R and Bioconductor Tools for Class Discovery Analysis: Example Analysis with Glioblastoma Multiforme (GBM) Data

Master's Thesis
Masters of science in bioinformatics
Faculty of Medicine and Life Sciences
University of Tampere, Finland
Samsudeen Sanusi
March 2017

MASTER'S THESIS

Place:	University of Tampere
	Faculty of Medicine and Life Sciences
Author:	Samsudeen Sanusi
Title:	R and Bioconductor Tools for Class Discovery Analysis:
	Example Analysis with Glioblastoma Multiforme (GBM) Data
Pages:	121
Supervisor:	Prof. Matti Nykter
Reviewer:	University Lecturer Juha Kesseli, Prof. Matti Nykter
Date:	March 2017

Abstract

Background The grade IV glioma tumor Glioblastoma multiforme (GBM) arises from a normal brain tissue, grows rapidly and is highly malignant. GBM causes an increase of blood vessels around the tumor containing dead cells. GBMs are common diagnosed in adults especially men aged between 45 and 56 years. Since the major cause of most brain tumor is unidentified, it is important to study the genes that play a role in glioblastoma development, hence the need for gene expression profiling. Gene expression profiling helps to identify/reveal molecular classes, which can never be detected by looking at GBM samples under the microscope (American Brain Tumor Association, 2016).

Aims: This study aims to use statistical methods on glioblastoma multiform (GBM) data obtained from microarray experiment to identify genes and samples subgroup, get differentially expressed genes, discover class membership, identify pathways, evaluate the aggressiveness of GBM sample groups and verify analysis outcome using an independent glioblastoma dataset.

Methods: The GBM data (Experimental data) used was obtained from The Cancer Genome Atlas (TCGA) via University of California Santa Cruz (UCSC) Genome browser as a zipped file that contains preprocessed data matrix, clinical data etc. The data matrix is used for class discovery, class comparison and class prediction analysis, while the clinical data is used for survival analysis. Samples subgroup, list of differentially expressed genes, sample membership, GBM aggressiveness prediction are results obtained from these analyses with the help of R and Biocondutor tools. In addition, biological interpretation is also one of the results obtained and it was done with PANTHER data analysis tool. PANTHER reveals the pathways and functional enrichment of the differentially expressed genes from experimental and validation datasets. Results from this study are validated with an independent GBM dataset (Validation data) obtained from Gene Expression Omnibus (GEO) by comparing the pathways and functional enrichment between the experimental dataset and the independent GBM dataset.

Results: A correlation matrix is obtained after filtering and scaling the gene expression data matrix. This correlation matrix is plotted on the heatmap to show the relationships that exist between samples, and on dendrogram to show how samples are clustered into groups. Class comparison analysis produces a new data matrix, which contains as rows the differentially expressed genes, obtained between the sample groups (two sample groups) identified with clustering analysis, which appear as the matrix column. The differentially expressed genes are obtained with fold change and t-test analysis. The top 50 differentially expressed genes from the new data matrix are interpreted with PANTHER in order to discover molecular functions, biological processes, cellular components and pathways in which these genes are active. The Kaplan-Mier survival curve obtained from GBM clinical is generated to evaluate survival difference between classes. Comparing molecular functions, biological processes, cellular components and pathways in which DEGs from both datasets play significant roles helps in validating analysis results.

Conclusion: Statistical methods are very important in analyzing microarray data since it gives insight into the data under study.

ACKNOWLEDGEMENTS

All praise and adoration to God, who is in infinite mercy, has showered his protection and guidance on me throughout my study in school.

I recognize and admire the role of my supervisor **Prof. Matti Nykter** in making this thesis a reality. You strongly stand by me in terms of dedication, understanding and consideration to make this thesis a success. I am grateful for the opportunity to do my thesis with the computational biology research group. I am also indebted to my study adviser, **Dr. Juha Kesseli,** thank you for the advice.

I acknowledge the impact of my parents **Mr. and Mrs. Sanusi**, for their moral, spiritual and financial supports, my brothers, sisters and fiancée for their motivation and support.

I also want to express my gratitude to the teachers, my course mates, relatives and friends; I appreciate you all for your support.

TABLE OF CONTENT

LIST OF ABBREVIATION

ANOVA	Analysis Of VAriance
BFGF	Basic Fibroblast Growth Factor
BPS	Base Pair
CDK	Cyclin-Dependent Kinase
cDNA	Complementary Deoxyribonucleic Acid
CT	Computed Tomography
DEG	Differentially expressed genes
DNA	Deoxyribonucleic Acid
ECM	Extracellular Matrix
EGFR	Epidermal Growth Factor Receptor
GBM	Glioblastoma multiforme
GEO	Gene Expression Omnibus
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
IGF	Insulin-like Growth Factor
KNN	K Nearest Neighbor
MEA	Modular Enrichment Analysis
MRI	Magnetic Resonance Imaging
mRNA	Messenger Ribonucleic Acid
MRS	Magnetic Resonance Spectroscopy
MSR	Mean-Square Residue
NCBI	National Center for Biotechnology Information
PANTHER	Protein ANalysis THrough Evolutionary Relationships
PDGFR	Platelet-Derived Growth Factor Receptor
RMA	Robust Microarray Average
SAM	Significance Analysis of Microarray
SEA	Singular Enrichment Analysis
SOM	Self-organizing maps

TCGA	The Cancer Genome Atlas
TGF	Transforming Growth Factor
TIFF	Tagged Image File Format
UniProt	Universal Protein Resource

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 BACKGROUND TO THE STUDY

Glioblastoma multiforme (GBM) is a grade IV glioma tumor, which arises from a normal brain tissue, grows rapidly and is highly malignant. There is an increase of blood vessels around GBM and it contains dead cells. GBMs are rare and most commonly occur in adults especially men aged between 45 to 65 years and very aggressive (American Brain Tumor Association, 2014). Been aggressive, it is important to predict the survival of patients with GBM.

Since the major cause of most brain tumor is unidentified, it is important to study the genes that play different roles in the development of glioblastoma. Hence, gene expression profiling is essential. In addition, molecular classes which can never be the detected by looking at GBM samples under the microscope has been revealed by gene expression profiling (American Brain Tumor Association, 2014).

In recent times, the advancement achieved in high-throughput microarrays has provided series of information relating to the biology behind glioma. Microarrays has helped to distinguish difference in the gene expression between normal and tumor (glioma) tissue (B.W. Kunkle et al 2013). Microarray experiment is a large-scale experiment that involves studying gene expression levels under a particular condition for thousand of genes concurrently. The process above is called gene expression analysis and sometimes referred to as gene expression profiling. Microarray technology has become a vey important tools employed by biologists to study an organisms genome wide expression levels of gene. It should however be noted that microarray measures the expression of genes in many ways but it most popular application is to compare gene expression levels maintained between two conditions such as healthy versus normal (M.M. Babu, 2004).

To achieve an outstanding results from microarray experiment we need to properly plan. A good plan will come in place if there are clear objectives, which must be able to answer biomedical questions before proceeding with the experiment. Hence, microarray

experiment objectives can be class comparism, class prediction and or class discovery. Class comparism involves identifying differentially expressed genes among predefined group of samples. Class prediction involves accurately predicting the biologic group a sample from a patient belongs based on the patient's gene expression profile with the help of a classifier. Class discovery involves discovering samples or gene groups that are similar based on their gene expression profile (R. Simon, 2003).

2. LITERATURE REVIEW

2.1 GLIOBLASTOMA MULTIFORME (GBM)

Glioma is a tumor that emanates from glial or the supportive tissue of the brain. An example of glioma is astrocytoma. Astrocytoma grows from astrocytes, which are star-shaped cells (American Brain Tumor Association, 2014). In addition, astrocytoma is the most common glioma with glioblastoma multiforme been their most threatening form. Researches conducted recently have helped to identify the basic biology behind GBM, though the major structure of the development of GBM remains unidentified. It has been difficult to understand the elementary molecular structure of pathophysiology that stimulates astrocytoma development and this has thus prevents the breakthrough that ought to have been achieved in astrocytoma treatment (B.W. Kunkle et al 2013).

Glioblastoma multiforme (GBM) is a grade IV glioma tumor, which arises from a normal brain tissue, grows rapidly, highly malignant (American Brain Tumor Association, 2014) and destructive (C.V. Neubeck et al 2015). Glioblastoma multiforme is the aggregation of tumors that emerges from glia or their precursors inside the central nervous system. Gliomas are of four grades with glioblastoma multiforme (GBM) been the most threatening of the gliomas. Most of the patients with GMB find it difficult to survive the disease in a period beyond one year and importantly with short survival. This has thus made these tumors popular (E.C. Holland 2000).

It is evident that lineage restricted progenitor and cells neural stem cells functions as the source of GBM or glioma-initiating cells (C.V. Neubeck et al 2015). It is a harsh brain cancer type defined by immense potential for growth and awful clinical result. It is an incurable form of cancer with a median survival of less than a year (M. Henriksen et al 2014). Since GBM occurs in different forms, necrosis and hemorrhage regions are excessively displayed. Also pleomorphic nuclei and cells, pseudopalisading necrosis and microvascular proliferation regions can be observed under the microscope. In addition, deletions, point mutations and amplification are also visible genetically in GBM, which leads to the amplification of signal transduction pathway activation downstream of platelet-derived

growth factor receptor (PDGFR) and epidermal growth factor receptor (EGFR) and cell cycle arrest pathway disruption by p53 mutations or INK4a-ARF loss (E.C. Holland 2000).

2.1.1 OCCURRENCE OF GBM

Glioblastoma multiforme (GBM) is a primary brain tumor that poses danger to life. It exists in 3 to 4 grown-up patients per 100,000 people living in Europe. GBM occurs mostly in adults aged 50 and above and it affects male than female. In addition, close to 9% of brain tumors that affect children are GBM (X. Xu et al 2011).

GBM is subdivided into primary and secondary. Primary GBM is also known as de novo GBM, they emerge quickly and their presence is easy to detect. In addition, they are the most common and aggressive form of GBM and occur in individual aged 55 and above (American Brain Tumor Association, 2014). In primary GBM, about 40 to 60% of these GBMs are characterized by overexpression and genetic mutation of epidermal growth factor receptor (EGFR) and its gene, which consecutively result in mutated, form of EGFR (M. Henriksen et al 2014). Secondary GBM on the other hand arises as low-grade astrocytoma, which will later develop into dangerous and actively growing glioblastoma (American Brain Tumor Association, 2014). It is characterized by continuous accumulation of mutations in p53, growth factor derived from platelets, and retinoblastoma gene (M. Henriksen et al 2014).

Furthermore, the larger parts of GBM emerge from primary glioblastoma while the remaining parts originate from lower grade astrocytoma. The molecular genetic differences between the benign minor astrocytoma (Grade I - II) and malignant major astrocytoma need to be determined so that characterization of these tumors can be achieved easily. Also, the genes and pathways involved in these tumor classes need to be determine so that better treatments could be provided for in years to come (B.W. Kunkle et al 2013).

2.1.2 CAUSE AND SYMPTONS GBM

The major cause of GBM and other forms of brain tumor is still unidentified. In addition, most brain tumors are not hereditary but some can sometimes be induced by syndrome such as Neurofibromatosis, Li-Frameni, Von Hippel-Lindau, Turcot and Tuberous Sclerosis that are inherited genetically. The genetically inherited syndromes are present in small number of patients (say 5% or lower) having this tumor. In recent time, scientists have traced the major causes of glioblastoma to deformities in genes of various chromosomes, which may be involved in the tumor development. Although, the major cause of the deformities that occur in genes remains unclear (American Brain Tumor Association, 2014).

The growth of this tumor in the brain causes a disruption of the brains normal function due to the fact that the skull size cannot be increased. This thus leads to increase pressure on brain, headache, seizures, loss of memory and behavioral changes. In addition, there is also loss in movement, cognitive deterioration and language dysfunction. Other symptoms are also possible depending on the size and location of the tumor in the brain (American Brain Tumor Association, 2014).

2.1.3 BIOLOGY BEHIND GBM

Over the last 20 years, knowledge of the molecular biology behind GBM has greatly advanced. Cells in GBM are subjected to different kind form cell deterioration, which makes them resistance to anti-GBM treatments. Intracellular events occur alongside tumor forming events and both together cause and sustain GBM (M. Nakada et al 2011).

The first of these events is cell cycle control loss. Glioma cells create a means for escaping the strict control that regulates normal cell progression. This thus makes glioma cells to gain growth benefit, which in turn leads to genetic defect in growth regulatory molecules within the glioma cells. These defects are common in malignant gliomas compared to low-grade gliomas. Alteration of at least one component of p16INK4a/cyclin-dependent kinase (CDK)-4/RB (retinoblastoma) 1 pathway, which is major pathways controlling G1–S phase

transition cell cycle checkpoint occur in several anaplastic astrocytomas and in the unlimited cases of GBMs (M. Nakada et al 2011).

The second event that cause and sustain GBM is the overexpression of growth factors and their receptors. Glioma cells set up an autocrine growth-promoting loop because they express both growth factor ligands and their receptors. Epidermal growth factor receptor (EGFR), platelet-derived growth factor (PDGF), basic fibroblast growth factor (bFGF, FGF-2), transforming growth factor (TGF)-α, and insulin-like growth factor (IGF)-1 are the growth factors that are overexpressed in GBM and they thus gives merits to neoplastic cells. Among the growth factors overexpressed in GBM, EGFR and PDGF are the most common (M. Nakada et al 2011).

Another event that play role in the cause and sustenance of GBM is angiogenesis. In GBM progression, sequence of angiogenic alterations appears like a ring-like contrast enhancement around the rapidly growing tumor. These alterations are seen to surround he tumor when viewed Magnetic Resonance Imaging (MRI) scan. Angiogenic molecules are present in malignant glioma especially GBMs. This is because malignant gliomas are vascular tumors of high grades as a result microvascular proliferation (M. Nakada et al 2011).

Invasion and migration is also a key feature that cause and sustain GBMs. Invasion and migration are influenced by expression of many extracellular matrix (ECM) molecules and cell surface receptors and this causes GBM to diffuse into their surrounding neural net (M. Nakada et al 2011).

Another event that create and sustain GBMs is abnormality in apoptosis. Apoptosis also known as cell death is usually characterized by non-inflammatory cellular condensation and it takes place in a programmed manner. Cells in glioma develop means for increased reproduction and to annul apoptosis. Mutation in p53 affects apoptotic response in normal glial, which normally accompany growth factor overexpression in low-grade glioma (M. Nakada et al 2011).

Genomic instability also plays a vital role in the cause and sustenance of GBM. Low-grade glioma progress to high-grade lesions quickly and such progression is related to the development of malignant clones. Many malignant clones are selected when genomic dame occurs as a result of genomic instability. Genomic instability causes tumor progression in genome with mutation in p53 genes (M. Nakada et al 2011).

2.1.4 PROGNOSIS AND TREATMENT OF GBM

Before the tumor can be treated, a proper diagnosis needs to be conducted on patients suspected of having this tumor. Diagnosis is commenced by first carrying out a neurological examination on the patient followed by a Magnetic Resonance Imaging (MRI), Computed Tomography (CT) or Magnetic Resonance Spectroscopy (MRS) scan. This scan thus helps to determine the location, size, tumor type and mineral and chemical level in the tumor, which in turn revealed whether the tumor is benign or malignant and also to know whether the patient has tumor or not (American Brain Tumor Association, 2014).

GBM is a tumor with complex characteristics, among which is the presence of subclones within the tumor cell population that makes the tumor to be genetically heterogeneous. Their complex characteristic has made them resistant to treatment interventions. For years, the conventional method of treating GBM has remained unchanged. Firstly a surgery is carried out on the patient in order to get rid of tumor, secondly is radiation therapy and lastly is chemotherapy. In most conditions, the mean survival of the patients with this disease only increases with just 9 to 10 months even after all the tumor seen on MRI scan has been surgically removed and the patients are fully treated with radiation and chemotherapy. This is so because the disease is diffuse topographically making the tumor and its location variable, which leads in improper resection of the tumor (E.C. Holland 2000). It is impossible to have a gross total resection without neurological and functional impairments such as motoric disorders, which thus have an adverse effect on quality of life (C.V. Neubeck et al 2015).

Before the last 10 years, the outcome of patents with GBM has been slightly improved despite the improvement in technology achieved in surgery, radiotherapy and also in

chemotherapy development. Despite applying intense treatments applied to GBM, it shows resistance to multimodal therapy and its survival time is still reported in months. The current method employed in treating GBM was proposed by the European Organization for Research and Treatment of Cancer (EORTC) and National Cancer Institute of Canada Clinical Trials Group (NCIC). This method involves surgery carried out to remove all tumors (debulking surgery) accompanied by fractionated radiotherapy alongside concomitant and adjuvant treatment of the cytostatic agent temozolomide (TMZ). This method has thus increased the median and the 2 years survival of patients with this disease to 14.6 months and 26.5% compared to the median and 2 years survival of patients treated with only radiotherapy, which is 12.1 months and 10.4% (C.V. Neubeck et al 2015).

2.2 RNA SEQUENCING VERSUS MICROARRAY IN GENE EXPRESSION PROFILING

An organism's transcriptome defines the whole range of transcripts present in that organism. The genes encode these transcripts as a phenotypic reply to the condition in which they occur. The ability to quantify the expression of thousand of genes at the same time has changed the face of biomedical research and this enables the analysis of gene expression pattern at a genome-wide scale. In the past 10 years, there has been a huge advancement in the improvement of methods used for analyzing and quantifying the expression level of gene transcriptome wise. RNA-seq and DNA microarray are exceptional and are the two most commonly used method for genome-wide gene expression quantification among the transcriptome profiling methods (S. Kogenaru et al 2012).

In RNA-seq, isolated transcripts are first converted into complementary DNA (cDNA). The resulting cDNA are then sequenced with a massive deep-sequencing approach. The gene expression levels in relation to the condition of interest or absolute level are quantified by mapping the resulting short sequencing reads to the reference genome. RNA-seq can be carried out on different platforms such as Illumina's Genome Analyzer, Roche 454 Genome Sequence, and Applied Biosystems' SOLiD. On the contrary, specimen target strands are hybridized unto the fastened complementary probe strands in microarray. In a two-color microarray, the extracted transcript from different conditions are labeled with specific

fluorescent dyes and converted to cDNA. The labeled transcripts are however hybridized to the fastened complementary probe strands in an array depicting the genes. The relative abundance of each transcript in the two different conditions is determined by measuring the intensity of light from the specific fluorescent dye. Microarray can be carried out on Affymetrix and Agilent platforms, which are the two common microarray technology platforms (S. Kogenaru et al 2012).

Recent literatures have compared two of the most widely used gene expression profiling method namely, gene expression microarray and RNA-seq documenting the utility and reproducibility of these methods. It is important to understand gene expression control as this makes understanding the relationship that exists between phenotype and genotype possible. Scientist developed DNA microarray technology with the aim of assessing transcript abundance in biological sample reliably. However, RNA-seq gives a more accurate measurement and absolute transcript abundance (K.J Mantione et al 2014).

Over the past 10 years, the analysis of microarray data has become easier for a beginner. The software packages used in analyzing microarray data are user friendly with most of them available for free. The protocols in each of these packages are universally applicable and can be compared across platforms. On the contrary, there are many data analysis method available in RNA-seq each with different protocols. The analysis of data from RNA-seq requires broad experience and bioinformatics skills required in processing the data files. The techniques used in analyzing data from RNA-seq differ in both the software employed for data transformation and different RNA-seq experiments. Data sharing and storage in RNA-seq is extremely difficult because an unprocessed RNA-Seq raw file is approximately 5GB compare to microarray with an unprocessed raw data of 0.7MB (K.J Mantione et al 2014).

In conclusion, the complex nature of data from RNA-seq will be reduced due to advancement in software and invention of newer techniques. Also the cost of running RNA-sequencing experiment will also drop with time. However, presently, it is more dependable

and cheaper to use microarray for gene-expression profiling in model organism compared with RNA-seq. In addition, gene expression microarray quickly and easily gives unique, useful and hidden information from examining the gene expression patterns across large number of samples. Microarrays has been used for clinical application for a longer period and might obtain regulatory approval for diagnostics purposes before RNA-seq gets its approval. In the future, RNA-seq will replace microarray but presently both techniques can complement each other (K.J Mantione et al 2014).

2.2.1 MICROARRAY

Microarrays give a good approach in understanding gene expression analysis that can be used for different experimental purposes. It has thus helped researchers to perform experiments that are impossible few years back, gives distinctive challenges in data analysis and experimental design (S. A. Ness 2006).

Microarrays are devices that identify and quantify the amount of mRNA transcript available in a cell (A. Sánchez and M. C. Ruíz de Villa 2008). They are usually made of glass slide, silicon chip or nylon membrane on to which DNA molecules are settled in a precise way at particular areas called spots. They may contain huge number of spots and every spot in turn contains a couple of million duplicates of indistinguishable genomic DNA molecule that compare to a gene (M.M. Babu 2004). These spots could also contain cDNAs, PCR products or chemically synthesized oligonucleotide that interestingly compare to a gene.
Spots in microarrays are imprinted on to the glass slide by a robot or are blended by the procedure of photolithography and are called probes (London school of Hygiene and Tropical medicine 2016).

Furthermore, microarrays plays roles in events like gene transcription, protein coding, mutation detection, copy number variation, and DNA methylation by measuring them and identifying where such event occurs in the human genome (C. Seidel 2008). It examines at the same time the expressions levels of more that hundreds of genes, gene relationship,

functions of genes and gene/samples classification in reply to some biological disorder (E. Naghieh and Y. Peng 2006).

2.2.1.1 MICROARRAY EXPERIMENT

Microarray experiments are very robust and it equips researchers with new and old methods of solving problems pan-genomic. Since microarrays contains probes for thousands of different genes, researchers can assess changes in all the genes in the genome simultaneously. However this experiment is very expensive, consumes time, complicated, and gives large and complex dataset that requires a lot of effort to analyze and validate. Hence, microarray experiment should not be performed without the researcher considering other options and checking for the right experimental design. New microarray users are however advised to consult with their domestic microarray main facility before preparing samples for the microarray experiment (S.A. Ness 2006). The steps involved in microarray experiment are as follow

A. TARGET EXTRACTION

mRNA is extracted from cells or tissues grown in two condition such that A is the reference condition, B is the test condition (M.M. Babu 2004). The mRNA in this case is called a target (A.D. Tarca et al. 2006, D.P. Berrar et al. 2003).

B. TRANSCRIBTION AND LABELLING

The mRNA molecules in the extract are transcribed in a reversed manner into cDNA and labeled with fluorescent dyes. Label the cDNA from the cell grown in condition A with red dye and that from condition B with blue dye (M.M. Babu 2004).

C. HYBRIDIZATION

Differentially labeled samples are allowed to hybridize onto the same glass slide. The cDNA sequences in the sample thus hybridize to spots on the slide that contains their complementary sequence. The number of cDNA attached to a particular spot is proportional to the original amount of RNA molecules for the gene in question in both

samples (M.M. Babu 2004). The hybridization step is done with great care in order to minimize cross-hybridization between genes that are similar (A.L. Tarca et al. 2006). After completing the hybridization reaction, target and any reference materials that couldn't find a probe partner are washed off (D.P. Berrar et al. 2003).

D. SCANNING

The hybridized spot in the microarray are excited by a laser and scanned in order to identify the red and the green dyes. The amount of florescent that is emitted after excitation is proportional to the amount of nucleic acid that is bounded (M.M. Babu 2004).

E. DETECT AND EVALUATE mRNA ABUNDANCE

After the scanning process, we obtain an image, which contain spots that compare to a gene (M.M. Babu 2004). The image obtained is stored as a 16-bit tagged image file format (TIFF) file, which are in pairs and each fluorescent dye corresponds to one TIFF file (S. Dudoit et al. 2002).

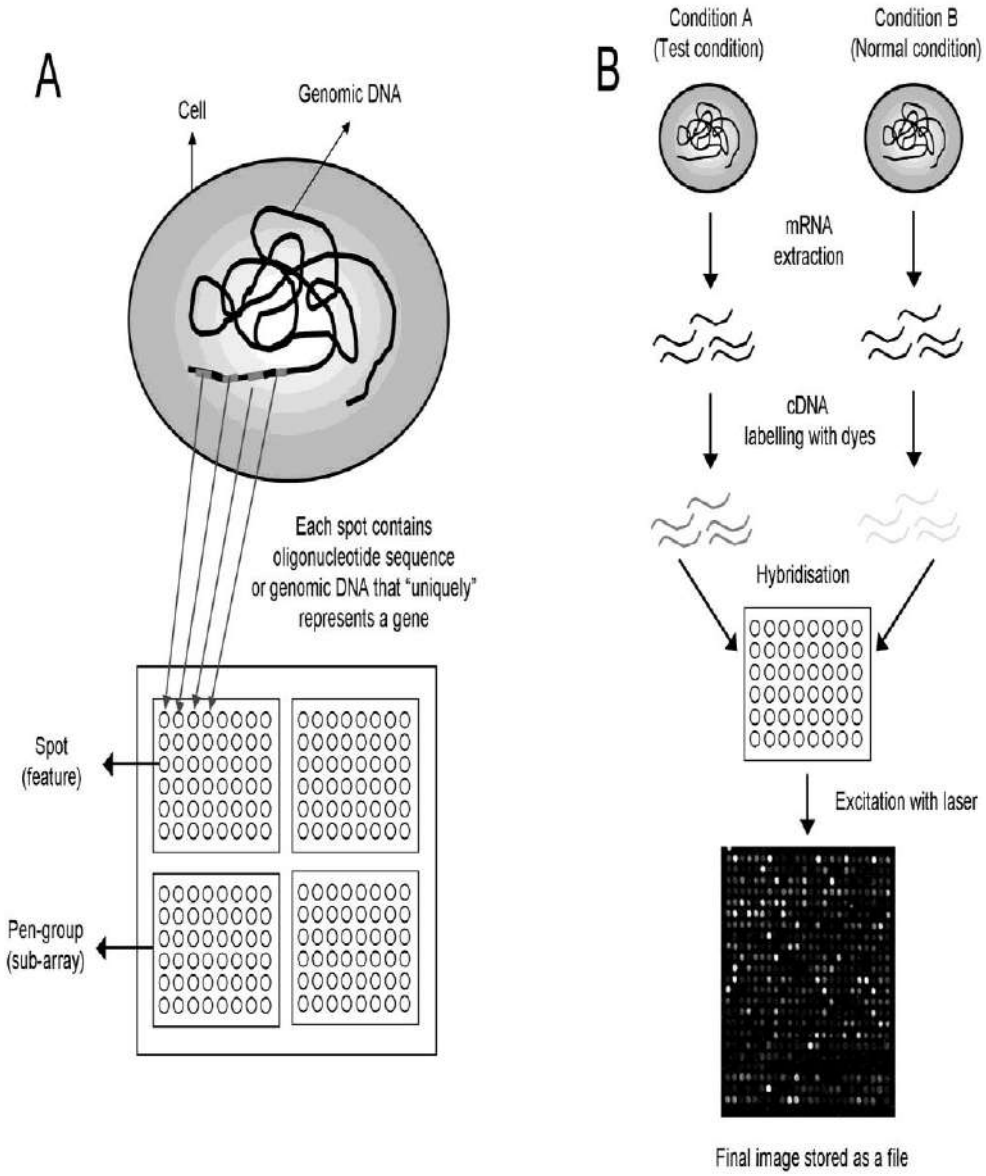

Figure 2.1: (A) A microarray (B) Steps in microarray experiment (Figure adapted from M.M Babu, 2004)

2.2.1.2 WHY MICROARRAY EXPERIMENT

Microarray experiment can be carried for the following reasons

A. Comparing patient samples

Microarray experiment helps to compare gene expression profiles of leukemia samples from patients in order to identify samples that might respond to a particular type of treatment or to identify the likely cause of the disease in question (S.A. Ness 2006).

B. Genetic difference analysis

Here microarray experiment are used to compare genetic expression pattern changes in the cell of an organism with the same genetic makeup (isogenic organism) which differs by mutation or overexpression of genes at particular gene location (S.A. Ness 2006).

C. Comparing treatment

The comparism of a cell line before or after some defined treatment is applied to a particular disease can be studied with the aid of a microarray experiment. Here, we expect the cell line to behave in a certain way after the treatment and as a result this type of study is easy and straightforward (S.A. Ness 2006).

2.3 TYPES OF MICROARRAY

Microarray can be classified by probe length, method of manufacturing and number of samples profiled on one array (A.L. Tarca et al. 2006).

2.3.1 PROBE LENGTH

Array can be classified according to the length of its probe as:

A. Complementary DNA (cDNA) arrays

This array uses long probes with length up to hundreds or thousand base pairs (A.L. Tarca et al. 2006). Due to their long length, cDNA array might detect the entire transcript produced through different RNA splicing. This array type is quite cheaper with robust hybridization. Its major disadvantages are high cost and it produces large libraries of

purified cDNAs that are always difficult to assemble (S.A. Ness, 2006). This type of array has a lot of variability in its design, more flexible, and easier to analyze with appropriate experimental design (S. Drăghici 2011).

B. Oligonucleotide arrays

The probes here are short with length of 50 base pairs or less (A.L. Tarca et al. 2006). This array gives reliable data than cDNA array, which are prone to specificity and G-C content problems. This array type is however expensive and has reduced hybridization efficiency (S.A. Ness 2006).

2.3.2 METHOD OF MANUFACTURING

Arrays are either manufactured by deposition or in-situ.

2.3.2.1 DEPOSITION APPROACH

This method of manufacturing microarray involves depositing spots of ante-synthesized nucleic acid on array support surface (S. Drăghici 2011). This can be

A. Deposition of PCR-amplified cDNA clones

Here, DNA is prepared at a distance from the chip. Thin pins are the dipped into the DNA material with the aid of robots and the pins are touched onto the array surface. A spot is then formed which results from the deposition of small DNA quantity on the array surface (S. Drăghici 2011).

B. Printing of synthesized oligonucleotides

Here, short, synthesized oligonucleotides are attached to the solid support. Since the probes used are short, the detection of splice variants is possible. This approach is called printed microarray (S. Drăghici 2011).

2.3.2.2 IN-SITU SYNTHESIS APPROACH

This method of manufacturing microarray involves photo-chemically synthesizing the probes on the chips. Probes are selected based on sequence information and hence each of the synthesized probes is recognized. This method monitors and differentiates genes that are related since it can avoid sequences that are identical among gene family members. Also this method does not introduce noise into the cDNA system, as there are no cloning, PCR reaction and spotting (S. Drăghici 2011). The approaches in this method are

A. Photolithographic approach

This approach uses a photolithographic mask for each probe. This mask however has a hole that takes in a probe with a given base. The next masks will the construct the sequences base by base. This approach thus produces arrays of very high densities but the DNA sequences constructed has limited length. Examples of array produce with this method are Affymetrix, Santa Clara, CA etc. (S. Drăghici 2011).

B. Ink-jet technology

This approach utilizes the technology used in ink-jet color printer. Cartridges are loaded with A, C, G and T nucleotides. Nucleotide deposition occurs when print head moves across the array substrate. Examples of array that uses this method are Agilent, Protogene etc. (S. Drăghici 2011). This method is thus fast, versatile and high yielding but gives a low resolution (R.D. Egeland and E.M. Southern 2005).

C. Electrochemical synthesis

This method makes use of small electrodes, which are planted into the microarray substrate in order to control individual reaction sites. These electrodes are activated in the required positions in a fixed sequence after solutions with specific bases are washed over the substrate. This will thus construct the required sequences base wise. Examples of arrays here are CombiMatrix, WA, Mukilteo, Bothel (S. Drăghici 2011).

2.3.3 NUMBER OF SAMPLES PROFILED ON ONE ARRAY

Another array classification is based on the number of samples that can be profiled on one array. This can be

A. Single-channel array

This is a type of array in which one sample is hybridized at a time and it is less common (J. Kesseli 2015).

B. Multi-channel array

Also called two-channel array, it hybridizes two sample at a time with each of the sample having is own fluorescent dye. The fluorescent dye thus allows the separation of the samples. This type of array is common in microarray experiments (J. Kesseli 2015).

2.4 APPLICATIONS OF MICROARRAYS

Microarray finds applications in the following areas

2.4.1 DIFFERENTIAL GENE EXPRESSION ANALYSIS

This analysis involves comparing gene expression levels between different experimental conditions say phenotypes i.e. comparing the expression discrepancy of a single gene expression profile against the experimental conditions. It usually considers one sample as the control or reference and the other sample as the experiment. Examples are healthy versus disease, treated versus untreated etc. (D.P. Berrar et al. 2003).

2.4.2 CO-REGULATION OF GENES ANALYSIS

This analysis involves comparing gene profiles of two or more genes with the purpose of identifying genes which expression measures differs in a corresponding manner throughout the experimental conditions. Two genes are said to co-regulate positively if the expression measure of one gene increases as that of the other and co-regulate negatively if the expression measure of one gene decreases as that of the other (D.P. Berrar et al. 2003).

2.4.3 GENE FUNCTION IDENTIFICATION

The function of a unique gene can be obtained from a detailed microarray experiment. This is done by comparing the expression profiles of previously studied genes with known function with that of the unique gene's expression profile under different conditions. The functions of the previously studied genes with highly similar expression profile can be used to predict the functions of the unique gene in question (D.P. Berrar et al. 2003).

2.4.4 PATHWAYS AND GENE REGULATORY NETWORK IDENTIFICATION

Microarray experiment helps to identify pathways and genes regulatory network when a cell is stimulated by finding genes that are turned on and off at different time points. Pathway identification analysis helps to show the paths and procedures where genes and their products function in cell, tissue and organism while gene regulatory network regulates gene expression (D.P. Berrar et al. 2003).

2.4.5 SEQUENCE VARIATION STUDIES

Microarray experiments also helps to study variations that occur in sequences. Sequence variation studies aims to uncover DNA sequence variations that corresponds with change in phenotype e.g. diseases. Examples of sequence variation can be single nucleotide polymorphism (SNP), insertions and deletion etc. (D.P. Berrar et al. 2003).

2.4.6 CLINICAL DIAGNOSTICS

Microarray also finds application in clinical diagnostics where it helps to reveal the different pattern in expression measures that are feature for a particular type of disease and also to deduce unknown disease subtypes from previously known diseases (D.P. Berrar et al. 2003).

2.5 ISSUES WITH MICROARRAY ANALYSIS

The major issues associated with microarray user are as follows

A. NOISE

Noise is one of the major issues associated with microarray experiment because it is introduced at every step in the experiment. Since microarrays are noisy, repeating an experiment more than once using the same materials and same preparations as done in the previous experiment, many genes gives different quantification values after scanning and image processing steps as a result of noise (S. Drăghici 2011).

B. EXPERIMENTAL DESIGN

Experimental design is the most important phase microarray analysis process but frequently ignored. It refers to the test or series of tests that a scientist influences the input variable of a process with the aim of detecting and recognizing the effect that change has on the process outcome. If the microarray experiment to be carried out is not well designed, then the experiment will give a false result (S. Drăghici 2011).

C. HUGE NUMBER OF GENES

The microarray technology is capable of examining thousands of genes at the same time. When the number of genes is too large say tens of thousand in a microarray experiment, experimental quality and designed method might change and this might lead to a problem in the analysis process (S. Drăghici 2011).

D. ASSESSMENT OF ARRAY QUALITY

Assessing the quality of array should be the aims of the data analysis process in a microarray experiment. This assessment thus helps in rejecting the data from faulty array and to determine why a microarray process may fail. Hence it is very important and if not giving proper attention, might after result from the experiment (S. Drăghici 2011).

2.6 MICROARRAY ANALYSIS PROCESS

The processes involved in microarray analysis are shown in the figure 2.2 below

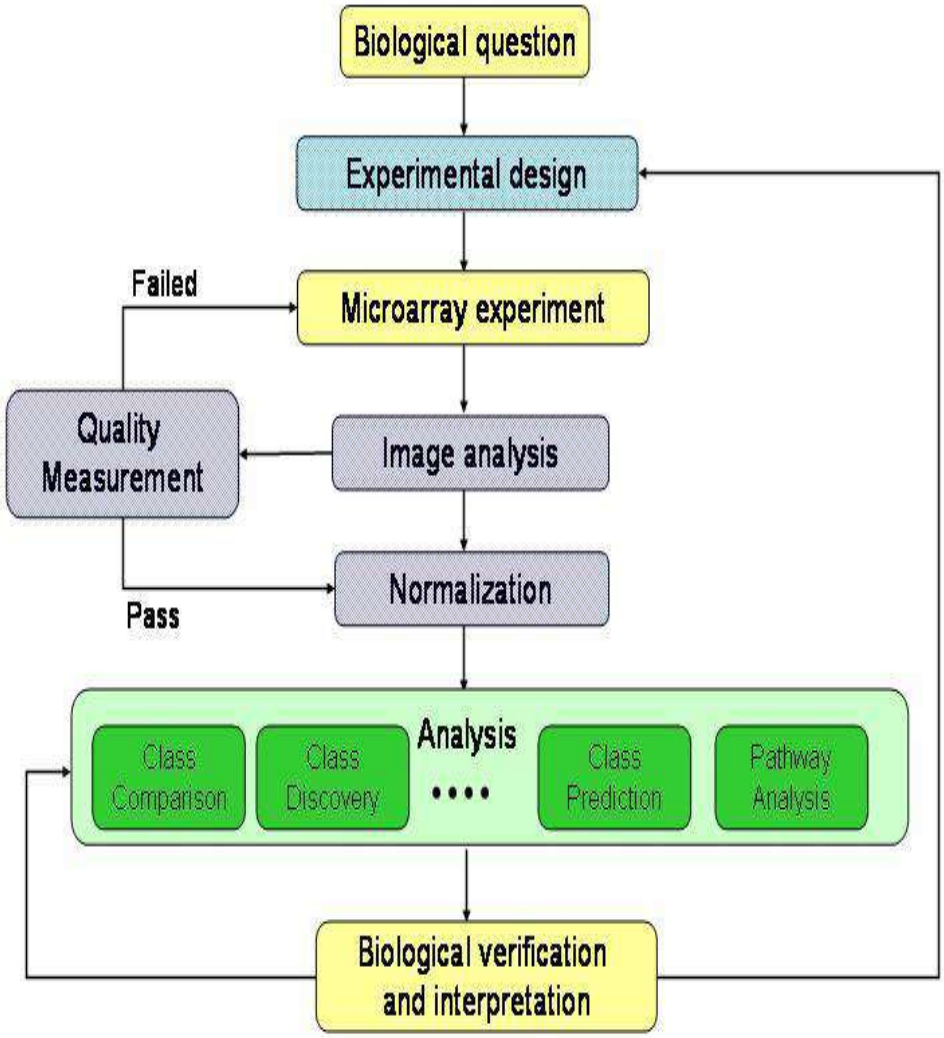

Figure 2.2: Microarray analysis processes (Figure adapted from A. Sánchez and M. C. Ruíz de Villa 2008).

2.6.1 BIOLOGICAL QUESTIONS

This is the first phase of microarray analysis process, which tends to define the biological effect of the microarray experiment we are about to start. Here we formulate hypothesis and we check relevant literatures that will back the biological findings we intend to make (D.P. Berrar 2003).

2.6.2 EXPERIMENTAL DESIGN

Experimental design is the most important phase microarray analysis process but frequently ignored. It refers to the test or series of tests that a scientist influences the input variable of a process with the aim of detecting and recognizing the effect that change has on the process outcome. To obtain a viable conclusion, experiment must be properly designed (S. Drăghici 2011). For a successful experiment, the following should be consider

2.6.2.1 EXPERIMENTAL DESIGN PRINCIPLES

This refers to the rules which when adapted, leads to a successful experiment. The rules are replication, randomization and blocking.

A. Replication

It is a process of repeating an experiment more than once. If introduced in a microarray experiment, it helps to recognize and limit the noise introduced in the hybridization step of the microarray experiment and to constrain biological variability (S. Drăghici 2011). In a microarray experiment, replication can be technical or biological replication (A. Sánchez and M. C. Ruíz de Villa 2008).

B. Randomization

Randomization is a process of using a random choice for all factors that are unimportant but can have influent the experimental result. This can be done in microarray analysis process by printing replicated spot at random location throughout the array or by using different batches of microarray slide in differential gene expression (i.e. comparing a

reference group versus a treatment group). It thus helps to deal with nuisance factors (S. Drăghici 2011).

C. Blocking

This is a technique of creating similar microarray slides of data in which factor of interest is left to vary while nuisance factor is kept constant. It helps to remove variability due to the difference between microarray slides (S. Drăghici 2011).

2.6.2.2 EXPERIMENTAL DESIGN GUIDELINES

When panning a microarray experiment, the following can be taken into account:

i. Define research problem
ii. Select the type of microarray to be used.
iii. Seek experts opinion on experimental design and data analysis
iv. Choose factors of interest in according to level of importance
v. Identify possible nuisance factors
vi. Choose significance level and desired power
vii. Design your experiment
viii. Execute the experiment and collect data
ix. Analyze data
x. Extract biological meaning from data analytics result (S. Drăghici 2011).

2.6.3 MICROARRAY EXPERIMENT

The steps involved in microarray experiment have been described in in section 2.2.1.1. Once the experiment as been performed, we proceed to the next step in the analysis process.

2.6.4 IMAGE ANALYSIS

Image analysis involves measuring spots intensities and calculating gene expression values based on these intensities. This analysis assists in evaluating data reliability, helps to produce warning showing possible issues during hybridization and array production phases (S. Drăghici 2011).

Furthermore, this analysis forms the basis of any further analysis as it represents the basic data collection step. The images from microarray experiment are seen as spots, which are arranged in an orderly manner into sub-grid and this orderly arrangement thus makes spot detection easy (M.M Babu 2004). The stages in image processing are explained in the subsections below.

2.6.4.1 SPOT DETECTION

This stage involves detecting signal spots in the images and computing the size of each spot. The method employed for spot detection can be manual, semiautomatics and automatic (S. Drăghici 2011).

2.6.4.2 IMAGE SEGMENTATION

This is a method of dividing image into regions, which do not overlap but whose unification represents the whole image. Segmentation aims to break the image down into spot and background in other to measure spot signal and evaluate the intensity of the background.
Image segmentation step is introduced once the spot has been discovered, because it helps to determine which pixels represents the spot and should be used in signal calculation, which pixels represents the background and which pixel represents noise and should be removed. Segmentation methods can be pure spatial-based, intensity based, Mann-Whitney, etc. (S. Drăghici 2011).

2.6.4.3 SPOT QUANTIFICATION

Spot quantification involves calculating a unique value for each gene on the chip, which might be proportional to the quantity of mRNA present in the solution that hybridized the chip. It aims to marge pixel intensity values into a distinctive value that represents the expression level of a gene saved on each spot (S. Drăghici 2011).

Spot quantification can be achieved by taking total, mean, median, mode, volume and ratio signal intensities across the two channels. These intensities value are taken for the entire pixel within the area taken into consideration for the spot (S. Drăghici 2011).

2.6.4.4 SPOT QUALITY ESTIMATION

Spot quality estimation helps to assess spot in images with questionable values. In microarray analysis, the following quality measures are important

i. Ratio between spot signal area and total spot area
ii. Shape regularity
iii. Spot area to perimeter ratio
iv. Displacement
v. Spot uniformity (S. Drăghici 2011).

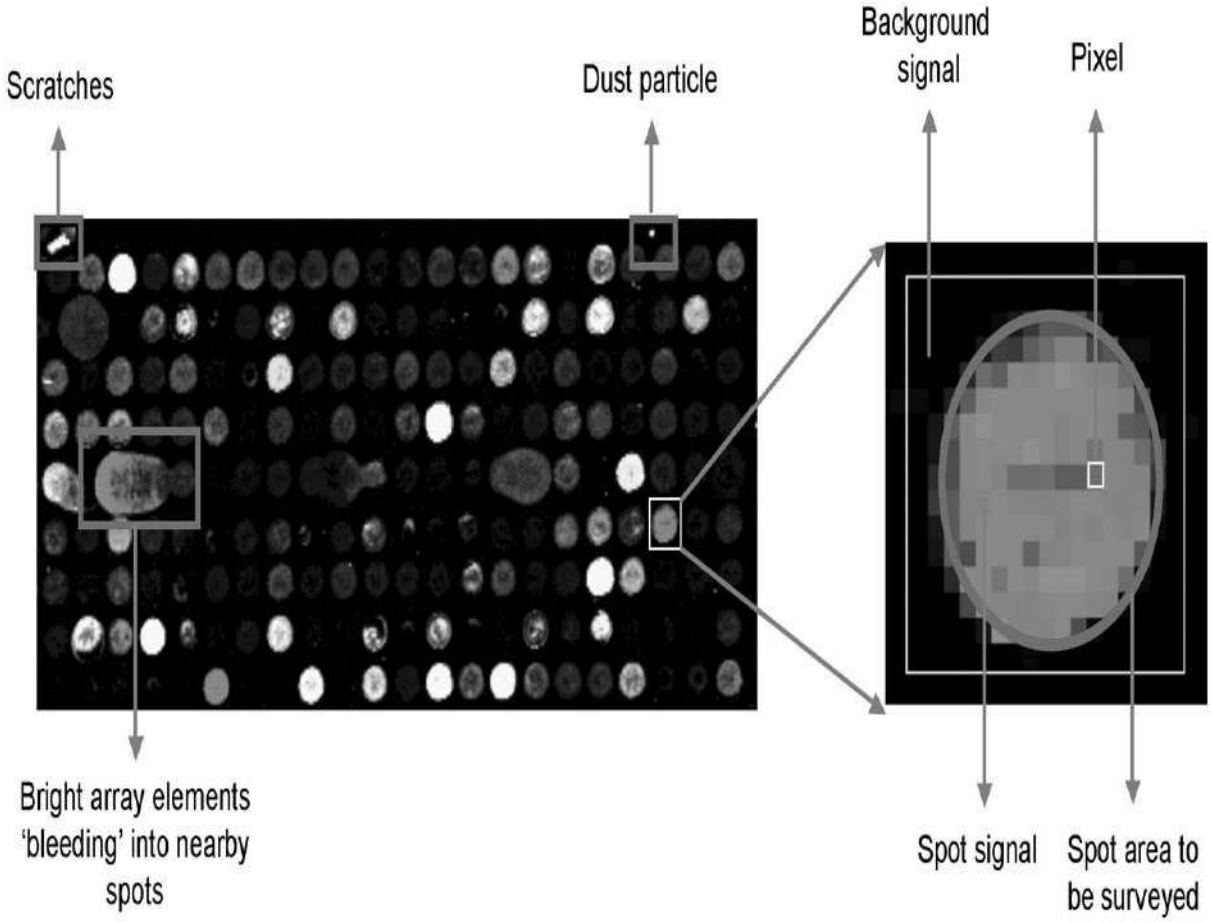

Figure 2.3: Microarray image spot and background (Figure adapted from M.M. Babu 2004).

2.6.5 QUALITY ASSESSMENT

Quality assessment is a crucial step after obtaining the raw data in a microarray experiment. It involves assessing the quality of data obtained for correct data interpretation and to proceed to the next step in the analysis. Since a single error or few weird arrays can disrupt the results of the data analysis process, it is mandatory that the data be of good quality before proceeding to the next step in the analysis (S. Drăghici 2011). It thus helps to know whether the whole microarray experiment has worked well so that the data used can be considered trustworthy (A. Sánchez and M. C. Ruíz de Villa 2008).

Quality assessment starts by visual inspection of the images followed by plots of the raw data (N.J. Armstrong and M.A. Van de Wiel 2004). For two channel arrays, quality assessment is based on image inspection and plots while for one channel array, it is based on image inspection, degradation plots, quality control metrics estimation such as average background, present calls, scale factors, hybridization quality etc. (A. Sánchez and M. C. Ruíz de Villa 2008).

2.6.6 PREPROCESSING

Preprocessing is a step in microarray analysis process that extracts meaningful data characteristics from dataset obtained from image analysis step (A.L. Tarca et al. 2004). The dataset (gene expression dataset) can be represented by a **real-valued matrix *I*** where I_{ij} is the measured expression level of gene i in experiment (condition) j. The ***i-th*** row of the matrix represents the expression pattern of gene ***I*** and the ***j-th*** column represents the expression profile of gene ***j*** (A. Ben-Dor et al. 1999).

The gene expression matrix can thus be represented as absolute measurement, expression ratio (relative measurement), log2 expression ratio, discrete value or vectors i.e. representing expression profiles as vectors (M.M Babu 2004). Since the gene expression matrix obtained contains noise, missing values and systemic variation dues to experimental procedure, hence it must be preprocessed before any further data analysis (D. Jiang et al. 2004). Preprocessing can be carried out by background correction, logarithmic transformation, and normalization (S. Drăghici 2011, A.L. Tarca et al. 2006).

2.6.6.1 BACKGROUND CORRECTION

It is the first preprocessing step in microarray analysis (S. Drăghici 2011). It aims to obtain an intensity value, which is proportional to the expression level by determining the level of hybridization that occurs between the targets and the samples (A. Sánchez, M.C. Ruíz de Villa 2004). Background correction can be local, sub-grid, group, blank spots, and control spots background correction (S. Drăghici 2011).

2.6.6.2 LOGARITHMIC TRANSFORMATION

Logarithmic transformation is the first technique used to preprocess microarray data. This technique provides data that are more meaningful and easy to interpret biologically (S. Drăghici 2011). It also makes statistical distribution symmetrical and almost normal, and very convenient. Hence, logarithmic transformation is called transformation to normality (D.P. Berrar 2003).

2.6.6.3 NORMALIZATION

Normalization is a step in data preprocessing that aims to correct systemic differences (error) between genes or array in a microarray experiment. The systemic error, which may be due to variability in sample preparation, experimental bias (A.L. Tarca et al. 2006), labeling efficiencies, scanner settings and spatial effect can occur at numerous stages during microarray experiment, hence the need for data normalization (S. Dudoit 2002)

The specific normalization techniques to use for data preprocessing depend on the array technology used. The Biocondutor projects has various algorithms which includes MAS 5.0, Robust Microarray Average (RMA), GC-RMA (for one channel array), and LOESS normalization (for two-channel array) which are used for microarray data preprocessing (A.L. Tarca 2006).

2.6.7 STATISTICAL ANALYSIS

Once the raw data has been preprocessed, statistical analysis of the data matrix can be carried out. In this study, the following analysis will be examined.

2.6.7.1 CLASS DISCOVERY

Class discovery involves analyzing a set of gene expression profiles with the sole aim of discovering subgroups that share similar features. This analysis helps to make meaningful biological inference about the set of genes or samples, identify different stages of disease severity and identify groups of gene that may behave alike in a disease state (A.L. Tarca et al. 2006). Class discovery is also known as clustering analysis (A. Sánchez and M.C. Ruíz de Villa 2008)

Clustering is an unsupervised analysis because there is no prior knowledge about the data and is currently the most frequently used technique in analyzing gene expression data (S. Drăghici 2011). It is the process of grouping objects into different classes called clusters, so that objects within a class are more similar to each other while objects from different classes are dissimilar (D. Jiang et al. 2004).

To group items (genes or samples) that are similar together successfully, one needs to define a good measure of similarity called metrics or distance. There are different measures of similarity that can be applied to clustering and this includes Euclidean, Manhattan, Chebychev, Correlation, Mahalanobis, Minkowski distances etc. It should be noted that the choice of the distance metric to use depends on the array technology in use as different array technology represents expression matrix in different format (S. Drăghici 2011) as explained section 2.1.5.6.

Clustering can be gene-based, sample-based or subspace clustering depending on the researchers aims and objectives. In gene-based clustering, genes are the objects, while the samples are the features. Gene-based clustering helps to determine co-regulated genes, recognize temporary expression patterns and to reduce the prediction model redundancy. Sample-based clustering however treats samples as objects and the genes as the features.

It helps to identify new tumor classes and to detect experimental artifact. Subspace clustering treats both samples and genes uniformly, so that either samples or genes can be objects or features and it captures clusters formed by a subset of genes across a subset of samples. Clustering results are usually shown in a diagram called dendogram (D. Jiang et al. 2004).

Clustering algorithms applied to microarray data may be hierarchical or non-hierarchical (M.M. Babu 2004) as shown in figure 1.4 below. It should be noted that clustering algorithms are dependent on distance metric used and the relationship between patterns within the clusters is independent of their position (S. Drăghici 2011).

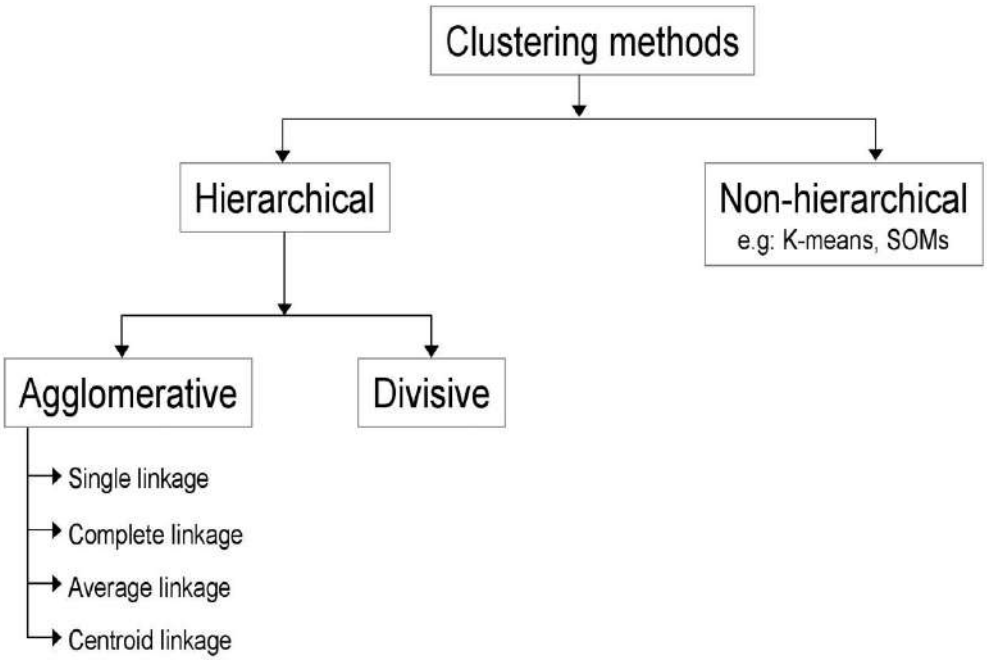

Figure 2.4: Clustering classifications (Figure adapted from M.M. Babu 2004).

2.6.7.1.1 HIERARCHICAL CLUSTERING ALGORITHM

Hierarchical clustering algorithm is the oldest clustering algorithm used in the analysis of microarray data (S. Drăghici 2011). It creates a hierarchical series of nested clusters (based on degree of similarity), which can be represented graphically by a tree like structure, called dendrogram (D. Jiang et al. 2004).

Furthermore, hierarchical clustering result may also be shown with a heatmap (A.L. Tarca et al. 2006). Heatmap is a color image plot, which is made up of a rectangular array of colored block with each block representing the expression level of one gene on the array (A. Sánchez and M. C. Ruíz de Villa 2008). Hierarchical clustering algorithm can be agglomerative or divisive clustering depending on the method used in drawing the dendrogram (D. Jiang et al. 2004).

A. AGGLOMERATIVE CLUSTERING

Agglomerative clustering is also known as the bottom-up approach. Here each object is considered as a cluster and all objects are successfully fused until a single cluster is formed. Agglomerative clustering thus lacks robustness and it expensive computationally (S. Drăghici 2011). The fusion of all objects (clusters) into a single cluster is based on the pairwise distance between them, hence object that are similar are first clustered and the process continues until a single cluster is formed. The pairwise distances used in agglomerative clustering are single, complete, average linkage and centroid linkage (M.M babu 2004).

a. Single linkage clustering

This computes the minimum distance between all possible pairs of objects; one from each cluster and it chooses clusters that are closest together. This method is insensitive to outliers and is also known as minimum distance or nearest neighbor linkage (M.M babu 2004). The distance here is given as

$$D(A, B) = min_{\vec{x}\in A, \vec{y}\in B} \| \vec{x} - \vec{y} \| \qquad \text{Eqn 1}$$

b. Complete linkage clustering

It also called maximum distance or farthest neighbor linkage, it chooses the clusters that are farthest apart by computing the maximum distance between all possible pairs of objects, one from each cluster. The method major disadvantage is that it is sensitive to outliers (M.M babu 2004). Complete linkage distance is given as

$$D(A, B) = max_{\vec{x} \in A, \vec{y} \in B} \| \vec{x} - \vec{y} \| \qquad \text{Eqn 2}$$

c. Average linkage clustering

Average linkage clustering evaluates the distance between two clusters as the average of the distances between all possible pairs of objects in the two clusters (M.M babu 2004). Distance here is given as

$$D(A, B) = Average_{\vec{x} \in A, \vec{y} \in B} \| \vec{x} - \vec{y} \| \qquad \text{Eqn 3}$$

d. Centroid linkage clustering

It finds the distance between the centers of the cluster. It calculates an average expression profile called centroid by first finding the mean in each dimension profile for all objects in the cluster and it then measures the distance between the clusters as the average expression profile of the two clusters (M.M babu 2004).

B. DIVISIVE CLUSTERING

Divisive clustering is also known as top-down approach. It is the reverse of agglomerative clustering as the entire set of object is seen as a single cluster and split is performed repeatedly until a single cluster of each object remains. Divisive clustering is not a popular clustering method unlike agglomerative clustering (M.M babu 2004) though it is faster and requires less computation (S. Drăghici 2011).

2.6.7.1.2 NON-HIERARCHICAL CLUSTERING

Non-hierarchical clustering algorithm is a partitioning-based clustering algorithm that decomposes data into a set of dismembered clusters (D. Jiang et al. 2004). The number of clusters is predetermined in this algorithm and the existing objects are grouped into these predefined clusters. Non-hierarchical clustering can be K-means clustering, Self-organizing maps (SOMs) etc. (M.M babu 2004). Details of k-means and SOMs are explained in the subsections below

A. K-means clustering

K-means clustering is the most wide used non-hierarchical clustering methods. It is simple, fast and involves grouping objects arbitrarily into a predetermined number of K clusters (S. Drăghici 2011). The predetermined number clusters can be obtained from hierarchical clustering results or chosen randomly. A centroid is then calculated for each cluster and objects are re-grouped from one cluster to the other depending on the centroid closer to the gene. This procedure of calculating centroid and re-grouping is repeated iteratively until the composition of clusters remains unaltered by further iteration (M.M babu 2004).

A shortcoming in K-means clustering is the specification of the number of clusters (K) before the algorithm is run. In a situation where there is no fix value for K, user has to try different values of K, which is not a good approach (F.M. Al-Akwaa 2012).

B. Self-organizing maps (SOMs)

SOMs are also known as Self-organizing feature maps or Kohonen map, was developed by Tuevo Kohonen in late 80s. It is a clustering method that tends to divide input patterns into group of patterns that are similar are plotted next to one another. SOMs are a grid neural network, which can be one (a string), two (an array) or three (a cube) dimensional (S. Drăghici 2011). Further details about SOMs and other non-hierarchical clustering algorithms can be found in (S. Drăghici 2011).

2.6.7.2 CLASS COMPARISON

Class comparison involves detecting the differences in gene expression levels between predefined groups of samples in patients e.g. tumor and normal sample (A.L. Tarca 2006). The genes selected are thus called differentially expressed genes while the process is called differential expression analysis (A. Sánchez, M.C. Ruíz de Villa 2004).

For a successful class comparison analysis, definite experimental design, hypothesis testing, sample size estimation (A.L. Tarca 2006) and performance assessment (S. Drăghici 2011) should all be considered.

2.6.7.2.1 HYPOTHESIS TESTING

Hypothesis testing helps in selecting the real genes that are differentially expressed between the groups under study. A null and alternative hypothesis needs to be developed with the null hypothesis stating that a given gene on the array is not differentially expressed between the two groups under study and the alternative hypothesis stating that the expression level of that is different between the two groups. This can be achieved by calculating the t-statistic of the expression value of the gene under study measured in the two groups. This value will then be compared with a chosen significant value and particular threshold (A.L. Tarca 2006). Further details on hypothesis testing can be found in (S. Drăghici 2011).

2.6.7.2.2 SAMPLE SIZE

This refers to the number of measurements in an experiment. Its computation required knowledge about the minimum fold change to be detected, gene expression variances within each group and statistical power. Large number of samples results in high fold change value, more expression variability and high statistical power, which will in turn give more differentially expressed genes, hence a large sample size is encouraged. However due to the high cost of microarray experiment, a certain number of replicate such as 5 samples per group are usually adopted (A.L. Tarca 2006).

2.6.7.2.3 PERFORMANCE ASSESSMENT

Performance assessment of each gene selection method helps to determine the best approach since there are several approaches used in selecting genes that are differentially expressed (DE Genes). Gene selection methods performance can be computed in terms of accuracy, specificity, sensitivity, positive predicted value (PPV), and negative predicted value (NPV). These measures have values between 0 and 1 and are sometimes expressed in percentage (S. Drăghici 2011). They are defined as follows

$$\textbf{Accuracy} = (T.P + T.N)/ (T.P + F.P + T.N + F.N) \qquad \textbf{Eqn. 4}$$

$$\textbf{Specificity} = T.N / (T.N + F.P) \qquad \textbf{Eqn. 5}$$

$$\textbf{Sensitivity} = T.P / (T.P + F.N) \qquad \textbf{Eqn. 6}$$

$$\textbf{PPV} = T.P / (T.P + F.P) \qquad \textbf{Eqn. 7}$$

$$\textbf{NPV} = T.N / (T.N + F.N) \qquad \textbf{Eqn. 8}$$

Furthermore, they all play a major role in selecting the best approach for selecting genes that are differentially expressed and they all depends on prevalence, which is given as

$$\textbf{Prevalence} = (T.P + F.N) / (T.P + F.P + T.N + F.N) \qquad \textbf{Eqn. 9}$$

Using change/unchanged as the condition, the following definitions hold: T.P (True positives) are the true change that are reported as change, F.P (False positives) are unchanged reported as change, F.N (False negatives) truly change reported as unchanged, and T.N (True negatives) are the truly unchanged that are reported. A perfect gene selection approach will produce value of 1 for accuracy, specificity, sensitivity, positive predicted value (PPV), and negative predicted value (NPV) (S. Drăghici 2011).

2.6.7.2.4 METHODS OF SELECTING DIFFERENTIALLY EXPRESSED GENES

Several methods of selecting differentially expressed genes have been developed but the methods employed in this study will be discussed briefly.

A. Fold change method

The fold change approach is the simplest and the most natural method used in finding genes that are differentially expressed between the two conditions (reference and experiment). An easy way to achieve this is by calculating the ratio between the two expression levels for each gene (S. Drăghici 2011). This ratio is then log-transformed (log2) and the transformation helps to improve the symmetry of the data distribution by giving a mean log-ratio of zero (A.L. Tarca et al. 2006).

To select the genes that are differentially expressed, a histogram which horizontal axis represents the log-ratio can be plotted. A threshold will then be set and the differentially expressed genes (DEG's) are those outside this threshold. This method's major drawback is that the fold threshold is chosen arbitrarily, which might be inappropriate and it doesn't consider the variance of the expression value measured (S. Drăghici 2011).

B. Unusual ratio method

This method involves the selection of genes for which experiment and reference values ratio is a certain distance (±2 multiplied by standard deviation) from the mean experiment/control ratio. This can be achieved by applying a z-transform to the log ratio value. The z-transform thus subtracts the mean and divides by the standard deviation. This method is also simple, intuitive and more superior than the fold change method as it adjust the cut-off threshold automatically, hence it picks the most affected genes irrespective of the number and extent the genes regulated. However, this method report 5% of the genes as differentially regulated even if there is none and 5% of the genes even if there are more differentially regulated genes (S. Drăghici 2011).

C. T-Test Method

This method employs classical hypothesis testing methodology along with some correction for multiple comparison (Bonferroni, Benjamini-Horcberg etc.) in selecting genes that are differentially expressed (S. Drăghici 2011).

Classical hypothesis testing (statistical hypothesis) refers to the assumption made about the genes involved the microarray experiment in other to make statistical decision about those genes on the basis of sample information. These assumptions are statements about genes probability distributions. The major steps in hypothesis testing are as follows

a. Define the problem clearly
b. Generate null and alternative hypothesis.
   Null hypothesis is the hypothesis that occurs by chance while an alternative hypothesis is that hypothesis that differs from the given null hypothesis.
c. Choose level of significance (S).
   Level of significance refers to the probability of rejecting a true null hypothesis.
d. Calculate statistics based on data and compute p-value.
   P-value is defined as the probability of drawing the wrong deduction by rejecting a null hypothesis that is true.
e. Compare p-value and significant level in other to accept or reject null hypothesis.
   If p-value is less S, we reject a true null hypothesis (Type I error) otherwise, we accept false null hypothesis (Type II error) (S. Drăghici 2011, M.R. Spiegel et al. 2001).

This method is conservative and it assumes that genes in the experiment are independent which is not always true in any real data set (S. Drăghici 2011).

In the past, many publications of microarray experiment determine differentially expressed genes solely from fold-change analysis by using a 2-fold cutoff. This method however does not consider variability in the data; hence its results are not reliable. T-test or the Wilcoxon test approach is later introduced in order to correct the defects in fold change analysis but it was found out that the result obtained with this test produces a lot False Discovery Rates

(FDRs) in small number of samples. In addition, the result obtained is poorly related to that obtained from the fold change analysis (D.J. McCarthy and G.K. Smyth 2009).

In recent years, modern statistical tests for microarray analysis have been developed. These tests derive their information between genes adopting empirical Bayes along side other statistical methods. In addition, the tests have thus performed better than the t-test and give results that are related with fold change analysis. Since these modern statistical tests allows genes with small fold change to be considered significant statically. Hence, it is now common practice that differentially expressed genes satisfy at the same time both p-value and fold-change thresholds. Peart et al. (2005) and Raouf et al. (2008) obtained differentially expressed genes by setting a fold-change cutoff of at least 1.5 and a p-value less than 0.05 after performing multiple testing. In addition, Huggins et al. (2008) adopts a fold change of 1.3 and a p-value less than 0.2 to find differentially expressed genes. It was observed that the combination of fold change and p-value threshold gives differentially expressed genes that are more meaningfully biologically than using the p-value alone (D.J. McCarthy and G.K. Smyth 2009).

Other methods used for selecting differentially expressed genes are Analysis Of VAriance (ANOVA), Noise sampling, Model-based maximum likelihood estimation, Significance Analysis of Microarray (SAM), moderated t-statistics, Use of Biocondutor packages such as Limma (S. Drăghici 2011), DESeq, DESeq2 etc.

2.6.7.3 CLASS PREDICTION

Class prediction analysis involves discovering class membership of a sample based on its gene expression profile. Its main goal is to accurately predict the class membership (e.g. phenotype) of a new individual by building a classifier that can distinguish between the classes. This goal can be achieved by first learning the existing complex relationship between the class membership and the expression values of the genes (A.L. Tarca et al 2006). Class prediction analysis involves three major components namely:

A. FEATURE SELECTION

Feature selection is the first component of class prediction analysis and it defines the number of genes to include in the predictor to be built. Feature selection is of great importance in microarray studies due to the fact that the number of informative variables used for differentiating our classes of interest is usually very small when compared to the overall genes on the microarray. Hence it is of great importance to choose informative genes to be use by the class predictor (R. Simon 2003).

Furthermore, future selection employs class comparism analysis, as it's most common approach i.e. selecting and identifying differentially expressed genes among gene classes when the classes are treated one after the other (R. Simon 2003).

B. SELECTING PREDICTOR FUNCTION (CLASSIFIER)

Classifiers are functions that take as input the values of several features (variables that are independent) from an example set of independent variable values and presume the category that the example set belongs i.e. a classifier is a function that produce a class label prediction from an example input. Classifiers are usually trained so as to learn some useful parameters from the training data and this turns the classifier to a model that reveals the relationship between variables and the training set class labels. The relationship between the variables and class labels in the training set are tested using the test data (F. Pereiraa et al. 2009).

Defining a predictor function that will for any expression vector gives a class label is the second component of class prediction analysis. Predictor function can be nearest neighbor predictor, linear discriminant analysis, logistic regression, support vector machine etc. (R. Simon 2003).

The K nearest neighbor classifier, also called case based reasoning, example based reasoning, memory based reasoning, instance based learning and lazy learning is a classifier that categorize samples based on the similarity measure of the class of their nearest neighbor. It is a very popular, easy to understand, effective and very efficient algorithm used in classification, statistical estimation and pattern recognition (M.A. Jabbar et al. 2013).

Furthermore, KNN is not expensive computationally and hence it is useful for data that changes quickly. KNN algorithm starts by first measuring the distance of a new sample with all the samples in the training set. The calculated euclidean distance are arranged in ascending order and the K sample, which distance is closest to the new sample are called the k-nearest neighbors. The K-nearest neighbors obtained are then used to categorize the new sample to the predefined class and this depends on the type of data in use. If data contain categorical variables, then a voting (simple or weighted) is adopted. However, for data continuous or quantitative variables, mean, geometric mean or median is adopted (Z. Jan et al. 2008).

The K value selected for KNN algorithm is important in evaluating the model effectiveness as this helps to decide the best way to use the available data so as to generalize KNN algorithm. If the value of k is large, then the variance value is reduced as a result of noisy data. However, this large value of k is disadvantageous since it causes bias. The bias thus makes the learner to ignore pattern, which might have useful insights. The value of k is usually taken as the square root of the number of observation in the data (Analytics Community 2016).

C. CLASSIFIER PERFORMANCE ASSESSMENT

The final component of class prediction analysis involves assessing the performance of the built classifier for future samples. This is carried out by determining how well the classifier predicts unseen sample and reliability of the prediction. Accuracy of class predictor can be estimated with split-sample and cross-validation method (R. Simon 2003).

The split-sample method is the most straightforward method of estimating the accuracy of future prediction of samples. It involves splitting the current set of data into training and test set with the test set emulating the future samples, which class label is to predicted.
It should be noted that samples within the test set shouldn't be used for developing the prediction model and choosing the genes to be used in the model building (R. Simon 2003).

Cross-validation method on the other hand also involves splitting data into training and test set. The test set in this case consist of a single sample, which is placed aside and not used in class prediction model development, while the training set is used for selecting informative genes and model parameter are adapted to the data. (R. Simon 2003).

Cross-validation can be leave-one-out cross-validation, k-fold cross validation, hold-out set and bootstrapping. Details of different cross-validation techniques can be found in (J.H. Friedman et al. 2001).

The assessment result of the built classifier is usually displayed on a confusion matrix. A confusion matrix refers to a table that classify foresighted and actual elements according to how they match with the foresighted elements vertical and the original elements shown horizontally or vice-versa. A confusion matrix can be two by two; three by three etc. depended on the numbers of classes available in the built model as shown in the figure below (B. Lantz 2015). From the table, performance is measure by accuracy and error rate and other parameters as described in section 2.6.7.2.3.

		FORESIGHTED VALUE	
		Negative	**Positive**
ORIGINAL VALUE	**Negative**	**T.N** **True Negative**	**F.P** **False Positive**
	Positive	**F.N** **False Negative**	**T.P** **True Positive**

Table 2.1: A two by two confusion matrix

2.6.8 BIOLOGICAL VERIFICATION AND INTERPRETATION

The result of microarray analysis irrespective of the platform and the analysis method employed is, in most cases the list of differentially expressed genes. This list of differentially expressed genes needs to be transformed to the underlying biological phenomena for better understanding. To achieve this, the list needs to be to be translated into functional profile, which will give insight into the cellular mechanism relevant in the condition under study. Many tools have been developed to help with this task among which are Gene Ontology (GO), Onto-Express, DAVID etc. The ontology tools are thus similar in their approaches but are greatly different in many respects, which influence the nature of the results of the analysis (P. Khatri, and S. Drăghici 2001). Details of different ontology tools, their capabilities, statistical model and annotation databases are explained in (P. Khatri, and S. Drăghici 2001). The interpretation of the list of differentially expressed genes can be analyzed with a method called enrichment analysis.

2.6.8.1 ENRICHMENT ANALYSIS

Enrichment analysis is defined as the method of obtaining biological significance in term of statistical significance and its main objective is to interpret gene group in order to identify biological information for the event under study. Gene Ontology (GO) database is very efficient in performing enrichment analysis (M. Mayo and J. Luís 2014). This analysis thus helps to interpret the results obtained from the statistical analysis of microarray data in other to get biological meaning into the data under study.

2.6.8.2 METHODS USED IN ENRICHMENT ANALYSIS

Several methods have been developed in the past with the aim of seeking biological meaning based on enrichment analysis. All the methods are similar and they work in two steps. First, they consider list of genes of interest from a gene population, which results from an experiment and maps each gene in the list to the annotation terms that are correlated with it. Second, enrichment of each gene annotated in each category is computed. Comparing the proportion of genes of interest assigned to such category against the genes from the population that were in that same category does the computation. Methods in enrichment analysis are Singular Enrichment Analysis, Modular Enrichment analysis and Gene Set Enrichment analysis (M. Mayo and J. Luís 2014). Details of each method are explained the subsections below.

2.6.8.2.1 SINGULAR ENRICHMENT ANALYSIS

Singular Enrichment Analysis (SEA) takes a list of genes that are differentially expressed from obtained from class comparison analysis by performing a statistical test and then uses the list to query different annotation terms one after the other. A threshold of significance is set; terms with p-values lower than this threshold is ordered by the enrichment p-value. Binomial probability, Hypergeometric distribution, Fisher's Exact Test, or Chi-square is the common statistical method used for SEA (M. Mayo and J. Luís 2014).

2.6.8.2.2 MODULAR ENRICHMENT ANALYSIS

Modular Enrichment Analysis (MEA) is an analysis that uses list of differentially expressed genes to test multiple annotation term at once considering the relationship between each pair of terms. Since MEA considers the relationship between each pair of term, it is possible to obtain unique biological meaning, which cannot be shown by a single annotation term (M. Mayo and J. Luís 2014).

2.6.8.2.3 GENE SET ENRICHMENT ANALYSIS

Gene Set Enrichment Analysis (GSEA) is a method used to find out whether a gene set display statistical significant difference between phenotypes by considering the measured difference that exist between them for individual genes from high-throughput experiment. It takes in a list of differentially expressed genes and a set of gene expression values measured from the two phenotypes under study. GSEA depends on thousands of predefined gene sets assembled from GO, KEGG etc. (M. Mayo and J. Luís 2014).

Furthermore, GSEA's statistical method involves testing some gene set from the high-throughput experiment for enrichment by arranging all the genes in the entire list based on the correlation of the gene expression pattern with the phenotype. The arrangement thus gives the fraction of genes that appear in the gene set called hits and the fraction of genes that doesn't appear in the gene set called misses, at a particular position in the entire list of genes. The result here is then used to compute enrichment score with Kolmogorov-Smirnov statistic, z-score or t-test (M. Mayo and J. Luís 2014).

2.6.8.3 GENE LIST ANALYSIS WITH PANTHER

Gene list analysis can be achieved with PANTHER ((Protein Analysis THrough Evolutionary Relationships). PANTHER is a commonly used online tool employed for evolutionary and efficient classification of protein and analysis of biological data on a large scale. It makes use of data set from UniProt Reference Proteomes, which are arranged into families of similar genes. Results from PANTHER is shown on a phylogenetic tree, which is built for each family reveals the evolutionary relationship between all the genes in the family and also transcription, speciation, gene duplication and horizontal transfer processes. Phylogenetic tree from PANTHER thus deduce all the processes above for all the available protein coding genes in organisms and helps to determine groups of genes that are similar within each gene family (H. Mi et al. 2016).

Furthermore, PANTHER gathers proteins, which have been further divided into families that are functionally related using human knowledge. Accurate association with function such as ontology terms and pathways is possible because the subfamilies of proteins in PANTHER model these functions divergence within the families of protein (H. Mi et al. 2005). In addition, PANTHER helps to translate protein sequence relationship to functional relationship in an perfect and flexible way. It is made up two main parts namely; PANTHER library and PANTHER index. PANTHER library refers to different books, with each book serving as a protein family in terms of a Hidden Markov Model (HMM), multiple sequence alignment, and a family tree while PANTHER index and it thus depicts compressed ontology and it performs the function of shortening and maneuvering molecular functions and biological processes associated with the protein families and their subfamilies (P.D. Thomas et al. 2003).

2.7 OVERVIEW OF SURVIVAL ANALYSIS

Survival analysis also known as failure analysis, event history analysis, hazard analysis, duration analysis, or transition analysis is an analysis developed by biostatisticians to predict the occurrence of death. It deals with gathering different statistical methods, which are used to interpret an event, its occurrence and timing. The event to be interpreted can be deaths, births, marriages, divorces, job terminations, migrations etc. (P.D. Allison 2012). The occurrence of an event of interest say death is described with the term "failure", which in some cases might represent a successful event e.g. therapy recovery. The amount of time it take for an event of interest to occur is referred to as the survival time (M. Stevenson 2009).

Furthermore survival analysis deals with analyzing survival time data, which can be obtained by different approaches. The approaches used in obtaining the survival time data thus affect its analysis. The following sampling (data collection) processes are used in generating survival time data (S.P. Jenkins 2005).

a. Stock Sampling Approach

Standard sampling approach is an approach, which involves the random sampling of individuals that are in a particular state of interest (e.g. Tumor relapse). Not all individuals in this state are interviewed and the time at which they entered the state of interest is determined (S.P. Jenkins 2005).

b. Inflow Sampling Approach

This approach involves random sampling of all individuals that are in a defined state of interest and each of the individual in the study are followed until a predefined date or end of the study (S.P. Jenkins 2005).

c. Outflow Sampling Approach

This approach generates survival data by taking random sample of individuals that leaves state of interest. In this approach, the start date of the state of interest is determined (S.P. Jenkins 2005).

d. Population Sampling Approach

This approach involves the overall sampling of the entire population. Sampling in this approach is not related to the state of interest. In this approach, individuals are asked about their current and past state [(S.P. Jenkins 2005).

It should be noted that a survival data can be generated by the combination of any or all of the sampling method stated above depending on purpose the data is meant for (S.P. Jenkins 2005).

2.7.1 CONCEPTS IN SURVIVAL ANALYSIS

In analyzing survival data, it is important to explain the following concepts

2.7.1.1 EVENT

Events refers to a change that takes place at a particular point in time, hence it is dependent on time. Examples of events are deaths, marriages, promotions etc. In understanding survival analysis, it is quite important to note that: firstly, the event to be modeled must be properly defined before proceeding with the analysis of data (P.D. Allison 2012).

Secondly, one must decide whether to treat all events as the same or to distinguish them. The distinctions between different events types depend on the available data, provided it gives enough information for the distinction. Distinguishing events helps to reveal the different effects that predictor variables used in our model has on different event types (P.D. Allison 2012).

Thirdly, when each individual have the different repeated events associated with them, one must decide whether to pick the first event or to employ a technique that merge all the events together (P.D. Allison 2012).

2.7.1.2 TIME TO EVENT

Time to event or survival time is the amount of time it takes for an event to take place. It is usually given in days, weeks, months, years etc. and the following terms helps in describing survival time.

A. Instantaneous Failure Rate

If the count of the length of time taken for an event to occur is plotted as a function of time on a frequency histogram, a curve fitted to this histogram gives the instantaneous failure rate of the event in question. It is denoted by *f(t)* and is shown in figure 1.6 below. Failure function *F(t)* signifies the fraction of individual who have died as a function of time t and it is obtained by setting the area under curve of fig 1.6 to be equal to 1. Then it is observed that for any time *t*, the area under the curve to the left of time t represents the failure function. Hence individual in this portion have experienced the event of interest (death) (M. Stevenson 2009).

B. Survival Function

Survival function is denoted by *S(t)* and is defined as the fraction of individual in a population who have survived to time *t.* It is denoted by the area under the instantaneous failure curve to the right of time as shown in figure 1.6 and it gives a survival curve when plotted against time *t* (M. Stevenson 2009). It is given as

$$\mathbf{S(t) = Probability\ (an\ individual\ survives\ longer\ than\ t)} \qquad \textbf{Eqn 12(a)}$$

$$\mathbf{S(t) = P(T > t)} \qquad \textbf{Eqn 12(b)}$$

Since the area to the left of the curve in figure 1.6 represents the failure function F(t), then, S(t) can also be given as

S(t) = 1-Probability (an individual fails before t) **Eqn 13(a)**

S(t) = 1 - F(t) **Eqn 13(b)**

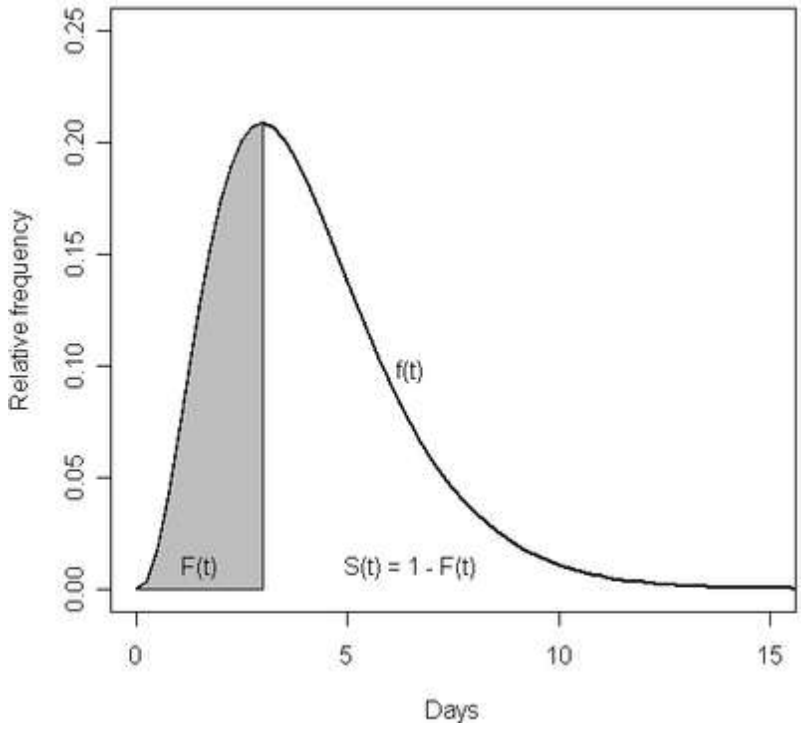

Figure 2.5: Line plot of instantaneous failure rate. Figure adapted from (M. Stevenson 2009).

C. Hazard Function

Hazard function *h(t)* is defined as the instantaneous time at which an individual that is selected randomly are alive at time (*t* -1) dies at time *t.* (E.T. Lee and J.W. Wang 2003). Also known as the conditional failure rate, instantaneous failure rate, conditional mortality rate, or force of mortality, gives the failure probability during a small interval provided the individual has survived to the start of the interval (M. Stevenson 2009).

2.7.1.3 PERIOD OF OBSERVATION

The period of observation in survival analysis refers to the process of observing each individual in the study over an interval of time. The time is thus recorded provided an event occurred, otherwise the observation continues and observation is stopped when an occurred event doesn't repeat itself. For a proper and accurate analysis of survival data, the origin and termination time results should be accounted for and validated (P.D. Allison 2012).

2.7.1.4 CENSORING AND TRUNCATION IN SURVIVAL ANALYSIS

2.7.1.4.1 CENSORING

An event is said to censored if it doesn't occur during the period of observing an individual but the total time in which the event doesn't occur is known. Individuals in this category are called censored observation and the data contributed by censored observations are included in the study until they are excluded from the risk set (M. Stevenson 2009).

However, an event of interest is said to be exact or uncensored if it occurs during the period of observing an individual and the number of days from the start to the end of event is known (E.T. Lee and J.W. Wang 2003). Censoring can be divided into the following

A. RIGHT CENSORING

In right censoring, an event of interest usually occurs after the recorded follow-up period (M. Stevenson 2009). This means that an observation is ended before an individual experiences an event of interest. It is the most conventional type of censoring (P.D. Allison 2012) and it can be sub-divided into the following (E.T. Lee and J.W. Wang 2003).

a. Type I Censoring

In type I censoring, all censored observation corresponds to the length of the study period provided there are no accidental losses such as death (E.T. Lee and J.W. Wang 2003).

b. Type II Censoring

In type Ii censoring, all censored observation corresponds to the length of the largest uncensored observation provided there are no accidental losses such as death. Type I and type II censored observations are also known as singly censored data since the study period here is not fixed and individuals in this study starts at the same time (E.T. Lee and J.W. Wang 2003).

c. Type III Censoring

Here, the study period is fixed and individual can enter the study at different time. It is also called progressively censored data with different censored times (E.T. Lee and J.W. Wang 2003).

In solving right censoring problems, it is best to assume that the censoring doesn't give enough information. For the assumption above to hold, the censoring time of the entire individual in the study must be the same. However, it should be noted that an informative censoring is one in which its censoring time occurs at varying time due to individuals leaving the study at different time (P.D. Allison 2012).

B. LEFT CENSORING

In left censoring, an event of interest usually occurs before the recorded follow-up period (M. Stevenson 2009). It is the least common among censoring types and the exact time before the event of interest occur is unknown (P.D. Allison 2012).

C. INTERVAL CENSORING

In interval censoring, an event of interest usually occurs between two times, but the exact time is unknown. It means that one can know that an event occurs between date A and B but there is no clear knowledge abut the date (M. Stevenson 2009).

2.7.1.4.2 TRUNCATION

Truncation on the contrary means that the outcome of an event of interest cannot probably occur. Truncation is divided into two namely:

A. RIGHT TRUNCATION

Right truncation occurs when an individual leaves a study after the study has commenced and progressed to a particular stage. Hence, individuals here cannot experience the event of interest at the time he left the study (M. Stevenson 2009).

B. LEFT TRUNCATION

Left truncation on the contrary occurs when an individual enters a study after the follow-up period has commenced. Hence, individuals cannot experience the event of interest at the time they enter the study because the follow-up period has already started (M. Stevenson 2009).

2.7.2 DESCRIBING SURVIVAL DATA

Once the survival data has been collected, the next step is to describe it graphically with a survival curve. The plot of this curve helps to reveal the patterns in the survival data, determine data distribution. Data pattern and its distribution are thus important in describing survival data. Methods used in describing data can be non-parametric or parametric (M. Stevenson 2009).

2.7.2.1 NON-PARAMETRIC METHOD

Non-parametric methods of describing survival data are quite simple to understand and implement. They are effective than parametric method when there are no known theoretical distributions and less effective when the time to event follow a theoretical distribution. Hence, it is encourage analyzing survival data with non-parametric methods before attempting to fit a theoretical distribution. These methods thus give estimates and graphs that help in selecting a distribution if we are to find a model for the data at hand (E.T. Lee and J.W. Wang 2003).

Non-parametric method is a method that is used when there is no theoretical distribution that fits the survival data. This method is also known as semi-parametric or and is usually used in epidemiology. The popular non-parametric methods used in describing survival data are Kaplan-Mier method and the life-table method (M. Stevenson 2009).

A. KAPLAN-MIER METHOD

Kaplan-Mier method is a method developed Kaplan and Mier in the year 1958, also called product-limit method and is based on the survival times of an individual. It estimates survival function, probabilities and graphical representation of survival distribution. It is the most wide used non-parametric method of analyzing survival data and is applicable to small, medium and large samples (E.T. Lee and J.W. Wang 2003). The characteristics of Kaplan-Mier method are as follows

a. It assumes that censoring times are not dependent on survival time i.e. a censored observation is independent of the state of interest (E.T. Lee and J.W. Wang 2003).

b. The survival curve produced by this method gives the median survival time when $S(t)$ = 0.5 at time t. If there are times t_1 and t_2 when at $S(t)$ = 0.5, then the median survival time equals any t value between t_1 and t_2 (E.T. Lee and J.W. Wang 2003).

c. Median survival time from this method is unavailable if number of censored observation exceeds the number of uncensored observation (E.T. Lee and J.W. Wang 2003).

d. It is based on individual survival times; hence it is less prone to bias (M. Stevenson 2009).

Kaplan-Mier estimator of survival at time *t* is given as

$$\mathbf{S(t)} = \prod_{k:t_k \leq t} \frac{(m_k - d_k)}{m_k} \qquad \textbf{Eqn 14}$$

Where $\mathbf{t_k}$, k = 1, 2, .., n is the total set of failure times, $\mathbf{d_k}$ is the number of failures at time $\mathbf{t_k}$, m_k is the number of individuals at risk at time t_k and $\boldsymbol{m_k - d_k}$ is the number of individual in the sample who will survive longer than t_k (M. Stevenson 2009).

B. LIFE-TABLE METHOD

It is one of the ancient methods employed for measuring death and interpreting populations' survival experience. Actuaries, demographers, governmental agencies, and medical researchers have made use of the life-table to study survival, population growth, fertility, migration, length of married life, and length of working life etc. Clinical life-tables are developed by epidemiologist to investigate patients with a particular disease who have been followed for a period of time in a study. Other examples of life-table are population life-table and survival life-table (E.T. Lee and J.W. Wang 2003). The life-table method assumes that

a. There are random selections of individual throughout each interval. This assumption leads to bias when intervals are long, however with short intervals, there is no bias (M. Stevenson 2009).

b. Failure rate within an interval is equal for all individuals and doesn't dependent on the probability of survival at other time periods (M. Stevenson 2009).

Furthermore, the life-table method is similar to Kaplan-Mier method but its survival times are grouped into intervals and thus efficient for large data sets (E.T. Lee and J.W. Wang 2003).

2.7.2.2 PARAMETRIC METHOD

Parametric methods are employed in describing survival data when the distribution of the survival function follows predictable pattern. This method is sometimes preferred to non-parametric method because it evaluates survival time easily even after the occurrence of the event of interest (M. Stevenson 2009). Parametric methods used to describe survival time are exponential, Weibull, Gompertz, lognormal, log-logistic, and gamma (P.D. Allison 2012).

2.7.3 COMPARING SURVIVAL CURVE

It is a usual practice to compare the survival between two or more groups. For instance one may compare if it takes a longer time for a particular disease to develop in a one region of a country compered with another region, if a treated patient survive longer than expected etc. The survival curve gives an insight on how the time to event differs among the groups under study. In addition, different survival curves for different group helps to recognize the factors that play important role in deciding survival. The curve thus serves as a tool for effective screening these important factors, which influence can then be tested with multivariate analysis after they are screened (M. Stevenson 2009). The comparism of survival curve for uncensored observation can be done with non-parametric test. The non-parametric tests are Gehan's generalized Wilcoxon test, the Cox-Mantel test, the log-rank test, Peto and Peto's generalized Wilcoxon test and Cox's F-test (E.T. Lee and J.W. Wang 2003).

The log-rank test, which is also called the Mantel log-rank test, the Cox Mantel log-rank test, and the Mantel- Haenszel test, appears to be the most popular test employed in comparing survival curves. It finds application in data with continuous censoring and it assigns identical weight to failure that occur early and that that occur lately. Log-rank test works with the assumption that hazard functions are parallel for the two groups. With this assumption, it picks each time point when a failure event takes place and creates a two by two table, which contains as columns the number of deaths and the overall number of subjects that are under follow up. The observed death, expected death and the variance of

the expected number are calculated and summed over all table to yield a chi-square statistics having a degree of freedom of 1 (E.T. Lee and J.W. Wang 2003).

Furthermore, for each group, the log-rank test also calculates the ratio of observed and expected number of death. This values thus relates the number of death observed to the expected number during follow up with a null hypothesis that he survival curve for the separate groups is the same as that for the combined data (E.T. Lee and J.W. Wang 2003).

3. AIMS OF THE STUDY

The aims of this study are as follows

- Download Glioblastoma Multiforme (GBM) preprocessed data matrix from microarray platform
- Classify genes and samples into subgroups (clusters) with similar features from a set of gene expression profile
- Identify differentially expressed genes among predefined phenotypic group of samples
- Discover class membership of samples from patients based on their gene expression profile
- Interpret differentially expressed genes with PANTHER for pathway and functional enrichment analysis
- Analyze GBM cluster aggressiveness by performing survival analysis on GBM's clinical data
- Verify pathways and functional enrichment analysis result obtained from differentially expressed genes with independent glioblastoma dataset

4. MATERIALS AND METHODS

4.1 SOURCE OF DATA

4.1.1 EXPERIMENTAL DATA

The data was obtained from The Cancer Genome Atlas (TCGA) via the University of Califonia Santa Cruz (UCSC) Cancer Browser as a zipped file named "TCGA_GBM_exp_u133a-2015-02-24.tgz". This means that the GBM gene expression data is obtained from microarray experiment (Affymetrix HT Human Genome U133a). The unzipped file thus contains a preprocessed data matrix, clinical and other relevant information used in this study. The preprocessed data matrix has 12042 rows (genes) and 539 columns (samples) i.e. there are 12042 genes and 539 tumor samples all from patients with GBM. The clinical data has 147 columns, which contains clinical information of GBM patients.

The National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI) collaborated to create TCGA, which has helped to create global and complex maps of the main genomic changes in different types of cancer. Its datasets are available publicly and have been employed in different types of research playing different roles in over one thousand studies of cancer by independent researchers and in the publications from the TCGA research network (The Cancer Genome Atlas 2016).

Furthermore, TCGA has developed a genomic data analysis pipeline, which on a large scale can accurately assemble, pick, and evaluate tissues from human for genomic alterations.
The success enjoyed by the TCGA project louds the effect of teamwork in science and it can be used as a prototype for future project. The Center for Cancer Genomics (CCG) is an NCI initiative that will replace the TCGA in 2017 and it will rely on the success enjoyed by TCGA by making genomic data available publicly using similar model collaboration for extensive genomic analysis (The Cancer Genome Atlas 2016).

In addition, the cancer browser by the University of Califonia Santa Cruz (UCSC) offers a series of web-based tools used to visualize, harmonize and examine cancer genomics data and its corresponding clinical data. The UCSC Cancer Browser is developed and maintained by the collaborative effort of the UCSC Cancer Genomics and UCSC Genome Browser groups. Both groups operate in the Center for Biomolecular Science and Engineering (CBSE) located at the University of California Santa Cruz (University of California Santa Cruz Genome Browser 2016).

4.1.2 VALIDATION DATA

The GBM dataset employed for result validation is obtained from the Gene Expression Omnibus (GEO) database via the National Center for Biotechnology Information (NCBI) website as a zipped file named "GSE20736_RAW". The unzipped file contains the raw CEL files, which are used in this study for validation.

GEO is a global archive, available to the general public. . GEO stores and distributes freely, next-generation sequencing, microarray, and different high-throughput functional genomic data, obtained from various research communities. Its main goals are to provide a database that is flexible, functional, and can efficiently store high-throughput functional genomic data. In addition, the archive provides a database that allows users to examine, detect, inspect and download datasets from experiments of their interest; and beyond it offers easy submission methods as well as schemes that back complete and well annotated data deposits (National Center for Biotechnology Information 2016)

4.2 MATERIALS USED IN DATA ANALYSIS

The major tool used in the analysis of the data obtained form The Cancer Genome Atlas (TCGA) is R and packages from Biocondutor.

The programming language R is a GNU project with an environment developed by John Chambers and colleagues at the Bells laboratories (Lucent Technologies). The programming language is used for analytical computing and graphical visualization. It is a highly flexible

programming language, which can perform series of statistical analysis such as nonlinear and linear modeling, standard statistical tests, time-series analysis, classification, class discovery, regression analysis, and graphical techniques (R Project for Statistical Computing 2016).

Furthermore, R gives a standard graphical plots when used for graphical display and it also has in built mathematical symbols and formulae. It is available as free software and runs on t UNIX platforms, FreeBSD, Linux, Windows and MacOS. The R environment includes efficient data manipulation and storage facilities, collection of standard tools for data analysis and graphical visualization, and operators for matrices manipulation (R Project for Statistical Computing 2016).

On the other hand, Bioconductor is an open development software project, which is available freely online. It offers the tools for extensive analysis of high throughput genomic data using R programming language. In addition, it contains many metadata packages, which provides pathway, organism, microarray and other annotations. Bioconductor has both development and release versions with the later updated two time in a year and relevant for most users and the former has new features and packages, which are added before the incorporation of the release version (Bioconductor 2016).

Furthermore, Bioconductor project aims to give public access to many efficient statistical and graphical techniques for genomic data analysis, enable the addition of biological metadata in genomic data analysis, produce excellent documentation and reproducible research for improve scientific findings and train researchers on computational and statistical techniques employed in genomic data analysis etc. (Bioconductor 2016).

4.3 CLASS DISCOVERY ANALYSIS

This analysis aims to find subgroup of genes and samples that are identical. Both hierarchical clustering and non-hierarchical clustering was performed on the gene expression data matrix. Since there are 12042 rows (genes) and 539 columns (samples), clustering the whole gene expression data matrix might be tedious.

4.3.1 HIERARCHICAL CLUSTERING

It is however important to first perform some analysis on the data matrix before carrying out hierarchical clustering. First, the data is filtered. Filtering is done with the gene filter package from the Bioconductor. It helps to remove genes that are uninteresting, and this reduces the number of rows (genes) in the data matrix while the number of columns (samples) remains the same. Next the correlation between samples are computer after finding the variance of the rows in the matrix, ordering the filtered data with the variance-calculated to get the top 100 row (genes) with the highest variance and scaling the matrix. Scaling makes the data matrix to have mean of zeros and standard deviation of ones. The correlation heatmap is plotted to compare samples that are similar.

The heat map is a fusion of a lot of different graphical display, developed by statisticians over 100 years ago. Also called cluster heat map is a diagram that shows at the same time the hierarchical cluster structures of row and column in data matrix. It consist of a colored rectangular blocks, which are shaded by a predefined color code and represents the values of the matrix element. The rows and columns of the shaded rectangular blocks are arranged such that rows and columns that are similar are next to one another and the block also has hierarchical cluster trees on its vertical and horizontal margins (L. Wilkinson and M. Friendly 2009).

The correlation matrix is converted to a “dist” object and a dendogram is drawn with “ward.D” hierarchical clustering method. The dendrogram thus helps to group the samples and the genes into two groups (i.e. group A and group B samples), which will be useful for class comparison analysis.

4.3.2 CLUSTER VALIDATION

The result of a clustering analysis can be validated with an R package called “clValid”. The “clValid” R package contains several method employed in clustering result validation. It has three validation methods namely internal, stability, and biological validation measures. The ”clValid” package main function is given as **clVAlid(),** this function thus allows the simultaneous selection of numerous clustering algorithms, validation measures, clusters number in a one time call in order to determine the best clustering method to use and the perfect number of clusters for the dataset. In addition, clValid package calculates biological validation measures automatically by employing the biological information the Gene Ontology (GO) database through Bioconductor annotation packages (G. Brock et al 2008).

Cluster validation is of three different types namely internal, stability and biological validation. Internal validation is a clustering validation measure that takes as input the dataset and the number of clusters to assess clustering quality using the underlying information in the data. In addition, the compactness, connectedness and separation of the cluster partitions are measured with internal validation. Compactness determines how homogenous clusters are by observing intra-cluster variance while separation estimates the distance between the centroids of the clusters and hence it evaluates separation between clusters. Compactness and separation are both combined in Dunn index and Silhouette Width, along side compactness and the three are displayed as internal validation measures result. The ratio of the minimum distance that exist between observations that doesn’t belong to the same cluster to the largest intra-cluster distance is given as the Dunn Index while the Silhouette Width measures an observation’s Silhouette value i.e. Silhouette value assign a value close to 1 for well clustered observations and a value close to -1 for badly clustered observations (G. Brock et al 2008).

Stability validation measures compares how consistent clustering results are by comparing this result with the cluster obtained when the column is deleted one after the other. It is a special form of internal validation measure and it worked well for highly correlated data, such as that obtained from high-throughput experiment. The measures in stability

validation are the average proportion of non-overlap (APN), the average distance (AD), the average distance between means (ADM), and the figure of merit (FOM) (G. Brock et al 2008).

Biological validation measures the competence of a clustering algorithm to produce a result that is relevant biologically. This measures finds application in microarray data, where the observations are represented as genes by using biological homogeneity index (BHI) and biological stability index (BSI). BHI examines clusters biological homogeneity while BSI investigates the clustering consistencies for genes that perform similar biological functions (G. Brock et al 2008).

4.4 CLASS COMPARISON ANALYSIS

This analysis involves selecting genes that are differentially expressed from the gene expression data. This was done with fold change and t-test method.

4.4.1 THE FOLD CHANGE METHOD

The fold change method involves calculating the mean expression level for each condition and then selecting genes that has fold changes that are greater than an approximate threshold. The filtered data matrix is used to calculate the arithmetic means and the log fold change, which corresponds to the log ratio of the mean intensities of two groups of samples. The two groups of samples are obtained from the hierarchical clustering results since the samples are not predefined (S. Drăghici 2011).

The result obtained here can be shown on a scatter plot, which is similar to a Minus-Add plot and has two lines that correspond to the threshold used. The scatter plots can be plot of log fold changes versus log intensities and log intensities of one group versus that of the other group (S. Drăghici 2011).

4.4.2 T-TEST METHOD

The T-Test Method is a parametric method of selecting genes that are differentially expressed. This method deals with multiple testing issues which normally arises during microarray analysis by first filtering the genes from the gene expression data matrix. Gene filtering is done with the gene filter package and it helps to get rid of array spike control, remove probes with little irregularity among samples or those with low level of expression. Probes removal thus makes the number of hypothesis to be tested to reduce and also reduces the number of uninteresting, differentially expressed genes (S. Drăghici 2011).

The t-test is carried out between the sample groups with the null hypothesis there is no particular gene on the array that is differentially expressed and an alternative hypothesis that the gene in question has a different expression level between the sample groups (A.L. Tarca et al 2008). After performing the t-test between the sample groups from the data matrix, the p-values obtained are plotted on a histogram. Multiple adjustments depend on the shape of the plot of p-value.

In this study, differentially expressed genes are selected by combining the fold change and the t-test method using different thresholds for each method. This thus helps to get the genes that are truly differentially expressed. The data matrix obtained contains the truly differentially expressed genes, as its row and the samples remains its column.

4.5 CLASS PREDICTION ANALYSIS

Class prediction analysis comprise of creating of a classification function that can correctly deduce a patient biologic group or his diagnostics category based on the expression pattern of a tissue from the patient in question. The phenotype classes are stated priori and this does not depend on the gene expression data, hence it a supervised analysis. The major steps in class prediction analysis are feature selection, selecting predictor function (classifier) and assessing classifier performance (R. Simon 2003).

The future selection step involves selecting genes that are differentially expressed from the expression data. These genes are informative and they are used in building the classifier (R. Simon 2003). Differentially expressed genes are obtained with the methods describes in section 3.3 above and the result from this section is adapted as the informative genes for this analysis.

The selection of a predictor function (classifier) is the next step after feature selection. Several classifiers such as linear discriminant analysis, nearest neighbor predictor, logistic regression or support vector machine can be adopted depending on the problem at hand (R. Simon 2003). It should be noted that data must be first divided before a classifier can be used to build a model, thus classifier performance assessment is incorporated into the model-building algorithm.

The KNN classifier is a straightforward algorithm that assembles all cases available and proceeds to categorize them using the majority vote of its k nearest neighbor. It thus classifies uncategorized data into groups that are well defined. The algorithm starts by creating the training and test set from the entire expression data set by dividing the data into two portions with larger path of the data as the training set and the rest as the test set. The KNN classifier is then used to train the model using classes from training sets and specifying the value of k as the number of samples in the expression set. KNN classifier selects the nearest neighbor using euclidean distance as its distance metric (B. Lantz 2015).

Furthermore, the performance of the built model is displayed in a confusion matrix. The confusion matrix is build by the table() function in R. However the CrossTable() function from the gmodel package in R gives a more detailed confusion matrix called the contingency table. From this table, accuracy and error rate is computed in percentage. A high value of accuracy and low value of error rate, say 95% and 0.05% implies that the built classifier is nearly perfect. However, a low accuracy value and a high error rate indicate a poor classifier and the classifier has to be re-built (B. Lantz 2015).

4.6 ENRICHMENT ANALYSIS

Enrichment analysis helps in obtaining biological significance in term of statistical significance and to interpret gene group in order to identify biological information for the event under study. Modular enrichment analysis is the method of enrichment analysis is employed and it involves using a list of differentially expressed genes to test multiple annotation term at once considering the relationship between each pair of terms (Mosquera Mayo and José Luís 2014).

Gene list analysis is performed on PANTHER by first pasting or uploading a gene list into the database. The pasted gene list needs to be similar with the supported IDs (e.g. Ensemble Gene ID, Entrez Gene ID etc.) on PANTHER database and the uploaded list has to be a file, which is in a format recognized by the PANTHER database. The list in question may be an identifier list, previously exported text search or a PANTHER generic mapping file. The next step involves selecting homosapiens as the organism of interest since the gene list is from human. Next, the analysis carried out is selected on the database as functional classification viewed in pie chart or bar chat. The result here is viewed as PANTHER GO-slim Biological Process which is the default Annotation and other Annotations such as PANTHER GO-slim Molecular Function, PANTHER GO-slim Cellular Component, PANTHER Protein Class and Pathway are obtained by using the drop down menu in the close to the Annotation data set. The chart showing the different 100% of the categories involves in each annotation analysis (PANTHER user manual 2015).

4.7 SURVIVAL ANALYSIS

Survival analysis is carried out using the non-parametric method i.e. Kaplan-Mier method and plotting the resulting survival curve.

The Kaplan-Mier method is the most wide used non-parametric method of analyzing survival data and is applicable to small, medium and large samples. It estimates survival function, probabilities and graphical representation of survival distribution by assuming that the censoring time does not depend on survival time (E.T. Lee and J.W. Wang 2003).

First the clinical data is read into R and the columns that contain the important parameter for this analysis are noted. These columns (event, time to event and sample ID) are extracted and used for plotting the Kaplan-Mier survival curve. The median survival times of the curves are computed and the curves are compared statistically with log-rank test. The comparison aims to determine if these curves are the same or not. In comparing survival curve with log-rank test, the following steps are involved.

a. Define hypothesis and determine significant level
    - H_o (Null hypothesis) - The survival curves from the two groups are identical
    - H_1 (Alternate hypothesis) - The survival curves from the two groups are not identical (i.e. α = 0-05) (Lisa Sullivan 2016).

b. Select appropriate statistics (Lisa Sullivan 2016).
    - The "survdiff" function from the survival package in R gives the result as chi-square statistics with degree of freedom and resulting p-value.

c. Define the decision rule to adopt
    - The critical value of chi-square distribution from chi-square table is for degree of freedom 1 (i.e. two group comparison) and α = 0-05. The chi-square critical value is 3.841 and the decision rule is to reject H_0 if chi-square critical value is greater than > 3.841 (Lisa Sullivan 2016).

d. Compute the test statistics by running the "survdiff" function and make conclusions.

4.8 RESULT VERIFICATION USING INDEPENDEDNT GBM DATASET

The verification dataset is obtained from the Gene expression Omnibus (GEO) database of the National Center for Biotechnology Information (NCBI). The dataset has an accession number of GSE20736_RAW and contains 6 GBM tumor samples, which are obtained from the Affymetrix Human Genome U133 Plus 2.0 Array platform. The file containing these samples are downloaded as zipped filled, unzipped and read into R as raw CEL files. Probe effects (RLE & NUSE plots), intensity plot, and box plot are used to access the qualities of the 6 raw CEL files. These plots reveal that the CEL files are raw in nature.

Furthermore, the CEL files are preprocessed using the rma() function, which performs background correction and normalization on the raw CEL files to give a normalized data. for The expr() function is applied to the normalized data to obtain the data matrix needed. The data matrix is in turn filtered in other to remove irrelevant genes. Correlation matrix is first created to define relationships that exist between samples and dendrograms are drawn for both samples and genes to perform hierarchical clustering on samples and genes correlation matrices. Differentially expressed genes are then obtained in a similar way as explained in section 4.4 above. The verification process is carried out by transforming the differentially expressed genes into biological meaning with the help of the PANTHER database. The results from experimental data are compared with the one from here in terms of molecular function, biological process, cellular components and pathways in which both differentially expressed genes play different roles.

5. FINDINGS

5.1 CLASS DISCOVERY ANALYSIS

5.1.1 HIERARCHICAL CLUSTERING

The data matrix derived from the gene expression matrix is clustered so that similar samples and genes are grouped together. It will be difficult to cluster the entire data matrix since the matrix has 12042 rows (genes) and 539 columns (samples).

A filtered data with of 3128 rows (genes) and 539 columns (samples) is obtained after filtering. A sample-sample Pearson's correlation matrix with 539 rows and 539 columns, that shows correlation between samples is obtained. This is done by finding and selecting the top 100 rows (genes) from the matrix with the highest variance. The result is ordered with the filtered data and finally scaled. A correlation heatmap is plotted using hierarchical clustering with ward's method to compare samples that are similar. In figure 5.1 below, the red color specifies the samples that correlate positively with each other and it implies that there are highly correlating genes between this samples. On the other hand, the green color indicated samples that correlate. Positive correlation also indicates a high similarity in sample's expressions within a group while negative correlation is the inverse. In addition, figure 5.1 also shows that samples are divided into two distinct clusters with each cluster having two sub-clusters each.

Dendrograms are drawn for both genes and samples by performing hierarchical clustering on the correlation matrices that relate to the genes and samples respectively. Both dendrograms are shown in figures 5.2a and 5.2b respectively. From figure 5.2a, there are two sample groups with group A having 253 samples and group B having 286 samples. Figure 5.2b shows two gene groups with group A having 60 genes and group B having 40 genes. Samples and gene group numbers are shown in table 4.1 below

GROUP	SAMPLES	GENES
A	253	60
B	286	40

Table 5.1: Samples and Genes groups

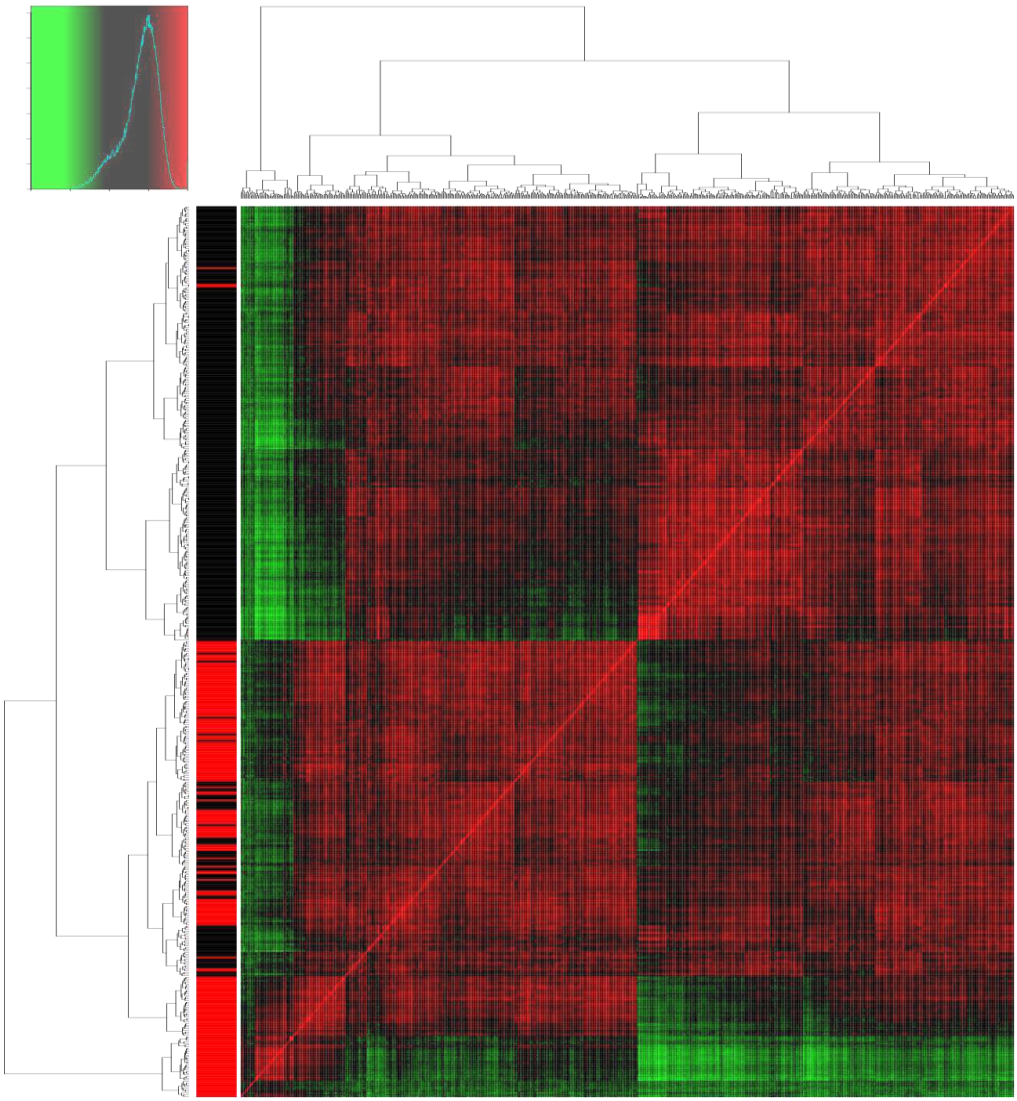

Figure 5.1: Correlation heatmap of data matrix

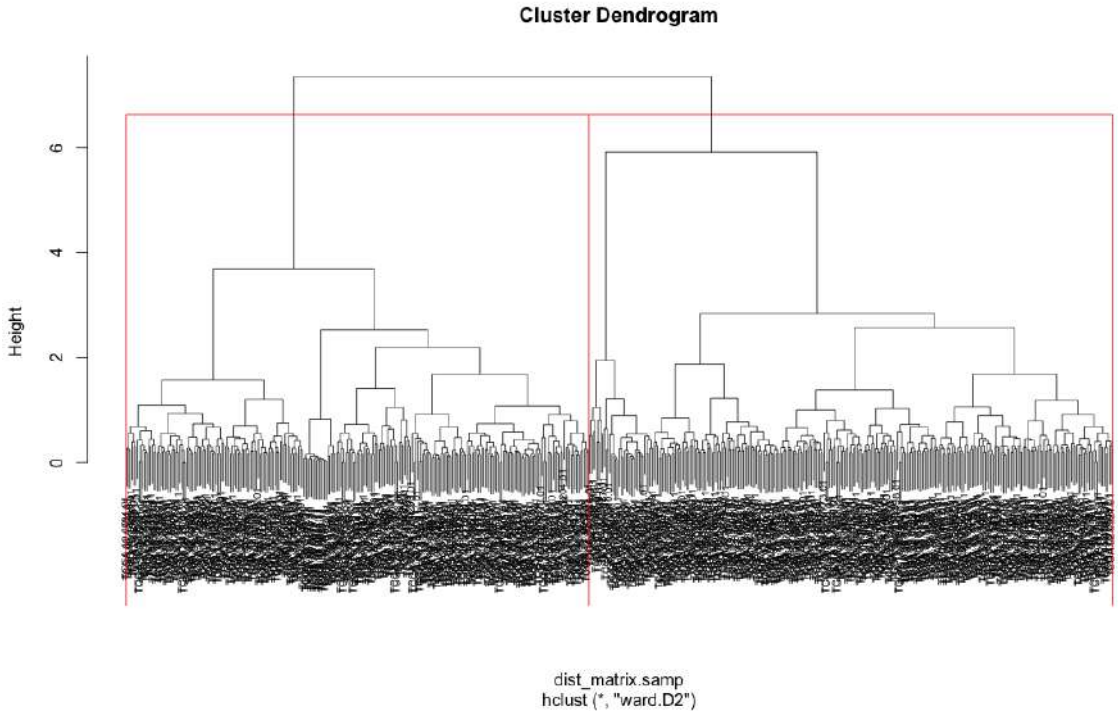

Figure 5.2A: Sample dendrogram

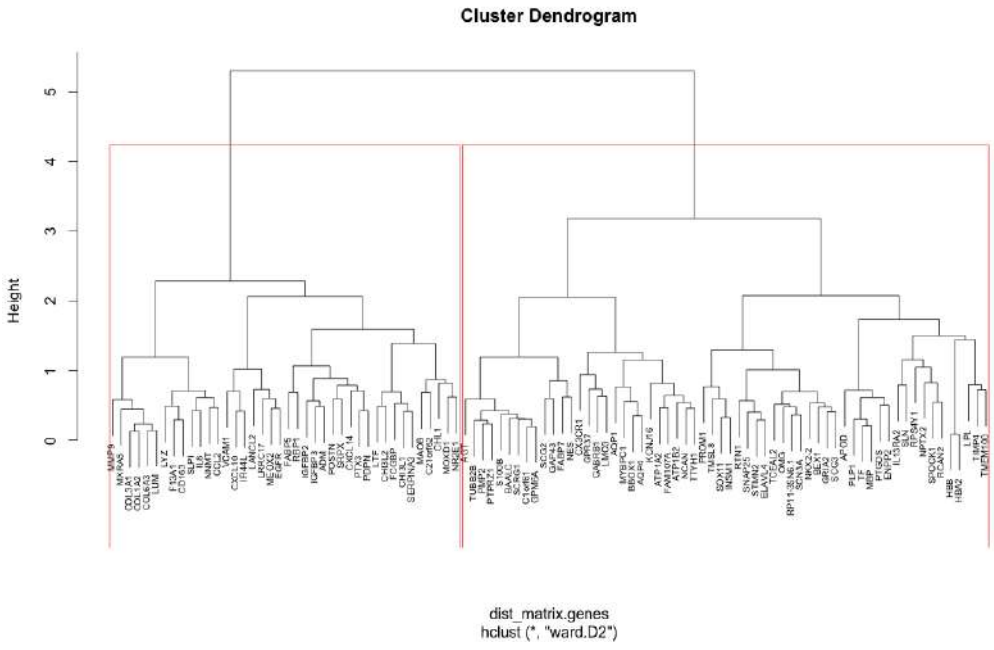

Figure 5.2B: Gene dendrogram

5.1.2 CLUSTER VALIDATION WITH "CLVALID" R PACKAGE

The result of samples and genes clustering is validated with an R package called "clValid". For sampled clustering, the scaled data that has genes as its rows and samples as its column is transposed. The transpose is carried out since the "clValid" package performs row clustering. Hence the need to have a transposed scaled data, which has the samples as its rows and genes as its column. The clValid" function test for hierarchical and k-means clustering for both the samples and genes from the expression data matrix. The validation considers two to five clusters and the results are shown in table 5.2 and 5.4 below.

S/N	CLUSTERING METHOD	VALIDATION PARAMETERS	NUMBER OF CLUSTERS AND SCORES			
			2	3	4	5
1	HIERARCHICAL CLUSTERING	Connectivity	11.2060	14.9933	17.8683	25.1520
		Dunn	0.4382	0.4382	0.4382	0.4382
		Silhouette	0.2589	0.2410	0.2056	0.2024
2	K-MEANS CLUSTERING	Connectivity	157.0353	145.5774	160.9516	167.7873
		Dunn	0.35492	0.3645	0.3780	0.3780
		Silhouette	0.1316	0.1372	0.1278	0.125

Table 5.2: Sample cluster validation result

From table 5.2 above, a table showing the clustering method, numbers of clusters, validation parameters, and their scores is generated.

S/N	VALIDATION PARAMETERS	SCORE	CLUSTERING METHOD	NUMBER OF CLUSTERS
1	Connectivity	11.2060	Hierarchical	2
2	Dunn	0.4382	Hierarchical	2
3	Silhouette	0.2589	Hierarchical	2

Table 5.3: Optimal scores for sample cluster validation

From table 5.3 above, the three internal validation parameters give two clusters and suggest hierarchical clustering as the appropriate clustering technique for the data. Larger values of Dunn index and Silhouette width indicates a good cluster while a lower connectivity value indicates a good cluster. This means that the Dunn Index and Silhouette width internal validation parameters are maximized while the connectivity internal validation parameter is minimized. Hence, hierarchical clustering with two clusters gives a better performance. In addition, it is clear that the samples that appeared as the column in the transposed version of the scaled data has two clusters and hierarchical clustering is the appropriate clustering technique that should be employed in clustering the data. Hence, the result obtained earlier with the dendrogram in section 5.1.1 has been validated.

For gene clustering result validation, the same method is employed. The data matrix used here is the scaled data that has genes as its rows and samples as its column. The result of genes clusters validation testing for hierarchical and k-means clustering for two to five clusters are shown in table 5.4 below.

S/N	CLUSTERING METHOD	VALIDATION PARAMETERS	NUMBER OF CLUSTERS AND SCORES			
			2	3	4	5
1	HIERARCHICAL CLUSTERING	Connectivity	2.9290	13.9171	16.8460	19.7750
		Dunn	0.5756	0.2383	0.2383	0.2383
		Silhouette	0.2246	0.2342	0.1876	0.1727
2	K-MEANS CLUSTERING	Connectivity	30.7845	33.0607	43.9794	50.0595
		Dunn	0.2637	0.2637	0.3438	0.3328
		Silhouette	0.2379	0.2381	0.1799	0.1616

Table 5.4: Gene cluster validation result

From table 5.4 above, a table showing the clustering method, numbers of clusters, validation parameters, and their scores is generated. Larger values of Dunn index and Silhouette width indicates a good cluster while a lower connectivity value indicates a good cluster. This means that the Dunn Index and Silhouette width internal validation parameters are maximized while the connectivity internal validation parameter is minimized. It is clear from table 5.5 below that Connectivity and Dunn validation parameters give two clusters and suggest hierarchical solution as the appropriate clustering technique while the Silhouette validation parameter gives three clusters and suggest Kmeans clustering as the appropriate clustering technique that should be adopted. Since two validation parameters gives two clusters and suggest hierarchical clustering as the appropriate clustering, it implies that hierarchical clustering with two clusters gives a better performance compared to the Kmeans clustering with five clusters. Hence, the result obtained from dendrogram in section 5.1.1 has been validated.

S/N	VALIDATION PARAMETERS	SCORE	CLUSTERING METHOD	NUMBER OF CLUSTERS
1	Connectivity	2.9290	Hierarchical	2
2	Dunn	0.5756	Hierarchical	2
3	Silhouette	0.2381	Kmeans	3

Table 5.5: Optimal scores for genes cluster validation

5.2 CLASS COMPARISON ANALYSIS

5.2.1 FOLD CHANGE METHOD

Fold change method is used to find the differentially expressed genes between the two sample groups obtained from hierarchical clustering analysis. With an absolute fold change greater than 1.0, the top 20 differentially expressed obtained are shown in table 5.6 below.

4.2.2 T-TEST METHOD

The t-test method is used to find the differentially expressed genes between sample groups obtained from hierarchical clustering analysis. The p-values obtained from t-test are plotted on the histogram as shown in figure 5.3 and the plot thus determines the need for p-value adjustment (multiple testing). The plot in figure 5.3 shows the p-value obtained from the t-test analysis with the horizontal axis showing the p-values and the vertical axis showing he number of times each p-value occurs. It is clear from this figure that the set of p-values obtained gives us evidence against the null hypothesis. This is because the flat distribution to the right of the figure contains the entire null p-values (null hypotheses), which are distributed uniformly between 0 and 1. In addition, the peak to the left (close to zero) contains the alternative hypotheses. Since majority of the null hypotheses are obtained at low p-values, then all p-values less than 0.05 are called significant. The top 20 genes with a p-value less than 0.05 are shown in table 5.6 below.

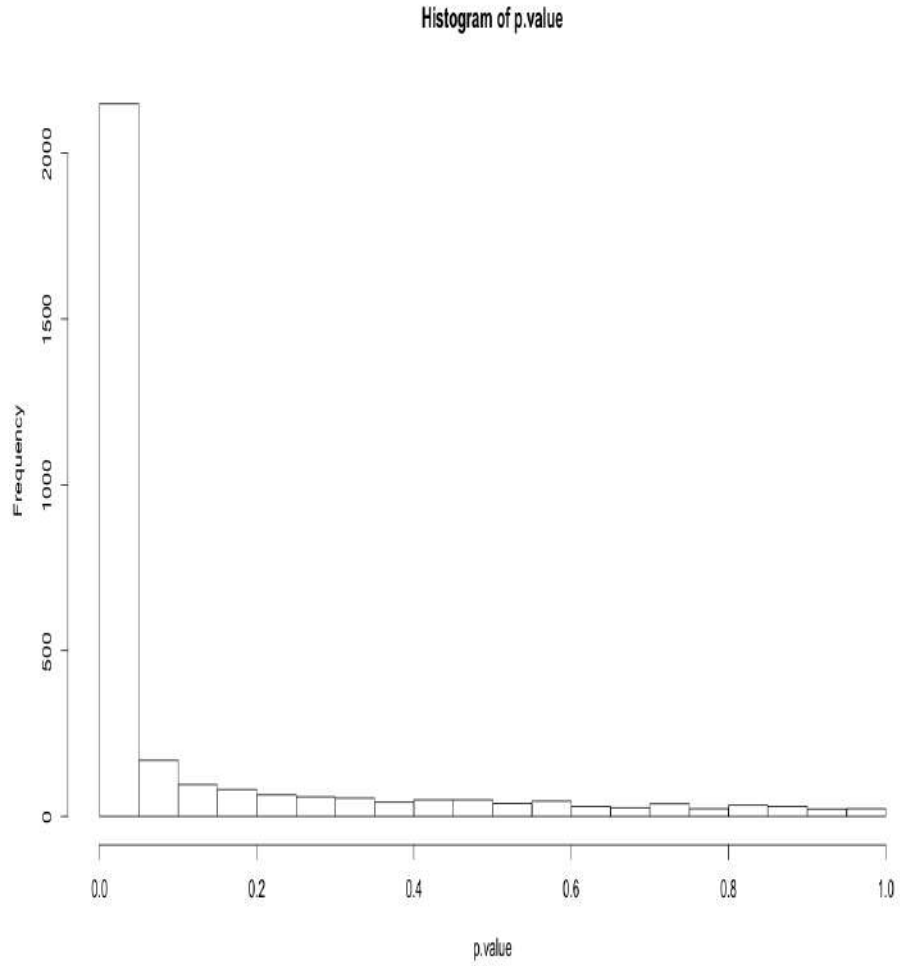

Figure 5.3: Histogram of p-values from t-test analysis

However, genes that are truly differentially expressed are obtained by a fold-change cutoff greater than 1.0 and a p-value less than 0.05. A total of 1351 genes are obtained as differentially expressed genes between the two sample groups. The top 20 differentially expressed genes are shown in table 5.6 below. From the table, some of the genes appeared as fold change genes and t-test genes simultaneously. Hence, this combination gives a better result because both technical and biological significance are considered at the same time.

S/N	FOLD CHANGE GENES	T-TEST GENES	FOLD CHANGE AND T-TEST DE GENES
1.	UBE2Q1	RNF11	RNF14
2.	RNF10	RNF13	UBE2Q1
3.	RNF11	PMM1	RNF10
4.	RNF13	ASS1	RNF11
5.	NDP	NID2	RNF13
6.	PMM1	ZC3H14	NDP
7.	ZC3H15	DHX9	PMM1
8.	ZC3H14	GRINA	ASS1
9.	RNF111	NUP93	NID2
10.	NACAP1	OPA1	ZNF706
11.	DHX9	RAB40B	ZC3H15
12.	XPC	KIAA0831	ZC3H14
13.	GRINA	UGCG	RNF111
14.	SP3	ATP2A2	NACAP1
15.	NUP93	RIT1	DHX9
16.	GOLIM4	ITGA7	XCP
17.	OPA1	DENND4B	GRINA
18.	RAB40B	SWAP70	SP3
19.	KIAA0831	PHLDA1	NUP93
20.	ATP2A2	GAP43	GOLIM4

Table 5.6: Differentially expressed genes from experimental data

For the validation data, the same method used in section 5.2.1 and 5.2.2 was repeated and the lists of differentially expressed genes are shown in table 5.7 below.

S/N	FOLD CHANGE GENES	T-TEST GENES	FOLD CHANGE AND T-TEST DE GENES
1.	DDR1	EPHB3	DDR1
2.	RFC2	C4orf33	RFC2
3.	PAX8	C6orf141	PABX
4.	UBA7	JAK1	UBA7
5.	EPHB3	CARD16	EPHB3
6.	ESRRA	CARD16	ESRRA
7.	SCARB1	CASP1	SCARB1
8.	TTLL12	TLR4	TTLL12
9.	MAPK1	RAB46	MAPK1
10.	PXK	ALKBH5	PXK
11.	C9orf30	NFX1	PXK
12.	AFG3L1	STK38	C9orf30
13.	PIGX	FOXC1	AFG3L1
14.	SLC39A13	NPB	PIGX
15.	NEXN	ZNF791	SLC39A13
16.	MFAP3	ZNF791	NEXN
17.	CNOT7	TMTC4	MFAP3
18.	CRYZL1	CLDND1	CNOT7
19.	LEAP2	CENPL	CRYZL1
20.	C4orf33	TIPRL	LEAP2

Table 5.7: Differentially expressed genes for independent GBM data (validation data).

5.3 CLASS PREDICTION

Class prediction analysis intends to build a model to predict the class membership of the sample groups obtained from hierarchical clustering analysis. First, features are selected, then a classifier is chosen. The chosen classifier is used to build a model, and finally the built model evaluated by computing model assessment parameters such as accuracy, error, sensitivity, specificity etc. This analysis gives a contingency table (confusion matrix) as shown in table 5.8 below.

		FORESIGHTED VALUE	
		Negative	**Positive**
ORIGINAL VALUE	**Negative**	**T.N** **True Negative**	**F.P** **False Positive**
	Positive	**F.N** **False Negative**	**T.P** **True Positive**

Table 5.8: Contingency table

This analysis is done with the data matrix obtained from class comparison analysis. Since the data matrix contains the differentially expressed genes as its rows and the samples as its column, then it is transposed in other to have the samples as columns and the differentially expressed genes as rows. Samples are first splitted into training and testing sets. Two-third of the total number of samples (539 samples) are used to create the training set (359 samples) while one-third of the samples give the testing set (180 samples). Then a KNN classifier is adopted and the value of K used is the odd number close to the square root of the training set size (i.e. √359 ≈ 19). Training sets are used in building the model while the training sets are used in testing the built model. Table 5.9 below shows the True Negative (TN), False Negative (FN), True Positive (TP) and False Positive (FP) values obtained with K= 19. It should be noted that the total number of items in table 5.9 equals the number of items in the training set.

	KNN PREDICTION	
TESTING CLASS	**SAMPLE A**	**SAMPLE B**
SAMPLE A	64	31
SAMPLE B	6	79

Table 5.9: KNN Contingency table for K of 19

Other K values are also tested in the other to check for the K value that gives an improved model. Their results are shown in table 5.10 below.

S/N	K-VALUES	TESTING CLASS/ KNN PREDICTION		
1	17		**SAMPLE A**	**SAMPLE B**
		SAMPLE A	66	29
		SAMPLE B	6	79
2	15		**SAMPLE A**	**SAMPLE B**
		SAMPLE A	69	26
		SAMPLE B	7	78
3	13		**SAMPLE A**	**SAMPLE B**
		SAMPLE A	69	24
		SAMPLE B	6	79
4	11		**SAMPLE A**	**SAMPLE B**
		SAMPLE A	71	24
		SAMPLE B	5	80
5	9		**SAMPLE A**	**SAMPLE B**
		SAMPLE A	69	26
		SAMPLE B	7	78
6	7		**SAMPLE A**	**SAMPLE B**
		SAMPLE A	69	26
		SAMPLE B	6	79
7	5		**SAMPLE A**	**SAMPLE B**
		SAMPLE A	68	27
		SAMPLE B	9	76

Table 5.10: Contingency table for different values of K.

The best K value is however selected form table 5.10 above and it is used for model performance assessment. It is clear from table 5.10 below that the K value of 11 gives a better result since it gives more true negatives and true positives, as well as lower false negative and false positives, when compared with other K values results. Hence it is adopted and used for model performance assessment. The contingency table for model with K value of 11 is shown in figure 5.11 below.

Cell Contents

N
N / Row Total
N / Col Total
N / Table Total

Total Observations in Table: 180

	KNN PREDICTION		
TESTING CLASS	**SAMPLE A**	**SAMPLE B**	**ROW TOTAL**
SAMPLE A	**71** 0.747 0.934 0.394	**24** 0.253 0.231 0.133	**95** 0.528
SAMPLE B	**5** 0.059 0.066 0.028	**80** 0.941 0.767 0.444	**85** 0.472
COLUMN TOTAL	**76** 0.422	**104** 0.578	**180**

Table 5.11: Contigenct table for K value of 11

The following model performance parameters are computed from the values obtained in table 4.11 above.

A. ACCURACY

Accuracy = (T.P + T.N) / (T.P + T.N + F.N + F.P)

= (80 + 71) / (80 + 71 + 5 + 24) = 151/180

= 0.839

= 83.9%

This implies that the classifier is 80% accurate.

B. ERROR RATE

Error rate = 1 - Accuracy

= 1 – 0.839

= 0.161

= 16.1%

The error rate here is 16.1% and it can be attributed to system bias.

C. SENSITIVITY

Sensitivity = T.P / (T.P + F.N)

= 80 / (80 + 5)

= 0.941

= 94.1%

This implies that 94.1% of sample A were correctly classified. This mean that 80 samples out of 85 samples classified as sample A are correctly classified while 5 samples are incorrectly classified.

D. SPECIFICITY

Specificity = T.N / (T.N + F.P)

= 71 / (71 + 24)

= 0.747

= 74.7%

This implies that 74.7% of sample B were correctly classified. This mean that 71 sample out of 95 samples classified as sample B is correctly classified while 24 samples are incorrectly classified.

E. PRECISION (Positive Predictive Value)

Precision = T.P / (T.P + FP)

= 81 / (81 + 24)

= 0.771

= 77.1%

Since precision defines how exact a classifier is, it is clear that a classifier with a precision value of 0.771 (77.1%) is almost perfect and has a lot of true positives.

F. RECALL (SENSITIVITY)

Recall = T.P / (T.P + F.N)

= 80 / (80 + 5)

= 0.941

= 94.1%

Since recall measures how complete a classifier is, it implies that the classifier above is 94.1% complete.

G. F-MEASURE

F-Measure = (2 * PRECISION * RECALL) / (PRECISION + RECALL)

= (2 * 0.771 * 0.941) / (0.771 + 0.941)

= 0.848

= 84.8%

This implies that the performance of the classifier is 84.8%, which is nearly effective.

In summary, all the measures above have values close to 1. Based on this analysis, separation of two classes is established.

5.4 ENRICHMENTS AND FUNCTIONAL ANALYSIS

The list of differentially expressed genes is translated into biological terms with PANTHER for both the experimental and verification data and the results obtained are shown below.

5.4.1 MOLECULAR FUNCTION (MF)

The molecular functions in which the differentially expressed genes from both data perform are shown in the table 5.12 below. It is observed that 32 genes out of the total 50 differentially expressed genes from the experimental data performs molecular functions while 30 genes out of the total 50 DE genes from the validation data performs molecular function. Also the DE genes from both data perform similar molecular functions (i.e. five). From the results displayed in from table 5.12, analysis result is validated.

5.4.2 BIOLOGICAL PROCESS (BP)

The biological processes in which the differentially expressed genes from both data play roles are shown in the table 5.13 below. It is observed that 80 genes out of the total 50 differentially expressed genes from the experimental data perform molecular functions while 60 genes out of the total 50 DE genes from the validation data play role in biological processes. It implies that some genes are involved in more than one biological process for both data. Also the DE genes from both data are involved in similar biological processes (i.e. ten) except for locomotion (GO:0040011) biological process, which is performed by

NEXN and JAK1 genes in the validation data. From the results displayed in from table 5.13, analysis result is validated.

5.4.3 CELLULAR COMPONENT

The differentially expressed genes from both data are part of the cellular components shown in the table 5.14 below. It is observed that 40 out of 50 differentially expressed genes from the experimental data are part of cellular components and it implies that some genes are part of different cellular components. For the validation data, 31 out of 50 DE genes are part of different cellular components shown in table 5.14 below. Also the DE genes from both data are part of similar cellular components (i.e. five), except for Synapse (GO:0045202) cellular component, which is present in the experimental data and extracellular region (GO:0005576) present in validation data. From the results displayed in from table 5.14, analysis result is validated.

5.4.5 PATHWAY

The pathways in which the differentially expressed genes from both data play roles are shown in the table 5.15 below. It is observed that 29 genes out of the total 50 differentially expressed genes from the experimental data play roles in the pathway while 38 genes out of the total 50 DE genes from the validation data play roles in the pathway. Also the DE genes from both data play roles in similar pathways (i.e. seven) most of which are signalling pathways. In addition, the other pathways from both data overlap as shown in table 5.15. The major pathways in GBM reported by M. Nakada et al 2011 are all present in the pathways discovered in the experimental data column of table 5.15. Hence the result from this analysis is validated.

DATA	EXPERIMENTAL DATA		VALIDATION DATA (INDEPENDENT GBM DATASET)	
S/N	MF	GENES	MF	GENES
1	Binding (GO:0005488)	SP3, OPA1, SMAD4, CKS1B, PREB, SMAD1, FBL, RIT1, XPC, RNF14, RIT1, SWAP70, NACAP1, ZC3H15, ZC3H14	Binding (GO:0005488)	NEXN, TRNT1, PAX8, MEGF11, JAK1, ESRRA, RFC2, PLCD3, STX6
2	Catalytic activity (GO:0003824)	ASS1, DHX9, SPTLC1, OPA1, CKS1B, PMM1, FBL, CHST7, RIT1, UBE2Q1, UGCG, CHST2, ATP2A2	Catalytic activity (GO:0003824)	CASP1, HS6ST2, GAMT, TRNT1, CRYZL1, POLR2J3, JAK1, MAPK1, LACTB, TTLL12, CNOT7, RFC2, PLCD3, ATAD3A, UBA7, PRUNE2
3	Receptor activity (GO:0004872)	GRINA	Receptor activity (GO:0004872)	TLR4, SCARB1
4	Structural molecule activity (GO:0005198)	OPA1, NUP93	Structural molecule activity (GO:0005198)	TTLL12, MFAP3
5	Transporter activity (GO:0005215)	ATP2A2	Transporter activity (GO:0005215)	SLC39A13

Table 5.12: Molecular functions and their genes for experimental and validation data

DATA	EXPERIMENTAL DATA		VALIDATION DATA (INDEPENDENT GBM DATASET)	
S/N	**BP**	**GENES**	**BP**	**GENES**
1	Biological adhesion (GO:0022610)	ITGA7, RIT1, COL4A1, COL4A2, ITGAV	Biological adhesion (GO:0022610)	SCARB1
2	Biological regulation (GO:0065007)	GRINA, RNF14, RAB40B, ATP2A2	Biological regulation (GO:0065007)	NEXN, CASP1, SLC39A13, JAK1
3	Cellular component organization or biogenesis (GO:0071840)	OPA1, FBL, COL4A1, COL4A2, RAB40B, NUP93	Cellular component organization or biogenesis (GO:0071840)	NEXN, STX6
4	Cellular process (GO:0009987)	ASS1, DHX9, SP3, OPA1, SMAD4 CKS1B, RNF11, SMAD1, FBL, CHST7, RNF111, RIT1, COL4A1, XPC, ATP6V0E2, RNF14, COL4A2, UGCG, RIT1, CHST2, RAB40B, ZC3H14, ATP2A2, NUP93	Cellular process (GO:0009987)	NEXN, CASP1, HS6ST2, GAMT, SCL39A13, TLR4, MEGF11, POLR2J3, JAK1, MAPK1, ESRRA, SCARB1, CNOT7, RFC2, PLCD3, MFAP3, UBA7, PRUNE2, STX6
5	Developmental process (GO:0032502)	COL4A1, UBE2Q1, RIT1	Developmental process (GO:0032502)	NEXN, CASP1, CRYZL1, PAXB, JAK1, EPHB3, PRUNE2
6	Immune system process (GO:0002376)	COL4A1, COL4A2, SWAP70	Immune system process (GO:0002376)	JAK1, MAPK1
7	Localization (GO:0051179)	OPA1, PREB, RIT1, RAB40B, NUP93	Localization (GO:0051179)	NEXN, JAK1, KLC3, STX6
8	Metabolic process (GO:0008152)	ASS1, DHX9, SPTLC1, SP3, SMAD4, RNF11, PREB, PMM1, SMAD1, FBL, CHST7, RNF111, DENND4B,	Metabolic process (GO:0008152)	HS6ST2, GAMT, TRNT1, CRYZL1, PAXB, POLR2J3, JAK1, TTLL12, CNOT7, RFC2, PLCD3, UBA7,

		UBE2Q1, UGCG, RIT1, SWAP70, NACAP1, CHST2, ZC3H15 ZC3H14, ATP2A2		PRUNE2
9	Multicellular organismal process (GO:0032501)	RIT1, COL4A1, RAB40B	Multicellular organismal process (GO:0032501)	NEXN, TLR4
10	Response to stimulus (GO:0050896)	RNF111, XCP, RIT1	Response to stimulus (GO:0050896)	CASP1, JAK1, MAPK1, RFC2
			Locomotion (GO:0040011)	NEXN, JAK1

Table 5.13: Biological Processes and their genes for experimental and validation data

DATA	EXPERIMENTAL DATA		VALIDATION DATA (INDEPENDENT GBM DATASET)	
S/N	**CC**	**GENES**	**CC**	**GENES**
1	Cell part (GO:0044464)	ASS1, DHX9, GAP43, SP3, OPA1, FBL, RNF111, XPC, RNF14, RIT1, RAB40B, ZC3H14, ATP2A2, NUP93	Cell part (GO:0044464)	NEXN, CASP1, GAMT, SLC39A13, POLR2J3, JAK1, TTLL12, CNOT7, RFC2, UBA7, PRUNE2, STX6
2	Extracellular matrix (GO:0031012)	COL4A1 COL4A2	Extracellular matrix (GO:0031012)	MEGF11
3	Macromolecular complex (GO:0032991)	DHX9, FBL, XPC, RNF14, NUP93	Macromolecular complex (GO:0032991)	CASP1, POLR2J3, CNOT7, RFC2, STX6
4	Membrane (GO:0016020)	CHST7, FAM134A, CHST2, RAB40B, ATP2A2, NUP93	Membrane (GO:0016020)	SLC39A13, JAK1, STX6
5	Organelle (GO:0043226)	GAP43, SP3, OPA1, FBL, RNF111, XPC, RIT1, RAB40B, ZC3H14, ATP2A2, NUP93	Organelle (GO:0043226)	NEXN, GAMT, SLC39A13, TTLL12, KLC3, RFC2, UBA7
6	Synapse (GO:0045202)	GAP43, RBA40B	Extracellular region (GO:0005576)	TLR4, MEGF11, MFAP3

Table 5.14: Cellular components and their genes for experimental and validation data

DATA	EXPERIMENTAL DATA		VALIDATION DATA (INDEPENDENT GBM DATASET)	
S/N	**PATHWAY**	**GENES**	**PATHWAY**	**GENES**
1	Apoptosis signaling pathway (P00006)	IGF2R	Apoptosis signaling pathway (P00006)	MAPK1
2	CCKR signaling map (P06959)	SP3, ITGAV	CCKR signaling map (P06959)	MAPK1
3	Gonadotropin-releasing hormone receptor pathway (P06664)	SMAD4, SMAD1	Gonadotropin-releasing hormone receptor pathway (P06664)	MAPK1, ESRRA
4	Histamine H1 receptor mediated signaling pathway (P04385)	HRH1	Histamine H1 receptor mediated signaling pathway (P04385)	PLCD3
5	Insulin/IGF pathway-mitogen activated protein kinase kinase/MAP kinase cascade (P00032)	IGF2R	Insulin/IGF pathway-mitogen activated protein kinase kinase/MAP kinase cascade (P00032)	MAPK1
6	Integrin signalling pathway (P00034)	COL4A5, ITGA7, COL4A1, COL4A2, ITGAV	Integrin signalling pathway (P00034)	MAPK1
7	TGF-beta signaling pathway (P00052)	SMAD4, SMAD1	TGF-beta signaling pathway (P00052)	MAPK1

8	ALP23B signaling pathway (P06209)	SMAD4	5HT2 type receptor mediated signaling pathway (P04374)	PLCD3
9	Activin beta Signaling pathway (P06210)	SMAD4	Alpha adrenergic receptor signaling pathway (P00002)	STX6
10	BMP/activin signaling pathway-drosophila (P06211)	SMAD4	Angiogenesis (P00005)	JAK1, MAPK1, EPHB3
11	DPP signaling pathway (P06213)	SMAD4	EGF receptor signaling pathway (P00018)	MAPK1
12	DPP-SCW signaling pathway (P06213)	SMAD4	Endothelin signaling pathway (P00019)	MAPK1
13	Insulin/IGF pathway-protein kinase B signaling cascade (P00033)	IGF2R	Interleukin signaling pathway (P00036)	JAK1, MAPK1
14	Wnt signaling pathway (P00057)	SMAD4, SMAD1	JAK/STAT signaling pathway (P00038)	JAK1
15	Mannose metabolism (P02752)	PMM1	PDGF signaling pathway (P00047)	JAK1, MAPK1

Table 5.15: Pathways and their genes for experimental and validation data

5.5 SURVIVAL ANALYSIS

The data frame used for this analysis is obtained by merge the clinical data, which contain the patient information and data matrix, which contains the expression values. The two data thus has a column (Sample ID) in common, hence the merging was successful. The resulting data frame thus contains as its columns the sample ID, event and time to event and as rows the patient sample ID. From this data frame, a survival object is created with event, time to event and sample ID.

A survival curve is plotted with a Kaplan-Meier estimate of 95% confidence bounds as shown in figure 5.4 below. The curve estimates survival throughout event time and event till the patients drop out of the study. The vertical line of this curve gives the proportion of people surviving while the horizontal axis represents the time (in days) after the beginning of the experiment. It is clear from figure 5.4 below that the survival curves for patients with sample A tumor and patients with sample B tumor are almost smooth, which implies that the curve accurately explains the death time.

Also the curves have median survival times (i.e. the time at which the proportion surviving is 50%) of 372 days (approximately 12 months) for sample A tumor and 298 days (approximately 10 months) for sample B tumor as shown in table 5.16 below. This thus implies that the survival of patients with GBM is below a year and this thus confirms that the tumor is very aggressive. Hence, the chance of surviving GBM is very low.

Sample ID	Events	Events	Median	0.95 LCL	0.95 UCLL
Sample A	245	245	372	327	432
Sample B	279	279	298	268	345

Table 5.16: Median survival time of GBM patients samples

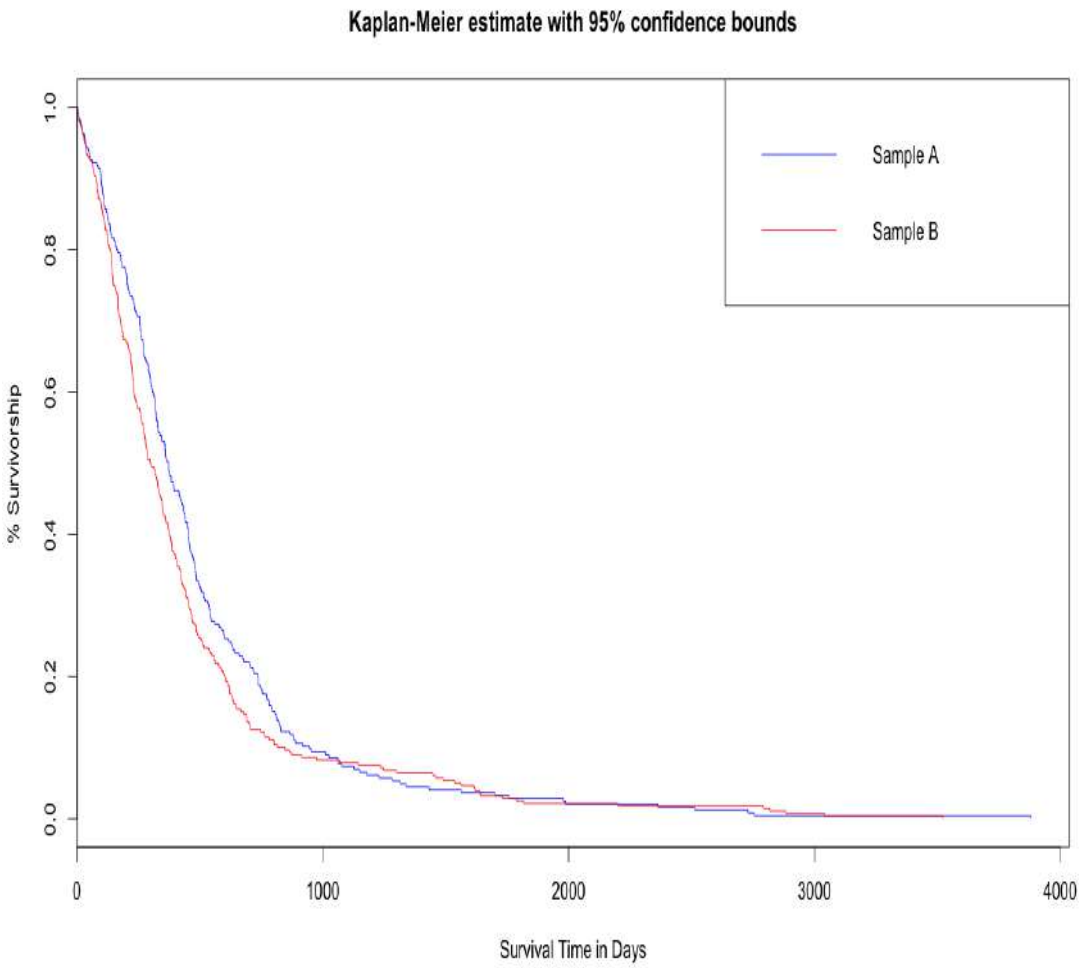

Figure 5.4: Kaplan-Mier Survival Plot

5.6.1 SURVIVAL CURVE COMPARISON

Observing the survival curves physically and computing their median survival is not enough to compare survival curve from two groups. A statistical approach, which involves the use of log-rank test, is employed to compare the curves in question. The result obtained after running the "survdiff" function from the survival package in R is shown in table 5.17 below.

Sample ID	Events	Observed	Expected	(O-E)^2/E	(O-E)^2/V
Sample A	245	245	265	1.48	3.04
Sample B	279	279	259	1.52	3.04

Table 5.17: Log-rank test result

Also the function gives a chi-square value of 3.000 on 1 degrees of freedom and an associated p-value of 0.0814. The chi-square value is less than 3.841, hence the null hypothesis that the survival curves for the two sample groups are identical. This implies that there is no statistical significant evidence at $\alpha = 0.05$ that the survival time between the two sample groups (sample A and sample B) are different.

6. DISCUSSIONS

6.1 EVALUATION OF STUDY METHODS

This study aims to obtain sample and gene subgroup from data matrix, identify genes that are differentially expressed between samples, predict sample membership. In addition, it also aims to transform differentially expressed genes into biological meaning and verify analysis results with an independent GBM dataset.

To achieve this, class discovery analysis (clustering) was performed on a filtered data matrix. This analysis helps in grouping the samples and or genes into subgroup (clusters) and the result from here can be adopted in class comparison and class prediction analysis. Since clustering works for any data even if there are no relationships between the terms in the data, it is important to perform this analysis in the best possible way. In order to get a perfect cluster from a data matrix, the correction between data column (samples) are evaluated and the resulting matrix is a sample-sample correlation matrix. The dendrogram plot of this correlation matrix gives the real relationship that exists between samples and the subgroups are identified easily along side the number of elements in each subgroup. The same approach is used to obtain a gene-gene correlation matrix. Unlike the use of heatmap for hierarchical clustering, the correlation matrix approach is very reliable since it gives a better result than the former.

The result from class discovery analysis was verified by the “clValid” R package with the aim of determining best number of cluster and assessing the performance of the method adopted. The package thus gives an internal validation measure of two-cluster solution and it suggests hierarchical clustering as the best method to be adopted in both sample and gene clustering. The result from the “clValid” R package conforms with that obtained from the class discovery analysis carried out in this study. The class discovery analysis reveals that GBM was grouped into two distinct clusters and each cluster has two sub-clusters. This indicates that GBM are of four subtypes.

Class comparison analysis aims to find genes that are differentially expressed between two conditions (samples) in a data matrix. This analysis somewhat depends on class discovery analysis, which gives insight into the subgroups that might be present in the data matrix. The analysis was carried out with fold change and t-test methods using a filtered data matrix. Filtering is done with gene filter package in R with the aim of removing genes with low expression value. The fold change method evaluates the mean of the expression level between the two-sample subgroups defined in class discovery analysis. The genes with a fold change greater than an absolute fold change (a.f.c of 1.0) threshold are selected as the genes that are differentially expressed between both sample subgroups. The value of absolute fold change mean is chosen arbitrarily and different value of it can be tested. It should be noted that the bigger the absolute fold change value, the lower the number of genes that will be selected as differentially expressed.

The t-test method selects differentially expressed genes between the sample subgroups obtained in class discovery analysis by calculating their p-values. Since a filtered matrix is used for this analysis, the problem of multiple testing has been dealt with. The histogram of the p-value obtained from this analysis reveals that there is evidence against the null hypothesis. The differentially expressed genes are obtained by calling all p-values less than 0.05 significant. However, the truly differentially expressed genes are obtained by a fold-change cutoff greater than 1.0 and a p-value less than 0.05. A total of 1351 genes are obtained as differentially expressed genes between the two sample groups.

Class prediction analysis predicts the membership of the sample group obtained from class discovery analysis by using the data matrix obtained after class comparison analysis. This matrix was transposed so that the samples appear as the rows and the DEGs appear as the column. KNN classifier was used to build the model and different K values are used to search for the best performing model. The performance of the selected model is assessed by computing accuracy, error rate, sensitivity, specificity, precision, recall etc.

Functional and pathway enrichment analysis gives the biological interpretation to the differentially expressed genes obtained from class comparison analysis. These genes are pasted into PANTHER database and the Molecular Function (MF), Biological Process (BP), Cellular Component (CC) and Pathways in which these genes plays significant role are obtained.

Results from the functional and pathway enrichment analysis in this study were validated with an independent GBM dataset from GEO. The same analysis steps were repeated for the independent dataset and the functional and pathway enrichment analysis was compared. The two results shows similar Molecular Function (MF), Biological Process (BP), Cellular Component (CC) and Pathways and this thus validates the results obtained in this study.

The aggressiveness of GBM was predicted by computing the survival of patients using both the clinical data matrix, which contains vital patient information and expression data matrix, which contains patients sample ID. Non-parametric survival method that involves the plot of Kaplan-Mier curve is adopted. First a new data frame, which has as its columns; event, time to event, and patients sample ID is obtained from the clinical data. A survival object is then created from the data frame with event, time to event and patient sample ID with the aim of predicting the aggressiveness of GBM using patient sample ID classification obtained from hierarchical clustering analysis. It was observed that median survival time for patients with sample A is 12 months and that for patients with samples B is 10 months. This implies that median survival of patients with GBM is under 1 year and this thus confirms the tumor aggressiveness. The survival curve for patients with samples A is compared with the curve for patients with samples B using log-rank test. The test indicates that the survival curves are identical.

6.2 ANALYSIS RESULTS VERSUS RESULTS FROM PREVIOUS STUDIES

The results from this study are broadly consistence with those achieved from previous studies. Sample-sample correlation matrix from this study gives two distinct clusters, which are made up of two sub-clusters each and this indicates that GBM are of four subtypes. Pearson correlation analysis was performed on 173 GBM samples from an mRNA expression dataset obtained TCGA website and the result gives two distinct clusters, which are made up of four different sub-clusters and that suggests that GBM is made of four distinct subtype (D. Renu et al 2015). In addition, it was also establish that GMB has four subtypes according to R.G.W. Verhaak et al (2010).

Differentially expressed genes in this study were selected by combining fold change and T-test method with different threshold. Although the issue of plotting p-value histogram to determine whether to perform multiple testing and to select p-value threshold has not been proper addressed. This study however suggests that the shape of the p-value histogram obtained goes a long way in determining the p-value threshold to be employed. Huggins et al. (2008) adopts a fold change of 1.3 and a p-value less than 0.2 to find differentially expressed genes and this is similar to what is done in this study.

The pathways in which the differentially expressed genes from this analysis play significant roles have been reported in previous studies. Apoptosis signaling (cell death), integrin pathway (Angiogenesis), TGF and IGF signalling pathways, MAPK pathway are the pathways found in GBM reported by M. Nakada et al (2011). In addition, Marina M. Marelli et al (2009) reported that Receptors for gonadotropin-releasing hormone (GnRH) is present in glioblastoma tissue. Yeri Lee et al (2016) reported that the deregulation of WNT signalling is related to glioblastoma.

This study also indicates that the median survival of GBM patients irrespective of the patients sample ID is under 1 year and this thus confirms the aggressiveness of GBM. Previous studies such as Michael Henriksen et al (2014) reported that GBM patients has median survival of less than 1 year while Azizul Haque et al (2011) reported that the

median survival of glioblastoma patients is between 10-12 months. These statements thus support the result obtained from this study.

6.3 FURTHER RESEARCH WORK

The analysis of GBM data has mainly addressed class discovery, class comparison, class prediction, and survival analysis while the identification of the molecular subtype in GBM has been neglected. In the future, analysis of GBM data should aim to identify and distinguish molecular subtype in GBM with R and Bioconductor tools.

Also, in the selection of differentially expressed genes with t-test method, the shape of the p-value histogram play a huge role in determining the significant level to adopt and it also give information on the need for multiple testing. Future research work should consider the need for P-value histogram after t-test analysis so as to ascertain the need for multiple testing and to determine the accurate significance level.

7. CONCLUSION

The true relationship that exists between samples and or genes in a gene expression data matrix can be best revealed with a sample-sample (or gene-gene) correlation matrix. The methods used in selecting differentially expressed genes from the experimental and validation datasets perform better because combining two methods gives better results. The use of PANTHER database helps to transform DEGs into biological information. The information from both datasets are compared in term of Molecular Functions, Biological Processes, Cellular Components, and Pathways. It was observed that the results are 95% similar and this validates the result obtained in this study. A KNN classifier achieves sample membership prediction by building a model, which is later assessed by performance parameters such as accuracy, sensitivity, and specificity etc. Also the Kaplan-Mier survival curve produced from GBM clinical data confirms that GBM is aggressive irrespective of sample types because it gives a median survival of less than 1 year. With the descriptions above, it can be concluded that the aims in this study were all met. However there could still be improvements to the methods adopted in this study for a better analysis outcome.

8. LIST OF REFERENCES

Adi L. Tarca, Roberto Romero and Sorin Draghici (2006). Analysis of microarray experiment of gene expression profiling. American Journal of Obstetrics and Gynecology. 195(2): 373 388; 2006.

Ahmed Sadeque, Nicola VL Serão, Bruce R Southey, Kristin R Delfino and Sandra L. Rodriguez-Zas (2012). Identification and characterization of alternative exon usage linked glioblastoma multiforme survival. BMC Medical Genomics 2012, 5:59.

Akdes Serin. Gene Expression Signature Analysis: Connecting small molecules, drugs and genes. International Max Plank Research School for Computational Biology and Scientific Computing.

Alex Sánchez, M and Carme Ruíz de Villa (2008). A Tutorial Review of Microarray Data Analysis. Departament d'Estadística. Universitat de Barcelona. Facultat de Biologia. Avda Diagonal 645. 08028 Barcelona. Spain. 2008

American Brain Tumor Association: Brain Tumor Information (2016). Chicago IL. American Brain Tumor Association, 2016.

Analytics Vidhya (2015). Best way to learn kNN Algorithm using R Programming.

Ashish Ghosh and Bijnan Parai (2008). Protein secondary structure prediction using distance based classifiers. International Journal of Approximate Reasoning 47 (2008), 37–44.

Azizul Haque, Naren L. Banik, and Swapan K. Ray (2011). Molecular Alterations In Glioblastoma: Potential Targets For Immunotherapy. Prog Mol Biol Transl Sci. 2011 ; 98: 187–234.

Ben-Dor A., Shamir R. and Yakhini Z (1999). Clustering gene expression patterns. Journal of Computational Biology, 6(3-4): 281-297; 1999

Brett Lantz (2015). Machine Learning with R, Second Edition. Packt Publishing, 2015.

Bioconductor (2016)

Bioinformatics and Functional Genomics Research Group: Genomic And Transcriptomic Explorer (GATExplorer).

Chris Seidel (2008). Introduction to DNA Microarrays. Chapter 1; pp 1-26, 2008

Churchill G.A (2002). Fundamentals of experimental design for cDNA microarrays. Nat Genet (Supplement), 32:490-495, 2002.

Daniel P. Berrar, Wener Dubitzky and Martin Granzow (2003). A practical approach to microarray data analysis. Kluwer Academic Publishers. Dordrecht 2003.

David M Diez (2013). Survival Analysis in R. OpenIntro 2013.

David Peck, Emily D Crawford, Kenneth N. Ross, Kimberly Stegmaier, Todd R Golub, and Justin Lamb (2006). A method for high-throughput gene expression signature analysis. Genome Biol. 2006; 7(7):R61.

Davis J. McCarthy and Gordon K. Smyth (2009). Gene expression: Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics 25(6), pp. 765–771.

Daxin Jiang, Chun Tang, and Aidong Zhang (2004). Cluster Analysis for Gene Expression Data: a survey," in IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 11, pp. 1370-1386, Nov. 2004.

Dobbin K and Simon R (2002). Comparison of microarray designs for class comparison and class discovery. Bioinformatics 2002; 18:1438-45.

D. Renu et al (2015). Molecular Subtypes in Glioblastoma Multiforme: Integrated Analysis Using Agilent GeneSpring and Mass Profiler Professional Multi-Omics Software. Agilent 2015.

Elisa T. Lee, John Wenyu Wang. Statistical Methods for Survival Data Analysis – 3rd Edition,2003, Wiley Interscience publication.

Erfaneh Naghieh and Yonghong Peng (2006). Microarray Gene Expression Data Mining (Clustering Analysis Review). Department of Computing, University of Bradford. 2006

Eric C. Holland (2000). Glioblastoma multiforme: The terminator. PNAS 97(12), pp. 6242-6244.

Fadhl M. Al-Akwaa (2012). Analysis of Gene Expression Data Using Biclustering Algorithms, Functional Genomics, Dr. Germana Meroni (Ed.), InTech, DOI: 10.5772/48150.

Francisco Pereiraa, Tom Mitchellb, and Matthew Botvinicka Machine learning classifiers and fMRI: a tutorial overview. Neuroimage. 2009 March ; 45(1 Suppl): S199–S209.

G. Jay Kerns (2011). Introduction to Probability and Statistics Using R. First edition, 2011.

Giacomini CP, Leung SY, Chen X, Yuen ST, Kim YH, Bair and Pollack J.R. (2005). A gene expression signature of genetic instability in colon cancer. Cancer Res. 2005 Oct 15; 65(20): 9200-5.

Huaiyu Mi, Betty Lazareva-Ulitsky, Rozina Loo, Anish Kejariwal, Jody Vandergriff, Steven Rabkin, Nan Guo, Anushya Muruganujan, Olivier Doremieux, Michael J. Campbell, Hiroaki Kitano and Paul D. Thomas(2005). The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Research, 2005, Vol. 33, D284–D288.

Jarno Tuimala (2008). DNA microarray data analysis using Biocondutor. The Finnish IT Center for Science (CSC). 2008.

Jarno Tuimala, M. Minna Laine (2003). DNA Microarray Data Analysis. The Finnish IT Center for Science (CSC).

Janez Demsarb (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7 (2006) 1–30.

Jerome H. Friedman, Robert Tibshirani and Trevor Hastie (2001). The Elements of Statistical Learning, Data Mining, Inference, and Prediction. Springer series in statistics, 2001.

John Fox (2014). Introduction to Survival Analysis - McMaster University, Hamilton, Ontario, Canada. 2014.

Juha Kesseli (2015). Institute of Biosciences and Medical Technology (BioMediTech), Department of Bioinformatics, University of Tampere, Tampere Finland.

J. A. Hartigan (1972). Direct clustering of a data matrix. Journal of the American Statistical Association (JASA), 67(337):123–129, 1972.

Kimberly L Cook and Gary S Saylery (2003). Environmental application of array technology. (Promise, Problems and Practicalities). Biotechnology. 14:311–318; 2003.

Kim Y.W., Koul D., Kim S.H., Lucio-Eterovic A.K. , Freire PR, Yao J, Wang J, Almeida J.S., Aldape K. and Yung W.K. (2013). Identification of prognostic gene signatures of glioblastoma: a study based on TCGA data analysis. Neuro Oncol. 2013 Jul; 15(7): 829-39.

Kunkle B.W, Yoo C, Roy D (2013). Reverse Engineering of Modified Genes by Bayesian Network Analysis Defines Molecular Determinants Critical to the Development of Glioblastoma. PLoS ONE 8(5)..

Kirk J Mantione, Richard M. Kream, Hana Kuzelova, Radek Ptacek, Jiri Raboch, Joshua M. Samuel and George B. Stefano (2014). Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq. Med Sci Monit Basic Res, 2014; 20: 138-141.

Leland Wilkinson and Michael Friendly (2009). The History of Cluster Heat Map. The American Statistician, May 2009, Vol. 63, No. 2.

Lisa Sullivan (2016). Survival Analysis. Boston University, School of Public Health 2016.

London school of Hygiene and Tropical medicine (2015): Microarray. Genome Resource Facility (2015).

Mark Stevenson (2009). An Introduction to Survival Analysis. EpiCentre, IVABS, Massey University. June 4, 2009.

Marina M. Marelli, Roberta M. Moretti, Stefania Mai, Oliver MüLler, Johan C. Van Groeninghen and Patrizia Limonta (2009). Novel insights into GnRH receptor activity: Role in the control of human glioblastoma cell proliferation. Oncology reports 21: 1277-1282, 2009.

Michael b. Eisen, Paul t. Spellman, Patrick O. Brown and David Botstein (1998). Cluster analysis and display of genome-wide expression patterns. Genetics. Vol. 95; pages 14863–14868; 1998.

Michael Henriksen, Kasper Bendix Johnsen, Hjalte Holm Andersen, Linda Pilgaard, and Meg Duroux (2014). MicroRNA Expression Signatures Determine Prognosis and Survival in Glioblastoma Multiforme—a Systematic Overview. Mol Neurobiol (2014) 50:896–913.

Mitsutoshi Nakada, Daisuke Kita, Takuya Watanabe, Yutaka Hayashi, Lei Teng, Ilya V. Pyko and Jun-Ichiro Hamada (2011). Aberrant Signaling Pathways in Glioma. Cancers 2011, 3, 3242-3278.

Mosquera Mayo and José Luís (2014). Doctoral Thesis: Methods and Models for the Analysis of Biological Significance Based on High-Throughput Data. Universitat de Barcelona, Dec 2014.

M.Akhil jabbar, B.L Deekshatulua and Priti Chandra (2013). Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm. International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013. Procedia technology 10 (2013) 85 – 94.

M. Kathleen Kerr, and Gary A. Churchil (2001)l. Statistical design and the analysis of gene expression microarray data. Genet. Res. 77:123–128; 2001.

M. Madan Babu (2004). An Introduction to Microarray Data Analysis. MRC Laboratory of Molecular Biology 2004, Chp 11, pp 225–249.

Nicola J. Armstrong and Mark A. Van de Wiel (2004). Microarray data analysis: From hypotheses to conclusions using gene expression data. Cellular Oncology. 26:279–290; 2004.

National Center for Biotechnology Information

Paul D. Allison (2012). Survival Analysis. Chp 31, pp. 413-424.

Paul D. Thomas, Michael J. Campbell, Anish Kejariwal, Huaiyu M.i, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, and Apurva Narechania (2003). PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. Protein Informatics, Genome Research 2003, 13:2129–2141.

Pedro Domingos. A Few Useful Things to Know about Machine Learning. Department of Computer Science and Engineering. University of Washington Seattle, WA 98195-2350, U.S.A.

Pierre Baldi, and Anthony D. Long (2001). A Bayesian framework for the analysis of microarray expression data, regularized t-test and statistical inferences of gene changes. Bioinformatics. 17:509-519; 2001.

P. Khatri and S. Draghici (2005). Ontological analysis of gene expression data: current tools, limitations, and problems. Bioinformatics, 18:3587-3595, 2005.

Richard O. Duda, Peter E. Hart and David (2000). Stork. Pattern Classification, 2nd Edition November 2000.

Richard Simon (2003). Using Microarray for Diagnostic and Prognostic Prediction. Expert Rev. Mol. Diagn. 3(5), 587-595 (2003).

R.G.W. Verhaak et al (2010). An integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR and NF1. Cancer Cell. 2010 January 19; 17(1): 98.

R Project for Statistical Computing (2016).

Ryan D. Egeland and Edwin M. Southern (2005). Electrochemically directed synthesis of oligonucleotides for DNA microarray fabrication. Nucleic Acid Res., Vol. 33(4). 2005.

Sandrine Dudoit, Yee Hwa Yang, Terence P. Speed, and Matthew J. Callow (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica, 12:111–139; 2002.

Scott A. Ness (2006). Basic Microarray Analysis. Methods in Molecular Biology (Bioinformatics and drug discoveries), vol. 316: Chapter 2, pages 13-33. 2006.

Sorin Drăghici (2011). Statistics and Data Analysis for Microarrays Using R and Biocondutor. Second Edition, 2011.

Song JH, Kim HJ, Lee CH, Kim SJ, Hwang SY, Kim TS (2006). Identification of gene expression signatures for molecular classification in human leukemia cells. Int Journals of Oncology 2006 Jul; 29(1):57-64.

S. Draghici (2003). Data Analysis Tools for DNA Microarrays. Chapman & Hall/CRC, 2003.

Stephen P. Jenkins (2005). Survival Analysis. Penn State University, USA. 2005.

Sunitha Kogenaru, Qing Yan, Yinping Guo and Nian Wang (2012). RNA-seq and microarray complement each other in transcriptome profiling. BMC Genomics 2012, 13:629.

The Cancer Genome Atlas - TCGA (2016)

Tsung-Hsien Chiang, Hung-Yi Lo and Shou-De Lin (2012) A Ranking-based KNN Approach for Multi-Label Classification. MLR: Workshop and Conference Proceedings 25:81–96, 2012.

Tuomas Tanner and Hannu Toivonen (2010). Predicting and preventing student failure using the k-nearest neighbour method to predict student performance in an online course environment. Department of Computer Science, University of Helsinki, Finland.

University of California Santa Cruz Genome Browser (2016).

Von Neubeck C, Seidlitz A, Kitzler H.H, Beuthien-Baumann B, Krause M. (2015). Glioblastoma multiforme: emerging treatments and stratification markers beyond new drugs. Br J Radiol 2015; 88: 20150354.

Wim P. Krijnen (2009). Applied Statistics for Bioinformatics using R. 2009.

Wolfgang Huber, Anja von Heydebreck and Matin Vingron (2003). Analysis of microarray gene expression data. 2003.

Xun Xu, Florian Stockhammer and Michael Schmitt (2011). Cellular-Based Immunotherapies for Patients with Glioblastoma Multiforme. Clinical and Developmental Immunology 2012(764213), pp. 1-15.

Yeri Lee, Jin-Ku Lee, Sun Hee Ahn, Jeongwu Lee and Do-Hyun Nam (2016). WNT signaling in glioblastoma and therapeutic opportunities. Laboratory Investigation (2016) 96, 137–150.

Yohan Lee, Adrienne C Scheck, Timothy F Cloughesy, Albert Lai, Jun Dong, Haumith K. Farooqi, Linda M Liau, Steve Horvath, Paul S. Mischel, Stanley F. Nelson (2008). Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age. BMC Medical Genomics, 2008; 1:52.

Zahoor Jan, M. Abrar, Shariq Bashir, and Anwar M. Mirza (2008). Seasonal to Inter-annual Climate Prediction Using Data Mining KNN Technique. IMTIC 2008, CCIS 20, pp. 40–51, 2008.