ANALYSIS OF TYPE-II DIABETES (T2D) MICROARRAY DATA

A PROJECT REPORT SUBMITTED IN THE PARTIAL FULFILLMENT FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE IN BIOINFORMATICS

SUBMITTED BY:

CHINMAYI KUMARI SAHU
REGD.NO. 07 / BI / 05

DEPARTMENT OF BIOINFORMATICS,

CENTRE FOR POST GRADUATE STUDIES

ORISSA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY,
BHUBANESWAR, ORISSA

DEDICATED TO MY PARENTS

CERTIFICATE - I

This is to certify that the project report entitled: "**ANALYSIS OF TYPE 2 DIABETES (T2D) MICROARRAY DATA**" submitted to Centre for Post Graduate Studies, Orissa University of Agriculture & Technology, Bhubaneswar, Orissa for the partial fulfillment of the requirements for the award of **Master of Science in Bioinformatics**, is a record of original work done by **Miss Chinmayi Kumari Sahu** during her study in Centre for Post Graduate Studies, Orissa University of Agriculture & Technology, Bhubaneswar, Orissa under my supervision in Institute of Bioinformatics and Applied Biotechnology (IBAB). This project has not formed the basis for the award of any Degree/Diploma/Associate ship/Fellowship or other similar title to any candidate in any University.

SIGNATURE
Dr. Narayan Behera
Project Guide
Faculty Scientist
IBAB, Bangalore

CERTIFICATE- II

This is to certify that the project report entitled: **"ANALYSIS OF TYPE 2 DIABETES (T2D) MICROARRAY DATA"** is submitted by Miss Chinmayi Kumari Sahu, Reg. no. 07/BI/05 to Department of Bioinformatics, Centre for Post Graduate Studies, Orissa University of Agriculture & Technology, Bhubaneswar, Orissa for the partial fulfillment of the requirements for the award of Master of Science in Bioinformatics, is a bona fide record of work carried by her. The project has not formed the basis for the award of any Degree/Diploma/Associate ship/Fellowship or other similar title to any candidate in any University.

PROJECT GUIDE
Dr. Narayan Behera
IBAB, Bangalore **SIGNATURE**

Submitted for the viva-voce examination held on _____ at Centre for Post Graduate Studies, Orissa University of Agriculture & Technology, Bhubaneswar-3, Orissa.

HEAD OF DEPARTMENT
Dr. P. N. Jagdev,
Dept. of Bioinformatics,
OUAT, BBSR **SIGNATURE**

EXTERNAL EXAMINER

SIGNATURE

INTERNAL ADVISOR

SIGNATURE

DECLA RATION

I, **Chinmayi Kumari Sahu, Reg. no. 07/BI/05** affirm that my project report entitled: "**ANALYSIS OF TYPE 2 DIABETES (T2D) MICROARRAY DATA**" is based on the original work carried out by me under the guidance of **Dr. Narayan Behera**, Faculty Scientist, Institute of Bioinformatics and Applied Biotechnology (IBAB), Bangalore in the partial fulfillment of requirement for the award of **Master of Science in Bioinformatics**, Centre for Post Graduate Studies, Orissa University of Agriculture & Technology, Bhubaneswar, Orissa. The project has not formed the basis for the award of any Degree/Diploma/Associate ship/Fellowship or other similar title to any candidate in any University.

Place: Bhubaneswar
Date:

Signature of the Candidate
(CHINMAYI KUMARI SAHU)

ACKNOWLEDGEMENT

It gives me immense pleasure to place on record my heartfelt thanks and acknowledgement to **Dr. Narayan Behera, Faculty Scientist, Institute of Bioinformatics and Applied Biotechnology (IBAB), Bangalore** for being my mentor in successful completion of my project. It is a great fortune and pride, to get an opportunity to work under his potential guidance.

I sincerely acknowledge and put in record the guidance and support given by **Dr. P.N. Jagdev**, Head of the department, Department of Bioinformatics, Orissa University of Agriculture and Technology, Bhubaneswar. Also I sincerely thank and acknowledge all my professors and academicians who have helped to make my project a great success.

It is an honour to show my gratitude to **Mr. Abdullah Khan**, Systems Administrator, IBAB for his kind assistance for successful completion of project.

I wish to express my whole hearted thanks to my parents and relatives whose inspiration helped me to come to this position. Last but not the least I would like to thank my friends for their timely help and moral encouragement. Finally, I would like to thank all whose direct and indirect support helped me completing my project in time.

CONTENTS

Subject **Page no.**

I. Supplementary Data ……
List of Tables ……
List of Chart ……..
List of Pathways ……..
List of Figure ……

II. Abstract ……..…..

1. Introduction ……..

2. Review of literature ……..

2.1 Insulin ……..
2.1.1 Regulation of insulin ……....
2.2 Prediabetes ……..
2.3 Diabetes ……
2.3.1 Brief overview ……....
2.3.2 Understanding diabetes ……...
2.3.3 Types of diabetes ……..
2.3.4 Brief study on Type 2 diabetes ……
2.3.4.1. General idea ……..
2.3.4.2. Prevalence and related risk factor …….
2.3.4.3. Stages of development Type 2 Diabetes ……...
2.3.4.4. Factors predisposing to Type 2 Diabetes ……...
2.3.4.5. Symptoms of Type 2 diabetes ……
2.3.4.6. Diagnosis of Type 2 diabetes …….
2.3.4.7. Risk factors for diabetes …….
2.3.4.8. Complications of diabetes ……..
2.3.4.9. Prevention or delay of onset of type 2 diabetes …..
2.3.4.10. Treatment for type 2 diabetes ……

2.4 Expression profiling ……..
2.4.1 Microarray ……..
2.4.1.1 Introduction ……
2.4.1.2 Technique ……..
2.4.1.3 Applications of microarray ……...

2.5. Analysis of Expression data ……
2.5.1 Slide scanning ……
2.5.2 Image processing and numerical data collection …….

2.5.3 Normalization ……..

2.5.3.1 Data cleaning and transformation …….

2.5.3.2 Within array normalization ……

2.5.3.3 Between array normalization …….

2.5.4. Statistical analysis ……

2.5.4.1 Types of statistical method ……

2.5.4.2 Steps of statistical testing ……..

2.5.4.3 Multiple testing …….

2.5.4.4 Empirical *p*-Value …….

2.5.4.5 Statistical conclusion on microarray data ……..

2.5.5. Collection of differentially expressed gene ……

2.5.6. Clustering the data ……..

2.5.6.1 Distance and similarity measure ……...

2.5.6.2 Types of clustering ……...

2.5.6.3 Application of clustering ……...

2.5.7. Pathway analysis ……..

2.5.7.1. Pathways and genes related to Type 2 Diabetes ……

2.5.8. Interaction network ……..

2.5.9. Promoter analysis …….

3. Materials and Methods …….....

3.1 Materials

3.1.1 Datasheet Used

3.1.2 Softwares Used

3.1.3 Algorithms Used

3.2 Methods

3.2.1 Normalization

3.2.2 Data filtering and statistical analysis

3.2.3 Clustering and differential expression test

3.2.4 Visualization of the pathway

3.2.5 Network regulation

3.2.6 Promoter analysis

4. Final Result...……...

4.1 Result after normalization

4.2 Result after data filtering and statistical analysis

4.3 Result after clustering and differential expression test

4.4 Result of visualization of pathways
4.5 Result from network regulation
4.6 Result from promoter analysis

5. Discussion ……....

6. Summery …….

7. Reference …….....

LIST OF TABLES

Page

Table 3.1 Description of the data ……

Table 3.2 Types of statistical analysis basing on type of data ……

Table 3.3 Regulation property …….

Table 4.1 Result (The differentially expressed gene with related pathway) …...

LIST OF CHARTS

Page

Chart 4.1 Number of gene vs. Pathway ……...

Chart 4.2 Venn-Diagram of pathways …….

LIST OF PATHWAYS

Page

Pathway 4.1 Pathway map of Calcium regulation cardiac cell …….

Pathway 4.2 Pathway map of Apoptosis ……...

Pathway 4.3 Pathway map of Glycogen metabolism ……

LIST OF FIGURES

Page

Fig 2.1 Regulation of insulin in liver ……...
Fig 2.2 Cell Conditions ……....
Fig 2.3 Diabetes in different country ……...
Fig 2.4 Stages of development of T2D ……
Fig 2.5 Complication of diabetes …….
Fig 2.6 Expression gene Profiling ……
Fig 2.7 Spotting Robot with pins …….
Fig 2.8 Overview of Microarray data analysis method ……
Fig 2.9 Type 2 Diabetes (T2D) Mellitus Pathway (KEEG) …….
Fig 2.10 Interaction network ……
Fig 4.1 Raw Data Box plot Fig ……
Fig 4.2 Raw Data Density plot ……....
Fig 4.3 Hierarchical image from dChip …….
Fig 4.4 Legend ……....
Fig 4.5 Direct interaction chain …….
Fig 4.6 Shortest pathway (EGF, transcription) ……..
Fig 4.7 FOS expanded subnet ……
Fig 4.8 PTPN11 expanded subnet …….
Fig 4.9 Network regulator …….
Fig 4.10 Promoter sites ……..

ABBREVIATIONS

NIIDM: Non Insulin Dependent Diabetes Mellitus

AODM: Adult Onset Diabetes Mellitus

CAD: Coronary Artery Disease

MI: Myocardial Infarction

FABP2: Fatty-Acid Binding Protein 2

LpL: Lipoprotein Lipase

PPAR: Peroxisome Proliferator-Activated Receptor

OGTT: Oral Glucose Tolerance Test

HDL: High-Density Lipoprotein

MAD: Median and Absolute Deviation

FDR: False Discovery Rate

HCL: Hierarchical Clustering

SOM: Self-Organizing Map

IRS: Insulin Receptor Substrates

PI3K: Phosphoinositide 3-Kinase

PDK-1: Phosphate-Dependent Kinase-1

ROS: Reactive Oxygen Species

RMA: Robust Multiarray Analysis

RLE: Relative Log Expression

NUSE: Normalized Unscaled Standard Error

CALM1: Calmodulin 1 (phosphorylase kinase, delta)

GNG11: Guanine nucleotide binding protein (G protein), gamma 11

PRKAR2A: Protein Kinase, cAMP-dependent, Regulatory, type II, alpha

PRKAR2B: Protein Kinase, cAMP-dependent, Regulatory, type II, beta

PRKCH: Protein kinase C, eta

JUN: v-jun sarcoma virus 17 oncogene homolog (avian)

EGF: Epidermal growth factor (beta-urogastrone)

FOS: v-fos FBJ murine Osteosarcoma viral oncogene homolog

GO: Gene Ontology

CTNNB1: catenin (cadherin-associated protein), beta 1, 88kDa

AR: Androgen Receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulbar muscular atrophy; Kennedy disease)

PTPN11: Protein tyrosine phosphatase, non-receptor type 11 (Noonan syndrome 1)

CALM1: Calmodulin 1 (phosphorylase kinase, delta)

PHKB: Glycogenin 2 (**GYG2)**, phosphorylase kinase, beta

PHKG1: Phosphorylase kinase, gamma 1 (muscle)

ABSTRACT

Any malfunctioning in body called as disease. Researcher always want understand it and explore its cause by different technologies. When the disease becomes a leading cause in world wide necessity of understanding it doubles. Now days there are so many technologies available which helps in this aspect.

In the past few years the gene expression microarray technology has become a central tool in the field of functional genomics. This field deals with exploring the functions of different gene products, the control mechanisms regulating their activity, their expression levels and their interactions. In the gene expression microarray technology, the expression levels of thousands of genes in a biological sample are determined in a single experiment. Thus this technology is preferred for analysis of any disease.

Cardiovascular disease acts as a risk factor for the multifactor diabetes. Three type of diseased subgroup from GEO GDS268 data for studied for understanding the disease. It is found that there are about 148 genes which are significantly disregulated resulting in diabetes. There are genes like **EIF5A** which show negative effect (-1.24) in morbidly obese condition where as in obese it is up regulated (1.1). Same way for the gene **EIF2S3,** show negative effect (-1.09) in obese condition where as in morbidly obese it is up regulated (1.06). From interaction study concluded that **JUN**, **EGF**, **FOS**, **CTNNB1**, **AR**, **PTPN11** participate in direct interaction. These genes including other differentially expressed genes seems to be relate with the different pathways as apoptosis, insulin signaling, fatty acid bio-synthesis, MAPK- signaling etc which are directly or indirectly deals with different disease. These are classified into clusters and found the co-expressed gene. Further data mining study and literature survey has been done.

We can conclude there is not single gene or any single factor effect this disease. So the name which is given for Type 2 Diabetes (T2D) as multifactor disease it appropriate for it. We can also say that the genes which relate with Type 2 Diabetes also relate with some other diseases causing direct and indirect effects.

1. INTRODUCTION

This project started with the **Aim** to perform microarray data analysis on Type 2 Diabetes (T2D) data, To find out the differentially expressed gene among control and diseased set, Among differentially expressed genes, the related gene with Type 2 Diabetes (T2D), The related gene influencing which all pathways, Interaction chain study for advance analysis and the literature for differentially expressed gene related to Type 2 Diabetes (T2D).

Any condition that impairs the normal functioning of an organism can be called a **disease** (1). It can be classified into three general groups as infectious, noninfectious and disease of unknown origin. Some disease falls under any one category but there are some diseases which fall under many categories with many causes, affecting many organs of the body. Example of such type of disease is Diabetes. When we speak of diabetes, generally correlate it with TYPE 2 DIABETES MELLITUS also called as Non Insulin Dependent Diabetes Mellitus (NIDDM) or Adult Onset Diabetes Mellitus (AODM). This is sixth leading cause of disease in world wide (2). So it became necessary to understand about disease, found the gene related and solution to overcome by manipulating the gene related with diabetes and its risk factors. For this, it is important to experiment over mammalian gene (Homo sapiens, Mus musculus and Rattus norvegicus). As we know to experiment over huge amount of gene, microarray experiment is the best solution. One can adapt statistical and pathway analysis of microarray expression data which are available in the net by using many freely available software. Before these steps we have to understand some terms related to diabetes.

Insulin, the "hunger hormone" produced by beta cells in the pancreatic islets, is the key regulator of carbohydrate metabolism (3). Its major function is to counter the concerted action of a number of hyperglycemia-generating hormones and to maintain low blood glucose levels. Because there are numerous hyperglycemic hormones, untreated disorders associated with insulin generally lead to severe hyperglycemia and shortened life span (4). In addition to its role in regulating glucose metabolism, insulin stimulates lipogenesis, diminishes lipolysis, and increases amino acid transport into cells. Insulin also modulates transcription, altering the cell content of numerous mRNAs. These all directly linked with diabetes.

Diabetes is of many types. Type 1-diabetes results from an autoimmune destruction of the beta-cells, and a total lack of insulin (5). **Type 2-diabetes**, the most common type of diabetes, typically affecting over-40's adults, on the other hand, is associated with an increased insulin resistance in the liver, skeletal muscles and adipose tissue. Pathologically it is characterized by impairment in both insulin secretion and insulin action. It is most common metabolic disorder. When the beta cell cannot compensate for the increasing demand for insulin, type 2 diabetes develops. Health care providers are finding more and more children are now a day gets affected with type 2 diabetes (T2D). The epidemics of obesity and the low level of physical activity among young people, as well as exposure to diabetes in utero, may be major contributors to the increase in type 2 diabetes during childhood and adolescence (6).

The relationship between T2DM and **obesity** is complex (7). The two conditions may have some common genetic predispositions and endocrinological features (8). Obesity, an unhealthy diet and physical inactivity are the main cause for the current increase in prevalence of type 2 diabetes. It stands out as a common and most popular risk. Increased adiposity plays a key role in the progression of insulin resistance in insulin sensitive tissues, namely skeletal muscle, liver, pancreatic beta cells and adipose tissue. NIDDM is not a disease which produces single effect on any organ but it affects many part of body.

Diabetes is considered a '**coronary artery disease** (CAD) equivalent' because patients with diabetes without known CAD have a similar cardiac event rate to patients without diabetes who had a prior myocardial infarction (MI) (9). Fat oxidation may be reduced in obese and morbidly obese individuals. Provides understanding of how obesity contributes to cardiovascular disease via insulin resistance (10). Most of the cardiovascular complications related to diabetes have to do with the way the heart pumps blood through the body. Diabetes can change the chemical makeup of some of the substances found in the blood and this can cause blood vessels to narrow or to clog up completely. This is called atherosclerosis, or hardening of the arteries, and diabetes seems to speed it up. Diabetes also plays an important role for causing blindness (retinopathy), kidney disease (nephropathy), nerve disease (neuropathy), amputation. To know about change in regulation of gene about above mentioned disorder at a time, microarray experiment has been implemented.

DNA microarray is a device that measures the expression of many thousand genes in parallel. The purpose of a microarray is to detect the presence and abundance of labeled nucleic acid in a biological sample. There are many sites which stores the experimental microarray data with complete annotation. Affymetrix is a leading group in it. We can get the data from NCBI's GEO site which is a public microarray repository with expression value in tabular form. These can be used on many available software such as RMAExpress, MeV, dChip, Cluster and Tree view, ArrayAssist, PathwayArchitect, Advance Pathway painter, Cytoscape for statistical and meta analysis. Once we can understand about the software and the data format required by the software, we can find some candidate genes for type 2 diabetes (T2D) which are differentially expressed.

To promote awareness about diabetes the World Diabetes Day celebrated every year on **November 14.** It was established by IDF and WHO in 1991 with the aim of coordinating diabetes advocacy worldwide. The World Diabetes Day 2007 campaign will focus on the impact of diabetes on the lives of children and adolescents worldwide.

2. REVIEW OF LITERATURE

2.1 Insulin

Insulin is synthesized as a **preprohormone** in the beta cells of the islets of Langerhans (3). Its signal peptide is removed in the cisternae of the endoplasmic reticulum and it is packaged into secretory vesicles in the Golgi, folded to its native structure, and locked in this conformation by the formation of two disulfide bonds. Specific protease activity cleaves the center third of the molecule, which dissociates as C peptide, leaving the amino terminal B peptide disulfide bonded to the carboxy terminal A peptide.

2.1.1 Regulation of insulin

Insulin secretion from beta cells is regulated by plasma glucose levels (11). Increased uptake of glucose by pancreatic beta-cells leads to a concomitant increase in metabolism which leads to an elevation in the ATP/ADP ratio (3). This in turn leads to an inhibition of an ATP-sensitive K^+ channel. The net result is a depolarization of the cell leading to Ca^{2+} influx and insulin secretion (12). Chronic increases in numerous other hormones, such as growth hormone, placental lactogen, estrogens, and progestin, up-regulate insulin secretion, probably by increasing the preproinsulin mRNA and enzymes involved in processing the increased preprohormone which leads to Type 2 Diabetes (T2D) (3).

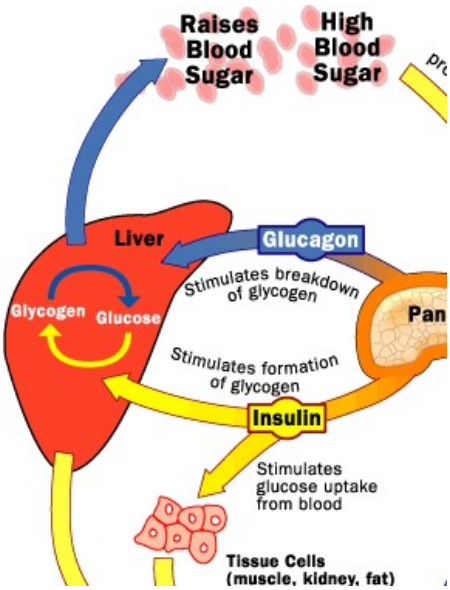


Fig 2.1 Regulation of insulin in liver

Insulin is directly infused via the portal vein to the liver **(Fig 2.1.)**, where it exerts profound metabolic effects. These effects are the response of the activation of the insulin receptor which belongs to the class of cell surface receptors that exhibit intrinsic tyrosine kinase activity (3). The insulin receptor is a heterotetramer of 2 extracellular, a-subunits disulfide bonded to 2 transmembrane b-subunits (13). With respect to hepatic glucose homeostasis, the effects of insulin receptor activation are specific phosphorylation events that lead to an increase in the storage of glucose with a concomitant decrease in hepatic glucose release to the circulation as diagrammed below (only those responses at the level of glycogen synthase and glycogen phosphorylase are represented) (3).

2.2 Prediabetes

In prediabetes, blood glucose levels are higher than normal but not high enough to be defined as diabetes (14). However, many people with prediabetes develop Type 2 Diabetes (T2D) within 10 years, states the National Institute of Diabetes and Digestive and Kidney Diseases. Prediabetes increases the risk of heart disease and stroke. With modest weight loss and moderate physical activity, people with prediabetes can delay or prevent Type 2 Diabetes (T2D).

2.3 Diabetes

2.3.1 Brief overview

Diabetes is a condition where the body is unable to automatically regulate blood glucose levels, resulting in too much glucose (a sugar) in the blood (15). Glucose comes from foods that contain carbohydrate (starches and sugars), travels to the muscles and other organs where it is used as fuel. Excess glucose is detoured to the liver where it may be stored for future use. The blood glucose level is regulated with the help of insulin, a hormone (or chemical messenger) made in the pancreas.

Diabetes develops when the pancreas stops producing insulin (Type 1 diabetes) or when the body does not respond properly to insulin (Type 2 Diabetes). The normal blood glucose level ranges between 3.5-7.8 m mol/l. Over time, high blood glucose levels may damage blood vessels and nerves. These complications of diabetes can cause damage to eyes, nerves and kidneys and increase the risk of heart attack, stroke, impotence and foot problems. This damage can happen before an individual knows if they have diabetes (15). Studies have shown that if blood glucose and cholesterol levels, and blood pressure are kept within normal limits, the risk of damage to the body is reduced. Therefore, it is important to know if a person has diabetes.

2.3.2 Understanding diabetes

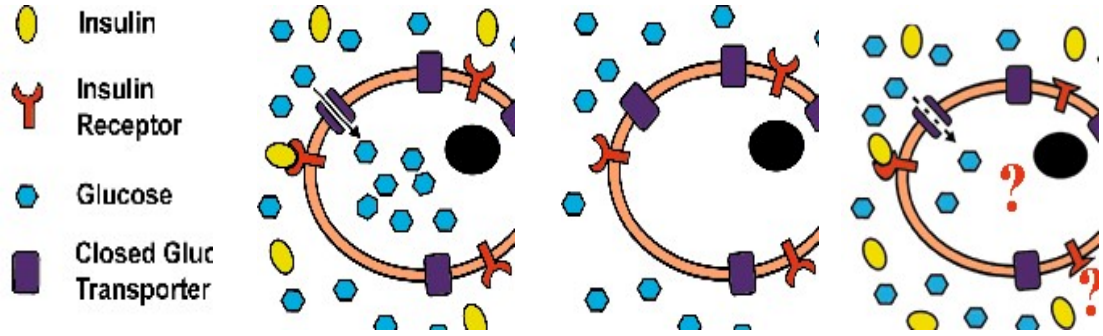
To understand diabetes, it is important to first understand the normal process of food metabolism. Several things happen when food is digested:

(a) A sugar called glucose enters the bloodstream, acting as source of fuel for the body.

(b) An organ called the pancreas makes insulin. The role of insulin is to move glucose from the bloodstream into muscle, fat, and liver cells, where it can be used as fuel. People with diabetes have high blood glucose. This is because their pancreas does not make enough insulin or their muscle, fat, and liver cells do not respond to insulin normally, or both.

2.3.3 Types of diabetes:

- **Type 1 diabetes**: - It is an autoimmune disease **(Fig 2.2.c)**. The body makes little or no insulin. It is found only about 5-10% of cases (3). Symptoms of type 1 diabetes usually develop over a short period, although beta cell destruction can begin years earlier. Symptoms may include increased thirst and urination, constant hunger, weight loss, blurred vision, and extreme fatigue. If not diagnosed and treated with insulin, a person with type 1 diabetes can lapse into a life-threatening diabetic coma, also known as diabetic ketoacidosis.
- **Type 2 Diabetes (T2D)**: - It is far more common than type 1 and makes up 90% or more of all cases of diabetes. It usually occurs in adulthood. Here, the pancreas does not make enough insulin to keep blood glucose levels normal, often because the body does not respond well to the insulin **(Fig 2.2.d)**. Type 2 Diabetes (T2D) is becoming more common due to the growing number of older Americans, increasing obesity, and failure to exercise.



(a) (b) (c) (d)

Fig 2.2 Cell conditions (a) Legend. (b) The normal activity of the cell for insulin action (c) Type I Diabetes cell with autoimmunity (d) Type II Diabetes (T2D) cell with insulin resistance

- **Gestational diabetes**: - In this high blood glucose develops at any time during pregnancy in a person who does not have diabetes. Women with gestational diabetes may not experience any symptoms.

- **Other specific types** of diabetes result from specific genetic syndromes, surgery, drugs, malnutrition, infections, and other illnesses. Such types of diabetes may account for 1 to 2 %.

2.3.4 Brief study on Type 2 Diabetes (T2D)

2.3.4.1 General idea

A subclass of DIABETES MELLITUS is NON-INSULIN-responsive or dependent (**NIDDM**). This is characterized initially by INSULIN RESISTANCE and HYPERINSULINEMIA; and eventually by GLUCOSE INTOLERANCE; HYPERGLYCEMIA; and overt diabetes. Type 2 Diabetes (T2D) begins when the body develops a resistance to insulin and no longer uses the insulin properly. As the need for insulin rises, the pancreas gradually loses its ability to produce sufficient amounts of insulin to regulate blood sugar.

2.3.4.2 Prevalence of Type 2 Diabetes (T2D) and related risk factor

While diabetes occurs in people of all ages and races, some groups have a higher risk for developing Type 2 Diabetes (T2D) than others as African Americans, Latinos, Native Americans, and Asian Americans/Pacific Female suffer more from diabetes in old age than male.

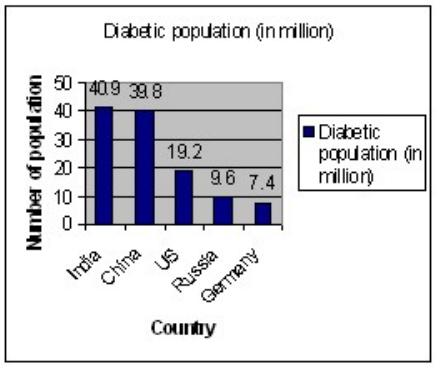


Fig 2.3. Diabetes in different country

Prevalence of **Type 2 Diabetes (T2D)** in urban Indian adults has increased from <3% in the 1970s to > 12% in 2000 (16).India has the highest number of diabetes patient of any country i.e. about 40.9 million **(Fig 2.3)** with comparison to China (39.8 million), the United States (19.2 million), Russia (9.6 million) and Germany (7.4 million). Over the past 25 year, the prevalence of **coronary heart disease** (CHD) in Indian adult has increased from <2% to ~ 10% (17). It is predicted that by 2025 India will have >60 million diabetic patients and that CHD will be the leading cause of death in adults (18, 19). In other words, one in five diabetic patients in the world will be Indian (19).

2.3.4.3 Stages of development Type 2 Diabetes (T2D)

- The first stage in Type 2 Diabetes (T2D) is the condition called **insulin resistance (Fig 2.4 Stage 1)**. Although insulin can attach normally to receptors on liver and muscle cells, certain mechanisms prevent insulin from moving glucose (blood sugar) into these cells where it can be used. Most patients with Type 2 Diabetes (T2D) produce variable, even normal or high, amounts of insulin.
- Over time, the pancreas becomes unable to produce enough insulin to overcome resistance. In Type 2 Diabetes (T2D), the initial effect of this stage is usually an abnormal rise in blood sugar right after a meal **(Fig 2.4 Stage 2 with limited glucose tolerance)** called postprandial hyperglycemia. This effect is now believed to be particularly damaging to the body.

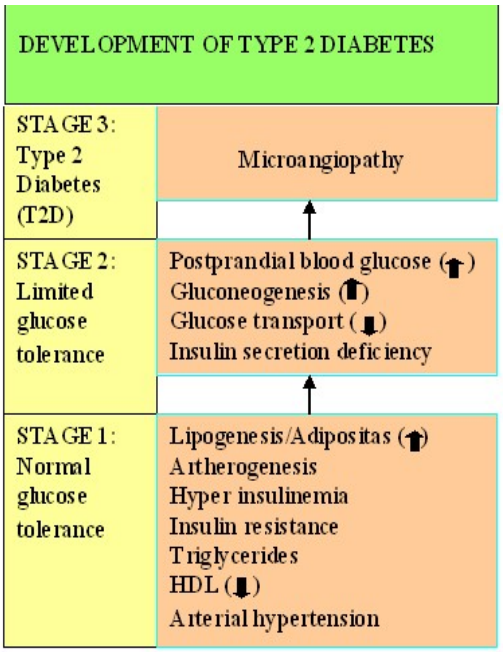


Fig 2.4. Stages of development of T2D

- Eventually, the cycle of elevated glucose further impairs and possibly destroys beta cells, thereby stopping insulin production completely and causing full-blown diabetes **(Fig 2.4 stage 3)**. This is made evident by fasting hyperglycemia, in which elevated glucose levels are present most of the time.

The risk factors work significantly to develop insulin resistance in the body. This leads to Impaired Glucose Tolerance (IGT). The proliferation of beta cell takes place for production of insulin resulting exhaustion of beta cell after some time. This leads to development of Type 2 Diabetes (T2D) Mellitus.

2.3.4.4 Factors predisposing to Type 2 Diabetes (T2D)

Type 2 Diabetes (T2D) is so called multifactorial disease is caused by a complicated interplay of genes, environment, insulin abnormalities, increased glucose production in the liver, increased fat breakdown, and possibly defective hormonal secretions in the intestine. The recent dramatic increase indicates that lifestyle factors (obesity and sedentary lifestyle) may be particularly important in triggering the genetic elements that cause this type of diabetes.

I. Genetic Factors

Genetic factors play an important role in Type 2 Diabetes (T2D), but the pattern is complicated, since both impairment of beta cell function and an abnormal response to insulin are involved. The problem with diabetes is that, even if the exact same mutation caused it in everyone, it would look different from person to person and family to family, depending on environmental influences, the genetic background, and modifier genes. Its expression would be variable. Diabetes is not simple; it's genetically complex, involving multiple genes, and multiple gene-environment interactions.

- ✓ A defective fatty-acid binding protein 2 (FABP2) gene may result in higher levels of unhealthy fat molecules (particularly triglycerides), which may be critical in the link between obesity and insulin resistance in some people with T2D.
- ✓ A defective lipoprotein lipase (LpL) gene may pose a risk for coronary artery disease and Type 2 Diabetes (T2D) in people who have it.
- ✓ A variation in a gene that regulates a protein called calpain-10 is proving to affect insulin secretion and action and may play a role in diabetes type 2.
- ✓ Defective genes that regulate a molecule called peroxisome proliferator-activated receptor (PPAR) gamma may contribute to both Type 2 Diabetes (T2D) and high blood pressure in some patients.
- ✓ A defective gene has been detected that reduces activity of a protective substance called beta$_3$-adrenergic receptor, which is found in *visceral* fat cells (those occurring around the abdominal region). The result is a slow-down in metabolism and an increase in obesity. The defective gene has been found in Pima Indians and other populations with a very high incidence of Type 2 Diabetes (T2D) and obesity.

II. Genotype by Environment Interaction

Since Type II diabetes essentially did not exist 100 years ago, it's obvious that a change in the environment has created the disease. Genetic factors combine with environmental factors diet and exercise cause diabetes. A group of Pima Indians in Mexico are believed to be genetically the same as the Pimas in Arizona, but live in area with no refrigeration, no trucks, no roads, no electricity, and no remote controlled TVs. The Mexican Pimas have no incidences of diabetes.

The high incidence of diabetes is not confined to the Pima tribe. Southwestern Native Americans, including Navajo and Pueblo, have the highest rates of diabetes in the world. Years ago, there was no diabetes on Indian reservations. Over the past five years, there has been a four-fold increase in Type II diabetes in kids. A high incidence of diabetes is seen in Hispanics because of admixture of Native American genes.

Anthropologist Robert Ferrell hypothesized that Type II diabetes is a New World syndrome. He has postulated the concept of a "Thrifty Genotype"; since native Americans went through cycles of feast and famine, they needed a gene that conserves blood sugar (perhaps so that mothers could conserve blood sugar for their babies), and this genotype was selected as a survival mechanism.

The Thrifty Gene: One theory suggests that some cases of Type 2 Diabetes (T2D) and obesity are derived from normal genetic actions that were once important for survival. A gene called as "thrifty" gene, regulates hormonal fluctuations to accommodate seasonal changes. In certain nomadic populations, hormones are released during seasons when food supplies have traditionally been low, which results in resistance to insulin and efficient fat storage. Because modern industrialization has made high-carbohydrate and fatty foods available all year long, the gene no longer serves a useful function and is now harmful because fat, originally stored for famine situations, is not used up.

2.3.4.5 Symptoms of Type 2 Diabetes (T2D):

- Frequent infections that are not easily healed
- High levels of sugar in the blood when tested
- High levels of sugar in the urine when tested
- Unusual thirst
- Frequent urination
- Extreme hunger but loss of weight
- Blurred vision
- Nausea and vomiting
- Extreme weakness and fatigue
- Irritability and mood changes
- Dry, itchy skin
- Tingling or loss of feeling in the hands or feet

Some people who have Type 2 Diabetes (T2D) exhibit no symptoms or symptoms may be mild and unnoticeable, or easy to confuse with signs of aging. It may also resemble other conditions or medical problems.

2.3.4.6 Diagnosis of Type 2 Diabetes (T2D)

The fasting blood glucose test is the preferred test for diagnosing diabetes in children and non pregnant adults because of the ease of measurement and the considerable time commitment of formal glucose tolerance testing, which can take two hours to complete. It is most reliable when done in the morning (20). However, a diagnosis of diabetes can be made based on any of the following test results, confirmed by retesting on a different day:

- A blood glucose level of 126 milligrams per deciliter (mg/dL) or more after an 8-hour fast. This test is called the fasting blood glucose test.
- A blood glucose level of 200 mg/dL or more 2 hours after drinking a beverage containing 75 gram of glucose dissolved in water. This test is called the oral glucose tolerance test (OGTT).
- A random (taken at any time of day) blood glucose level of 200 mg/dL or more, along with the presence of diabetes symptoms.

By current definition, two fasting glucose measurements above 126 mg/dL or 7.0 m mol/l are considered diagnostic for diabetes mellitus.

Patients with fasting sugars between 6.1 and 7.0 m mol/l (i.e., 110 and 125 mg/dL) are considered to have "impaired fasting glycemia" and patients with plasma glucose at or above 140mg/dL or 7.8 m mol/l two hours after a 75 g oral glucose load are considered to have "impaired glucose tolerance". "Prediabetes" is either impaired fasting glucose or impaired glucose tolerance; the latter in particular is a major risk factor for progression to full-blown diabetes mellitus as well as cardiovascular disease.

2.3.4.7 Risk factors for diabetes

A risk factor is anything that may increase a person's chance of developing a disease. Although these factors can increase a person's risk, they do not necessarily cause the disease. Some people with one or more risk factors never develop the disease, while others develop disease and have no known risk factors.

But, knowing risk factors to any disease can help to guide us into the appropriate actions, including changing behaviors and being clinically monitored for the disease.

List of some risk factor are:

1. Age
2. People over the age of 45 are at higher risk for diabetes.
3. Family history of diabetes
4. Being overweight
5. Not exercising regularly
6. Acean dethnicity
7. High blood levels of triglycerides (a type of fat molecule), blood pressure and blood cholesterol level.
8. History of gestational diabetes, or giving birth to a baby that weighed more than 9 pounds.
9. A low level HDL (high-density lipoprotein - the "good cholesterol")
10. A high triglyceride level

2.3.4.8 Complications of diabetes

1. **Heart disease**

- Heart disease is the leading cause of diabetes-related deaths. Adults with diabetes have heart disease death rates about 2 to 4 times as high as those of adults without diabetes **(Fig 2.5)**.

2. **Blindness**

- Diabetes is the leading cause of new cases of blindness in adults 20 to 74 years old.
- Diabetic retinopathy causes from 12,000 to 24,000 new cases of blindness each year **(Fig 2.5)**.

3. Stroke

- The risk of stroke is 2 to 4 times higher in people with diabetes **(Fig 2.5)**.

4. Kidney disease

- Diabetes is the leading cause of end-stage renal disease, accounting for about 40 percent of new cases.
- 27,851 people with diabetes developed end-stage renal disease in 1995 **(Fig 2.5)**.

5. High blood pressure

- An estimated 60 to 65 percent of people with diabetes have high blood pressure.

6. Nervous system disease

- About 60 to 70 percent of people with diabetes have mild to severe forms of nervous system damage (which often includes impaired sensation or pain in the feet or hands, slowed digestion of food in the stomach, carpal tunnel syndrome, and other nerve problems).
- Severe forms of diabetic nerve disease are a major contributing cause of lower extremity amputations.

7. Amputations

- More than half of lower limb amputations in the United States occur among people with diabetes.
- From 1993 to 1995, about 67,000 amputations were performed each year among people with diabetes.

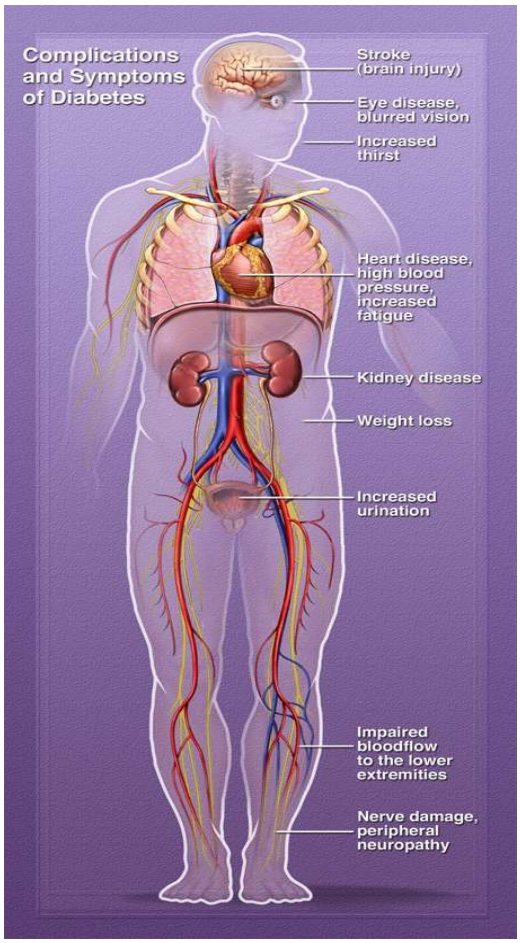


Fig 2.5 Complications of diabetes

8. Dental disease

- Periodontal disease (a type of gum disease that can lead to tooth loss) occurs with greater frequency and severity among people with diabetes. Periodontal disease has been reported to occur among 30 percent of people age 19 years or older with type 1 diabetes.

9. Complications of pregnancy

- The rate of major congenital malformations in babies born to women with preexisting diabetes varies from 0 to 5 percent among women who receive preconception care to 10 percent among women who do not receive preconception care.
- Between 3 and 5 percent of pregnancies among women with diabetes result in death of the newborn; the rate for women who do not have diabetes is 1.5 percent.

10. Other complications

- Diabetes can directly cause acute life-threatening events, such as diabetic ketoacidosis* and hyperosmolar nonketotic coma.*
- People with diabetes are more susceptible to many other illnesses.

*Diabetic ketoacidosis and hyperosmolar nonketotic coma are medical conditions that can result from biochemical imbalance in uncontrolled diabetes.

"PREVENTION IS BETTER THEN CURE"

2.3.4.9 Prevention or delay of onset of Type 2 Diabetes (T2D):

Type 2 Diabetes (T2D) may be prevented or delayed by following a program to eliminate or reduce risk factors - particularly losing weight and increasing exercise.

2.3.4.10 Treatment for Type 2 Diabetes (T2D):

Treatment typically includes diet control, exercise, home blood glucose testing, and, in some cases, oral medication and/or insulin.

Specific treatment for Type 2 Diabetes (T2D) will be determined by physician based on:

- Age, overall health, and medical history
- Extent of the disease
- Tolerance for specific medications, procedures, or therapies
- Expectations for the course of the disease
- Opinion or preference

The goal of treatment is to keep blood sugar levels as close to normal as possible. Emphasis is on control of blood sugar (glucose) by monitoring the levels, regular physical activity, meal planning, and routine healthcare. Treatment of diabetes is an ongoing process of management and education that includes not only the person with diabetes, but also healthcare professionals and family members. Often, Type 2 Diabetes (T2D) can be controlled through losing weight, improved nutrition, and exercise alone. However, in some cases, these measures are not enough and either oral medications and/or insulin must be used.

2.4 Expression profiling

Recent advances in bioinformatics and high-throughput technologies such as microarray analysis are bringing about a revolution in understanding of the molecular mechanisms underlying normal and dysfunctional biological processes. These stimulate the discovery of new targets for the treatment of disease which is aiding drug development, immunotherapeutics and gene therapy (21). Gene expression profiling or microarray analysis **(Fig 2.6)** has enabled the measurement of thousands of genes in a single RNA sample. There are a variety of microarray platforms that have been developed to accomplish this and the basic idea for each is simple: a glass slide or membrane is spotted or "arrayed" with DNA fragments or oligonucleotides that represent specific gene coding regions. Purified RNA is then fluorescently- or radioactively labeled and hybridized to the slide/membrane. In some cases, hybridization is done simultaneously with reference RNA to facilitate comparison of data across multiple experiments. After thorough washing, the raw data is obtained by laser scanning or autoradiographic imaging. At this point, the data may then be entered into a database and analyzed by a number of methods.

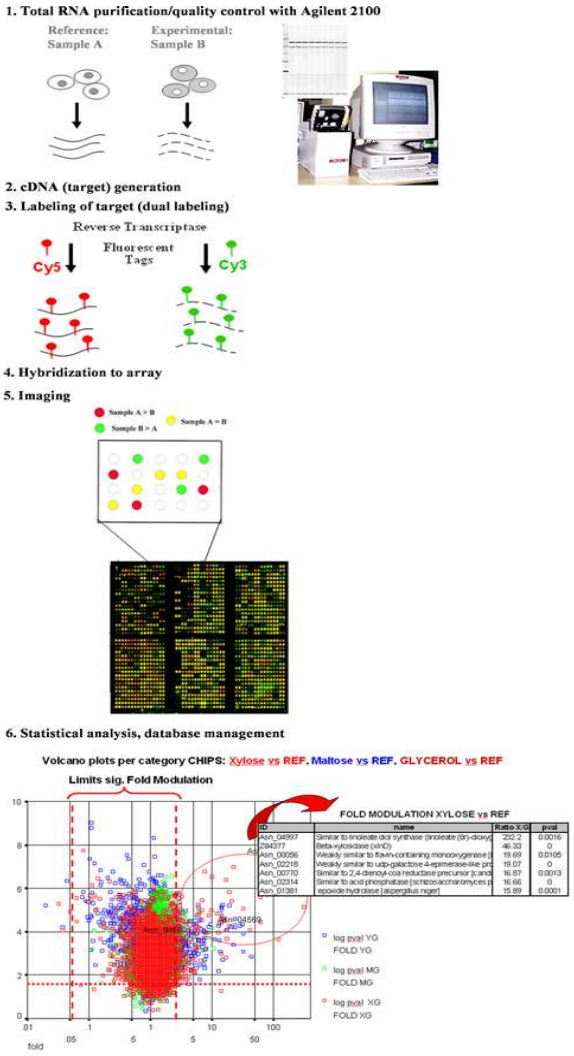


Fig 2.6 Expression gene Profiling

2.4.1 Microarray

Microarray expression analysis has become one of the most widely used functional genomics tools. Efficient application of this technique requires the development of robust and reproducible protocols. The name itself suggests that analysis of large number of data present in array with minute quantity. There are different types of microarray as DNA microarray, Protein microarray, Tissue microarray, Transfection microarray/Cell microarray, Chemical compound Microarray, Antibody microarray. The detail of DNA microarray is give below.

2.4.1.1 Introduction

Microarray consists of a solid surface, usually a microscope slide, onto which sample molecules have been chemically bonded. The purpose of a microarray is to detect the presence and abundance of labeled nucleic acids in sample. In the majority of microarray experiments, the nucleic acids are derived from the mRNA, thus microarray measures gene expression. There are two technologies for making microarray. These are:-

(A) Robotic spotting

This is the technology by which first microarray was manufactured. The array is made using a spotting robot **(Fig 2.7)** via three main steps.

1. Making the DNA probes to put on the array.
2. Spotting the DNA onto the glass surface of the array with the spotting robot.
3. Post spotting processing of the glass slide.

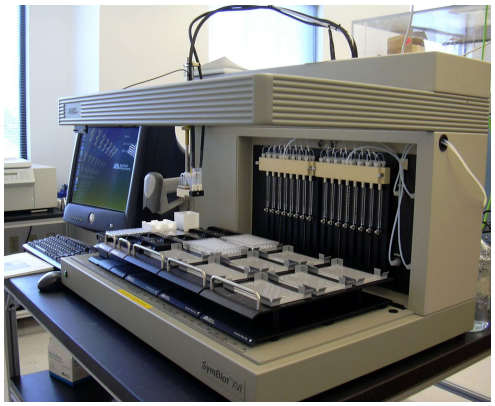


Fig 2.7 Spotting Robot with pins

The attachment chemistry may be covalent or non-covalent. With **covalent attachment**, a primarily aliphatic amine (NH_2) group is added to the DNA probe and the probe is attached to the glass by making a covalent bond between this group and chemical linkers on the glass. With oligonucleotide probes, the amine group can be added to either end of the oligonucleotide during synthesis, although it is more usual to add it to the 5 prime end of the oligonucleotide. With cDNA probes, the amine group is added to the 5 primer, thus the cDNA probes are always attach from the 5 prime ends. With non-covalent attachment, the bonding of the probe to the array is via electrostatic attraction between the phosphate backbone of the DNA probe and NH_2 groups attached to surface of the glass. The interaction takes place at several locations along the DNA backbone, so that the probe is tethered to the glass at many points. Because most oligonucleotide probes are shorter than cDNAs, these interactions are not strong enough to anchor oligonucleotide probes to glass. Therefore, non-covalent attachment is usually only used for cDNA microarray. The typical printing process follows steps:-

1. Pins are dipped into the wells to collect the first batch of DNA.
2. This DNA is spotted onto a number of differentially, depending on the number of array being made and the amount of liquid pin can hold.
3. The pins are washed to remove any residual solution and ensure no contamination of the next sample.
4. The pins are dipped into the next set of well. Then return to the above step 2 and repeat until the array is complete.

(B) In-Situ Synthesized Oligonucleotide Arrays

These arrays are fundamentally different from spotted arrays: instead of presynthesising oligonucleotides, oligos are built up base-by-base on the surface of the array. This takes place by covalent reaction between the 5 prime hydroxyl group of the sugar of the last nucleotide to be attached and the phosphate group of the next nucleotide. Each nucleotide added to the oligonucleotide on the glass has protective group on its 5 prime position to prevent the addition of more than one base during each round of synthesis. The protective group then converts to a hydroxyl group either with acid or with light before next round of synthesis. The different methods for detection lead to the three main technologies for making in-situ synthesized arrays:

1. Photodeprotection using masks: this is the basis of the Affymetrix technology.

2. Photodeprotection without masks: this is the method used by Nimblegen and Febit.

3. Chemical deprotection with synthesis via inkjet technology: this is the method used by Rosetta, Aligent and Oxford Gene Technology.

2.4.1.2 Technique

Efficient expression analysis using microarray requires the development and successful implementation of a variety of laboratory protocols and strategies for fluorescence intensity normalization. The process of expression analysis can be broadly divided into Array Fabrication, Probe Preparation and Hybridization, Data Collection, Normalization and Analysis.

2.4.1.3 Applications of microarray

1. Determination of transcriptional programs of cells for a given cellular function (e.g., cell function, cell differentiation, etc.) or when they are exposed to certain conditions leading to activation, inhibition or apoptosis.

2. Compare and contrast transcriptional programs to aid diagnosis of diseases, predict therapeutic response and provide class discovery and sub-classification of diseases.

3. Identification of genome-wide binding sites for transcriptional factors that regulate the transcription of genes.

4. Prediction of gene function and evaluation of gene expression.

5. Identification of new therapeutic targets (target identification, target validation, and drug toxicity).

6. Development of public databases that will help us understanding of the functioning of complex biological systems.

2.5 Analysis of Expression data

Before describing all the methods let us look the flowchart **(Fig 2.8)** for brief idea. To get idea about data source and type of data in data bases we can visit the following website: - http://ihome.cuhk.edu.hk/~b400559/arraysoft_public.html.

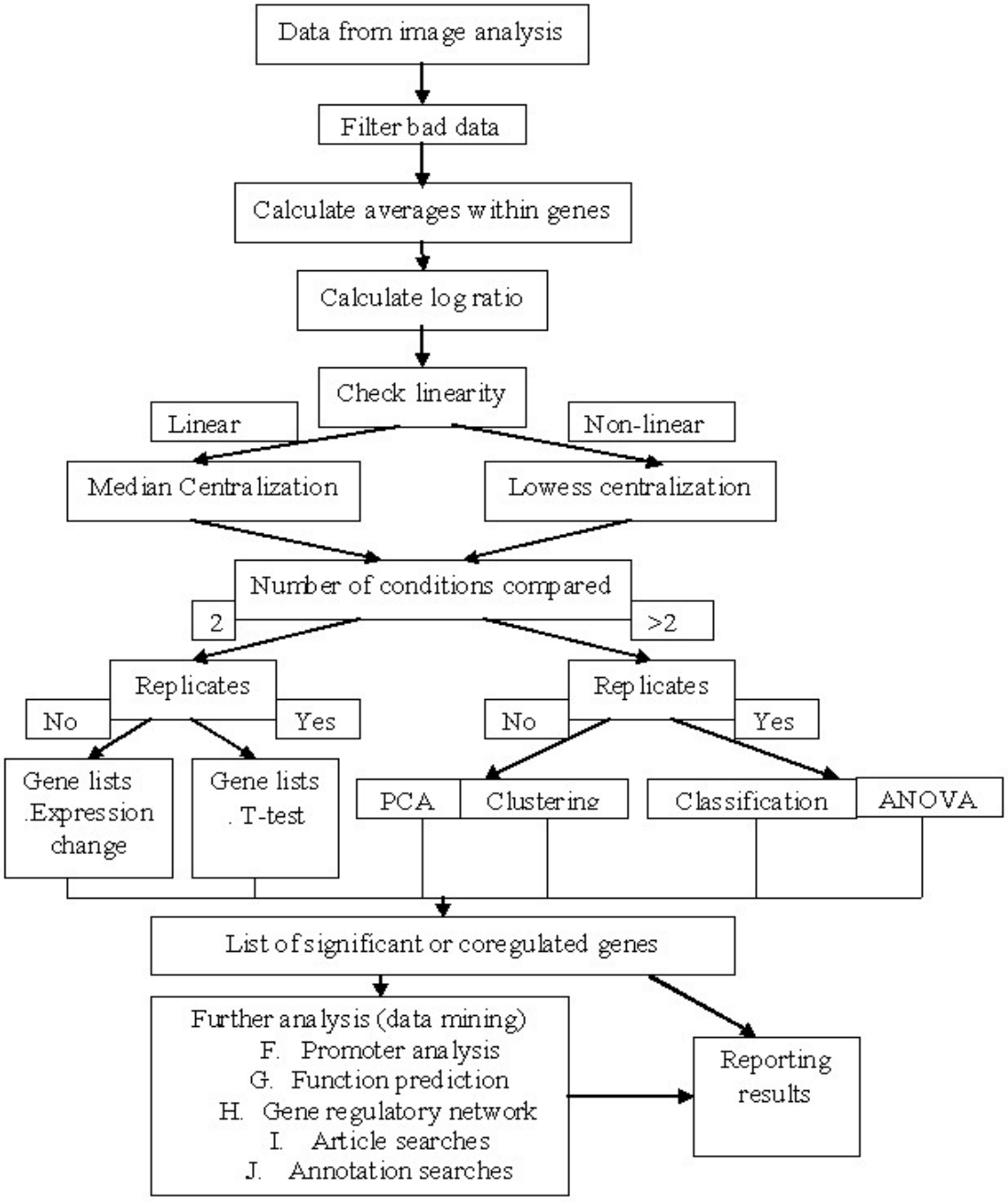


Fig 2.8 Overview of Microarray data analysis method

2.5.1 Slide scanning

Differential gene expression is assessed by scanning the hybridized arrays using a confocal laser scanner capable of interrogating both the Cy3- and Cy5-labeled probes and producing separate TIFF images for each. As is the case with arraying robots, there are a number of manufacturers that produce scanners capable of detecting Cy3 and Cy5 and most are planning to release instruments capable of detecting additional dyes. The image of the microarray generated by the scanner is the raw data of the experiment. A computer algorithm, known as feature extraction software, converts the image into the numerical information that quantifies gene expression. The image processing involved in feature extraction has a major impact on the quality of the data and interpretation.

2.5.2 Image Processing and numerical data collection

The first step in the analysis of microarray data is to process the image. Most manufacturers of microarray scanners provide their own software; however, it is important to understand how data is actually being extracted from images, as this represents the primary data collection step and forms the basis of any further analysis.

2.5.3 Normalization

Normalization is collection of methods that are directed at resolving the systematic error and bias introduced by microarray experimental platform. The purpose of normalization is to identify and remove the effects of systematic variation, other than differential expression, in the measured fluorescence intensities. It is necessary to normalize the fluorescence intensities before any analysis which involves comparing expression levels within or between slides (e.g. classification, multiple testing), in order to ensure that differences in intensities are indeed due to differential expression and not experimental artifacts. This can be done by following methods.

2.5.3.1 Data cleaning and transformation

There are three stages in data cleaning and transformation

1. **Removing flagged feature**

   There are four flagged feature is mostly known. These are bad feature, negative feature, dark feature and manually flagged feature. There are two approaches to deal with flagged features:-

   (a) Removing of flagged feature from the data set. This is a straight forward method which is easy but of disadvantage of removing potentially valuable data sometimes. **(b)** Back to the original image of every flagged feature and to identify the problem that has resulted in flagging. This procedure has the disadvantage of requiring time and resources, and may not be always practical.

2. Background subtraction

The background signal is thought to represent the contribution of non-specific hybridization of labeled target to the glass, as well as the natural fluorescence of the glass slide itself. This approach is good when the feature intensity is greater then the background intensity but in reverse condition produces negative number. There are three approaches that deals with this situation are:

- Remove the feature
- Use the lowest available signal intensity measurement as the background subtracted intensity which is typically 1.
- Use more sophisticated (Bayesian) algorithm to estimate the true feature intensity.

3. Taking logarithms

The easiest approach for normalization is log transformation. The objectives of it are:

- There should be a reasonably even spread of features across the intensity range.
- The variability should be constant at all intensity level.
- The distribution of experimental error should be approximately normal.
- The distribution of intensities should be approximately bell-shaped.

Some methods used in data cleaning and transformations are:

- **Log2 Transform**

  This is fairly self-evident, just taking the log2 transform of every element in the matrix. We can convert a natural logarithm to a logarithm to base 2 via following equation:

  Log (to base 2) x = log (natural) x / log (natural) 2.

- **Normalize Genes/Rows**

  This will transform values using the mean and the standard deviation of the row of the matrix to which the value belongs, using the following formula:

  Value = [(Value) – Mean (Row)]/ [Standard deviation (Row)]

- **Divide Genes/Rows by RMS**

  This will divide the value by the root mean square of the current row.

  Root mean square = square root [$\Sigma (x_i)^2 / (n-1)$]; Here x_i is the i^{th} element in the row consisting of n elements.

- **Divide Genes/Rows by SD**

  This will divide each value by the standard deviation of the row it belongs to.

- **Mean Center Genes/Rows**

  This will replace each value by [value – Mean (row that value belongs to)].

- **Median Center Genes/Rows**

  This will replace each value by [value – Median (row that value belongs to)].

- **Digital Genes/Rows**

  This will divide up the interval between the minimum and the maximum values in a row into a number of equal-sized "bins". Each value is now replaced by an integer value of zero or greater, denoting which bin it belongs to (e.g., the minimum value is assigned to bin "zero", indicating it belongs to the lowest bin; the maximum value is assigned to the highest bin, and the rest of the values fall in the intermediate bins).

- **Sample/Column Adjustment**

  These function in the same way as their corresponding options on genes/rows, except that the current column values, rather than the current row values, are used in the computation.

- **Log10 to Log2**

  This assumes that the current data are log 10 transformed, and transforms them to log base 2, i.e., it assumes that the input data is in the form log10x, and it outputs log2x.

- **Log2 to Log10**

  This assumes that the current data are log 2 transformed, and transforms them to log base 10, i.e., it assumes that the input data is in the form log2x, and it outputs log10x.

- **Unlog2 transformation**

  This assumes that the current data are log2 transformed, and removes the log2 transformation.

2.5.3.2 With in array normalization

The main aim of microarray experiment is to find out differentially expressed gene. We know that in the considered control sample very less number of the gene are differentially expressed. Thus when we measures the differentially expressed gene between samples, we need to ensure that the experiment represent true differential gene expression and not the samples. We can assume three methods for correction of different responses. These are:

1. Linear regression of Cy5 against Cy3
2. Linear regression of log ratio against average intensity.
3. Non-linear (Loess i.e. Locally Weighted Polynomial Regression) regression of log ratio against average intensity.

We can control the smoothing factor by using smooth option in SAS loess procedure. Typically 0.2 – 0.4 used to specify the fraction, or the procedure can apply the default fraction.

2.5.3.3 Between Array normalization

Between arrays normalization is necessary when comparing or correlating the microarray data between patients or different arrays. Two basic methods, scaling and centering, are used in the current program. Box-plots are the most preferable way to visualize the data. The other method is Distribution normalization.

- **Centering**

Centering makes the means and the standard deviations of all the distributions equal. The method is similar to scaling: for each measurement on the array, subtract the mean value of the array and divide by the standard deviation. Alternatively in place of mean we can use median and absolute deviation from the median (MAD). This has advantage of being more robust to outliers than the mean and standard deviation, but has the disadvantage of not producing a distance matrix when using Pearson correlation. Centering is very commonly used for comparing multiple arrays.

- **Scaling**

Scaling scales the data so that the means of all the distributions are equal. It is simply done by subtracting the mean value of all data on the array from each measurement on the array. The mean of the measurement on each array will be zero after scale normalization. With out mean we can use median which provides more robust value.

- **Distribution normalization**

Data distribution normalized to ensure that the distribution of the data on each of the array is identical. The methodologies are:

1. Center the data
2. For each array, order the centered measurements from lowest to highest.
3. Compute a new distribution whose lowest value is the average of the values of the lowest expressed gene on each of the array; whose second lowest values is the average of the second-lowest values from each of the arrays; and so on until the highest value is the average value of the highest values from each of the array.
4. Replace each measurement on each array with the corresponding average in the distribution.

2.5.4 Statistical analysis

Statistical analysis plays a very important role in understanding the relation between the genes. It tries to answer the question: "Can the difference between these observations be explained by chance alone?" or "How significant is this difference?" Statistical testing can also be viewed as hypothesis testing, where two different hypotheses are compared. For example, we can test whether diabetes patients and their healthy controls differ statistically significantly for the expression of the gene X. By statistical analysis we can conclude the co expressive gene.

2.5.4.1 Types of most widely used Statistical methods

- **T-test**

The t-test assesses whether the means of two groups are *statistically* different from each other. This analysis is appropriate whenever we want to compare the means of two groups. The formula for the t-test is a ratio. The top part of the ratio is just the difference between the two means or averages. The bottom part is a measure of the variability or dispersion of the scores. This formula is essentially another example of the signal-to-noise metaphor in research: the difference between the means is the signal that, in this case, we think our program or treatment introduced into the data; the bottom part of the formula is a measure of variability that is essentially noise that may make it harder to see the group difference. The t-value will be positive if the first mean is larger than the second and negative if it is smaller. Once you compute the t-value you have to look it up in a table of significance to test whether the ratio is large enough to say that the difference between the groups is not likely to have been a chance finding.

- **ANOVA: Analysis of Variance**

ANOVA is an extension of the t-test to more than two experimental conditions. It picks out genes that have significant differences in means across three or more groups of samples. The user is initially required to enter the number of groups, following which a sample grouping panel similar to the t-test panel, with the appropriate number of groups, is created. Samples can be assigned to any group or excluded from the analysis. F-statistics are calculated for each gene, and a gene is considered significant if p-value associated with its F-statistic is smaller than the user-specified alpha or critical p-value. Currently, p-values are computed only from the F-distribution.

2.5.4.2. Steps of statistical testing

A suitable statistical test can solve our question. Statistical testing is divided into five sub steps as:

1. Select an appropriate statistical test.
2. Select a threshold for *p*-value and Form the pair of hypotheses you want to compare.
3. Calculate the test statistic and degrees of freedom.
4. Compare the test statistic to the critical values table of the test statistic distribution.
5. Draw conclusions.

1. Select an appropriate statistical test.

Usually the means of certain groups are compared in the microarray experiments. For example, we can test whether the expression of a certain gene is higher in the diabetes patients than in their healthy controls. When choosing a test, there are two essential questions, which need to be answered: Is there more than two groups to compare, and should we assume that the data is normally distributed? If two groups are compared, there are two applicable tests, the t-test and Mann-Whitney U test. If more than two groups are compared, an analysis of variance (ANOVA) or a Kruskal-Wallis test is used. If the data is normally distributed the t-test for two groups, or the ANOVA for multiple groups can be used for comparisons.
If the data is not normally distributed, and each group has at least five observations, the Mann-Whitney U test or Kruskal-Wallis test can be applied.

2. Threshold for *p*-value and Hypothesis pair

The *p*-value is usually associated with a statistical test, and it is the risk that we reject the null hypothesis, when it actually is true. Before testing, a threshold for *p*-value should be decided. This is a cut-off below which the results are statistically significant. Often a threshold of 0.05 is used. This means that every 20th time we conclude by chance alone that the difference between groups is statistically significant, when it actually isn't.
If the compared groups are large enough, even the tiniest difference can get a significant *p*-value. In such cases it needs to be carefully weighted whether the statistical significance is just that, statistical significance, or is there real biological phenomenon acting in the background. Before applying the test to the data, a hypothesis pair should be formed. A hypothesis pair consists of a null hypothesis ($H0$) and an alternative hypothesis ($H1$).

3. Calculation of test statistic and degrees of freedom

The test statistic is a standardized numerical description of the differences between the group means. Depending on the test type, the actual mathematical formula for the calculation of the test statistic is different.

4. Critical values table

A critical values table contains the standardized values of a distribution. The critical value table for t-tests contains the standardized values of the t-distribution. The shape of the t-distribution was partly defined by degrees of freedom. Therefore, dfs are also essentially needed when reading the critical values table for the t-distribution.

5. Drawing conclusions

Compare the test statistic and the critical value from the table. If the calculated test statistic is larger, then the null hypothesis is rejected, and we can conclude that there is a difference between the groups with a certain *p*-value threshold.

2.5.4.3 Multiple testing

The more analyses are performed on a data-set, the more the results will meet the conventional significance level by chance alone. One commonly used correction is the Bonferroni correction, where the original *p*-value is adjusted by the number of comparisons to create a new corrected *p*-value against which those comparisons should be tested. **False discovery rate** (FDR) controls for the expected proportion of false positives instead of the chance of any false positives as Bonferroni correction does. FDR threshold is calculated from the distribution of observed *p*-values, but FDR is not interpreted in a similar fashion as *p*-value. In DNA microarray studies FDR is often reported as a q-value, which is actually a "posterior Bayesian *p*-value". FDR seems to be more appropriate than Bonferroni corrected *p*-value for large datasets.

2.5.4.4 Empirical *p*-value

Permutation tests are a family of methods that can be used for calculation of empirical *p*-values. An empirical t-test *p*-value for a gene discriminating between two groups of samples, *e.g.*, diabetic patients and their healthy controls can be calculated as follows. First a t-test statistic for the original dataset is calculated. Then the sample labels, *i.e.*, the group the individual samples belong to, are randomized very many times (10 000). A t-test statistic is recalculated for every of the randomized datasets, and the empirical p-value is the percentage of randomized dataset that got a larger t-test statistic than the original dataset. The genes having the smallest empirical *p*-values best discriminate the groups from each other. However, if the number of individuals in both groups is small (<10), it might be better to use critical values table for calculation of p-values.

2.5.4.5 Statistical conclusion on microarray data

1. Scatter plot

The scatter plot is an important graphical tool for studying the linearity of the data. In its simplest form, two variables are plotted along the axes, and marks are drawn according to these coordinates.

2. Box plot

It is also an important tool which allows us to visualize distribution of the data simultaneously. A box plot shows a distribution as central box bracketed by horizontal line known as whiskers. So this plot also known as Whisker's box plot. The line through the centre of the box represents the mean of the distribution. The box itself represents the standard deviation of the distribution. The horizontal line bracketing the box represents the extreme value of the distribution.

3. Correlation

The correlation of two variables represents the degree to which the variables are linearly related. When two variables are perfectly linearly related, the points in the scatter plot fall on a straight line. Two summary measures or correlation coefficients, Pearson's correlation and Spearman's rho, are most commonly used. Both of these measure range from -1 to 1. Spearman's correlation is based on ranks. Before calculating the value of Spearman's correlations all the observations are ranked. If the correlation is perfect, the smallest observations of variable *X* corresponds to the smallest observation of variable *Y*, and so forth until the largest observations.

4. Linear regression

Linear regression is used for describing how much the dependent variable (the predicted variable) changes if the independent variable (the predictor variable) changes. Before applying linear regression, the linearity of the data should be checked from a scatter plot. In addition, the normality of the dependent variable should be checked before running the analysis (see the following sections). Linear regression fits a straight line through the data points so that the square sum deviation of the line from the data points is minimized.

2.5.5 Collection of differentially expressed gene

After applying all the method we have to choose the differentially expressed gene which is our gene of interest, by calculating the fold change value. The value 1 also keeps significant because we are considering the log transformed data. The value of 1 in real expression is 2 fold change of expression value.

2.5.6 Clustering the data

Clustering is the process of organizing objects into groups whose members are similar in some way. A *cluster* is therefore called as a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. The similarity criterion is *distance*: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called *distance-based clustering*. Another kind of clustering is *conceptual clustering*: two or more objects belong to the same cluster if this one defines a concept *common* to all that objects.

2.5.6.1. Distance and similarity measure

Calculation of distance or similarity between two expression vectors is fundamental of clustering algorithms. By considering the distance we can say which genes are co-expressed and which are different from a particular group. In other word the similar genes have difference zero. These distances can be based on a single dimension or multiple dimensions. If we had a two- or three-dimensional space this measure is the actual geometric distance between objects in the space. However, the joining algorithm does not "care" whether the distances that are "fed" to it, are actual real distances, or some other derived measure of distance that is more meaningful. Some type of distance measures are:

- **Euclidean distance.** This is most commonly chosen type of distance. It simply is the geometric distance in the multidimensional space. Euclidean distances are usually computed from raw data. This method has certain advantages. However, the distances can be greatly affected by differences in scale among the dimensions from which the distances are computed.
- **City-block (Manhattan) distance.** This distance is simply the average difference across dimensions. In most cases, this distance measure yields results similar to the simple Euclidean distance. However, in this measure, the effect of single large differences (outliers) is dampened (since they are not squared).
- **Pearson's correlation (r).** This is used to calculate degree of association.
- **Spearman Rank-Order correlation:** It is a distribution-free analog of correlation analysis mentioned. Like regression, it can be applied to compare two independent random variables

2.5.6.2. Types of clustering

1. Hierarchical Clustering (HCL)

HCL is an agglomerative clustering method which joins similar genes into groups. The iterative process continues with the joining of resulting groups based on their similarity until all groups are connected in a hierarchical tree. During construction of the hierarchy, decisions must be made to determine which clusters should be joined. The distance or similarity between clusters must be calculated. The rules that govern this calculation are linkage methods.

Linkage Methods: Linkage methods are rules or metrics that return a value that can be used to determine which elements (clusters) should be linked. Three linkage methods that are commonly used are:

Single Linkage

Cluster-to-cluster distance is defined as the minimum distance between members of one cluster and members of another cluster. Single linkage tends to create ‘elongated’ clusters with individual genes chained onto clusters.

Average Linkage

Cluster-to-cluster distance is defined as the average distance between all members of one cluster and all members of another cluster. Average linkage has a slight tendency to produce clusters of similar variance.

Complete Linkage

Cluster-to-cluster distance is defined as the maximum distance between members of one cluster and members of another cluster. Complete linkage tends to create clusters of similar size and variability.

2. k means clustering

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function.

2.5.6.3 Application of clustering

There are a lot of applications of clustering not only in the biological field but also in day to day life. Just to cite some example:-

1. Similarity searching in Medical Image Database.

2. Data Mining purpose

Data mining is the process of discovering meaningful new correlation, patterns and trends by sifting through large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques. Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation. Additional analyses using standard analytical and other data mining techniques can determine the characteristics of these segments with respect to some desired outcome.

2.5.7 Pathway analysis

Biological pathways **(Fig 2.9)** of Type 2 Diabetes (T2D) represent complex reactions at the molecular level in living cells, and interactions between bio-molecules to accomplish a biological function. Based on the overall effect they have on the functioning of an organism, pathways may be divided into several different categories. Three major categories are: metabolic pathways (24), transcription and protein synthesis pathways (25), and signal transduction pathways (26). As the requirements to analyze different kinds of pathways are similar, unless explicitly stated otherwise, a pathway for this discussion refers collectively to all.

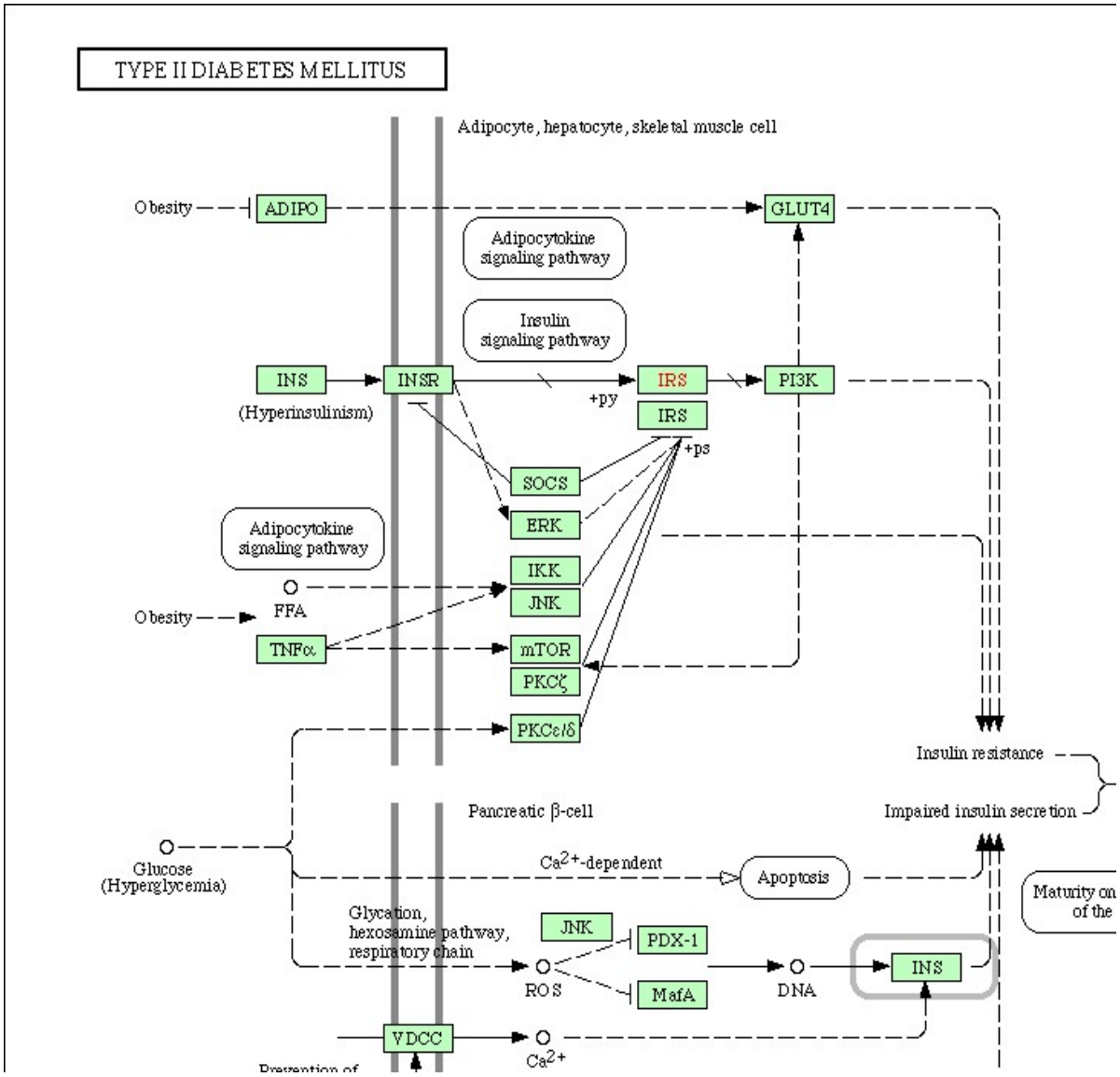


Fig 2.9 Type 2 Diabetes (T2D) Mellitus Pathway (KEEG)

Pathways serve as a focal point to integrate other diversely related information, such as literature citations, experimental data, research notes, etc. To facilitate usage and exploratory analysis of complex pathways, visual representations of pathways are necessary. Some diagrams are manually generated such as found in textbooks (27), KEGG (28) and Biocarta (29), and others are generated by interactive visualization software such as GenMapp (30) and PathwayAssist (31). Editor tools such as Pathway Editor (32) and Knowledge Editor (33) allow users to create pathway visualizations manually. A large number of systems such as PathwayAssist (31), Pathway Finder (34), PubGene (35), GENIES (36), Omniviz (37), etc. use Natural Language Processing (NLP) algorithms to generate pathways automatically from research articles retrieved from search engines. Systems such as GenePath (38) infer pathways from microarray data. VectorPathBlazer (39) provides functionalities to create pathways by combining information from different reference databases such as KEGG (28), and BIND (40).

2.5.7.1 Pathways and genes related to Type 2 Diabetes (T2D)

The genes which are directly related to Type 2 Diabetes (T2D) in different tissue are ADIPO, GLUT4, INS, INSR, TNF alpha, VDCC, SUR1, Kir6.2, GLUT2, SOCS, ERK, IKK, JNK, mTOR, GK, IRS, PDX-1, MafA, PYK, PI3K, TNFR1, TNFR2, LEPR, ADIPOR, LEP, TRADD, JAK, SHP-2, AMPKK, TRAF2, STAT3, PPARalpha, PEPCK, ACC2, AGRP, NPY, FACS etc.

There are so many pathways which are related with the Type 2 Diabetes (T2D) directly and indirectly. These includes Adipocytokine signaling pathway, Insulin signaling pathway, MAPK signaling pathway, Phosphatydyl insitol signaling pathway, Apoptosis, Starch and Sucrose metabolism, Pyruvate metabolism, Glycolysis/Gluconeogenesis, beta alanine metabolism, Fatty acid biosynthesis (Path I), Fatty acid metabolism, Wnt signaling pathway etc.

1. Insulin signaling pathway: Signaling through the insulin pathway is critical for the regulation of intracellular and blood glucose levels and the avoidance of diabetes. Insulin binds to its receptor leading to the autophosphorylation of the b-subunits and the tyrosine phosphorylation of insulin receptor substrates (IRS). IRS phosphorylates the SH2 domain of Shp2, a tyrosine phosphatase, and the SH3 domain of the adaptor molecule Grb2. Activated Grb2 recruits Sos1 that, in turn, activates the Ras signaling pathway and gene transcription. IRS also activates phosphoinositide 3-kinase (PI3K) through its SH2 domain, thus increasing the intracellular concentration of PIP_2 and PIP. This, in turn, activates phosphatidyl Inositol phosphate-dependent kinase-1 (PDK-1), that subsequently activates Akt/PKB. This results in the translocation of the glucose transporter (GLUT4) from cytoplasmic vesicles to the cell membrane.

Impairment in insulin secretion: The reduction of β – cell function develops early in the pathogenesis of type 2 diabetes. B-cell dysfunction is present already in normoglycemic offspring of diabetic patients, in subjects with IFG and IGT, and is prominent at the time of the diagnosis of diabetes. The observation that prolonged and acute hyperglycemia can lead to decreased insulin secretion gave rise to the concept of

glucose toxicity. The toxic effects of glucose thought to be mediated via increased production of reactive oxygen species (ROS), down-regulation glucose transporters, and induction of β-cell apoptosis (41). On the other hand, prolonged exposure to elevated Free Fatty Acids, commonly observed in patients with type 2 diabetes, may also lead to decreased glucose – stimulated insulin secretion a β-cell apoptosis. Possible explanations are accumulation long – chain coenzyme A (CoA), reduced formation of ATP, and subsequent opening of K_{ATP} channels. Furthermore, insulin secretory capacity may be affected by impaired function of incretins, fetal and infant malnutrition, and various genetic causes influencing glucose metabolism, energy transduction and β-cell apoptosis and regenration (42).

Tissue – specific insulin resistance: Skeletal muscle is the main tissue of insulin – dependent glucose metabolism in humans, where as adipocytes account for only 10% of total glucose disposal (43). Nevertheless, muscle – specific ablation of Insulin Resistance has very modest effects on systemic glucose homeostasis, as shown by tissue – specific knock out studies in mice. Loss of Insulin Receptor in fat does not affect the over all tolerance, products from weight gains, and even extends life span. How ever, deletion of Insulin Resistance in muscle and fat causes impaired glucose tolerance, probably indicating a synergistic and mutual metabolic compensation between the two tissues (44)

In contrast, liver – specific and β-cell – specific disruption of insulin signaling result in more drastic metabolic consequences. It increases hepatic glucose production, impairs glucose – dependent insulin secretion and leads to the development of glucose intolerance and diabetes (45). More over, brain – specific Insulin Resistance knockout causes hyperphagia, moderate obesity and reduced fertility, suggesting the role of insulin in the control of appetite and reproduction.

2. Apoptosis: Apoptosis is a likely mechanism mediating the increase in cell death. Based on the some experiment of b-cell mass and replication rate measurements, the measured changes in b-cell mass were due not to a failure of b-cell proliferation but to a change in the balance between b-cell neogenesis and death from apoptosis. This change in balance involves a decrease in the rate of neogenesis, an increase in the rate of cell death, or both. The excessive b-cell death most likely results from an increased rate of b-cell apoptosis and raises the importance of defining factors that promote b-cell survival and those that promote b-cell death in models of NIDDM.

3. Wnt signaling pathway: The Wnt genes have been reported to play a pivotal role in embryonic development and oncogenesis. Recently, one isoform of the WNT family has been reported to play an important role in adipogenesis (Ross et al. 2000; Bennett et al. 2002), and low-density lipoprotein receptor–related protein 5 (*LRP5* [MIM 603506]), which is known as a mediator for WNT signaling, has also been shown to be involved in glucose-induced insulin secretion (Fujino et al. 2003). Therefore, genes related to the Wnt signaling pathway could be considered as candidate genes for conferring susceptibility to Type 2 Diabetes (T2D).

2.5.8. Interaction Network

After all above analysis the main advance analysis step becomes interaction network. By combining gene expression analysis, perturbations or treatments, and mutations of genes we can study processes like signal transduction and metabolism yielding information on molecular effects or functions of specific genes. By this gene expression data permits us to go beyond the traditional line of research and to study finer structures of molecular pathways exposing causal regulation relations between genes. It not only describes the nature of interactions between genes, inhibition or activation, but also exemplifies direct and indirect effects of genes. For example, is gene **A** regulating gene **B** or vice versa, is the regulation direct or indirect where there is a mediating gene **C** (or maybe many) so that **A** regulates **C** and then **C** regulates **B**. Inspecting the finer structures, which are called regulatory networks, gives us a more intricate view of molecular interactions offering further possibilities for medical interventions. Example:

A directed edge E(i , j) between vertices Vi and Vj represents regulation between genes attached to vertices Vi and Vj , and the direction of the edge, denoted by Vi →Vj , reflects the order of regulation where gene *Vi* regulates gene Vj (up- or down regulation) **(Fig 2.10)**. Node Vi is called a parent of Vj and node Vj is called a child of Vi. To present logical structures of group regulations, which appear extensively in nature, hyper edges should be used, where an edge is adjacent to several genes, but these are used quite rarely because of the convenience of simple drawings.

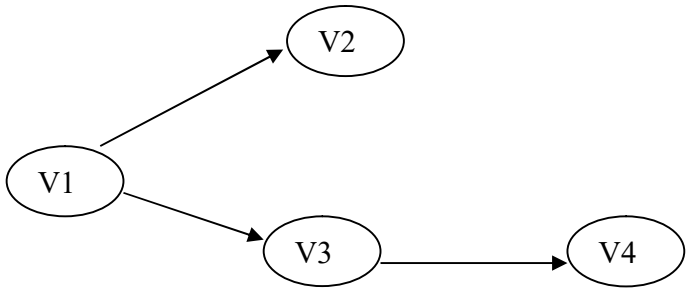


Fig 2.10. Interaction network

To infer a regulatory network, several distinct expression samples of the genes of discourse are needed. In a time series analysis, the expression levels are recorded along fixed time points, while in a perturbation analysis, the expression levels are recorded separately for each manual perturbation (over/under, deletion) of specific genes. A fundamental difference between perturbation data and time series data is that perturbation allows a firm inferring order of regulation while time series data can only reveal the probable regulation direction. Because the number of possible networks increases super-exponentially on the number of genes, prominent regulatory networks are usually searched for by using supervised approaches, where some intrinsic partial knowledge about regulations between genes are formulated by implicit or explicit rules. For the mathematical modeling of regulation inside a network, there are many approaches like Bayesian network, Boolean network and its generalization, ordinary and partial differential equations, qualitative differential equations, stochastic master equations, Petri nets, transform grammars, process algebra, and rule-based formalisms etc.

2.5.9 Promoter Analysis

A **promoter** is a regulatory region of DNA located upstream (towards the 5' region) of a gene, providing a control point for regulated gene transcription. The promoter contains specific DNA sequences that are recognized by proteins known as transcription factor (TF). These factors bind to the promoter sequences, recruiting RNA polymerase, the enzyme that synthesizes the RNA from the coding region of the gene. As promoters are typically immediately adjacent to the gene of interest, positions in the promoter are designated relative to the transcriptional start site, where transcription of RNA begins for a particular gene. A major challenge in the analysis of microarray gene expression data is to mine meaningful biological knowledge out of the huge volume of data. It is meaningful in case of disease study. The basic idea for promoter analysis is to study the change in condition of those genes which are differentially expressed. The change in the condition is in the term of chromosomal, hereditary and all other function. This helps us to know about the cause of change in the disease.

A **cis-regulatory element** or cis-element is a region of DNA or RNA which regulates the expression of genes located on that same strand. These cis-regulatory elements are often binding site of one or more trans-acting factors. This may be located in the promoter region 5' to the gene it controls, in the intron, or in the 3' region. In contrast, trans-regulatory elements are species which may modify the expression of genes distant from the gene that was originally transcribed to create them.

A **TATA box** (also called **Goldberg-Hogness box**) is a DNA sequence (Cis-regulatory element) found in the promoter region of most genes (it is considered to be the core promoter sequence) in eukaryotes. It is the binding site of either transcription factors or histones (binding of a transcription factor blocks binding of a histone and vice versa) and is involved in the process of transcription by RNA polymerase. It has the core DNA sequence 5'-TATAAA-3' or a variant, which is usually followed by three or more adenine bases and has been highly conserved through evolution.

It is normally bound by the TATA Binding Protein (TBP) in the process of transcription, which unwinds the DNA, and bends it through 80°. The AT-rich sequence facilitates easy unwinding (due to 2 hydrogen bonds between bases as opposed to 3 between GC pairs). The TBP is an unusual protein in that it binds the minor groove and binds with a β sheet.

The TATA box is usually found as the binding site of RNA polymerase II. The transcription factor TFIID binds to the TATA box, followed by TFIIA binding to the upstream part of TFIID. TFIIB can then bind to the downstream part of TFIID. The polymerase can then recognize this multi-protein complex and bind to it, along with various other transcription factors such as TFIIF, TFIIE and TFIIH. Transcription is then initiated, and the polymerase moves along the DNA strand, leaving TFIID and TFIIA bound to the TATA box. These can then facilitate the binding of additional RNA polymerase II molecules.

3. MATERIALS and METHODS

3.1 Materials

3.1.1 Datasheet Used:

There are many sites from where we can download the expression data. In this experiment I have considered the gene experiment data of GEO data set from NCBI. The data set is of Type 2 Diabetes with coronary problem, a risk factor of diabetes. The accession number of the data set is GDS268.

3.1.2 Softwares Used:

There is hundreds of software for microarray data analysis which may or may not be freely available. Softwares used for this experiment are:-

1. RMAExpress

RMAExpress is a standalone GUI program for Windows (and Linux) to compute gene expression summary values for Affymetrix Genechip® data using the Robust Multichip Average expression summary. RMA is the Robust Multichip Average. It consists of three steps: a background adjustment, quantile normalization (see the Bolstad et al reference) and finally summarization. It can be downloaded from following website.

http://rmaexpress.bmbolstad.com/.

2. Multi Experiment Viewer (MeV)

MeV: MeV is a microarray data analysis tool present in TM4 microarray software suite, incorporating sophisticated algorithms for clustering, visualization, classification, statistical analysis and biological theme discovery. MeV can handle several input file formats. It generates informative and interrelated displays of expression and annotation data from single or multiple experiments. It can be downloaded from following website.

http://www.tm4.org/mev.html

3. ChipInfo

ChipInfo software is designed for retrieving annotation information from online databases such as NetAffx and Gene Ontology (GO) and organizing such information into easily interpretable tabular format outputs. As companion software to dChip and GoSurfer, ChipInfo enables users to independently construct the information resource files of these software packages. This can be downloaded from following website.

http://biosun1.harvard.edu/complab/chipinfo/

4. DNA Chip (dChip)

DNA-Chip Analyzer (dChip) is a Windows software package for probe-level (e.g. Affymetrix platform) and high-level analysis of gene expression microarray and SNP microarray. High-level analysis in dChip includes comparing samples, hierarchical clustering, view expression and SNP data along chromosome, LOH and copy number analysis of SNP arrays, and linkage analysis. In these functions the gene information and sample information are correlated with the analysis results. Then by considering this result we can do further analysis. This software can be downloaded freely from website: -

www.dchip.org

5. Advance pathway painter (APP)

This freeware program visualizes pathways. The user has the possibility to display any kind of quantitative data from gene and protein experiments (e.g. microarray from affymetrix) directly within the pathways (colors represent the value). The linking between the pathway items and the experiment data is done over the gene or protein names and their accession numbers. Furthermore the user has a quick overview on the gene/protein with the collected links in the web-interface (direct links). It can be downloaded from following website.

http://www.gsa-online.de/eng/app.html

6. Cladist

CLADIST is an open-source and integrated platform that provides gene-gene / protein-protein correlation, nearest neighborhood, clustering and visualization tools to allow groups of genes or proteins with similar co-expression patterns to be directly projected onto networks of protein-protein interactions and transcriptional regulatory associations in pSTIING knowledgebase, thus providing additional functional context to coexpression networks. This can be downloaded from following website.

http://pstiing.licr.org/software/cladist_run.jsp

7. Cytoscape

Cytoscape is an open-source community software project for integrating biomolecular interaction networks with high-throughput expression data and other molecular states into a unified conceptual framework. Cytoscape is most powerful when used in conjunction with large databases of protein-protein, protein-DNA, and genetic interactions that are increasingly available for humans and model organisms. This software can be downloaded from following website.

http://www.cytoscape.org/download_list.php

8. ArrayAssist: ArrayAssist® Software powered by avadis™, is easy to use. It offers a full spectrum of advanced analytical features such as statistical tests (multi-way ANOVA and non-parametric statistical tests), data mining tools (classifiers for machine learning-based prediction) and sophisticated, dynamic, interactive visualization options. A new scripting engine is featured in all products of the ArrayAssist line. The engine allows scripting of complex functions and user-defined workflows and is fully integrated with external programs such as the R statistical software. The scripting capabilities enable creation and deployment of automated workflows significantly extending ArrayAssist expression software functionality, ease of use, and its usefulness for the research team. ArrayAssist Expression software is integrated with the PathwayArchitect™ software.

http://www.stratagene.com/tradeshows/feature.aspx?fpId=110

9. PathwayArchitect

Stratagene's PathwayArchitect® software, powered by avadis™, is the new standard in pathway analysis. With its easy to use biology-focused workflow and a vast database of over two million biological interactions, PathwayArchitect® software is designed with the life scientist. The PathwayArchitect® software also features a new method to explore biological interactions—the Relevance Interaction Network. Relevance Interaction Network analysis begins with a protein list and identifies the network of proteins and small molecules, which are most statistically related to the biology of the protein list. This algorithm makes it possible to easily identify binding complexes, transcriptional regulation networks, and even small molecules that regulate the biology of a group of proteins. It can be down loaded from following website.

http://www.stratagene.com/tradeshows/feature.aspx?fpId=90

10. Promoter Integration in Microarray Analysis (PRIMA)

PRIMA is a program for finding transcription factors (TFs) whose binding sites are enriched in a given set of promoters. It is aimed at the identification of TFs that take part in these networks. The basic biological assumption is that genes that are co-expressed over multiple biological conditions are regulated by common TFs, and therefore are expected to share common regulatory elements in their promoters. By utilizing human genomic sequences and models for binding sites (BSs) of known TFs, PRIMA identifies TFs whose BSs are significantly over-represented in a given set of promoters.

http://www.cs.tau.ac.il/~rshamir/expander/expander.html

3.1.3 Algorithms Used:

1. Robust Multiarray analysis (RMA).
2. Hierarchical Clustering algorithm.
3. K means Clustering algorithm.

3.2 Methods

Data sheet

- **Procedure for downloading the data:-**

1. NCBI website opened with typing URL:- www.ncbi.nlm.nih.gov/
2. From search GEO Datasets selected with Type 2 diabetes in the "for" space.
3. Several options come from which GDS268 is chosen for experiment.
4. From this page the supplementary file i.e. .CEL files download and unzipped into a folder.
5. By clicking GDS268 record we get another page. In this page data option is present, clicking it four option comes from which "dataset soft file" is choose and that file is saved to a folder after unzip.
6. The .CDF file, a channel file of HG-U133A is downloaded from following website, which is necessary for processing .CEL file in dChip Software. http://chip.dfci.harvard.edu/biostat/cdf%20files/

- **Description of the data:-**

The normalized expression data of obesity contributing to cardiovascular disease via Type 2 diabetes from National Center for Biotechnology Information (NCBI)'s Gene Expression Ominbus (GEO) site with GEO ID: - GDS268 and citation "Park JJ, Berggren JR, Hulver MW, Houmard JA et al. GRB14, GPD1, and GDF8 as potential network collaborators in weight loss-induced improvements in insulin action in human skeletal muscle. *Physiol Genomics* 2006 Oct 11; 27(2):114-21" and pub med id 16849634 is considered as datasheet for this experiment. The sample organism is Homo sapiens. The platform for this experiment is GPL96 describing Affymetrix Gene Chip Human Genome U133 Array Set **HG-U133A.** In this dataset samples are of human female skeletal muscles and divided into three types of subgroup as:-

Description	Data subset
Non-obese (Control)	GSM3987, GSM3988, GSM3989, GSM3990, GSM3991, GSM3992, GSM3993, GSM3994
Obese	GSM3995, GSM3996, GSM3997, GSM3998, GSM3999, GSM4000, GSM4001, GSM4002
Morbidly obese	GSM3979, GSM3980, GSM3981, GSM3982, GSM3983, GSM3984, GSM3985, GSM3986

Table 3.1 Description of the data

From this **(Table 3.1)** we can conclude the microarray data is on ***case control study*** and ***no replicates*** with ***more than two groups (i.e. three)***.Each group having more than five sub sets. Data set linked with diabetes, obesity and cardiovascular disease. In this sample data first condition is Non-obese which act as control in this study. Other two are diseased set. One is obese and the other one is morbidly obese. In morbidly obese condition the disease is sever. There are several steps through which data analysis done.

3.2.1 Normalization

After downloading the .CEL files, unzipped into a single folder. The HG-U133.CDF file also placed in that folder. RMA Express software is used for normalization of the data by applying robust multiarray analysis (RMA) algorithms.

Procedure:-

From the file menu "Read unprocessed file" is selected. It opens a window for inputting .CDF file and .CEL files. Once it is loaded the datasheet is ready for normalization. Then the "Compute RMA measure" is clicked. This opens a new window with several option as Background Adjust, Normalization (Quantile, None), Summarization Method (median Polish, PLM), with unchecked Store residue. All can be remain default but we have to check the Store residue checkbox and click ok button. This gives us the log transformed normalized data which is used for statistical analysis.

3.2.2 Data filtering and Statistical analysis

Here we have chosen MeV of TM4 software suite for statistical analysis. Before doing any statistical analysis we have keep in mind the data type and which test should be applied. The table 3.2 shows brief idea on it.

Analysis Type	**Parametric**	**Non-Parametric**
Single Group	t-Test against 0	Mann Whitney against 0
Multiple Groups, Unpaired, Pair wise Analysis	t-Test, Unpaired	Mann Whitney, Unpaired
Multiple Groups, Paired, Pair wise Analysis	t-Test, Paired	Mann Whitney, Paired
Multiple Groups, Unpaired, All Together	One-Way ANOVA	Kruskal-Wallis
Multiple Groups, Paired, All Together	Repeated Measures	Repeated Measures (Friedman)
Multiple Factors, Multiple Groups, Unpaired, All Together	n-Way ANOVA	None
Multiple Factors, Multiple Groups, Paired, All Together	Repeated Measures	None

Table 3.2 Types of statistical analysis basing on type of data

Procedure:-

The resulted log transformed file saved as tab delimited format and loaded into the MeV software. It results a graphical image immediately. The rows are called as gene and the column is called as sample. We can get expression image of each gene individually. We have to filter the bad data. There are so many methods from which I have chosen variance filtering for filtering the bad data. Then we can choose statistics from analysis menu for statistical analysis.

As the dataset GDS268 is having three conditions so, I have chosen the ANOVA test **(Table 3.2)**. And find out the significant data by giving p cutoff value. Generally the p cutoff is taken as 0.05. Here after clicking the statistical test several options came from which we choose ANOVA. The new window opened asking the number of group and critical p value. The number of group is set to three and p value to 0.05. There is default parameter. We can change it or if not want can remain as it is. Then after clicking the ok button we will get the result. We can make hierarchical tree on significant gene.

3.2.3 Clustering and Differential expression test

The data is then to be tested for differential expression and differentially expressed gene related to which pathways. There is many more software which helps us for comparing the samples for differentially expressed gene. Let us take dChip for this analysis.

Procedure:-

In dChip software we can upload the raw .CEL files but it is better to load the resulted RMA algorithm applied data file for fold change regulation. It require gene information file. This downloaded from the dChip website or we can make it in Chip info software. After uploading the data clustering test has been done on both samples and genes. There are many clustering methods as hierarchical, k-means, SOM. But dChip has only hierarchical method option.

Clustering can be done on only samples or on genes or on both. On the right side of it the label comes. From the clustering we can find the sample correlation matrix. We have to compare a control sample set against the diseased one for fold change value. As here we have considered log value fold change 1 also keeps significance. By choosing compare sample we get the result which gene are differentially expressed. Then for pathway analysis, selected gene list by annotation from tool menu or if want to list all pathway at a time then can select classify gene. This will give the pathway list and genes corresponding to each pathway. This method applied twice as in data set we have to compare control against obese and morbidly obese. Result from this taken into single file for comparison and drawing conclusion for expression in different sets. There may be some genes which come in many Type 2 Diabetes related pathway. We can plot this as graph for easy view.

3.2.4 Visualization of the pathway

There is hundreds of software which help us to visualize the differentially expressed gene in the pathway map. Pathway visualization software, called as **Advance Pathway Painter (APP)**, is used for visualization of the differentially expressed gene.

Procedure:-

After loading the data into the software we have to click pathway. In the left panel many list will come from which we have to choose one. After clicking on it we will get pathway view. The red marks are the differentially expressed gene. We can click on each to find out related id or can click on list all to find all the ids which are related in differential expression.

3.2.5 Network regulation

The differentially expressed genes now can be use to find the regulatory network. There are many types of software which can be used in this purpose. The method is based on gene-gene interaction. Some software directly calculates the interaction as per its own database as in PathwayArchitect, but for other software we have to give gene interaction table, as in Cytoscape. The gene interaction table can be calculated by the Cladist software based on Pearson's coefficient. The widely used software for network regulation is Cytoscape. It has different plug-ins serving different purpose. The most important plug-in is "Bingo". And main disadvantage is it requires 2GB RAM. The easiest software is PathwayArchitect but the main problem with it is it is not freely available.

Procedure in PathwayArchitect:-

- The differentially expressed gene which extracted from the above software is then uploaded into the PathwayArchitect. In this out of 148 genes only 41 are selected as node. Then studied about the interaction chain.
- First looked for the "**direct interaction**". In direct interaction, the software looks for the protein and other molecules, which are directly related and construct the interaction chain.
- The "**shortest pathway network**" constructed between EGF and transcription.
- In this if we are interested for only two genes then, can be read as "**navigator**" feature. Just selected two genes as JUN and FOS and from right click selected the navigator. Then on each gene right clicked independently and selected "**show neighbors**". Then this new pathway can be saved in new window.
- Our area of interest is the genes which are differentially expressed in the Type 2 Diabetes (T2D). We have pre-knowledge that which genes are acting as "**marker gene**" for this metabolic disease. If we will not find the genes playing important role, are in differently expressed then go for the "**expand interaction**" feature by considering single gene. This will give us sub-network. There we can find the genes which all are related with Type 2 diabetes (T2D). Here for each gene the sub-net constructed by clicking expand network.

- If we wish to find the common genes which are regulating network we can go for network feature.
- A more advanced feature in PathwayArchitect is it directly connects to the "**GO browser**". We can select genes and click on GO browser. This will show the correspondence rank in each pathway type. We can save p value of all. Even give the cutoff p value to select the genes which all have low p value.
- "**Relevance list**" is important because it tally with marker to show which all important genes are related with Type 2 Diabetes (T2D). This can be uploaded and compared with all the genes which all are differentially expressed.
- We can go for "**advanced analysis**" also with all these features.
- For each of above the "**data report**" is saved. This is of different symbol.
- From this genes "**similar pathways**" can also be seen

Explanation of data report

- **Regulation**

If a gene **"A"** regulates **"B"** then we can show symbolically as A---> B, where this is regulation with unknown effect. Similarly we can show different symbolic explanation in a table **(Table 3.3).**

Change in A	Change in B	Effect	Representation
Up	Up	Positive	A --+ > B
Up	Down	Negative	A ---| B
Down	Down	Positive	A --+ > B
Down	Up	Negative	A ---| B

Table 3.3 Regulation property

- **Binding**

Two entities, A and B, participate in a "Binding" interaction, if they physically interact with each other. Since both A, B takes part equally in the interaction, there is no "Target" node and the text representation takes the form A ---- B.

- **Promoter Binding**

If protein A binds to the 5' upstream (or promoter) region of gene B, we represent the relation as a "Promoter Binding" relation, A ---> B. Transcription factor binding information can best be represented in the form of Promoter Binding interactions.

- **Expression**

If A, causes a quantitative change in the amount of protein/mRNA B, we represent the relation as an "Expression" relation.
Example: "Introduction of A caused the degradation of B" can be shown as A ---| B.

- **Metabolism**

Relations involving the quantitative change of small molecule entities are represented as "Metabolism" interactions.
Example: "Enzyme E catalyzed the conversion of small molecule A to B" can be represented as
1. A ---> B (Metabolism type), 2. E --+> (1) (Regulation type)

- **Transport**

The entities are transported between compartments. Relations involving the quantitative change of an entity are not because it is being synthesized (or degraded). These are represented by "Transport" interactions.

3.2.6 Promoter Analysis

The genes are which are differentially expressed and are meaningful are then used for promoter analysis. MatInspector in Genomatrix suite can be used for this purpose which gives all the detail information as promoter binding site, the sequence, literature for each gene relating it with all the signaling pathways and the 3D diagram of genes. This gives us better understanding about each gene. PRIMA related with EXPANDER used for promoter analysis, gives the long sequence with binding site.

Procedure of promoter analysis in EXPANDER:-

- After clicking the EXPANDER icon a window will opened. From file menu "New session -> Expression data" chosen. This gives a window to load the expression file.
- After loading the expression file K- means clustering was done with k = 3 to find co- expressed data.
- Then from group analysis, Promoter analysis had chosen. This pops with a new window with parameters. After adjusting all the parameter (as p value 0.05, the require sequence length) ok button was clicked.
- This will give the binding site bar chart.
- After clicking "view binding site" icon we can see the sequence with binding site for each cluster.

4. RESULT

In microarray analysis is not a single step procedure so result comes in various steps.

4.1 Result after normalization

This gives some graphical result and activates write result to file (log scale) option. Then we can save this normalized log transformed file to use further.

Graphical results from RMAExpress

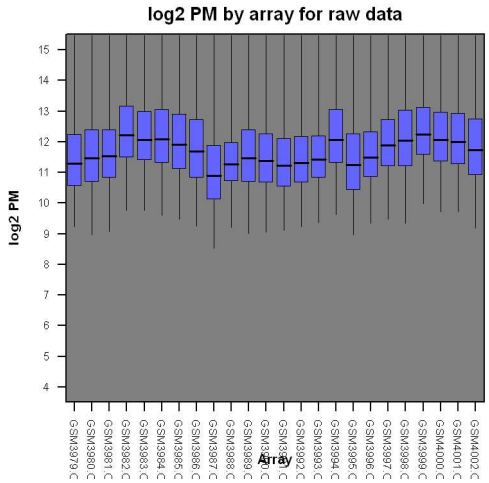


Fig 4.1 Raw Data Box plot

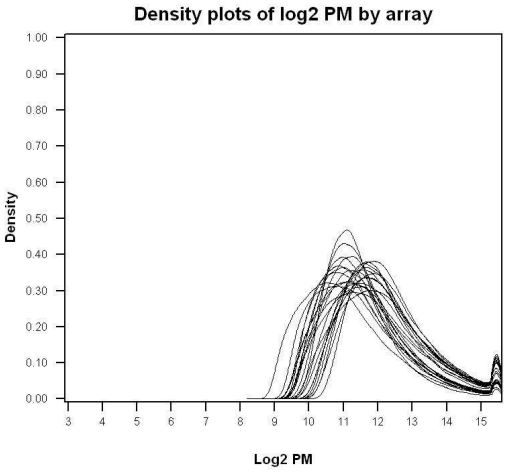


Fig 4.2 Raw Data Density plot

The raw data box plot **(Fig 4.1)** shows the distribution of raw values while the raw density plot **(Fig 4.2)** shows the expression value. This plot gives us the brief idea of quality of the data. If in box plot the value distribution is good then the box will not big and the box will be centralized.

4.2 Result after data filtering and statistical analysis

The result from the ANOVA test in MeV software is Expression images of significant and non-significant genes, Hierarchical Tree of significant genes, Centroid Graphs, Expression Graphs, Table view of significant genes and non-significant genes. Then we can set significant genes as data source for further analysis. Or after saving the significant gene list we can do differential expression test in other software. In this case when significance gene set as data source and saved that data got significant genes which is used for the further analysis.

4.3 Result after Clustering and Differential expression test

In dChip software when hierarchical clustering done it show the cluster of gene each labeled with its id in one side and in other side related pathway or any metabolic function **(Fig 4.3).**

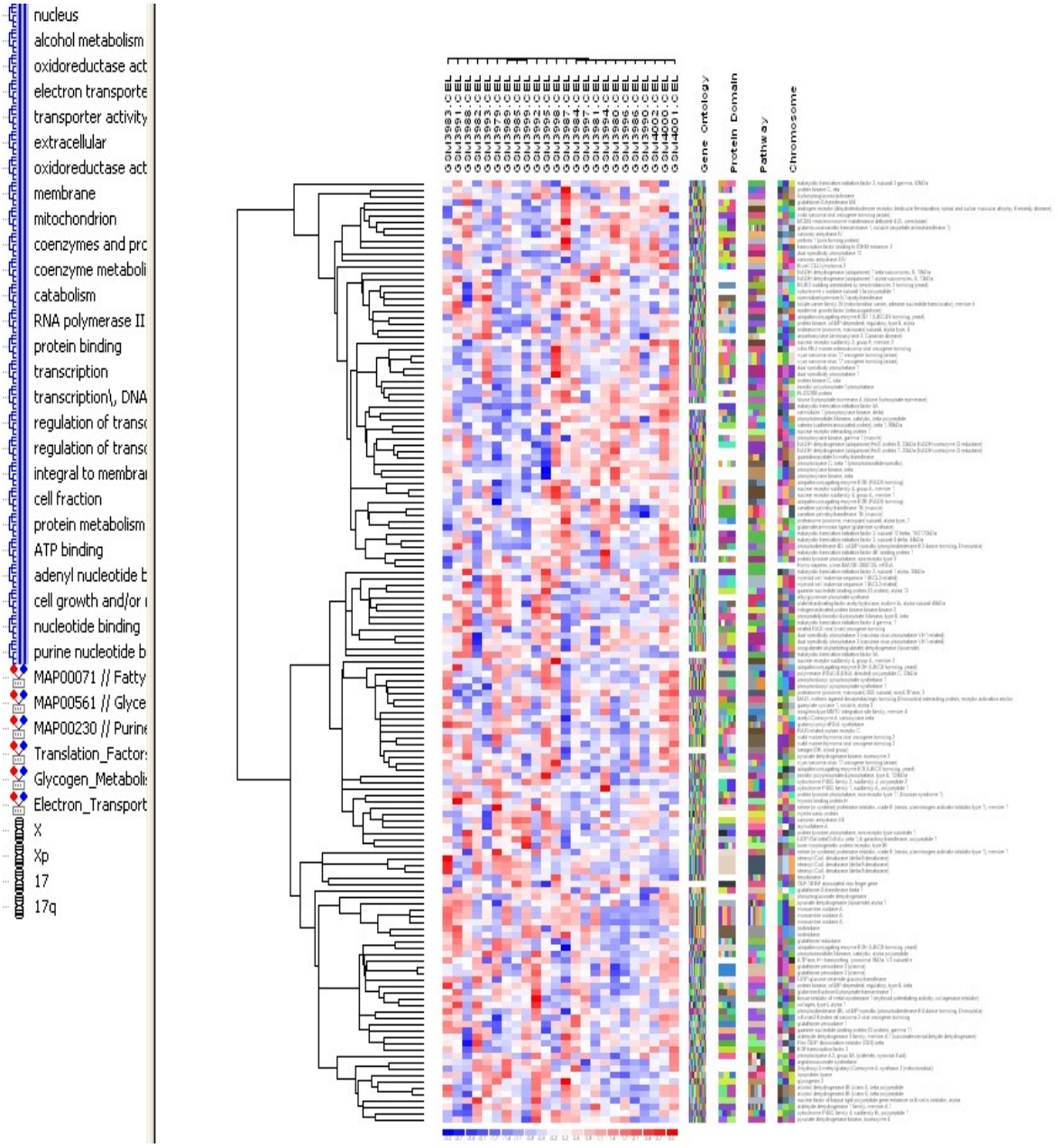


Fig 4.3 Hierarchical image from dChip

The data in the tabular format is also can be saved and carry out for the further analysis. Below is a table given **(Table 4.1)** for showing tabular format of disregulated genes.

1. DIFFERENTIALLY EXPRESSED GENE WITH CORRESPONDING PATHWAY ANALYSED IN dChip

Note: - XXX represent the gene is not differentially expressed.

Probe set	Gene Symbol	Fold change in morbidly obese sample	P value	Regulation	Fold change in obese sample	P value	Regulation
Proteasome Degradation							
201388_at	PSMD3	1.06	0.1046	**UP**	1.05	0.2757	**UP**
202334_s_at	UBE2B	-1.07	0.1046	**DOWN**	xxx	xxx	**xxx**
203396_at	PSMA4	xxx	xxx	**xxx**	-1.04	0.2757	**DOWN**
211763_s_at	UBE2B	-1.07	0.1046	**DOWN**	xxx	xxx	**xxx**
214590_s_at	UBE2D1	-1.08	0.1046	**DOWN**	xxx	xxx	**xxx**
216088_s_at	PSMA7	-1.06	0.1046	**DOWN**	-1.06	0.2757	**DOWN**
Phosphatidyl Insitol Signaling System							
201041_s_at	DUSP1	-1.11	0.1012	**DOWN**	-1.06	0.1276	**DOWN**
201044_x_at	DUSP1	-1.08	0.1012	**DOWN**	-1.06	0.1276	**DOWN**
201081_s_at	PIP5K2B	1.05	0.1012	**UP**	xxx	xxx	**xxx**
201537_s_at	DUSP3	xxx	xxx	**xxx**	1.05	0.1276	**UP**
201538_s_at	DUSP3	1.05	0.1012	**UP**	1.07	0.1276	**UP**
202794_at	INPP1	-1.06	0.1012	**DOWN**	xxx	xxx	**xxx**
202895_s_at	SIRPA	1.05	0.1012	**UP**	xxx	xxx	**xxx**
203997_at	PTPN3	-1.05	0.1012	**DOWN**	xxx	xxx	**xxx**
205868_s_at	PTPN11	1.06	0.1012	**UP**	xxx	xxx	**xxx**
205376_at	INPP4B	xxx	xxx	**xxx**	1.06	0.6601	**UP**
221563_at	DUSP10	xxx	xxx	**xxx**	-1.06	0.6601	**DOWN**
Fatty acid metabolism							
205073_at	CYP2J2	1.06	0.0965	**UP**	1.04	0.6601	**UP**
205749_at	CYP1A1	1.06	0.0965	**UP**	xxx	xxx	**xxx**
209613_s_at	ADH1B	-1.06	0.0965	**DOWN**	-1.04	0.6601	**DOWN**
210069_at	CPT1B	-1.05	0.0965	**DOWN**	xxx	xxx	**xxx**
210070_s_at	CPT1B	-1.05	0.0965	**DOWN**	xxx	xxx	**xxx**
210096_at	CYP4B1	-1.06	0.0965	**DOWN**	xxx	xxx	**xxx**
212224_at	ALDH1A1	1.08	0.0965	**UP**	xxx	xxx	**xxx**

Tryptophane metabolism

201282_at	OGDH	xxx	xxx	**xxx**	1.06	0.7613	**UP**
204388_s_at	MAOA	-1.1	0.1042	**DOWN**	xxx	xxx	**xxx**
204389_at	MAOA	-1.14	0.1042	**DOWN**	xxx	xxx	**xxx**
205073_at	CYP2J2	1.06	0.1042	**UP**	1.04	0.7613	**UP**
205749_at	CYP1A1	1.06	0.1042	**UP**	xxx	xxx	**xxx**
210096_at	CYP4B1	-1.06	0.1042	**DOWN**	xxx	xxx	**xxx**
212224_at	ALDH1A1	1.08	0.1042	**UP**	xxx	xxx	**xxx**
212741_at	MAOA	-1.09	0.1042	**DOWN**	xxx	xxx	**xxx**

Nuclear Receptor

202340_x_at	NR4A1	-1.1	0.4563	**DOWN**	xxx	xxx	**xxx**
204622_x_at	NR4A2	xxx	xxx	**xxx**	1.06	0.4529	**UP**
206419_at	RORC	1.06	0.4563	**UP**	1.08	0.4529	**UP**
209120_at	NR2F2	xxx	xxx	**xxx**	-1.05	0.4529	**DOWN**
211143_x_at	NR4A1	-1.07	0.4563	**DOWN**	xxx	xxx	**xxx**
211621_at	AR	-1.05	0.4563	**DOWN**	xxx	xxx	**xxx**

Apoptosis

200796_s_at	MCL1	1.11	0.3085	**UP**	1.09	0.0009	**UP**
200798_x_at	MCL1	1.06	0.3085	**UP**	1.05	0.0009	**UP**
201464_x_at	JUN	xxx	xxx	**xxx**	-1.04	0.0009	**DOWN**
201465_s_at	JUN	xxx	xxx	**xxx**	1.07	0.0009	**UP**
201466_s_at	JUN	xxx	xxx	**xxx**	-1.06	0.0009	**DOWN**
201502_s_at	NFKBIA	-1.06	0.3085	**DOWN**	-1.06	0.0009	**DOWN**
203685_at	BCL2	xxx	xxx	**xxx**	-1.05	0.0009	**DOWN**
214617_at	PRF1	-1.06	0.3085	**DOWN**	xxx	xxx	**xxx**

Translation Factor

200597_at	EIF3S10	-1.05	0.0167	**DOWN**	xxx	xxx	**xxx**
201123_s_at	**EIF5A**	**-1.24**	**0.0167**	**DOWN**	**1.1**	**0.1822**	**UP**
201143_s_at	EIF2S1	xxx	xxx	**xxx**	1.06	0.1822	**UP**
205321_at	**EIF2S3**	**1.06**	**0.0167**	**UP**	**-1.09**	**0.1822**	**DOWN**
208624_s_at	EIF4G1	1.09	0.0167	**UP**	1.07	0.1822	**UP**
208773_s_at	ANKHD1	-1.06	0.0167	**DOWN**	xxx	xxx	**xxx**
208887_at	EIF3S4	-1.04	0.0167	**DOWN**	xxx	xxx	**xxx**
213757_at	EIF5A	1.05	0.0167	**UP**	xxx	xxx	**xxx**
214919_s_at	EIF4EBP3	-1.05	0.0167	**DOWN**	xxx	xxx	**xxx**
221539_at	EIF4EBP1	-1.05	0.0167	**DOWN**	-1.04	0.1822	**DOWN**

G13 Signaling pathway

200655_s_at	CALM1	xxx	xxx	**xxx**	-1.04	0.1988	**DOWN**
201288_at	ARHGDIB	xxx	xxx	**xxx**	-1.06	0.1988	**DOWN**
204369_at	PIK3CA	1.07	0.3775	**UP**	xxx	xxx	**xxx**
206304_at	MYBPH	1.05	0.3775	**UP**	xxx	xxx	**xxx**
206917_at	GNA13	xxx	xxx	**xxx**	1.05	0.1988	**UP**
212688_at	PIK3CB	-1.07	0.3775	**DOWN**	xxx	xxx	**xxx**

Alanine Asparate metabolism

206030_at	ASPA	-1.06	0.033	**DOWN**	xxx	xxx	**xxx**
207076_s_at	ASS1	1.05	0.033	**UP**	1.06	0.4143	**UP**
208813_at	GOT1	-1.04	0.033	**DOWN**	xxx	xxx	**xxx**

G protein Signaling

200655_s_at	CALM1	xxx	xxx	**xxx**	-1.04	0.1613	**DOWN**
203680_at	PRKAR2B	xxx	xxx	**xxx**	-1.07	0.1613	**DOWN**
203708_at	PDE4B	1.05	0.434	**UP**	1.08	0.1613	**UP**
204115_at	GNG11	-1.08	0.434	**DOWN**	-1.08	0.1613	**DOWN**
204491_at	PDE4D	-1.04	0.434	**DOWN**	xxx	xxx	**xxx**
204842_x_at	PRKAR2A	-1.04	0.434	**DOWN**	xxx	xxx	**xxx**
206917_at	GNA13	xxx	xxx	**xxx**	1.05	0.1613	**UP**
212647_at	RRAS	xxx	xxx	**xxx**	1.05	0.1613	**UP**
213518_at	PRKCI	-1.06	0.434	**DOWN**	-1.07	0.1613	**DOWN**
214352_s_at	KRAS	-1.07	0.434	**DOWN**	xxx	xxx	**xxx**
218764_at	PRKCH	-1.08	0.434	**DOWN**	xxx	xxx	**xxx**

Wnt Signaling Pathway

201464_x_at	JUN	xxx	xxx	**xxx**	-1.04	0.1339	**DOWN**
201465_s_at	JUN	xxx	xxx	**xxx**	1.07	0.1339	**UP**
201466_s_at	JUN	xxx	xxx	**xxx**	-1.06	0.1339	**DOWN**
201533_at	CTNNB1	-1.05	0.3394	**DOWN**	-1.05	0.1339	**DOWN**
208606_s_at	WNT4	1.06	0.3394	**UP**	xxx	xxx	**xxx**
211547_s_at	PAFAH1B1	1.08	0.3394	**UP**	xxx	xxx	**xxx**
213518_at	PRKCI	-1.06	0.3394	**DOWN**	-1.07	0.1339	**DOWN**
218764_at	PRKCH	-1.08	0.3394	**DOWN**	xxx	xxx	**xxx**

ATP Synthesis

214149_s_at	ATP6V0E1	1.06	0.6121	**UP**	1.07	0.4926	**UP**

Electron Transport Chain gene list

200657_at	SLC25A5	-1.06	0.671	**DOWN**	xxx	xxx	**xxx**
200925_at	COX6A1	-1.05	0.671	**DOWN**	-1.05	0.5933	**DOWN**
201304_at	NDUFA5	xxx	xxx	**xxx**	-1.05	0.5933	**DOWN**
203189_s_at	NDUFS8	-1.05	0.671	**DOWN**	-1.05	0.5933	**DOWN**
211752_s_at	NDUFS7	-1.04	0.671	**DOWN**	-1.05	0.5933	**DOWN**

Glycogen Metabolism

200655_s_at	CALM1	xxx	xxx	**xxx**	-1.04	0.093	**DOWN**
202738_s_at	PHKB	xxx	xxx	**xxx**	-1.07	0.093	**DOWN**
207312_at	PHKG1	1.05	0.8655	**UP**	xxx	xxx	**xxx**
210964_s_at	GYG2	xxx	xxx	**xxx**	-1.05	0.093	**DOWN**

Glycerolopid metabolism

203548_s_at	LPL	1.05	0.0875	**UP**	1.05	0.2485	**UP**
203649_s_at	PLA2G2A	xxx	xxx	**xxx**	-1.06	0.2485	**DOWN**
205401_at	AGPS	1.06	0.0875	**UP**	1.05	0.2485	**UP**
209612_s_at	ADH1B	-1.06	0.0875	**DOWN**	-1.07	0.2485	**DOWN**
210069_at	CPT1B	-1.05	0.0875	**DOWN**	xxx	xxx	**xxx**
210070_s_at	CPT1B	-1.05	0.0875	**DOWN**	xxx	xxx	**xxx**
211547_s_at	PAFAH1B1	1.08	0.0875	**UP**	xxx	xxx	**xxx**
212224_at	ALDH1A1	1.08	0.0875	**UP**	xxx	xxx	**xxx**

Phospholipid degradation metabolism

203649_s_at	PLA2G2A	xxx	xxx	**xxx**	-1.06	0.4629	**DOWN**

Glutamate Metabolism

200841_s_at	EPRS	xxx	xxx	**xxx**	1.05	0.478	**UP**
205770_at	GSR	1.05	0.0631	**UP**	xxx	xxx	**xxx**
208813_at	GOT1	-1.04	0.0631	**DOWN**	xxx	xxx	**xxx**
215001_s_at	GLUL	-1.04	0.0631	**DOWN**	xxx	xxx	**xxx**
202722_s_at	GFPT1	1.09	0.0828	**UP**	xxx	xxx	**xxx**
203608_at	ALDH5A1	-1.07	0.0828	**DOWN**	-1.06	0.343	**DOWN**
208813_at	GOT1	-1.04	0.0828	**DOWN**	xxx	xxx	**xxx**

Nitrogen metabolism

203963_at	CA12	xxx	xxx	**xxx**	1.05	0.2282	**UP**
206208_at	CA4	xxx	xxx	**xxx**	-1.05	0.2282	**DOWN**
215001_s_at	GLUL	-1.04	0.3787	**DOWN**	xxx	xxx	**xxx**
219464_at	CA14	1.05	0.3787	**UP**	xxx	xxx	**xxx**

Glycolysis/Gluconeogenesis Pathway

200980_s_at	PDHA1	-1.04	0.3678	**DOWN**	xxx	xxx	**xxx**
202934_at	HK2	-1.07	0.3678	**DOWN**	xxx	xxx	**xxx**
209612_s_at	ADH1B	-1.06	0.3678	**DOWN**	-1.07	0.7512	**DOWN**
209613_s_at	ADH1B	-1.06	0.3678	**DOWN**	-1.04	0.7512	**DOWN**
212224_at	ALDH1A1	1.08	0.3678	**UP**	xxx	xxx	**xxx**

Pyruvate metabolism

200980_s_at	PDHA1	-1.04	0.5727	**DOWN**	xxx	xxx	**xxx**
212224_at	ALDH1A1	1.08	0.5727	**UP**	xxx	xxx	**xxx**

Starch and Sucrose metabolism

202934_at	HK2	-1.07	0.6257	**DOWN**	xxx	xxx	**xxx**

Tyrosine metabolism

204388_s_at	MAOA	-1.1	0.0038	**DOWN**	xxx	xxx	**xxx**
208813_at	GOT1	-1.04	0.0038	**DOWN**	xxx	xxx	**xxx**
209612_s_at	ADH1B	-1.06	0.0038	**DOWN**	-1.07	0.3673	**DOWN**
209613_s_at	ADH1B	-1.06	0.0038	**DOWN**	-1.04	0.3673	**DOWN**
212741_at	MAOA	-1.09	0.0038	**DOWN**	xxx	xxx	**xxx**

TGF Beta Signaling Pathway

201464_x_at	JUN	xxx	xxx	**xxx**	-1.04	0.0055	**DOWN**
201465_s_at	JUN	xxx	xxx	**xxx**	1.07	0.0055	**UP**
201466_s_at	JUN	xxx	xxx	**xxx**	-1.06	0.0055	**DOWN**
201533_at	CTNNB1	-1.05	0.0141	**DOWN**	-1.05	0.0055	**DOWN**
202627_s_at	SERPINE1	1.09	0.0141	**UP**	xxx	xxx	**xxx**
204893_s_at	ZFYVE9	1.11	0.0141	**UP**	xxx	xxx	**xxx**
206254_at	EGF	-1.05	0.0141	**DOWN**	xxx	xxx	**xxx**
206649_s_at	TFE3	-1.05	0.0141	**DOWN**	xxx	xxx	**xxx**
209189_at	FOS	-1.17	0.0141	**DOWN**	-1.17	0.0055	**DOWN**
204893_s_at	ZFYVE9	xxx	xxx	**xxx**	1.09	0.0055	**UP**
213755_s_at	SKI	-1.05	0.0141	**DOWN**	xxx	xxx	**xxx**
214761_at	ZNF423	xxx	xxx	**xxx**	-1.05	0.0055	**DOWN**

Fatty acid degradation

203548_s_at	LPL	1.05	0.3135	**UP**	1.05	0.7024	**UP**
210069_at	CPT1B	-1.05	0.3135	**DOWN**	xxx	xxx	**xxx**
210070_s_at	CPT1B	-1.05	0.3135	**DOWN**	xxx	xxx	**xxx**

Glutathione metabolism

200736_s_at	GPX1	-1.04	0.0336	**DOWN**	xxx	xxx	**xxx**
201348_at	GPX3	1.04	0.0336	**UP**	xxx	xxx	**xxx**
203815_at	GSTT1	1.05	0.0336	**UP**	xxx	xxx	**xxx**
210912_x_at	GSTM4	xxx	xxx	**xxx**	-1.05	0.5818	**DOWN**
214091_s_at	GPX3	1.07	0.0336	**UP**	xxx	xxx	**xxx**

Inflammatory Response Pathway

202310_s_at	COL1A1	1.06	0.8322	**UP**	xxx	xxx	**xxx**

Urea cycle and metabolism of amino group

205354_at	GAMT	-1.06	0.1905	**DOWN**	-1.05	0.0991	**DOWN**
207076_s_at	ASS1	1.05	0.1905	**UP**	1.06	0.0991	**UP**

Kreb TCA Cycle

200980_s_at	PDHA1	-1.04	0.5487	**DOWN**	xxx	xxx	**xxx**
201282_at	OGDH	xxx	xxx	**xxx**	1.06	0.3792	**UP**
205960_at	PDK4	-1.04	0.5487	**DOWN**	xxx	xxx	**xxx**
213724_s_at	PDK2	xxx	xxx	**xxx**	1.06	0.3792	**UP**

Fatty acid synthesis

200832_s_at	SCD	1.13	0.0026	**UP**	xxx	xxx	**xxx**
214584_x_at	ACACB	1.08	0.0026	**UP**	1.05	0.4143	**UP**

Cell Cycle

203693_s_at	E2F3	-1.06	0.9606	**DOWN**	xxx	xxx	**xxx**
209974_s_at	BUB3	xxx	xxx	**xxx**	-1.05	0.8971	**DOWN**
222036_s_at	MCM4	-1.08	0.9606	**DOWN**	-1.07	0.8971	**DOWN**

Ovarian infertility

202599_s_at	NRIP1	-1.05	0.803	**DOWN**	-1.07	0.4026	**DOWN**
210523_at	BMPR1B	xxx	xxx	**xxx**	1.07	0.4026	**UP**

Pentose phosphate pathway

201118_at	PGD	-1.05	0.2956	**DOWN**	xxx	xxx	**xxx**
208447_s_at	PRPS1	1.07	0.2956	**UP**	1.09	0.1634	**UP**
209440_at	PRPS1	xxx	xxx	**xxx**	1.04	0.1634	**UP**

Purine metabolism							
203708_at	PDE4B	1.05	0.9077	**UP**	1.08	0.7542	**UP**
204491_at	PDE4D	-1.04	0.9077	**DOWN**	xxx	xxx	**xxx**
208447_s_at	PRPS1	1.07	0.9077	**UP**	1.09	0.7542	**UP**
208996_s_at	POLR2C	xxx	xxx	**xxx**	1.04	0.7542	**UP**
209440_at	PRPS1	xxx	xxx	**xxx**	1.04	0.7542	**UP**
221942_s_at	GUCY1A3	1.05	0.9077	**UP**	xxx	xxx	**xxx**
S1P Signaling Pathway							
203809_s_at	AKT2	1.08	0.4404	**UP**	1.07	0.0728	**UP**
213222_at	PLCB1	xxx	xxx	**xxx**	-1.07	0.0728	**DOWN**
Insitol phosphate metabolism							
201081_s_at	PIP5K2B	1.05	0.0148	**UP**	xxx	xxx	**xxx**
202794_at	INPP1	-1.06	0.0148	**DOWN**	xxx	xxx	**xxx**
204369_at	PIK3CA	1.07	0.0148	**UP**	xxx	xxx	**xxx**
205376_at	INPP4B	xxx	xxx	**xxx**	1.06	0.5203	**UP**
212688_at	PIK3CB	-1.07	0.0148	**DOWN**	xxx	xxx	**xxx**
Pentose Phosphate pathway							
201118_at	PGD	-1.05	0.0001	**DOWN**	xxx	xxx	**xxx**
212973_at	RPIA	-1.04	0.0001	**DOWN**	xxx	xxx	**xxx**
218387_s_at	PGLS	-1.05	0.0001	**DOWN**	xxx	xxx	**xxx**
Biotin metabolism							
214117_s_at	BTD	-1.08	0.2531	**DOWN**	-1.07	0.0046	**DOWN**
Fatty acid biosynthesis (Path I)							
214584_x_at	ACACB	1.08	0.0114	**UP**	1.05	0.0029	**UP**
MAPK cascade							
201464_x_at	JUN	xxx	xxx	**xxx**	-1.04	6E-05	**DOWN**
201465_s_at	JUN	xxx	xxx	**xxx**	1.07	6E-05	**UP**
201466_s_at	JUN	xxx	xxx	**xxx**	-1.06	6E-05	**DOWN**
209072_at	MBP	xxx	xxx	**xxx**	1.05	6E-05	**UP**
212647_at	RRAS	xxx	xxx	**xxx**	1.05	6E-05	**UP**
221695_s_at	MAP3K2	xxx	xxx	**xxx**	1.06	6E-05	**UP**

Table 4.1 Result (The differentially expressed gene with related pathway)

Charts

(I) Bart Chart of number of differentially expressed genes in all pathways

Following is the bar chart of number of gene corresponding to each pathway.

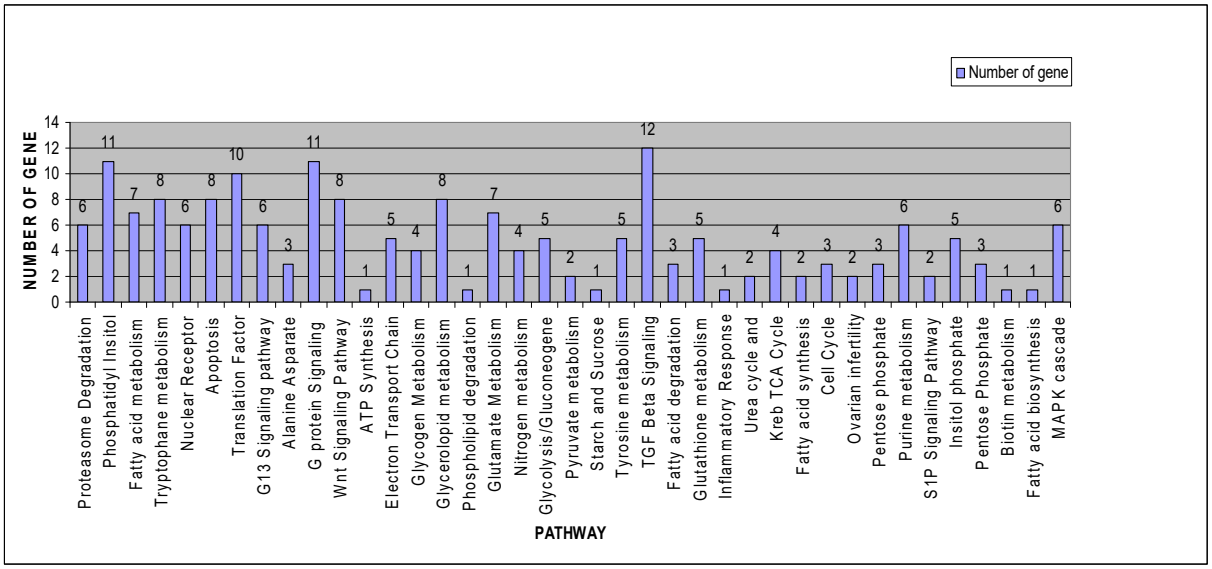


Chart 3.1 Number of gene vs. Pathway

This diagram **(Chart 3.1)** shows in TGF beta signaling has more number of differentially expressed genes i.e. 12 genes are differentially expressed.

(II) Venn-Diagram

We can make Venn-diagram of different pathways taking 3 at a time. Following is the example of a Venn-diagram with three pathways namely Wnt signaling pathway, Apoptosis and TGF beta signaling pathway. In this Venn-diagram I have taken gene as identifier. Thus the gene **JUN** expressed in all three pathways differentially **(Chart 3.2)**.

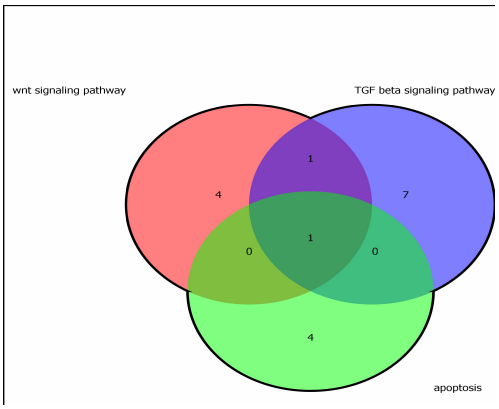


Chart 3.2 Venn-Diagram of pathways

Note: The above procedures are done in different softwares to get the robust result. There are softwares also available, where all the procedure can be done in single software as **ArrayAssist** with many good features but the main disadvantage is it is not freely available.

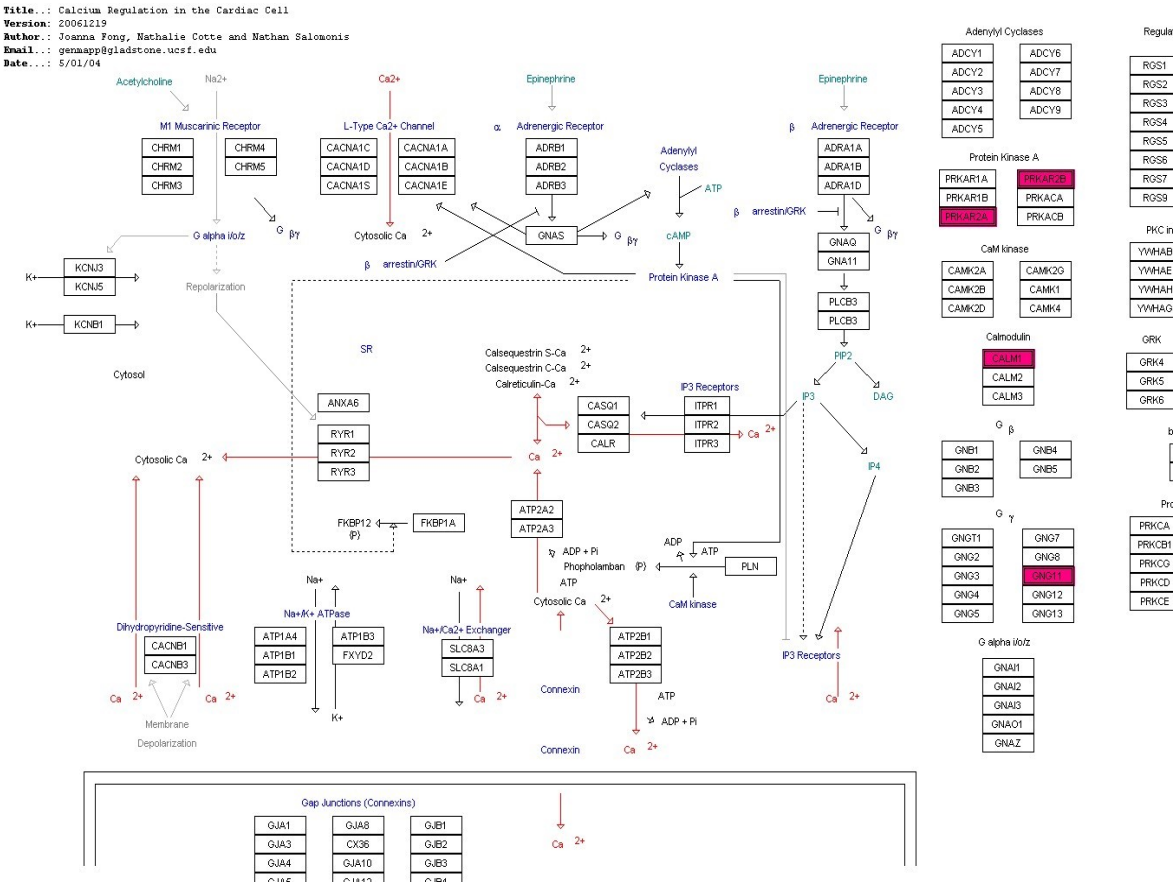
4.4 Result of Visualization of the pathways

There are so many types of pathways which may categorized into Pathway of physiological process, Pathway of cellular process and Pathway of metabolic process. Below given pathway diagram of each of these processes.

(I) Pathway of physiological process

1. Calcium regulation in cardiac cell

As we know this data is related with cardiac cycle so it is important to see the pathway of calcium regulation in cardiac cell **(Pathway 4.1).**

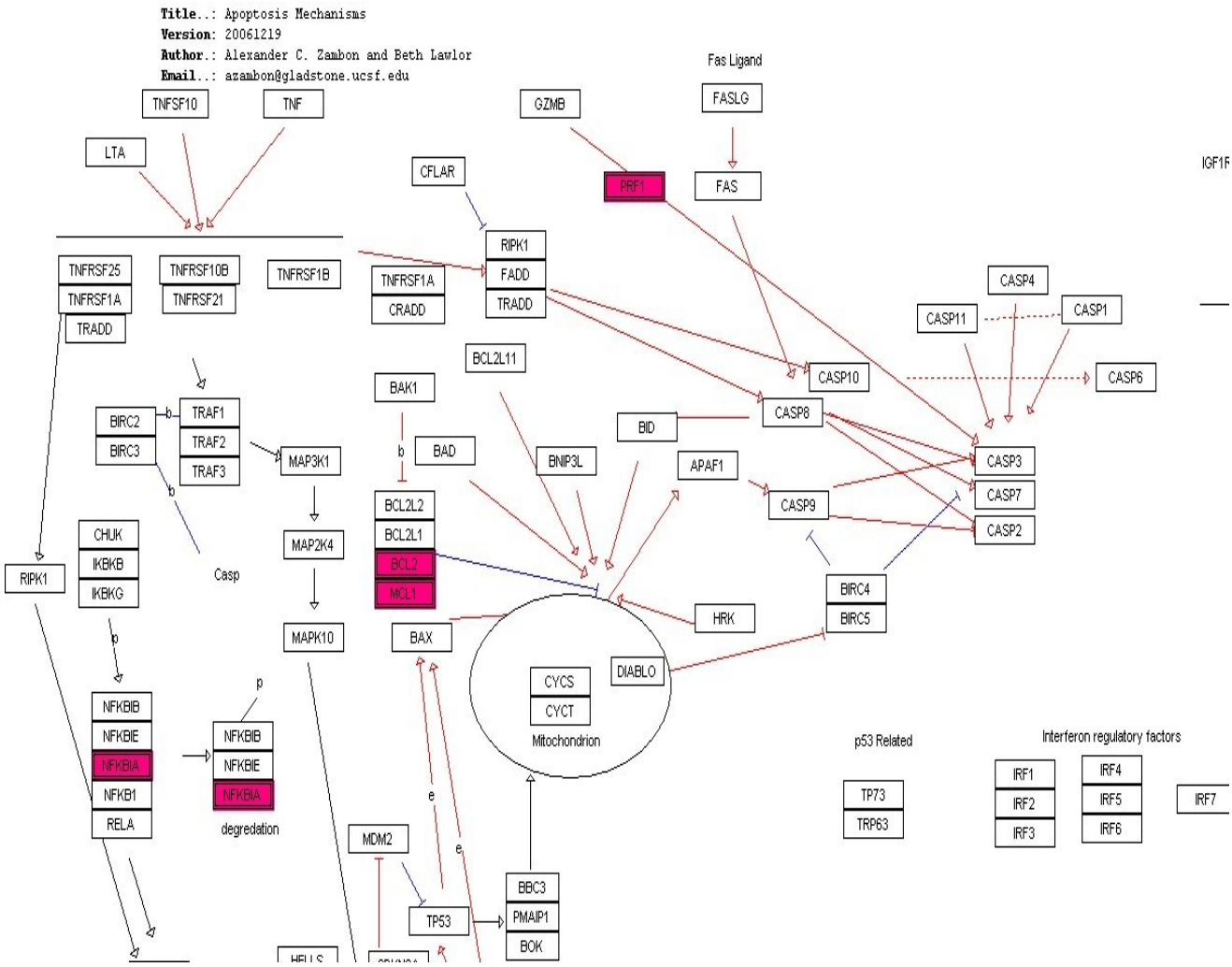


Pathway 4.1 Pathway map of Calcium regulation cardiac cell

The genes related in this pathway are calmodulin 1 (phosphorylase kinase, delta) (**CALM1**), guanine nucleotide binding protein (G protein), gamma 11 (**GNG11**), protein kinase, cAMP-dependent, regulatory, type II, alpha (**PRKAR2A**), protein kinase, cAMP-dependent, regulatory, type II, beta (**PRKAR2B)** and protein kinase C, eta (**PRKCH)** with 6 different ids. These genes are not only expressed in the human but in many cases expressed in mammals also resulting in the problem in different pathways and causing disease.

(II) Pathway of cellular process

1. Apoptosis

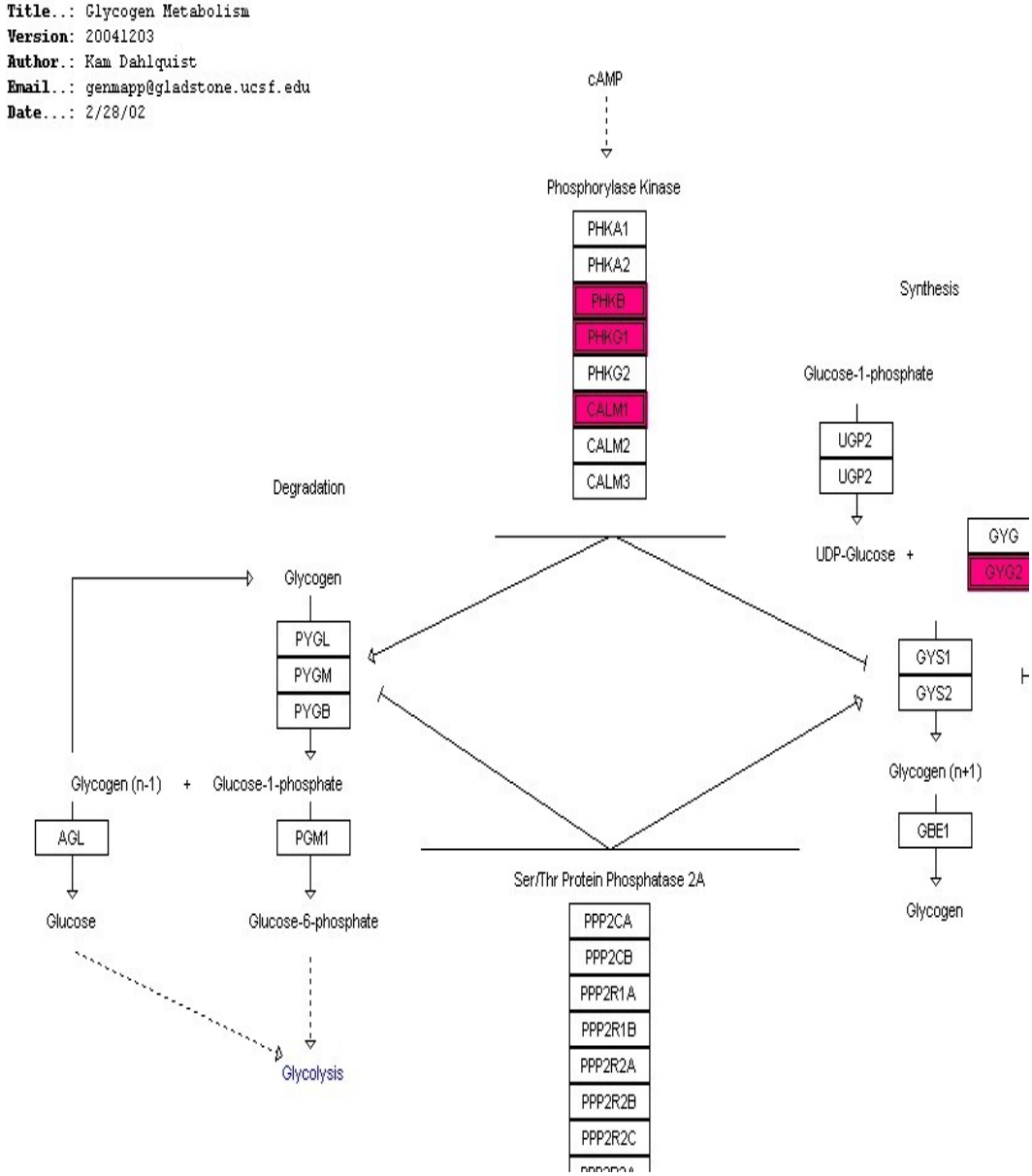


Pathway 4.2 Pathway map of Apoptosis

There are 5 genes related with the apoptosis **(Pathway 4.2)** having 8 different affymetrix ids. The genes are myeloid cell leukemia sequence 1 (BCL2-related) (**BCL2)**, v-jun sarcoma virus 17 oncogene homolog (avian) (**JUN)**, myeloid cell leukemia sequence 1 (BCL2-related) (**MCL1)**, nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha (**NFKBIA)**, perforin 1 (pore forming protein) (**PRF1)** with 8 different ids.

(III) Pathway of Metabolic process

1. Glycogen metabolism



Pathway 4.3 Pathway map of Glycogen metabolism

There are 4 genes with 5 different ids associate with modification of glycolysis pathway **(Pathway 4.3)** resulting Type 2 Diabetes (T2D). These are calmodulin 1 (phosphorylase kinase, delta) (**CALM1)**, glycogenin 2 (**GYG2)**, phosphorylase kinase, beta (**PHKB)** and phosphorylase kinase, gamma 1 (muscle) (**PHKG1)**.

4.5 Result from Network regulation

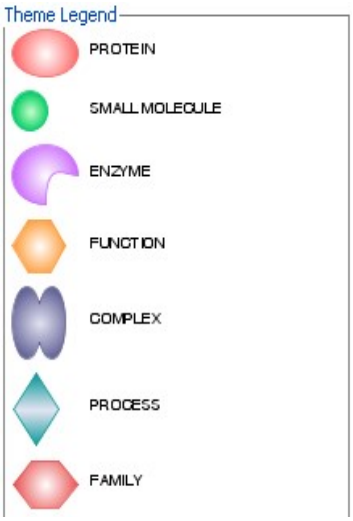


Fig 4.4 Legend

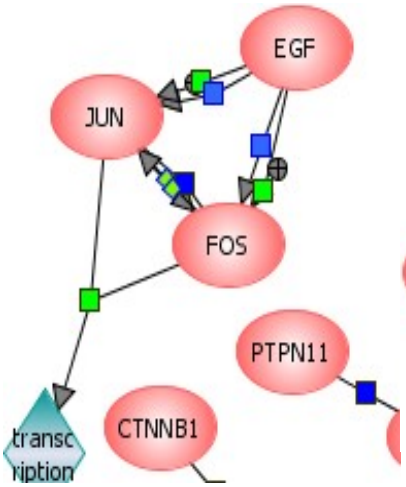


Fig 4.5 Direct interaction chain

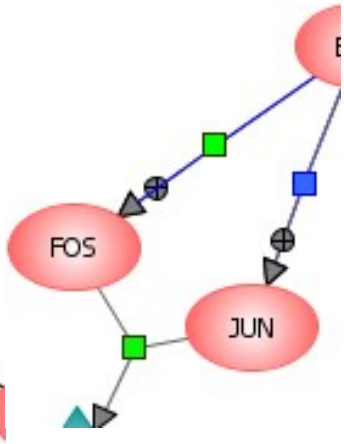


Fig 4.6 Shortest pathway (EGF, transcription)

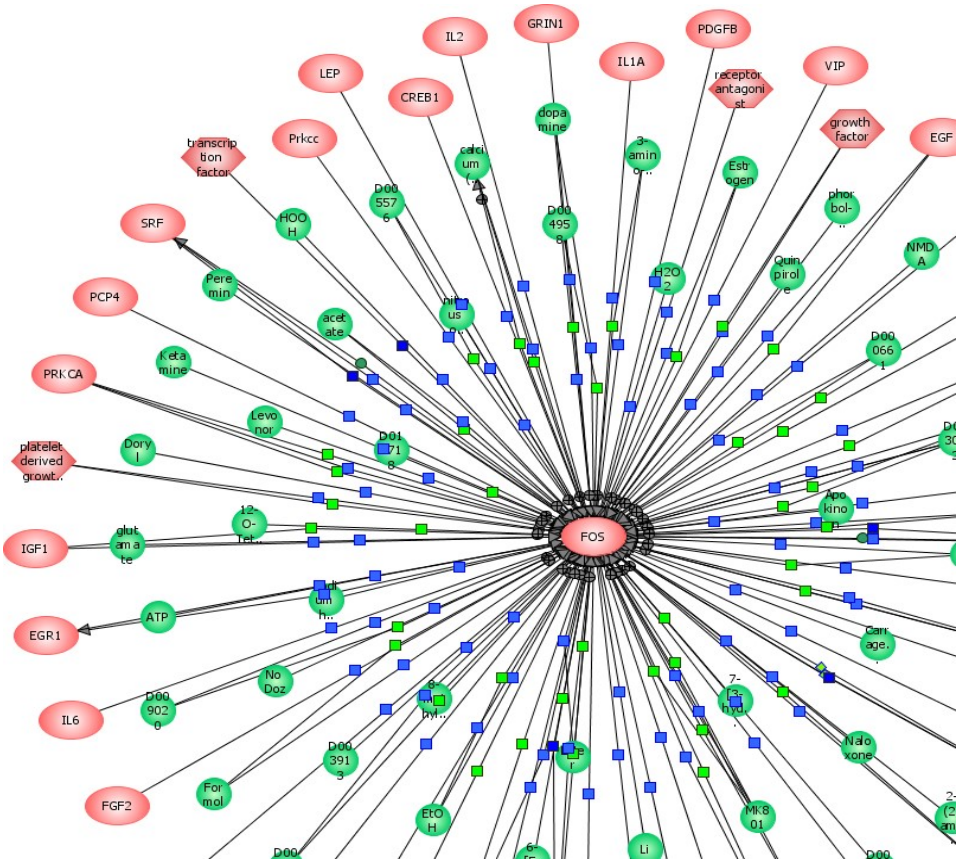


Fig 4.7 FOS expanded subnet

Fig 4.8 PTPN11 expanded subnet

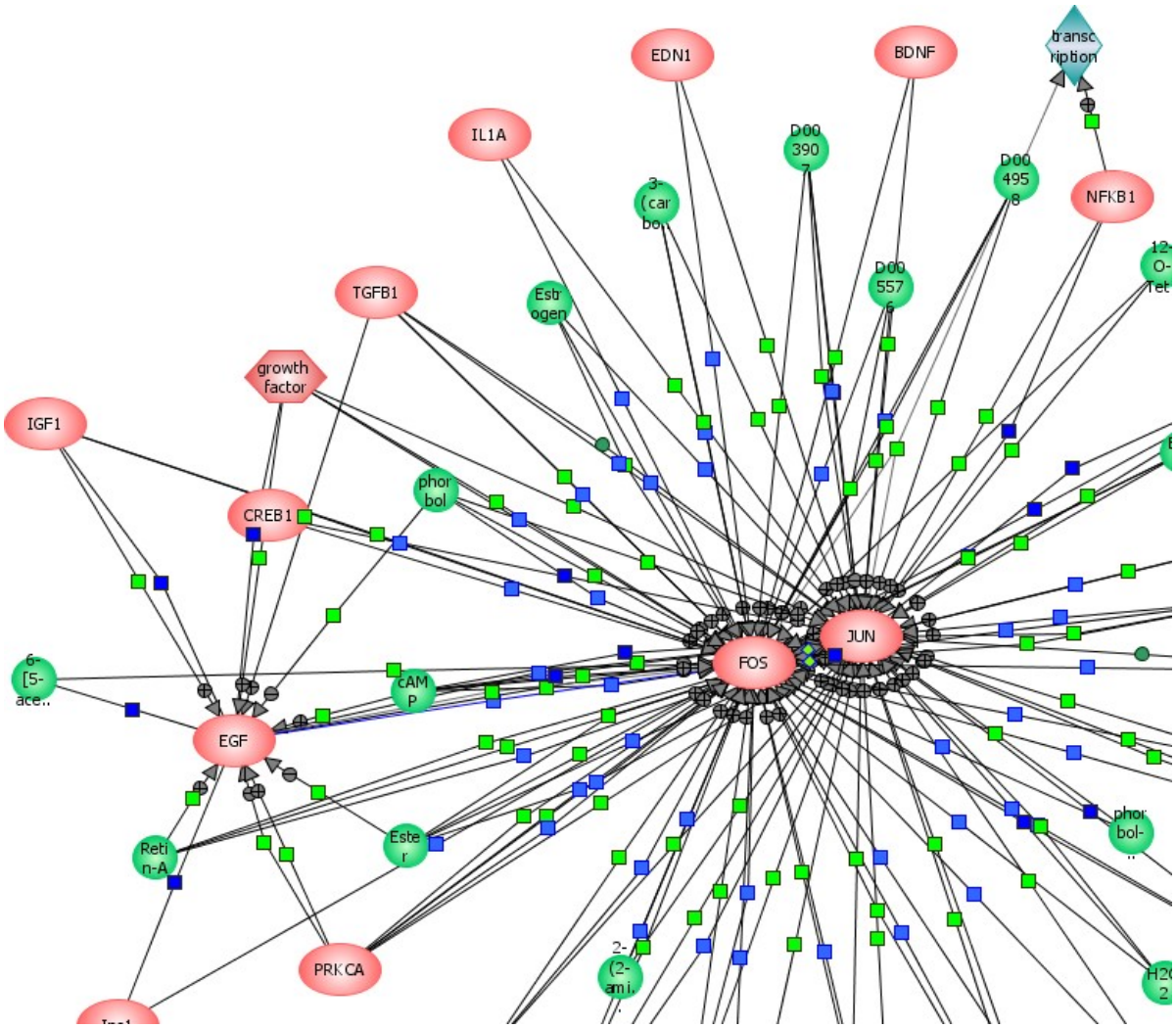


Fig 4.9 Network regulator

When the data is loaded and clicked the direct interaction the software gives the figure as Fig 4.5. These shows there are three different subnets. One combines JUN, FOS, EGF and transcription, second one with two protein i.e. PTPN11 and PTPNS1, third one also wit two protein CTNNB1 and AR **(Fig 4.5).** The shortest pathway between EGF and transcription involve all the four of first group **(Fig 4.6)**. Then each gene is individually expanded to see the other related genes which are related with their expression. Here expanded subnet of FOS **(Fig 4.7)** and PTPN11 **(Fig 4.8)** are shown. It includes many proteins, small molecules (In green color), Enzymes, Hormones, Function, Family etc. We then clicked on the network regulator by selecting all the genes. This shows us a diagram **(Fig 4.9)** how the genes are related to each other. Here we can find FOS, JUN, EGF acts as main network regulator. We can also find the similar pathway for these genes. These includes 3 proteins in SAPK-JNK Signaling, 2 proteins in TGF – Beta, 2 proteins in P53 Signaling, 2 proteins in G-protein-MAPK Activation and one protein in Wnt Signaling (Calcium).

4.6 Result from Promoter analysis

After running the k-means algorithm we got 3 clusters with 112 probes in first in first with homogeneity 0.189, 20 probes in second with homogeneity 0.26 and 16 probes in third set with 0.485 homogeneity values. The promoter binding site is noted from binding site window for each group of gene. As these are long sequences a part of it is represented here **(Fig 4.10)**.

Fig 4.10 Promoter sites

5. DISCUSSION

The genes which all are present in direct interaction selected from differentially expression are v-jun sarcoma virus 17 oncogene homolog (avian) (**JUN**), epidermal growth factor (beta-urogastrone) (**EGF**), v-fos FBJ murine osteosarcoma viral oncogene homolog (**FOS**), catenin (cadherin-associated protein), beta 1, 88kDa (**CTNNB1**), androgen receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulbar muscular atrophy; Kennedy disease) (**AR**), protein tyrosine phosphatase, non-receptor type 11 (Noonan syndrome 1) (**PTPN11**). All the above genes some how related with Type 2 Diabetes pathway or its related pathway. The description of each gene individually is given below.

1. V-jun sarcoma virus 17 oncogene homolog (avian) (JUN)

Description: It is a protein present only in mammal, located on chromosome 5, 5q34.1 (Rattus norvegicus), 4 44.6 cM (Mus musculus), chromosome 4, 4 44.6 cM, 4 C5-C7 (Mus musculus), **chromosome 1, 1p32-p31 (Homo sapiens).** This gene is the putative transforming gene of avian sarcoma virus 17. It encodes a protein which is highly similar to the viral protein, and which interacts directly with specific target DNA sequences to regulate gene expression. This gene is intron less and is mapped to 1p32-p31, a chromosomal region involved in both translocations and deletions in human malignancies. The cellular localization is Nucleus, Chromosome, Cell organelle. Its connectivity is 122. The UniGene id is Hs.525704. Entrez id is 3725. This gene is related with SAPK-JNK Signaling and G-protein-MAPK activation and grouped under ligand and Transcription factor.

GO Process: Leading edge cell differentiation, Cellular physiological process, Regulation of transcription, DNA-dependent, Regulation of progression through cell cycle, Transcription, Positive regulation of transcription from RNA polymerase II promoter.

2. Epidermal growth factor (beta-urogastrone) (EGF)

Description: It is a protein present only in mammal, located on **chromosome 4, 4q25 (Homo sapiens),** 3 65.2 cM (Mus musculus), chromosome 3, 3 65.2 cM, 3 G3 (Mus musculus), chromosome 2, 2q42-q43 (Rattus norvegicus). Epidermal growth factor has a profound effect on the differentiation of specific cells in vivo and is a potent mitogenic factor for a variety of cultured cells of both ectodermal and mesodermal origin. The EGF precursor is believed to exist as a membrane-bound molecule which is proteolytically cleaved to generate the 53-amino acid peptide hormone that stimulates cells to divide. The cellular localization is Plasma membrane, Nucleus, Membrane, Extracellular region and Cell Organelle. Its connectivity is 84. The UniGene id is Hs.419815. Entrez id is 1950.

GO Process: Regulation of protein secretion, Activation of MAPK activity, Branching morphogenesis, Peptidyl-tyrosine phosphorylation, Regulation of peptidyl-tyrosine phosphorylation, epidermal growth factor receptor signaling pathway, Protein kinase cascade, and Positive regulation of cell proliferation, DNA replication, Chromosome organization and biogenesis (sensu Eukaryota).

3. V-fos FBJ murine osteosarcoma viral oncogene homolog (FOS)

Description: It is a protein present only in mammal, located on **chromosome 14, 14q24.3 (Homo sapiens)**, 12 40.0 cM (Mus musculus), chromosome 12, 12 40.0 cM, 12 D2 (Mus musculus), chromosome 6, 6q31 (Rattus norvegicus). The Fos gene family consists of 4 members: FOS, FOSB, FOSL1, and FOSL2. These genes encode leucine zipper proteins that can dimerize with proteins of the JUN family, thereby forming the transcription factor complex AP-1. As such, the FOS proteins have been implicated as regulators of cell proliferation, differentiation, and transformation. In some cases, expression of the FOS gene has also been associated with apoptotic cell death. The cellular localization is Nucleus, Chromosome. Its connectivity is 122. The UniGene id is Hs.25647. Entrez id is 2353. This gene is related with G-protein-MAPK activation.

GO Process: Inflammatory response, Cellular physiological process, Regulation of transcription, DNA-dependent, DNA methylation, Regulation of transcription from RNA polymerase II promoter, Regulation of progression through cell cycle, Nervous system development.

4. Catenin (cadherin-associated protein), beta 1, 88kDa (CTNNB1)

Description: It is a protein present only in mammal, located on **chromosome 3, 3p21 (Homo sapiens),** chromosome 10, 10q24 (Rattus norvegicus), 9 72.0 cM (Mus musculus) chromosome 9, 9 72.0 cM, 9 F4 (Mus musculus). Beta-catenin is an adherens junction protein. Adherens junctions (AJs; also called the zonula adherens) are critical for the establishment and maintenance of epithelial layers, such as those lining organ surfaces. AJs mediate adhesion between cells, communicate a signal that neighboring cells are present, and anchor the actin cytoskeleton. In serving these roles, AJs regulate normal cell growth and behavior. At several stages of embryogenesis, wound healing, and tumor cell metastasis, cells form and leave epithelia. This process, which involves the disruption and reestablishment of epithelial cell-cell contacts, may be regulated by the disassembly and assembly of AJs. AJs may also function in the transmission of the 'contact inhibition' signal, which instructs cells to stop dividing once an epithelial sheet is complete. The cellular localization is Membrane, Cytoskeleton, Cytoplasm, Nucleus, Plasma membrane, Cell projection and Cell Organelle. The UniGene id is Hs.476018. Entrez id is 1499. Connectivity is 26. This gene is related with Wnt Signaling (Calcium).

GO Process: Embryonic digit morphogenesis, cell fate specification, cell-cell adhesion, regulation of cell differentiation, synaptic vesicle transport, cell maturation, negative regulation of transcription, DNA-dependent, osteoclast differentiation, embryonic hindlimb morphogenesis, cellular physiological process, forebrain development, Wnt

receptor signaling pathway, embryonic arm morphogenesis, negative regulation of cell differentiation, dorsal/ventral pattern formation, cell fate determination, regulation of cell proliferation, regulation of osteoblast differentiation, positive regulation of transcription, DNA-dependent, negative regulation of osteoclast differentiation, proximal/distal pattern formation, hemopoiesis, heart development, synaptic transmission, synapse organization and biogenesis, patterning of blood vessels, bone resorption, regulation of transcription, lung development, positive regulation of transcription, androgen receptor signaling pathway, positive regulation of transcription from RNA polymerase II promoter, gastrulation (sensu Mammalia), odontogenesis (sensu Vertebrata), regulation of transcription from RNA polymerase II promoter regulation of transcription, DNA-dependent, transcription, dorsal/ventral axis specification, positive regulation of osteoblast differentiation, cell adhesion, skeletal development, ectoderm development.

5. Androgen receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulbar muscular atrophy; Kennedy disease) (AR)

Description: It is a protein present only in mammal, located on chromosome X, X 36.0 cM, X C3 (Mus musculus), chromosome X, Xq22-q32 (Rattus norvegicus), **chromosome X, Xq11.2-q12 (Homo sapiens)**, X 36.0 cM (Mus musculus). The androgen receptor gene is more than 90 kb long and codes for a protein that has 3 major functional domains: the N-terminal domain, DNA-binding domain, and androgen-binding domain. The protein functions as a steroid-hormone activated transcription factor. Upon binding the hormone ligand, the receptor dissociates from accessory proteins, translocates into the nucleus, dimerizes, and then stimulates transcription of androgen responsive genes. This gene contains 2 polymorphic trinucleotide repeat segments that encode polyglutamine and polyglycine tracts in the N-terminal transactivation domain of its protein. Expansion of the polyglutamine tract causes spinal bulbar muscular atrophy (Kennedy disease). Mutations in this gene are also associated with complete androgen insensitivity (CAIS). Two alternatively spliced variants encoding distinct isoforms have been described. The cellular localization is Nucleus and cytoplasm. Its connectivity is 18. UniGene id is Hs.496240. Entrez id is 367. This gene is grouped under ligand, nuclear receptor and transcription factor.

GO process: Male gonad development, Transport, Male sex differentiation, Regulation of transcription, DNA-dependent, Cell growth, Transcription, Prostate gland development, Regulation of transcription, Cell-cell signaling, Cell proliferation, Signal transduction, Cellular physiological process, Embryonic development (sensu Mammalia), Androgen receptor signaling pathway, Sex differentiation

6. Protein tyrosine phosphatase, non-receptor type 11 (Noonan syndrome 1) (PTPN11)

Description: It is a protein present only in mammal, located on **chromosome 12, 12q24 (Homo sapiens),** chromosome 12, 12q16 (Rattus norvegicus), chromosome 5, 5 F (Mus musculus). The protein encoded by this gene is a member of the protein tyrosine phosphatase (PTP) family. PTPs are known to be signaling molecules that regulate a

variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation. This PTP contains two tandem Src homology-2 domains, which function as phospho-tyrosine binding domains and mediate the interaction of this PTP with its substrates. This PTP is widely expressed in most tissues and plays a regulatory role in various cell signaling events that are important for a diversity of cell functions, such as mitogenic activation, metabolic control, transcription regulation, and cell migration. Mutations in this gene are a cause of Noonan syndrome as well as acute myeloid leukemia. This gene is present in Plasma membrane, Cytoskeleton, Membrane, Cell Organelle. The UniGene id is Hs.506852. The Entrez id is 5781. Its connectivity is 5. This protein comes under ligands, phosphatases and receptors.

GO Process: protein amino acid dephosphorylation, activation of MAPK activity, DNA damage checkpoint, perception of sound, regulation of protein-nucleus export, signal transduction, axonogenesis, intracellular signaling cascade, regulation of protein export from nucleus, nerve growth factor receptor signaling pathway.

7. Protein tyrosine phosphatase, non-receptor type substrate 1 (PTPNS1)

Description: It is a protein present only in mammal, located on **chromosome 20, 20p13 (Homo sapiens)**, chromosome 3, 3q36 (Rattus norvegicus), 2 73.1 cM (Mus musculus), chromosome 2, 2 73.1 cM, 2 F3 (Mus musculus). The protein encoded by this gene is a member of the signal-regulatory-protein (SIRP) family, and also belongs to the immunoglobulin superfamily. SIRP family members are receptor-type transmembrane glycoproteins known to be involved in the negative regulation of receptor tyrosine kinase-coupled signaling processes. This protein can be phosphorylated by tyrosine kinases. The phospho-tyrosine residues of this PTP have been shown to recruit SH2 domain containing tyrosine phosphatases (PTP), and serve as substrates of PTPs. This protein was found to participate in signal transduction mediated by various growth factor receptors. CD47 has been demonstrated to be a ligand for this receptor protein. This gene and its product share very high similarity with several other members of the SIRP family. These related genes are located in close proximity to each other on chromosome 20p13. The UniGene id is Hs.128846. The Entrez id is 140885. Its connectivity is 3. This protein comes under ligands and Extracellular.

GO Process: phagocytosis, recognition phagocytosis, engulfment, cell migration, cell-matrix adhesion, actin filament organization, cytoskeleton organization and biogenesis, cell motility and positive regulation of phagocytosis.

6. SUMMERY

This project started with the aim to do microarray data analysis on Type 2 Diabetes data. The name diabetes takes us to the word Insulin which believes to main factor of Diabetes. Insulin is also called as "hunger hormone" secreted from pancreas of our body require to maintain the blood glucose level of the body. In absence or malfunction of it causes a disease called as diabetes, a metabolic disorder. It is of several type but people understand diabetes mean Type 2 Diabetes (T2D) or Non Insulin Dependent Diabetes Mellitus (NIDDM), a multifactor disease which is around 90%. There are some diagnosis and cure for it. But always prevention is better then cures. This disease is sixth leading cause of the death and one person in each five Indian is suffering from it. 14th November is declared as Diabetes day by World Health Organization.

Microarray data analysis plays a key role for a specific disease like Type 2 Diabetes (T2D). As this technique consider thousands of meaningful genes and do the analysis thus said to be a robust method for identifying differentially expressed gene.

The data analysis can be divided into some steps. The first step is Normalization which removes noise data. Most widely used method for normalization for Affymetrix data is by using Robust Multi-array Average algorithm. This does the background correction, normalization and summarization which result in producing less false negative data.

Then we can check the data type and go for the statistical analysis. In statistical analysis keeping in mind to data Analysis of Variance (ANOVA) method is applied which filter out others remaining significant data basing on the basis of p- value. The p-value is a value below which the data consider to be significant. In other word we can say low the p-value more the significant the data is.

The significance will be more meaningful when two different conditions will have differentially expressed gene. Very less number of genes comes under this test. These genes then correlated with the pathways for further study. Clustering done for finding the genes which are coregulated and which genes are different from other group.

An interaction study was done for further analysis. This results in an expanded knowledge about how one gene relates with the other and how one influences the activity of the other gene, the target molecule of one pathway. This may result in new pathways which have to validate by wet-lab technology.

In analysis method a step came after it is promoter analysis. This gives a clear idea about total sequence and the gene fragment (Exon), where the promoter attaches (binding site), and how one gene link to others. The literature study also had done basing on this result to validate the result.

7. REFERENCE

A. Journals

1. Dictionary definition of **disease** The American Heritage® Dictionary of the English Language, Fourth Edition Copyright © 2004, 2000 by Houghton Mifflin Company.

2. **Stanley R. Benedict, Emil Osterberg, and Isacc Neuwirth.** Studies in carbohydrate metabolism.

3. **Michael W. King,** Ph. D / IU School of Medicine /. Action of insulin in diet.

4. **Lepore M, Pampanelli S, Fanelli C, et al.** Pharmacokinetics and pharmacodynamics of subcutaneous injection of long-acting human insulin analog glargine, NPH insulin, and ultralente human insulin and continuous subcutaneous infusion of insulin lispro. Diabetes 2000; 49 (12):2142-8

5. **Diabetes project**

6. **Clayton D, McKeigue PM**. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 2001; 358: 1356-60.

7. **Bougnères P**. Genetics of obesity and T2DM. Tracking pathogenetic traits during the predisease period. Diabetes 2002; 51 (suppl 3): S295-S303.

8. **Haffner S M, Lehto S, Ronnemaa T, Pyorala K, Laakso M**,"Mortality from coronary heart disease in subjects with type 2 diabetes and in nondiabetic subjects with and without prior myocardial infarction", N. Engl. J. Med. (1998);339: pp. 229–234.

9. **Park JJ, Berggren JR, Hulver MW, Houmard JA et al.** GRB14, GPD1, and GDF8 as potential network collaborators in weight loss-induced improvements in insulin action in human skeletal muscle. *Physiol Genomics* 2006 Oct 11; 27 (2):114-21.

10. **World Diabetes Day 2006**

11. **Bharathi Raju and Philip E. Cryer.** Maintenance of the postabsorptive plasma glucose concentration: insulin or insulin plus glucagon?

12. Wiley Interscience Encyclopedia of Genetics, Genomics, Proteomics, and Bioinformatics, Part 3 Proteomics, **M. Dunn, ed., Section 3.8 Systems Biology, R. L. Winslow, ed., John Wiley & Sons,** Ltd. DOI: 10.1002/047001153X.g308213

13. **Bentley Cheatham and C. Ronald Khan**. Insulin Action and the Insulin Signaling Network

14. **National Diabetes Information Clearinghouse (NDIC)**

15. **International Diabetes Institute Diabetes Fact Sheet**

16. **Ramachandran, A., Snehalatha, C., Kapur, A., Vijay, V., Mohan, V., Das, A. K., Rao, P. V., Yajnik, C. S., Prasanna Kumar, K. M. & Nair, J. D.** (2001) High prevalence of diabetes and impaired glucose tolerance in India: National Urban Diabetes Survey. Diabetologia 9:1094-1101.

17. **Gupta, R. & Gupta, V. P.** (1996) Meta-analysis of coronary heart disease prevalence in India. Ind. Heart J. 48:241-245.[Medline]

18. **Reddy, K. S.** (1993) cardiovascular diseases in India. World Health Stat. Q. 46:101-107.[Medline]

19. **King, H. & Rewers, M.** (1998) Global burden of diabetes; 1995–2025: prevalence, numerical estimates and projections. Diab. Care 21:1414-1431.[Abstract].

20. **UK Prospective Diabetes Study Group,** 'Overview of Six Years' Therapy of Type 2 Diabetes: – A Progressive Disease (UKPDS 16)', *Diabetes*, Vol. 44, 1995, pp1249-58.

21. **Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R, Hughes JE, Snesrud E, Lee N, Quackenbush J** (2000) A concise guide to cDNA microarray analysis. *Biotechniques* 3, 548-554. [PUBMED]

22. **RMAExpress Manual.**
http://rmaexpress.bmbolstad.com/RMAExpress_UsersGuide.pdf

23. **MeV Manual.**
http://www.tm4.org/documentation/MeV_Manual_4_0.pdf

24. **Silva, P., (2002)** A general overview of the major metabolic pathways, Universidade Fernando Pessoa, http://www2.ufp.pt/~pedros/bq/integration.htm

25. Regulation of protein Synthesis, Natural Toxic Research Center, Texas A&M University – Kingsville, http://ntri.tamuk.edu/cell/regulation.html

26. **Baral, C.,** Signal Networks and Pathways: Computational Biology, http://www.public.asu.edu/~kkahol/images/Signal%20Networks%20and%20Pathways.pdf.

27. **Lodish, H., Berk, A., Zipursky, S. L.,** (2000) "Molecular Cell Biology", Palgrave Macmillian, ISBN: 071673706X

28. **KEGG,** Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/

29. **BIOCARTA,** Charting Pathways of life, www.biocarta.com.

30. **Dahlquist, K. D, Salomonis, N., Vranizan, K., Lawlor, S. C., Conklin, B. R., (2002),** "GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways", *Nat Genet.*, May, 31(1) pp: 19-20. (www.genmapp.org)

31. **PathwayAssist™**, http://www.ariadnegenomics.com/products/pathway.html

32. **Zhang, B., (2003)** "Pathway Editor: A Tool for Creating and Editing Biological Pathways Data", Masters Project, EECS Dept, CWRU, (http://nashua.cwru.edu/PathwaysRelease1/)

33. **Toyoda, T., Hirosawa, K., and Konagaya, A., (2003)**, "KnowledgeEditor: a new tool for interactive modeling and analyzing biological pathways based on microarray data" *Bioinformatics* 19(3):433-4. http://gscope.gsc.riken.go.jp/KEditor.exe.htm)

34. **YAO, D., Qu, K., Wang, J., Lu, Y., Noble, N., Sun, H., Zhu, X., Lin, N, Payan, D., and Li, M., (2004),** "PathwayFinder: paving the way towards automatic pathway extraction", Proceedings of the second conference on Asia-Pacific bioinformatics - v (29).

35. **PubGEne™**, http://www.pubgene.com/.

36. **Friedman, C., Kra, P., Yu, H; Krauthammer, M., Rzhetsky, A., (2001),** "GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles". Bioinformatics; 17 Suppl 1:S74-82.

37. **Omniviz™**, http://www.omniviz.com/applications/pathways.htm

38. **Zupan, B., Demsar, J., Bratko, I., Juvan, P., Halter, J., Kuspa, A., and Shaulsky, G., (2003),** "GenePath: a system for automated construction of genetic networks from mutant data", Bioinformatics. 12; 19 (3):383. (http://www.genepath.org/)

39. **Vector PathBlazer™,** Informax Inc. Solutions, http://register.informaxinc.com/solutions/pathblazer/

40**. Baderr, G., Donaldson, I., Wolting, C., Ouellette, B., Pawson, T., Hogue, C., (2001),** "Bind- the biomolecular interaction database", *Nucleic acids Research*, vol 29, no. 1, 242-245. (http://submit.bind.ca:8080/bind/).

41. **Edlund H**. Factors controlling pancreatic cell differentiation and function Diabetologia 2001

42. **Dinneen S, Gerich J, Rizza R.** Carbohydrate metabolism in non – insulin – dependent diabetes Mellitus. N Engl J Med 1992.**43)**.

43. **Bruning JC, Michael MD, Winnay JN, et a**l. A muscle – specefic insulin receptor knockout exhibits features of the metabolic syndrome of NIDDM with out altering glucose tolerance Mol cell 1998

44. **Kulkarni RN, Brunning JC, Winnay JN**. Tissue – specific knockout of the insulin receptor pancreatic beta cells creates an insulin secretory defect similar to that in type 2 diabetes. Cell 1999

45. **Accili D , Drago J , Lee EJ ,** Early neonatal death in mice homozygous for a null allele of the insulin receptor gene . Nature 1996

B. URLs

1. http://heartdiseasediabetes.suite101.com/article.cfm/diabetes.
2. www.americandiabetes.org
3. http://to-reverse-diabetes.blogspot.com/2007/04/malaise-symptom-or-side-effect.html
4. http://darwin.nmsu.edu/~molbio/diabetes/disease.html
5. http://www.patienthealthinternational.com/article/501584.aspx
6. http://adam.about.com/encyclopedia/Diabetes-risk-factors.htm
7. http://www.noble.org/medicago/GEP.html

C. Books

1. Microarray Bioinformatics (Dov Stekel)
2. DNA Microarray data analysis (Iiris Hovatta, Katja Kimppa, Antti Lehmussola, Tomi Pasanen, janna Saarela, Ilana Aaarikko, Juha Saharinen, Pekka Tiikainen, Teemu Toivanen, Martti Tolvanen, Mauno Vihinen and Garry Wong) Editors Jarono Tuimala and M. Minna Laine. CSC