M. Tech (Bioinformatics) [Sadiya Mirza] 2019

SCLERODERMA AND IT'S COMORBIDITIES - EXPLORING LINKS BY NETWORK APPROACH

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY

IN

BIOINFORMATICS

Submitted by:

SADIYA MIRZA

(2K17/BIO/06)

Under the supervision of

Dr. YASHA HASIJA

Associate professor

DEPARTMENT OF BIOTECHNOLOGY

DEPARTMENT OF MECHANICAL ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

JULY, 2019

DEPARTMENT OF BIOTECHNOLOGY,

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, New Delhi -110042

CANDIDATE'S DECLARATION

I, SADIYA MIRZA, Roll No. - 2K17/BIO/06 student of M.Tech (Bioinformatics), hereby declare that the project dissertation titled "Scleroderma and it's comorbidities - exploring links by network approach" which is submitted by us to the Department of Mechanical Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any degree, diploma associate ship, fellowship or other similar title or recognition.

Place: New Delhi	**SADIYA MIRZA**
Date:	**2K17/BIO/06**

DEPARTMENT OF MECHANICAL ENGINEERING,

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, New Delhi -110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "Scleroderma and it's comorbidities - exploring links by network approach" which was submitted by Sadiya Mirza, Roll No. - 2K17/BIO/06 Department of Biotechnology, Delhi Technological University, New Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma in this University or elsewhere.

Dr. YASHA HASIJA

Place: New Delhi SUPERVISOR

Date: Associate professor

Department of Biotechnology

Delhi Technological University

ACKNOWLEDGEMENT

The efficient and effective completion of the project is unfinished without giving the acknowledgement to everyone who made it feasible and the person whose guidance and support was very crucial.

I owe a lot to my supervisor, **Dr. YASHA HASIJA**, Associate Professor, Department of Biotechnology, for supporting and guiding to explore the depth in the search of knowledge and creativity. I would like to thank him for all the assistance and being a source of motivation whenever needed.

I am obliged to everyone who directly or indirectly lent his/her support for successful completion the task. I would like to thank the Department of Biotechnology for providing us advanced computational lab facilities for producing efficient work.

In the end, I thank my parents for always being supportive.

ABSTRACT

Scleroderma is an autoimmune disease characterized by indolent obliterative vasculopathy and widespread fibrosis. The two main morphological manifestations of the disease overlap and may make it difficult to separate activity from damage. Many patients, especially those with the limited subset of the disease, have an indolent course without clear-cut inflammatory manifestations. The aetiology of systemic scleroderma is currently an expanding area of study, since the exact nature of the events underlying this disease remains unclear. It is observed that along with scleroderma a number of other diseases could also occur simultaneously for example myocardial infarction, coronary artery disease, hypertension, osteoarthritis etc. The purpose of this study was to explore the relationship of comorbidities of scleroderma at molecular level with the help of shared genes, protein interactions and biological pathways. The expression profile data datasets were obtained from Gene Expression Omnibus. The differentially expressed genes (DEGs) were screened, followed by functional enrichment analysis, protein-protein interaction (PPI) network construction and analysis of significant network modules. A diseasome was constructed in this study and individual disease association was analyzed by protein interaction networks. At the end, hub genes were identified and further enrichment analysis was done.

Keywords Scleroderma. Differentially expressed genes . Protein-protein interaction network . Hub genes.

CONTENTS

CHAPTER 1: INTRODUCTION

Scleroderma is a rare connective tissue disease that is manifested by cutaneous sclerosis and variable systemic involvement. Two categories of scleroderma are known: systemic sclerosis, characterized by cutaneous sclerosis and visceral involvement, and localized scleroderma or morphea which classically presents benign and self-limited evolution and is confined to the skin and/or underlying tissues. The prevalence of systemic sclerosis (SSc) varies from 2 to 20 per lakh of the population in different parts of the world[5]. Epidemiologic data from India are scarce, the estimated prevalence being 12 per lakh general population[6]. Over 20% of patients with Limited scleroderma develop extracutaneous manifestations such as arthritis, seizures, and uveitis. Neurological complications are the most common association of systemic manifestations in LSsc 2(Mariana et. al.,). Even though current clinical and diagnostic utilities have led to a better understanding of the disease, its pathogenesis still remains unknown. There are also self-reports by patients with systemic sclerosis (SSc) of 5 common, chronic conditions hypertension, diabetes, cancer, depression, and osteoarthritis/back pain 3(Hudson et al., 2009). Despite the impact of the disease in survival and quality of life, data on the epidemiology and clinical expression of comorbidities in SSc so far are limited [7].

Because scleroderma can take so many forms and affect so many different areas of the body, it can be difficult to diagnose. It can be fatal, with a 3-year survival rate of 47-56% in cases of serious pulmonary or cardiac involvement, particularly PAH [7-9]; in fact, it is the single connective tissue disease with the worst survival prognosis .Currently, there is no medication that can cure or stop the overproduction of collagen that is characteristic of scleroderma but a variety of medications can help control scleroderma symptoms and prevent complications.

Although there are still many unanswered questions, the participation of the immune response cells and inflammatory mediators, fibroblasts, and other components of the extracellular matrix and the central role of endothelial damage have changed the paradigm of this disease that was previously considered as predominantly fibrotic. Now it is conceived as a complex syndrome with multiple pathogenic pathways that may be treated simultaneously. The ideal of “targeted therapy” will be an increasingly attainable objective insofar as our understanding of the disease improves.

A systematic exploration of the shared component hypothesis existing between the co-morbidities can shed light about the molecular connections aiding the prevention, early diagnosis and treatment of scleroderma.

The objective of this study include:

- To select, process and analyse microarray datasets from GEO and identify differentially expressed genes (DEGs) in Scleroderma.
- To determine associated diseases, biological terms and pathways related to differentially expressed genes of scleroderma.
- To associate comorbidities with scleroderma at the molecular level by common genes, proteins, biological processes and pathways.
- To contruct a protein protein interaction network and to determine and analyze hub genes.
- To determine the effectof these hub nodes on biological processes and pathways.

CHAPTER 2: REVIEW OF LITERATURE

Scleroderma is an autoimmune disease characterized by indolent obliterative vasculopathy and widespread fibrosis. The two main morphological manifestations of the disease overlap and may make it difficult to separate activity from damage[1]. It is classified as two separate but related entities, a localized form and a systemic form. The classification scheme for morphea presented divides morphea into five categories: plaque, generalized, bullous, linear, and deep[2]. Using this system, these authors estimated the incidence rate of localized scleroderma to be 27 new cases per million population per year. Overall survival was similar to that of the general population[2]. There was a preponderance of female cases (approximately 3:1) for all forms of morphea except for linear scleroderma, which had an even sex distribution. Systemic scleroderma is divided into limited and diffuse disease based on the extent of skin involvement. Recent estimates have placed the incidence rate of systemic sclerosis in the United States at 19 new cases per million adults per year, with an overall prevalence of 240/million adults. The female-to-male ratio is approximately 5:1. The prevalence of scleroderma varies by geographic region and ethnic background and is higher in the United States than in Europe or Japan. Although systemic sclerosis survival has improved over the past two decades, with 5-year survival over 80%, long-term survival is significantly lower than expected, and morbidity is considerable[3]. It is considered a rheumatic disease and a connective tissue disorder. It is also thought to be an autoimmune condition, in which the body's own immune system attacks the body's tissues.This results in an overproduction of collagen, the protein that forms the basis of connective tissue. The result is a thickening, or fibrosis, and scarring of tissue.

Scleroderma is not contagious. It may run in families, but it often occurs in patients without any family history of the disease. It ranges from very mild to potentially fatal. Up to 1 in 3 people with the condition develop severe symptoms[3]. It is not known what causes scleroderma, but it is thought to be an autoimmune condition that causes the body to produce too much connective tissue. This leads to a thickening, or fibrosis, and scarring of tissue. Genetic factors are thought to play a role, and possibly environmental factors[4], but this has not been confirmed. People with scleroderma often come from families in which another autoimmune disease exists.

In the case of systemic scleroderma may affect the connective tissue in many parts of the body. Systemic scleroderma can involve the skin, esophagus, gastrointestinal tract (stomach and bowels), lungs, kidneys, heart and other internal organs. It can also affect blood vessels, muscles and joints. The tissues of involved organs become hard and fibrous, causing them to function less efficiently. The term systemic sclerosis indicates that “sclerosis” (hardening) may occur in the internal systems of the body. There are two major recognized patterns that the illness can take - diffuse or limited disease. In diffuse scleroderma, skin thickening occurs more rapidly and involves more skin areas than in limited disease. In addition, people with diffuse scleroderma have a higher risk of developing “sclerosis” or fibrous hardening of the internal organs.

About 50 percent of patients have a slower and more benign illness called limited scleroderma[5]. In limited scleroderma, skin thickening is less widespread, typically confined to the fingers, hands and face, and develops slowly over years. Although internal problems occur, they are less frequent and tend to be less severe than in diffuse scleroderma, and are usually delayed in onset for several years. However, persons with limited scleroderma, and occasionally those with diffuse scleroderma, can develop pulmonary hypertension, a condition in which the lung’s blood vessels become narrow, leading to impaired blood flow through the lungs resulting in shortness of breath[4].

The pathological events in SSc may include impaired communication between endothelial cells, epithelial cells and fibroblasts; lymphocyte activation; autoantibody production; inflammation; and connective tissue fibrosis. These events result in an accumulation of constituents of the extracellular matrix (ECM), which replaces the normal tissue architecture, which in turn can culminate in organ failure[6]. Endothelial cell damage may be the initiating factor, but the precise triggering event(s) remain elusive. Angiogenesis also appears to be dysregulated[7]. Vasculopathy shows similarities in different organs (e.g. pulmonary arterial hypertension, renal disease, digital tip ulcers). Endothelin-1 is a potent mediator of vasculopathy, and hence represents a highly relevant target for intervention of vascular features in SSc[6]. The pathogenesis of SSc is complex and appears to involve endothelium, epithelium, fibroblasts and immunological mediators, resulting in dysregulated vascular remodelling and, ultimately, vasculopathy. Endothelial cell injury is an early and probably initiating event, but the precise aetiology remains unclear. There are similarities between the vasculopathies of different organs, including pulmonary arterial hypertension, scleroderma renal crisis and digital ulcers, of which ET-1 is believed to be an important mediator. ET-1 is over-expressed in patients with SSc, and its serum plasma concentrations correlate with markers of disease severity. ET-1 therefore represents an important molecular target for therapeutic intervention in the vascular disease manifestations of SSc[8].

SSc is a major medical challenge with high mortality and morbidity[9]. No drug has been so far labelled according to the properties to reduce skin fibrosis or organ involvements. Many trials failed in the past and one might question whether the failure is mostly driven by the wrong choice of the drug, the use of imprecise outcome measures or imperfect selection of the included patients[9]. It can be said that SSc is far more than a 'simple' inflammatory or autoimmune disease.

A census of SSc cases for the period 1989–1991 was conducted in the Detroit area, using multiple sources for case identification. Diagnoses were verified by medical record review. Capture-recapture analysis was used to estimate the total SSc population. Cases of localized scleroderma (morphea and linear disease) were excluded. Based on 706 verified cases of SSc, prevalence was initially estimated to be 242.0 cases per million adults (95% confidence interval [95% CI] 213– 274), with an annual incidence of 19.3 new cases per million adults per year (95% CI 12.4–30.2). Capturerecapture analysis, based on the degree of overlap of verified cases among multiple sources, resulted in a revised prevalence estimate of 276 cases per million adults (95% CI 245–310). This study establishes baseline estimates of SSc occurrence and characteristics in a large US cohort consisting primarily of black adults and white adults. These data should facilitate research regarding the role of geographic, ethnic, racial, and environmental factors for this disease in comparison populations[10].

Like other autoimmune diseases, systemic sclerosis occurs more frequently in women, with a peak of onset in the fifth decade of life[11]. A geoepidemiology studies suggest that systemic sclerosis (SSc) is more common, occurs only at a younger age, and is more severe in African Americans than Caucasians. The major differences noted were mainly in their clinical and serological phenotypes. These differences can be further related to different environmental exposure as well as immunogenetic makeup of these patients. Previous studies had shown that anti-centromere antibodies were strongly associated with renal dysfunction in lcSSc patients and CENP-B was a major target antigen reported for AECA in lcSSc patients indicating the association between anti-centromere antibodies and AECA autoantibodies leading to AECA mediated endothelial dysfunction due to an underlying autoimmune mechanism. This study throws light on a need for some biomarker antibodies and discovery of new target antigens among scleroderma patients and their immunodiagnostic potential[12].

The current era for the treatment of SSc is probably the most exciting of the last decades. Several explanations support the major interest in SSc by investigators and pharmaceutical companies. First, huge developments have been achieved in other autoimmune diseases such as rheumatoid arthritis, but further improvements in these areas are very challenging and require major investments[13]. Today, the SSc clinical spectrum is better known and defined, but unmet needs still remain a critical issue in this rare disease. Refocusing immunotherapies, already used by the rheumatologists in other diseases, makes sense and is in line with the genetic data that demonstrated the shared autoimmunity between several different autoimmune diseases[14]. Skin from SSc patients shows inflammatory infiltrates in which macrophages, T lymphocytes, and dendritic cells are the predominant cell types [15]. These cells produce cytokines and chemokines with proinflammatory and profibrotic activities, and it is very possible that they participate in the process of endothelium-mesenchymal and epithelial-mesenchymal transition, processes by which endothelial and epithelial cells are activated and acquire characteristics similar to myofibroblasts [16].

In the case of immune response, although the factors that promote the persistent activation of cells of the immune system are unknown, recent studies of various groups have pointed towards the Toll-like receptors as possibly responsible for interacting with their classical agonists or with other exogenous or endogenous agonists from damaged tissue to activate dendritic cells, which could secrete proinflammatory cytokines and present antigens to the T cells to activate them[17]. Overexpression of TLR4 and TLR2 has been found in skin and fibroblasts of patients with SSc [18, 19]. On the other hand, the TLR3, TLR7, TLR8, and TLR9, generally viral nucleic acid sensors, could be involved in inflammation in systemic sclerosis; the association between Epstein-Barr virus infection, overexpression of interferon-associated genes, transforming growth factor-beta (TGFβ), and other markers of fibroblast activation. In this sense, persistent damage after a viral infection could cause chronic inflammation and fibrosis in susceptible subjects. Activation of dendritic cells through TLRs generally leads to the production of several proinflammatory cytokines, particularly type I interferons, which have been found overexpressed in peripheral blood in patients with systemic sclerosis (interferon signature)[20].

To perform this literature review, a research through electronic resources was conducted (PubMed, ScienceDirect, Nature, Elsevier, BMJ, and Wiley Online), reviewing references in the English language from the last 10 years. We identified articles via general search of the terms “systemic sclerosis OR scleroderma” and “systemic sclerosis pathogenesis;” the first search yielded 6334 articles, which were handpicked according to relevance that was determined according to the article’s date of publication, ranging from 2008 to October 2018, and its direct relation to scleroderma pathogenesis and directed therapies. Subsequently, we directed a specific search of the terms “Bosentan”, “Macitentan”, “Ambrisentan”, “Selexipag”, “Riociguat”, “bardoxolone methyl”, “Infliximab”, “Adalimumab”, “Rituximab”, “Basiliximab”, “Efalizumab”, “Abatacept”, “AIMSPRO”, “Tocilizumab”, “AM095”, “SAR100842”, “Imatinib”, “Dasatinib”, “Nilotinib”, “CAT-192”, “GC-1008”, “FG-3019”, “P144”, “αvβ6 integrin”, “Pirfenidone”, and “Nintedanib”, which resulted in a range of 2 to 300 references per term[21].

CHAPTER 3: MATERIALS AND TOOLS

3.1 Gene expression omnibus database

GEO is a database which is supported by NCBI (National Center for Biotechnology Information) and provides an open design and flexible platform that possess the ability to submit, store, and retrieve various datasets with high-throughput gene expression and genomic hybridization experiments. The database can be employed to access data for thousands of studies and also provide various strategies and web-based tools that allow users to reach data relevant to their particular interests, as well as ease visualization and analysis of the data without the need of prerequisite computational or knowledge of softwares, and with time- efficient processing and download, thereby greatly increasing the utility of the data. Microarray data of scleroderma and other diseases was accessed and downloaded from GSE95065, GSE82107, GSE113439, GSE118370, GSE25724. The homepage is available at http://www.ncbi.nlm.nih.gov/geo/.

3.2 R/ Bioconductor

Bioconductor is a free and open source software project developed for the comprehension and analysis of high through put genomic data generated by various lab experiments in the field of molecular biology. It is primarily based on statistical R programming language. In this study the bioconductor package was used in R for microarray dataset analysis which resulted in the required differentially expressed genes for further analysis.

3.2 Human protein reference database (HPRD)

Human Protein Reference Database is a web-based and open source resource based on technologies to provide protein information about different aspects of human proteins including protein- protein interactions(PPI), enzyme and substrate relationships, and various disease associations. In this database, the proteins can be accessed by browsing or by using BLAST or by using the query page. Various fields can also be queried at the same time. One of the important and powerful tools in HPRD is it’s search method as any field can be used for searching purpose. The homepage is at http://www.hprd.org.

3.3 DAVID 6.8

DAVID (Database for annotation, visualization and integrated discovery) is a free and open source bioinformatics resource. The DAVID 6.8 Bioinformatics Resources consists of the DAVID Knowledgebase and five other integrated, web-based functional annotation integrated tools: DAVID Functional Annotation Tool, DAVID Gene Functional Classification Tool, DAVID Gene ID Conversion Tool, DAVID NIAID Pathogen Genome Browser and DAVID Gene Name Viewer. DAVID was used to measure the significance of certain genes associated with a specific pathways, Biological Process(BP), Molecular Function(MF) and/or Cellular Component(CC) based on a modified Fisher's exact test. It can be further accessed by http://david.abcc.ncifcrf.gov.

3.4 Cytoscape 3.6.0

Cytoscape integrates high-throughput expression data with biomolecular interaction networks and other molecular states into a single unified conceptual framework. It is an open source software project. In this study, it was used for visual investigation and for the construction of protein protein interaction(PPI) networks. It offers a flexible array of tools to construct biological networks, study the interactions and form a comprehensive set of various topological parameters using its various plug ins. The Network Analyser plug-in was used for analysis of the resulting PPI network in this study.

3.5 STRING

STRING is a widely used database containing both known and predicted protein-protein interactions. The data is curated from various sources including high-throughput wet lab experiments, automated text-mining and genomic context predictions. A PPI network was constructed using the stringApp plug-in in Cytoscape for this study, which accesses the data from STRING database and gives the output results in the form of a network. The STRING resource is available at https://string-db.org/ and the stringApp can be downloaded from http://apps.cytoscape.org/apps/stringApp.

3.7 CytoHubba

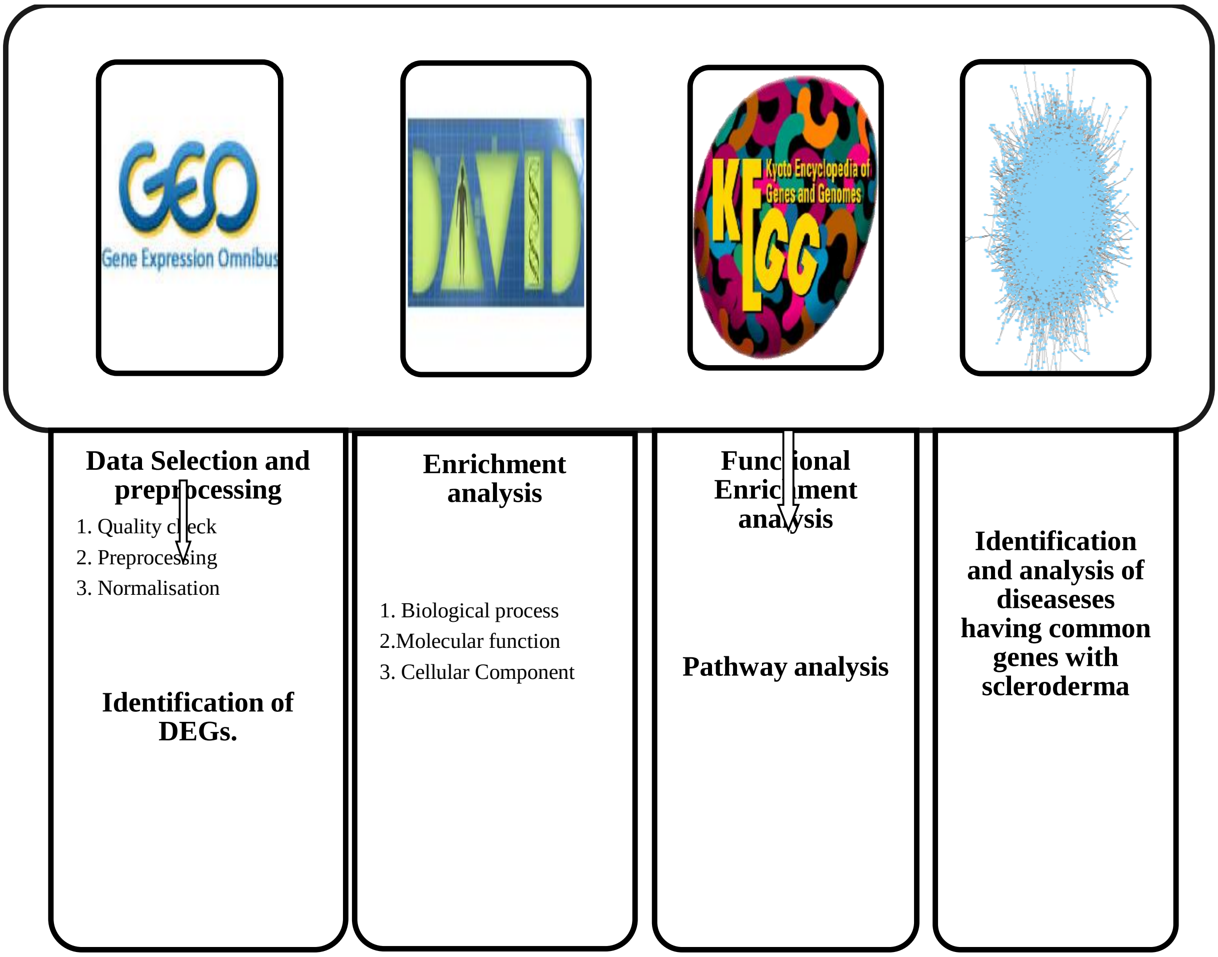
CytoHubba is a user friendly Cytoscape plug-in that allows an effective ranking of network nodes based on various network parameters. This plug-in provides various topological analysis methods which include EPC, Maximal Clique Centrality (MCC)

EcCentricity, Betweenness, Degree, Maximum Neighbourhood Component, and Stress centrality that can be used to identify hub nodes or key genes in any network. The plug-in is available online at http://apps.cytoscape.org/apps/cytohubba.

CHAPTER 4: METHODOLOGY

4.1 Identification of differentially expressed hub genes involved in scleroderma:

Figure 1 Methodology used to identify DEGs in scleroderma datset



4.1.1 Identification and selection of GEO datasets

In this study, gene expression CEL raw data files of dataset GSE95065 from GEO database were used to carry out the analysis. The series GSE95065 provides transcription data of 33 patients used out of which 19 were diseased and 15 were controls.

Table 1 Microarray dataset

Disease Accession number	Samples		Platform	Contributors
SCLERODERMA GSE95065	Total- 33	Disease- 19	GPL23080 [HG-U133A_2] Affymetrix Human Genome U133A 2.0 Array	Mantero JC, Lafyatis R
		Normal- 15		

4.1.2 Microarray gene expression processing

In order to analyze microarray data a specific R package was used named bioconductor. The process of collecting microarray data lead to a high possibility of variation in measured fluorescence intensities from different arrays that result from various steps such as the sample preparation and the hybridization of the samples on the array.

Non-specific binding and optical noise is a major issue which requires that the raw signals from the data are preprocessed before different statistical data analysis methods are applied to the data. In order to remove this variation, the microarray data was preprocessed and normalized using the Robust Multi-array Average (RMA) algorithm taking a log transformation of base 2, lower cut-off value as 1.0 and centering of the data was done around the mean.

4.1.3. Identification of differentially expressed genes

Various statistical tests which will rank the genes in the order of significance of differential expression were applied. The fold change was calculated and a threshold was set according to which genes were ranked. Unpaired t-test was performed in order to identify genes that were differentially expressed between the normal and diseased samples. Probes with expression values of $p < 0.05$ and corresponding values of False Discovery Rate (FDR) < 0.01 were considered to be statistically significant. For GSE95065, a threshold of fold change $|FC| > 1.2$ was considered statistically significant. The Bioconductor package limma was used for the differential expression analyses.

4.1.4 Gene enrichment analysis

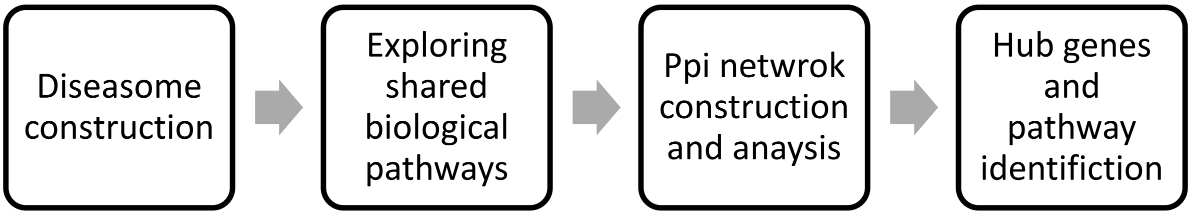
The differentially expressed probes were mapped to their corresponding genes and DEGs were processed through DAVID to identify prominent disease-related biological pathways and processes to better understand the underlying disease mechanism. A total of 236 genes which included 145 upregulated genes and 91 downregulated genes were used as the input gene set for enrichment. The diseases enriched were choosen according to the genes involved as well as their p-value and then their co occurrence with scleroderma was verified with the help of published literature. In order to understand the functional changes of the DEGs, their biological processes (BP), molecular function (MF) and cellular components (CC) were also analysed by DAVID. Only GO terms with p-value < 0.05 were considered statistically significant.

4.1.5 Pathway enrichment analysis

DAVID 6.8 was used for enrichment of pathways. The DEGs obtained were used to perform a Fisher's exact test which was then followed by multiple test correction using BH's False Discovery Rate correction. A threshold of p-value of < 0.05 was considered statistically significant so as to determine the overrepresentation of some biological pathways.

4.2 Linking scleroderma with comorbidities

Figure 2 Workflow of methodology for linking scleroderma with diseases


Diseasome construction
Exploring shared biological pathways
Ppi netwrok construction and anaysis
Hub genes and pathway identifiction

4.2.1 Identification and selection of GEO datasets

The disease enriched during functional annotation using DAVID were analyzed and according to the number of genes involved as well as p value four diseases were taken in order to study their relationship with scleroderma. The occurrence of these diseases with scleroderma was verified through published literature.

Individual processing of microarray datasets of these four disease was done in order to remove biasness. The raw files from different studies of these diseases were processed using the bioconductor packages in R. A threshold value of 1 fold change and a p-value< 0.5 were chosen as threshold for selecting the differentially expressed genes. For every disease, the dataset was selected only if it contained a minimum number of five samples in both control and disease category.

Table 2 Microarray datsets of comorbidity genes

Disease Accession number	Sample		Platform	Contributers
Osteoarthritis GSE82107	Total-17	Disease-10 Normal-7	GPL570	Broeren MG, de Vries M, Bennink MB, van Lent PL et al
Hypertension GSE113439	Total-26	Disease-15 Normal-11	GPL6244	Mura M
Lung cancer GSE118370	Total-12	Disease-6 Normal-6	GPL570	Xu L, Lu C, Huang Y, Zhou J et al.
Type 2 diabetes GSE25724	Total-13	Disease-6 Normal-7	GPL96	Dominguez V, RaimondiC, et al.

4.2.2. Diseasome construction

"Diseasome" was termed as the association between scleroderma and its comorbidities. The two diseases were considered linked if they share variations in similar set of genes. These genes were identified for respective diseases. An important and underlying cause of comorbidities could be the deregulation of common set of genes either in opposite or in similar direction. To obtain the molecular interpretation of the comorbidities, meta analysis was done to compare the direction of dysregulation of the genes shared by scleroderma and its comorbidities. The DEGs deregulated in the same direction can be considered as recognized signatures of the comorbidities while the DEGs deregulated in the opposite direction can indicate to inverse comorbidities.

4.2.3 Exploring the shared biological processes diseasome

To discover the biological functions shared by the scleroderma and its associated diseases, a functional enrichment analysis was performed based on the BP. Further, investigation of the pathways that were common between the scleroderma and its comorbidities was done.

4.2.4 PPI network construction and analysis

STRING database was used to construct a protein protein interaction (ppi) network on Cytoscape. Only query proteins were displayed and a minimum required interaction score of 0.4 (medium confidence) was kept as the primary parameter. using Network Analyser tool was used for analysis of properties such as degree and shortest path. Further, by using a Cytoscape plugin, MCODE module selection was done using their pre-set cut- off criteria.

4.2.5 DEG PPI Network Analysis to identify hub genes

cytoHubba, another Cytoscape plugin which uses five algorithms, namely, MCC, MNC, EPC, Degree and EcCentricity was used to categorize hub genes. The overlapped genes of the five algorithms mentioned above were further enriched using DAVID in terms of GO molecular functions, GO biological processes, GO cellular components and KEGG pathways.

CHAPTER 5: RESULTS

5.1 Identification of differentially expressed genes in normal skin samples and scleroderma patients

5.1.1 Selection of GEO datasets

A total of 20 data series from GEO were analyzed for Scleroderma using keywords, ‘Scleroderma’, ‘Homo sapiens’ and ‘Expression profiling by array’. On the basis of study design, treatment based studies and other dataset details 19 datasets were removed. Finally, 1 data series, GSE95065 was found suitable for the purpose of this study and was downloaded from GEO. 19 samples of scleroderma and 15 samples of normal skin were taken for this study.

5.1.2 Identification of differentially expressed genes

The series GSE95065 was used to identify DEGs between normal skin samples and scleroderma. A global normalisation, quantile normalisation, log-transformation and centring on the mean was performed on all the samples. Student’s t-test and BH correction was performed on the data. A threshold of Fold change of 1.2 and p-value less than 0.05 were used to identify the DEGs. A total of 235 genes were identified which included 145 upregulated genes and 91 downregulated genes.

Table 3 Up and Down regulated genes in scleroderma

Upregulated DEGs	Genes
	CHP1,SLC35A2,SGCG,GRIA2,RXRB,TOX4P1,IL1RN,SGCA,FGFR1,HBA2,HBA1,CASP2,PLEC,LIPE,AZGP1P1,TNXB,TNXA,MDM4,TBC1D3P1,DHX40P1,CDK2,PPIF,SOX9,NEAT1,GLRB,RASA4DP,AZGP1,EXPH5,C9orf16,LTBP4,PRKAB1,TPSB2,CDSN,TPSAB1,CES1P1,TAF9B,CDH1,RASA4CP,RAB35,IL17RC,SH3GLB2,GOLGA8A,NDRG2,ZEB1,APOE,TSPAN8,LGR5,CLPTM1,GOLGA8B,TRRAP,SIGMAR1,ACTB,CES1,ALDH3B2,EXOSC7,MFAP4,SLC29A1,CLEC3B,NUCB1,SERPINB13,PPAP2B,CAB39L,NAB1,PCOLCE2,DGKA,C

	CNG2,PRSS12,NPR2,RECK,FGFR3,LEPR,SRSF1,TOX4,DDX17,PLA2R1,LRRC2,MAT2A,FXYD1,AGTR1,NUMA1,ALDH1A1,RNF126,PTBP1,MAF,TUBGCP3,MCOLN3,ELF5,TMEM259,CD44,POLR2E,LAMP1,ENO1,TRIM29,GAS1,CLU,LPIN1,RNFT1,GM2A,UBE4B,PSEN1,F2RL1,FBLN5,PPP2R1A,WDR1,FBLN1,RASA4B,IGFBP6,NONO,WIF1,NXT2,MTMR1,HNRNPH1,RARG,OGT,CDK10,SVEP1,EPHB3,HMG20B,RASA4,ZFP36L2,SQLE,DAPK3,ZBTB7A,MAP2K2,PKP3,ZNF273,GSN,WISP2,OLFML3,ABHD6,CDA,NOVA1,MMP27,APOD,DEDD,KLHL20,DLG1,MVK
Downregulated DEGs	
	PECAM1,TDO2,PTPRC,CCL8,SERPINE2,FCGR1A,COL4A4,FCGR1C,CHN1,SFRP4,PRND,CACNB4,PTX3,PALLD,ISG15,SUGCT,CFI,GOLIM4,PXDN,MECOM,COL5A2,PAPSS2,ERG,PSD3,LAMP5,PDE1A,GEM,ZFR,ADAM12,DIO2,KERA,CCL19,KLRB1,C6,WNK1,LGALS2,CCR1,CD93,ANGPT2,DUSP1,MED18,MIS18BP1,ADAMDEC1,COL10A1,COL4A1,LYN,PENK,CYR61,THY1,SETD2,SERPINE1,EIF4G1,ASPN,EPHA4,KCNE4,SCG2,COL11A1,SULF1,CCL2,CCND2,TAF13,NID2,TNC,CTGF,PRSS23,CDH11,STMN2,APLN,RNOX4,IGSF6,LRRC15,AKR1B10,CALD1,KRA,SZBED2,FCGR1B,ARPC1B,FOS,IFI27,SLAMF8,ITGA8,SELP,SELE,FMO3,PRRC2C,TNFSF4,CEP350,PTGS2

5.1.3 Functional annotation using DAVID

The DEGs were analyzed by DAVID server and gene disease associations were obtained. On submission, a total of 190 disease categories were enriched out of which 4 (osteoarthritis, T2D, Hypertension and lung cancer) were chosen according to number of genes involved and p-value.

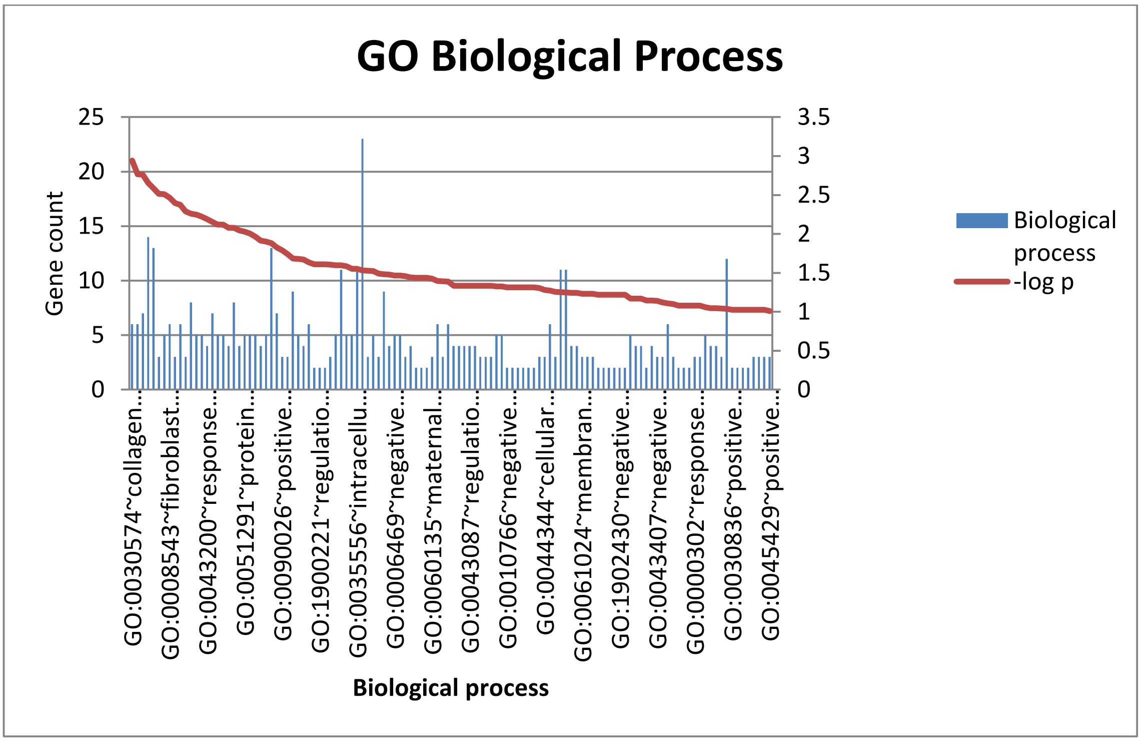
Table 4 Diseases enriched having common genes with scleroderma

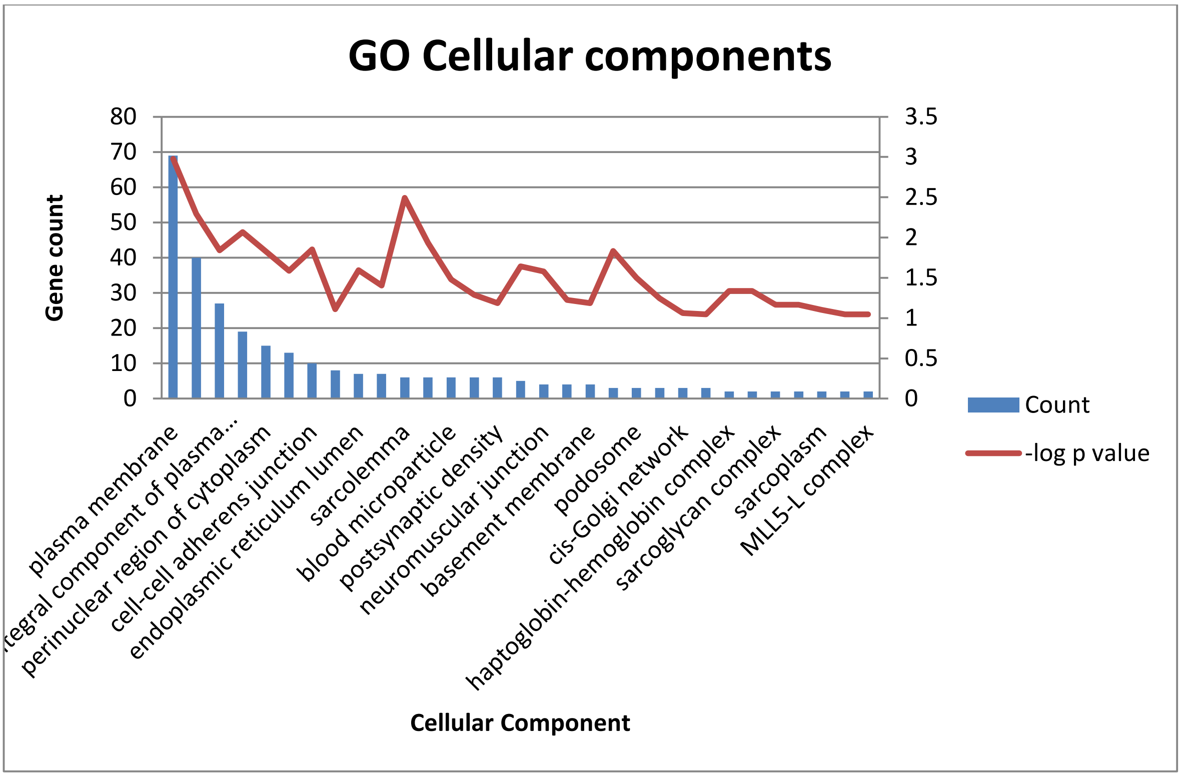
Disease	Gene count	p-value	Genes involved
Type 2 Diabetes| edema | rosiglitazone	47	0.00210706	CCL2, PTGS2, C6, LTBP4, CCR1, LEPR, F2RL1, IGFBP6, MMP27, SLC29A1, AGTR1, FOS, CD44, ISG15, APOD, APOE, PDE1A, SERPINE1, RNFT1, PTX3, COL11A1, SUGCT, RECK, NOX4, SELP, TNXB, RARG, TNFSF4, COL4A1, MAT2A, RXRB, LGALS2, IL1RN, WNK1, PRKAB1, NPR2, PALLD, COL5A2, CLPTM1, RGS1, DUSP1, SQLE, AKR1B10, PKP3, MDM4, SELE, LIPE
Chronic renal failure|Kidney Failure, Chronic	26	7.29E-04	CCL2, PTGS2, LEPR, CCR1, TNC, LTBP4, SLC29A1, ALDH1A1, FOS, AGTR1, KRAS, CD44, APOE, FMO3, SERPINE1, CDA, SELP, TNXB, IL1RN, WNK1, CDK2, DUSP1, CCND2, ITGA8, SELE, LIPE
Hypertension	23	3.07E-04	SELP, FGFR1, PXDN, COL4A1, CCL2, CES1, TNFSF4, FGFR3, PTGS2, LEPR, IL1RN, WNK1, NPR2, HBA1, LPIN1, AGTR1, APLNR, DIO2, APOE, PECAM1, FMO3, SERPINE1, SELE
Multiple Sclerosis	17	0.006977254	KLRB1, PTPRC, SELP, CCL2, PTGS2, CCR1, C6, IL1RN, CCL8, CDSN, RGS1, DUSP1, APOE, PECAM1, SERPINE1, SELE, CASP2
Coronary Artery Disease	16	0.004962461	SELP, FGFR1, COL4A1, CCL2, PTGS2, IL1RN, NID2, LGR5, AGTR1, APLNR, APOE, PECAM1, SERPINE1, SUGCT, SELE, CDH11
Osteoarthritis	8	3.44E-05	ASPN, DIO2, PTGS2, CLEC3B, IL1RN, ADAM12, COL11A1, PAPSS2
Lung cancer	15	0.012794905	CCL2, PTGS2, RXRB, LEPR, IL1RN, IGFBP6, CDH1, FOS, KRAS, SERPINE2, CTGF, APOE, SERPINE1, CDA, SELE
Type 2 diabetes	13	0.004448318	MAF, AGTR1, SELP, CCL2, PTGS2, APOE, LEPR, SERPINE1, IL1RN, TSPAN8, MVK, LGR5, SELE

5.1.4 Functional enrichment analysis

To understand the function of the DEGs, their biological processes (BP), molecular function (MF) and cellular components (CC) were analysed by DAVID. To understand the function of the DEGs, their biological processes (BP), molecular function (MF) and cellular components (CC) were analysed by DAVID. The DEGs were mainly enriched in positive regulation of apoptotic process(GO:0043065), collagen catabolic process(GO:0030574),signal transduction (GO:0007165), immune response(GO:0006955), inflammatory response (GO:0006954) in the BP group.

In CC group, the DEGs were mostly enriched in plasma membrane (GO: 0005886), cytosol (GO:0005829), extracellular exosome (GO: 0070062), internal component of plasma membrane (GO: 0005887), golgi apparatus (GO:0005794). Finally, in the MF group, the genes enriched in protein binding (GO: 0005515), calcium ion binding (GO: 0005509), heparin binding (GO: 0008201), protein kinase activity (GO: 0004672), cadherin binding involved in cell-cell adhesion (GO:0098641), extracellular matrix structural constituent (GO: 0005201), carbohydrate binding (GO:0030246).Only statistically significant terms, with p-values < 0.05, were considered for analysis.





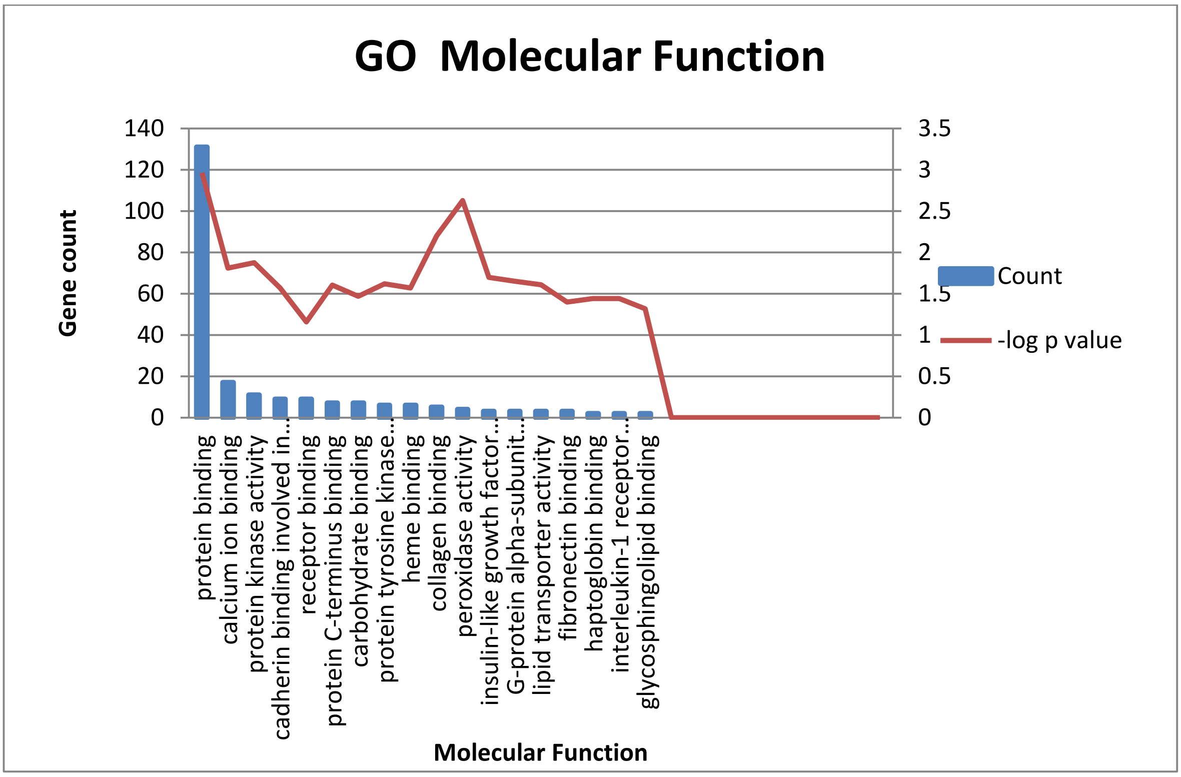


Figure 3 Bar graph representation of GO analysis and significantly enriched GO terms for DEGs in this dataset. GO analysis classified the DEGs into three groups (molecular function, biological process, and cellular component).

Table 5 Results of Table 5.1 A tabular representation of analysis of significant enrichment GO terms of DEGs

CC				
Term	Process	Count	PValue	Genes
GO:0005886	plasma membrane	69	0.001054699	LEPR, F2RL1, IL17RC, LGR5, AZGP1, AGTR1.
GO:0005829	cytosol	59	6.94E-04	FOS, TDO2, RASA4B, ISG15, CTGF
GO:0070062	extracellular exosome	57	2.78E-05	SRSF1, GM2A, LTBP4, IGFBP6, TSPAN8,
GO:0005576	extracellular region	50	3.21E-10	FGFR1, FGFR3, KERA, C6, LTBP6, TNC
GO:0016020	membrane	40	0.005075518	LEPR, CDH1, PRRC2C, DGKA, SLC29A1
GO:0005615	extracellular space	39	3.25E-07	PXDN, CCL2, KERA, TNC, LTBP4, IGFBP6
GO:0005887	integral component of plasma membrane	27	0.01441721	FXYD1, LGR5, CCR1, F2RL1,FGFR1
BP		Count	PValue	Genes
GO:0007165	signal transduction	23	0.029301358	ERG, CCL2, TNFSF4, LYN, IGFBP6,
GO:0007155	cell adhesion	21	1.07E-06	SELP, CCL2, SVEP1, TNXB, TNC, CCR1, ADAM12, MFAP4,
GO:0030198	extracellular matrix organization	18	3.34E-10	RECK, COL4A4, PXDN, COL4A1, TNC,

5.1.5 Pathway enrichment analysis

GO:0006955	immune response	14	0.002221576	PXDN, TNFSF4, CCL2, CCR1, IL1RN, CCL19,
GO:0001525	angiogenesis	13	2.15E-05	FGFR1, CCL2, PTGS2, LEPR, EPHB3, THY1,
GO:0006954	inflammatory response	13	0.002609795	NOX4, SELP, CCL2, LYN, PTGS2, CCR1,
GO:0008284	positive regulation of cell proliferation	13	0.013128214	EIF4G1, FGFR1, KRAS, FGFR3, RARG,
GO:0008285	negative regulation of cell proliferation	11	0.025310062	NOX4, AZGP1, RARG, SERPINE2, PTGS2,
MF				
GO:0005515	protein binding	131	0.001083491	PTGS2, LTBP4, F2RL1, NONO, AZGP1, AGTR1,
GO:0005509	calcium ion binding	17	0.015485702	ASPN, SVEP1, LTBP4, MMP27, CHP1, CDH1,
GO:0008201	heparin binding	12	4.56E-06	FGFR1, SELP, WISP2, CCL2, TNXB,
GO:0004672	protein kinase activity	11	0.013357583	EPHA4, CCL2, MAP2K2, PRKAB1, WNK1, CCL8,
GO:0005178	integrin binding	10	5.73E-06	FBLN1, WISP2, TNXB, LYN, CTGF,
GO:0098641	cadherin binding involved in cell-cell adhesion	9	0.026939683	EIF4G1, SH3GLB2, PKP3, TRIM29,
GO:0005201	extracellular matrix structural constituent	7	1.69E-04	COL4A4, FBLN1, PXDN, COL4A1,

The gene set of 235 DEGs was used to identify statistically significant pathways with p-value < 0.05. Significantly enriched pathways highlighted were Akt signaling pathway, pathways in cancer, Focal adhesion.

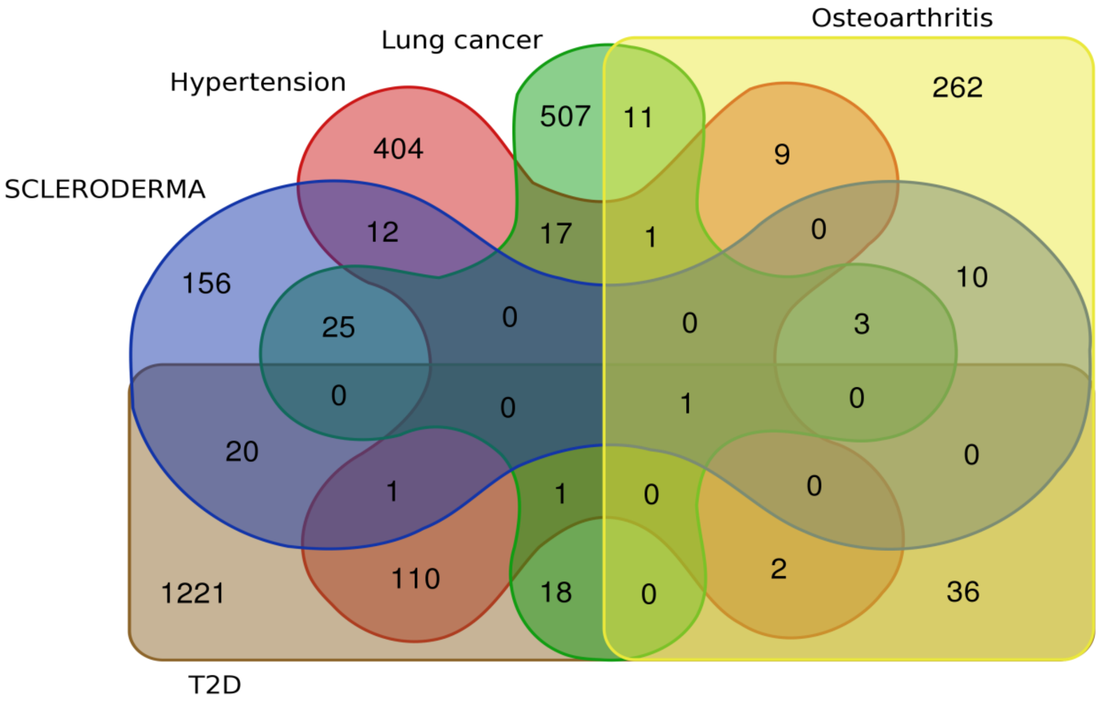
Table 6 Significant KEGG pathways identified for the DEGs

Term	Count	PValue	Genes
hsa04151:PI3K-Akt signaling pathway	15	0.001880653	COL4A4, PPP2R1A, FGFR1, COL4A1, TNXB, FGFR3, MAP2K2, TNC, COL5A2, CDK2, KRAS, CCND2, ITGA8, ANGPT2, COL11A1
hsa05200:Pathways in cancer	14	0.014627132	COL4A4, FGFR1, AGTR1, FOS, KRAS, FGFR3, COL4A1, PTGS2, RXRB, MAP2K2, CDH1, MECOM, DAPK3, CDK2
hsa05206:MicroRNAs in cancer	11	0.022104834	RECK, KRAS, FGFR3, TNXB, CD44, PTGS2, CCND2, MAP2K2, TNC, MDM4, ZEB1
hsa04510:Focal adhesion	9	0.022693919	ACTB, COL4A4, COL4A1, TNXB, CCND2, TNC, ITGA8, COL11A1, COL5A2
hsa04512:ECM-receptor interaction	8	6.21E-04	COL4A4, COL4A1, TNXB, CD44, TNC, ITGA8, COL11A1, COL5A2
hsa04921:Oxytocin signaling pathway	8	0.013130388	ACTB, FOS, KRAS, PTGS2, MAP2K2, PRKAB1, NPR2, CACNB4
hsa05144:Malaria	7	1.56E-04	KLRB1, SELP, CCL2, PECAM1, HBA2, HBA1, SELE
hsa04390:Hippo signaling pathway	7	0.041858879	ACTB, PPP2R1A, CCND2, CTGF, SERPINE1, CDH1, DLG1
hsa05410:Hypertrophic cardiomyopathy (HCM)	6	0.009930845	ACTB, SGCG, ITGA8, PRKAB1, CACNB4, SGCA

5.2 Diseasome construction

Scleroderma and it’s comorbidities had a number of genes in common which ranged from 10 to 25. Scleroderma had 10 genes in common with osteoarthritis, 25 with lung canser, 12 with hypertension, 20 with T2DM. We constructed a scleroderma specific interactome and utilized the knowledge obtained from the interactome to mine out the commonality between scleroderma and its comorbidities. The scleroderma diseasome constructed using the knowledge obtained above revealed that the comorbidities were connected with scleroderma through shared genes or proteins which interact to form the cellular interactome.

Figure 4 Venn diagram representing overlapping genes in diseasome



5.3. Construction of PPI network

A protein-protein interaction network was constructed using STRING database on Cytoscape. Only query proteins were displayed and a minimum required interaction score of 0.4 (medium confidence) was kept as the primary parameter.

5.3.1 Scleroderma and hypertension

A protein protein interaction network was build by stringApp of cytoscape and the interaction can be seen in the figure

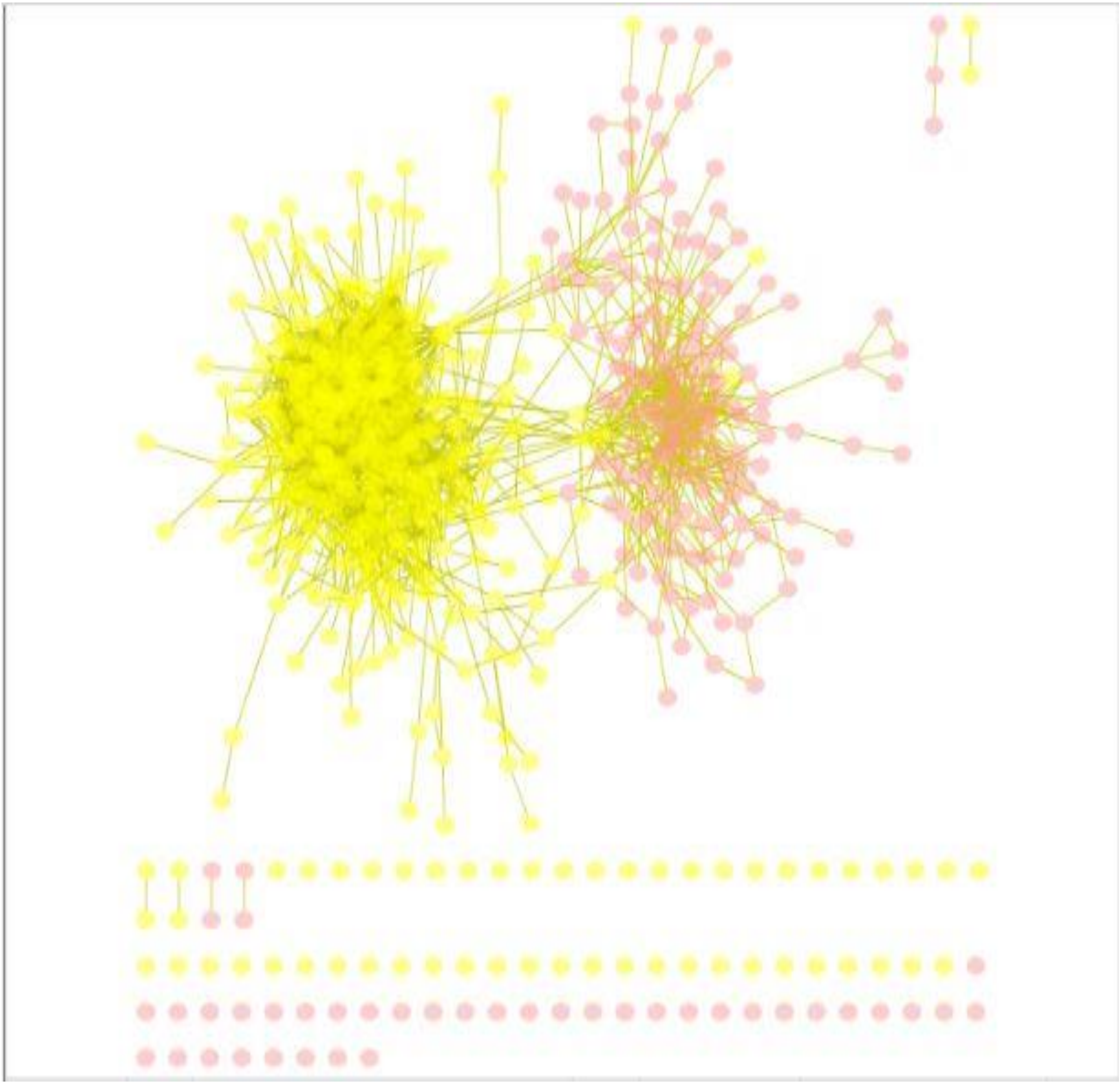
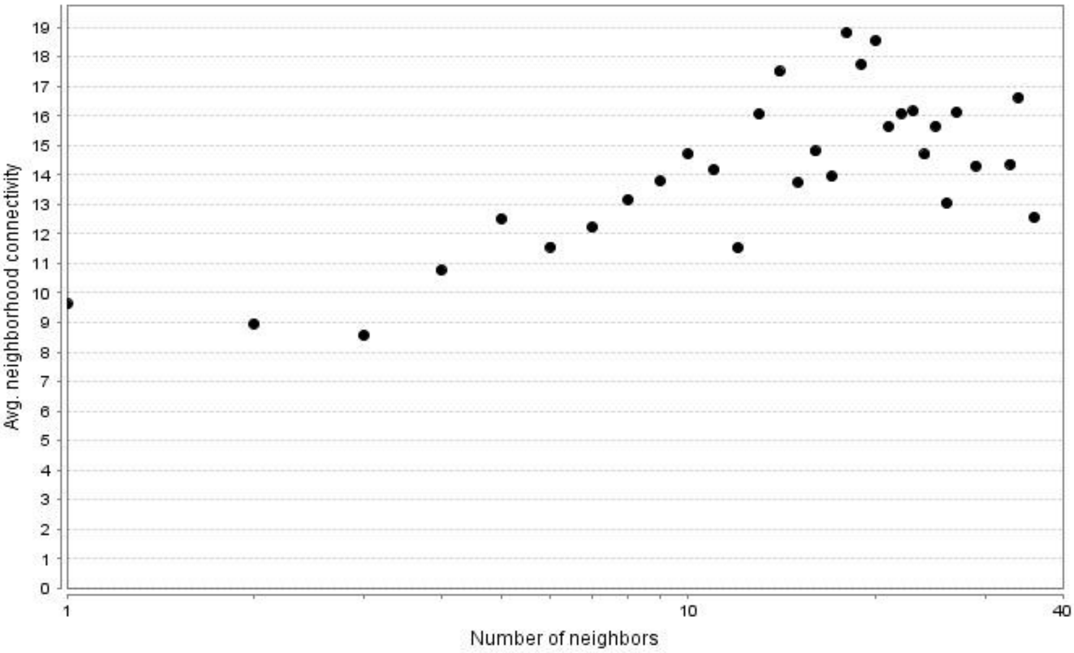
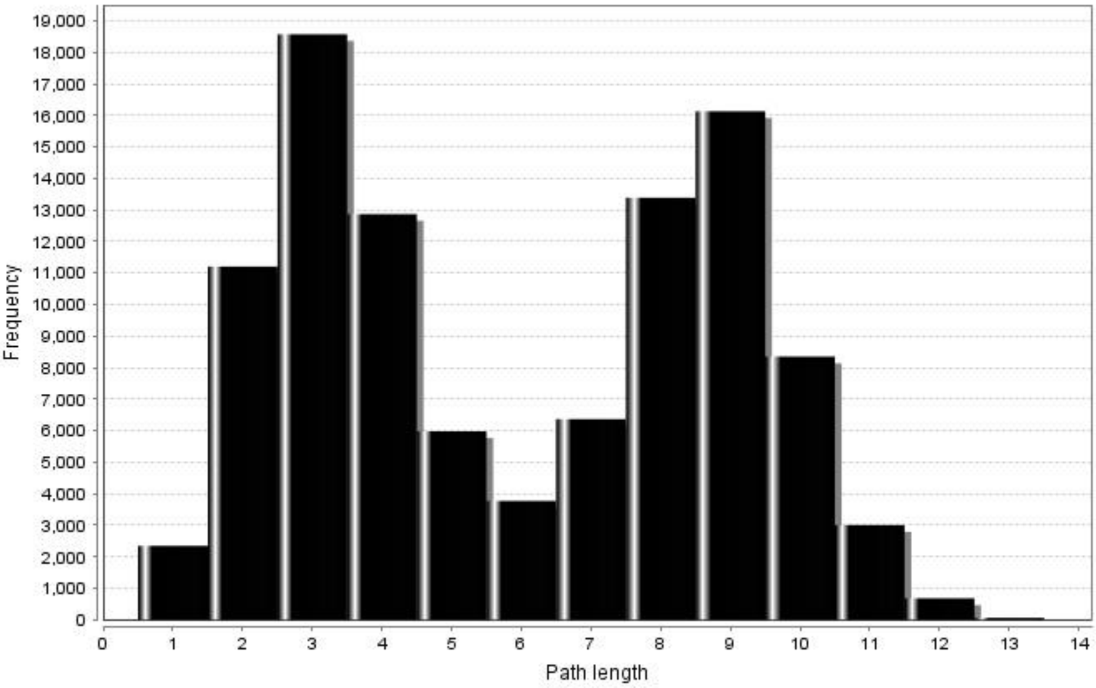


Figure 5 Yellow nodes—proteins involved hypertension, pink nodes— proteins involved in scleroderma, grey edges- protein protein interactions.

The network was analyzed by NetworkAnalyzer tool of cytoscape. The network had 330 nodes, 1172 edges with a clustering coefficient of 0.301. The following network properties of degree distribution, average neighbourhood connectivity, average aggregation coefficient and distribution of shortest path were also analysed as seen in Figure.


Avg. neighborhood connectivity
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
1
10
40
Number of neighbors


Frequency
19,000
18,000
17,000
16,000
15,000
14,000
13,000
12,000
11,000
10,000
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Path length

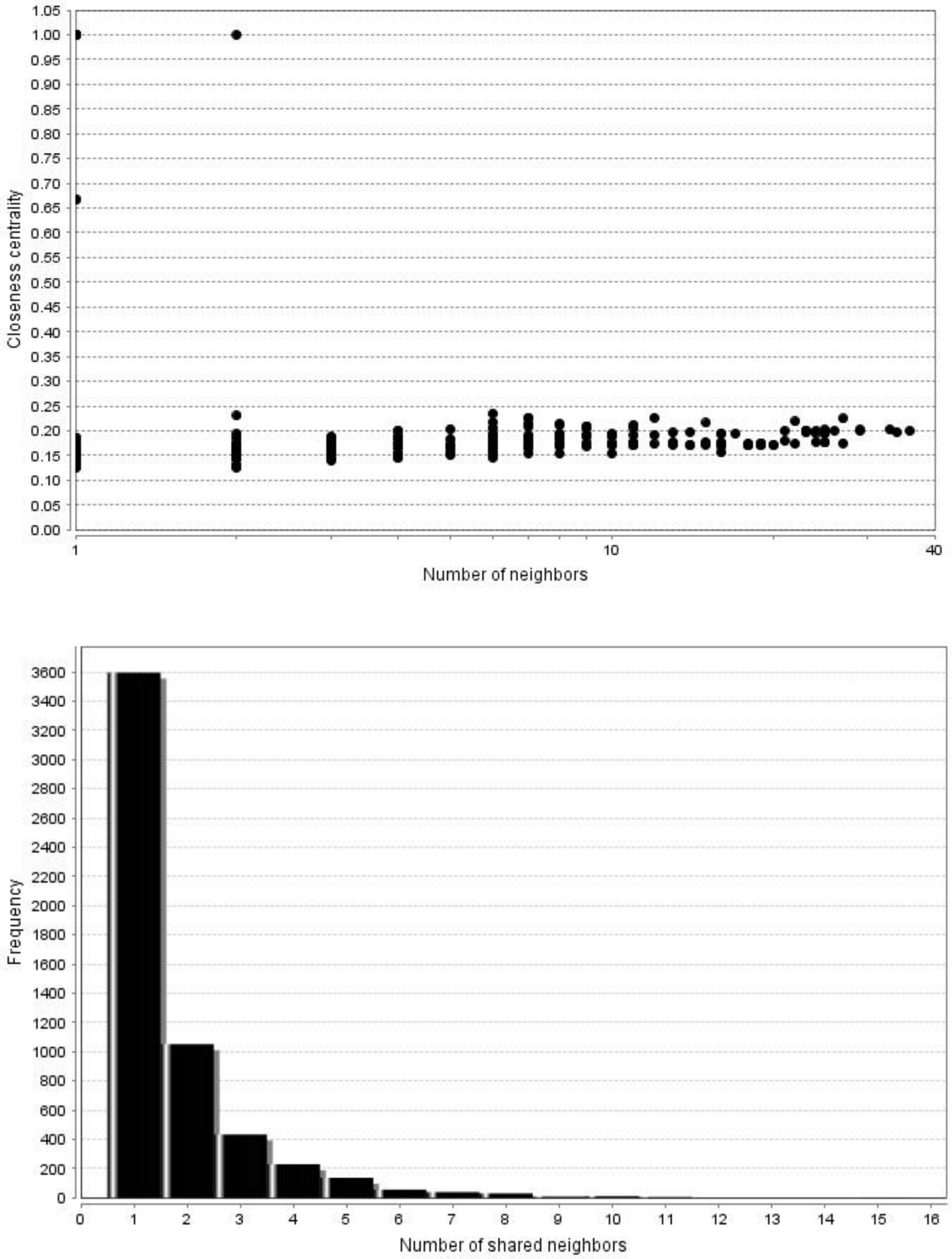
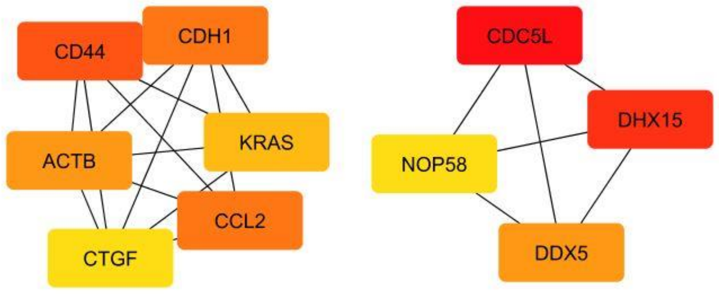


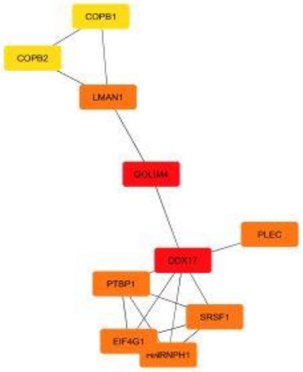
Figure 6 PPI network analysis of DEGs by NetworkAnalyser tool

5.3.1.1 Identification of hub genes through PPI network analysis DEGs

The cytoHubba application which uses 11 algorithms to give hub genes was employed to identify the top 20 genes evaluated using the following algorithms: Degree, MCC, MNC, EPC and EcCentricity as shown in the figure.

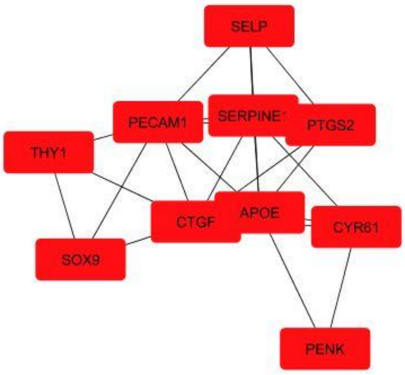


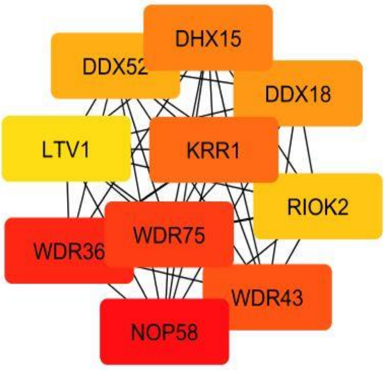
a) Degree



b) EcCentricity

c) EPC





d) MCC

e) MNC

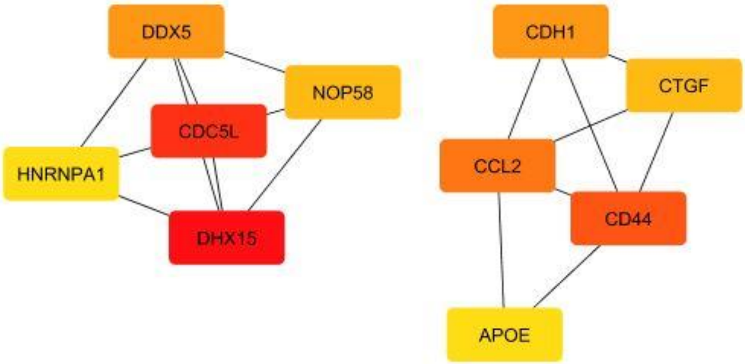
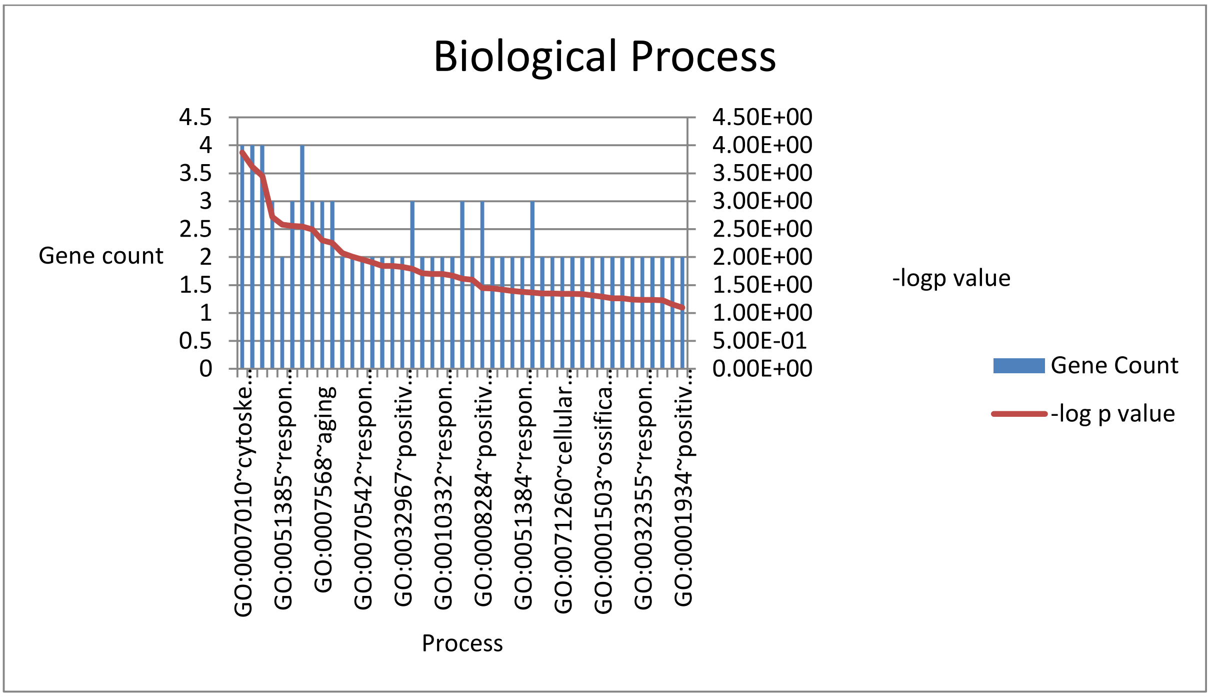
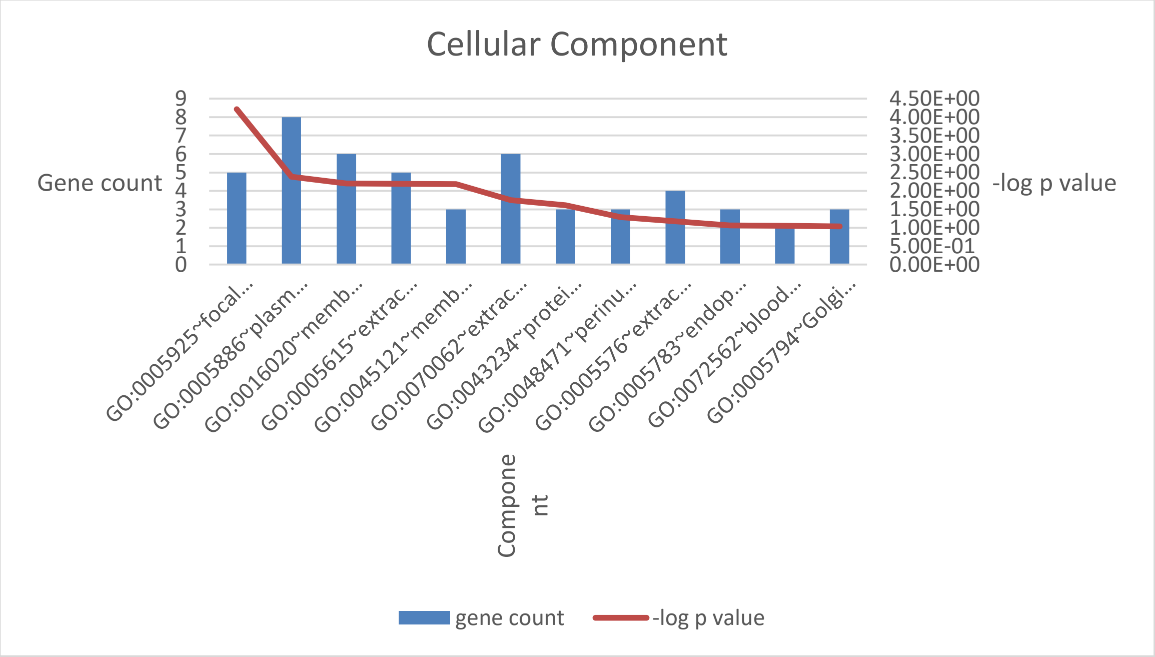


Figure 7 Top 10 genes evaluated in the PPI network using five calculation methods

Then, the hub genes were identified by taking the intersection of these results. These genes include ACTB, CD44, CDH1, KRAS, FOS, CCL2, SOX9, PTGS2, CTGF, PECAM1, APOE, PTPRC.

5.3.1.2 Pathway Analysis





Gene	MCC	Gene	MNC	Gene	Degree	Gene	EPC	Gene	EcCentricity
NOP58	1.25E+10	DHX15	33	CDC5L	36	CD44	160.014	DDX17	0.13896
WDR36	1.25E+10	CDC5L	32	DHX15	34	CDH1	160.014	GOLIM4	0.13896
WDR75	1.25E+10	CD44	31	CD44	33	CCL2	160.014	PTBP1	0.12159
WDR43	1.25E+10	CCL2	29	CDH1	29	ACTB	160.014	SRSF1	0.12159
KRR1	1.25E+10	DDX5	27	CCL2	29	KRAS	160.014	HNRNPH1	0.12159
DHX15	1.25E+10	CDH1	27	DDX5	27	APOE	160.014	EIF4G1	0.12159
DDX18	1.25E+10	NOP58	25	ACTB	27	CTGF	160.014	PLEC	0.12159
DDX52	1.25E+10	CTGF	25	KRAS	26	PTGS2	160.014	LMAN1	0.12159
RIOK2	1.25E+10	HNRNPA1	24	NOP58	25	PTPRC	160.014	ACTB	0.10808
LTV1	1.25E+10	APOE	24	EPRS	25	PECAM1	160.014	SOX9	0.10808
NOC3L	1.25E+10	EPRS	24	NCL	25	SOX9	160.014	GSN	0.10808
ESF1	1.25E+10	NCL	23	APOE	25	FOS	160.014	PPP2R1A	0.10808
SDAD1	1.25E+10	PECAM1	23	CTGF	25	SERPINE1	160.014	POLR2E	0.10808
MPHOSPH10	6.23E+09	PTPRC	23	HNRNPA1	24	COL4A1	160.014	COPB1	0.10808
DDX5	6.23E+09	ESF1	22	HSPA4	24	CYR61	160.014	HSPA5	0.10808
NCL	730936	PTGS2	22	PTGS2	24	THY1	160.014	ISG15	0.10808

RSRC1	40385	ACTB	21	RANBP2	23	LYN	160.014	COPB2	0.10808
CD44	37168	KRAS	21	PTPRC	23	COL5A2	160.014	USO1	0.10808
PECAM1	34310	SERPINE1	21	PECAM1	23	SELP	160.014	NOVA1	0.10808

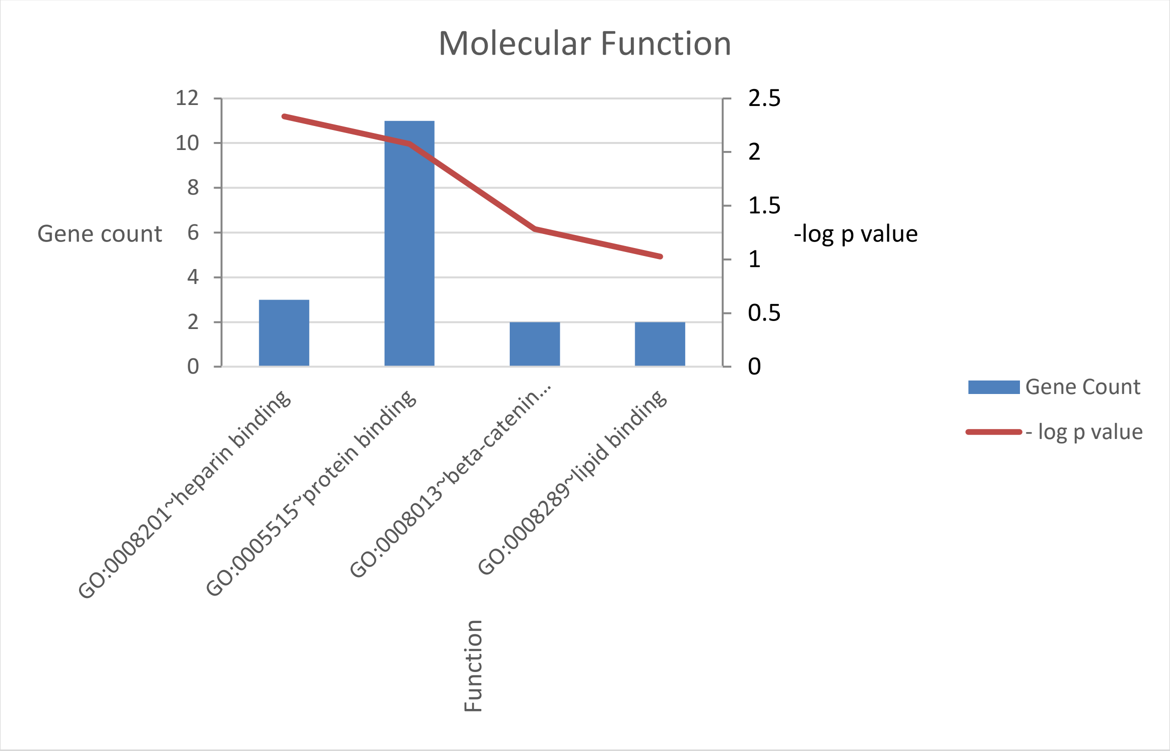


Figure 8 Bar graph representation of GO analysis depicting significantly enriched GO terms and pathways for DEGs in Cluster A. GO analysis classified the DEGs into three groups (molecular function, biological process, and cellular component).

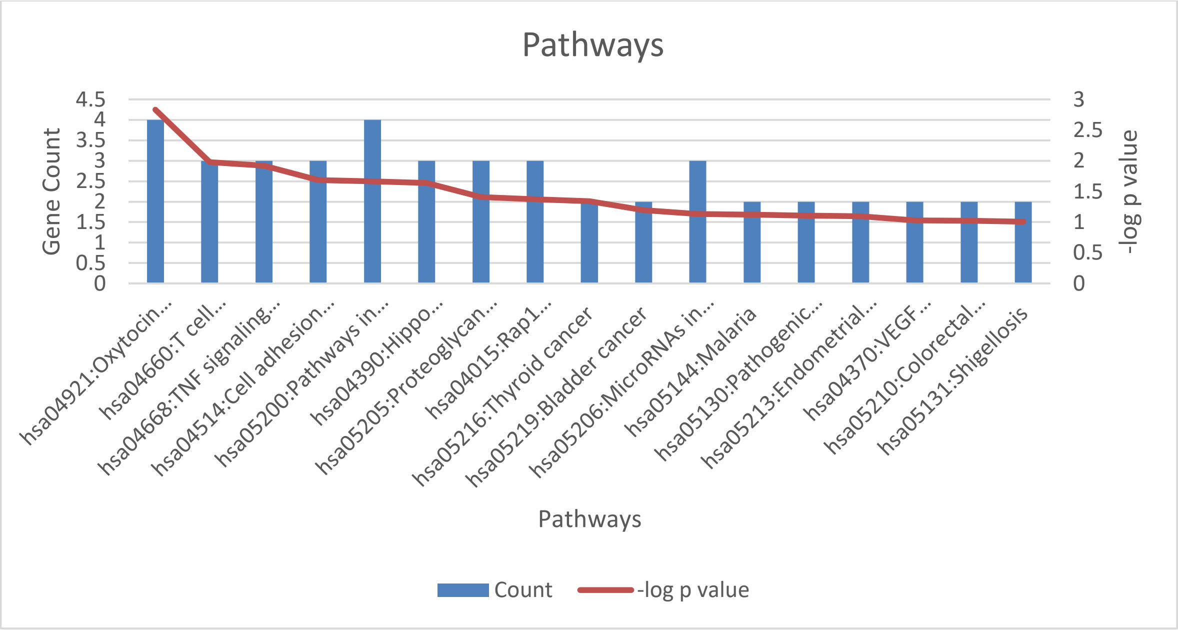


Figure 9 Enriched KEGG pathways

5.3.2 Scleroderma and Type 2 Diabetes mellitus

A protein protein interaction network was build by stringApp of cytoscape and the interaction can be seen in the figure.

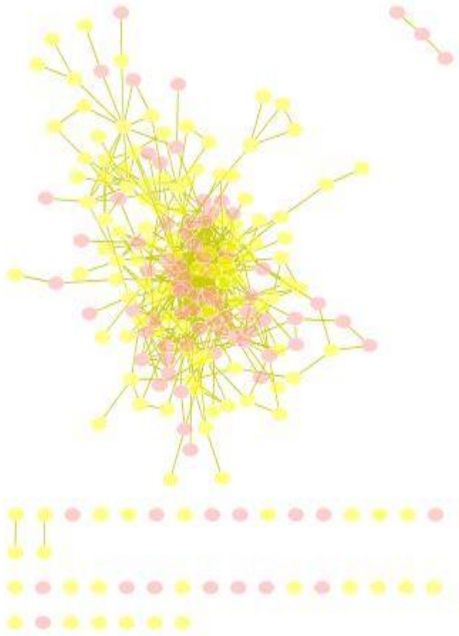
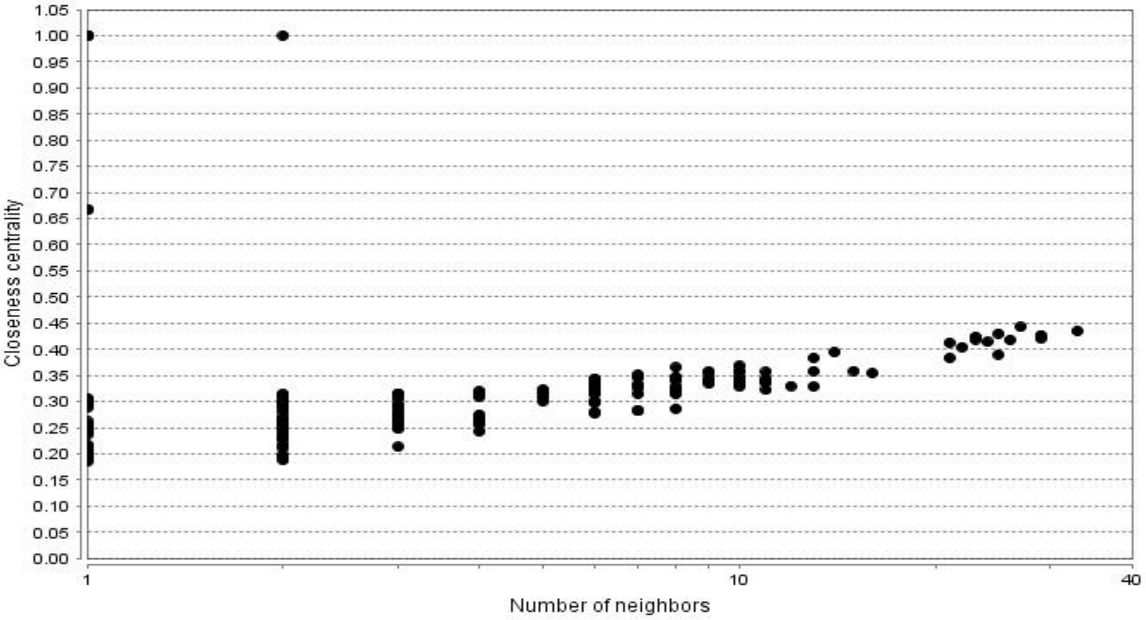
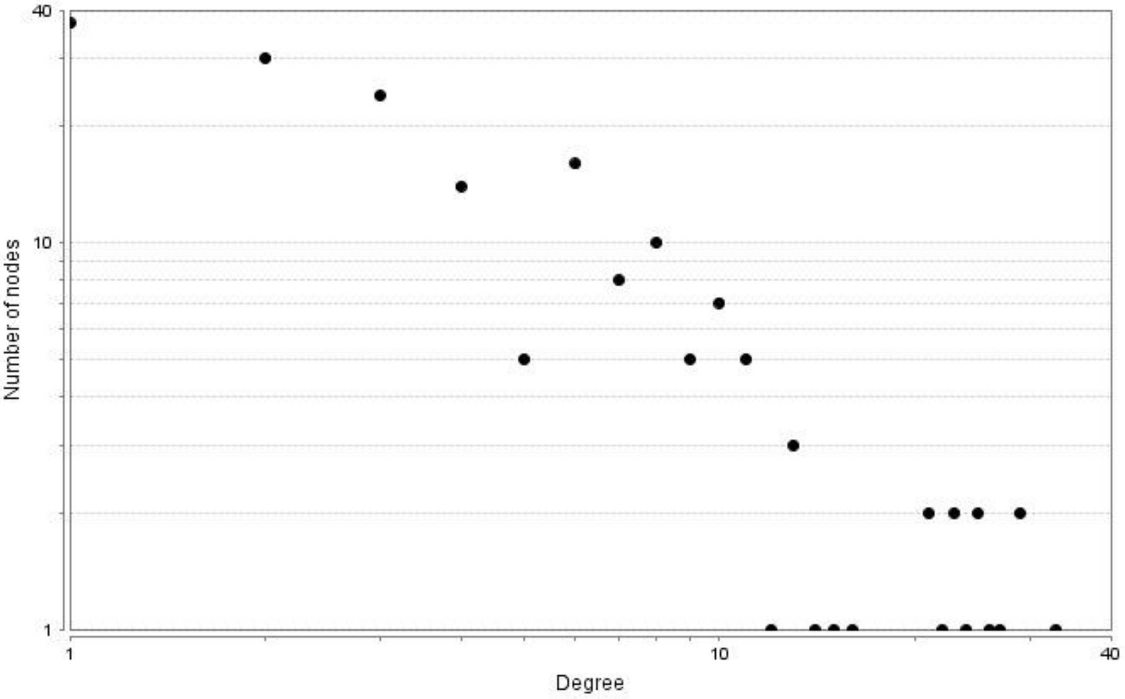


Figure 10 Yellow nodes- T2D, Pink nodes- Scleroderma

The network was analyzed by NetworkAnalyzer tool of cytoscape. The network had 218 nodes, 538 edges with a clustering coefficient of 0.236. The following network properties of degree distribution, average neighbourhood connectivity, average aggregation coefficient and distribution of shortest path were also analysed as seen in Figure.





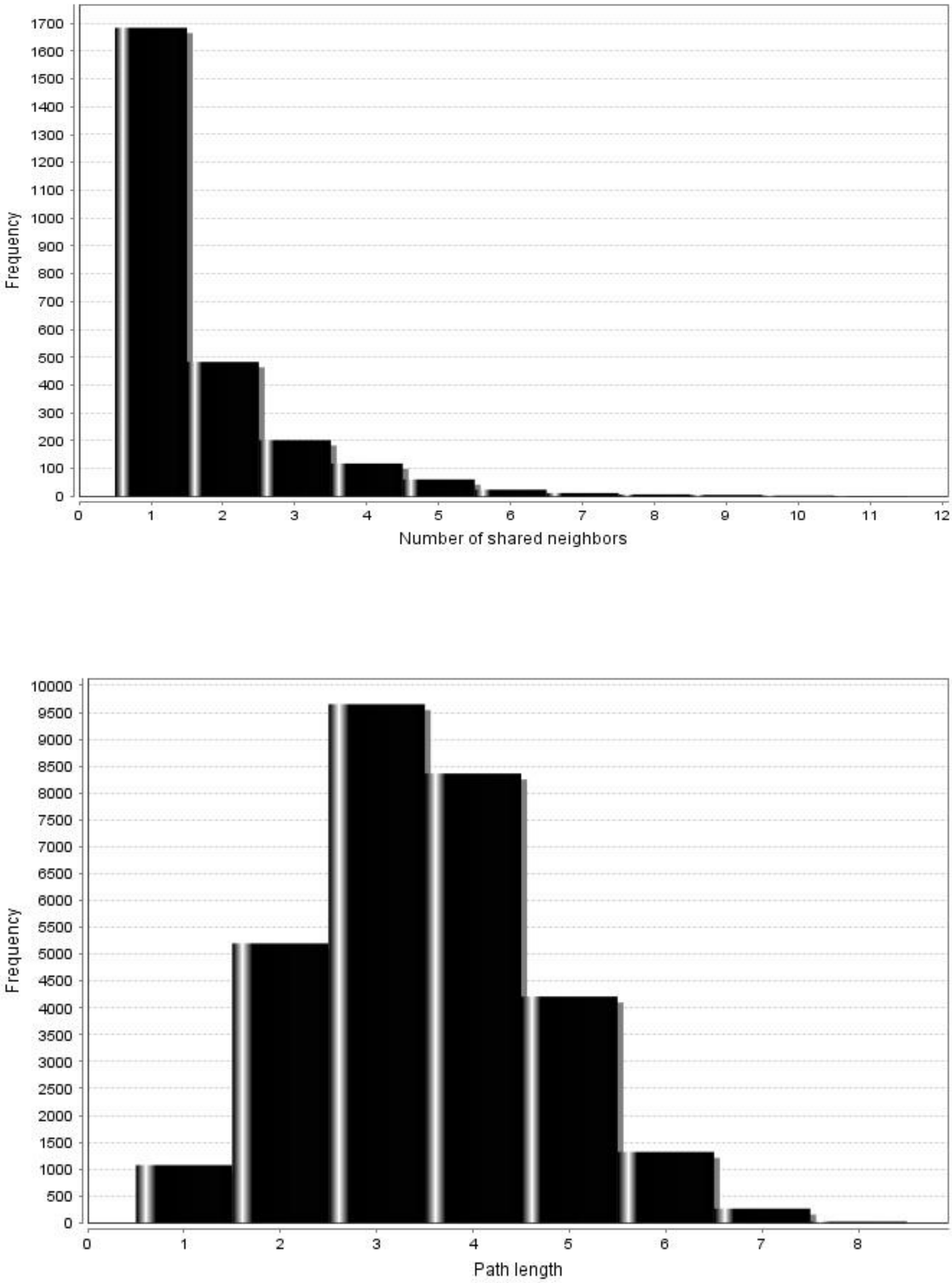
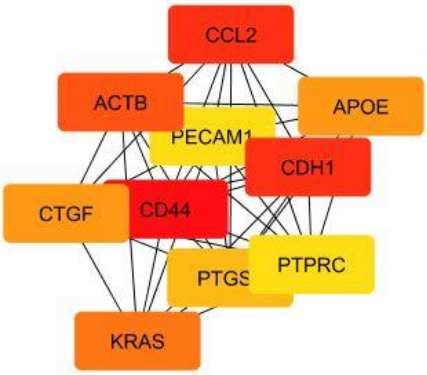


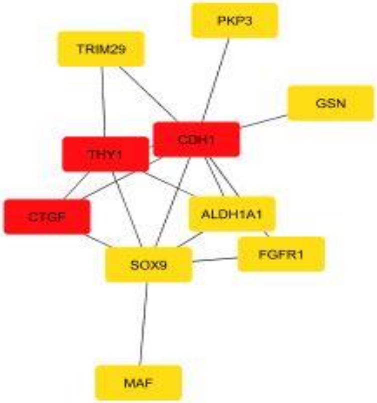
Figure 11 PPI network analysis of DEGs by NetworkAnalyser tool

5.3.2.2. Identification of hub genes through PPI network analysis DEGs

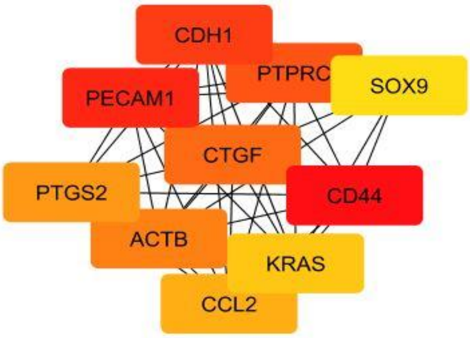
The cytoHubba application which uses 11 algorithms to give hub genes was employed to identify the top 20 genes evaluated using the following 5 algorithms: Degree, MCC, MNC, EPC and EcCentricity as shown in the figure.



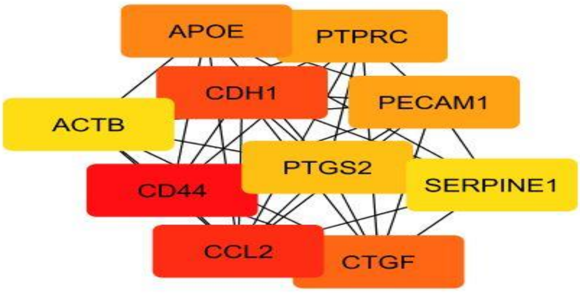
a) Degree



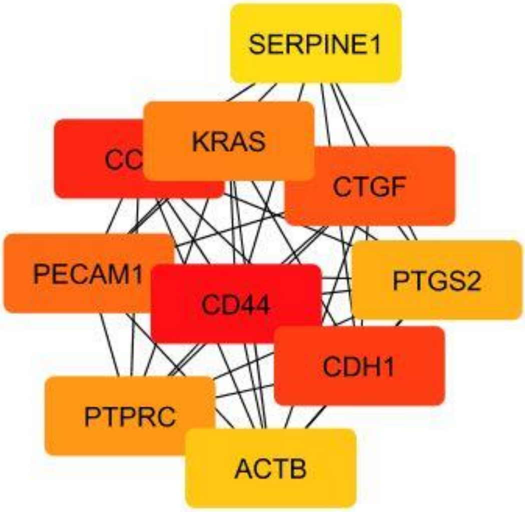
b) Eccentricity



c) MNC



d) MCC



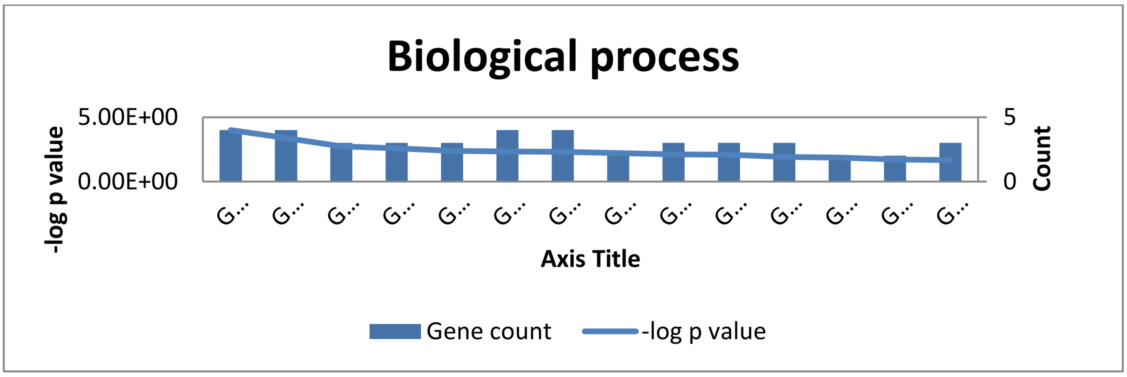
e) EPC

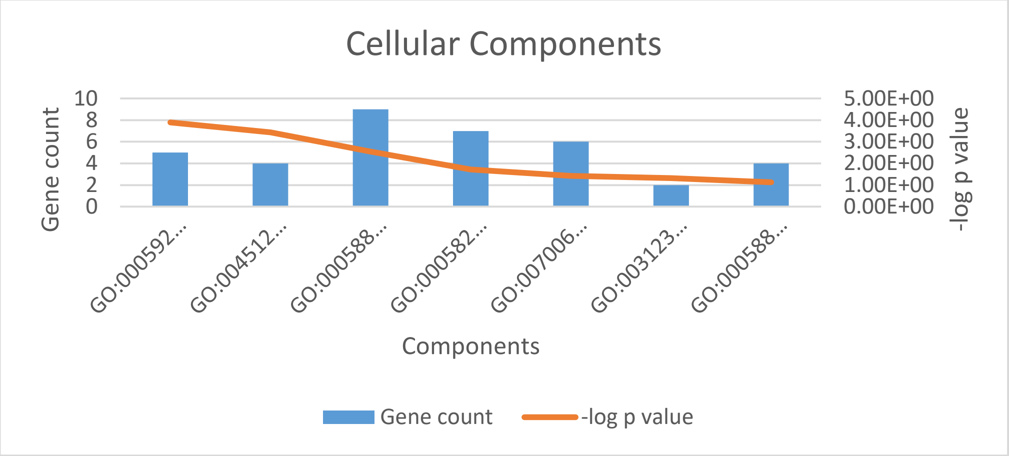
Table 7 Top 10 genes evaluated in PPI network using cytoHubba

Gene	Degree	Gene	EcCentricity	Gene	EPC	Gene	MNC	Gene	MCC
CD44	33	CDH1	0.24033	CD44	44.332	CD44	31	CD44	37168
CCL2	29	CTGF	0.24033	CCL2	43.292	CCL2	29	PECAM1	34310
CDH1	29	THY1	0.24033	CDH1	43.156	CDH1	27	CDH1	31782
ACTB	27	CD44	0.19227	CTGF	42.664	CTGF	25	PTPRC	30666
KRAS	26	CCL2	0.19227	PECAM1	42.201	APOE	24	CTGF	28486
CTGF	25	ACTB	0.19227	KRAS	42.158	PECAM1	23	ACTB	28284
APOE	25	KRAS	0.19227	PTPRC	41.788	PTPRC	23	PTGS2	20732
PTGS2	24	APOE	0.19227	PTGS2	41.468	PTGS2	22	CCL2	20228
PTPRC	23	PTGS2	0.19227	ACTB	41.278	ACTB	21	KRAS	12437
PECAM1	23	PTPRC	0.19227	SERPINE1	40.963	KRAS	21	SOX9	10728

Then, the hub genes were identified by taking the intersection of these results. The 14 genes include THY1 ACTB CD44 LYN ZEB1 AGTR1 CDH1 KRAS CCL19 PECAM1 CCND2 FOS CYR61 FGFR1.

5.3.2.3 Pathway analysis





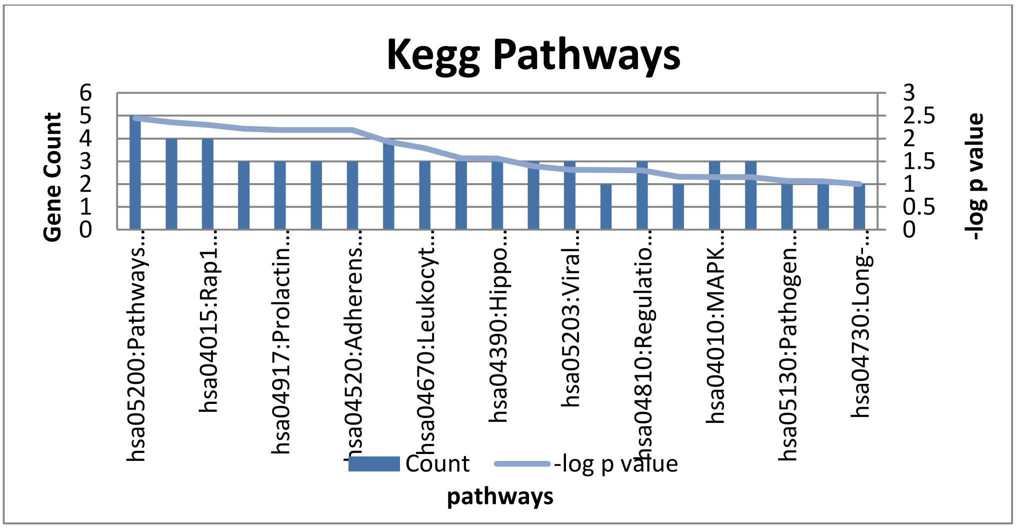


Figure 12 GO and Pathway enricment analysis results

5.3.3 Scleroderma and osteoarthritis

A protein protein interaction network was build by stringApp of cytoscape and the interaction can be seen in the figure.

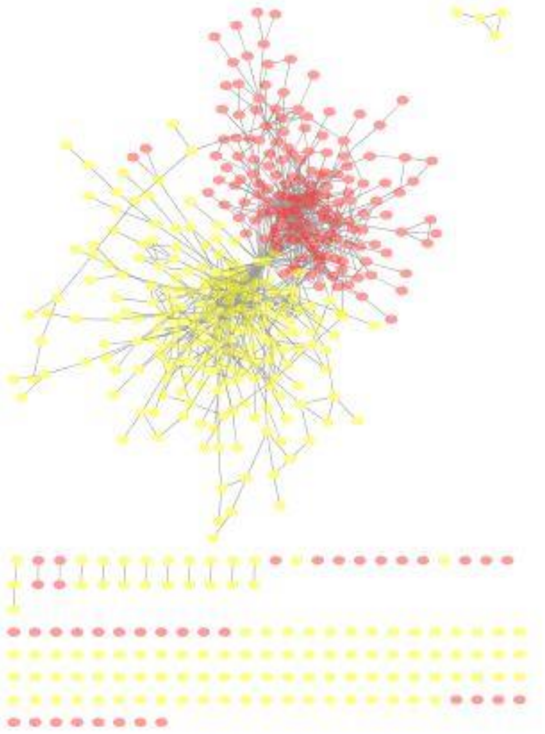
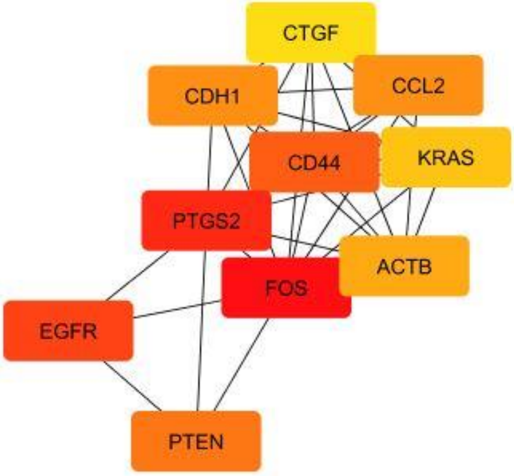


Figure 13 Yellow nodes- Osteoarthritis, Pink nodes- Scleroderma

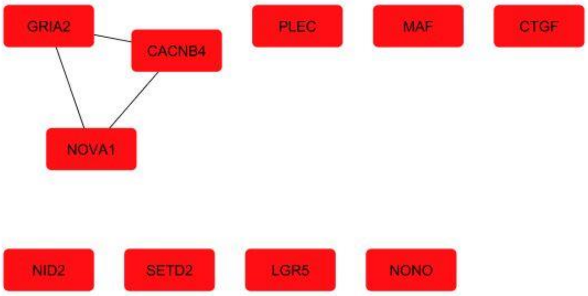
The network was analyzed by NetworkAnalyzer tool of cytoscape. The network had nodes 529, 1038 edges with a clustering coefficient of 0.190. The following network properties of degree distribution, average neighbourhood connectivity, average aggregation coefficient and distribution of shortest path were also analysed.

5.3.3.2. Identification of hub genes through PPI network analysis DEGs

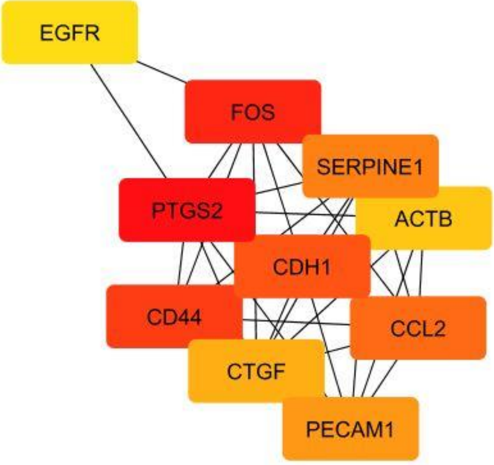
The cytoHubba application which uses 11 algorithms to give hub genes was employed to identify the top 20 genes evaluated using the following 5 algorithms: Degree, MCC, MNC, EPC and EcCentricity as shown in the figure.



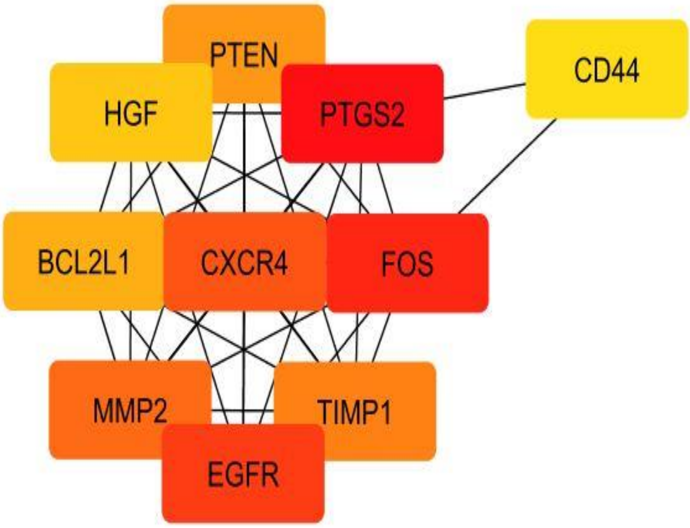
a) Degree



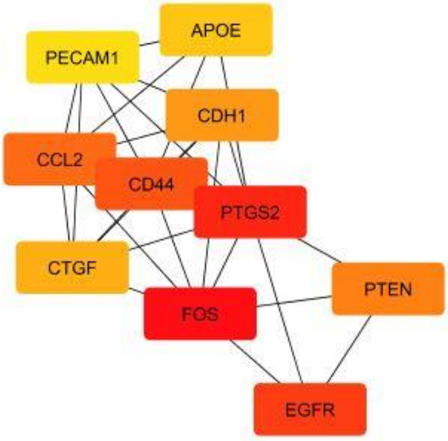
b) EcCentricity



c) EPC



d) MCC



e) MNC

Table 8 Top 10 genes evaluated in the PPI network using cytoHubba

Gene	Degree	Gene	EcCentricity	Gene	EPC	Gene	MCC	Gene	MNC
FOS	44	CD44	0.11614	PTGS2	66.26	PTGS2	69279	FOS	42
PTGS2	41	CDH1	0.11614	FOS	65.563	FOS	54908	PTGS2	38
EGFR	40	ACTB	0.11614	CD44	62.753	EGFR	54078	EGFR	36
CD44	33	KRAS	0.11614	CDH1	61.126	CXCR4	52812	CD44	31
PTEN	30	CTGF	0.11614	CCL2	61.025	MMP2	52714	CCL2	29

CDH1	29	PTPRC	0.11614	SERPINE1	60.421	TIMP1	51938	PTEN	28
CCL2	29	PECAM1	0.11614	PECAM1	59.904	PTEN	47454	CDH1	27
ACTB	27	SOX9	0.11614	CTGF	59.186	BCL2L1	46941	CTGF	25
KRAS	26	THY1	0.11614	ACTB	59.055	HGF	46467	APOE	24
APOE	25	POLR2E	0.11614	EGFR	58.97	CD44	37168	PECAM1	23

Then, the hub genes were identified by taking the intersection of these results. These genes include THY1 ACTB CD44 CDH1 KRAS PECAM1 SOX9 PTPRC CTGF.

5.3.3.3 Pathway analysis

Biological Pathway	Count	PValue	Genes
GO:0007155~cell adhesion	6	7.81E-07	VWF, CD44, PECAM1, CXCL12, CDH5, SPP1
GO:0030198~extracellular matrix organization	5	1.21E-06	VWF, CD44, PECAM1, CDH1, SPP1
GO:0050900~leukocyte migration	4	2.04E-05	CD44, PECAM1, TEK, ANGPT1
GO:0022617~extracellular matrix disassembly	3	5.56E-04	CD44, CDH1, SPP1
GO:0016337~single organismal cell-cell adhesion	3	9.80E-04	CD44, CDH1, CDH5
GO:0048014~Tie signaling pathway	2	0.001428656	TEK, ANGPT1
GO:0072012~glomerulus vasculature development	2	0.001904478	TEK, ANGPT1
GO:0070374~positive regulation of ERK1 and ERK2 cascade	3	0.002901631	CD44, TEK, ANGPT1
GO:0001525~angiogenesis	3	0.004663711	PECAM1, TEK, ANGPT1

Cellular Component	Count	PValue	Genes
GO:0005576~extracellular region	6	2.39E-04	VWF, TEK, ANGPT1, CDH1, CXCL12, SPP1
GO:0070062~extracellular exosome	7	2.83E-04	VWF, CD44, PECAM1, ANGPT1, CDH1, CXCL12, SPP1
GO:0045121~membrane raft	3	0.003404384	PECAM1, TEK, ANGPT1

GO:0009897~external side of plasma membrane	3	0.003634659	PECAM1, CXCL12, CDH5
GO:0005925~focal adhesion	3	0.011801874	CD44, TEK, CDH1
GO:0030054~cell junction	3	0.016027604	PECAM1, CDH1, CDH5
GO:0005615~extracellular space	4	0.01701686	PECAM1, ANGPT1, CXCL12, SPP1
GO:0005886~plasma membrane	6	0.017759255	CD44, PECAM1, TEK, ANGPT1, CDH1, CDH5
Molecular Function	Count	PValue	Genes
GO:0005515~protein binding	8	0.010325405	ACTB, PTPRC, KRAS, CD44, CTGF, PECAM1, CDH1, SOX9
GO:0008013~beta-catenin binding	2	0.033517079	CDH1, SOX9

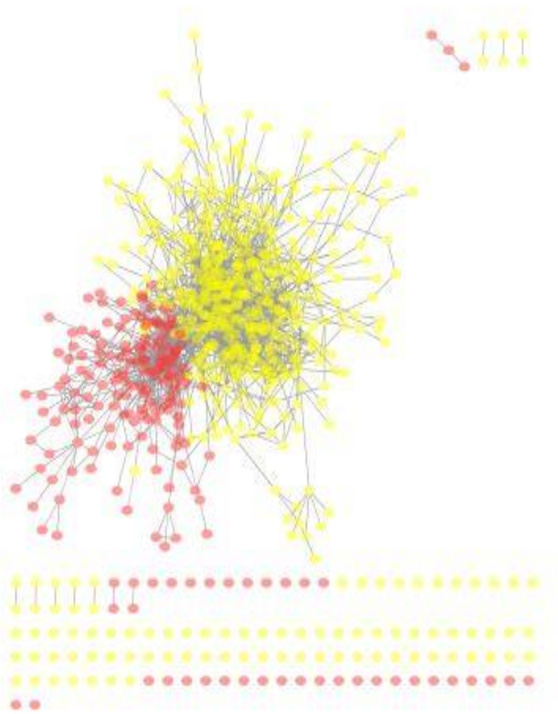
Table 9 GO and Pathway enrichment results

Pathways	Count	PValue	Genes
hsa04514:Cell adhesion molecules (CAMs)	3	0.008301628	PTPRC, PECAM1, CDH1
hsa04390:Hippo signaling pathway	3	0.009350095	ACTB, CTGF, CDH1
hsa05205:Proteoglycans in cancer	3	0.016040915	ACTB, KRAS, CD44
hsa04015:Rap1 signaling pathway	3	0.017602972	ACTB, KRAS, CDH1
hsa05216:Thyroid cancer	2	0.02915205	KRAS, CDH1
hsa05219:Bladder cancer	2	0.041000116	KRAS, CDH1
hsa05130:Pathogenic Escherichia coli infection	2	0.050778619	ACTB, CDH1
hsa05213:Endometrial cancer	2	0.051751752	KRAS, CDH1
hsa05131:Shigellosis	2	0.063362813	ACTB, CD44
hsa04520:Adherens junction	2	0.070079487	ACTB, CDH1
hsa05218:Melanoma	2	0.070079487	KRAS, CDH1
hsa05100:Bacterial invasion of epithelial cells	2	0.076754833	ACTB, CDH1
hsa04660:T cell receptor signaling pathway	2	0.097467808	PTPRC, KRAS

5.3.4 Scleroderma and lung cancer

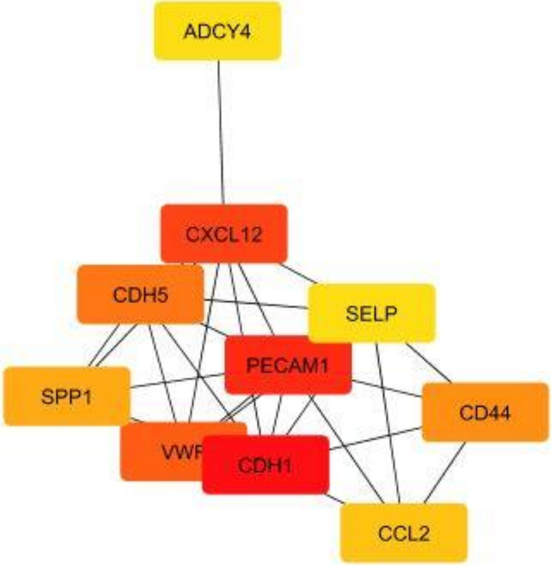
A protein protein interaction network was build by stringApp of cytoscape. The network was analyzed by NetworkAnalyzer tool of cytoscape. The network had 709 nodes, 1913 edges with a clustering coefficient of 0.210. The following network properties of degree distribution, average neighbourhood connectivity, average aggregation coefficient and distribution of shortest path were also analysed.

Table 10 Yellow nodes- Lung Cancer, Pink nodes- Scleroderma

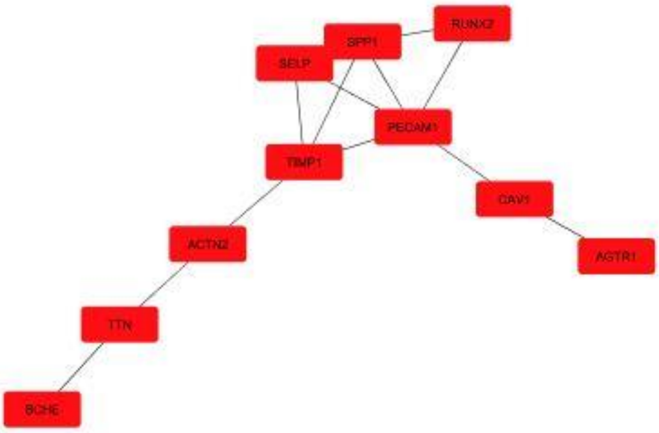


5.3.4.1. Identification of hub genes through PPI network analysis DEGs

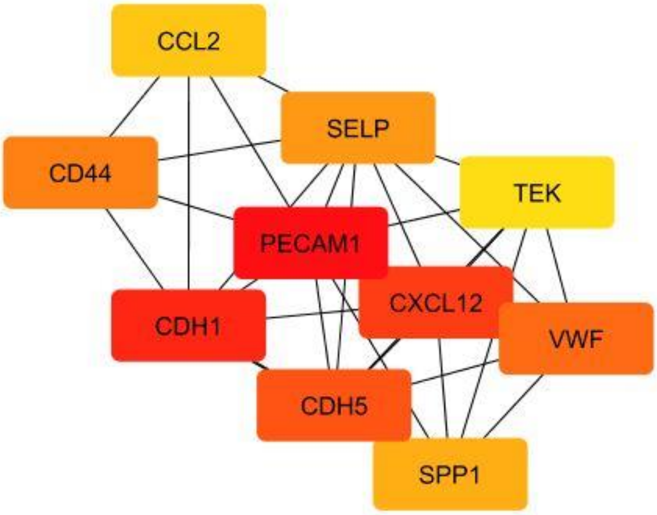
The cytoHubba application which uses 11 algorithms to give hub genes was employed to identify the top 20 genes evaluated using the following 5 algorithms: Degree, MCC, MNC, EPC and EcCentricity as shown in the figure.



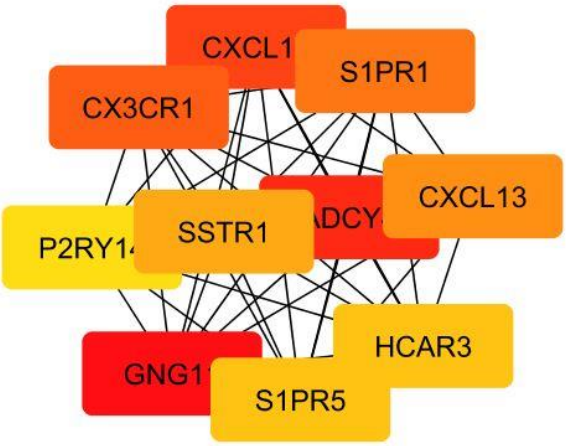
a) Degree



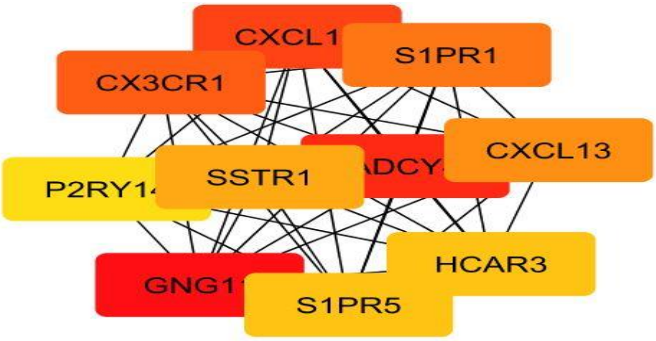
b) EcCentricity



c) EPC



d) MCC



e) MNC

Table 11 Top 10 genes evaluated by cytoHubba

Gene	Degree	Gene	EcCentricity	Gene	EPC	Gene	MCC	Gene	MNC

CDH1	70	CDH1	0.1603	PECAM1	60.107	GNG11	3991772	CDH1	62
PECAM1	61	PECAM1	0.1603	CDH1	57.787	ADCY4	3991763	PECAM1	60
CXCL12	42	SPP1	0.1603	CXCL12	53.348	CXCL12	3673812	CXCL12	40
VWF	41	SELP	0.1603	CDH5	51.518	CX3CR1	3628842	CDH5	36
CDH5	38	CAV1	0.1603	VWF	49.413	S1PR1	3628834	VWF	35
CD44	33	RUNX2	0.1603	CD44	47.161	CXCL13	3628813	CD44	31
SPP1	30	AGTR1	0.1603	SELP	44.848	SSTR1	3628802	CCL2	29
CCL2	29	TIMP1	0.1603	SPP1	43.392	S1PR5	3628801	SPP1	27
SELP	28	ACTN2	0.1603	CCL2	43.172	HCAR3	3628801	SELP	26
ADCY4	28	TTN	0.1603	TEK	43.161	P2RY14	3628800	GNG11	25

Then, the hub genes were identified by taking the intersection of these results. These genes include CD44 ANGPT1 CDH1 PECAM1 VWF CDH5 TEK SPP1 CXCL12.

5.3.4.2. Pathway analysis of resultant genes

Biological process	Count	PValue	Genes
GO:0007155~cell adhesion	6	7.81E-07	VWF, CD44, PECAM1, CXCL12, CDH5, SPP1
GO:0030198~extracellular matrix organization	5	1.21E-06	VWF, CD44, PECAM1, CDH1, SPP1
GO:0050900~leukocyte migration	4	2.04E-05	CD44, PECAM1, TEK, ANGPT1
GO:0022617~extracellular matrix disassembly	3	5.56E-04	CD44, CDH1, SPP1
GO:0016337~single organismal cell-cell adhesion	3	9.80E-04	CD44, CDH1, CDH5
GO:0048014~Tie signaling pathway	2	0.001428656	TEK, ANGPT1
GO:0072012~glomerulus vasculature development	2	0.001904478	TEK, ANGPT1
GO:0070374~positive regulation of ERK1 and ERK2 cascade	3	0.002901631	CD44, TEK, ANGPT1
GO:0001525~angiogenesis	3	0.004663711	PECAM1, TEK, ANGPT1

GO:0031589~cell-substrate adhesion	2	0.008072132	VWF, ANGPT1

Cellular Components	Count	PValue	Genes
GO:0005576~extracellular region	6	2.39E-04	VWF, TEK, ANGPT1, CDH1, CXCL12, SPP1
GO:0070062~extracellular exosome	7	2.83E-04	VWF, CD44, PECAM1, ANGPT1, CDH1, CXCL12, SPP1
GO:0045121~membrane raft	3	0.003404384	PECAM1, TEK, ANGPT1
GO:0009897~external side of plasma membrane	3	0.003634659	PECAM1, CXCL12, CDH5
GO:0005925~focal adhesion	3	0.011801874	CD44, TEK, CDH1
GO:0030054~cell junction	3	0.016027604	PECAM1, CDH1, CDH5
GO:0005615~extracellular space	4	0.01701686	PECAM1, ANGPT1, CXCL12, SPP1
GO:0005886~plasma membrane	6	0.017759255	CD44, PECAM1, TEK, ANGPT1, CDH1, CDH5
GO:0009986~cell surface	3	0.02194939	CD44, TEK, CDH5
GO:0005902~microvillus	2	0.024754441	TEK, ANGPT1
GO:0005911~cell-cell junction	2	0.073070753	TEK, CDH5

Molecular Functions	Count	PValue	Genes
GO:0005518~collagen binding	2	0.028088866	VWF, CD44
GO:0001948~glycoprotein binding	2	0.030398128	VWF, CDH1
GO:0008013~beta-catenin binding	2	0.038213749	CDH1, CDH5

Pathways	Count	PValue	Genes
hsa04512:ECM-receptor interaction	3	0.004213842	VWF, CD44, SPP1
hsa05323:Rheumatoid arthritis	3	0.004309322	TEK, ANGPT1, CXCL12
hsa04151:PI3K-Akt signaling pathway	4	0.005798812	VWF, TEK, ANGPT1, SPP1
hsa04670:Leukocyte transendothelial migration	3	0.007263795	PECAM1, CXCL12, CDH5

hsa04514:Cell adhesion molecules (CAMs)	3	0.010919832	PECAM1, CDH1, CDH5
hsa04015:Rap1 signaling pathway	3	0.023003258	TEK, ANGPT1, CDH1

Table 12 GO and Pathway enrichment analysis

CHAPTER 6: DISCUSSION

Scleroderma, particularly systemic scleroderma is a chronic connective tissue disease generally classified as one of the autoimmune rheumatic diseases. The disease has been called "progressive systemic sclerosis", though it varies from person to person. disease include crystalline silica, chlorine solvents, welding vapors, and various other solvents. Clustering within families indicates a role for genetic factors. Although concordance for the disease among identical twins is low, concordance for autoantibodies associated with systemic sclerosis and for fibroblast gene expression profiles is higher. Because multiplex families are rare, association and candidate gene strategies are the most appropriate methods for investigating the genetics of systemic sclerosis. The most consistent data relate to the involvement of fibrosis genes, most notably the TGF-β regulation pathway, secreted protein acid and rich in cysteine (SPARC) genes, and the fibrillin-1 gene (FBN1). In this study, one dataset was identified for scleroderma comparing the differences in mRNA expression in normal skin samples and disease samples. Eventually, a total of 235 DEGs were screened including 145 upregulated and 91 downregulated genes.functional enrichment analysis showed positive regulation of apoptotic process, collagen catabolic process, signal transduction, immune response, inflammatory response in the BP group.

In CC group, the DEGs were mostly enriched in plasma membrane, cytosol, extracellular exosome, internal component of plasma membrane, golgi apparatus. Finally, in the MF group, the genes enriched in protein binding, calcium ion binding, heparin binding, protein kinase activity, cadherin binding involved in cell-cell adhesion, extracellular matrix structural constituent, carbohydrate binding. The pathways that were enriched included PI3K-Akt signaling pathway, Focal adhesion, Oxytocin signaling pathway, Hypertrophic cardiomyopathy (HCM). When diseases were enriched, 164 results were obtained and after checking with the published literature and depending upon the p value four diseases were selected to obtain scleroderma diseasome. These diseases are often observed to occur with scleroderma.

It is evident that persistent overproduction of collagen is responsible for the progressive nature of tissue fibrosis in SSc. Up-regulation of collagen gene expression in SSc fibroblasts appears to be a critical event in this process. The coordinate transcriptional activation of numerous collagen genes suggests a fundamental alteration in the regulatory control of gene expression in SSc fibroblasts. Trans-acting nuclear factors which bind to cis-acting elements in enhancer (intronic) and promoter regions of the genes modulate the basal and inducible transcriptional activity of the collagen genes. The identification of the nuclear transcription factors that regulate normal collagen gene expression may provide promising approaches to the therapy of this incurable disease. Also, it has been found that downregulation of EB13 gene contributes to Type I Collagen Overexpression in Scleroderma Skin.

The construction of diseasome helped in gaining the insight about how various diseases are connected at molecular level through association between genes, proteins and pathways. The interactome was explored to better understand these relationships. Protein protein interactions were mapped and using various tools the resulting networks were analysed. Cytohubba plugin of cytoscape tool was used to find out the hub genes from the networks that are involved in various diseases and the effect of genes on different pathways and process was observed.

CHAPTER 7: CONCLUSION

Scleroderma is a rare autoimmune disease which may lead to major disabilities due to vascular complications, cardiopulmonary involvement, inflammatory myopathy, and arthritis; likewise, it can cause malnutrition due to gastrointestinal tract involvement, and it can decrease quality of life as a consequence of the psychological and social impact. It can be fatal, with a 3-year survival rate of 47-56% in cases of serious pulmonary or cardiac involvement, it is the single connective tissue disease with the worst survival prognosis. Though numerous scientific communities contribute to scleroderma research around the world, the etiology of the disease still remains unidentified.

Scleroderma shares common immunological features with many complex disorders such as cardiovascular disease, diabetes, obesity, depression and inflammatory arthritis. However, the patho-mechanism connecting these systemic comorbidities with scleroderma remains to be determined. A systematic exploration of the shared component hypothesis existing between the co-morbidities can shed light about the molecular connections aiding the prevention, early diagnosis and treatment of scleroderma. In this study, the interactomes were explored to identify the biological processes and pathways linking scleroderma with its comorbidities and identification of hug genes using a network approach along with their functional analysis, PPI network construction, module selection and module enrichment analyses.

Further research is needed to shed light on multimeric novel targets and pathways which can be targeted to offer diagnosis and/or cure for scleroderma along with its associated co-morbidities.

REFERENCES

[1] Allanore Y, Simms R, Distler O, et al. Systemic sclerosis. *Nat Rev Dis Primers* 2015;1:15002.

[2] Ferri C, Sebastiani M, Lo Monaco A, et al. Systemic sclerosis evolution of disease pathomorphosis and survival. Our experience on Italian patients' population and review of the literature. *Autoimmun Rev* 2014;13:1026–34.

[3] Maurer B, Graf N, Michel BA, et al. Prediction of worsening of skin fibrosis in patients with diffuse cutaneous systemic sclerosis using the EUSTAR database. *Ann Rheum Dis* 2015;74:1124–31.

[4] Bossini-Castillo L, López-Isac E, Martín J. Immunogenetics of systemic sclerosis: defining heritability, functional variants and shared-autoimmunity pathways. *J Autoimmun* 2015;64:53–65.

[5] HUDSON, M., SHARMA, A., BERNSTEIN, J., & BARON, M. (2009). Validity of Self-Reported Comorbidities in Systemic Sclerosis. *The Journal of Rheumatology, 36(7), 1477–1480.*doi:10.3899/jrheum.081134

[6] Gupta L, Phatak S, Edavalath S. Biomarkers in scleroderma: Current status. *Indian J Rheumatol* 2017;12, Suppl S1:149-55.

[7] Barnes J, Mayes MD. Epidemiology of systemic sclerosis: Incidence, prevalence, survival, risk factors, malignancy, and environmental triggers. *Curr Opin Rheumatol* 2012;24:165-70

[8] Iudici M, Codullo V, Giuggioli D, et al. Pulmonary hypertension in systemic sclerosis: prevalence, incidence and predictive factors in a large multicentric Italian cohort. *Clin Exp Rheumatol* 2013;31:31–6.

[9] Galluccio F, Walker UA, Nihtyanova S, et al. Registries in systemic sclerosis: a worldwide experience. *Rheumatology (Oxford)* 2011;50:60–8.

[10] Nikpour M, Hissaria P, Byron J, et al. Prevalence, correlates and clinical usefulness of antibodies to RNA polymerase III in systemic sclerosis: a cross-sectional analysis of data from an Australian cohort. *Arthritis Res Ther* 2011;13:R211.

[11] Harris E, Budd RC, Firestein GS, et al.Seibold J. Scleroderma. In: Harris E, Budd RC, Firestein GS, et al., editors. *Kelley's textbook of rheumatology*. 7th ed. Philadelphia: Elsevier; 2005.

[12] Frantz C, Avouac J, Distler O, et al. Impaired quality of life in 24. systemic sclerosis and patient perception of the disease: a large international survey. *Semin Arthritis Rheum* 2016.

[13] Merkel PA, Silliman NP, Clements PJ, et al. Patterns and predictors of change in outcome measures in clinical trials in scleroderma: an individual patient meta-analysis of 629 subjects with diffuse cutaneous systemic sclerosis. *Arthritis Rheum* 2012;64:3420–9.

[14] Ferri C, Sebastiani M, Lo Monaco A, et al. Systemic sclerosis evolution of disease pathomorphosis and survival. Our experience on Italian patients' population and review of the literature. *Autoimmun Rev* 2014;13:1026–34.

[15] de Groot V, Beckerman H, Lankhorst GJ, Bouter LM. How to measure comorbidity. A critical review of available methods. *J Clin Epidemiol* 2003;56:221–9.

[16] Sangha O, Stucki G, Liang MH, Fossel AH, Katz JN. The Self-Administered Comorbidity Questionnaire: a new method to assess comorbidity for clinical and health services research. *Arthritis Rheum* 2003;49:156–63.

[17] Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol* 1992;45:613–9.

[18] Wilson L. Cost-of-illness of scleroderma: the case for rare diseases. *Semin Arthritis Rheum 1997*;27:73–84.

[19] Radstake TR, Gorlova O, Rueda B, Martin JE, Alizadeh BZ, Palomino-Morales R, *et al.* Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. *Nat Genet* 2010;42:426-9.

[20] Clements P, Lachenbruch P, Siebold J, White B, Weiner S, Martin R, et al. Inter and intraobserver variability of total skin thickness score (modified rodnan TSS) in systemic sclerosis. *J Rheumatol* 1995;22:1281-5.

[21] Milano A, Pendergrass SA, Sargent JL, George LK, McCalmont TH, Connolly MK, *et al.* Molecular subsets in the gene expression signatures of scleroderma skin. *PLoS One* 2008;3:e2696.

[22] Brkic Z, van Bon L, Cossu M, van Helden-Meeuwsen CG, Vonk MC, Knaapen H, *et al.* The interferon type I signature is present in systemic sclerosis before overt fibrosis and might contribute to its pathogenesis through high BAFF gene expression and high collagen synthesis. *Ann Rheum Dis* 2016;75:1567-73.

[23] Farina G, Lafyatis D, Lemaire R, Lafyatis R. A four-gene biomarker predicts skin disease in patients with diffuse cutaneous systemic sclerosis. *Arthritis Rheum* 2010;62:580-8.