ARDInteract: Construction, Visualization and Analysis of ARD Networks

A Major Project report submitted

In partial fulfilment of the requirement for the degree of

Master of Technology (M. Tech)

In

Bioinformatics

Submitted by

Pooja Khurana

(2K14/BIO/11)

Under the supervision of

Dr. Yasha Hasija

Department of Biotechnology
Delhi Technological University
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road,
Delhi-110042, INDIA

CERTIFICATE

This is to certify that the M.Tech dissertation entitled **ARDInteract: Construction, Visualization and Analysis of ARD Networks** submitted by **Pooja Khurana (2K14/BIO/11)** in the partial fulfilment of the requirements for the award of the degree of Master of Technology, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the candidate's own work carried out by her under my guidance.

The information and data enclosed in this thesis is original and has not been submitted elsewhere for honouring of any other degree.

Dr. Yasha Hasija
(Project Mentor)
Department of Biotechnology
Delhi Technological University
Delhi – 110042

Dr. D. Kumar
Professor and Head
Department of Biotechnology
Delhi Technological University
Delhi - 110042

DECLARATION

I declare that the project report entitled **“ARDInteract: Construction, Visualization and Analysis of ARD Networks”,** submitted by me is in the partial fulfilment of the requirement for the award of the degree of Master of Technology in Bioinformatics, Department of Biotechnology, Delhi Technological University. It is a record of original research work carried out by me under the supervision of Dr. Yasha Hasija, Department of Biotechnology, Delhi Technological University, Delhi.

The matter reported in this project report is original and has not been submitted for the award of any other degree.

Pooja Khurana
2K14/BIO/11
Department of Biotechnology
Delhi Technological University
(Formerly Delhi College of Engineering)

Date:

ACKNOWLEDGEMENT

I would like to express my gratitude to my guide, Dr. Yasha Hasija, for her continuous encouragement, support and critical suggestions throughout the beginning of this work. There were times when she kept on having faith in me, although I, myself, was a little apprehensive about the progress of my work; I truly appreciate that.

Also, I would like to express my gratitude to other faculty members for the valuable knowledge that they have imparted through their courses. I would also acknowledge the support provided by the staff of the department.

I would also like to express my cordial appreciation to Ms. Isha Srivastava, Research Scholar, for providing me with the initial data for genetic variants of Age Related Disorders and also explaining me the different aspects to be considered when dealing with genetic variation data. Also, my sincere thanks to my seniors Mr. Lokesh, Ms. Navneet and Mr. Nitin for helping me in using different softwares/programs required.

And finally, my parents and friends, for always providing me emotional support and determination. It is only because of them whatever I am today and whatever good I will achieve tomorrow.

Pooja Khurana
(2K14/BIO/11)

CONTENTS

ARDInteract: Construction, Visualization and Analysis of ARD Networks

Pooja Khurana

*Delhi Technological University, Delhi, India

btpooja@gmail.com

ABSTRACT

Age-Related Disorders are the complex disorders associated with the process of ageing affecting the quality of life of our elderly population. Understanding the biology of ageing and age-related disorders has thus become one of the most important field of research in medicine and requires integration of the vast amount of biological data residing in the databases. Large amount of data on ARD associated variants, genes and proteins are contained in these databases but they lack connectivity which otherwise is important to decipher the biology of aging and associated diseases. Recently, studies have demonstrated that network approaches helps when applied to integrate biological information can provide novel insights and pave the way for understanding and curing complex diseases. Here we, therefore, propose a new platform ARDInteract that integrate different data sources and allow the creation of heterogeneous networks at various '-omic' levels - SNP, gene, protein, disease and drug interaction levels. ARDInteract provides a user-friendly interface to integrate, visualize and analyse genome-scale biological networks for ARDs ultimately allowing the research community to connect distinct spots in space to solve the puzzle of human aging and ARDs. We believe that by using holistic approach we can achieve our rationale of personalized medicine, and eventually healthy ageing.

LIST OF ABBREVIATIONS

ARD	Age Related Disorder
GO	Gene Ontology
GWAS	Genome-Wide Association Study
GWAS	Genome-wide association study
HPRD	Human protein reference database
KEGG	Kyoto encyclopedia of genes and genomes
OMIM	Online Mendelian Inheritance in Man
PPI	Protein-protein interaction.
SNP	Single Nucleotide Polymorphism
T2D	Type 2 Diabetes

LIST OF FIGURES

LIST OF TABLES

PART I: INTRODUCTION

1. INTRODUCTION

Aging is a multilevel and multifaceted process of getting older characterised by physical, physiological, psychological and social changes. It is an inevitable process and is usually marked by the negative deviation of the bodily structures and functions from the optimum. The decrease in ability to respond to stress, increase in homeostasis imbalance and increase in the rise of age-related diseases/disorders (ARDs) are some of the characteristic features of Aging. It has been found that nearly 100,000 people die per day of various ARDs such as cardiovascular diseases, cancer, arthritis, osteoporosis, type 2 diabetes, schizophrenia, Alzheimer's disease to name a few. It has been observed that incidence of ARDs increases rapidly with aging and hence there are growing efforts in aging research to understand the biology of aging with the ultimate goal of delaying/stopping aging and extending the healthy lifespan.

Aging and longevity are the consequences of the complex interplay of environmental and genetic factors. Number of biological pathways have been found to be play key role in aging and includes lipid/cholesterol metabolism, immune system processes, energy metabolism in mitochondria and insulin receptor signalling pathway. Researches in model organisms have shown that by manipulating a few genes the lifespan of the organism can be extended (Kenyon, 2010). One such example is demonstrated in *C. elegans* where inactivation in daf-2 gene increases the lifespan by approximately 100% (Sebastiani et. al., 2009). It has been estimated that nearly 30% of differences in life expectancy in humans are governed by genetic factors (Shi et. al., 2012). There is growing evidence that the genes involved in ARDs affect the human life expectancy. However, understanding the genetics of aging process and age-related diseases is complex and is currently one of the world's major scientific challenges.

Several approaches have been implemented to identify the genes associated with aging and ARDs in humans (Melzer et. al., 2007). One such approach is candidate gene approach in which scientists look for genes in humans that serve similar functions in the body as genes already associated with aging in model organisms and then looking for the variants of these genes which are found to be common in people with long lifespan. However, this method has its own shortcomings and cannot be used for determining all the genes associated with diseases. The other method that is used is linkage analysis. In this method scientists look for regions of the human genome shared more often than expected by chance between close relatives who also share exceptional longevity. However, one major drawback of this method is its inability to map gene of modest relative risk (OR<2). With the advancements in genome sequencing methods, yet another powerful approach that has become the main pipeline for the identification of variants associated with diseases is Genome-Wide Association Study. The rapid development and expanded use of microarray technologies, including oligonucleotide array comparative genomic hybridization and SNP genotyping arrays, as well as next-generation sequencing with paired-end methods, has enabled a whole-genome analysis with almost unlimited resolution. State-of-the art GWAS studies most often look for individual genes with large impacts on a single phenotype. However, this approach also suffers from a limitation as the impact of genetic variations cannot be studied in isolation. Predictive elements, such as single nucleotides (SNPs), loci, genes, or entire biological pathways interact at all levels of granularity. The pervasiveness and strength of bio-molecular interactions thus require a step back from reductionist biology and an acknowledgement of the importance of biological networks and pathways.

The field of network biology has recently emerged as a powerful paradigm to visualize and analyse large data ensembles in novel ways with unparalleled flexibility (Barabasi and Otavi, 2004). The application of this approach in the recent past has enabled a detailed look at the genetic landscape of complex human phenotypes (Ideker and Sharan, 2005). In 2007, Goh et al. reported the first human disease network and provided a novel view of the genetic relationship among diseases (Goh et. al., 2007). Another pioneering study summarized the application of protein networks for network-based classification of diseases (Ideker and Sharan, 2005) and integration of drug targets and disease gene products led to the field of systems pharmacology (Berger and Iyenagar, 2010; Yildrim et. al., 2007).

Therefore, in order to understand the common etiology of complex ARDs and the how genetic variants play role in various ARDs, we have developed a framework for the construction, visualization and analysis of ARD networks at different granularity levels by integrating information from genome-wide association studies, protein-protein interaction and gene ontology databases. In this work, we used SNPs as a smaller unit, and biological pathways as a larger unit. We took a bird's eye view of the effect of genetic variations on the ARDs. We also took into account two phenomena that illustrate the underlying complexity of these genetic variations: pleiotropy, when a single mutation affects several traits, and epistasis, when multiple mutations have synergetic, usually non-linear, effects on a single phenotype (Moore et. al., 2014). Moore et. al. (Moore et. al. 2009) asserted that epistasis and pleiotropy are not isolated occurrences, but ubiquitous and inherent properties of biomolecular networks. A systems level understanding of epistasis and pleiotropy is, thus, critical to furthering our understanding of human genetics and its contribution to common human disease.

Construction of ARD networks based on predictive elements of different scales, therefore, provide a deeper insight into how various diseases are associated together and allow us to identify the pleiotropic and epistatic interactions at the system's level. The analysis of ARD networks can be applied to discover patterns and predict causal genetic markers (nodes in the network), or active modules (referred to as sub-networks or pathways) to help understand the molecular basis of ARDs and their relationships that an facilitate early diagnosis, prognosis, prevention and treatment of ARDs.

The aim of the current study is:

- To construct Age-Related Disorder (ARD) networks at different levels of granularity, from common genetic variants to entire biological pathway.
- To explore the hypothesis that subsets of associated SNPs/genes characterize different pathways to ARDs.
- To develop an integrated platform "ARDInteract" that allows visualization and analysis of ARD networks.

PART II: REVIEW OF LITERATURE

2. REVIEW OF LITERATURE

This section introduces the commonly used network terminology and definitions followed by the applications of network biology in the medicine. The chapter ends with a review of recent computational work on network-based analysis of age-related diseases.

2.1 NETWORK THEORY

A network is a set of nodes connected by interactions, and can represent anything from social interactions between pupils in a school class, to traffic between airports. Organizing complex systems into networks can give a clear overview of the system, and effectively identify important components.

The analysis of complex networks constitutes a field of science known as network theory. Network theory is a powerful tool for systems biologists. Disease-associated genes can be organized into networks using various biological sources such as physical protein-protein interactions (PPIs), literature co-citation or co-expression. Those networks can then be dissected to identify disease pathways, mechanisms, clinical markers and drug targets. Two concepts are of particular interest in this kind of analysis: modularity and centrality.

2.1.1 Modularity

In a modular network, the nodes are divided into groups that share many interactions internally but few interactions with other groups. Many real world networks, such as social networks and gene networks tend to be highly modular. Identifying modules in a large, complex network can reduce thousands of nodes to a handful of modules, thereby drastically simplifying the analysis.

The most extreme version of module is termed clique. A clique is a set of nodes that are completely connected with each other. All cliques in a network that are not subsets of another clique, are termed as maximal cliques.

A method to determine if a gene is part of a module or not is termed clustering coefficient (see Table 1 for the description of properties of a node in a network). The clustering coefficient of a node is defined by how many of the node's interactors that are connected with each other. Average clustering coefficient of a network calculated as the mean of the clustering coefficients of all the nodes in the network gives a measure of cohesiveness in the network which is also commonly referred to as the extent of modularity. The higher the clustering coefficient greater is the modular nature of the network. To compare the extent of cohesiveness in a network often clustering coefficients of the real networks are compared with random networks with similar size and degree distribution.

Table 1. *Different local properties which can be defined for a node in complex networks.*

Property	Description
Indegree/Incoming degree	In directed networks where directionality of an interaction is taken into account, indegree refers to the number of incoming connections to a node of interest. In other words, indegree is the number of arrows that flow into the node under investigation.
Outdegree/Outgoing degree	Out degree refers to the number of edges which start from a node of interest and point to other nodes in the network and is valid for directed networks where there is direction associated with each edge represented.
Degree or Connectivity	Degree or connectivity of a node refers to the total number of interactions it has in a network – the higher the connectivity (i.e., hub nodes) the more the number of targets it interacts with. In directed networks degree simply corresponds to the sum of in and out degrees of a node.
Clustering Coefficient	Clustering coefficient of a node reflects the extent to which the neighbors of a given node are interconnected among themselves to what is expected theoretically and indicates the cohesiveness

	or local modularity of the network. An extension of this metric to the complete network defined as the average clustering coefficient of all nodes, tells whether the network is modular or is sparsely connected.
Betweenness	Betweenness centrality of a node measures the number of shortest paths between all pairs of nodes in the network that pass through a node of interest – the higher the number of paths that pass through a node, the more important it is.
Average path length	Average length of the shortest paths between all pairs of nodes in the network.
Closeness	Closeness centrality is defined as the inverse of the average length of all the shortest paths from a node of interest to all other nodes in the network - note that closeness centrality defined this way implies that higher the closeness value, the higher the importance (centrality) of a node.
Diameter	The diameter of a network is the length of the longest path among all the shortest paths defined between two nodes. It gives an estimation of the distance between the farthest nodes in the network.
Graph density	The density of a network is the ratio of the number of edges to the number of total possible edges.
Power law fit	Fitting a power-law distribution function to the degree distribution of the network to study whether the network is likely to exhibit a scale-free network structure.

2.1.2 Centrality

Centrality is a way to prioritize nodes within a network. A central node is well-connected with the rest of the network; removing it will strongly affect the integrity of the network. The measures of centrality are: Degree, Betweenness and Closeness.

a. Network degree

The most straightforward definition of centrality is degree (sometimes called connectivity), which corresponds to the number of interactions (i.e. the number of immediate neighbors) a node has. Calculation of the degree allows determining the degree distribution P(k), which gives the probability that a selected node has exactly k links. P(k) is obtained by counting the number of nodes N(k) with k = 1, 2, 3 ... links and dividing by the total number of nodes N. Determining the degree distribution allows distinguishing different kind of graphs. For instance, a graph with a peaked degree distribution (Gaussian distribution) indicates that the system has a characteristic degree with no highly connected nodes (most of the nodes have average degree). This is typical of random, non-natural, networks. By contrast, a power-law degree distribution indicates the presence of few nodes having a very high degree. And this behaviour has been shown to be non-random. Networks displaying a degree distribution approximating a power-law are called scale-free networks. Many networks have been identified to show the characteristics of the scale free networks.

Nodes with a high degree are commonly referred to as hubs and they hold together several nodes with lower degree. This behaviour gives these networks a kind of robustness against random node deletion/failure. Additionally it also gives an opportunity to identify few network influential nodes in disease related biological networks as potential therapeutic targets.

In biological terms, the degree allows an immediate evaluation of the regulatory relevance of the node. For instance, in signaling networks, proteins with very high degree are interacting with several other signaling proteins, thus suggesting a central regulatory role of the proteins, that is they are likely to be regulatory hubs.

However, degree is a local centrality measure. It is often important to know a node’s centrality with respect to the entire network. Two common global centrality measures are betweenness and closeness (Junker et al., 2006).

b. Betweenness Centrality

Betweenness is calculated by first identifying the shortest paths between all nodes in the network. The betweenness centrality of a node n is calculated considering couples of nodes (s and t) and counting the number of shortest paths linking node s and node t and passing through the node n . Then, the value is related to the total number of shortest paths linking s and t. Thus, a node can be traversed by only one path linking s and t, but if this path is the only connecting s and t the node i will score a higher betweenness value. Nodes with many shortest paths passing through them receive a high betweenness and are sometimes referred to as bottlenecks.

The betweenness centrality of a node in a biological network, for instance a protein interaction network, can indicate the relevance of a protein as functionally capable of holding together different communicating proteins. The higher the value the higher the relevance of the protein as organizing regulatory component. Betweenness centrality of a protein effectively indicates the capability of a protein to bring together communication between distant proteins. In signaling modules, proteins with high betweenness centrality are likely to be crucial in maintaining functionality and coherence of signaling mechanisms.

c. Closeness Centrality

Closeness is calculated in a similar way. A node’s closeness is a measure of its average distance to all other nodes in the network. The closeness centrality of a node is calculated by computing the sum of the shortest path between the node and all other nodes in the graph, and then dividing by the number of nodes. Once this value is obtained, its reciprocal is calculated, so higher values assume a positive meaning in term of node proximity. Notably, high values of closeness centrality should indicate that all other nodes are in proximity to the node. In contrast, low values of closeness should indicate that all other nodes are distant from the node.

The closeness of a node in a biological network, for instance a protein-signaling network, can be interpreted as a measure of the possibility of a protein to be functionally relevant for several other proteins, but with the possibility to be irrelevant for few other proteins. Thus, a protein with high closeness, compared to the average closeness of the network, will be easily central to the regulation of other proteins but with some proteins not influenced by its activity.

The inverse of closeness is called average shortest path length. Since closeness is based on the average distance to all other nodes in the network, some nodes could still be very distant from a node with high closeness. To compensate for this, closeness should be complemented with eccentricity, which is the distance from a given node to the farthest node in the network. Note that unlike for degree, betweenness and closeness, a low eccentricity value implies high centrality.

2.2 NETWORKS IN MOLECULAR BIOLOGY

Development of several high throughput approaches in the last decade have not only increased the amount of information that we could gather to reveal important insights on the transcriptional, post-transcriptional or functional organization of an organism but they have also enabled us to start our journey to uncover the principles which hold them together. This is mainly because of the extent of information that has been possible to be collected by interrogating the cell's environment at different levels of detail. For instance, availability of modern techniques now enable us to identify the set of protein-protein interactions, genetic interactions, metabolic maps and small molecule interactions at a whole-organism level. Perhaps the most common form of interaction graphs which have been studied since the early days of genome sequencing are protein interactions.

2.2.1 Protein-Protein Interactions

In order to fulfil their function, proteins interact with other substances (molecules, ions, DNA, etc.) or other proteins. Proteins interact in numerous different contexts and with different outcomes. Some proteins activate or deactivate other proteins by binding to them or by (de)phosphorylating them. In the process of (de)phosphorylation, a phosphate group is (removed)added from a protein, which activates or deactivates the protein. Some proteins bind to each other, creating so-called protein complexes. These have important roles in the entire cell. Another class of proteins bind to each other to create structural complexes which give the cell its 3-dimensional structure. Yet other proteins pass on signals by interacting with source and destination proteins in so-called signaling pathways. Transcription factors are proteins that bind to DNA to activate the transcription process of a gene. This activation often requires multiple transcription factors to interact and also bind to the DNA. Thus, the elucidation of protein interactions is a central problem in biology. Unless we understand the complex interaction patterns of the tens of thousands of proteins that constitute our proteome, we cannot hope to attempt to efficiently combat some of the most important diseases.

A number of different approaches have been in use towards reconstructing the interactions between proteins. The literature comprises studies that use high throughput experiments to find if there exist pairwise interactions between a large set of query proteins. Other studies use computational modeling to determine which proteins may bind to each other based on their (predicted) structural properties. Some of the most popular and widely used experimental and computational techniques are explained below:

a. Yeast two hybrid (Y2H) (Fields and Song, 1989) **:** It is the genetic method that uses the transcriptional activity as a measure of protein-protein interaction. Two hybrid proteins are created corresponding to the proteins A and B between which the interaction is to be identified. One protein(say A) is bind with the DNA binding domain. Other protein(B) is fused with the transcription activation domain. These two hybrids are expressed in a cell containing reporter genes. The positive expression of the reporter genes identifies the interaction between the proteins A and B.

b. Co-localization: These methods work on the hypothesis that the genes which physically interact should be present in physically close proximity in the genome.

c. Co-occurrence: This method exploits the co-occurrence of homologous pairs of genes across multiple genomes. The fact that a pair of genes express together across many different species suggest that these genes are functional associated or physically interacting.

As a result of the high throughput experiments and computational techniques, many databases have been designed and setup to store the protein-protein interaction data. These databases usually are the results of integration of diverse data-sets. Some of the databases providing information about the protein-protein interactions are given below:

1. Biomolecular Interaction Network Database - BIND
2. Database of Interacting Proteins - DIP
3. Search Tool for the Retrieval of Interacting Genes/Proteins - STRING
4. General Repository for Interaction Datasets - GRID
5. Human Protein Reference Database - HPRD
6. Molecular Interactions Database –MINT

PPI Networks

The protein-protein interaction network is generally represented as undirected graph G(V,E), where the set of nodes V are the proteins. An edge (p1, p2)->E is present if there is an interaction between the two proteins p1 and p2. Multiple sources of protein interaction networks from different studies and databases represent the protein interaction networks differently.

2.2.2 Metabolic Networks

Another class of networks which are commonly studied is that of metabolic networks. They comprise of representing the metabolites and enzymes involved in catalyzing metabolic reactions as the nodes and edges in a directed network. The metabolic network maps are likely the most comprehensive of all biological networks. Most of the work on understanding metabolic networks relies on either manually curated or semi-automated metabolic databases such as the kyoto encyclopedia of genes and genomes (KEGG) and Metacyc which are available for a wide range of model organisms (Caspi et al., 2008; Grossetete et al., ; Kanehisa et al., 2008). Recently, Duarte et al. published a comprehensive literature-based genome-scale metabolic reconstruction of human metabolism with 2,766 metabolites and 3,311 metabolic and transport reactions. An independent manual construction by Ma et al. contains nearly 3,000 metabolic reactions, organized into about 70 human-specific metabolic pathways.

2.2.3 Regulatory Networks

Mapping of the human regulatory network is in its infancy, making this network perhaps the most incomplete among all biological networks. Data generated by experimental techniques, such as ChIP-on-chip and ChIP-Sequencing, have started to be collected in databases such as Universal Protein Binding Microarray Resource for Oligonucleotide Binding Evaluation (UniPROBE) and JASPAR. Literature-curated and predicted protein-DNA interactions have been compiled in various databases, such as TRANSFAC and the B-cell interactome (BCI). Human post-translational modifications can be found in databases such as Phospho.ELM, PhosphoSite, and phosphorylation site database (PHOSIDA).

2.2.4 RNA networks

RNA networks can refer to networks containing RNA-RNA or RNADNA interactions. Recently, with the increased understanding of microRNAs' role in disease, microRNA-gene networks have been constructed using predicted microRNA targets available in databases such as TargetScan, PicTar, microRNA, miRBase, and miRDB. The number of experimentally supported targets is also increasing, which are now compiled in databases such as TarBase and miRecords.

Organisms respond to continuous variations in internal and external cellular conditions by orchestrating their responses depending on the environmental challenges they are faced with. This involves the usage of a complex network of interactions among different proteins, RNA, metabolites and several other cellular entities, which undergo rewiring when perturbed by small molecules such as chemicals or drugs. The interaction between different chemicals and cellular entities can be represented in the form of a network- so called Drug-Target network.

Table 2. *Summary of the aspects of network theory that pertain to biological networks*

Degree distribution and hubs	In a random network, most nodes have approximately the same number of links, and highly connected nodes (hubs) are quite rare. The fraction of links with a given degree, called the degree distribution, follows the well-known Poisson distribution. In contrast, many real networks, including human protein-protein interaction and metabolic networks are scale-free, which means that the degree distribution has a power-law tail, i.e., the degree distribution $P(k)$ with degree k follows $P(k) \sim k-\gamma$, where γ is called the degree exponent. The most noticeable consequence of this property is the presence of a few highly connected hubs that hold the whole network together. The biological role and dynamical behavior of hubs allowed their classification into “party” hubs, which function inside modules and coordinate specific cellular processes, and “date” hubs, which link together rather different processes and organize the interactome.
Small world phenomena	Most complex networks (including random networks) display the small world property, which means that there are relatively short paths between any pair of nodes. This observation means that most proteins (or metabolites) are only a few interactions (or reactions) from any other proteins (metabolites). Therefore, perturbing the state of a given node can affect the activity of most nodes in their vicinity as well as of the behavior of the network itself
Motifs	Some subgraphs (a group of nodes that link to each other forming a small subnetwork within a network) in biological networks appear more (or less) frequently than expected given the network’s degree distribution. Such subgraphs are often called motifs, and they are likely associated with some optimized biological function (e.g., negative feedback loop, positive feed forward loop).

Modules	Most networks display a high degree of clustering, implying the existence of topological modules, representing highly interlinked local regions in the network. While the identification of such modules can be computationally challenging, a wide array of network clustering tools have emerged in the past few years.

2.3 NETWORK BIOLOGY APPROACH TO COMPLEX DISORDERS

Networks have an important role in systems biology, as a method to organize large numbers of disease-associated genes, or analyze cellular signaling pathways (e.g. the insulin signaling pathway). Network approaches offer an improved understanding of the relationship between the genes implicated in diseases and may be a valuable resource to find candidate disease genes. It has been reported that the Mendelian component of complex diseases, such as for example breast cancer, represent less than 30% of its incidence. In the particular case of breast cancer and the BRCA1 and BRCA2 genes, it is a mere 5% of all cases. Furthermore, the recent results of the many GWAS undertaken in recent years have shown that a large amount of disease-causing genes are yet to be accounted for. To explain the missing causal factors of complex disease, it is suggested future investigations should focus not on the genes in and of themselves, but rather on the effect of the interaction at various –omics and environmental level.

Exploring associations between various diseases by using multi-omics information is expected to improve our current knowledge of disease relationships, which may lead to further improvements in diseases diagnosis, prognosis and treatment (Park et al., 2009). Recent research has increasingly demonstrated that many seemingly dissimilar diseases have common molecular mechanisms and strong associations among them (Yu et al., 2015). Because of the associations among diseases, multiple diseases occur together in a patient, which is called disease comorbidities. Comorbidity associations can be due to direct or indirect causal relationships and the shared risk factors among them (Tong and Stevenson, 2007). If two diseases have comorbidity association, the incidence of one of them in an individual may increase the likelihood of another disease occurring. Certain diseases, such as diabetes and obesity often co-occur in the same patient, sometimes one being considered a significant risk factor for the other (Lee et al., 2008). Disease comorbidities are increasingly placing a greater burden on individuals, societies and health care services. It is an important factor for better risk stratification of patients and treatment planning.

Diseases with similar molecular, environmental, and lifestyle risk factors may be comorbid in individuals or may be risk factors for another disorder (Davis et al., 2010). Shared genetic, environmental and lifestyle factors have similar consequences, increasing the co-occurrence of associated diseases in the same individual. So, a person diagnosed for a combination of disorders and exposed to particular environmental, lifestyle and genetic risk factors may be at an increased risk of developing several other genetically and environmentally associated diseases (Barabási et al., 2011). It is now well accepted that phenotypes are determined by genetic material under environmental influences. For instance, many well-known and influential lifestyle factors such as smoking, diet, and alcohol intake are actively related to diabetes type 1 and type 2, and obesity (Astrup, 2001). Moreover, many complex diseases, such as cancer and diabetes, are affected by an integrated effect of environment and epistasis among many genes (Davis et al., 2010).

Recently, genome-wide association studies (gwas) proved to be useful as a method for exploring phenotypic associations with diseases (Lewis et al., 2011). Single-nucleotide polymorphisms (SNPs), a variation of a single nucleotide, are assumed to play a major role in causing phenotypic differences between individuals. It has become possible to assess systematically the contribution of common SNPs to complex diseases.

Most of the research works focussed on a particular data type, for example gene expression, to find profiles that are associated with particular disease, prognosis and drug response. The integrative analysis of various omics data has become increasingly widespread because each approach has intrinsic caveats. For instance, important information may be missing because of false negatives or may be misleading because of false positives. In addition, by analyzing different types of data in isolation we may miss important information that results from the coordinated activity of biological components at various levels. Some studies indicated that these limitations can be mitigated by integrating two or more omics datasets. Several studies (Goh et al., 2007;

Lee et al., 2008; Lu et al., 2008; Hu and Agarwal, 2009; Liu et al., 2009;Park et al., 2009; Schadt, 2009; Jiang et al., 2010; Suthram et al., 2010) reported on the role of a single omic or phenotypic measure to represent disease-disease associations (such as shared pathways or gene ontology). But, one needs to study diverse sources of evidence including shared genetic factors, ontology, SNPs, and phenotypic manifestations for better understanding.

Since, diseases may share many different types of associations with varying levels of risk for disease comorbidities, a singular view of associations between diseases is not enough to predict comorbidities. As more and more ontology, phenotype, omics and environmental data sets become publicly available, it is beneficial to improve our understanding of human diseases and diseases comorbidities based on these new system-level biological data. Combination of multiple types of omics, phenotype and ontology data identifies integrative biomarkers for the stratification of patients with clinical outcome. Therefore, it is clear that network based methods and tools are of critical importance in the field of medicine.

Network approaches for the study of the aging and age-related disorders(ARDs)

Aging is one of the most multifactorial, complex processes of living organisms. In spite of this complexity, until very recently the majority of studies examined separate elements of the aging process. The multiplicity of approaches has contributed to the large number of aging theories and definitions (Simko et. al, 2009).

- According to the antagonistic pleiotropy theory of aging, genes, which are preferable during early development, become detrimental in the aged organism.

- The disposable soma theory of ageing highlights the relocation of resources from somatic maintenance towards increased fertility leading to a slow deterioration .

- The reliability theory gives a rather descriptive picture of aging emphasizing that aging is a phenomenon of increasing risk of failure with the passage of time.

- The network theory of ageing integrates many elements of the previous theories and describes the shifting balance between various types of damage and repair mechanisms in aging of cellular systems. The young state is characterized by well-repaired damage, while the aged organism cannot cope with the accumulated damage and gradually surrenders.

Aging is accompanied by a number of age-related diseases, cancer, atherosclerosis, and diabetes and neurodegenerative disorders, such as Alzheimer's disease, Parkinson's disease or others. For the understanding and complex treatment of these interrelated diseases we need novel approaches. The network approach proved to be a highly efficient tool to describe complex system behaviour including the aging process and age-related diseases. Networks provide a framework for the conceptualization of the aging process, but can also be used to understand aging in many ways. It is therefore not surprising that the analysis of the phenomenon of aging can be expanded by the application of a network approach which will give us several novel approaches to understand the aging process better and to cure age-related diseases (ARDs) in entirely novel ways.

We propose a computational framework to construct ARD networks that integrates genetic alteration (GWAS) data with standardized textual descriptions of gene functions and processes(GO) and their inter-relationships in order to characterize the mechanistic underpinnings of diseases. We have developed a user-friendly web-based interface that allows visualization of various ARD networks at various –omic levels providing detail insight into the comorbidity of various ARDs. The goal of developing this interface is to allow physicians and researchers to visualize relationship between various ARDs by incorporating disease interactions, omics, and ontology information. We believe that combination of different types of data will result in a much deeper understanding of age related changes as well as identify the causes for variability across individual.

PART III: MATERIALS/TOOLS USED

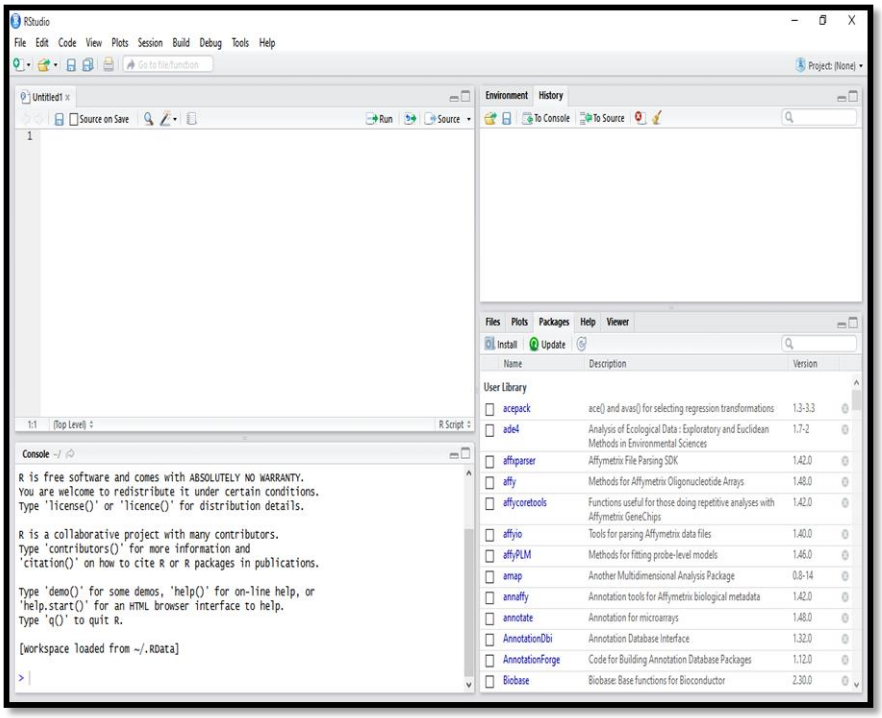
3. MATERIALS/TOOLS USED

3.1 R 3.2.2

R is a free software environment (publically available) used for graphics and statistical computing. It runs and compiles on variety of OS (UNIX platforms, MacOS and Windows. Before downloading R, preferred CRAN mirror was chosen (i.e. INDIA). R software is available at https://www.r-project.org/ .

3.2 RStudio

It is a set of integrated tools designed for the user which enable them to be more productive with R. It includes editor for syntax-highlighting (supports direct code execution), tools for plotting, history, a console, as well as provides debugging and workspace management (Figure 1).



```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> 
```

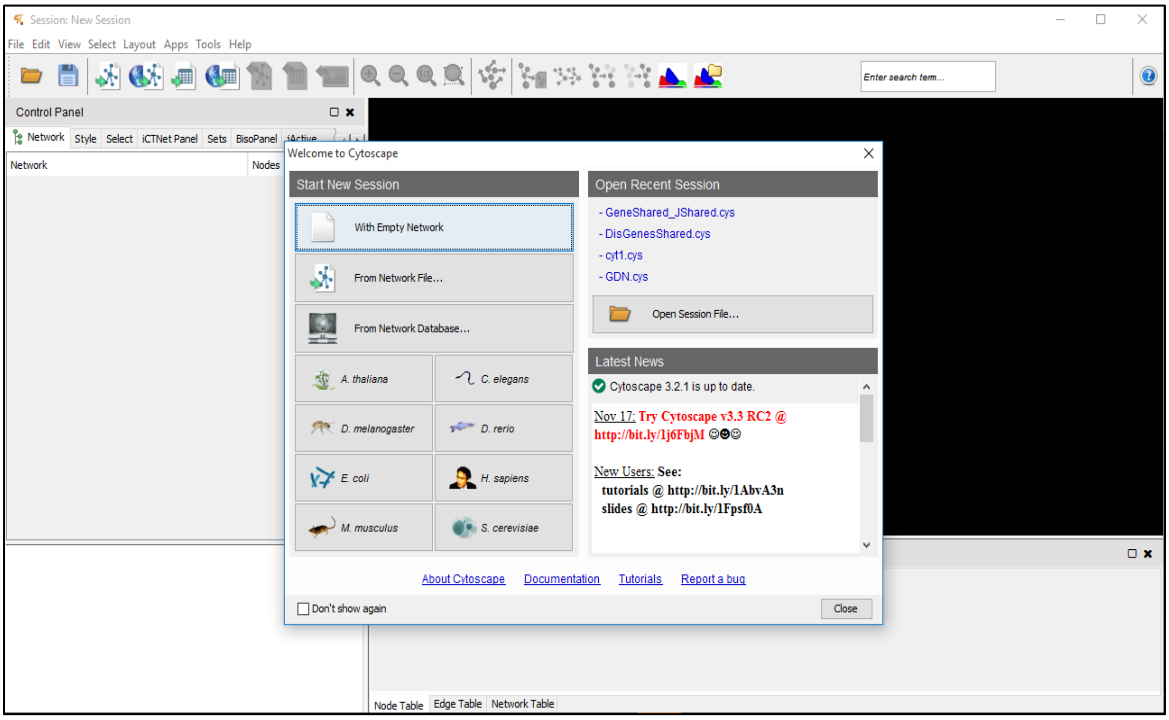
User Library

	Name	Description	Version
□	acepack	ace() and avas() for selecting regression transformations	1.3-3.3
□	ade4	Analysis of Ecological Data : Exploratory and Euclidean Methods in Environmental Sciences	1.7-2
□	affxparser	Affymetrix File Parsing SDK	1.42.0
□	affy	Methods for Affymetrix Oligonucleotide Arrays	1.48.0
□	affycoretools	Functions useful for those doing repetitive analyses with Affymetrix GeneChips	1.42.0
□	affyio	Tools for parsing Affymetrix data files	1.40.0
□	affyPLM	Methods for fitting probe-level models	1.46.0
□	amap	Another Multidimensional Analysis Package	0.8-14
□	annaffy	Annotation tools for Affymetrix biological metadata	1.42.0
□	annotate	Annotation for microarrays	1.48.0
□	AnnotationDbi	Annotation Database Interface	1.32.0
□	AnnotationForge	Code for Building Annotation Database Packages	1.12.0
□	Biobase	Biobase: Base functions for Bioconductor	2.30.0

Figure 1. *Window of R studio showing different tools. Upper left corner is provided with the tools for writing the source codes, and there is console at the right bottom of window, left upper corner is for history and the left end corner is for plots, packages, help etc.*

3.3 Cytoscape

Cytoscape (Figure 2) is an open source bioinformatics software platform for visualizing molecular interaction networks and integrating with gene expression profiles and other state data. Additional features are available as plugins. Plugins are available for network and molecular profiling analyses, new layouts, additional file format support and connection with databases and searching in large networks.



Cytoscape 3.2.1 is up to date.

Nov 17: Try Cytoscape v3.3 RC2 @ http://bit.ly/1j6FbjM ☺☻☺

New Users: See:
tutorials @ http://bit.ly/1AbvA3n
slides @ http://bit.ly/1Fpsf0A

Figure 2. *Window of Cytoscape, Different panels for different purposes viz visualization, control panel, table panel and result panel.*

3.4 NetBox 1.0 Software

NetBox is a Java-based software tool for performing network analysis on human interaction networks. It is pre-loaded with a Human Interaction Network (HIN) derived from four literature curated data sources, including the Human Protein Reference Database (HPRD), Reactome, NCI-Nature Pathway Interaction (PID) Database, and the MSKCC Cancer Cell Map. Currently, NetBox provides the analyzeNet.py method that provides a simple command line interface for connecting genes into a network, identifying statistically significant "linker" genes, partitioning the network into modules, and executing two random background models. Results are then made available to the end user as an HTML web page and a series of network and attribute files, which can be loaded into Cytoscape for visualization and further analysis.

3.5 FunRich

FunRich is a stand-alone software tool used mainly for functional enrichment and interaction network analysis of genes and proteins. Besides, the results of the analysis can be depicted graphically in the form of Venn, Bar, Column, Pie and Doughnut charts. Currently, FunRich tool is designed to handle variety of gene/protein data sets irrespective of the organism. Users can not only search against default background database, but can also load customized database against which functional enrichment analysis can be carried out.

FunRich database option currently supports the enrichment analysis of the following categories:

- Biological process
- Cellular component
- Molecular function
- Protein domains
- Site of expression (normal tissues, cancer tissues, cell types and cell lines)
- Biological pathways
- Transcription factors
- Clinical synopsis phenotypic terms

PART IV: METHODOLOGY

4. METHODOLOGY

The workflow for the methodology is shown in Figure 3 and the detail of each step is described as follows.

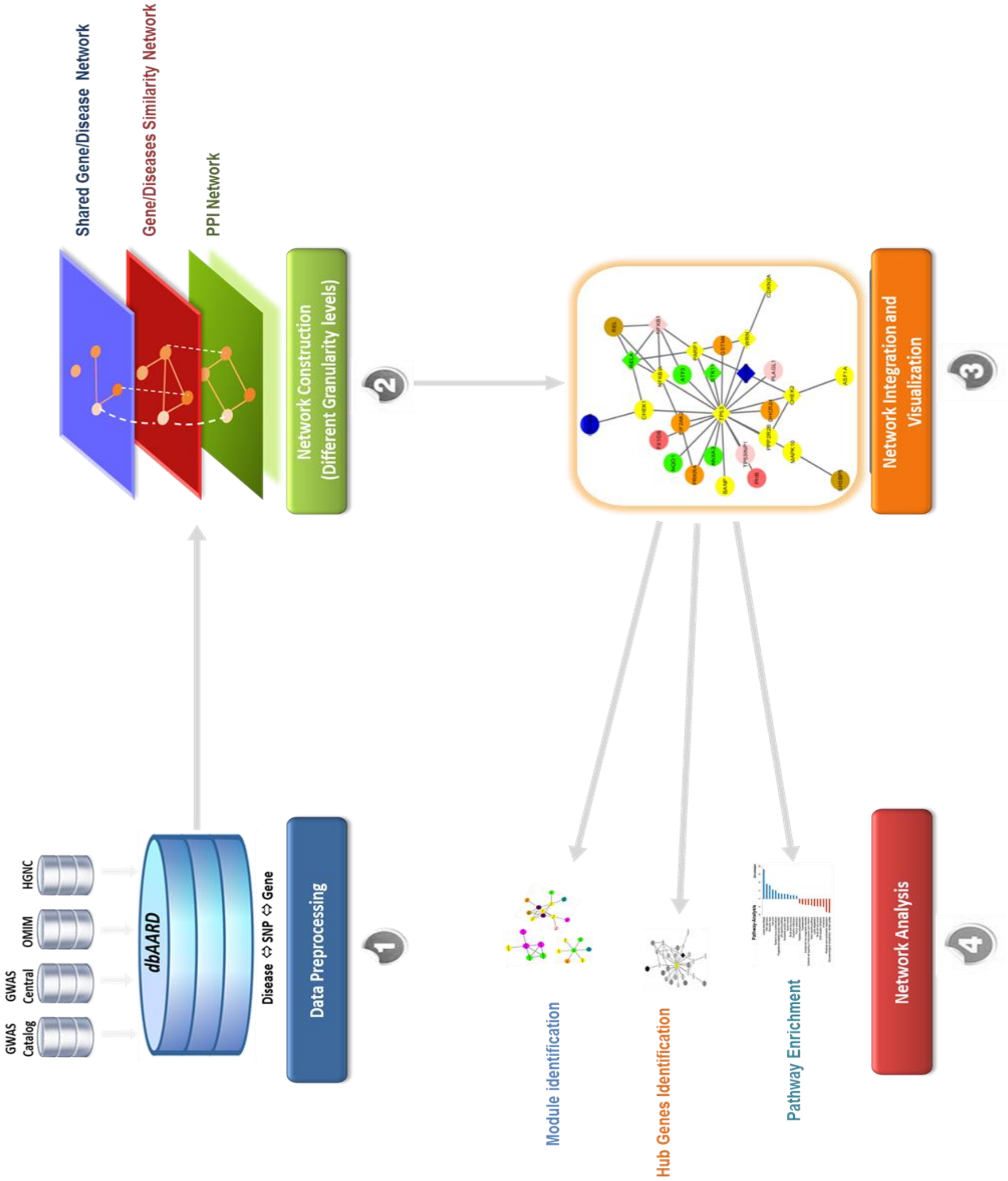


Figure 3. *Workflow for the methodology adopted.*

4.1 Data Collection and Preprocessing

4.1.1 Disease-Gene-SNP associations

The initial data for the study was obtained from the freely available comprehensive database – dBAARD (http://genomeinformatics.dtu.ac.in/dbAARD/) (Srivastava et. al., 2016). dbAARD is an interactive database that contains information on human age-related disorders and the associated human SNPs with supporting evidence from different sources (NHGRI GWAS Catalogue, GWAS Central, OMIM, HGMD and others). dbAARD catalogues information about human genetic variants associated with various ARDs like Alzheimers' disease, Parkinson's disease, Diabetes, Cardiovascular disorders and Cancers etc. Each entry in dbAARD contains the information on SNP, the associated disease, p-value assigned to the association, the odd ratio of the disease-SNP association and the literature reference. The database can be queried individually or in combination of disease class, disease name, gene and rsIDs.

The data obtained from dbAARD comprised of 3197 SNPs across 1297 genes associated with 53 ARDs falling under 12 classes. Analysis of SNPs associated with ARDs depicted 75.33% as noncoding; 1.02 % as coding –synonymous, 20.52 % SNPs as missense, and the rest were those SNPs whose locations were unknown in dbSNP. A total of 1762 of intragenic SNPs associated were included with p-values<0.05.

Host genes of intragenic SNPs were assigned using dbSNP. In spite of the rare instances where a SNP could lead to two distinct genes, in dbSNP each SNP is uniquely mapped to a single host gene. As a result, 533 disease host genes were mapped from the 1762 distinct disease-associated intragenic SNPs and were analysed further to construct ARD networks at various level of information.

Table 3: *Summary of information in disease-SNP list*

	dbAARD	After processing step
Diseases	53	53
SNPs	3197	1762
Genes	1297	533
Diseases Classes	12	12

4.1.2 GO Annotations

Genes are annotated with GO terms to represent their biological properties. We downloaded the ontology file and annotations of Homo sapiens from the Gene Ontology database (http://www.geneontology.org). We removed annotations with evidence code 'Inferred from Electronic Annotation' (IEAs), since IEAs are computationally inferred annotations which have not been reviewed by curators. GO terms classified as 'GO biological processes' (GO:BP) was identified and analyzed as semantic similarity metric as described below.

4.1.3 PPI Network

To construct a human interactome, we obtained 35,021 protein-protein interactions (PPIs) pertaining to 9462 proteins from the Human Protein Reference Database (HPRD) database (release 7), as it is known to be one of the most reliable databases for PPI data.

4.2 Construction and Visualization of ARD Networks at different levels of granularity

In recent years there has been a trend toward studying disease through network based analysis of various systems of connections between diseases. The linkage of a gene to various diseases often indicates that these diseases have a common genetic origin. Motivated by this hypothesis, Goh et al. used the gene–disease associations that are collected in the OMIM database to build a network of diseases that are linked if they share one or more genes (Goh et al., 2007). This resulted in the Human Disease Network (HDN). The nodes in the HDN represented human genetic disorders and the edges represented shared genes between disorders. The underlying connections of the HDN contributed to the understanding of the basis of disorders, which in turn led to a better understanding of human disease. However, the study by Goh et. al., was limited in analysis to the genes shared by different diseases. Another study by Li et al. traced the SNPs connecting disease traits (Li et. al., 2011). In 2009, Barrenas et al. studied genetic architecture of complex diseases by doing a GWAS, and found that complex disease genes are less central than the essential and monogenic disease genes in the human interactome (Barrenas et. al., 2009). In this work we explored methods of building the ARD networks that go beyond previously mentioned gene-centric disease network approaches.

4.2.1 Network construction using Shared Genetic Architecture Hypothesis

In our study, we started by building bipartite networks, consisting of two disjoint sets of nodes. The nodes are connected in such a way that the nodes of one set will have no connections between them, but can only be connected to nodes of the other set. The use of a bipartite network is natural when dealing with two different types of data sets (Figure 4b), in our case diseases (e.g. the rectangles) and SNPs or genes (e.g. the circles). This type of network gives us three distinct degree distributions, one for each projection, and one for the bipartite network. Each degree distribution shows how many links each node has. Nodes in a projection of a bipartite network are connected if they share at least one node in the other group. This gives us the ability to see the interactions within each set.

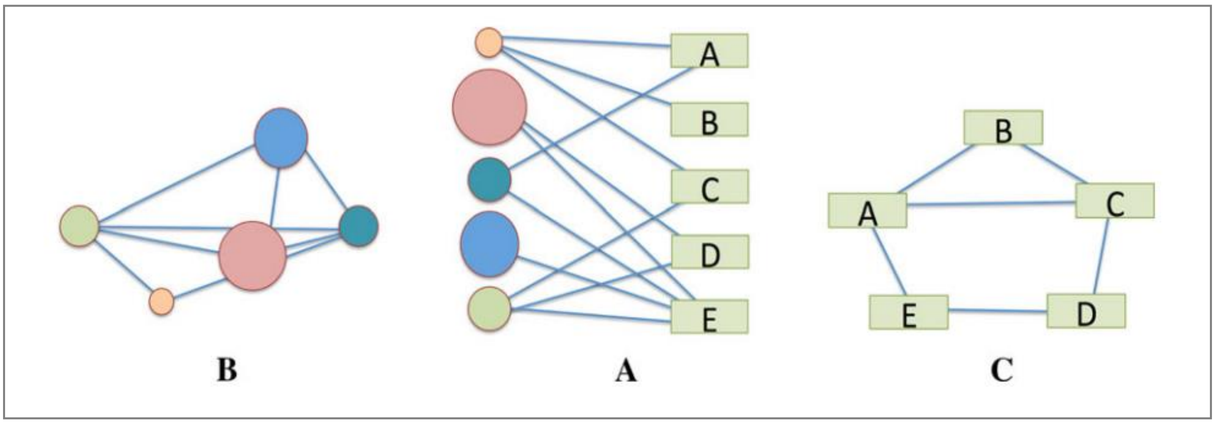


Figure 4. *Bipartite Network schematic. A bipartite network (b) made of 2 data sets the "circles", and the "rectangles". Projections in the "circle" space (a) and in "rectangle" space (c).*

The data from the bipartite network can be projected onto either data space (Figure 4a, c). In both cases, the nodes are connected to one another through a vertex of the other space. By ignoring the different types of data, all network properties described above remain valid on the bipartite network (as a single data set network) and on either projection.

The following sections present our methods for building the ARD Networks based on different sharing elements. We start at the smallest element, the SNP, then move on to SNP clusters, and finally to genes. These offer varying density of the information contained with both the bipartite network and the projected disease networks. The networks were visualized using Cytoscape 3.2.1.

a. Genetic Variations (SNP) based ARD Networks

We started with a bipartite graph of SNP-Disease consisting of two disjoint sets of nodes. One set corresponded to all known genetic disorders, whereas the other set corresponded to all known variants in the human genome. A disorder and a SNP were then connected by a link if the disease-SNP association were present in data fetched from dbAARD.

Linking diseases that share at least one SNP, we built SNP-ADN. The disorders were linked based only on shared genetic variants, i.e. overlapping SNPs. Two diseases were connected if they share at least one SNP that is statistically significant dysregulated to the disease related gene. The resulting ARD variome will allow us to establish connections between diseases/traits that share blocks i.e. that have overlapping SNPs.

b. Gene based ARD Networks

Starting from the ARDome bipartite graph constructed of two disjoint sets of nodes in which one set of nodes represented ARDs and the other set represented genes, we generated two biologically relevant network projections – ADN (Age-related Disorders Network) and AGN (ARD associated Gene Network). In the ADN nodes represented disorders, and two disorders were connected to each other if they share at least one gene in which mutations were associated with both disorders. In the AGN nodes represented disease genes, and two genes were connected if they were associated with the same disorder.

4.2.2 Network construction using Semantic Similarity Approach

Measuring similarity between diseases plays an important role in disease-related molecular function research. Functional associations between disease-related genes and semantic associations between diseases are often used to identify pairs of similar diseases from different perspectives. The quantitative measurement of similarity between diseases based on qualitative association plays an important role in predicting disease-causing genes, inferring microRNA function associations and identifying novel drug indications (Cheng et. al., 2014).

Several methods have been developed for calculating gene/disease similarity. The most commonly used are Semantic-based methods. Semantic-based methods are widely used for measuring similarity between terms of Gene Ontology (GO) (Ashburner et. al., 2000 ; Pesquita et. al., 2009) in the biomedical and bioinformatics domain. The use of semantic similarity between biological processes to estimate disease association could enhance the identification and characterization of disease association besides identifying novel biological processes involved in the diseases. Graph-based methods using the topology of GO graph structure is used to compute semantic similarity. We adapted the approach for computing the functional similarity of GO terms from Li et.al. (2012). To determine functional similarity among genes, we used the Biological Process category of Gene Ontology classification.

The similarity of two genes is conceptually defined by the similarity of their GO annotations as measured by their shared information content. Specifically, the semantic similarity of two terms is defined as the information content of their minimal ancestor in GO (common ancestor with maximal information content) divided by the average information content of the two terms, where the information content of a single term is the probability of the term and its sub-terms being selected randomly in GO (Li et. al., 2012). There are many measures of semantic similarity and one of the popular method is Lin method. Thus, the calculation of gene similarity score was done using Lin metric as defined below. Firstly the semantic similarity between the GO terms as

$$Sim(a,b)=2*sim(ms(a,b))sim(a)+sim(b)$$

$$ic(a)=-\log(|G(a)||G(A)|)$$

where $ic(a)$ is the information content of GO term a, $ms(a,b)$ is the minimal ancestor of terms a and b, $G(a)$ is the sub-graph of GO rooted at a, A is the root term of the GO, and the function '$|G(a)|$' is the cardinality of $G(a)$ measured as the count of distinct terms in this sub-graph.

The information similarity of two genes (Gene_Sim), such as two SNP host genes, was then measured by the average best-matching pair similarity between their annotated GO terms. For any GO term annotated to a gene, its best-matching term pair in another gene's GO annotation list is the one with the maximum term–term similarity as compared with all other terms from the other gene. Furthermore, only the most reliable subset of the best-matching term pairs across the two term sets is retained in the calculation, while all other term pairs are ignored because of the annotation noise in GO. Mathematically, the information similarity of two genes is defined as:

$$Gene\_Sim(\alpha,\beta)=2\times\sum_{(ai,bi)\in\pi,ITS(ai,bi)\geq t} ITS(ai,bi)/\ (|\alpha|+|\beta|)$$

where α and β are two genes being annotated to two term sets, the included best-matching term pairs are represented as a relationship π with pairs a_i and b_i, and t is the similarity threshold for any term pair to be included in the calculation for further annotation noise reduction (set as 0.7 in our implementation). The similarity of two genes is based on their number of shared GO terms and, if the terms were not identical, the term proximity in the GO graph. The information similarity of two genes is normalized to the range of 0 to 1, corresponding to genes with no similar annotations and genes with equivalent annotations,

There are many online tools, standalone R based packages that allow the measure of semantic similarity between GO terms using different metrics like GOssTO, FunSim, GoSemSim, InteGO2 etc (Caniza et. al, 2014 ; Yu et, al., 2010 ; Schlicker et. al., 2010 ; Peng et. al., 2014). Since the calculation of semantic similarity score is computationally intensive we used GOSemSim Package (Yu et. al., 2010) in R studio to calculate the semantic similarity between the ARD gene pairs using Lin metric. Steps to install and run GOSemSim are described below:

1. Download the latest version of R and R studio.
2. Start R studio and type the following commands in the console window.

```
>source("https://bioconductor.org/biocLite.R")
>biocLite("org.Hs.eg.db")
>biocLite("GOSemSim")
>library(GOSemSim)
#mgeneSim to calculate semantic similarity among multiple gene products.
>res<-
mgeneSim(c("28","59","105","108","202","288","308","326","341","348","354","463","472","490","4
93","501","627","629","636","638","640","717","718","775","783","841","868","923","945","965","9
94","999","1002","1012","1021","1029","1116","1136","1191","1201","1235","1277","1282","1295","
1308","1312","1378","1380","1394","1404","1488","1493","1559","1586","1594","1609","1636","174
0","1767","1778","1788","1804","1830","1952","2037","2068","2099","2104","2153","2162","2200","
2201","2212","2246","2255","2262","2263","2494","2524","2580","2646","2700","2859","2897","291
5","2917","2972","3075","3084","3092","3107","3117","3122","3482","3556","3559","3570","3575","
3606","3613","3643","3646","3655","3658","3662","3663","3684","3687","3700","3710","3753","376
7","3782","3784","3798","3827","3911","3988","4018","4041","4046","4088","4092","4137","4157","
4211","4233","4295","4354","4439","4477","4482","4485","4507","4524","4600","4609","4642","464
4","4745","4794","4811","4853","4855","4864","4905","5071","5122","5125","5314","5328","5334","
5357","5468","5607","5636","5649","5654","5789","5793","5819","5890","5894","5906","5915","593
7","6014","6125","6205","6239","6310","6366","6403","6499","6568","6581","6597","6622","6711","
6720","6774","6775","6925","6927","6928","6934","7015","7018","7074","7077","7091","7096","712
8","7132","7140","7148","7185","7187","7253","7297","7299","7332","7369","7410","7444","7466","
7472","7473","7532","7709","8000","8131","8224","8398","8477","8613","8626","8631","8638","880
7","8863","8897","8924","9031","9037","9103","9333","9373","9425","9429","9467","9497","9531","
9586","9659","9842","9948","10133","10144","10198","10217","10257","10279","10318","10347","1
0452","10466","10497","10611","10644","10665","10666","10714","10758","11046","11062","11116
","11138","11262","22808","22834","22848","22853","22891","22913","22926","22998","23025","23
043","23095","23112","23180","23263","23274","23301","23321","23534","23544","23704","23788",
"25780","25861","25902","25970","25976","26147","26191","26228","26797","27074","27185","273
28","27347","29086","29945","29951","29994","50807","51131","51151","51196","51230","51555","
51654","53339","53942","54209","54535","54622","54749","54790","54894","54899","54901","5497
1","55016","55024","55054","55133","55553","55600","55692","55759","55819","55892","55937","5
5973","56244","56606","56776","56913","56916","56922","57111","57178","57474","57492","57608
","57628","57661","57705","58504","60468","60678","63027","63826","63892","63977","63982","64
127","64135","64167","64170","64231","64478","64710","64754","79054","79068","79083","79258",
"79660","79728","79774","79925","80003","80129","80279","80736","81037","81492","81579","818
47","84286","84446","84515","84618","84619","84624","84628","84660","84668","84700","84722","
84765","84898","84942","85415","91752","91828","94241","114134","114781","114803","114815","
114818","114876","115106","115352","116085","116113","116285","116985","120892","121260","1
22402","123624","126549","126859","127700","128869","133522","140733","140766","144406","14
4811","145781","149233","150084","150962","152189","152330","154442","159296","160777","163
059","163486","164312","164656","168620","169026","196740","201266","204010","204801","2049
62","219790","220164","220416","221692","221895","222546","253461","254428","255738","25653
6","257194","266722","284996","285362","285600","285830","341880","344148","375056","387694
","387715","388650","389170","390928","400954","414236","503835","613227","643714","644192",
"647121","728597","729967","729993","100048912","100128977","100129583","101929777","10272
3475"), ont="BP", organism="human", measure="Lin", combine="BMA")
>write.table(res,file="result.txt”)
```

3. The result.txt file stores the semantic similarity score for each gene pair in the matrix form. The file was converted into the readable network format using C code in which gene pairs with score > 0.7 were selected and then used for the construction of network using Cytoscape.

To construct semantic similarity based ARD Networks, we measured disease-disease similarity by using the shared information between the host genes of the associated intragenic SNPs, specifically the average similarity of reciprocal best-matching host gene pairs from GWAS. The best-matching pair of a host gene (γ) with respect to another trait is the host gene (δ) of the other trait with maximum gene–gene similarity with the first host gene (γ). The disease-disease similarity was defined as below (Li et al, 2012)

$$DisSim(U,V)=2\times\sum_{(\alpha i,\beta i)\in\pi'}Gene\_Sim(\alpha i,\beta i)/(|U|+|V|)$$

where U and V are two diseases representing two sets of host genes, and the reciprocal best-matching host gene pairs are represented as a relationship π' with pairs α_i and β_i. The information similarity of two traits ranges from 0 (for two traits with totally dissimilar host genes) to 1 (for two traits with identical or equivalently annotated host genes).

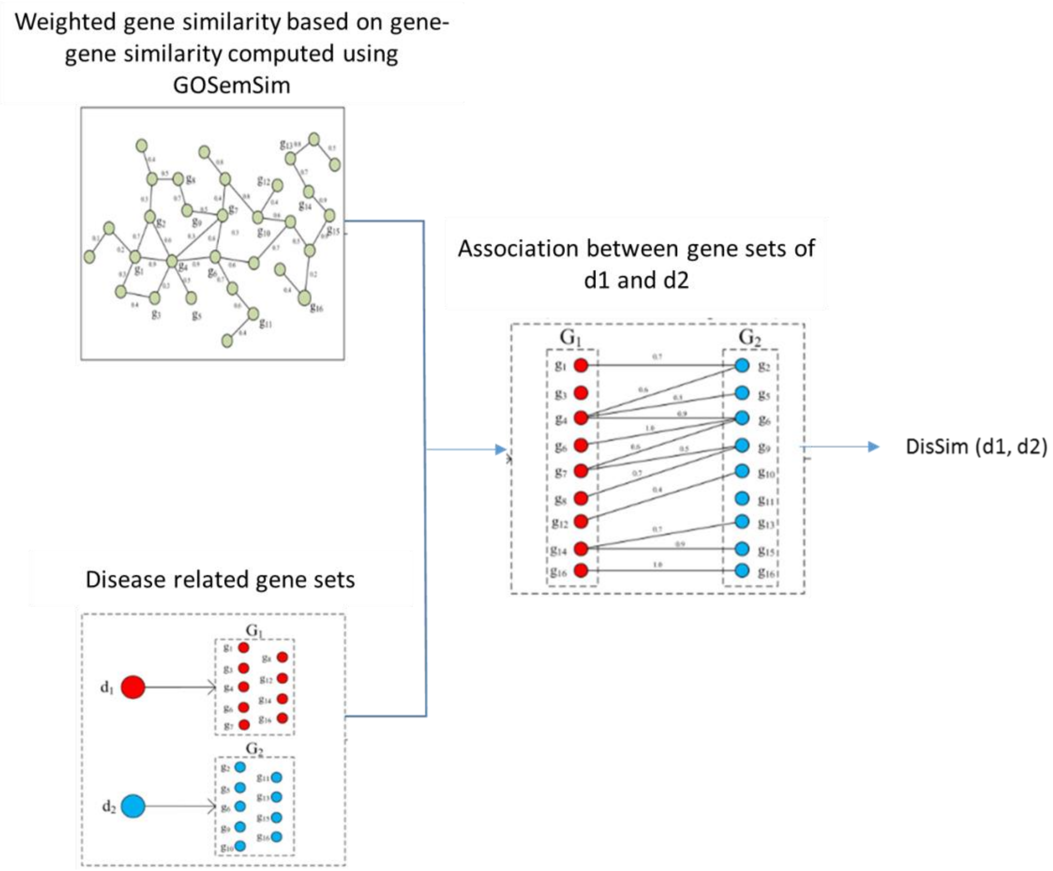

Weighted gene similarity based on gene-gene similarity computed using GOSemSim
Association between gene sets of d1 and d2
G_1
G_2
DisSim (d1, d2)
Disease related gene sets
G_1
d_1
G_2
d_2

Figure 5. *Overview of steps involved in disease similarity calculation.*

Since GOSemSim doesn't have module for disease similarity calculation, the DisSim score was calculated by writing a code in C that took gene-gene similarity score matrix as input along with the the list that specifies which genes were present in a particular diseases.

A disease–disease network was thus constructed from pairwise similarity scores directly subjected to a certain threshold, where nodes in the network represented complex ARDs, and links represented the significant biological similarity between two

ARDs. The sizes of the nodes and edges were proportional to the number of host genes and the strength of the diseases similarities, respectively.

4.2.3 Network construction using Human PPI data from HPRD

Human Protein Reference Database (HPRD) is one of the most extensively used database for research purposes and all the enclosed interaction are experimental, so it was selected for fetching PPIs of genes of interest. The PPI information was utilized in our research work for construction of various required networks at different steps. Construction of global network was achieved by mapping ARDs protein on human PPI.

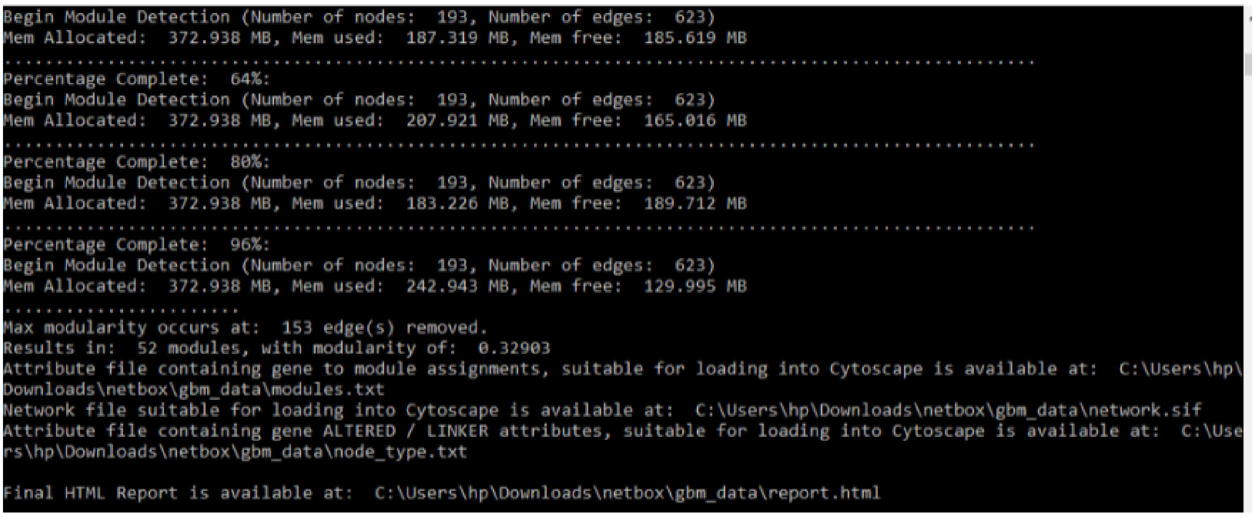
4.3 Analysis of ARD Networks

4.3.1 Functional Enrichment Analysis of the ARD associated genes

To further explore the hypothesis that the associate SNP host genes characterize different pathways to ARDs, the Functional enrichment analysis was performed using FunRich.

4.3.2 Module Identification

To identify the highly enriched modules in the ARD gene network, the NetBox tools was used (Cerami et. al., 2010). The tool NetBox provides the method to overlay the disease genes on the Human Protein-Protein Interaction network and determine the functional modules based on the cut-off shortest path length and p-value.



```
Begin Module Detection (Number of nodes:  193, Number of edges:  623)
Mem Allocated:  372.938 MB, Mem used:  187.319 MB, Mem free:  185.619 MB
...............................................................................
Percentage Complete:  64%:
Begin Module Detection (Number of nodes:  193, Number of edges:  623)
Mem Allocated:  372.938 MB, Mem used:  207.921 MB, Mem free:  165.016 MB
...............................................................................
Percentage Complete:  80%:
Begin Module Detection (Number of nodes:  193, Number of edges:  623)
Mem Allocated:  372.938 MB, Mem used:  183.226 MB, Mem free:  189.712 MB
...............................................................................
Percentage Complete:  96%:
Begin Module Detection (Number of nodes:  193, Number of edges:  623)
Mem Allocated:  372.938 MB, Mem used:  242.943 MB, Mem free:  129.995 MB
......................
Max modularity occurs at:  153 edge(s) removed.
Results in:  52 modules, with modularity of:  0.32903
Attribute file containing gene to module assignments, suitable for loading into Cytoscape is available at:  C:\Users\hp\
Downloads\netbox\gbm_data\modules.txt
Network file suitable for loading into Cytoscape is available at:  C:\Users\hp\Downloads\netbox\gbm_data\network.sif
Attribute file containing gene ALTERED / LINKER attributes, suitable for loading into Cytoscape is available at:  C:\Use
rs\hp\Downloads\netbox\gbm_data\node_type.txt

Final HTML Report is available at:  C:\Users\hp\Downloads\netbox\gbm_data\report.html
```

Figure 6. *Run of NetBox for module identification*

4.3.3 Identification of linker genes

Netbox also helps to find the genes that are not there in the disease network but through PPI interaction studies seems to be closely associated with the genes in the disease network thus indicating the likelihood of the involvement of those genes in the diseases and hence helps to predict new markers of age-related diseases (ARDs).

4.3.4 Hub Genes Identification

Hubs were identified by two approaches. One of them was identifying hubs based on degree, betweeness centrality, bottleneck and maximum clique component (MCC) score. In this approach, the top 10 hubs were identified which can be targeted by drugs and cure more than one ARDs "polyphormacology".

4.4 ARDInteract: Online Platform for ARD Interaction Networks

Finally we integrated all the information obtained from different network studies and made that available as a web interface.

PART V: RESULTS & DISCUSSIONS

5. RESULTS & DISCUSSION

5.1 ARD Networks constructed using shared genetic architecture hypothesis

5.1.1 Disease-SNP bipartite network

Disease-SNP bipartite network was constructed such that the circular nodes represented ARDs and the triangular nodes represented SNPs implicated in diseases. Figure 7 represents disease-SNP network. As it can be clearly observed that many SNPs are only connected to single diseases. However, there are few SNPs which connect more than one disease and hence can be concluded that they have pleiotropic effects. And multiple SNPs associated with single disease exhibit epistatic effect on the development of that disease.

As expected, high number of mutations is observed to be linked with cancer especially prostate and breast cancer. Myopia and Parkinson disease share many variants, which reflect some biological relationship leading to the co-manifestation of both the diseases.

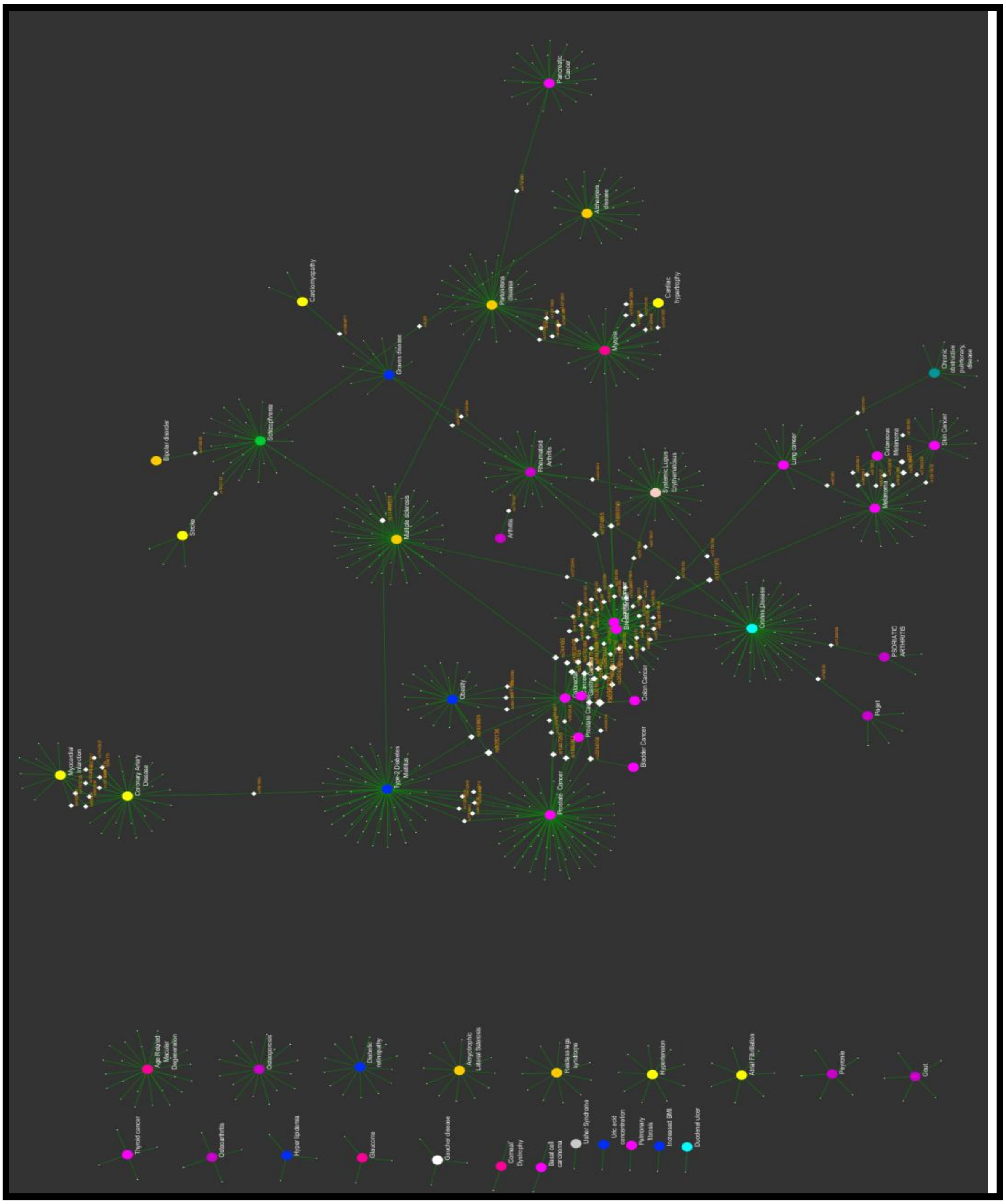


Figure 7. *Disease-SNP bipartite network (ARD- circles; SNPs – squares). The nodes representing diseases are color coded based on the disorder class. And the size of the nodes representing SNPs is proportional to the number of diseases to which SNP is associated.*

5.1.2 Disease-Gene bipartite network

Figure 8 represents disease-gene bipartite network. In order to enhance the visualization, the p-values of mapped genes were set < 10^{-7}. Similar to disease-SNP network, this network also exhibits similar properties. Few genes are responsible for many diseases and many genes together are implicated for a single disease.

a. ARD Gene Projection Network – AGN

The AGN consisted of 522 nodes and 8265 edges. A small number of genes were associated with multiple diseases, most of them connecting diseases in the giant component. Genes such as NR, HLA-DQA1, CDKN2B, IL23 were associated with more than two diseases (Figure 9(a)). Most of these genes were involved in the biological regulation and metabolic processes. HLA-DQA1 is involved in the recognition of foreign pathogens (Figure 9(b) and 9(c)).

b. ARD Disease Projection Network – ADN

ADN network provides a broader overview of interconnectedness of various ARDs. The ADN consisted of 53 nodes and 120 edges; 44 of the diseases had any interactions with other diseases. In other words, many diseases, including many types of cancer, did not share any genes with other diseases. Since GWAS generally detect genetic variants with a high minor allele frequency and large effect size (referred to as high-profile variants), the number of genes associated with a given disease is indicative of the genetic architecture of the disease. The number of genes associated with each disease varied greatly; prostate cancer and breast cancer were found to be associated with 59 and 42 genes respectively while bipolar disorder was associated with very few genes. The reason for variation in the number of genes associated could be biological or it could be due to lack of GWAS studies on the particular disease. Furthermore, some diseases of the same class were highly interconnected as in the case of metabolic disorders indicating common biological origin and some diseases belonging to the same class were not connected (for example, cardiovascular diseases).

ADN revealed some interesting connections between diseases of different classes. For example, Type 2 Diabetes Mellitus and Obesity are strongly connected to cancer.

Crohn's disease with highest degree is associated with 4 different classes of ARDs implicating the common biological mechanism or pathways in these diseases.

Functional Clustering of HDN and DGN. To probe how the topology of the ADN and AGN deviates from random, we randomly shuffled the associations between disorders and genes, while keeping the number of links per each disorder and disease gene in the networks unchanged. Interestingly, the clustering coefficient of the randomized disease networks is 0.279+/-0.041 significantly smaller than the clustering coefficient of the AND (0.51.5, p-value<10^{-4}). Similarly, the clustering coefficient of the randomized gene networks is 0.131+/-0.003, significantly smaller than the actual clustering cofficient of the AGN (0.925, p-value10^{-4}). These differences suggest important pathophysiological clustering of disorders and disease genes.

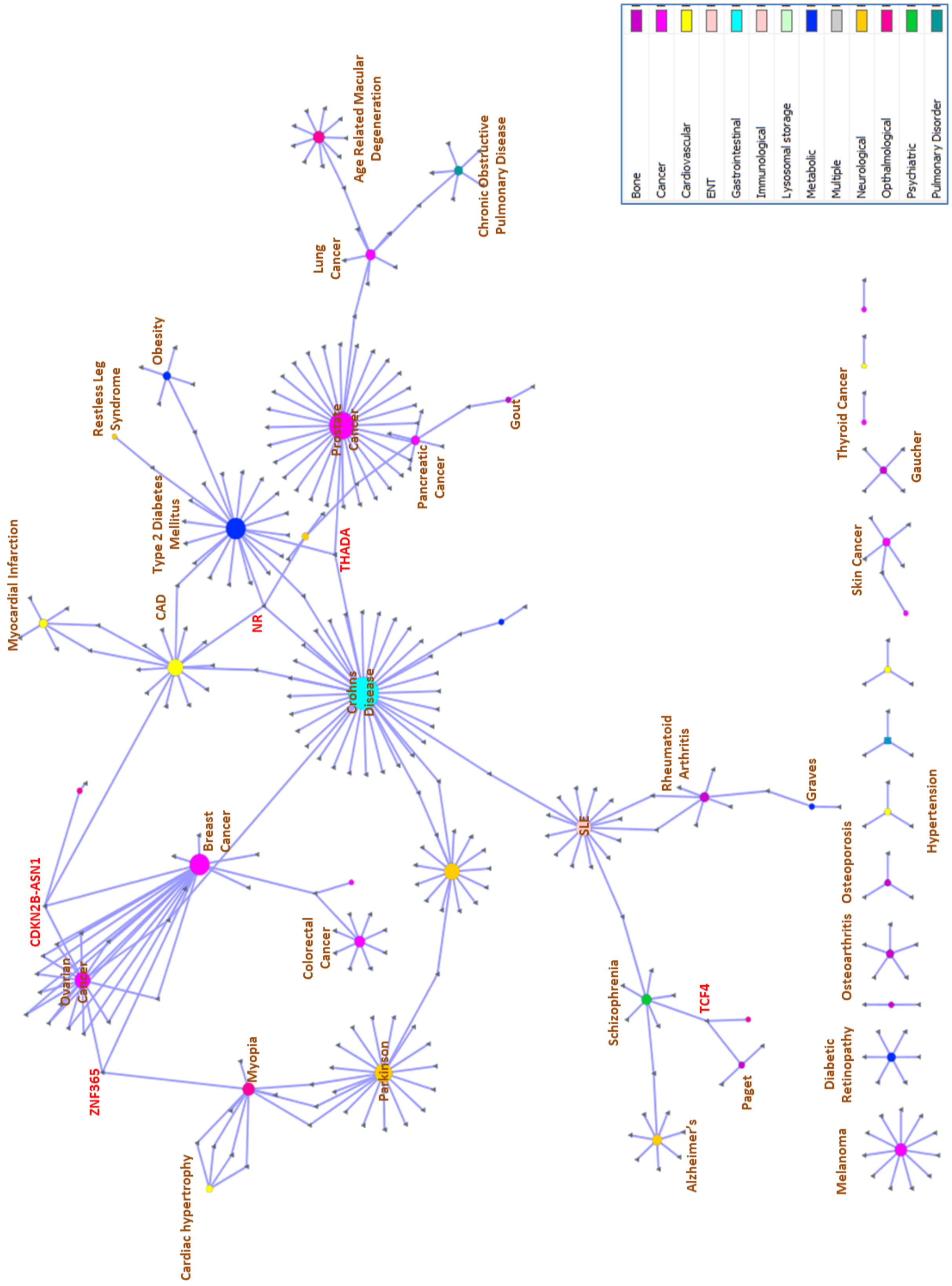


Figure 8. *Disease-Gene bipartite network.*

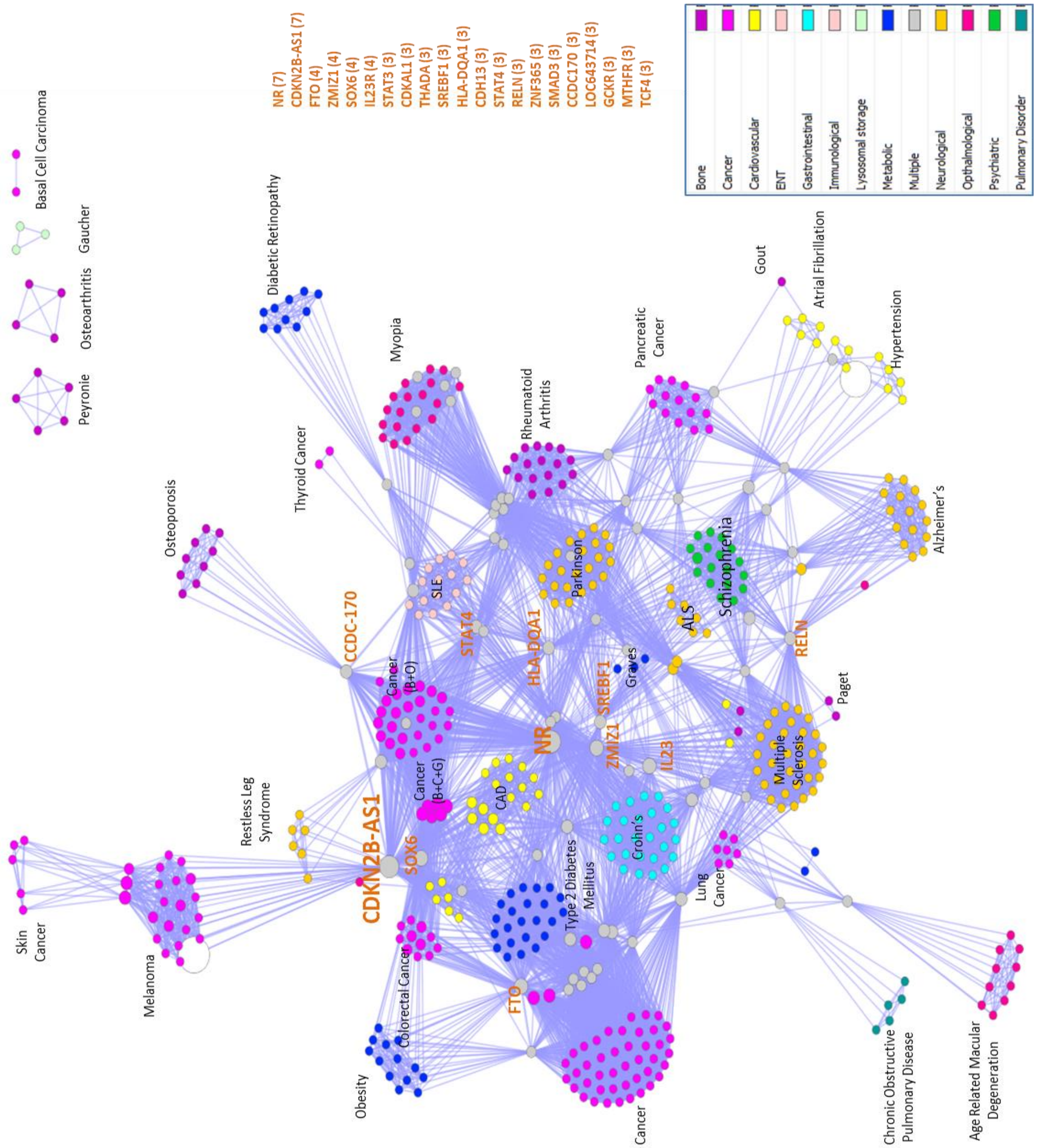
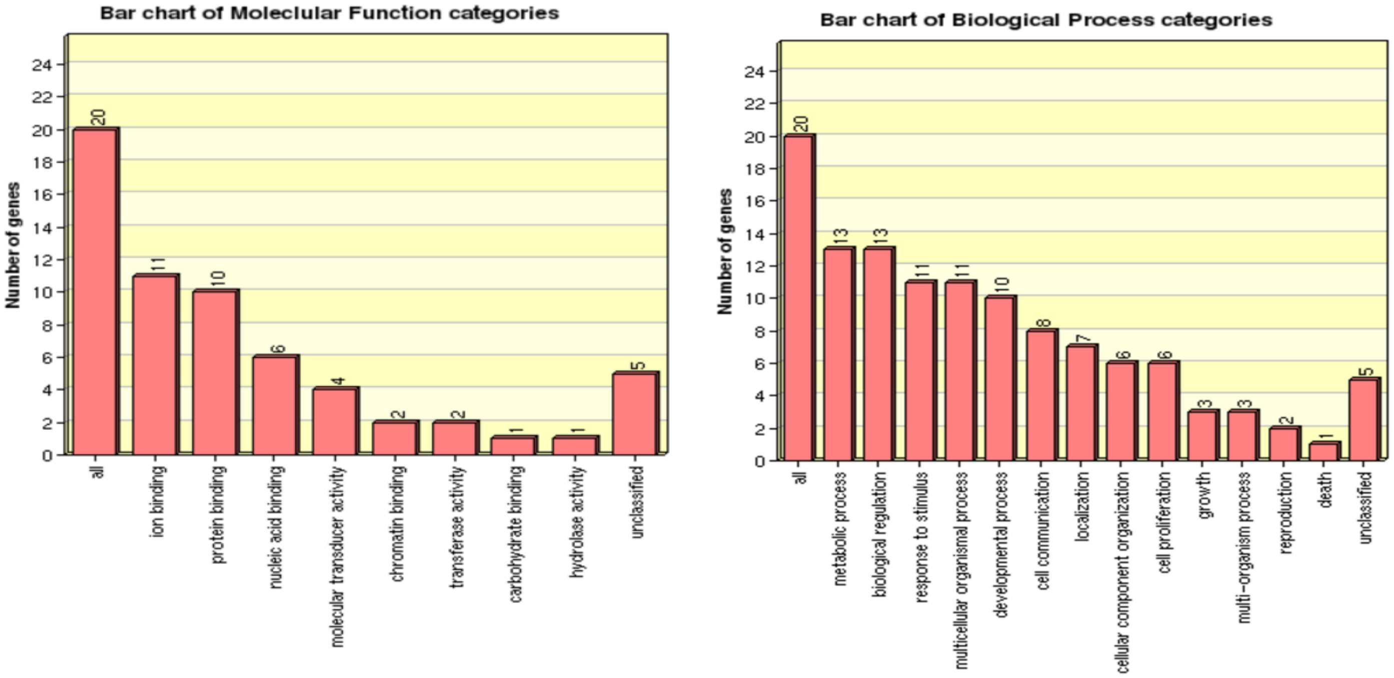


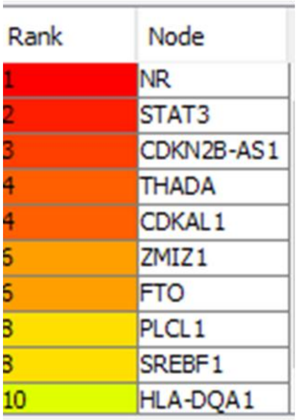
Figure 9(a). ARD Gene Projection Network (AGN). Gene network constructed by overlaps of involvement of the genes in same ARD. Circles in the figure represent genes whose sizes are proportional to their number of shared diseases. Grey lines represent shared ARDs common between gene nodes. Line thicknesses are proportional to number of common diseases.



(b)

(c)

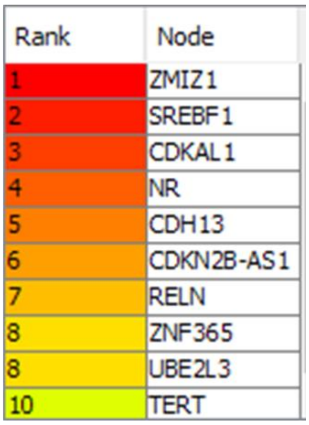
Degree



Rank	Node
1	NR
2	STAT3
3	CDKN2B-AS1
4	THADA
4	CDKAL1
6	ZMIZ1
6	FTO
8	PLCL1
8	SREBF1
10	HLA-DQA1

(d)

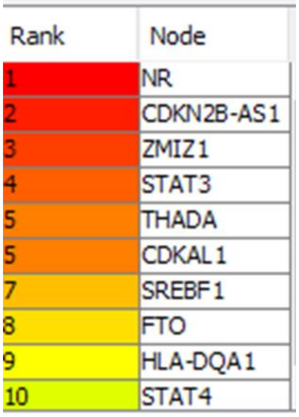
Bottleneck



Rank	Node
1	ZMIZ1
2	SREBF1
3	CDKAL1
4	NR
5	CDH13
6	CDKN2B-AS1
7	RELN
8	ZNF365
8	UBE2L3
10	TERT

(e)

Closeness



Rank	Node
1	NR
2	CDKN2B-AS1
3	ZMIZ1
4	STAT3
5	THADA
5	CDKAL1
7	SREBF1
8	FTO
9	HLA-DQA1
10	STAT4

(f)

Figure 9(b)GO Enrichment Analysis of the genes involved in more than 3 ARDs. (d-f) Hub gene analysis of AGN.

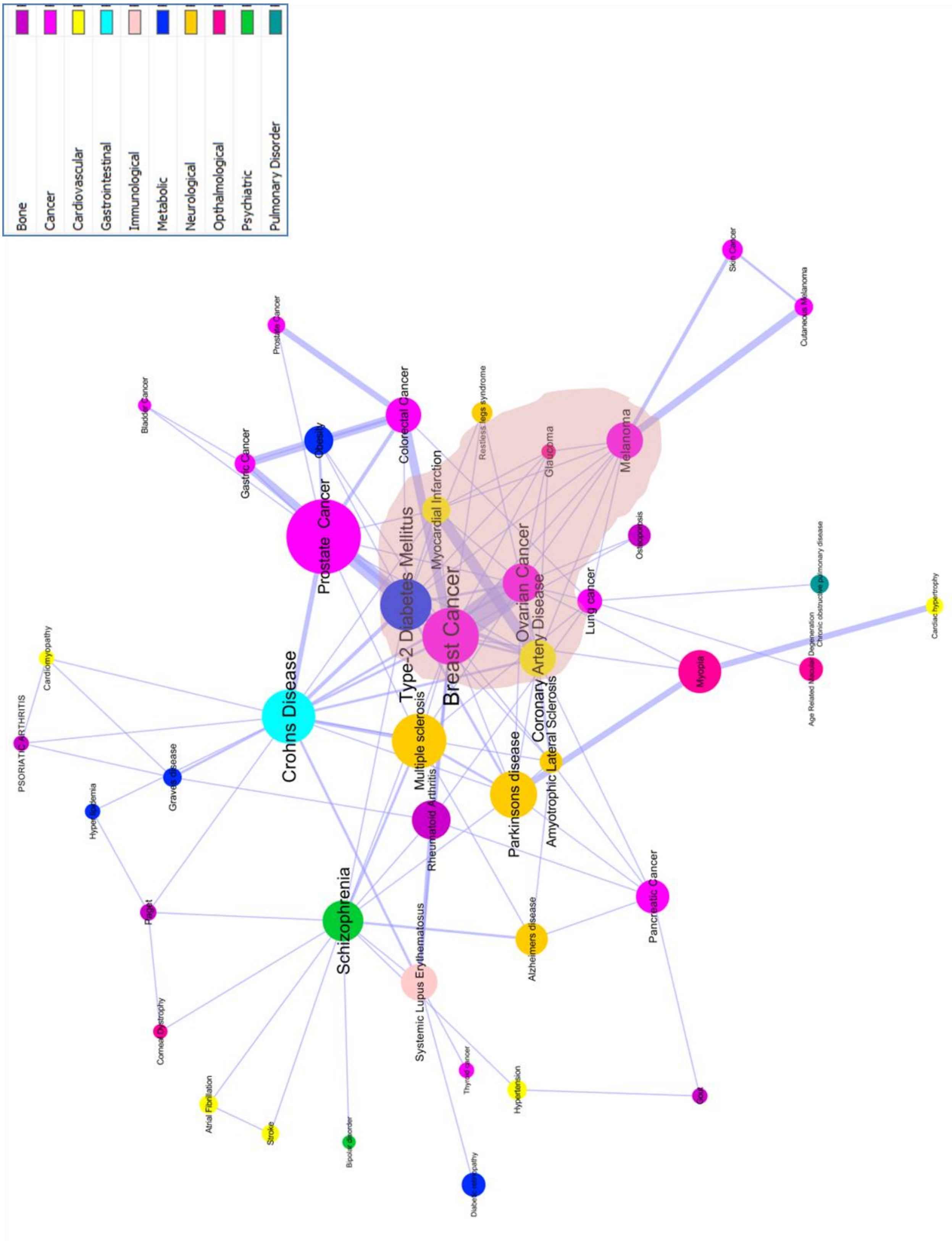


Figure 10(a). ARD Disease Projection Network (ADN). Disease network constructed by overlaps of the host genes of ARD-associated intragenic SNPs. Circles in the figure represent diseases whose sizes are proportional to their number of associated intragenic SNP host genes.

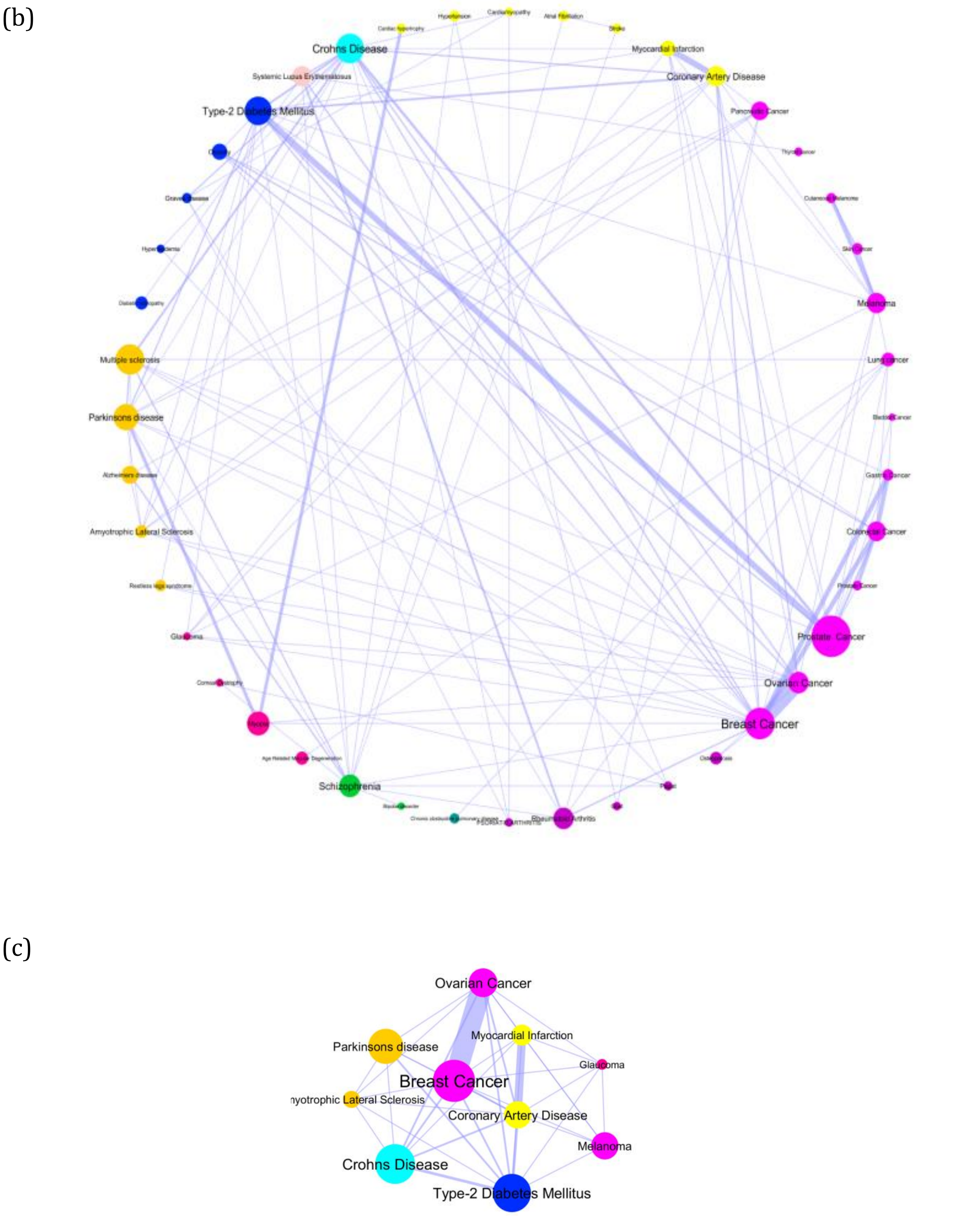


Figure 10(b) and 10(c). ADN (b) Circular layout of ADN. (c) Highly connected disease cluster in ADN (Highlighted with pink in (a)).

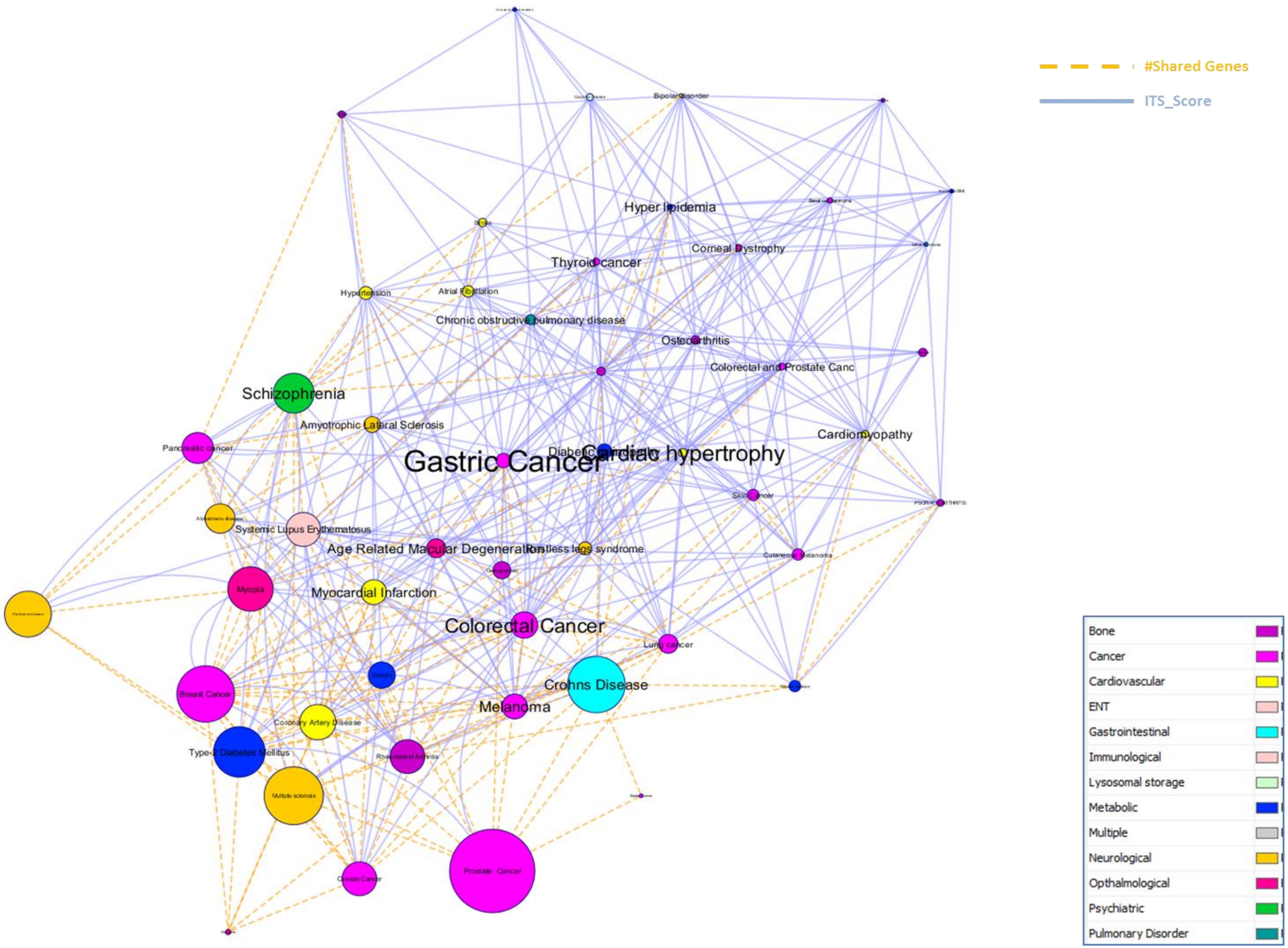

#Shared Genes
ITS_Score
Bipolar disorder
Hyper lipidemia
Corneal Dystrophy
Thyroid cancer
Hypertension
Atrial Fibrillation
Chronic obstructive pulmonary disease
Osteoarthritis
Colorectal and Prostate Canc
Schizophrenia
Amyotrophic Lateral Sclerosis
Pancreatic cancer
Gastric Cancer
Diabetic
Cardiac hypertrophy
Cardiomyopathy
Skin cancer
PSORIATIC ARTHRITIS
Systemic Lupus Erythematosus
Age Related Macular Degeneration
Restless legs syndrome
Cutaneous Melanoma
Myopia
Myocardial Infarction
Colorectal Cancer
Lung cancer
Breast Cancer
Crohns Disease
Melanoma
Coronary Artery Disease
Type-2 Diabetes Mellitus
Rheumatoid Arthritis
Multiple sclerosis
Ovarian Cancer
Prostate Cancer
Bone
Cancer
Cardiovascular
ENT
Gastrointestinal
Immunological
Lysosomal storage
Metabolic
Multiple
Neurological
Opthalmological
Psychiatric
Pulmonary Disorder

Figure 11(a). Disease similarity network was calculated using Genome Ontology biological process similarity of the host genes of ARD-associated intragenic SNPs with similarities ≥0.2 and an empirical p value <0.05. Therefore, this figure illustrates that information theoretic similarity method has found non-trivial relationships that would not have been found by conventional methods. Circles represent diseases or traits whose sizes are proportional to their number of associated intragenic SNP host genes. Blue lines represent biological process similarities that are ≥0.2 and have a p value <0.05. Pink lines represent shared SNP host genes between diseases if their DisSim is ≥0.2 (in other words overlapping connections between our information theoretic method and conventional gene overlapping method). Line thicknesses are proportional to DisSim similarity values or number of shared genes.

5.2 Comparison of ARD Networks constructed using shared gene and semantic similarity approaches

Gene networks were constructed using the conventional shared gene method using intragenic SNP diseases shared genes and also by gene semantic similarity method using GO ontology (Biological Processes). The gene network created using Shared genes had less number of edges as compared to the network created using Gene_Sim Score. The networks varied in the topological and clustering parameters. Both the networks and their calculated parameters are shown in the Figure 9, 10 and 11.

The projection network created using shared genes approach clearly showed distinct cliques being formed by diseases and hence clearly shows that the genes of a particular disease are associated closely. However, the interconnectivity between diseases belonging to different class was less evident which otherwise clearly became evident in the ARD disease network constructed using semantic similarity approach.

The comparative results also suggested that similarity network contains many potential connections among ARDs that have not yet been discovered by GWAS, and thus demonstrate that semantic similarity based method is able to capture non-trivial relationships that would not have been otherwise found by conventional methods.

5.3 ARD Protein-Protein Interaction Network

ARD-PPI Network constructed by mapping ARD genes on human PPI exhibited scale-free and small world property. That is, most nodes are not neighbours of one another but most nodes can be reached from every other node by a small number of hops/steps. ARD-PPI network formed the central part of the human PPI network with most of the ARD genes residing in the centre of the network.

5.4 Analysis of ARD-PPI Network

5.4.1 Functional enrichment analysis of ARD-PPI Network

Functional enrichment analysis of ARD-PPI network revealed that many genes were involved in the cell communication and signalling pathways.

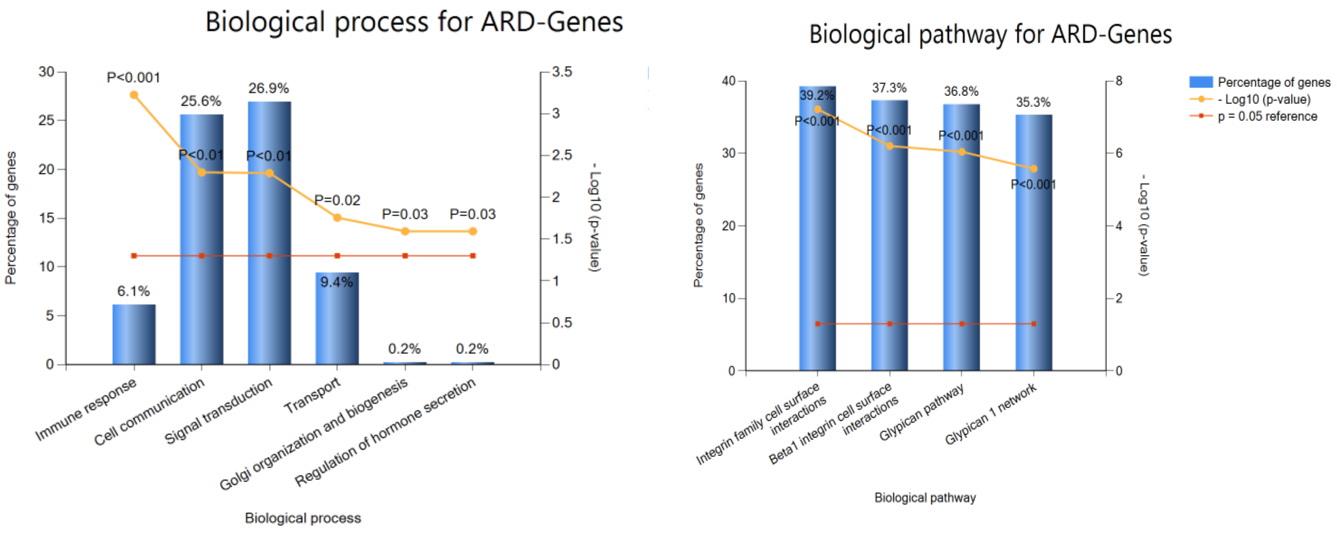
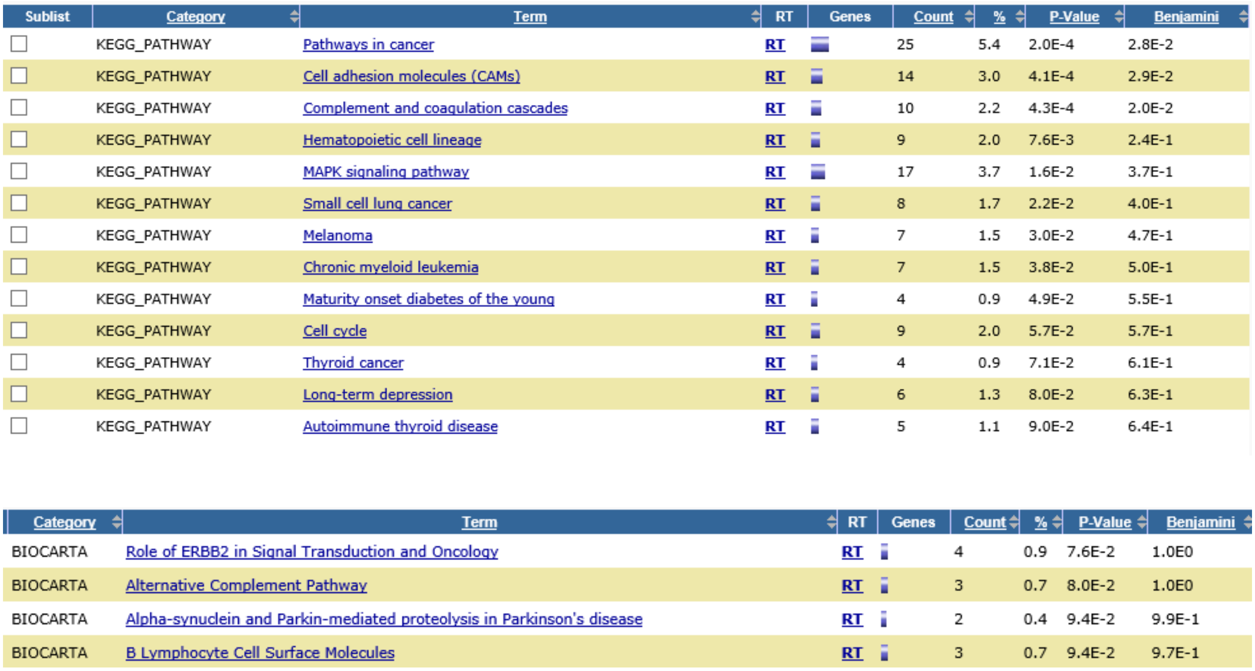


Figure 12. Functional Enrichment of ARD Genes

5.4.2 Disease Gene Network Analysis for Identification of "Hub" genes

Almost all of the protein protein interaction network consist of few dense nodes that exhibits high number of neighbour node directly connected. Finding out hub proteins and targeting them for the cure disorders is an effective method but one should keep the fact in mind that these hub proteins should not any essential protein, which is highly significant for survival and growth of an organism. Hub proteins are potential targets due to their location in the human PPI network. Betweenness is one of the most important topological properties of a network. It measures the number of shortest paths going through a certain node. Therefore, nodes with the highest betweenness control most of the information flow in the network, representing the critical points of the network. We thus call these nodes the ‘‘bottlenecks’’ of the network. Betweenness can thus be used as a measure to find the potential targets. Other parameters such as degree, closeness centrality, eccentricity etc are also important topological measures for finding the hub proteins. Therefore, to prioritize hubs in the ARD-PPI network, we captured essential proteins using HUBBA. HUBBA allows to find the hub proteins based on various topological parameters. Figure represents the hub proteins identified in the AGN network and Figure represents the hub in the ARD-PPI Network. As it can be seen that most hubs are common in both AGN and ARD-PPI network, but few interesting hubs are revealed from the ARD-PPI Network which otherwise were missing in the AGN due to lack of GWAS studies.

Hub Genes : HLA-DRA, ZNF315, STAT3, ZMIZ1, CCD170, SOX6, FTO,MTHFR, NF1A, BDNF



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	KEGG_PATHWAY	Pathways in cancer	RT	▬	25	5.4	2.0E-4	2.8E-2
□	KEGG_PATHWAY	Cell adhesion molecules (CAMs)	RT	▬	14	3.0	4.1E-4	2.9E-2
□	KEGG_PATHWAY	Complement and coagulation cascades	RT	▬	10	2.2	4.3E-4	2.0E-2
□	KEGG_PATHWAY	Hematopoietic cell lineage	RT	▬	9	2.0	7.6E-3	2.4E-1
□	KEGG_PATHWAY	MAPK signaling pathway	RT	▬	17	3.7	1.6E-2	3.7E-1
□	KEGG_PATHWAY	Small cell lung cancer	RT	▬	8	1.7	2.2E-2	4.0E-1
□	KEGG_PATHWAY	Melanoma	RT	▬	7	1.5	3.0E-2	4.7E-1
□	KEGG_PATHWAY	Chronic myeloid leukemia	RT	▬	7	1.5	3.8E-2	5.0E-1
□	KEGG_PATHWAY	Maturity onset diabetes of the young	RT	▬	4	0.9	4.9E-2	5.5E-1
□	KEGG_PATHWAY	Cell cycle	RT	▬	9	2.0	5.7E-2	5.7E-1
□	KEGG_PATHWAY	Thyroid cancer	RT	▬	4	0.9	7.1E-2	6.1E-1
□	KEGG_PATHWAY	Long-term depression	RT	▬	6	1.3	8.0E-2	6.3E-1
□	KEGG_PATHWAY	Autoimmune thyroid disease	RT	▬	5	1.1	9.0E-2	6.4E-1

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
BIOCARTA	Role of ERBB2 in Signal Transduction and Oncology	RT	▬	4	0.9	7.6E-2	1.0E0
BIOCARTA	Alternative Complement Pathway	RT	▬	3	0.7	8.0E-2	1.0E0
BIOCARTA	Alpha-synuclein and Parkin-mediated proteolysis in Parkinson's disease	RT	▬	2	0.4	9.4E-2	9.9E-1
BIOCARTA	B Lymphocyte Cell Surface Molecules	RT	▬	3	0.7	9.4E-2	9.7E-1

Figure 13 Hub Genes Functional Enrichment Analysis. Showing the involvement in pathways responsible for diseases.

5.4.3 Identification of modules and new markers for ARDs

The NetBox generated 52 modules. The top modules generated by the tool are listed in Table 3. The table also shows that in some modules there were linker genes which were tightly associated with other genes in the ARD network, thus establishing the possible involvement of these genes into the ARD. Thus, these genes can be considered as putative ARD associated genes (Table 4)

Table 4 *Modules Generated using NetBox.*

Module ID	Number of Genes	Genes
0	3	TRIM2 MYO5A MLPH
1	9	FBN2 TMPRSS6 HTRA1 MATN2* FBN1 NID1 CDYL COL1A1 COL4A1
2	7	MAPT STUB1* APOE SNCA UBE2L3 MPP1 PARK2
3	48	PRKCD* IL2RA JAK1* NOS2* IL18RAP RELA* STAT4 TERT INSR IL18R1* JAK2* STAT3 FYN* IL13* PPARG PLAU RELN GADD45G* STAT5A* NFKB1* TYK2 PIK3CA* GADD45B* IL23R IL18 IL12RB2* IL6R ITGA3* IL12RB1* IL23A* RIPK2* PIK3R1* ZBTB17 IL7R HLX* CTLA4 IL12B* IFNG* HLA-DRA MYC FOS* IL1RAP IL12A* TSHR RAF1 SMAD3 ESR1 IL2*
4	10	CFI* CR2 C4B* CFB C4A* CFP* CR1 C2 C3 CFH
5	4	TNIP1 EIF3E RPS11 RPL5
6	1	SKAP1
7	2	FCGR2A BLK
8	5	TCF4 GRM5 CALM3* GRM7 CALM2*
9	4	CASP8 TNFRSF1A TRAF1 TNFAIP3
10	1	RASGRP3
11	13	SELP FGG* ITGAM CD226 ITGAX FCER2* LPA PVRL2 IRF4 FGB* ITGB2* F13A1 KNG1

* Linker gene was not present in the original input list, but is significantly connected to members of the input list.

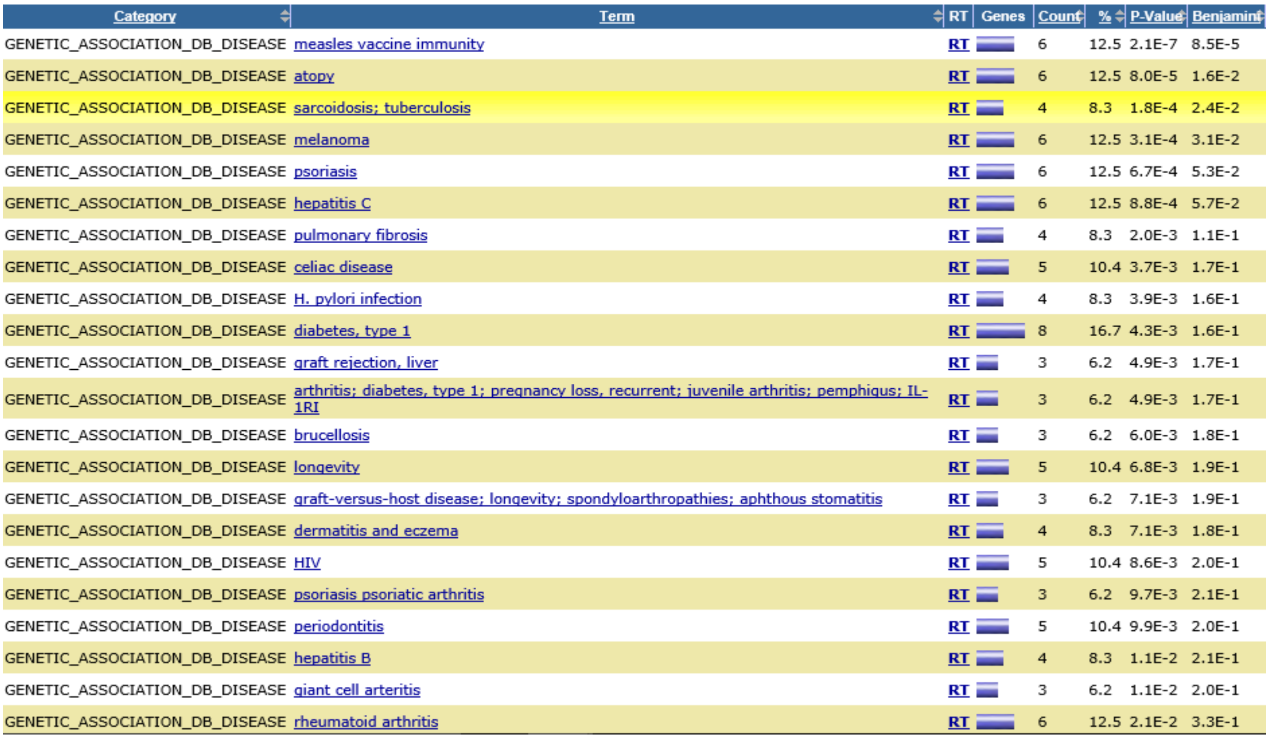
Table 5 *Linker Gene Details: Based on Global Network with: 9264 genes and 68111 edges*

Gene Symbol	Local Degree	Global Degree	Unadjusted P-Value	FDR Adjusted P-Value
PIK3R1	22	202	4.32E-07	0.0004
JAK2	17	174	3.40E-05	0.014
NFKB1	14	126	4.14E-05	0.014
IL18R1	6	24	8.27E-05	0.0193
IFNG	10	75	0.0001	0.0193
RIPK2	7	36	0.0001	0.0193
FOS	13	128	0.0002	0.0222
HLX	4	10	0.0002	0.0222
IL2	12	112	0.0002	0.0222
ERBB2IP	5	19	0.0003	0.0232
ITGB4	7	41	0.0003	0.0232
PRKCD	12	117	0.0003	0.0232
YWHAH	8	55	0.0003	0.0232
MATN2	4	12	0.0004	0.029
PLAUR	6	32	0.0004	0.029
PIK3CA	12	125	0.0005	0.0295
GADD45B	5	22	0.0005	0.0295
NOS2	4	13	0.0006	0.0295
IL23A	6	34	0.0006	0.0295

NFYC	3	6	0.0006	0.0295
RAPGEF6	3	6	0.0006	0.0295
SMARCD3	4	14	0.0008	0.0312
FGG	4	14	0.0008	0.0312
FGB	4	14	0.0008	0.0312
BRAF	5	24	0.0008	0.0312
GADD45G	6	36	0.0008	0.0312
JAK1	11	116	0.0009	0.0312
CALM3	5	25	0.001	0.0312
LAMC1	4	15	0.001	0.0312
C4B	3	7	0.001	0.0312
C4A	3	7	0.001	0.0312
CFP	2	2	0.001	0.0312
CFI	2	2	0.001	0.0312
RELA	13	155	0.001	0.0312
STAT5A	8	67	0.0011	0.0313
ITGB2	8	71	0.0015	0.0429
TNFRSF18	3	8	0.0016	0.0429
CALM2	5	28	0.0016	0.0429
IL12RB2	7	56	0.0017	0.0433
ITGA3	5	29	0.0019	0.0469

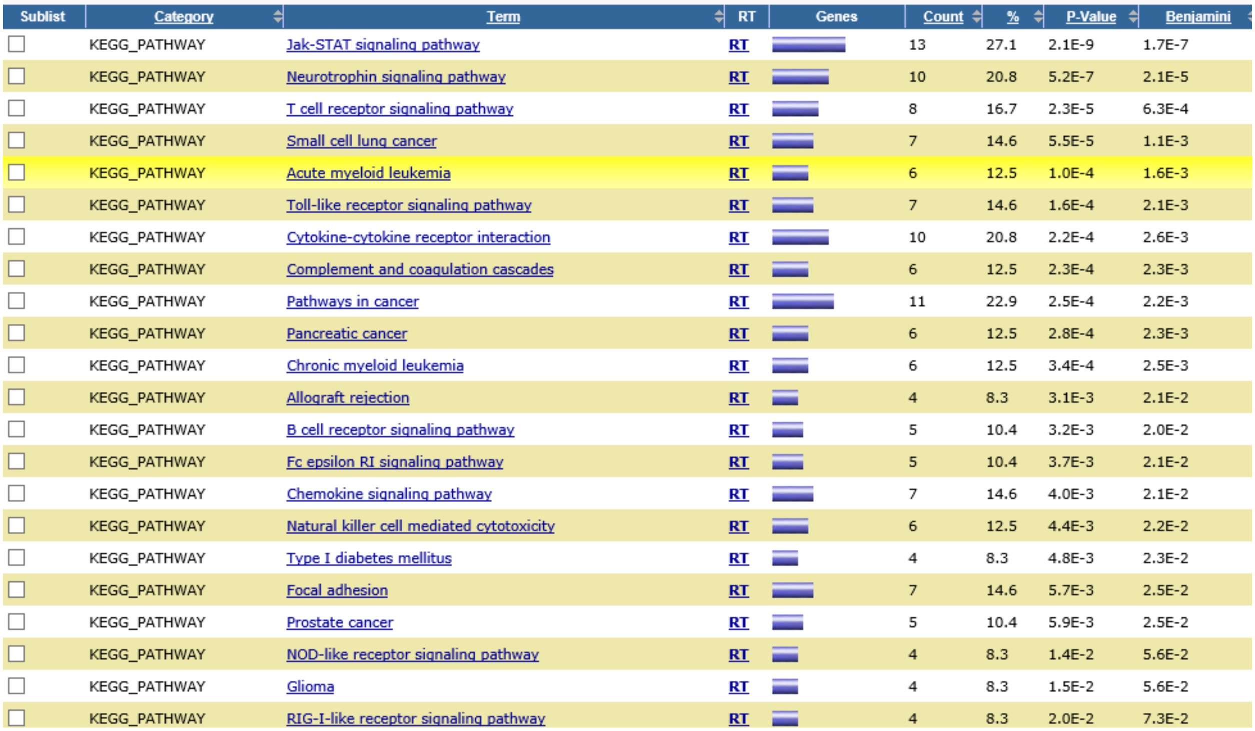
FYN	14	187	0.0019	0.0469
IL12B	8	74	0.002	0.0469
IL12RB1	8	74	0.002	0.0469
GTF2I	4	18	0.0021	0.0481
STUB1	5	30	0.0022	0.0481
IL12A	7	59	0.0022	0.0481
NTF3	3	9	0.0023	0.0481
IL13	3	9	0.0023	0.0481
FCER2	3	9	0.0023	0.0481

To confirm the association of these genes with diseases or pathways involved in the complex diseases, the functional enrichment of the genes listed in Table was done using DAVID Bioinformatics Resources 6.7. The result of DAVID clearly indicated the involvement of these genes in ARDs (Figure 14) as well as pathways associated with aging and age-related diseases (Figure15).



Category	Term	RT	Genes	Count	%	P-Value	Benjamini
GENETIC_ASSOCIATION_DB_DISEASE	measles vaccine immunity	RT	▬	6	12.5	2.1E-7	8.5E-5
GENETIC_ASSOCIATION_DB_DISEASE	atopy	RT	▬	6	12.5	8.0E-5	1.6E-2
GENETIC_ASSOCIATION_DB_DISEASE	sarcoidosis; tuberculosis	RT	▬	4	8.3	1.8E-4	2.4E-2
GENETIC_ASSOCIATION_DB_DISEASE	melanoma	RT	▬	6	12.5	3.1E-4	3.1E-2
GENETIC_ASSOCIATION_DB_DISEASE	psoriasis	RT	▬	6	12.5	6.7E-4	5.3E-2
GENETIC_ASSOCIATION_DB_DISEASE	hepatitis C	RT	▬	6	12.5	8.8E-4	5.7E-2
GENETIC_ASSOCIATION_DB_DISEASE	pulmonary fibrosis	RT	▬	4	8.3	2.0E-3	1.1E-1
GENETIC_ASSOCIATION_DB_DISEASE	celiac disease	RT	▬	5	10.4	3.7E-3	1.7E-1
GENETIC_ASSOCIATION_DB_DISEASE	H. pylori infection	RT	▬	4	8.3	3.9E-3	1.6E-1
GENETIC_ASSOCIATION_DB_DISEASE	diabetes, type 1	RT	▬	8	16.7	4.3E-3	1.6E-1
GENETIC_ASSOCIATION_DB_DISEASE	graft rejection, liver	RT	▬	3	6.2	4.9E-3	1.7E-1
GENETIC_ASSOCIATION_DB_DISEASE	arthritis; diabetes, type 1; pregnancy loss, recurrent; juvenile arthritis; pemphigus; IL-1RI	RT	▬	3	6.2	4.9E-3	1.7E-1
GENETIC_ASSOCIATION_DB_DISEASE	brucellosis	RT	▬	3	6.2	6.0E-3	1.8E-1
GENETIC_ASSOCIATION_DB_DISEASE	longevity	RT	▬	5	10.4	6.8E-3	1.9E-1
GENETIC_ASSOCIATION_DB_DISEASE	graft-versus-host disease; longevity; spondyloarthropathies; aphthous stomatitis	RT	▬	3	6.2	7.1E-3	1.9E-1
GENETIC_ASSOCIATION_DB_DISEASE	dermatitis and eczema	RT	▬	4	8.3	7.1E-3	1.8E-1
GENETIC_ASSOCIATION_DB_DISEASE	HIV	RT	▬	5	10.4	8.6E-3	2.0E-1
GENETIC_ASSOCIATION_DB_DISEASE	psoriasis psoriatic arthritis	RT	▬	3	6.2	9.7E-3	2.1E-1
GENETIC_ASSOCIATION_DB_DISEASE	periodontitis	RT	▬	5	10.4	9.9E-3	2.0E-1
GENETIC_ASSOCIATION_DB_DISEASE	hepatitis B	RT	▬	4	8.3	1.1E-2	2.1E-1
GENETIC_ASSOCIATION_DB_DISEASE	giant cell arteritis	RT	▬	3	6.2	1.1E-2	2.0E-1
GENETIC_ASSOCIATION_DB_DISEASE	rheumatoid arthritis	RT	▬	6	12.5	2.1E-2	3.3E-1

Figure 14 DAVID Results (Disease Enrichment for the linker genes)



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	KEGG_PATHWAY	Jak-STAT signaling pathway	RT	▬	13	27.1	2.1E-9	1.7E-7
□	KEGG_PATHWAY	Neurotrophin signaling pathway	RT	▬	10	20.8	5.2E-7	2.1E-5
□	KEGG_PATHWAY	T cell receptor signaling pathway	RT	▬	8	16.7	2.3E-5	6.3E-4
□	KEGG_PATHWAY	Small cell lung cancer	RT	▬	7	14.6	5.5E-5	1.1E-3
□	KEGG_PATHWAY	Acute myeloid leukemia	RT	▬	6	12.5	1.0E-4	1.6E-3
□	KEGG_PATHWAY	Toll-like receptor signaling pathway	RT	▬	7	14.6	1.6E-4	2.1E-3
□	KEGG_PATHWAY	Cytokine-cytokine receptor interaction	RT	▬	10	20.8	2.2E-4	2.6E-3
□	KEGG_PATHWAY	Complement and coagulation cascades	RT	▬	6	12.5	2.3E-4	2.3E-3
□	KEGG_PATHWAY	Pathways in cancer	RT	▬	11	22.9	2.5E-4	2.2E-3
□	KEGG_PATHWAY	Pancreatic cancer	RT	▬	6	12.5	2.8E-4	2.3E-3
□	KEGG_PATHWAY	Chronic myeloid leukemia	RT	▬	6	12.5	3.4E-4	2.5E-3
□	KEGG_PATHWAY	Allograft rejection	RT	▬	4	8.3	3.1E-3	2.1E-2
□	KEGG_PATHWAY	B cell receptor signaling pathway	RT	▬	5	10.4	3.2E-3	2.0E-2
□	KEGG_PATHWAY	Fc epsilon RI signaling pathway	RT	▬	5	10.4	3.7E-3	2.1E-2
□	KEGG_PATHWAY	Chemokine signaling pathway	RT	▬	7	14.6	4.0E-3	2.1E-2
□	KEGG_PATHWAY	Natural killer cell mediated cytotoxicity	RT	▬	6	12.5	4.4E-3	2.2E-2
□	KEGG_PATHWAY	Type I diabetes mellitus	RT	▬	4	8.3	4.8E-3	2.3E-2
□	KEGG_PATHWAY	Focal adhesion	RT	▬	7	14.6	5.7E-3	2.5E-2
□	KEGG_PATHWAY	Prostate cancer	RT	▬	5	10.4	5.9E-3	2.5E-2
□	KEGG_PATHWAY	NOD-like receptor signaling pathway	RT	▬	4	8.3	1.4E-2	5.6E-2
□	KEGG_PATHWAY	Glioma	RT	▬	4	8.3	1.5E-2	5.6E-2
□	KEGG_PATHWAY	RIG-I-like receptor signaling pathway	RT	▬	4	8.3	2.0E-2	7.3E-2

Figure 15 DAVID Results (Pathway Enrichment for the linker genes)

5.5 ARDInteract

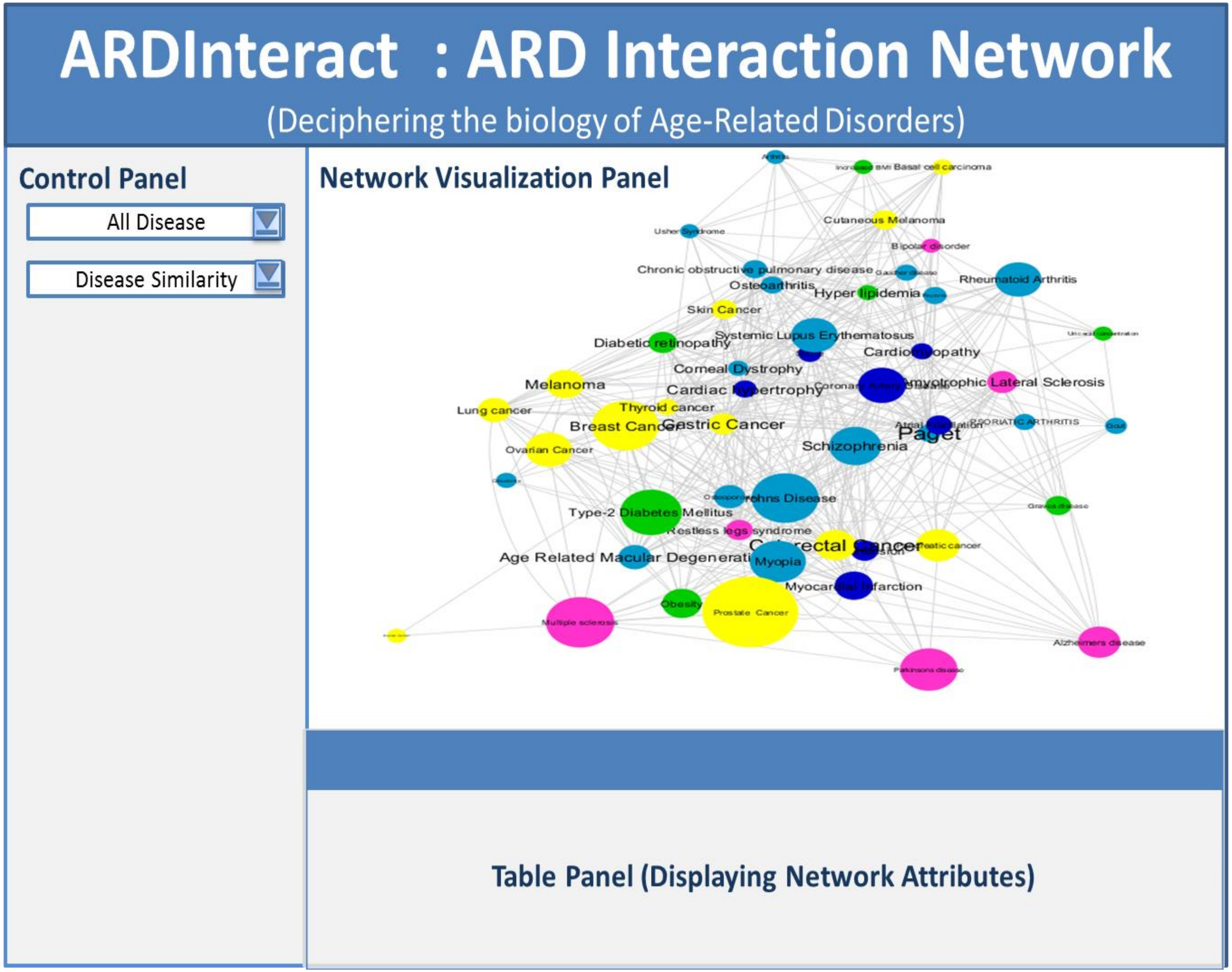
The layout of the user-interface of ARDInteract is shown in the figure below:

ARDInteract have 3 panels as shown in the Figure.

Control Panel: The control panel allow the user to select the diseases he/she wants to consider for network construction. And then user can select the type of network it wants to display – SNP interaction, gene interaction, disease similarity etc. All these networks are displayed on the basis of the data stored in the MySQL database connected to the front end.

Visualization Panel: The visualization panel display the network depending on the choice of the user. Display panel have features like zoom, drag etc and also when user select on node or edge a new window pops-up giving more description about the node or the edge. Also the new window have link to other databases to provide more comprehensive information about the node to the user.

Table Panel: The table panel display the information about the nodes and the edges in the tabular format. It displays the attributes of the nodes and the edges as computed during the network construction earlier.



ARDInteract : ARD Interaction Network

(Deciphering the biology of Age-Related Disorders)

Control Panel

All Disease

Disease Similarity

Network Visualization Panel

Table Panel (Displaying Network Attributes)

Figure 16: ARDInteract User Interface

PART VI: CONCLUSION

6. CONCLUSION

Aging is an inevitable process and is one of the major risk factors for Age-related diseases(ARDs). The complete biology of aging and ARDs is still known and researches are being carried out to understand the genetics and environmental factors that affect the process leading to complex diseases. Many of the ARDs have been known to share common genetic factors, identification of these shared genetic components and the underlying biological mechanism is of utmost importance in order to slow down the process of aging or rather for the successful aging. With the lots of data being generated from GWAS, there has been a need for statistical tools and methods that can combine the genetic variation and trait association data from different sources and determine the underlying common pathways. In order to facilitate this with respect to ARDs, it is proposed that the present work can provide deep insight into SNPs associated with ARDs and the underlying biological pathways.

ARDInteract available as an online platform l allow to visualize and analyse the ARD networks at different '-omics' levels in a user-friendly environment. ARDInteract also allow the construction of multi-partite network for analysis and visualization to uncover genetic similarities among various ARDs and also direct the users to connect to other relevant databases/websites to retrieve more meaningful information.

In summary, by exploring the genes and genetic variations associated with ARDs in context of network and pathways, we can unravel the unknown molecular mechanisms associated with ARDs and can identify the novel genes and variants which are still unknown. The study will provide links between various age-related disorders, which can provide valuable perspective to physicians, counsellors and biomedical researchers. Moreover, the integration of this study with gene/miRNA expression profiles could further highlight the key players linking various ARDs.

Our multidimensional approach using network and pathway based analysis can be applicable to other diseases as well. However, there are certain limitations of the present study. In the present approach, we are only including the intragenic SNPs whereas it is known that even intergenic SNPs also exert their roles on the biological pathways and hence the progression of the complex diseases. But the methods to take into consideration those intergenic SNPs are very few. Also, by exploiting the power of Bayesian networks, more statistically powerful enrichment of network can be done which is missing from the proposed plan but may be done in the future.

REFERENCES

- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25, 25–29.
- Baranzini, S. E., Galwey, N. W., Wang, J., Khankhanian, P., Lindberg, R., Pelletier, D., Barnes, M. R. (2009). Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Human Molecular Genetics, 18(11), 2078–2090
- Burcu Bakir-Gungor, Osman Ugur Sezerman, and Joaquín Dopazo. (2011) A New Methodology to Associate SNPs with Human Diseases According to Their Pathway Related Context . PLoS ONE, 6(10)
- Cheng L, Li J, Ju P, Peng J, Wang Y (2014) SemFunSim: A New Method for Measuring Disease Similarity by Integrating Semantic and Gene Functional Association. PLoS ONE 9(6)
- Guo X, Liu R, Shriver CD, Hu H, Liebman MN (2006) Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics, 22, 967–973.
- H. Schwender, K. Ickstadt,(2008). Identification of SNP interactions using logic regression. Biostatistics, 9, 187-198.
- Horacio Caniza, Alfonso E. Romero, Samuel Heron, Haixuan Yang, Alessandra Devoto, Marco Frasca, Giorgio Valentini and Alberto Paccanaro, (2014). GOssTo: a user-friendly stand-alone and web tool for calculating semantic similarities on the Gene Ontology., Bioinformatics.
- Jeck, W. R., Siebold, A. P., & Sharpless, N. E. (2012). Review: A Meta-Analysis of GWAS Studies and Age-Associated Diseases. Aging Cell, 11(5), 727–731.
- Kenyon CJ. (2010). The genetics of aging. Nature.464(7288), 504-12
- Kogelman et. al. (2014). Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA Sequencing in a porcine model, BMC Medical Genomics, 4, 7:57
- Kogelman, LJA and Kadarmideen, HN. (2014). Weighted Interaction SNP Hub (WISH) network method for building genetic networks for complex diseases using whole genome genotype data. BMC Systems Biology 8
- Lesnick T.G., Papapetropoulos S., Mash D.C., Ffrench-Mullen J., Shehadeh L., de Andrade M., Henley J.R., Rocca W.A., Ahlskog J.E., Maraganore D.M. (2007) A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. PLoS Genet.3:e98.
- Li J, Gong B, Chen X, Liu T, Wu C, et al. (2011) DOSim: an R package for similarity between diseases based on Disease Ontology. BMC Bioinformatics, 12

- Melzer D, Hurst AJ, Frayling T.(2007) Genetic variation and human aging: progress and prospects. J Gerontol A Biol Sci Med Sci.62(3), 301–307
- Peng , J., Li, H., Jiang, Q., Wang, Y., & Chen, J. (2014). An integrative approach for measuring semantic similarities using gene ontology. BMC Systems Biology, 8(Suppl 5), S8.
- Pesquita C, Faria D, Falcao AO, Lord P, Couto FM (2009) Semantic similarity in biomedical ontologies. PLoS Comput Biol 5: e1000443. doi: 10.1371/journal.pcbi.1000443
- Schlicker A., Albrecht M. (2010) FunSimMat update: new features for exploring functional similarity. Nucleic Acids Research, 38
- Sebastiani et al. (2009) RNA editing genes associated with extreme old age in humans and with lifespan in C. elegans. PLoS ONE, 4(12):e8210
- Shi, H., Belbin, O., Medway, C., Brown, K., Kalsheker, N., Carrasquillo, Morgan, K. (2012). Genetic variants influencing human aging from late-onset Alzheimer's disease (LOAD) genome-wide association studies (GWAS).Neurobiology of Aging, 33(8), 1849.e5–1849.18.
- Wan et. al. (2010). Predictive rule inference for epistatic interaction detection in genome-wide association studies, Bioinformatics, 26 (1), 30-37.
- Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF (2007) A new method to measure the semantic similarity of GO terms. Bioinformatics, 23, 1274–1281
- X. Wan, C. Yang, Q. Yang, H. Xue, N. L. S. Tang, W. C. Yu. (2010). Predictive rule inference for epistatic interaction detection in genomewide association studies. Bioinformatics, 26, 30-37.
- Y. Liu, M. K. Ng. (2010). Shrunken Methodology to Genome-wide SNPs Selection and Construction of SNPs Networks. BMC Systems Biology.

APPENDIX

A1. List of Age-Related Disease and the class included in the work

ID	Disease	Class
T1	Age Related Macular Degeneration	Opthalmological
T2	Alzheimers disease	Neurological
T3	Amyotrophic Lateral Sclerosis	Neurological
T4	Arthritis	Bone
T5	Atrial Fibrillation	Cardiovascular
T6	Basal cell carcinoma	Cancer
T7	Bipolar disorder	Neurological
T8	Bladder Cancer	Cancer
T9	Breast Cancer	Cancer
T10	Cardiac hypertrophy	Cardiovascular
T11	Cardiomyopathy	Cardiovascular
T12	Chronic obstructive pulmonary disease	Pulmonary Disorder
T13	Colon Cancer	Cancer
T14	Colorectal Cancer	Cancer
T15	Corneal Dystrophy	Opthalmological
T16	Coronary Artery Disease	Cardiovascular
T17	Crohns Disease	Gastrointestinal
T18	Cutaneous Melanoma	Cancer
T19	Diabetic retinopathy	Metabolic
T20	Duodenal ulcer	Gastrointestinal
T21	Gastric Cancer	Cancer
T22	Gaucher disease	Lysosomal storage
T23	Glaucoma	Opthalmological
T24	Gout	Bone
T25	Graves disease	Metabolic
T26	Hyper lipidemia	Metabolic
T27	Hypertension	Cardiovascular
T28	Increased BMI	Metabolic
T29	Lung cancer	Cancer
T30	Melanoma	Cancer

T31	Multiple sclerosis	Neurological
T32	Myocardial Infarction	Cardiovascular
T33	Myopia	Opthalmological
T34	Obesity	Metabolic
T35	Osteoarthritis	Bone
T36	Osteoporosis	Bone
T37	Ovarian Cancer	Cancer
T38	Paget	Bone
T39	Pancreatic cancer	Cancer
T40	Parkinsons disease	Neurological
T41	Peyronie	Bone
T42	Prostate Cancer	Cancer
T43	PSORIATIC ARTHRITIS	Bone
T44	Restless legs syndrome	Neurological
T45	Rheumatoid Arthritis	Bone
T46	Schizophrenia	Psychiatric
T47	Skin Cancer	Cancer
T48	Stroke	Cardiovascular
T49	Systemic Lupus Erythematosus	Immunological
T50	Thyroid cancer	Cancer
T51	Type-2 Diabetes Mellitus	Metabolic
T52	Uric acid concentration	Metabolic
T53	Usher Syndrome	ENT

A2. Detailed information about 53 disease-disease similarity connections selected with ITS_Score > 0.2

Disease1	Disease2	No. of genes in Disease1	No. of genes in Disease2	No. of overlapping genes	ITS_Score_BP
T14	T21	16	8	7	0.583333
T9	T37	36	21	19	0.521719
T26	T38	2	4	1	0.519
T7	T52	1	1	0	0.51
T15	T38	2	4	1	0.496333
T28	T53	1	1	0	0.492
T35	T38	4	4	0	0.4845
T18	T30	6	15	6	0.47619
T18	T47	6	6	2	0.468833
T16	T32	22	15	9	0.439568
T12	T19	5	8	0	0.436308
T10	T15	4	2	0	0.433333
T26	T50	2	3	0	0.4216
T1	T29	11	11	1	0.418364
T13	T38	3	4	0	0.415143
T13	T50	3	3	0	0.407333
T10	T12	4	5	0	0.404889
T19	T38	8	4	0	0.3975
T4	T15	1	2	0	0.392
T12	T35	5	4	0	0.391333
T12	T38	5	4	0	0.386444
T7	T28	1	1	0	0.384
T14	T34	16	16	2	0.379562
T11	T50	3	3	0	0.375333
T6	T10	2	4	0	0.374
T13	T26	3	2	0	0.3676
T26	T52	2	1	0	0.367333
T7	T26	1	2	0	0.366667
T10	T38	4	4	0	0.3665
T21	T38	8	4	0	0.3635
T13	T19	3	8	0	0.360909
T24	T52	3	1	0	0.36
T4	T28	1	1	0	0.355
T10	T13	4	3	0	0.352857
T35	T48	4	4	0	0.34975
T38	T50	4	3	0	0.344857
T24	T27	3	7	1	0.3436
T6	T38	2	4	0	0.343333
T11	T28	3	1	0	0.3425
T28	T43	1	3	0	0.3425
T6	T35	2	4	0	0.339333
T30	T47	15	6	3	0.338857
T22	T35	3	4	0	0.338286

T9	T14	36	16	7	0.336077
T38	T47	4	6	0	0.3354
T38	T44	4	7	0	0.334545
T44	T47	7	6	0	0.334
T11	T43	3	3	1	0.333333
T4	T53	1	1	0	0.331
T26	T35	2	4	0	0.330333
T11	T15	3	2	0	0.322
T33	T34	28	16	0	0.321273
T10	T41	4	4	0	0.321
T7	T50	1	3	0	0.3205
T40	T46	29	25	1	0.318943
T27	T48	7	4	0	0.316182
T10	T33	4	28	4	0.312562
T3	T47	9	6	0	0.308933
T35	T44	4	7	0	0.308909
T3	T50	9	3	0	0.305667
T7	T24	1	3	0	0.3055
T7	T53	1	1	0	0.305
T14	T30	16	15	0	0.304452
T6	T28	2	1	0	0.304
T11	T35	3	4	0	0.304
T5	T12	6	5	0	0.302909
T6	T26	2	2	0	0.3025
T5	T38	6	4	0	0.3022
T11	T13	3	3	0	0.300667
T22	T26	3	2	0	0.3
T1	T2	11	18	0	0.29531
T1	T38	11	4	0	0.293067
T14	T49	16	21	0	0.292108
T45	T49	21	21	3	0.292048
T1	T14	11	16	0	0.291407
T4	T26	1	2	0	0.290667
T28	T50	1	3	0	0.2905
T1	T36	11	10	0	0.289619
T3	T14	9	16	0	0.28896
T43	T53	3	1	0	0.2885
T21	T47	8	6	0	0.288143
T21	T29	8	11	0	0.287368
T27	T38	7	4	0	0.285818
T12	T44	5	7	0	0.285333
T15	T43	2	3	0	0.2852
T11	T41	3	4	0	0.284857
T1	T49	11	21	0	0.282438
T14	T38	16	4	0	0.2823
T11	T44	3	7	0	0.281
T11	T47	3	6	0	0.279333
T15	T26	2	2	0	0.2785
T13	T35	3	4	0	0.278

T26	T28	2	1	0	0.278
T13	T41	3	4	0	0.277429
T1	T47	11	6	0	0.277059
T18	T35	6	4	0	0.277
T50	T52	3	1	0	0.277
T1	T46	11	25	0	0.276444
T11	T19	3	8	0	0.276
T15	T22	2	3	0	0.276
T1	T21	11	8	0	0.274842
T24	T26	3	2	0	0.2748
T25	T38	6	4	0	0.2748
T15	T35	2	4	0	0.274667
T15	T48	2	4	0	0.274333
T3	T12	9	5	0	0.273143
T9	T21	36	8	7	0.272727
T15	T28	2	1	0	0.272667
T32	T38	15	4	0	0.272526
T5	T21	6	8	0	0.271429
T27	T35	7	4	0	0.270727
T11	T53	3	1	0	0.2695
T11	T38	3	4	0	0.268571
T6	T15	2	2	0	0.2685
T1	T10	11	4	0	0.268133
T14	T29	16	11	1	0.267926
T14	T45	16	21	0	0.267892
T3	T27	9	7	0	0.267625
T42	T51	54	32	9	0.267326
T38	T43	4	3	0	0.266286
T45	T46	21	25	1	0.266
T13	T36	3	10	0	0.265846
T38	T53	4	1	0	0.2652
T26	T27	2	7	0	0.264222
T4	T7	1	1	0	0.264
T5	T47	6	6	0	0.2635
T12	T26	5	2	0	0.263429
T1	T45	11	21	0	0.263375
T4	T6	1	2	0	0.262
T48	T50	4	3	0	0.262
T13	T44	3	7	0	0.2618
T10	T19	4	8	0	0.2615
T10	T36	4	10	0	0.260714
T14	T33	16	28	0	0.260591
T3	T21	9	8	0	0.259882
T1	T34	11	16	0	0.259852
T5	T48	6	4	1	0.2598
T9	T49	36	21	2	0.259649
T3	T30	9	15	0	0.258583
T21	T48	8	4	0	0.2585
T14	T51	16	32	1	0.258167

T5	T15	6	2	0	0.25725
T12	T15	5	2	0	0.256857
T18	T48	6	4	0	0.2566
T1	T3	11	9	0	0.2563
T12	T13	5	3	0	0.25625
T12	T50	5	3	0	0.25525
T16	T49	22	21	0	0.254977
T15	T21	2	8	0	0.254
T43	T47	3	6	0	0.253556
T19	T47	8	6	0	0.253286
T5	T13	6	3	0	0.252889
T6	T7	2	1	0	0.25
T18	T38	6	4	0	0.25
T26	T47	2	6	0	0.25
T33	T40	28	29	5	0.25
T10	T44	4	7	0	0.248545
T10	T22	4	3	0	0.248286
T44	T50	7	3	0	0.2482
T10	T30	4	15	0	0.247895
T17	T49	36	21	2	0.247714
T10	T28	4	1	0	0.2476
T1	T33	11	28	0	0.247077
T6	T50	2	3	0	0.2468
T1	T13	11	3	0	0.246
T19	T26	8	2	0	0.246
T14	T50	16	3	0	0.245789
T6	T13	2	3	0	0.2448
T9	T33	36	28	1	0.244344
T1	T18	11	6	0	0.243882
T19	T32	8	15	0	0.243391
T5	T50	6	3	0	0.243333
T26	T53	2	1	0	0.242667
T21	T50	8	3	0	0.241818
T14	T44	16	7	0	0.241652
T38	T41	4	4	0	0.241
T29	T30	11	15	0	0.240923
T25	T44	6	7	0	0.240308
T3	T5	9	6	0	0.24
T13	T28	3	1	0	0.24
T15	T53	2	1	0	0.24
T5	T19	6	8	0	0.239857
T21	T26	8	2	0	0.2392
T29	T38	11	4	0	0.2388
T33	T45	28	21	0	0.238571
T30	T49	15	21	0	0.2385
T19	T27	8	7	0	0.238267
T1	T30	11	15	0	0.238231
T39	T46	19	25	0	0.237864
T30	T39	15	19	0	0.236706

T2	T46	18	25	2	0.235767
T30	T35	15	4	0	0.235579
T4	T38	1	4	0	0.2352
T14	T19	16	8	0	0.235167
T3	T32	9	15	0	0.234917
T9	T45	36	21	2	0.234421
T30	T36	15	10	0	0.23424
T12	T49	5	21	0	0.234231
T15	T19	2	8	0	0.2338
T29	T35	11	4	0	0.233733
T30	T32	15	15	1	0.233733
T1	T19	11	8	0	0.233579
T2	T44	18	7	0	0.23352
T2	T3	18	9	1	0.233481
T17	T31	36	37	3	0.233278
T3	T24	9	3	0	0.233167
T30	T45	15	21	0	0.233111
T2	T38	18	4	0	0.232727
T10	T18	4	6	0	0.2324
T1	T5	11	6	0	0.231529
T1	T9	11	36	0	0.231021
T4	T11	1	3	0	0.231
T37	T49	21	21	0	0.231
T24	T48	3	4	0	0.230857
T18	T36	6	10	0	0.23025
T6	T53	2	1	0	0.23
T40	T49	29	21	0	0.229796
T13	T18	3	6	0	0.229778
T10	T29	4	11	0	0.2288
T1	T32	11	15	0	0.228385
T31	T46	37	25	2	0.227935
T3	T44	9	7	0	0.227875
T19	T33	8	28	0	0.227444
T34	T44	16	7	0	0.227391
T1	T37	11	21	0	0.227375
T19	T21	8	8	0	0.227375
T2	T27	18	7	0	0.22712
T4	T13	1	3	0	0.2265
T44	T51	7	32	1	0.225846
T7	T13	1	3	0	0.2255
T10	T34	4	16	0	0.2255
T3	T34	9	16	0	0.22504
T36	T44	10	7	0	0.224706
T7	T15	1	2	0	0.224667
T17	T45	36	21	0	0.224536
T21	T27	8	7	0	0.224133
T10	T32	4	15	0	0.224105
T7	T35	1	4	0	0.224
T27	T50	7	3	0	0.2238

T1	T27	11	7	0	0.223778
T10	T50	4	3	0	0.223714
T19	T29	8	11	0	0.223368
T12	T21	5	8	0	0.222769
T11	T25	3	6	1	0.222222
T25	T43	6	3	1	0.222222
T2	T32	18	15	0	0.221333
T12	T14	5	16	0	0.221333
T6	T11	2	3	0	0.2212
T33	T39	28	19	0	0.220894
T16	T44	22	7	0	0.22069
T34	T37	16	21	0	0.220432
T38	T52	4	1	0	0.2204
T33	T46	28	25	0	0.220113
T7	T11	1	3	0	0.22
T7	T38	1	4	0	0.22
T29	T31	11	37	1	0.219667
T32	T50	15	3	0	0.219556
T6	T21	2	8	0	0.219
T14	T27	16	7	0	0.21887
T19	T43	8	3	0	0.218364
T19	T34	8	16	0	0.21825
T18	T29	6	11	0	0.218118
T27	T46	7	25	1	0.218063
T3	T39	9	19	1	0.217
T50	T53	3	1	0	0.217
T34	T39	16	19	0	0.216629
T15	T18	2	6	0	0.2165
T14	T35	16	4	0	0.2159
T11	T29	3	11	0	0.215857
T26	T44	2	7	0	0.215778
T37	T45	21	21	0	0.215524
T6	T47	2	6	0	0.2155
T15	T47	2	6	0	0.2155
T33	T51	28	32	0	0.2155
T12	T52	5	1	0	0.215
T5	T36	6	10	0	0.214875
T14	T18	16	6	0	0.214455
T4	T35	1	4	0	0.2132
T18	T43	6	3	0	0.213111
T21	T49	8	21	0	0.212966
T2	T45	18	21	0	0.212718
T14	T16	16	22	0	0.212632
T9	T30	36	15	1	0.212471
T22	T38	3	4	0	0.212
T22	T48	3	4	0	0.212
T14	T37	16	21	0	0.211892
T11	T18	3	6	0	0.211556
T3	T35	9	4	0	0.211385

T3	T38	9	4	0	0.211077
T19	T44	8	7	0	0.210933
T12	T33	5	28	0	0.210788
T13	T14	3	16	2	0.210526
T13	T21	3	8	0	0.21
T19	T35	8	4	0	0.21
T27	T49	7	21	0	0.209643
T21	T30	8	15	0	0.209391
T7	T12	1	5	0	0.208667
T32	T33	15	28	0	0.208651
T22	T52	3	1	0	0.2085
T12	T29	5	11	0	0.208375
T10	T25	4	6	0	0.208
T33	T49	28	21	0	0.207755
T1	T51	11	32	0	0.207535
T27	T36	7	10	0	0.207294
T11	T30	3	15	0	0.207111
T10	T21	4	8	0	0.206833
T10	T27	4	7	0	0.206727
T5	T22	6	3	0	0.206222
T13	T47	3	6	0	0.206222
T3	T16	9	22	0	0.206194
T14	T47	16	6	0	0.206
T46	T51	25	32	1	0.205895
T18	T41	6	4	0	0.2058
T38	T39	4	19	0	0.205739
T5	T29	6	11	0	0.205647
T19	T22	8	3	0	0.205636
T32	T48	15	4	0	0.205579
T14	T31	16	37	0	0.204377
T16	T38	22	4	0	0.203769
T44	T49	7	21	0	0.203714
T3	T51	9	32	0	0.20361
T2	T34	18	16	0	0.203412
T11	T26	3	2	0	0.2028
T21	T36	8	10	0	0.202333
T17	T19	36	8	0	0.202093
T5	T24	6	3	0	0.202
T21	T39	8	19	0	0.202
T35	T41	4	4	0	0.2015
T12	T41	5	4	0	0.200889
T10	T53	4	1	0	0.2008
T25	T47	6	6	0	0.2005
T14	T42	16	54	2	0.200171
T16	T45	22	21	0	0.200047
T10	T14	4	16	0	0.2
T10	T47	4	6	0	0.2