genes being hub nodes in the network, as their roles in various diseases are revealed in the literature and recorded in disease databases.



TP53

Literature

Database	Count	Description
Books	40	books and reports
MeSH	17	ontology used for PubMed indexing
NLM Catalog	16	books, journals and more in the NLM Collections
PubMed	13,569	scientific & medical abstracts/citations
PubMed Central	27,715	full-text journal articles

Health

Database	Count	Description
ClinVar	1,142	human variations of clinical significance
dbGaP	358	genotype/phenotype interaction studies
GTR	401	genetic testing registry
MedGen	34	medical genetics literature and links
OMIM	342	online mendelian inheritance in man
PubMed Health	86	clinical effectiveness, disease and drug reports

Chemicals

Database	Count	Description
BioSystems	3,874	molecular pathways with links to genes, proteins and chemicals

PAX6

Literature

Db	Count	Description
Books	31	books and reports
MeSH	11	ontology used for PubMed indexing
NLM Catalog	2	books, journals and more in the NLM Collections
PubMed	2,783	scientific & medical abstracts/citations
PubMed Central	9,193	full-text journal articles

Health

Db	Count	Description
ClinVar	180	human variations of clinical significance
dbGaP	27	genotype/phenotype interaction studies
GTR	120	genetic testing registry
MedGen	12	medical genetics literature and links
OMIM	84	online mendelian inheritance in man
PubMed Health	8	clinical effectiveness, disease and drug reports

Chemicals

Db	Count	Description
BioSystems	723	molecular pathways with links to genes, proteins and chemicals

FGFR2

Literature

Db	Count	Description
Books	127	books and reports
MeSH	7	ontology used for PubMed indexing
NLM Catalog	2	books, journals and more in the NLM Collections
PubMed	2,62	scientific & medical abstracts/citations
PubMed Central	6,686	full-text journal articles

Health

Db	Count	Description
ClinVar	206	human variations of clinical significance
dbGaP	82	genotype/phenotype interaction studies
GTR	172	genetic testing registry
MedGen	48	medical genetics literature and links
OMIM	68	online mendelian inheritance in man
PubMed Health	7	clinical effectiveness, disease and drug reports

Chemicals

Db	Count	Description
BioSystems	1,87	molecular pathways with links to genes, proteins and chemicals

PTEN

Literature

Db	Count	Description
Books	23	books and reports
MeSH	25	ontology used for PubMed indexing
NLM Catalog	18	books, journals and more in the NLM Collections
PubMed	13,131	scientific & medical abstracts/citations
PubMed Central	48,762	full-text journal articles

Health

Db	Count	Description
ClinVar	945	human variations of clinical significance
dbGaP	101	genotype/phenotype interaction studies
GTR	370	genetic testing registry
MedGen	32	medical genetics literature and links
OMIM	164	online mendelian inheritance in man
PubMed Health	45	clinical effectiveness, disease and drug reports

Chemicals

Db	Count	Description
BioSystems	3,225	molecular pathways with links to genes, proteins and chemicals

MSH2

Literature

Db	Count	Description
Books	17	books and reports
MeSH	11	ontology used for PubMed indexing
NLM Catalog	2	books, journals and more in the NLM Collections
PubMed	3,73	scientific & medical abstracts/citations
PubMed Central	7,37	full-text journal articles

Health

Db	Count	Description
ClinVar	2,177	human variations of clinical significance
dbGaP	2	genotype/phenotype interaction studies
GTR	341	genetic testing registry
MedGen	9	medical genetics literature and links
OMIM	56	online mendelian inheritance in man
PubMed Health	53	clinical effectiveness, disease and drug reports

Chemicals

Db	Count	Description
BioSystems	944	molecular pathways with links to genes, proteins and chemicals

Figure 37: NCBI statistics for Top 5 gene in terms of diseases

A similar analysis has been done for top genes in terms of the number of associations with the affected systems (as opposed to the previous analysis, which was done for top genes in terms of the number of associations with diseases). The top 5 genes in terms of associated phenotypes are displayed under Figure 38. Also, in Figure 4-30, the hitogram of genes vs. phenotypes is illustrated. The gene with the highest rank in the histogram is Pten, which has 520 associated phenotypes. In Table 10, the genes are grouped according to their phenotype association frequencies. The total related

phenotype column shows the number of phenotypes and the second column titles total gene number show how many genes are associated to those phenotypes. For example, 41 genes have just 1 phenotype association.

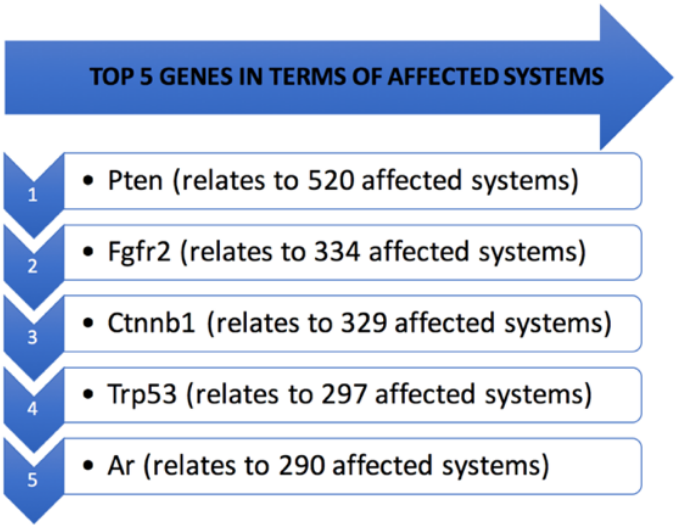


Figure 38: Top5 genes in terms of the total number of their associated affected systems (i.e. phenotypes).

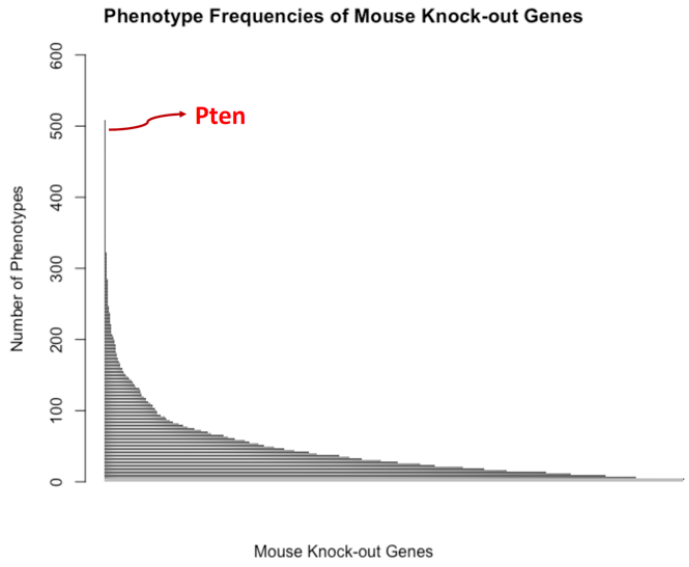


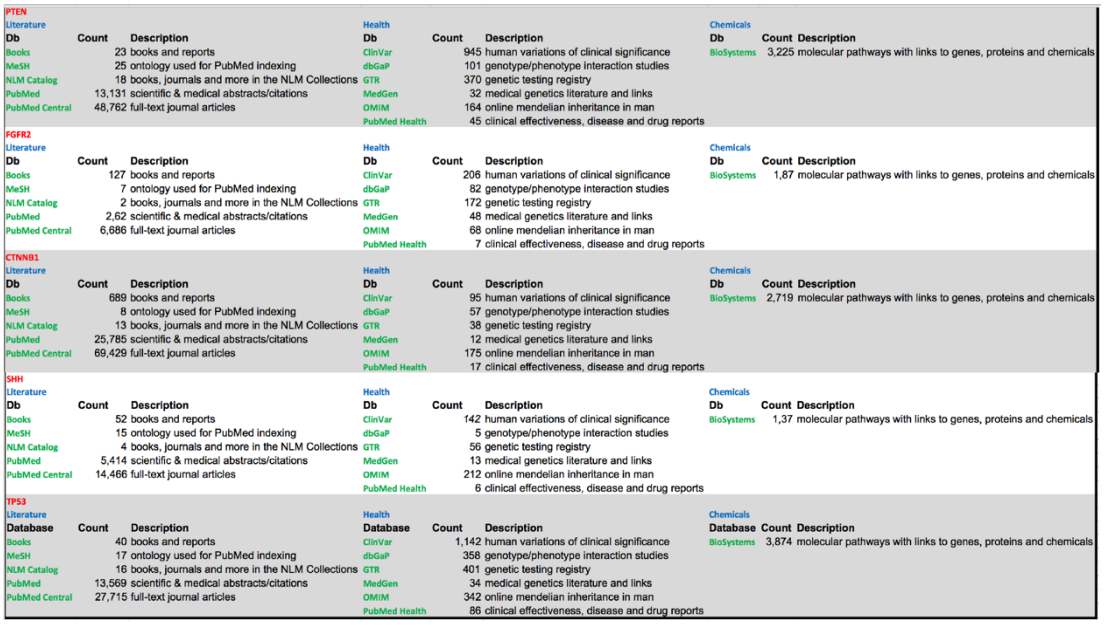
Figure 39: Histogram plot of all mouse knock-out genes in terms of their associated affected systems (i.e. phenotypes).

Table 10: Phenotype frequencies for genes in total

Total Related Phenotypes	Total Gene Number	Total Related Phenotypes	Total Gene Number	Total Related Phenotypes	Total Gene Number	Total Related Phenotypes	Total Gene Number	Total Related Phenotypes	Total Gene Number
1	41	20	23	39	8	58	11	77	8
2	36	21	21	40	11	59	7	78	4
3	45	22	17	41	10	60	7	79	8
4	36	23	22	42	16	61	8	80	5
5	41	24	18	43	17	62	5	81	4
6	26	25	16	44	8	63	8	82	5
7	35	26	22	45	13	64	6	83	4
8	27	27	21	46	12	65	13	84	5
9	35	28	23	47	9	66	13	85	3
10	30	29	18	48	9	67	3	86	2
11	32	30	8	49	8	68	5	87	3
12	36	31	16	50	6	69	3	88	4
13	33	32	15	51	10	70	9	89	2
14	31	33	11	52	6	71	5	90	3
15	25	34	16	53	9	72	10	91	5
16	34	35	24	54	8	73	8	92	5
17	21	36	14	55	6	74	5	93	2
18	23	37	19	56	4	75	7	94	4
19	26	38	12	57	10	76	3	95	2
Total Related Phenotypes	Total Gene Number	Total Related Phenotypes	Total Gene Number	Total Related Phenotypes	Total Gene Number	Total Related Phenotypes	Total Gene Number	Total Related Phenotypes	Total Gene Number
98	1	121	2	148	1	184	2	277	1
99	2	123	1	149	3	193	1	290	1
100	1	125	1	150	3	194	1	297	1
102	1	126	2	152	1	196	2	300	1
103	2	129	1	153	1	202	1	329	1
105	4	131	1	154	1	204	1	334	1
106	1	132	2	155	2	206	1	520	1
107	1	133	3	157	1	208	1		
108	3	134	4	158	2	209	1		
109	3	135	2	159	1	211	2		
111	3	137	1	161	1	219	1		
112	1	138	3	162	1	223	1		
113	2	139	2	163	2	225	1		
114	3	140	2	166	2	228	1		
115	1	143	1	168	1	239	1		
117	1	144	4	172	3	241	1		
118	1	145	2	174	2	242	1		
119	3	146	1	176	1	252	2		
120	4	147	1	180	1	254	1		

A database search was done from the NCBI resources for the human orthologues of top5 mouse knock-out genes in terms of phenotypes. The finding in terms of literature,

health and chemicals statistics are shown in Figure 40. Ctnnb1 and Shh genes exist in the top5 list, different from the previous list shown in Figure 37. Epithelial-mesenchymal transition (EMT) and the related gene CTNNB1 plays an important role for the regulation of cancer signaling and stem cell pluripotency (Tanabe *et al.*, 2016). SHH gene in other words "Sonic Hedgehog" gene encodes a protein that is crucial in the early embryo stage, adult organ homeostasis and organ repair. It has been stated that it provides a key inductive signal ventral neural tube, the anterior-posterior limb axis, and the ventral somite (Petrova & Joyner, 2014).



PTEN

Literature

Db	Count	Description
Books	23	books and reports
MeSH	25	ontology used for PubMed indexing
NLM Catalog	18	books, journals and more in the NLM Collections
PubMed	13,131	scientific & medical abstracts/citations
PubMed Central	48,762	full-text journal articles

Health

Db	Count	Description
ClinVar	945	human variations of clinical significance
dbGaP	101	genotype/phenotype interaction studies
GTR	370	genetic testing registry
MedGen	32	medical genetics literature and links
OMIM	164	online mendelian inheritance in man
PubMed Health	45	clinical effectiveness, disease and drug reports

Chemicals

Db	Count	Description
BioSystems	3,225	molecular pathways with links to genes, proteins and chemicals

FGFR2

Literature

Db	Count	Description
Books	127	books and reports
MeSH	7	ontology used for PubMed indexing
NLM Catalog	2	books, journals and more in the NLM Collections
PubMed	2,62	scientific & medical abstracts/citations
PubMed Central	6,686	full-text journal articles

Health

Db	Count	Description
ClinVar	206	human variations of clinical significance
dbGaP	82	genotype/phenotype interaction studies
GTR	172	genetic testing registry
MedGen	48	medical genetics literature and links
OMIM	68	online mendelian inheritance in man
PubMed Health	7	clinical effectiveness, disease and drug reports

Chemicals

Db	Count	Description
BioSystems	1,87	molecular pathways with links to genes, proteins and chemicals

CTNNB1

Literature

Db	Count	Description
Books	689	books and reports
MeSH	8	ontology used for PubMed indexing
NLM Catalog	13	books, journals and more in the NLM Collections
PubMed	25,785	scientific & medical abstracts/citations
PubMed Central	69,429	full-text journal articles

Health

Db	Count	Description
ClinVar	95	human variations of clinical significance
dbGaP	57	genotype/phenotype interaction studies
GTR	38	genetic testing registry
MedGen	12	medical genetics literature and links
OMIM	175	online mendelian inheritance in man
PubMed Health	17	clinical effectiveness, disease and drug reports

Chemicals

Db	Count	Description
BioSystems	2,719	molecular pathways with links to genes, proteins and chemicals

SHH

Literature

Db	Count	Description
Books	52	books and reports
MeSH	15	ontology used for PubMed indexing
NLM Catalog	4	books, journals and more in the NLM Collections
PubMed	5,414	scientific & medical abstracts/citations
PubMed Central	14,466	full-text journal articles

Health

Db	Count	Description
ClinVar	142	human variations of clinical significance
dbGaP	5	genotype/phenotype interaction studies
GTR	56	genetic testing registry
MedGen	13	medical genetics literature and links
OMIM	212	online mendelian inheritance in man
PubMed Health	6	clinical effectiveness, disease and drug reports

Chemicals

Db	Count	Description
BioSystems	1,37	molecular pathways with links to genes, proteins and chemicals

TP53

Literature

Database	Count	Description
Books	40	books and reports
MeSH	17	ontology used for PubMed indexing
NLM Catalog	16	books, journals and more in the NLM Collections
PubMed	13,569	scientific & medical abstracts/citations
PubMed Central	27,715	full-text journal articles

Health

Database	Count	Description
ClinVar	1,142	human variations of clinical significance
dbGaP	358	genotype/phenotype interaction studies
GTR	401	genetic testing registry
MedGen	34	medical genetics literature and links
OMIM	342	online mendelian inheritance in man
PubMed Health	86	clinical effectiveness, disease and drug reports

Chemicals

Database	Count	Description
BioSystems	3,874	molecular pathways with links to genes, proteins and chemicals

Figure 40: NCBI statistics for Top 5 gene in terms of phenotypes

4.2.3 AFFECTED SYSTEM STATISTICS

The phenotype (a.k.a. affected system) statistics for the network generated using different versions of Dataset 2 (i.e. ''child'', ''parent'', ''2 steps before the root'' and ''higher level affected systems'' versions) were analyzed. The reason of using 4 different versions for Dataset 2 was explained in section 3.2.1.2 under Phenotype Levels title explicitly. Here are the results for child and high-level affected systems

version are given, since the calculations for mid-levels are not straightforward (e.g. the same term can both be a parent and a child phenotype annotation in different cases), leading to biased results.

4.2.3.1 CHILD AFFECTED SYSTEM STATISTICS

The phenotype annotations directly collected from the MGI database are named here as child affected systems (i.e. asserted annotations) and their frequencies are calculated and the top 5 child affected systems are shown in Figure 44. Premature death phenotype is the mostly connected term and mapped to 440 genes in total. Child phenotype systems histogram plot in Figure 45 gives the distribution of phenotypes in terms of the connected genes.

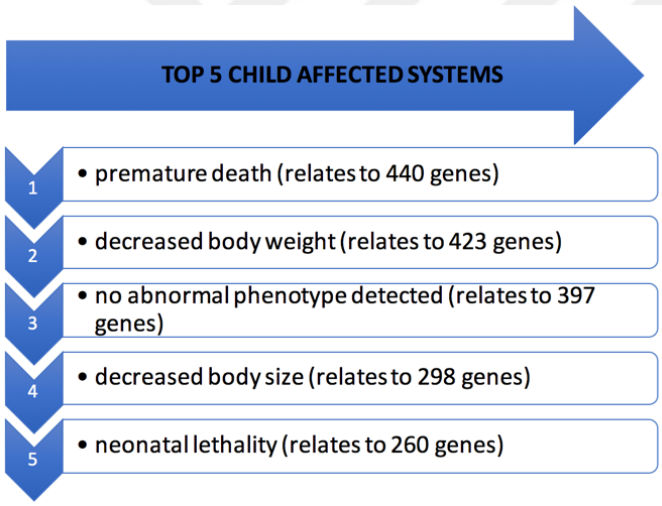


Figure 41: Top 5 child affected systems in terms of the number of gene associations

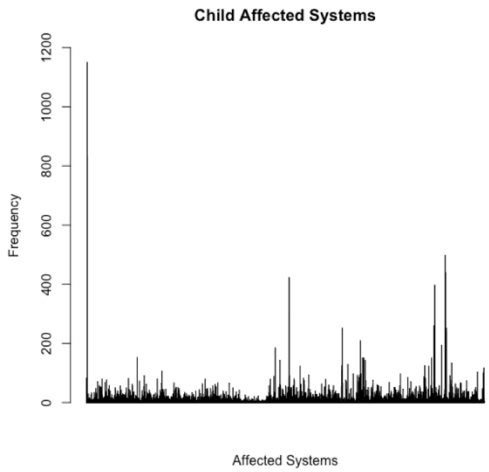


Figure 42: Child affected system histogram plot.

4.2.3.2 HIGH-LEVEL AFFECTED SYSTEM STATISTICS

There are 30 high-level phenotypes under the root term ‘‘Mammalian Phenotype’’ in MPO. The list of all high-level terms is given in Table 11and the top 5 high-level affected systems are shown in Figure 44. Mortality/aging is the most connected phenotype and it is mapped to 2,538 genes in total.

Table 11: The list of high-level phenotypes

HIGH LEVEL AFFECTED SYSTEM
endocrine/exocrine gland phenotype
liver/biliary system phenotype
respiratory system phenotype
cardiovascular system phenotype
reproductive system phenotype
normal phenotype
immune system phenotype
growth/size/body region phenotype
integument phenotype
homeostasis/metabolism phenotype
embryo phenotype
behavior/neurological phenotype
hematopoietic system phenotype
mortality/aging
neoplasm
digestive/alimentary phenotype
nervous system phenotype
cellular phenotype
skeleton phenotype
muscle phenotype
renal/urinary system phenotype
craniofacial phenotype
pigmentation phenotype
vision/eye phenotype
no phenotypic analysis
hearing/vestibular/ear phenotype
obsolete other phenotype
adipose tissue phenotype
limbs/digits/tail phenotype
taste/olfaction phenotype

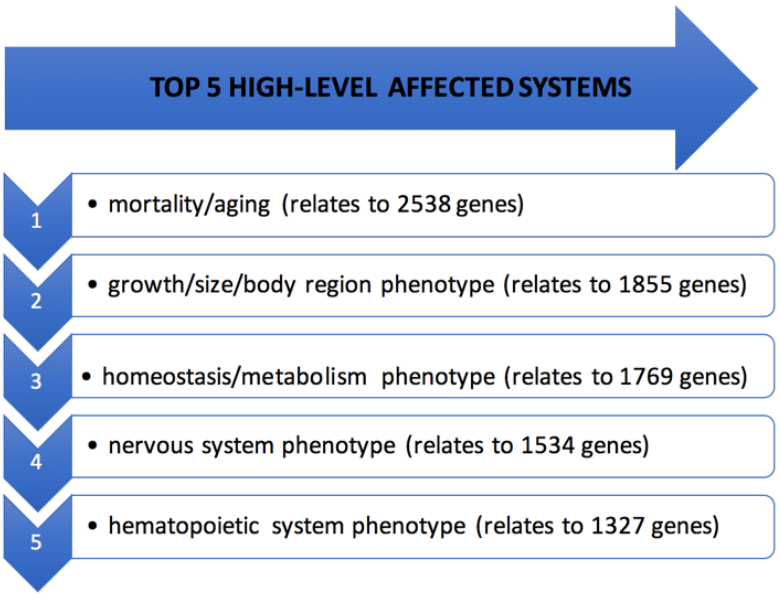
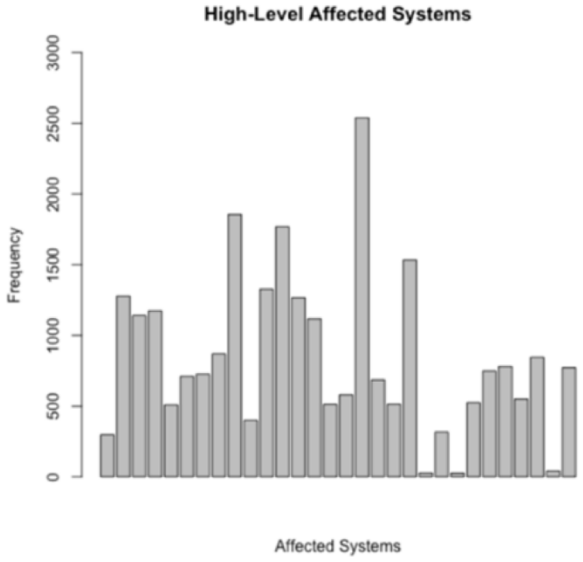


Figure 43: Top 5 high-level affected systems in terms of the number of gene associations

Figure 43 referred to the statistics for high-level phenotypes when all phenotype annotations are propagated through the root of MPO. Figure 44 shows the statistics for the same 30 high-level systems when only direct annotations to these terms are considered (i.e. no propagation from more specific terms). In Figure 44, the most frequent term is “no abnormal phenotype detected” and the most frequent systemic phenotype is the “nervous system phenotype”.



High-Level Systems	Frequency
no abnormal phenotype detected	397
nervous system phenotype	214
homeostasis/metabolism phenotype	210
behavior/neurological phenotype	185
mortality/aging	152
immune system phenotype	151
cardiovascular system phenotype	143
vision/eye phenotype	101
growth/size/body region phenotype	93
hearing/vestibular/ear phenotype	72
neoplasm	69
skeleton phenotype	67
endocrine/exocrine gland phenotype	65
renal/urinary system phenotype	62
digestive/alimentary phenotype	58
cellular phenotype	56
muscle phenotype	48
integument phenotype	47
respiratory system phenotype	45
liver/biliary system phenotype	41
embryo phenotype	38
limbs/digits/tail phenotype	21
adipose tissue phenotype	20
craniofacial phenotype	20
pigmentation phenotype	16
taste/olfaction phenotype	16
TOTAL	2407

Figure 44: High-level affected system histogram plot (only considering direct annotations).

4.3 MOUSE2HUMANNET WEB-SERVICE

Mouse2HumanNet open access web-service was constructed for both Genes-Node and Genes-Edge versions using Gephi web exporter tool gexf Js master (Velt, R. 2011). Detailed informationabout the web-service usage is given under Appendix A.

4.4 TERM SIMILARITY CALCULATIONS WITH CASE STUDIES

Users can utilize the Genes-Edge network especially when they are interested in disease vs. affected system relations. For example, the diseases and genes connected to two target affected systems can be compared, to observe their similarity. The user can collect both the gene and disease information by searching the name of the target affected systems via the search box. In our example, “Increased mean systemic arterial blood pressure” and “decreased systemic arterial diastolic blood pressure” phenotype comparison was made. Related disorders, together with the gene symbols are shown in Table 12. In this example, phenotypes with opposite effects were chosen, and as expected, the overlap between the connected disorders are quite low. The same analysis can be made for comparing two diseases, as well.

In any comparison, connections (disorders, phenotypes or genes) can also provide a quantitative measure for the similarity of the compared terms. This can be achieved by calculating the number of shared nodes between terms. This calculation can yield a similarity measure between 0 and 1, zero meaning no similarity and one meaning 100% similarity. In an example shown in Table 12, the similarity between the target phenotypes are calculated as (2 * 3) / (25 + 16) = 0.146. This similarity calculation can also be formulated by taking the disorder classes and the phenotype hierarchy which would yield a better estimate about the actual similarities.

To test the idea of “similar diseases can have similar phenotypic traits” we carried out various quantitative analyses. First, two diseases “Macrothrombocytopenia” and “Factor X deficiency”, which belong to the same disorder class “hematological” were analyzed. The associated phenotypes for these diseases were extracted and compared. It was observed that there were 86 phenotype annotations for the first disease and 95 for the second one, and 13 of these phenotype annotations were shared between the two target diseases, which led to a similarity score of 0.14. Some of the non-shared phenotype annotations between these diseases could be from the same hierarchy in the MPO DAG, which means that they are similar. As a result, compared phenotype

annotations should not be counted as totally dissimilar if there is a parent-child relationship in-between. To take this into account, we propagated the asserted phenotype annotations for these diseases to high-level phenotypes. This resulted in 13 different high-level phenotype annotations for the “Macrothrombocytopenia” disease and 18 for the “Factor X deficiency” disease. It was observed that 9 of these terms were shared between the target diseases. These high-level phenotype term annotations are shown in Table 13, with star symbols next to the shared ones. The same similarity calculation using the high-level phenotypes yielded 0.58. As observed, the similarity is increased from 0.14 to 0.58 with the inclusion of the phenotype relations into account.

In order to find a more sophisticated way to express similarities, we generated a measure called relative ratios (i.e. relative frequencies) for each high-level phenotype term annotation by calculating what portion of the asserted phenotype annotations for that disease lead to the corresponding high-level phenotype term. For example, it is observed in Table 13 that, nearly 9% of the “Macrothrombocytopenia” disease’s asserted phenotype annotations belong to mortality/aging high-level phenotype class. This calculation is made for all high-level phenotypes for both target diseases. To calculate the total ratio of shared high-level phenotype annotations between these diseases (i.e. those 9 phenotypes marked with a star in Table 13), their relative ratios (i.e. frequencies) are summed. As a result, it can be inferred that 87% (total frequency: 0.87) of the phenotype annotations of “Macrothrombocytopenia” is similar to “Factor X deficiency”, and 68% of the phenotype annotations of “Factor X deficiency” is similar to “Macrothrombocytopenia”.

Table 12: High- level phenotype levels and relative frequencies for the target diseases in the first analysis.

Macrothrombocytopenia	Relative Ratio	Factor X deficiency	Relative Ratio
hematopoietic system phenotype*	0,427272727	hematopoietic system phenotype*	0,214285714
immune system phenotype*	0,172727273	immune system phenotype*	0,150793651
mortality/aging*	0,090909091	homeostasis/metabolism phenotype*	0,119047619
skeleton phenotype*	0,081818182	renal/urinary system phenotype	0,079365079
embryo phenotype	0,036363636	mortality/aging*	0,079365079
liver/biliary system phenotype	0,036363636	cardiovascular system phenotype*	0,079365079
cellular phenotype*	0,027272727	vision/eye phenotype	0,079365079
homeostasis/metabolism phenotype*	0,027272727	nervous system phenotype	0,055555556
respiratory system phenotype	0,027272727	reproductive system phenotype	0,047619048
cardiovascular system phenotype*	0,018181818	digestive/alimentary phenotype	0,023809524
growth/size/body region phenotype*	0,018181818	growth/size/body region phenotype*	0,015873016
integument phenotype	0,018181818	endocrine/exocrine gland phenotype	0,015873016
normal phenotype*	0,009090909	craniofacial phenotype	0,007936508
		pigmentation phenotype	0,007936508
		cellular phenotype*	0,007936508
		normal phenotype*	0,007936508
		skeleton phenotype*	0,007936508

In the second analysis, the same similarity calculations were repeated for diseases belong to different disorder classes which are “Macrothrombocytopenia” in hematological class and “Hearth Block” in “Cardiovascular” class (note that the target diseases chosen in the previous analysis were from the same disorder class), to observe whether the diseases from the same classes would have higher similarities. The direct (i.e. asserted) phenotype annotation comparison gave 0.08 similarity and the high-level class comparison raised this similarity to 0.45. In Table 14, high level phenotype annotations (stars on the shared ones) are displayed with the relative ratios. It was found that 16% (total frequency: 0.16) of the phenotype annotations of “Macrothrombocytopenia” is similar to “Hearth Block”, and 85% of the phenotype annotations of “Hearth Block” is similar to “Macrothrombocytopenia”.

As a result, comparisons of “Macrothrombocytopenia” disease with another disease from the same disorder class and a disease from a different disorder classes yielded high-level similarity values of 0.58 and 0.45, respectively; and relative ratio similarities with 87% and 16%, respectively. This indicates that the diseases from the same disorder class have more similar phenotypic traits, compared to diseases from different disorder classes. Thus, different forms of phenotypic trait similarity calculations can be used as an indicator to compare diseases with each other. In a

disease similarity network such as the one generated in the *Diseasome* study, the edges can be weighted according to phenotypic similarities, to capture the relations more accurately.

Table 13: High- level phenotype levels and ratios for chosen diseases

Macrothrombocytopenia	Relative Ratio	Hearth Block	Relative Ratio
hematopoietic system phenotype	0,427272727	cardiovascular system phenotype*	0,645833333
immune system phenotype	0,172727273	mortality/aging*	0,125
mortality/aging*	0,090909091	muscle phenotype	0,083333333
skeleton phenotype	0,081818182	embryo phenotype*	0,041666667
embryo phenotype*	0,036363636	behavior/neurological phenotype	0,020833333
liver/biliary system phenotype	0,036363636	growth/size/body region phenotype*	0,020833333
cellular phenotype	0,027272727	nervous system phenotype	0,020833333
homeostasis/metabolism phenotype	0,027272727	normal phenotype*	0,020833333
respiratory system phenotype	0,027272727	vision/eye phenotype	0,020833333
cardiovascular system phenotype*	0,018181818		
growth/size/body region phenotype*	0,018181818		
integument phenotype	0,018181818		
normal phenotype*	0,009090909		

The same similarity calculations can be done to compare the phenotype terms, by using disease annotation similarities in-between. For this analysis, fistly two phenotypes that belong to same high-level class (i.e. “oligozoospermia” and “azoospermia”) were compared to each other. After that, two phenotypes that belong to different high-level class phenotypes (i.e. “azoospermia” and “decreased skeletal muscle mass”) were compared to each other.

First, the similarities were calculated using the direct disease term annotation matches between the target phenotypes. The similarity ratio was found as 0.28 and 0.15 for “oligozoospermia” vs. “azoospermia” and for “azoospermia” vs. “decreased skeletal muscle mass”, respectively.

Secondly, the high-level term similarity concept was applied for the phenotype comparisons (similar to the disease comparison analysis done using high-level phenotypic term annotations, mentioned above), in terms of comparing the disorder

classes of the annotated diseases. In Table 15, disorder classes for “oligozoospermia” and “azoospermia” phenotype terms are shown and the common disorder classes between these two phenotype terms are marked with star, and the relative ratios are given. Table 14shows the same results for “azoospermia” vs. “decreased skeletal muscle mass” comparison. similarities were calculated using the high-level term annotations (i.e. disorder classes) and the similarity values were found as 0.88 and 0.67 for the same high-level class phenotypes and for different high-level class phenotypes, respectively. Also, the ratios for the matched disorder classes have been summed up for each phenotype. According to this calculation, “azoospermia” term’s annotated disease similarity to “oligozoospermia” term (the same high-level class phenotype) was found as 100%; whereas, “azoospermia” term’s annotated disease similarity to “decreased skeletal muscle mass” term (a different high-level class phenotype) was found as 77%. These results indicated that, phenotypes from the same high-level class have more similar disease annotations compared to the phenotypes from different high-level classes.

Table 14: Disorder classes of the annotated diseases for “oligozoospermia” and “azoospermia” phenotype terms. Common disorder classes are marked with stars

oligozoospermia	Relative ratio	azoospermia	Relative ratio
Cancer*	0,21686747	Cancer*	0,333333333
Endocrine*	0,13253012	Endocrine*	0,106060606
Metabolic*	0,108433735	Hematological*	0,090909091
Neurological*	0,084337349	Multiple*	0,090909091
Developmental*	0,072289157	Neurological*	0,090909091
Hematological*	0,072289157	Cardiovascular*	0,060606061
Multiple*	0,060240964	Dermatological*	0,045454545
Ophthamological*	0,048192771	Metabolic*	0,045454545
Renal*	0,048192771	Muscular*	0,03030303
Ear,Nose,Throat	0,036144578	Ophthamological*	0,03030303
Bone	0,024096386	Connective Tissue Disorder*	0,015151515
Cardiovascular*	0,012048193	Developmental*	0,015151515
Connective Tissue Disorder*	0,012048193	Gastrointestinal*	0,015151515
Dermatological*	0,012048193	Renal*	0,015151515
Gastrointestinal*	0,012048193	Unclassified*	0,015151515
Muscular*	0,012048193		
Respiratory	0,012048193		
Skeletal	0,012048193		
Unclassified*	0,012048193		

Table 15: Disorder classes of the annotated diseases for “azoospermia” and “decreased skeletal muscle mass” phenotype terms. Common disorder classes are marked with stars.

azoospermia	Relative ratio	decreased skeletal muscle mass	Relative ratio
Cancer*	0,333333333	Cancer*	0,423076923
Endocrine*	0,106060606	Muscular*	0,134615385
Hematological	0,090909091	Multiple*	0,096153846
Multiple*	0,090909091	Neurological*	0,076923077
Neurological*	0,090909091	Endocrine*	0,057692308
Cardiovascular*	0,060606061	Ophthamological*	0,057692308
Dermatological	0,045454545	Bone	0,038461538
Metabolic*	0,045454545	Metabolic*	0,038461538
Muscular	0,03030303	Cardiovascular*	0,019230769
Ophthamological*	0,03030303	Immunological	0,019230769
Connective Tissue Disorder	0,015151515	Nutritional	0,019230769
Developmental	0,015151515	Renal*	0,019230769
Gastrointestinal	0,015151515		
Renal*	0,015151515		
Unclassified	0,015151515		

CHAPTER 5

DISCUSSION

5.1 SUMMARY

This section presents a discussion over the obtained results and a conclusion of this study including its output Mouse2HumanNet web-service, and suggests further improvements and applications for studying human biology using graph theory concepts and network analysis.

- This study aims to create a mapping between human diseases and the abnormal phenotypes observed in mouse experiments. Mouse is the most widely used model organism to study the mammalian physiology and diseases. Therefore, relating the observations from the mouse gene knock-out experiments to human diseases may provide novel information about both the symptoms of diseases and the potential affected systems. This information can be utilized for research in medical diagnostics and for novel treatment options.
- The proposed mapping has been structured as biological networks in two different forms. Genes-Node version network is composed of nodes of diseases, phenotype terms and genes and connections (i.e. edges) indicating direct relations; and ii) Genes Edge version is composed of nodes of diseases and phenotype terms and connections (i.e. edges) in between indicating relations through shared genes. The generated networks are published in an open-access web-service with an easy to use interface, where the users can

display either the whole or the relevant parts of the networks and can download the corresponding information.

- One of the objectives of this project is to aid laboratory scientists to design by selecting the most relevant mouse knock-out models for studying a specific human disease. This can be trivial while studying single gene diseases, as the researcher can directly obtain thee information from relevant biological databases. However, when multiple genes are associated with a specific disease, network approach may provide multiple alternative models.
- Another objective of this study is to enrich the associations between abnormal phenotypes observed in animal studies and human diseases by connecting these two via mouse/human orthologous genes. Novel relations may both aid the studies on medical diagnostics (since the phenotypes contain symptoms) and discovering the systems affected due to a certain disease. In this sense, this project will also aid computational researchers working on ontological systems (e.g. the Human Phenotype Ontology project – HPO) to find and record new disease-phenotype-gene relations. To illustrate this with a very simple example, the difference between the phenotype annotations of human TP53 gene (from the HPO project) can be compared to the annotations of its orthologue in mouse Trp53 (from the MGI project). Both annotation tables are given in Appendix C.1. and its observed from these tables that mouse phenotype annotations are richer compared to human due to extensive animal studies on the mouse. This information may be used to annotate human TP53 with additional phenotypic abnormalities, which in turn can be utilized to aid the development of novel treatments to hereditary/genetic diseases.
- Multiple usage scenarios can be derived for Mouse2HumanNet. For example, a user who is interested in two different genes can do a simple search on our service with the symbol of the corresponding genes. This will return the connected phenotypes and diseases for the target genes. In a hypothetical case that these target genes share high number of phenotype connections, and the first one have a certain disease association but the second one does not. This may lead the user to do more investigation to see if it would be possible for the

second gene be also related to the same disease (relation here is defined as a certain mutation in the corresponding gene would lead to the formation of the corresponding disease), due to high phenotype similarity with the first gene. A similar methodology can be followed to compare two diseases, as well.

- Different types of term similarity approaches have been tested and explained in the Results section. According to these analyses it was observed that, intersections of phenotypic traits can be a good indicator for disease similarities, and similarly, intersections of disease annotations can be a good indicator for phenotypic term similarities. Another finding was that, it would be possible to improve the similarity measures by including the relations between phenotype terms and between the diseases, to the similarity calculation.
- In the light of this information, it can be said that integrating multiple types of information to the similarity calculation would yield more accurate results. In other words, using just one type of information (e.g. only considering the asserted phenotype annotations for comparing 2 diseases) can be misleading due to both the incomplete information in the biological databases and the inconsistencies between the data sources. For example, “Azoospermia” is a disease record under “Endocrine” disorder class in the OMIM database. However, “azoospermia” is also a phenotype term under the “reproductive system phenotype” and “cellular phenotype” high-level phenotypes in the MPO and mostly mapped to the cancer, endocrine and metabolic disorder classes.
- As a case study, we have investigated selected connections of “azoospermia” phenotype from Mouse2HumanNet. There is a connection between “azoospermia” phenotype and “Diabetes Mellitus” disorder, which is interesting to discuss. In order to investigate this connection, we carried out a literature search. According to the literature, Diabetes mellitus (DM) is a chronic disorder that can change carbohydrate, protein, and fat metabolism and caused by the absence of insulin secretion in the body. Obesity is highly correlated with the insulin resistance and pancreatic β-cell dysfunction;

therefore, there is a strong link between obesity and DM (Al-Goblan, Al-Alfi, & Khan, 2014). Especially in obese people, the amount of non-esterified fatty acids, glycerol, hormones, cytokines, and other substances that play a role in the development of insulin resistance, is increased. In women, early stages of obesity take favors the development of menses irregularities, chronic oligo-anovulation and infertility during the adult life. The main factors may be insulin excess and insulin resistance that implicates the association between fertility and the obesity. Furthermore, in men, obesity is correlated with low testosterone levels. Obese individuals usually have reduced spermatogenesis associated with hypotestosteronemia, which can cause infertility (Pasquali, Patton, & Gambineri, 2007). These findings indicate a possible link between infertility in males and DM. However, a more detailed literature search and a structured research study should be conducted to discover whether there is biological mechanism behind it or not.

- Another interesting case study would be considering the connection between the "azoospermia" phenotype and cancer class disorders. There are various surveys in the literature indicating the relations between cancer risks and infertility. Although some studies have found eminent risks for some cancer types connected to infertility, the underlying biological reasons stands unclear. In the study handled by Brinton *et al.*, in 2005, authors found that women diagnosed with infertility have 23% higher risks of uterine and ovarian cancers compared to the control group (Brinton *et al.*, 2005). Furthermore, a retrospective cohort study was carried out to investigate the incidence of chronic medical conditions of men who have infertility (Eisenberg *et al.*, 2016). Men diagnosed with male factor infertility had an important risk of developing chronic conditions such as hypertension, diabetes, hyperlipidemia, renal diseases, pulmonary disease, testis and prostate cancers etc. in the following years (Jacobsen *et al.*, 2000; Walsh *et al.*, 2009 & Walsh *et al.*, 2010; Eisenberg *et al.*, 2015). The findings again suggest a connection between different types of cancer and infertility, which requires immediate mechanistic studies.

- As discussed in the case studies above, it is possible to find previously non-reported links between various genes, phenotypes and diseases in Mouse2HumanNet, which may lead researchers to do a detailed literature search or even to design new experiments to test the biologically interesting links they've observed in our networks. This way, laboratory scientists can benefit from Mouse2HumanNet to select targeted knock-out models to study a specific human disease by observing the genes of interest together with the related phenotypic traits and the affected systems.

5.2 FUTURE DIRECTIONS

- We can divide the future directions of this study into two groups: i) potential projects to infer biological insight using Mouse2HumanNet, and ii) technical modifications to add new functions to the tool that would benefit the users. In terms of the first group of directions, we plan to investigate the generated networks to observe novel for selected genes and diseases. In the wet-lab laboratory of our group (i.e. Cancer Systems Biology Laboratory – CanSyl, METU) the focus is on liver cancers, especially hepatocellular carcinoma, and its related pathways such as the PI3K/AKT/mTOR pathway (Ersahin *et al.*, 2015). One of the aims of CanSyl is investigating novel genes/proteins to target hepatocellular carcinoma and repurposing drugs for this purpose. We plan to employ Mouse2HumanNet to search the associated phenotypic traits to liver cancers and their connected mouse genes. In the case of discovering interesting novel connections, first a literature search will be performed and this may be followed by the construction of an experimental setup to test the candidates. The same methodology can be followed for other cancer group diseases, their phenotypic traits and potential target genes.
- First of the planned technical modifications is the generation of mono-partite disease-disease, gene-gene and phenotype-phenotype similarity networks. Disease-disease and gene-gene networks have previously been proposed in the

Diseasome study (Goh *et al.*, 2007); however, here we plan to weight the edges between genes and between diseases using their hierarchical phenotypic similarities discussed in the results section. This way, more accurate networks can be obtained. Apart from that, phenotype-phenotype similarity networks are proposed for the first time as far as we are aware. These networks will give an idea regarding the similarities between abnormal traits such as symptoms and can be used in research on medical diagnosis.

- A secondary modification for adding new functionalities can be the addition of different types of nodes to the network along with the current disease, phenotype and gene nodes. For this purpose, we plan to add nodes correspond to pathways/systems, terms of the other ontological systems such as the Gene Ontology (GO) and HPO, and drug molecules. The connections between pathways and genes will correspond to membership of those genes in the corresponding pathways. The edges between pathway and disease nodes will mean that those pathways are affected during those diseases. This will add a redundancy to the networks as the MPO phenotype terms also include the system information, however, addition of multiple systems will increase the information coverage. The connections between diseases and drugs will display which drug compounds are currently used to treat which diseases, and the connections between drugs and genes will tell us what are the targeted genes/proteins of those drugs. Finally, addition of other ontological annotations will enrich the information stored in our network. As a result, researchers using our system will find comprehensive information regarding diseases, traits, genes and drugs that are used for treatment.
- The third possible direction is the construction of other networks similar to Mouse2HumanNet, this time using other model organisms and human. These networks can provide further insight regarding the human diseases and their phenotypic reflections, especially where the mouse models remain insufficient. These model organisms will most probably be more distant to the human compared to mouse from an evolutionary point of view; however, the construction of the networks can be done over evolutionarily highly conserved

functions. The comparison of these networks against Mouse2HumanNet would produce interesting results. Candidate model organisms can be animals such as drosophila or even bacteria such as E. coli.

- Another possible technical modification for Mouse2HumanNet web-service may be adding a functionality to display the phenotypes at the desired phenotypic level (i.e. parent or high-level phenotypes instead of asserted/child phenotypes, which is the only options now). This can either be achieved by generating an independent network for each level or just displaying the ancestor terms of a phenotype in the network, when the user clicks or just drags the cursor over the corresponding node. This modification will serve two purposes: i) the number of nodes on the networks will be reduced, which will provide a better perceptibility, and ii) the information about phenotypes will be condensed into more generic classes of phenotypic traits, to be able to analyze the relations on higher systemic levels.
- As the final potential direction, functional associations can be made for diseases, in the form of when function X is lost from the relevant genes due to mutations, disease Y occurs. These kinds of associations could be extremely useful to aid disease mechanism studies. This can be done by selecting all the genes connected to a disease, and carrying out a functional enrichment analysis to observe the properties share between all or at least most of the genes in this list. If the resulting highly enriched property is a pathway, then we can conclude that the corresponding disease is caused by disruptions in this pathway. If the enriched property is a subcellular location, then we can infer that the corresponding disease is particularly effective in that location inside the body. If the enriched property is a biological process GO term, then we can say that the disease could be affecting this high-level system in the organism. In order to test this idea in a small-scale analysis, all of the genes annotated with Leukemia disease were downloaded from Mouse2HumanNet and analyzed with DAVID Functional Annotation Tool (Huang *et al.*, 2009). The enrichment results were investigated and it was observed that many of these genes shared the same or similar annotations. For example, among the most

highly enriched ones were “Acute myeloid leukemia”, “Pathways in cancer” and “hemopoiesis” as expected. Along with those, the terms “protein binding”, “negative regulation of cell proliferation”, and “Acetylation” and many others were also enriched. This may indicate that the corruptions in these functions, due to mutations, may contribute to the appearance of Leukemia. Both the gene list and the significantly enriched annotations can be observed in the tables under Appendix C.2. One option as a future direction would be automatizing this process, to associate highly enriched functional properties with the diseases in the network. This type of analyses can also be used to discover novel candidate disease genes, by finding the other genes (i.e. the genes that were not connected to the corresponding disease in the first place), which were annotated with the corresponding disease associated functional properties.

REFERENCES

Adie, E. a, Adams, R. R., Evans, K. L., Porteous, D. J., & Pickard, B. S. (2005). Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics, 6, 55. http://doi.org/10.1186/1471-2105-6-55

Al-Goblan, A. S., Al-Alfi, M. A., & Khan, M. Z. (2014). Mechanism linking diabetes mellitus and obesity. Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy, 7, 587–591. http://doi.org/10.2147/DMSO.S67400

Austin, C. P., Battey, J. F., Bradley, A., Bucan, M., Capecchi, M., Collins, F. S., … Zambrowicz, B. (2004). The Knockout Mouse Project. Nature Genetics, 36(9), 921–924. http://doi.org/10.1038/ng0904-921

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. Third International AAAI Conference on Weblogs and Social Media, 361–362. http://doi.org/10.1136/qshc.2004.010033

Blackburn, A. C., & Jerry, D. J. (2002). Knockout and transgenic mice of Trp53: what have we learned about p53 in breast cancer? Breast Cancer Research, 4(3), 101. http://doi.org/10.1186/bcr427

Blondel, V.D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008) 'Fast unfolding of communities in large networks', Journal of Statistical Mechanics: Theory and Experiment, 2008(10), p. P10008. doi: 10.1088/1742-5468/2008/10/p10008.

Botstein, D., & Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature Genetics, 33 Suppl(march), 228–37. http://doi.org/10.1038/ng1090

Brinton, L. A., Westhoff, C. L., Scoccia, B., Lamb, E. J., Althuis, M. D., Mabie, J. E., & Moghissi, K. S. (2005). Causes of infertility as predictors of subsequent cancer risk. Epidemiology (Cambridge, Mass.), 16(4), 500–7. http://doi.org/10.1097/01.ede.0000164812.02181.d5

Capecchi, M. (2008). The first transgenic mice: an interview with Mario Capecchi. Interview by Kristin Kain. Disease Models & Mechanisms, 1(4–5), 197–201. http://doi.org/10.1242/dmm.001966

Chaffer, C. L., Dopheide, B., Savagner, P., Thompson, E. W., & Williams, E. D. (2007). Aberrant fibroblast growth factor receptor signaling in bladder and other cancers. Differentiation. http://doi.org/10.1111/j.1432-0436.2007.00210.x

Chen, C. K., Mungall, C. J., Gkoutos, G. V., Doelken, S. C., Kohler, S., Ruef, B. J., … Smedley, D. (2012). Mousefinder: Candidate disease genes from mouse phenotype data. Human Mutation, 33(5), 858–866. http://doi.org/10.1002/humu.22051

Chen, J., Xu, H., Aronow, B.J. and Jegga, A.G. (2007) 'Improved human disease candidate gene prioritization using mouse phenotype', BMC Bioinformatics, 8(1), p. 392. doi: 10.1186/1471-2105-8-392

Cherven, K. (2013). Network graph analysis and visualization with Gephi: Visualize and analyze your data swiftly using dynamic network graphs built with Gephi.

Comfort, A. 1959. Natural aging and the effects of radiation. Radiat. Res. (Suppl. 1): 216-234.

Czechanski, A., Byers, C., Greenstein, I., Schrode, N., Donahue, L.R., Hadjantonakis, A.-K. and Reinholdt, L.G. (2014) 'Derivation and characterization of mouse embryonic stem cells from permissive and nonpermissive strains', Nature Protocols, 9(3), pp. 559–574. doi: 10.1038/nprot.2014.030.

Dailey, L., Ambrosetti, D., Mansukhani, A., & Basilico, C. (2005). Mechanisms underlying differential responses to FGF signaling. Cytokine & Growth Factor Reviews, 16(2), 233–247. http://doi.org/10.1016/j.cytogfr.2005.01.007

Davis, A. P., Murphy, C. G., Johnson, R., Lay, J. M., Lennon-Hopkins, K., Saraceni-Richards, C., … Mattingly, C. J. (2013). The comparative toxicogenomics database: Update 2013. Nucleic Acids Research, 41(D1). http://doi.org/10.1093/nar/gks994

Ersahin, T., Tuncbag, N., & Cetin-Atalay, R. (2015). The PI3K/AKT/mTOR interactive pathway. Molecular Biosystems, 11(7), 1946-1954.

Eisenberg, M. L., Li, S., Cullen, M. R., & Baker, L. C. (2016). Increased risk of incident chronic medical conditions in infertile men: Analysis of United States claims data. Fertility and Sterility, 105(3), 629–636. http://doi.org/10.1016/j.fertnstert.2015.11.011

Eisenberg ML, Li S, Brooks JD, Cullen MR, Baker LC. Increased risk of cancer in infertile men: analysis of U.S. claims data. J Urol 2015; 193:1596–601.

Feenstra, I., Fang, J., Koolen, D. a, Siezen, a, Evans, C., Winter, R. M., … Schinzel, a. (2006). European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations (ECARUCA); an online database for rare chromosome abnormalities. European Journal of Medical Genetics, 49(4), 279–291. http://doi.org/10.1016/j.ejmg.2005.10.13

Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., … Carter, N. P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. American Journal of Human Genetics, 84(4), 524–533. http://doi.org/10.1016/j.ajhg.2009.03.010

Fortunato, S. and Barthélemy, M. (2007) ‘Resolution limit in community detection’, Proceedings of the National Academy of Sciences, 104(1), pp. 36–41. doi: 10.1073/pnas.0605965104.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A.-L. (2007). The human disease network. Proceedings of the National Academy of Sciences of

the United States of America, 104(21), 8685–8690.
http://doi.org/10.1073/pnas.0701361104

Grose, R., & Dickson, C. (2005). Fibroblast growth factor signaling in tumorigenesis. Cytokine & Growth Factor Reviews, 16(2), 179–186.
http://doi.org/10.1016/j.cytogfr.2005.01.003

Guénet, J. L. (2005). The mouse genome. Genome Research.
http://doi.org/10.1101/gr.3728305

Hall, B., Limaye, A., & Kulkarni, A. B. (2009). Overview: Generation of Gene Knockout Mice. Current Protocols in Cell Biology / Editorial Board, Juan S. Bonifacino ... [*et al.*], CHAPTER, Unit–19.1217.
http://doi.org/10.1002/0471143030.cb1912s44

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic acids research, 33(suppl_1), D514-D517.

Han, J.-D.J., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J.M., Cusick, M.E., Roth, F.P. and Vidal, M. (2004) 'Evidence for dynamically organized modularity in the yeast protein-protein interaction network: Abstract: Nature', Nature, 430(6995), pp. 88–93. http://doi: 10.1038/nature02555

Hanneman, R. a., & Riddle, M. (2005). Introduction to social network methods. Network, 149. http://doi.org/10.1016/j.cell.2011.03.009

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics, 6(2), 95–108.
http://doi.org/10.1038/nrg1521

Hoehndorf, R., Schofield, P. N., & Gkoutos, G. V. (2011). PhenomeNET: A whole-phenome approach to disease gene discovery. Nucleic Acids Research, 39(18).
http://doi.org/10.1093/nar/gkr538

Hoehndorf, R., Schofield, P. N., & Gkoutos, G. V. (2015). Analysis of the human *Diseasome* using phenotype similarity between common, genetic, and infectious diseases. Scientific Reports, 5(October 2014), 10888. http://doi.org/10.1038/srep10888

Hu, Y. (2005). Efficient, High-Quality Force-Directed Graph Drawing. Mathematica Journal, 10(1), 37–71. http://doi.org/10.3402/qhw.v6i2.5918

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols, 4(1), 44.

Human, T., & Project, G. (2001). Human Genome Project. Genome, 26(October 1990), 8–10. http://doi.org/10.1080/19371918.2011.579488

Jacobsen R, Bostofte E, Engholm G, Hansen J, Olsen JH, Skakkebaek NE, *et al.* Risk of testicular cancer in men with abnormal semen characteristics: cohort study. BMJ 2000; 321:789–92.

Jensen, R. A. (2001). Orthologues and paralogs - we need to get it right. Genome Biology, 2(8), interactions1002.1–interactions1002.3.

Jin, E. M., Girvan, M., & Newman, M. E. J. (2001). Structure of growing social networks. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, 64(4 II), 461321–461328. http://doi.org/10.1103/PhysRevE.64.046132

Jordan, I. K., Mariño-Ramírez, L., & Koonin, E. V. (2005). Evolutionary significance of gene expression divergence. Gene, 345(1 SPEC. ISS.), 119–126. http://doi.org/10.1016/j.gene.2004.11.034

Justice, M. J., & Dhillon, P. (2016). Using the mouse to model human disease: increasing validity and reproducibility. Disease Models & Mechanisms, 9(2), 101–103. http://doi.org/10.1242/dmm.024547

Kann, M. G. (2009). Advances in translational bioinformatics: Computational approaches for the hunting of disease genes. Briefings in Bioinformatics, 11(1), 96–110. http://doi.org/10.1093/bib/bbp048

Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., … Robinson, P. N. (2014). The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. Nucleic Acids Research, 42(D1). http://doi.org/10.1093/nar/gkt1026

Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., … Robinson, P. N. (2014). The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. Nucleic Acids Research, 42(D1). http://doi.org/10.1093/nar/gkt1026

Köhler, S., Doelken, S.C., Ruef, B.J., Bauer, S., Washington, N., Westerfield, M., Gkoutos, G., Schofield, P., Smedley, D., Lewis, S.E., Robinson, P.N. and Mungall, C.J. (2013) 'Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research', F1000Research, http://doi: 10.12688/f1000research.2-30. v1.

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., … Robinson, P. N. (2009). Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. American Journal of Human Genetics, 85(4), 457–464. http://doi.org/10.1016/j.ajhg.2009.09.003

Labs, N. (2012) Information epidemics and synchronized viral social contagion. Available at: http://noduslabs.com/research/information-epidemics-viral-social-contagion/ (Accessed: 17 February 2017).

Lek, M., Karczewski, K., Minikel, E., Samocha, K., Banks, E., … MacArthur, D. (2016). Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv. Retrieved from http://biorxiv.org/content/early/2016/05/10/030338.abstract

LESLIE, N. R., & DOWNES, C. P. (2004). PTEN function: how normal cells control it and tumour cells lose it. Biochemical Journal, 382(1), 1–11. http://doi.org/10.1042/BJ20040825

Liao, B. Y., & Zhang, J. (2006). Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol, 23(3), 530–540. http://doi.org/10.1093/molbev/msj054

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2011). Entrez gene: Gene-centered information at NCBI. Nucleic Acids Research, 39(SUPPL. 1). http://doi.org/10.1093/nar/gkq1237

Majzoub, J.A. and Muglia, L.J. (1996) 'Knockout mice',New England Journal of Medicine, 334(14), pp. 904–906. doi: 10.1056/nejm199604043341407.

Newman, M., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical Review E, 69(2), 1–16. http://doi.org/10.1103/PhysRevE.69.026113

Nuutila, E., & Soisalon-Soininen, E. (1994). On finding the strongly connected components in a directed graph. Information Processing Letters, 49(1), 9-14. doi:10.1016/0020-0190(94)90047-7

Orphanet. (2014). Orphanet Report Series Prevalence of rare diseases: Bibliographic data. Rare Diseases, (1), 1–29.

Pal, C., Papp, B., & Hurst, L. D. (2001). Highly Expressed Genes in Yeast Evolve Slowly. Genetics, 158(2), 927–931. Retrieved from http://http//www.genetics.org%5Cnpapers://d2952c50-9509-4ba2-9a03-22fbc04267d4/Paper/p639

Pasquali, R., Patton, L., & Gambineri, A. (2007). Obesity and infertility. Current Opinion in Endocrinology, Diabetes, and Obesity, 14(6), 482–7. http://doi.org/10.1097/MED.0b013e3282f1d6cb

Pereira, C. S., Oliveira, M. V. M. De, Barros, L. O., Bandeira, G. A., Santos, S. H. S., Basile, J. R., … De Paula, A. M. B. (2013). Low expression of MSH2 DNA repair protein is associated with poor prognosis in head and neck squamous cell carcinoma. Journal of Applied Oral Science : Revista FOB, 21(5), 416–21. http://doi.org/10.1590/1679-775720130206

Piñero, J., Queralt-Rosinach, N., Bravo, À., Deu-Pons, J., Bauer-Mehren, A., Baron, M., … Furlong, L. I. (2015). DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. Database, 2015. http://doi.org/10.1093/database/bav028

Riggs, E. R., Jackson, L., Miller, D. T., & Van Vooren, S. (2012). Phenotypic information in genomic variant databases enhances clinical care and research: The international standards for cytogenomic arrays consortium experience. Human Mutation, 33(5), 787–796. http://doi.org/10.1002/humu.22052

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., & Mundlos, S. (2008). The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. American Journal of Human Genetics, 610–615. http://doi.org/10.1016/j.ajhg.2008.09.017

Rocha, E. P. C., & Danchin, A. (2004). An Analysis of Determinants of Amino Acids Substitution Rates in Bacterial Proteins. Molecular Biology and Evolution, 21(1), 108–116. http://doi.org/10.1093/molbev/msh004

Rodriguez-Caso, C., Medina, M. A., & Solé, R. V. (2005). Topology, tinkering and evolution of the human transcription factor network. FEBS Journal, 272(24), 6423–6434. http://doi.org/10.1111/j.1742-4658.005.05041.x

Rosenthal, N., & Brown, S. (2007). The mouse ascending: perspectives for human-disease models. Nature Cell Biology, 9(9), 993–999. http://doi.org/10.1038/ncb437

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y. W. W., Mazaitis, M., Felix, V., … Kibbe, W. A. (2012). Disease ontology: A backbone for disease semantic integration. Nucleic Acids Research, 40(D1). http://doi.org/10.1093/nar/gkr972

Seymour, P., Schrijver, A., & Diestel, R. (2005). Graph Theory. Oberwolfach Reports. http://doi.org/10.4171/OWR/2005/03

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., … Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology, 25(11), 1251–1255. http://doi.org/10.1038/nbt1346

Smith, C.L. and Eppig, J.T. (2009) 'The mammalian Phenotype Ontology: Enabling robust annotation and comparative analysis', 1(3).

Thakurela, S., Tiwari, N., Schick, S., Garding, A., Ivanek, R., Berninger, B., & Tiwari, V. K. (2016). Mapping gene regulatory circuitry of Pax6 during neurogenesis. Cell Discovery, 2, 15045. http://doi.org/10.1038/celldisc.2015.45

Vandamme, T. F. (2014). Use of rodents as models of human diseases. J Pharm

Bioallied Sci, 6(1), 2–9. http://doi.org/10.4103/0975-7406.124301

Vulto-van Silfhout, A. T., van Ravenswaaij, C. M. A., Hehir-Kwa, J. Y., Verwiel, E. T. P., Dirks, R., van Vooren, S., … De Leeuw, N. (2013). An update on ECARUCA, the european cytogeneticists association register of unbalanced chromosome aberrations. European Journal of Medical Genetics, 56(9), 471–474. http://doi.org/10.1016/j.ejmg.2013.06.010

Walsh TJ, Croughan MS, Schembri M, Chan JM, Turek PJ. Increased risk of testicular germ cell cancer among infertile men. Arch Intern Med 2009;169:351–6.

Walsh TJ, Schembri M, Turek PJ, Chan JM, Carroll PR, Smith JF, *et al.* Increased risk of high-grade prostate cancer among infertile men. Cancer 2010;116:2140–7.

Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M., & Lewis, S. E. (2009). Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biology, 7(11). http://doi.org/10.1371/journal.pbio.1000247

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Mouse Genome Sequencing, C. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature, 420(6915), 520–562. http://doi.org/10.1038/nature01262

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., … Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research, 42(D1). http://doi.org/10.1093/nar/gkt1229

White, J. K., Gerdin, A., Karp, N. a, Ryder, E., Buljan, M., Bussell, J. N., … Steel, K. P. (2013). Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for various genes. Cell, 154(2), 452–64. http://doi.org/10.1016/j.cell.2013.06.022

Winslow, T. (2017) Embryonic stem cells. Available at: http://stemcells.nih.gov/info/Regenerative_Medicine/2006Chapter1.htm (Accessed: 11 January 2017).

Zhang, R., Ou, H.-Y., & Zhang, C.-T. (2004). DEG: a database of essential genes. Nucleic Acids Research, 32(Database issue), D271-2. http://doi.org/10.1093/nar/gkh024

Zhang, J., & He, X. (2005). Significant impact of protein dispensability on the instantaneous rate of protein evolution. Molecular Biology and Evolution, 22(4), 1147–1155. http://doi.org/10.1093/molbev/msi101

http://www.pnas.org/content/suppl/2007/05/03/0701361104.DC1/01361Table1.pdf

http://www.pnas.org/content/suppl/2007/05/03/0701361104.DC1/01361Table2.pdf

http://www.pnas.org/content/suppl/2007/05/03/0701361104.DC1/01361Table3.pdf

APPENDICES

APPENDIX A

USER MANUAL FOR THE MOUSE2HUMANNET WEB-SERVICE

Mouse2HumanNet

Mouse2HumanNet is an open source of bioinformatics platform / web-service for the visualization and manipulation of the networks indicating the relationships between diseases, disease causing genes, and abnormal phenotypic traits in mouse and human organisms. In this network, diseases, phenotype terms and genes correspond to nodes (genes are modeled as edges in the alternative version of the network) and the pairwise relationships between these entities correspond to edges between the nodes. The first one of the networks is called “Genes-Node version network” and it can be accessed from the link: “https://nilaycan.github.io/mousepheno”. The second version, which is called “Genes-Edge version network” can be accessed from the link: “https://nilaycan.github.io/mousepheno/edges/”.

Mouse2HumanNet Manual

1-) Introduction

Mouse2HumanNet is a JavaScript based web-service to visualize the relations between human diseases, human/mouse orthologues genes and abnormal phenotypes (i.e. affected systems). The web based viewer is a modified version of Gexf js master tool ((Velt, R. (2011), Gexf-js Gephi visualisation plugin, Github. Retrieved [04.06.2017] from [https://github.com/raphv/gexf-js]). Users may find various options on the interface such as zooming in and out, magnifying and displaying only selected nodes and its connections using the sub-network selector, a search bar to type the names of nodes, a small network (shown at the bottom right of the screen) functioning as a navigator to help user to find the current location while zoomed in, a color code table displaying different types of nodes in the network together with the number of nodes for the corresponding node types, ability to temporarily highlight the connections of nodes by dragging cursor onto them (without clicking), exporting the connections of any chosen node in .xlsx format and exporting the selected sub-networks in .png format. The interface of Mouse2HumanNet is shown in Figure 45 with explanations.

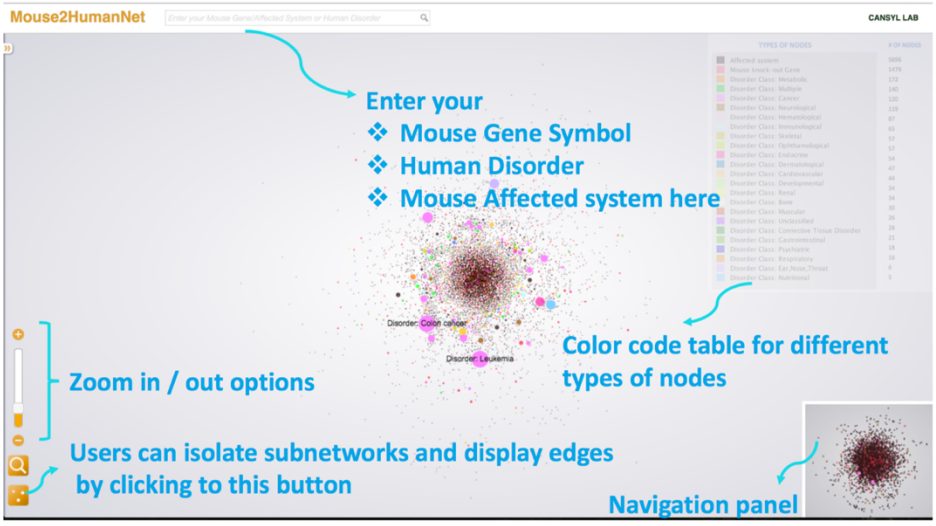


Figure 45: Interface of Mouse2HumanNet (Genes-Node version network).

2-) Interpretation of Mouse2HumanNet Interface

The first property that will be explained in this section is the search bar. When any letters are typed into the search field, all possible results (in terms of the node names) are listed in a window. Search column is not case sensitive. In Figure A.2 and Figure A.3, nodes names are displayed for example searches.

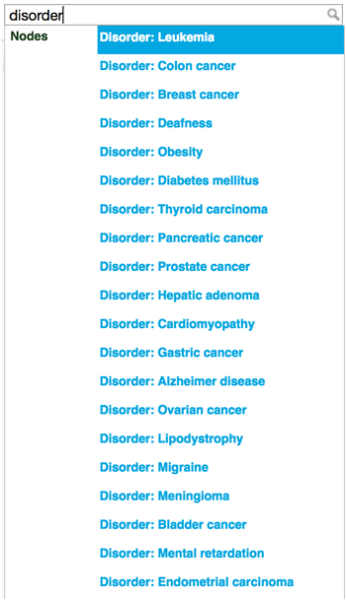


Figure A.2: Diseases are displayed with “disorder” term at the beginning.

Figure A.3: Affected system (left) and gene (right) names as they appear in search window.

In Genes-Node version network (http://nilaycan.github.io/mousepheno), all biological entities (i.e. diseases, phenotypes and genes) are represented as nodes. When any mouse/human orthologue gene symbol is searched and the corresponding node is selected by clicking the gene symbol, the user can see its connected diseases and affected systems as listed on the left pane. If a disease term is typed into the search column and the corresponding node is selected, its related genes are revealed on the left pane. Also, if an affected system was searched and selected, its related genes are listed on the left pane. For any selected disease node, Total number of connected nodes can be seen under degree information on the left pane. The connections of a selected node can be exported with the “export nodes” button in .xlsx format. Furthermore, the displayed network can be exported in .png format with “export png” button.

The size (area) of the nodes is proportional to the number of connections it possesses. The colors of the nodes are given according to the color table at the right side of the screen (Figure 9). There are 24 distinct node types in the Genes-Node network, which are mouse/human orthologues genes, mouse phenotypes (i.e. affected systems) and 22 different human disorder classes. Mouse genes and affected systems are coded as red and black, respectively.

The edge weights are given to only some of the node types. The weights between disorders and genes are constant; however, the weights between genes and affected systems change according to the number of diseases that gene is connected to, and visualized as edge thicknesses. Constant edge weights between the diseases and genes in Genes-Node network is illustrated under Figure A.4 and variable edge weights between affected systems and genes are shown in Figure A.5.

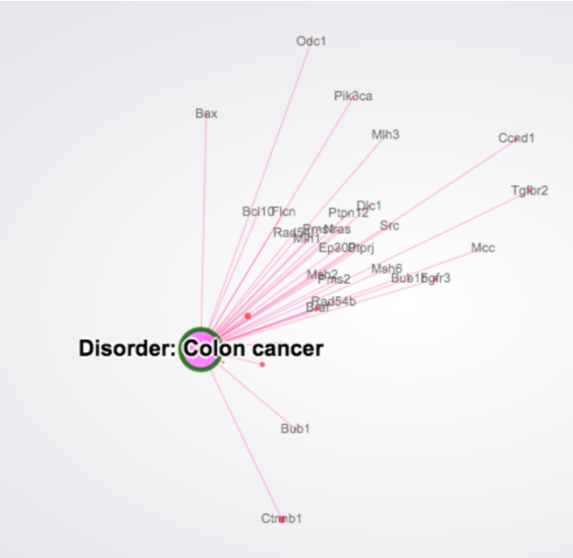


Figure A.4: Edge weights between genes and diseases in Genes-Node network.

Figure A.5: Edge weight between affected systems and genes

In Genes-Edge version (https://nilaycan.github.io/mousepheno/edges/), affected systems and diseases are represented as nodes and genes correspond to edges that connect affected system and disease nodes. Users can only search diseases and affected systems in this version. When any disease term is typed into the search column and

the corresponding node is selected, its connected affected systems are displayed on the left pane and the connected gene names appear in a list. Also, if any affected system is searched and the node is selected, its related diseases are displayed.

There are 23 distinct node types in the Genes-Edge network, which are mouse phenotypes / affected systems and 22 different human disorder classes. Affected systems and genes are coded with black and red colors, respectively. Disorder and affected system node sizes are selected with respect to their corresponding number of connections.

Different edge weights are applied according to the number of genes shared between a disease and an affected system. The edge weight property is visualized for "premature death" phenotype in Figure A.6.

Figure A.6: Edge weights between premature death phenotype and its connected diseases.

3-) A use-case example

Suppose a user is interested in breast cancer and starts typing the disease name in the search box in the Genes-Node version network. This will reveal the related disorder names in a window, from where the user selects “Disorder: Breast cancer” by clicking on it. This will open the page for the corresponding disease, where the Disease Id, disorder class, the degree (total number of connected nodes) and the symbols of the connected genes are listed on the left pane (Figure A.7, left). The user isolates only breast cancer and its connections on the network by clicking the subnetwork button (otherwise, the whole network will be displaed in the background). The user can also temporarily visualize the connections of genes (in terms of phenotypes and other diseases) that is connected to breast cancer node, by dragging cursor onto the connected gene nodes without clicking. For example, while not clicking any node, dragging the cursor over the “Trp53” gene node will temporarily visualize its connections (Figure A.8). So, it can be said that the temporarily displayed connections (i.e. phenotypes and other diseases connected to Trp53) are in-directly connected to the breast cancer node, which can provide additional insight while investigating breast cancer.

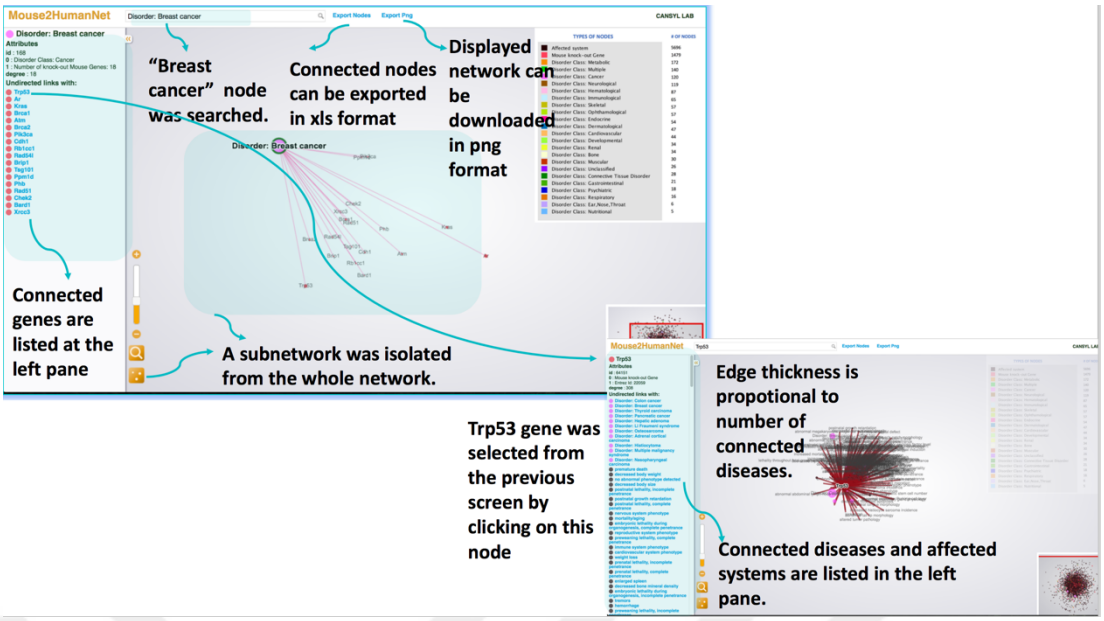


Figure A.7: "Disorder: Breast Cancer" page (left), "Trp53" gene page (right) in the Genes-Node version network.

Figure A.8: Temporarily visualizing the connections of Trp53 gene node while on the ''Disorder: Breast Cancer'' node.

When any node on the screen is clicked, the dedicated connections of the clicked node will appear on the screen. If the selected node is a gene in Genes-Node network, its connected affected systems and diseases will appear on the left pane. This is not the case for selected diseases and affected systems since these node types are not directly connected to each other on the Genes-Node network (they are directly connected in the Genes-Edge version). In Figure A.7 (right side), "Trp53" (orthologue of TP53 human gene) node was clicked from the list of breast cancer's connections. Users can see the node type, Entrez Id and the degree information of the chosen gene node.

Edge thickness between genes and affected systems change according to the number of diseases connected to that gene. This is illustrated in Figure A.9 (top screenshot), as the edges are thicker between "Trp53" and its connections since "Trp53" is densely connected to diseases. In figure A.9 (bottom screenshot), edge thicknesses between "premature death" phenotype and gene nodes change with respect to the number of diseases that gene is connected to. This property indicates how critical a gene is, in terms of disease relations.

In our example so far, the user has found out that Trp53 gene (human orthologue: TP53) is critical for breast cancer disorder and Trp53/TP53 is a hub gene that has been associated with many diseases along with breast cancer. Now the user can move on to investigate phenotypic traits for this gene. "Premature death" is an abnormal phenotype that is associated with Trp53, as shown in the left pane (the first black colored node on the list in Figure A.9, top screenshot). To have better idea, the user clicks "Premature death" link on the left pane which directs to the dedicated sub-network for this phenotype (Figure A.9, bottom screenshot). The user now can observe the other gene nodes connected to "Premature death" phenotype, both on the left pane as a list, and as a network on the main window. The inference here is that, "Premature death" can be caused by many other genes along with Trp53. The user can further move on with selecting another interesting gene. Up to this point, one simple conclusion is that, the formation of breast cancer can be related to mutations in Trp53/TP53 gene, which can lead to premature death.

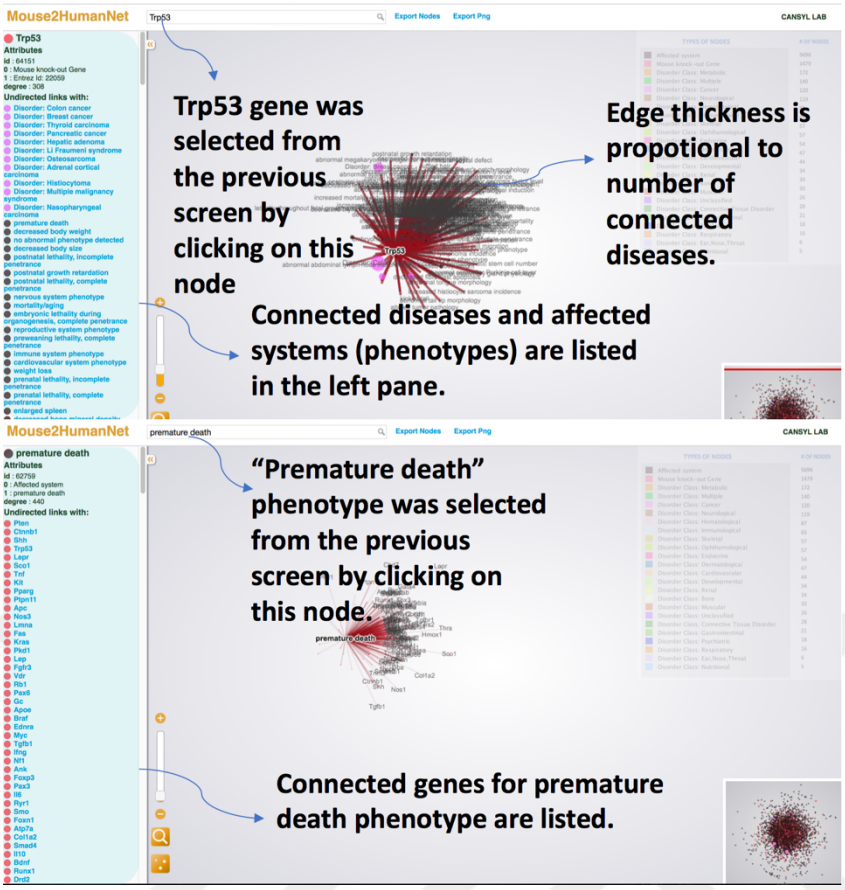


Figure A.9: "Trp53" gene (top) and "Premature death" affected system (bottom) pages in Genes-Node version network.

Figure A.10 displays the interface for Genes-Edge version network. In this version, users can only search human disorders and mouse affected systems, because the genes are coded as edges and not as clickable as nodes, in this version.

As an example, case, "Disorder: Papillary serous carcinoma of the peritoneum" disease was typed into search box and its connections are shown (Figure A.11, left). Disease Id, disorder class, total number of connected nodes to "Papillary serous carcinoma of the peritoneum" as degree and the connecting genes' symbols (serving as edges in this network) are listed on the left pane. The user isolates only this disease's connections by clicking subnetwork button. The only connected gene is Brca1. Among the many listed connected phenotypes, "Uterus hyperplasia" was selected by clicking the

corresponding link (i.e. the name of the phenotype) as shown on the right side of Figure A.11. Now, the related diseases of "Uterus hyperplasia" phenotype (including Papillary serous carcinoma of the peritoneum) appears on the left pane. Here, the edge thicknesses between disease and affected system nodes change according to the total number of genes shared between them. At this point, the user can directly observe which diseases are related to Papillary serous carcinoma of the peritoneum over the connections with Uterus hyperplasia phenotype. At a very basic level, the user has learnt that the disease "Papillary serous carcinoma of the peritoneum" may cause "Uterus hyperplasia", which is the increased uterus size, and the biological mechanism behind this process may lie within certain mutations in the BRCA1 gene.

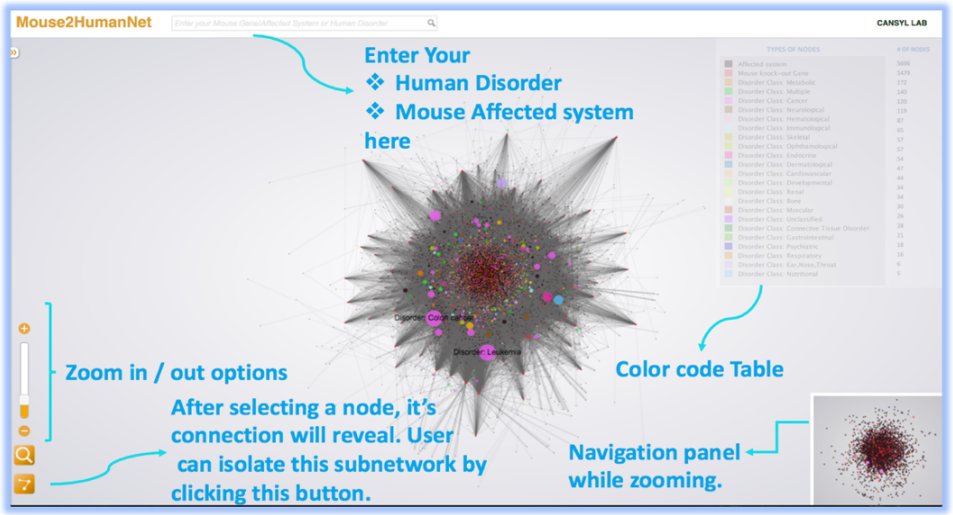


Figure: A.10: Interface of the Genes-Edge version network in Mouse2HumanNet with explanations for features and options.

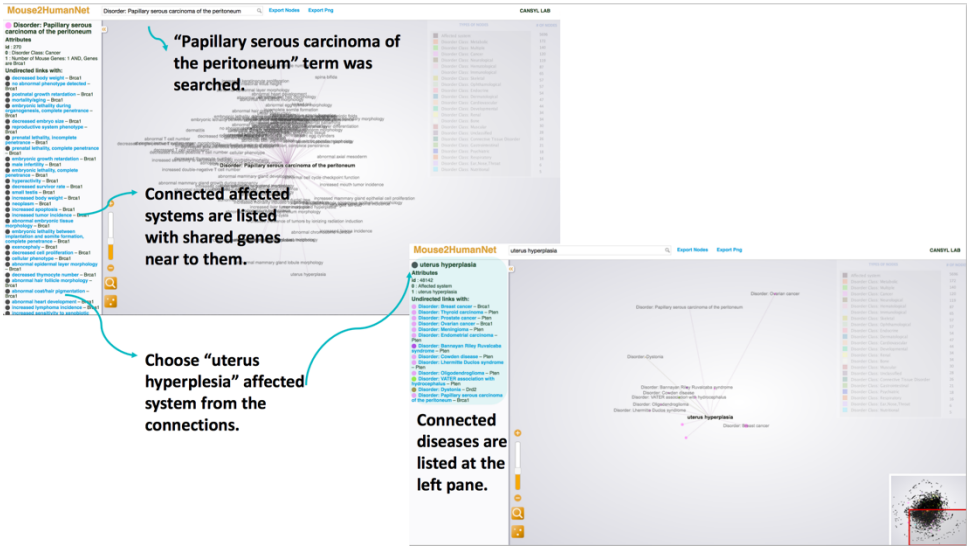


Figure A.11: "Papillary serous carcinoma of the peritoneum" disease's interface (left), and "Uterus hyperplasia" phenotype's interface (right) on the Genes-Edge version network.

APPENDIX B

SOURCE CODES

B.1 GENES-EDGE VERSION

```
import os
import sys
import time
import math
import xlrd
from xlrd import open_workbook
import random
from sys import stdout
from time import sleep
#width2height=2.0
area_density=0.05
def bgr(minimum, maximum, value):
    minimum, maximum = float(minimum), float(maximum)
    ratio = 2 * (float(value)-minimum) / (maximum - minimum)
    b = int(max(0, 255*(1 - ratio)))
    r = int(max(0, 255*(ratio - 1)))
    g = 255 - b - r
    bgr=[]
    bgr.append(b)
```

```
    bgr.append(g)

    bgr.append(r)

    return bgr

def write_xml_header(file):

    file.write('<?xml    version="1.0"    encoding="UTF-
8"?>\n')

    #file.write('<gexf  xmlns="http://www.gephi.org/gexf"
xmlns:viz="http://www.gephi.org/gexf/viz">\n')

    file.write('<gexf
xmlns="http://www.gexf.net/1.2draft"        version="1.2"
xmlns:viz="http://www.gexf.net/1.2draft/viz"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.gexf.net/1.2draft
http://www.gexf.net/1.2draft/gexf.xsd">\n')

    #file.write('<graph type="static">\n')

    file.write('       <meta   lastmodifieddate="2014-01-
30">\n')

    file.write('    <creator>Gephi 0.8.1</creator>\n')

    file.write('    <description></description>\n')

    file.write('  </meta>\n')

    file.write('      <graph   defaultedgetype="directed"
mode="static">\n')

    file.write('              <attributes   class="node"
mode="static">\n')

    file.write('          <attribute  id="0"  title="Type"
type="string"/>\n')

    #file.write('        <attribute id="1" title="disease
class" type="string"/>\n')

    file.write('        <attribute id="1" title="Remarks"
type="string"/></attributes>\n')

    #file.write('<attributes                class="node"
type="static">\n')

    #file.write('<attribute      id="0"      title="type"
type="string"/>\n')
```

```
    #file.write('<attribute    id="1"    title="disclass"
type="string"/>\n')

    #file.write('<attribute    id="2"    title="Polygon"
type="integer"/>\n')

    #file.write('</attributes>\n')

    file.write('    <nodes>\n')

def
write_a_node(file,_node_id,_label,_att1,_att2,_pos,_col,_
size):

    file.write('                <node   id="'+_node_id+'"
label="'+_label+'"\n>')

    file.write('        <attvalues>\n')

    #if _att1=="gene":

    #    att1="Mouse Gene"

    #if _att1=="humandisease":

    #    att1="Disease of Class: "+

    #    _att3="3"

    file.write('                          <attvalue  id="0"
value="'+_att1+'"></attvalue>\n')

    file.write('                          <attvalue  id="1"
value="'+_att2+'"></attvalue>\n')

    #file.write('                          <attvalue  id="2"
value="'+_att3+'"></attvalue>\n')

    file.write('        </attvalues>\n')

    file.write('            <viz:position  x="'+_pos[0]+'"
y="'+_pos[1]+'" z="0.0"></viz:position>\n')

    file.write('              <viz:color  b="'+_col[0]+'"
g="'+_col[1]+'" r="'+_col[2]+'"></viz:color>\n')

    file.write('                                  <viz:size
value="'+_size+'"></viz:size>\n')

    #file.write('<viz:shape value="triangle"/>\n')

    file.write('      </node>\n')
```

```
    #file.write('\n')

print "undirected graph generation human diseases to child
affected names via genes"

#can: the name of the workbook, be caution about the format
of the book

first_book_name="first_table_corrected_new.xls"

first_book = xlrd.open_workbook(first_book_name)

#can: the name of the active worksheet

ws=first_book.sheet_by_name("Sheet1")

#can:

#can: a python dictionary conneting the genes (as keys=)
to the humandiseases (=as values)

#gene2humandis={}

humandis2gene={}

dis_class={}

#can: flag for the while loop to recognize end of the excel
sheet

first_line=True

#can: the start row

current_row=1

gene2entrez={}

while first_line:

    try:

        current_entrezid=ws.cell(current_row,0).value

    except:

        first_line=False

    else:

        try:


current_gene=ws.cell(current_row,1).value.strip()
```

```
        except:
            pass
        else:
            gene2entrez[current_gene]=current_entrezid

current_humandis=ws.cell(current_row,3).value.strip()
            current_disclass=ws.cell(current_row,4).value
            dis_class[str(current_disclass)]=1.0

#current_tuple=(current_humandis,current_disclass,current
_row)

current_tuple=(current_gene,current_disclass,current_row)
            #if current_gene not in gene2humandis.keys():
            #    gene2humandis[current_gene]=[]

#gene2humandis[current_gene].append(current_tuple)
            if       current_humandis       not       in
humandis2gene.keys():
                humandis2gene[current_humandis]=[]

humandis2gene[current_humandis].append(current_tuple)
        current_row=current_row+1
for l in range(len(dis_class.keys())):
    dis_class[dis_class.keys()[l]]=float(l)
second_book_name="second_with_parenting.xlsx"
second_book = xlrd.open_workbook(second_book_name)
ws=second_book.sheet_by_name("DENE")
gene2affected={}
mp2id={}
second_line=True
```

```
prev_row=current_row
current_row=1
while second_line:
    try:
        current_mp=ws.cell(current_row,0).value
    except:
        second_line=False
        print "here you have some problem"+str(current_row)
    else:
        try:
current_gene=ws.cell(current_row,2).value.strip()
        except:
            pass
        else:
            current_affected=ws.cell(current_row,1).value
#current_parent1_mp=str(ws.cell(current_row,3).value)
#current_parent1_affected=str(ws.cell(current_row,4).value)
      #current_parent2_affected="noparent"
            #try:
            #current_parent2_mp=str(ws.cell(current_row,5).value)
            #except:
      #   current_parent2_mp="noparent"
            #else:
```

```
#current_parent2_affected=str(ws.cell(current_row,6).valu
e)


#affected_tuple=(current_affected,prev_row+current_row,cu
rrent_parent1_mp,current_parent1_affected,current_parent2
_mp,current_parent2_affected)


affected_tuple=(current_mp,current_affected)#,current_par
ent1_mp,current_parent1_affected,current_parent2_mp,curre
nt_parent2_affected)


            if current_gene not in gene2affected.keys():
                gene2affected[current_gene]=[]

gene2affected[current_gene].append(affected_tuple)
            if current_gene=="Flt3":
                print "here:"+str(affected_tuple)
            mp2id[current_mp]=prev_row+current_row
            current_row=current_row+1
print "numberofrows:"+str(current_row)


gene2geneId={}
prev_row=max(mp2id.values())+1
current_row=1
for gene in gene2affected.keys():
    gene2geneId[gene]=prev_row+current_row
    current_row=current_row+1


no_notaffected_genes=0
max_radius=0
max_disease="some"
```

```
total_area=0.0

for gene_tuple_array in humandis2gene.values():

    for gene_tuple in gene_tuple_array:


current_gene,current_disclass,current_row=gene_tuple


current_gene,current_disclass,current_row=str(current_gen
e),str(current_disclass),str(current_row)

        #print current_gene,gene_tuple

        try:

            x=len(gene2affected[current_gene])

        except:

            no_notaffected_genes=no_notaffected_genes+1

        else:

            #if max_radius<x:

            #    max_radius=x

            #    for humandis in gene2humandis[gene]:

            #        max_disease, dummy1, dummy2=humandis

            total_area=total_area+float(x)*float(x)*3.14

#print "the maximum radius and the corresponding disease
(number of affected connected to disease) is:
"+str(max_radius)+" of "+max_disease

print "total area has been found to be: "+ str(total_area)

#print no_notaffected_genes

#compute heigth

r=math.sqrt(total_area/area_density/3.14)

print "the maximum radius of the window has been found to
be "+str(r)

outfile=open("dis2affected.gexf","w")

write_xml_header(outfile)
```

```
debugfile=open("debugfile.log","w")
nodedic={}
missing=[]
for humandisease in humandis2gene.keys():
    prev_missing_length=len(missing)
    current_size=0.0
    current_gene_list = []
    for gene_tuple in humandis2gene[humandisease]:

current_gene,current_disclass,current_row=gene_tuple
        current_gene_list.append(current_gene)
        try:

current_size=current_size+len(gene2affected[current_gene]
)
        except:
            missing.append(current_gene)
    current_size=str(current_size)
    first_gene_tuple=humandis2gene[humandisease][0]

current_gene,current_disclass,current_row=first_gene_tupl
e
    current_gene,                              current_att2,
current_node_id=str(current_gene),str(current_disclass),s
tr(current_row)
    current_label="Disorder: "+humandisease
    current_att1="Disorder Class: "+current_disclass
    current_att2="Number      of      Mouse      Genes:
"+str(len(humandis2gene[humandisease]))+"  AND,  "+"Genes
are " + ', '.join(current_gene_list)
    #angle=random.uniform(-3.14*2,3.14*2)
    #radius=random.uniform(0,r)
```

```
#current_pos=str(radius*math.cos(angle)),str(radius*math.
sin(angle))


#current_col=bgr(0,len(dis_class.keys()),dis_class[curren
t_att2])


#current_col[0],current_col[1],current_col[2]=str(current
_col[0]),str(current_col[1]),str(current_col[2])

    seperator="**##"


writenodestring=current_node_id+seperator+current_label+s
eperator+current_att1+seperator+current_att2+seperator+cu
rrent_size


nodedic[current_node_id+seperator+current_label]=writenod
estring



print      "Total      #      of      nodes      appended:
"+str(len(nodedic.keys()))

print    "#    of    nodes    without    MP    Id    appended:
"+str(len(missing))

for k in missing:

    debugfile.write(k+"\n")



from collections import Counter

from itertools import chain



def rowToPairs(sheet, row):

  """covert a sheet row to (affected_system, disease) pairs"""

    affected_system = sheet.cell(row, 1).value.strip()

    diseases    =    [d.strip()    for    d    in    sheet.cell(row,
3).value.split(',')]
```

```
    aff_sys_disease_pairs = [(affected_system, disease)
for disease in diseases]

    return aff_sys_disease_pairs


def sheet_to_pairs(sheet):

  """convert the sheet to (affected_system, disease) pairs iterable"""

    return (rowToPairs(sheet, row) for row in range(0,
sheet.nrows))


def count_affected_in_sheet(sheet):

    unique_pairs                                        =
set(chain.from_iterable(sheet_to_pairs(sheet)))

    return Counter(aff_sys for (aff_sys, disease) in
unique_pairs)


#                          doc                          =
open_workbook('second_disease_added.xlsx').sheet_by_index
(0)

counter = count_affected_in_sheet(ws)

for humandisease in humandis2gene.keys():

    for gene_tuple in humandis2gene[humandisease]:

        gene,current_disclass,current_row=gene_tuple

        try:

            dummy=len(gene2affected[gene])

        except:

            pass

        else:

            for affected in gene2affected[gene]:

                #current_label,
current_node_id,mp1,affected1,mp2,affected2=affected
```

```
                current_mp,current_affected=affected
                # current_size=str(50.0)
                current_size                                    =
str(counter.get(current_affected))
        #current_att1="Child affected system"
                current_att1="Affected system"
                current_att2=current_affected
                angle=random.uniform(-3.14*2,3.14*2)
                radius=random.uniform(0,r)

current_pos=str(radius*math.cos(angle)),str(radius*math.s
in(angle))
                current_col=str(68),str(68),str(238)
                seperator="**##"

current_target_node_id=str(mp2id[current_mp])
                current_affected=current_att2

writenodestring=current_target_node_id+seperator+current_
affected+seperator+current_att1+seperator+current_att2+se
perator+current_size

nodedic[current_target_node_id+seperator+current_mp]=writ
enodestring
                '''
                try:

current_parent1_target_node_id=str(mp2id[mp1])
                except:
                    pass
                else:
                    angle=random.uniform(-3.14*2,3.14*2)
```

```
                    radius=random.uniform(0,r)

current_pos=str(radius*math.cos(angle)),str(radius*math.s
in(angle))
                    current_col=str(68),str(68),str(238)
                    seperator="**##"
                    current_att1="Parent affected system"
                    current_att2=affected1
                    mp1="Parent     Affected     System1:
"+mp1.strip('\"')

writenodestring=current_parent1_target_node_id+seperator+
mp1+seperator+current_att1+seperator+current_att2+seperat
or+current_size

nodedic[current_parent1_target_node_id+seperator+mp1]=wri
tenodestring

#write_a_node(outfile,current_parent1_target_node_id,mp1,
current_att1,current_att2,current_pos,current_col,current
_size)
                if mp2!="noparent":
                    try:

current_parent2_target_node_id=str(mp2id[mp2])
                    except:
                        pass
                    else:
                        angle=random.uniform(-
3.14*2,3.14*2)
                        radius=random.uniform(0,r)

current_pos=str(radius*math.cos(angle)),str(radius*math.s
in(angle))
```

```
current_col=str(68),str(68),str(238)

                        seperator="**##"

                        current_att1="Parent     affected
system"

                        current_att2=affected2

                        mp2="Parent   Affected   System2:
"+mp2.strip('\"')


writenodestring=current_parent2_target_node_id+seperator+
mp2+seperator+current_att1+seperator+current_att2+seperat
or+current_size


nodedic[current_parent2_target_node_id+seperator+mp2]=wri
tenodestring


#write_a_node(outfile,current_parent2_target_node_id,mp2,
current_att1,current_att2,current_pos,current_col,current
_size)


                '''

#beneath is outcommented on purpose, since the genes will
not be considered as nodes for this version

'''

for gene in gene2affected.keys():

    current_label="Mouse knock-out Gene: "+gene


    current_node_id=str(gene2geneId[gene])

    current_att1="Mouse knock-out Gene"

    try:

        current_att2="Entrez                              Id:
"+str(int(gene2entrez[gene]))

    except:
```

```
            current_att2="Entrez Id missing"
        #angle=random.uniform(-3.14*2,3.14*2)
        #radius=random.uniform(0,r)

#current_pos=str(radius*math.cos(angle)),str(radius*math.
sin(angle))
        #current_col=str(168),str(68),str(238)
        seperator="**##"
        current_size=str(len(gene2affected[gene]))

writenodestring=current_node_id+seperator+current_label+s
eperator+current_att1+seperator+current_att2+seperator+cu
rrent_size
        nodedic[gene+seperator+current_label]=writenodestring
'''


for n in sorted(nodedic.keys()):
        seperator="**##"
        current_node_string=nodedic[n]

current_node_id=current_node_string.split(seperator)[0]
        current_label=current_node_string.split(seperator)[1]
        current_att1=current_node_string.split(seperator)[2]
        current_att2=current_node_string.split(seperator)[3]
        current_size=current_node_string.split(seperator)[4]
        angle=random.uniform(-3.14*2,3.14*2)
        radius=random.uniform(0,r)

current_pos=str(radius*math.cos(angle)),str(radius*math.s
in(angle))
        if current_att1.startswith("Disease Class:")==True:
```

```
        current_disclass=current_att1.split("Disease
Class: ")[1]


current_col=bgr(0,len(dis_class.keys()),dis_class[current
_disclass])


current_col[0],current_col[1],current_col[2]=str(current_
col[0]),str(current_col[1]),str(current_col[2])

    else:

        if current_att1=="Mouse knock-out Gene":

            current_col=str(9),str(9),str(9)

        else:

          current_col=str(68),str(68),str(238)


write_a_node(outfile,current_node_id,current_label,curren
t_att1,current_att2,current_pos,current_col,current_size)

outfile.write("    </nodes>\n")



edgedic={}

edge_counter=1

humdis_counter=1

for humandisease in humandis2gene.keys():

    first_gene_tuple=humandis2gene[humandisease][0]


first_gene,current_disclass,current_row=first_gene_tuple

    first_gene,                                  current_att2,
current_dis_source_node_id=str(first_gene),str(current_di
sclass),str(current_row)

    #above   is   only   required   for   determining   the
current_dis_source_node_id

    #item="finished                                percentage:
"+str(float(humdis_counter/len(humandis2gene.keys()))*100
.0)
```

```
    #print  item, "\r",
    #sys.stdout.flush()
    #sleep(1)
    humdis_counter=humdis_counter+1
    print humdis_counter
    for gene_tuple in humandis2gene[humandisease]:
        gene,current_disclass,current_row=gene_tuple
        #if humandisease=="Leukemia":
        #    print gene
        try:
            dummy=len(gene2affected[gene])
        except:
            pass
        else:
            for affected in gene2affected[gene]:
                current_mp,current_affected=affected

current_mp,current_affected=str(current_mp),str(current_a
ffected)
                #if humandisease=="Leukemia":
                #    print "affected:"+current_affected

current_target_node_id=str(mp2id[current_mp])
                #current_edge_string='              <edge
id="'+str(edge_counter)+'"
source="'+current_dis_source_node_id+'"
target="'+current_target_node_id+'" label="'+gene+'">\n'
                #current_edge_string='              <edge
id="'+str(edge_counter)+'"
source="'+current_dis_source_node_id+'"
target="'+current_target_node_id+'" label=allgenes">\n'
```

```
                current_edge_string='"
source="'+current_dis_source_node_id+'"
target="'+current_target_node_id+'" label="allgenes">\n'


#current_edge_string=current_edge_string+'
<attvalues></attvalues>\n      </edge>\n'

        #edge_counter=edge_counter+1

                if     current_edge_string     not     in
edgedic.keys():

                    edgedic[current_edge_string]=[]

                edgedic[current_edge_string].append(gene)



outfile.write("    <edges>\n")

length_of_gene_keys={}

for e in sorted(edgedic.keys()):

    outfile.write('      <edge id="'+str(edge_counter))

    genestring=""

    for st in range(len(edgedic[e])-1):

        genestring=genestring+edgedic[e][st]+","

    genestring=genestring+edgedic[e][len(edgedic[e])-1]

    outfile.write(e.replace("allgenes",str(genestring)))

    length_of_gene_keys[len(edgedic[e])]=True

    outfile.write('             <attvalues></attvalues>\n
</edge>\n')

    edge_counter=edge_counter+1

    #outfile.write(e)

print max(length_of_gene_keys)

outfile.write('    </edges>\n')

outfile.write('  </graph>\n')
```

```
outfile.write('</gexf>\n')


print "finished"
```

B.2 GENES-NODE VERSION

```
import os

import sys

import time

import math

import xlrd

from xlrd import open_workbook

import random

area_density=0.05

def bgr(minimum, maximum, value):

    minimum, maximum = float(minimum), float(maximum)

    ratio = 2 * (float(value)-minimum) / (maximum -
minimum)

    b = int(max(0, 255*(1 - ratio)))

    r = int(max(0, 255*(ratio - 1)))

    g = 255 - b - r

    bgr=[]

    bgr.append(b)

    bgr.append(g)

    bgr.append(r)

    return bgr

def write_xml_header(file):

    file.write('<?xml    version="1.0"    encoding="UTF-
8"?>\n')

    file.write('<gexf
xmlns="http://www.gexf.net/1.2draft"        version="1.2"
xmlns:viz="http://www.gexf.net/1.2draft/viz"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
```

```
xsi:schemaLocation="http://www.gexf.net/1.2draft
http://www.gexf.net/1.2draft/gexf.xsd">\n')

    file.write('       <meta   lastmodifieddate="2014-01-
30">\n')

    file.write('    <creator>Gephi 0.8.1</creator>\n')

    file.write('    <description></description>\n')

    file.write('  </meta>\n')

    file.write('      <graph   defaultedgetype="directed"
mode="static">\n')

    file.write('              <attributes   class="node"
mode="static">\n')

    file.write('          <attribute  id="0"  title="Type"
type="string"/>\n')

    file.write('       <attribute id="1" title="Remarks"
type="string"/></attributes>\n')

    file.write('    <nodes>\n')

def
write_a_node(file,_node_id,_label,_att1,_att2,_pos,_col,_
size):

    file.write('                <node   id="'+_node_id+'"
label="'+_label+'"\n>')

    file.write('        <attvalues>\n')



    file.write('                          <attvalue  id="0"
value="'+_att1+'"></attvalue>\n')

    file.write('                          <attvalue  id="1"
value="'+_att2+'"></attvalue>\n')

    file.write('        </attvalues>\n')

    file.write('            <viz:position x="'+_pos[0]+'"
y="'+_pos[1]+'" z="0.0"></viz:position>\n')

    file.write('              <viz:color  b="'+_col[0]+'"
g="'+_col[1]+'" r="'+_col[2]+'"></viz:color>\n')

    file.write('                                  <viz:size
value="'+_size+'"></viz:size>\n')
```

```
    file.write('      </node>\n')
    print "undirected graph generation human diseases to
child affected names via genes"
first_book_name="first_table_corrected_new.xls"
first_book = xlrd.open_workbook(first_book_name)
ws=first_book.sheet_by_name("Sheet1")
humandis2gene={}
dis_class={}
first_line=True
current_row=1
gene2entrez={}
while first_line:
    try:
        current_entrezid=ws.cell(current_row,0).value
    except:
        first_line=False
    else:
        try:

current_gene=ws.cell(current_row,1).value.strip()
        except:
            pass
        else:
            gene2entrez[current_gene]=current_entrezid

current_humandis=ws.cell(current_row,3).value.strip()
            current_disclass=ws.cell(current_row,4).value
            dis_class[str(current_disclass)]=1.0
```

```
current_tuple=(current_gene,current_disclass,current_row)
            if       current_humandis       not       in
humandis2gene.keys():
                humandis2gene[current_humandis]=[]

humandis2gene[current_humandis].append(current_tuple)
        current_row=current_row+1
for l in range(len(dis_class.keys())):
    dis_class[dis_class.keys()[l]]=float(l)
second_book_name="second_with_parenting.xlsx"
second_book = xlrd.open_workbook(second_book_name)
ws=second_book.sheet_by_name("DENE")
gene2affected={}
mp2id={}
second_line=True
prev_row=current_row
current_row=1
while second_line:
    try:
        current_mp=ws.cell(current_row,0).value
    except:
        second_line=False
    else:
        try:

current_gene=ws.cell(current_row,2).value.strip()
        except:
            pass
```

```
        else:
            current_affected=ws.cell(current_row,1).value
            affected_tuple=(current_mp,current_affected)

            if current_gene not in gene2affected.keys():
                gene2affected[current_gene]=[]

gene2affected[current_gene].append(affected_tuple)
            mp2id[current_mp]=prev_row+current_row
            current_row=current_row+1
gene2geneId={}
prev_row=max(mp2id.values())+1
current_row=1
for gene in gene2affected.keys():
    gene2geneId[gene]=prev_row+current_row
    current_row=current_row+1
    no_notaffected_genes=0
max_radius=0
max_disease="some"
total_area=0.0
for gene_tuple_array in humandis2gene.values():
    for gene_tuple in gene_tuple_array:

current_gene,current_disclass,current_row=gene_tuple
current_gene,current_disclass,current_row=str(current_gen
e),str(current_disclass),str(current_row)
        #print current_gene,gene_tuple
        try:
            x=len(gene2affected[current_gene])
```

```
        except:
            no_notaffected_genes=no_notaffected_genes+1
        else:
            total_area=total_area+float(x)*float(x)*3.14
print "total area has been found to be: "+ str(total_area)
r=math.sqrt(total_area/area_density/3.14)
print "the maximum radius of the window has been found to
be "+str(r)
outfile=open("dis2affected.gexf","w")
write_xml_header(outfile)
debugfile=open("debugfile.log","w")
nodedic={}
missing=[]
for humandisease in humandis2gene.keys():
    prev_missing_length=len(missing)
    current_size=0.0
    current_gene_list = []
    for gene_tuple in humandis2gene[humandisease]:

current_gene,current_disclass,current_row=gene_tuple
        current_gene_list.append(current_gene)
        try:

current_size=current_size+len(gene2affected[current_gene]
)
        except:
            missing.append(current_gene)
    current_size=str(current_size)
    first_gene_tuple=humandis2gene[humandisease][0]
```

```
current_gene,current_disclass,current_row=first_gene_tupl
e

    current_gene,                              current_att2,
current_node_id=str(current_gene),str(current_disclass),s
tr(current_row)

    current_label="Disorder: "+humandisease

    current_att1="Disorder Class: "+current_disclass

    #current_att2="#        of        Mouse        Genes:
"+str(len(humandis2gene[humandisease]))+"             AND
"+str(len(missing)-prev_missing_length)+" of them do/does
not have any MP Id on Jax Database"+". "+" Genes are " +
', '.join(current_gene_list)

    current_att2="This network aims to reveal the OMIM
disorders and mouse affected system connections by using
ortholog mouse knock out Diseasome genes as edges. The
size of disease and affected system nodes are respectively
proportional to the number of genes in it and number of
disease it connects. Black nodes represent affected
systems. Red nodes represents mouse genes and other colours
show distinct 22 disorder classes."

seperator="**##"

writenodestring=current_node_id+seperator+current_label+s
eperator+current_att1+seperator+current_att2+seperator+cu
rrent_size

nodedic[current_node_id+seperator+current_label]=writenod
estring

print      "Total      #      of      nodes      appended:
"+str(len(nodedic.keys()))

print   "#   of   nodes   without   MP   Id   appended:
"+str(len(missing))

for k in missing:

    debugfile.write(k+"\n")

from collections import Counter

from itertools import chain

def rowToPairs(sheet, row):
```

```
    affected_system = sheet.cell(row, 1).value.strip()

    diseases = [d.strip() for d in sheet.cell(row,
3).value.split(',')]

    aff_sys_disease_pairs = [(affected_system, disease)
for disease in diseases]

    return aff_sys_disease_pairs

def sheet_to_pairs(sheet):

    return (rowToPairs(sheet, row) for row in range(0,
sheet.nrows))


def count_affected_in_sheet(sheet):

    unique_pairs                                        =
set(chain.from_iterable(sheet_to_pairs(sheet)))

    return Counter(aff_sys for (aff_sys, disease) in
unique_pairs)



counter = count_affected_in_sheet(ws)

for humandisease in humandis2gene.keys():

    for gene_tuple in humandis2gene[humandisease]:

        gene,current_disclass,current_row=gene_tuple

        try:

            dummy=len(gene2affected[gene])

        except:

            pass

        else:

            for affected in gene2affected[gene]:

                current_mp,current_affected=affected

                current_size                            =
str(counter.get(current_affected))
```

```
                current_att1="Affected system"
                current_att2=current_affected
                angle=random.uniform(-3.14*2,3.14*2)
                radius=random.uniform(0,r)

current_pos=str(radius*math.cos(angle)),str(radius*math.s
in(angle))
                current_col=str(68),str(68),str(238)
                seperator="**##"

current_target_node_id=str(mp2id[current_mp])
                current_affected=current_att2
writenodestring=current_target_node_id+seperator+current_
affected+seperator+current_att1+seperator+current_att2+se
perator+current_size
nodedic[current_target_node_id+seperator+current_mp]=writ
enodestring
for gene in gene2affected.keys():
    current_label="Mouse knock-out Gene: "+gene

    current_node_id=str(gene2geneId[gene])
    current_att1="Mouse knock-out Gene"
    try:
        current_att2="Entrez                          Id:
"+str(int(gene2entrez[gene]))
    except:
        current_att2="Entrez Id missing"
    seperator="**##"
    current_size=str(len(gene2affected[gene]))
writenodestring=current_node_id+seperator+current_label+s
eperator+current_att1+seperator+current_att2+seperator+cu
rrent_size
```

```
    nodedic[gene+seperator+current_label]=writenodestring
for n in sorted(nodedic.keys()):
    seperator="**##"
    current_node_string=nodedic[n]

current_node_id=current_node_string.split(seperator)[0]
    current_label=current_node_string.split(seperator)[1]
    current_att1=current_node_string.split(seperator)[2]
    current_att2=current_node_string.split(seperator)[3]
    current_size=current_node_string.split(seperator)[4]
    angle=random.uniform(-3.14*2,3.14*2)
    radius=random.uniform(0,r)

current_pos=str(radius*math.cos(angle)),str(radius*math.s
in(angle))
    if current_att1.startswith("Disease Class:")==True:
        current_disclass=current_att1.split("Disease
Class: ")[1]

current_col=bgr(0,len(dis_class.keys()),dis_class[current
_disclass])

current_col[0],current_col[1],current_col[2]=str(current_
col[0]),str(current_col[1]),str(current_col[2])
    else:
        if current_att1=="Mouse knock-out Gene":
            current_col=str(9),str(9),str(9)
        else:
          current_col=str(68),str(68),str(238)

write_a_node(outfile,current_node_id,current_label,curren
t_att1,current_att2,current_pos,current_col,current_size)
```

```
outfile.write("    </nodes>\n")


edgedic={}

edge_counter=1

for humandisease in humandis2gene.keys():

    first_gene_tuple=humandis2gene[humandisease][0]


first_gene,current_disclass,current_row=first_gene_tuple

    first_gene,                                  current_att2,
current_dis_source_node_id=str(current_gene),str(current_
disclass),str(current_row)

    for gene_tuple in humandis2gene[humandisease]:

        gene,current_disclass,current_row=gene_tuple

        try:

            dummy=len(gene2affected[gene])

        except:

            pass

        else:


current_gene_target_node_id=str(gene2geneId[gene])

            current_edge_string='                    <edge
id="'+str(edge_counter)+'"
source="'+current_dis_source_node_id+'"
target="'+current_gene_target_node_id+'"
label="'+gene+'">\n'

            current_edge_string=current_edge_string+'
<attvalues></attvalues>\n      </edge>\n'

            edge_counter=edge_counter+1

            edgedic[current_edge_string]=1.0


current_source_node_id=current_gene_target_node_id

            for affected in gene2affected[gene]:
```

```
                current_mp,current_affected=affected


current_mp,current_affected=affected=str(current_mp),str(
current_affected)


current_target_node_id=str(mp2id[current_mp])

                current_edge_string='                <edge
id="'+str(edge_counter)+'"
source="'+current_source_node_id+'"
target="'+current_target_node_id+'" label="'+gene+'">\n'

                current_edge_string=current_edge_string+'
<attvalues></attvalues>\n      </edge>\n'

                edge_counter=edge_counter+1

                edgedic[current_edge_string]=1.0

outfile.write("    <edges>\n")

for e in sorted(edgedic.keys()):

    outfile.write(e)

outfile.write('    </edges>\n')

outfile.write('  </graph>\n')

outfile.write('</gexf>\n')


print "finished"
```

B.3 REMOVING NUMBERS

```
Function RemoveNumbers(t As String)
  Dim i As Long, newString As String
    For i = 1 To Len(t)
    If Not IsNumeric(Mid(t, i, 1)) Then
            newString = newString & Mid(t, i, 1)
    End If
    Next i
    RemoveNumbers = newString
     End Function
```

APPENDIX C

C.1 HUMAN AND MOUSE PHENOTYPES FOR TP53 AND TRP53 ORTHOLOGUE GENES

HUMAN PHENOTYPES FOR TP53 GENE

HPO id	HPO label
HP:0011974	Myelofibrosis
HP:0001658	Myocardial infarction
HP:0001413	Micronodular cirrhosis
HP:0002669	Osteosarcoma
HP:0100576	Amaurosis fugax
HP:0000007	Autosomal recessive inheritance
HP:0000006	Autosomal dominant inheritance
HP:0002667	Nephroblastoma
HP:0002665	Lymphoma
HP:0002863	Myelodysplasia
HP:0006744	Adrenocortical carcinoma
HP:0002861	Melanoma
HP:0001250	Seizures
HP:0006740	Transitional cell carcinoma of the bladder
HP:0030448	Soft tissue sarcoma
HP:0001939	Abnormality of metabolism/homeostasis
HP:0001428	Somatic mutation
HP:0004936	Venous thrombosis
HP:0006716	Hereditary nonpolyposis colorectal carcinoma
HP:0100543	Cognitive impairment
HP:0001425	Heterogeneous
HP:0002315	Headache
HP:0001744	Splenomegaly
HP:0003401	Paresthesia
HP:0003003	Colon cancer
HP:0100787	Prostate neoplasm
HP:0006753	Neoplasm of the stomach
HP:0003002	Breast carcinoma
HP:0100749	Chest pain
HP:0004375	Neoplasm of the nervous system
HP:0005584	Renal cell carcinoma
HP:0004374	Hemiplegia/hemiparesis
HP:0011875	Abnormal platelet morphology
HP:0010982	Polygenic inheritance
HP:0012125	Prostate cancer
HP:0009919	Retinoblastoma
HP:0030078	Lung adenocarcinoma
HP:0100273	Neoplasm of the colon
HP:0000505	Visual impairment
HP:0002448	Progressive encephalopathy
HP:0100630	Neoplasia of the nasopharynx
HP:0002326	Transient ischemic attack
HP:0002488	Acute leukemia
HP:0001276	Hypertonia
HP:0005513	Increased megakaryocyte count
HP:0004420	Arterial thrombosis
HP:0003010	Prolonged bleeding time
HP:0008069	Neoplasm of the skin
HP:0006572	Subacute progressive viral hepatitis
HP:0200022	Choroid plexus papilloma
HP:0100641	Neoplasm of the adrenal cortex
HP:0001402	Hepatocellular carcinoma
HP:0000238	Hydrocephalus
HP:0002018	Nausea
HP:0002013	Vomiting
HP:0002894	Neoplasm of the pancreas
HP:0002891	Uterine leiomyosarcoma
HP:0001085	Papilledema

MOUSE PHENOTYPES FOR Trp53 GENE

MP ID	Term
MP:0009625	abnormal abdominal lymph node morphology
MP:0001648	abnormal apoptosis
MP:0000297	abnormal atrioventricular cushion morphology
MP:0006126	abnormal cardiac outflow tract development
MP:0001544	abnormal cardiovascular system physiology
MP:0003077	abnormal cell cycle
MP:0004045	abnormal cell cycle checkpoint function
MP:0000313	abnormal cell death
MP:0003091	abnormal cell migration
MP:0005621	abnormal cell physiology
MP:0000350	abnormal cell proliferation
MP:0004097	abnormal cerebellar cortex morphology
MP:0000857	abnormal cerebellar foliation
MP:0000875	abnormal cerebellar Purkinje cell layer
MP:0004023	abnormal chromosome number
MP:0000367	abnormal coat/ hair morphology
MP:0006410	abnormal common myeloid progenitor cell morphology
MP:0002661	abnormal corpus epididymis morphology
MP:0000428	abnormal craniofacial morphology
MP:0002123	abnormal definitive hematopoiesis
MP:0000569	abnormal digit pigmentation
MP:0008057	abnormal DNA replication
MP:0004896	abnormal endometrium morphology
MP:0008884	abnormal enterocyte apoptosis
MP:0003743	abnormal facial morphology
MP:0000377	abnormal hair follicle morphology
MP:0010763	abnormal hematopoietic stem cell physiology
MP:0001545	abnormal hematopoietic system physiology
MP:0000026	abnormal inner ear morphology
MP:0000527	abnormal kidney development
MP:0002109	abnormal limb morphology
MP:0002619	abnormal lymphocyte morphology
MP:0008280	abnormal male germ cell apoptosis
MP:0000628	abnormal mammary gland development
MP:0006270	abnormal mammary gland growth during lactation
MP:0001881	abnormal mammary gland physiology
MP:0002413	abnormal megakaryocyte progenitor cell morphology
MP:0006301	abnormal mesenchyme morphology
MP:0003720	abnormal neural tube closure
MP:0009937	abnormal neuron differentiation
MP:0002864	abnormal ocular fundus morphology
MP:0001330	abnormal optic nerve morphology
MP:0002177	abnormal outer ear morphology
MP:0000696	abnormal Peyer's patch morphology
MP:0008872	abnormal physiological response to xenobiotic
MP:0005195	abnormal posterior eye segment morphology
MP:0002792	abnormal retinal vasculature morphology
MP:0001325	abnormal retina morphology
MP:0002113	abnormal skeleton development
MP:0001156	abnormal spermatogenesis
MP:0004174	abnormal spine curvature
MP:0002362	abnormal spleen marginal zone morphology
MP:0000689	abnormal spleen morphology
MP:0002357	abnormal spleen white pulp morphology
MP:0013177	abnormal tail tip morphology
MP:0003155	abnormal telomere length
MP:0009258	abnormal thymocyte apoptosis
MP:0000703	abnormal thymus morphology
MP:0003763	abnormal thymus physiology
MP:0000558	abnormal tibia morphology
MP:0000762	abnormal tongue morphology
MP:0002019	abnormal tumor incidence
MP:0010307	abnormal tumor latency
MP:0003448	abnormal tumor morphology
MP:0000865	absent cerebellum vermis
MP:0001333	absent optic nerve
MP:0002386	adipose tissue inflammation
MP:0000414	alopecia
MP:0000343	altered response to myocardial infarction
MP:0010639	altered tumor pathology
MP:0001577	anemia
MP:0004024	aneuploidy
MP:0001393	ataxia
MP:0001625	cardiac hypertrophy
MP:0005385	cardiovascular system phenotype
MP:0005384	cellular phenotype
MP:0000851	cerebellum hypoplasia
MP:0008866	chromosomal instability
MP:0000111	cleft palate
MP:0005170	cleft upper lip
MP:0003269	colon polyps
MP:0005409	darkened coat color
MP:0006043	decreased apoptosis
MP:0006043	decreased apoptosis
MP:0001265	decreased body size
MP:0001262	decreased body weight
MP:0000333	decreased bone marrow cell number
MP:0000063	decreased bone mineral density
MP:0000774	decreased brain size
MP:0000352	decreased cell proliferation
MP:0003207	decreased cellular sensitivity to gamma-irradiation
MP:0004228	decreased cellular sensitivity to ionizing radiation
MP:0008411	decreased cellular sensitivity to ultraviolet irradiation
MP:0004701	decreased circulating insulin-like growth factor I level
MP:0008813	decreased common myeloid progenitor cell number
MP:0002875	decreased erythrocyte cell number
MP:0008973	decreased erythroid progenitor cell number
MP:0013502	decreased fibroblast apoptosis
MP:0002028	decreased hematocrit
MP:0004810	decreased hematopoietic stem cell number
MP:0011903	decreased hematopoietic stem cell proliferation
MP:0004502	decreased incidence of tumors by chemical induction
MP:0003918	decreased kidney weight
MP:0000221	decreased leukocyte cell number
MP:0003402	decreased liver weight
MP:0003010	decreased mortality induced by ionizing radiation
MP:0004981	decreased neuronal precursor cell number
MP:0003204	decreased neuron apoptosis
MP:0003204	decreased neuron apoptosis
MP:0000222	decreased neutrophil cell number
MP:0008944	decreased sensitivity to induced cell death
MP:0009764	decreased sensitivity to induced morbidity/mortality
MP:0001204	decreased sensitivity to skin irradiation
MP:0004819	decreased skeletal muscle mass
MP:0008477	decreased spleen red pulp amount
MP:0004953	decreased spleen weight
MP:0009341	decreased splenocyte apoptosis
MP:0008844	decreased subcutaneous adipose tissue amount
MP:0008770	decreased survivor rate
MP:0006414	decreased T cell apoptosis
MP:0005095	decreased T cell proliferation
MP:0004852	decreased testis weight
MP:0009542	decreased thymocyte apoptosis
MP:0003447	decreased tumor growth/size
MP:0002052	decreased tumor incidence
MP:0010308	decreased tumor latency
MP:0008009	delayed cellular replicative senescence
MP:0000423	delayed hair regrowth
MP:0002908	delayed wound healing
MP:0010699	dilated hair follicles
MP:0000284	double outlet right ventricle
MP:0008008	early cellular replicative senescence
MP:0011098	embryonic lethality during organogenesis, complete penetrance
MP:0011108	embryonic lethality during organogenesis, incomplete penetrance
MP:0005379	endocrine/exocrine gland phenotype
MP:0003927	enhanced cellular glucose import
MP:0002724	enhanced wound healing
MP:0000274	enlarged heart
MP:0000599	enlarged liver
MP:0000691	enlarged spleen
MP:0000914	exencephaly
MP:0001661	extended life span
MP:0000240	extramedullary hematopoiesis
MP:0003299	gastric polyps
MP:0005378	growth/size/body region phenotype
MP:0010700	hair follicle comedo
MP:0005397	hematopoietic system phenotype
MP:0001914	hemorrhage
MP:0002628	hepatic steatosis
MP:0005387	immune system phenotype
MP:0001525	impaired balance
MP:0009308	increased adenocarcinoma incidence
MP:0010383	increased adenoma incidence
MP:0005601	increased angiogenesis
MP:0006042	increased apoptosis
MP:0010277	increased astrocytoma incidence
MP:0002023	increased B cell derived lymphoma incidence
MP:0005238	increased brain size
MP:0009277	increased brain tumor incidence
MP:0002038	increased carcinoma incidence
MP:0000351	increased cell proliferation
MP:0004227	increased cellular sensitivity to ionizing radiation
MP:0008410	increased cellular sensitivity to ultraviolet irradiation
MP:0008577	increased circulating interferon-gamma level
MP:0008623	increased circulating interleukin-3 level
MP:0008626	increased circulating interleukin-5 level
MP:0008596	increased circulating interleukin-6 level
MP:0008553	increased circulating tumor necrosis factor level
MP:0004868	increased endometrial carcinoma incidence
MP:0008885	increased enterocyte apoptosis
MP:0011703	increased fibroblast proliferation
MP:0010363	increased fibrosarcoma incidence
MP:0009317	increased follicular lymphoma incidence
MP:0000575	increased foot pad pigmentation
MP:0010278	increased glioma incidence
MP:0010306	increased hamartoma incidence
MP:0002833	increased heart weight

MP:0002947	increased hemangioma incidence
MP:0003667	increased hemangiosarcoma incidence
MP:0004809	increased hematopoietic stem cell number
MP:0003331	increased hepatocellular carcinoma incidence
MP:0010344	increased hibernoma incidence
MP:0009321	increased histiocytic sarcoma incidence
MP:0004499	increased incidence of tumors by chemical induction
MP:0004500	increased incidence of tumors by ionizing radiation induction
MP:0001846	increased inflammatory response
MP:0002404	increased intestinal adenoma incidence
MP:0002035	increased leiomyosarcoma incidence
MP:0002026	increased leukemia incidence
MP:0010343	increased lipoma incidence
MP:0008019	increased liver tumor incidence
MP:0002027	increased lung adenocarcinoma incidence
MP:0002048	increased lung adenoma incidence
MP:0008014	increased lung tumor incidence
MP:0012431	increased lymphoma incidence
MP:0001883	increased mammary adenocarcinoma incidence
MP:0010299	increased mammary gland tumor incidence
MP:0006283	increased medulloblastoma incidence
MP:0001272	increased metastatic potential
MP:0001658	increased mortality induced by gamma-irradiation
MP:0003992	increased mortality induced by ionizing radiation
MP:0012043	increased myoepithelioma incidence
MP:0003789	increased osteosarcoma incidence
MP:0008000	increased ovary tumor incidence
MP:0009153	increased pancreas tumor incidence
MP:0002013	increased pilomatricoma incidence
MP:0008186	increased pro-B cell number
MP:0010287	increased reproductive system tumor incidence
MP:0002036	increased rhabdomyosarcoma incidence
MP:0002032	increased sarcoma incidence
MP:0008943	increased sensitivity to induced cell death
MP:0009766	increased sensitivity to xenobiotic induced morbidity/mortality
MP:0002051	increased skin papilloma incidence
MP:0009704	increased skin squamous cell carcinoma incidence
MP:0010300	increased skin tumor incidence
MP:0010367	increased spindle cell carcinoma incidence
MP:0009318	increased splenic marginal zone lymphoma incidence
MP:0009336	increased splenocyte proliferation
MP:0004207	increased squamous cell carcinoma incidence
MP:0011276	increased tail pigmentation
MP:0002024	increased T cell derived lymphoma incidence
MP:0002627	increased teratoma incidence
MP:0004478	increased testicular teratoma incidence
MP:0009541	increased thymocyte apoptosis
MP:0002020	increased tumor incidence
MP:0002020	increased tumor incidence
MP:0009828	increased tumor latency
MP:0010289	increased urinary system tumor incidence
MP:0010771	integument phenotype
MP:0000512	intestinal ulcer
MP:0008011	intestine polyps
MP:0000585	kinked tail
MP:0000160	kyphosis
MP:0010249	lactation failure
MP:0011099	lethality throughout fetal growth and development, complete penetrance
MP:0011109	lethality throughout fetal growth and development, incomplete penetrance
MP:0005202	lethargy
MP:0003046	liver cirrhosis
MP:0005141	liver hyperplasia
MP:0000600	liver hypoplasia
MP:0001860	liver inflammation
MP:0000162	lordosis
MP:0001861	lung inflammation
MP:0010768	mortality/aging
MP:0002269	muscular atrophy
MP:0012400	nail dystrophy
MP:0002006	neoplasm
MP:0003631	nervous system phenotype
MP:0002169	no abnormal phenotype detected
MP:0002700	opacity of vitreous body
MP:0006219	optic nerve degeneration
MP:0006221	optic nerve hypoplasia
MP:0003751	oral leukoplakia
MP:0000067	osteopetrosis
MP:0000066	osteoporosis
MP:0002016	ovary cysts
MP:0013236	ovary degeneration
MP:0003674	oxidative stress
MP:0003717	pallor
MP:0005152	pancytopenia
MP:0002633	persistent truncus arteriosis
MP:0004025	polyploidy
MP:0001732	postnatal growth retardation
MP:0011085	postnatal lethality, complete penetrance
MP:0011086	postnatal lethality, incomplete penetrance
MP:0003786	premature aging
MP:0002083	premature death
MP:0011091	prenatal lethality, complete penetrance
MP:0011101	prenatal lethality, incomplete penetrance
MP:0011100	preweaning lethality, complete penetrance
MP:0011110	preweaning lethality, incomplete penetrance
MP:0009908	protruding tongue
MP:0006050	pulmonary fibrosis
MP:0003446	renal hypoplasia
MP:0005367	renal/urinary system phenotype
MP:0005389	reproductive system phenotype
MP:0010715	retina coloboma
MP:0000420	ruffled hair
MP:0002064	seizures
MP:0004351	short humerus
MP:0000088	short mandible
MP:0001212	skin lesions
MP:0001116	small gonad
MP:0002989	small kidney
MP:0000692	small spleen
MP:0001147	small testis
MP:0000706	small thymus
MP:0000416	sparse hair
MP:0000693	spleen hyperplasia
MP:0000694	spleen hypoplasia
MP:0004029	spontaneous chromosome breakage
MP:0001940	testis hypoplasia
MP:0006099	thin cerebellar granule layer
MP:0001244	thin dermal layer
MP:0003179	thrombocytopenia
MP:0000745	tremors
MP:0010537	tumor regression
MP:0001864	vasculitis
MP:0010402	ventricular septal defect
MP:0001263	weight loss

C.2 LIST OF GENES ASSOCIATED WITH LEUKEMIA DISEASE & DAVID FUNCTIONAL ENRICHMENT RESULTS

Gene symbols
Abl1
Arhgap26
Arhgef12
Arnt
Bcl2
Bcr
Ccnd1
Cebpa
Chic2
Flt3
Gata1
Hoxd4
Kit
Kras
Lpp
Nbn
Nf1
Npm1
Nqo1
Numa1
Nup214
P2rx7
Picalm
Pml
Ptpn11
Runx1
Stat5b
Tal1
Tal2
Whsc1l1
Zbtb16

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
KEGG_PATHWAY	Acute myeloid leukemia	RT		9	29.0	5.4E-13	5.0E-11
KEGG_PATHWAY	Pathways in cancer	RT		14	45.2	8.1E-13	3.8E-11
GOTERM_BP_DIRECT	homeostasis of number of cells within a tissue	RT		6	19.4	4.3E-9	3.4E-6
GOTERM_BP_DIRECT	hemopoiesis	RT		7	22.6	5.3E-9	2.1E-6
KEGG_PATHWAY	Chronic myeloid leukemia	RT		7	22.6	1.9E-8	5.8E-7
GOTERM_MF_DIRECT	protein binding	RT		22	71.0	3.3E-8	5.1E-6
GOTERM_BP_DIRECT	negative regulation of cell proliferation	RT		9	29.0	1.5E-7	3.9E-5
UP_KEYWORDS	Ubl conjugation	RT		13	41.9	2.0E-7	2.0E-5
UP_KEYWORDS	Proto-oncogene	RT		6	19.4	2.3E-7	1.2E-5
UP_KEYWORDS	Acetylation	RT		17	54.8	3.5E-7	1.2E-5
GOTERM_BP_DIRECT	liver development	RT		6	19.4	4.6E-7	9.2E-5
GOTERM_BP_DIRECT	regulation of cell cycle	RT		6	19.4	1.0E-6	1.7E-4
UP_KEYWORDS	Phosphoprotein	RT		24	77.4	1.7E-6	4.4E-5
GOTERM_CC_DIRECT	nucleus	RT		23	74.2	1.8E-6	1.9E-4
GOTERM_BP_DIRECT	embryonic hemopoiesis	RT		4	12.9	5.4E-6	7.1E-4
UP_KEYWORDS	Nucleus	RT		18	58.1	1.0E-5	2.1E-4