

# **PREDICTIVE MODELLING USING HADOOP & STATISTICAL ANALYSIS OF GENE EXPRESSION DATA FOR BREAST CANCER**

Project report submitted in partial fulfillment of the requirement for the degree of  
Bachelor of Technology

By

**SOMYA JAISWAL 141829**

Under the supervision of

**Dr. Tiratha Raj Singh**

**&**

**Dr. P.K. Gupta**

to



MAY 2018

**DEPARTMENT OF BIOTECHNOLOGY & BIOINFORMATICS**

**Jaypee University of Information Technology Waknaghat, Solan-173234,  
Himachal Pradesh**

## DECLARATION BY THE SCHOLAR

I hereby declare that the work reported in the B.Tech. project report entitled **“Predictive Modelling using Hadoop & Stastistical Analysis of Gene Expression Data for Breast Cancer”** submitted at **Jaypee University of Information Technology, Wagnaghat, India**, is an authentic record of my work carried out under the supervision of **Dr. Tiratha Raj Singh & Dr. P.K. Gupta**. I have not submitted this work elsewhere for any other degree or diploma.

(Signature of the Scholar)

Somya Jaiswal (141829)

Department of Biotechnology and Bioinformatics

Jaypee University of Information Technology, Wagnaghat, India

Date:

## **SUPERVISOR’S CERTIFICATE**

This is to certify that the work reported in the B.Tech. project report entitled **“Predictive Modelling using Hadoop & Stastistical Analysis of Gene Expression Data for Breast Cancer”**, submitted by **Somya Jaiswal (141829)** at **Jaypee University of Information Technology, Waknaghat, India** is a bonafide record of her original work carried out under my supervision. The work has not been submitted elsewhere for any other degree or diploma.

**(Signature of Supervisors)**

**Dr. Tiratha Raj Singh**

**Dr. P.K. Gupta**

**Associate Professor(BI)**

**Associate Professor(CSE)**

**Date:**

**Date:**

## **ACKNOWLEDGEMENT**

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. A special gratitude I give to my final year project supervisor, Dr. Tiratha Raj Singh & Dr. P.K. Gupta whose contribution in stimulating suggestions and encouragement helped me to coordinate my project especially in writing this report. Furthermore, I would also like to acknowledge with much appreciation the crucial role of the staff of Bioinformatics lab, who gave the permission to use all the required equipments and the necessary material to complete the task. Last but not least, I have to appreciate the guidance given by other faculty member of Bioinformatics and the evaluation panels especially in our project presentation that has improved our communication skills.

## **List of Acronyms and Abbreviations**

BC – Breast Cancer

DNA- di-oxyribo nucleic acid

DCIS- Ductal Carcinoma

AT- Adjuvant Therapy

ER- Estrogen Receptors

PR- Progesterone Receptors

## LIST OF FIGURES

Figure Number	Caption
1.1	Estimated Prostate Cancer Worldwide in 2012. Copyright GLOBOCAN 2012
1.2	The frequency distribution of Breast Cancer data in India.
2.1	miRNA synthesis
3.1	Architecture of Hadoop
3.2	Hadoop Installation in process
3.3	Hadoop Installation in process
3.4	Model Development using Hadoop & R
4.1	NCBI GEO homepage
4.2	Feeding the data into the GEO2R
4.3	Creating groups in GEO2R
4.4	Top results of the query
4.5	Sorted value on the basis of logFC value
4.6	Division of the result on the basis of their expression level
4.7	Division of the result on the basis of their expression level
4.8	Updation of the list
4.9	Second dataset loaded into the GEO2R
4.10	Division of the samples into group
4.11	Sorted table based on the F value
4.12	Updation of the table with information from various sources
4.13	F value greater than 5 of the results were extracted out.
5.1	PTHLH structure
5.2	GEN1 structure
5.3	PTHLH (5744) structure
5.4	CLEC4A structure
5.5	MPZL3 structure

5.6	FAM155A structure
5.7	Expression studies of has-mir-92a-1

## LIST OF FLOWCHARTS

Flowchart Number	Caption	Page Number
3.1	Workflow	16



## CONTENTS

	Page No
<b>Chapter 1</b>	<b>1</b>
Introduction	
<b>Chapter 2</b>	<b>5</b>
Role of miRNA in Breast Cancer	
<b>Chapter 3</b>	<b>9</b>
Method & Results	
• HADOOP	
<b>Chapter 4</b>	<b>19</b>
Methods & Results	
• GEO2R	
<b>Chapter 5</b>	<b>29</b>
Discussion	
<b>Chapter 6</b>	<b>37</b>
Conclusion	
<b>REFERENCES</b>	<b>38</b>

## **INTRODUCTION**

The most common cancer of all found in the women is BC. It is found that one of every eight women suffers from BC in United States. The number of people suffering from BC was found to be 1.7 million in the year 2012. In India, one out of every two women diagnosed with BC dies.

There are different types of BC that have the capability to metastasize to other body organs and tissues other than those of breasts. BC also affects not only to women but also is found in some men. Changes in the DNA of the breast cells lead to the development of cancer cells. The mutation of the proto-oncogenes helps in the progression of the cells to mutate and develop in an uncontrolled manner.

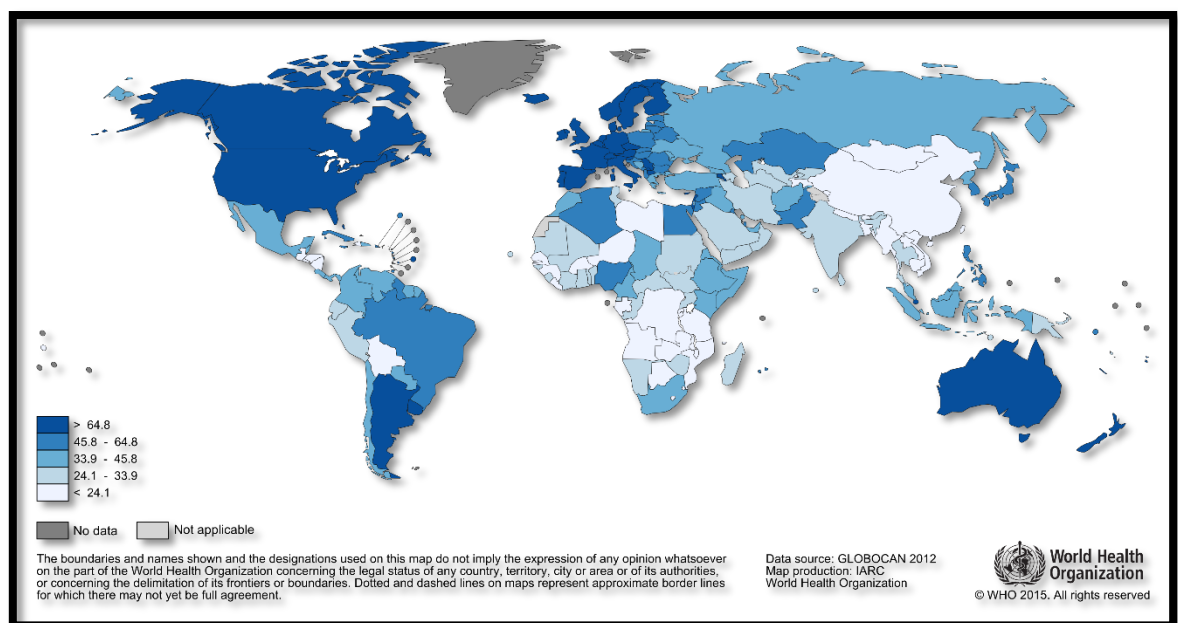


Figure 1.1 Estimated Prostate Cancer Worldwide in 2012. Copyright GLOBOCAN 2012[14]

There are increasing number of BC cases in India in the younger age groups as well of 30s and 40s as well. The reason for the same is the population pyramid, as the age group of the wider part of the pyramid constitutes of the younger age group and the narrower part constitutes of the older age group. BC in India is most common in cities as well as in rural areas also.

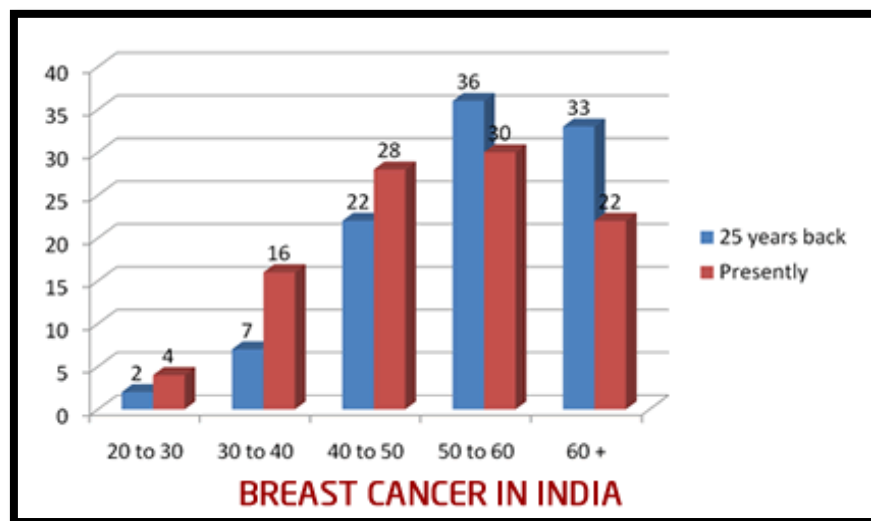


Figure 1.2 The frequency distribution of Breast Cancer data in India. [32]

The BC develops in 5 stages in total which can also be listed from 0-4. These are:

#### Stage 0:

In this stage, the malignant cells are present in the linings of the breast ducts but have not invaded beyond the duct lining of the breasts. The development of abnormal cells starts also called as (DCIS).

Early detection and treatment leads to a good survival rate.

#### Stage 1:

The cells starts to invade the breast cells outside the duct lining in this stage. The tumour size is very small in this case of about 2cm or less.

The chances of local recurrence is 3% and distant metastatic is 8% of total.

#### Stage 2:

The tumour may have grown between 2cm-5cm in size. Or may have been remained of the same size but may have been relocated to one or many lymph nodes.

The treatment for Stage II BC can be done by the method of AT.

#### Stage 3:

This stage occurs when the tumour gets bigger than 5cm in size or the size remains 5cm but the tumour shows significant metastasis.

Regular chemotherapy is done on the patient and the survival rate is 70% with regular chemotherapy sessions.

Stage 4:

The tumour may have also spread to other parts of the body like liver, bones or any other parts of the body.

The survival rate is low. Quality of life can only be achieved during the course of treatment.

The treatment of tumour can also be done in different ways depending on the type of tumour i.e. ER, PR, HER2, the stage of the cancer, genomic markers, and the age of the person, mental health or any of the known mutation that may have occurred.

Small cancers can be removed through surgery whereas for larger tumour, chemotherapy and hormonal therapy can also be done after surgery. Hence, the treatment includes surgery, lymph nodes removal and analysis, radiation therapy, systemic therapy, chemotherapy, hormonal therapy, targeted therapy.

## **ROLE OF miRNA IN BREAST CANCER**

miRNAs are endogenous molecules. They are found in most eukaryotes including humans. The human genome consists of 1-5% of miRNAs. The total number of miRNA that are discovered are 940 in number in the whole human genome. They are highly conserved and are involved in the regulation of gene expression. RNA polymerase II and III transcribe miRNA. The prevention of protein production is done by miRNA binding to specific mRNA.

As miRNA and siRNA cannot be distinguished clearly because of the same function and are also similar biochemically. Hence miRNAs can be distinguished by their origin. miRNA are derived from hairpin precursor while siRNA are derived from double stranded RNA. The precursors of miRNA are found in clusters. These clusters are found in many areas of the genome like the junk DNA. miRNA can usually be located in their respective intronic regions.

The function of gene regulation is done by miRNA which is a small, single stranded non-protein RNA that are highly conserved. miRNA plays a major role in the development of in the growth, development and death of a cell. The expression level of almost 20-30% of the genes is regulated by miRNAs in eukaryotes including humans too. Change in the expression pattern of these miRNA cause progression of many diseases miRNAs play an important role in the progression of tumour suppressor genes which are due to upregulation or downregulation of the miRNAs.

miRNA pair to mRNA of endogenous protein coding genes to redirect their post-transcriptional repression. They modulate more than 30% of the proteins. Hence the dysregulation of these miRNA can impose changes to genes that are responsible in many diseases including variety of cancers. BC is also included in such types of cancer. In the research it is seen that miRNA play an important role in the cancer progression process like proliferation, metastasis, apoptosis and EMT.

Cancer is a disease caused by the changes in the oncogenes and tumour suppressor genes during their expression. The involvement of miRNA is observed in every type of cancer present including the lung, prostate, and colon cancer as well. miRNA are observed to act as tumour suppressor genes when their expression is down-regulated and as oncogenes when they are over-expressed.

In humans, more than 200 miRNAs are found to show progression in cancer development.

The roles of miRNA:

*Synthesis of miRNA:*

The pathway consists of synthesis consists of two events where the cleavage occurs. The first consists of cleavage of nucleus and the other consists of cleavage of cytoplasm. There are many enzymes responsible in the process of synthesis.

The synthesis of miRNA is done through transcription of pri-miRNA by the help of RNA polymerase II. The pri-miRNA is then converted into pre-miRNA. The pre-miRNAs are then cleaved into a doubled stranded mature mi-RNA. The translation of miRNA then follows. It then identifies its complementary mRNA.

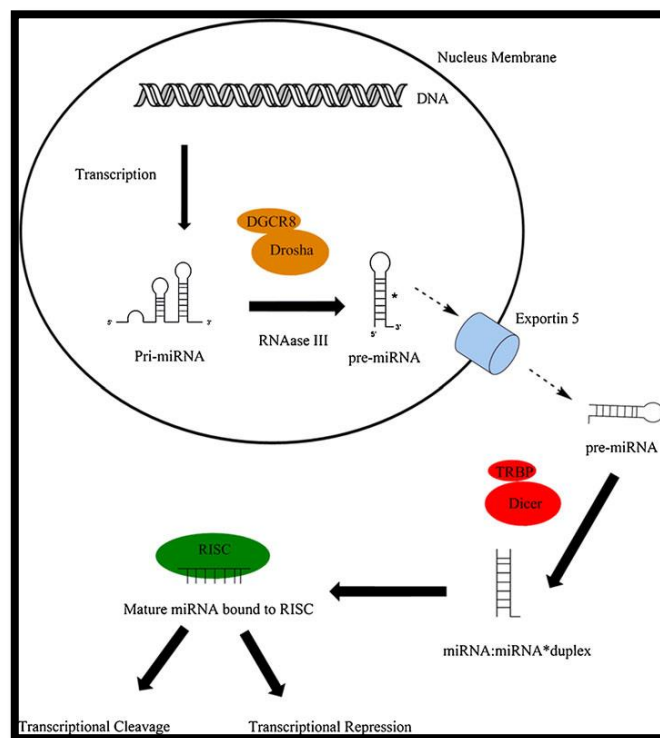


Figure 2.1 : miRNA synthesis [33]



*miRNA in cellular processing:*

miRNA plays an important role in the processing of the various cell processes like development of various tissues, heart and skeletal muscles too. miRNA's role is also observed in insulin secretion and apoptosis.

## METHODS USED

### Model Development using Hadoop:

In this world of technology and information there is lot of data that is generated. Managing this continuous huge datasets is a huge problem from decades. BIG DATA has come to a rescue where analysis and storage of data is of great importance. Data obtained from various experimental sources can be unstructured, semi-structured or structured data. Data from various sources can be in the form of text, images, videos or any other type of source like social media posts etc.

There are various complexities that arise with big data:

#### 1. *Size and Scale of Data:*

Due to the emergence of Cloud Technology, the amount of data that has been generated is becoming more and more complex and enormous in amount.

#### 2. *Speed:*

Another complexity that comes with this amount of data size is the reduction of speed. The larger the amount of data, the slower will be the processing system of the software.

#### 3. *Privacy:*

The privacy of data is very important. The private data of any firm should not be

accessible to every or any person without permission. The misuse or mishandling of data can occur, hence the privacy of the data is very important.

#### *4. Technology:*

Almost every organisation of the top level is practising Big Data now. Organisation like Facebook, Instagram are handling information of more than 50 billion photos of its users.

#### **HADOOP (a big data technology)**

It is a programming framework used for processing of large datasets of any type having any formats. It processes large datasets in distributed computing environment i.e. this software works on multiple computers which has a main server for computing and processing of the data files and running the software. It can be used to develop various applications in various fields.

Hadoop consists of 4 models:

##### **1. Hadoop Distributed File System:**

It allows data to be stored in an easier and accessible manner. The data is stored across various storage devices.

##### **2. Map Reduce Technology:**

This module is carried out in two basic steps, the first one is reading and fetching the data from the database, mapping the data into readable and accessible format for the

software to work. The second one is reducing, which is removing all the duplicates and redundant data from the file and performing the mathematical calculations on them.

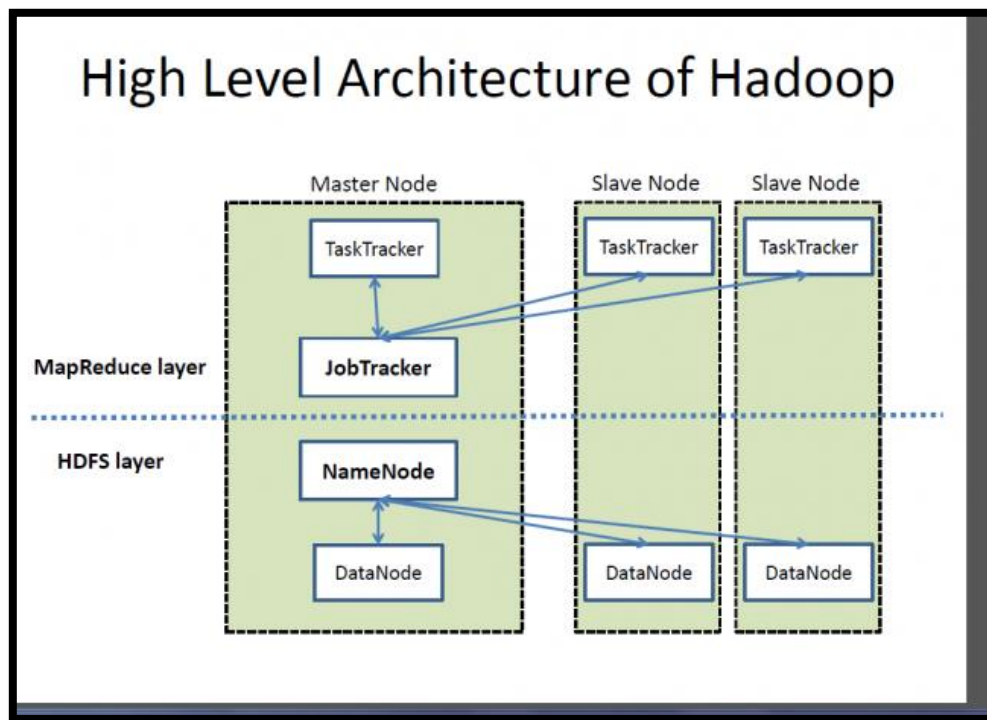


Figure 3.1 Architecture of Hadoop [34]

### 3. Hadoop Common:

This module consists of tools which are made in Java that are needed to be performed on user's computer systems. This module is needed to read the files from the HDFS.

#### 4. YARN:

It has multiple data processing engines like SQL, batch processing etc. It works in respect to scheduling the jobs that are run on Hadoop and cluster resource management also.

### **ADVANTAGES OF HADOOP:**

It has the following benefits:

#### *1. Scalable:*

It can store and distribute huge amount of datasets across various servers which run in parallel.

#### *2. Cost Effective:*

The traditional database management system is very costly.

#### *3. Flexible:*

It can work on any type of data whether structured, unstructured or semi-structured. The data can also be in a format of video, photos, and textual too.

#### *4. Fast:*

As the tools for the data processing are in the same server, the processing of the data is done at a very fast rate. Hadoop efficiently processes terabytes of data.

#### *5. Resilient to failure:*

It has an important feature known as fault tolerance. Hadoop makes many clusters of the data and then distributes into various nodes.

#### APPLICATION IN BIOINFORMATICS:

The traditional tools in Bioinformatics work hard on the large scale data that is generated from high-throughput sequencing. MapReduce framework and HDFS, helps bioinformatics researchers opportunities to obtain a scalable, efficient and much more reliable computing performance on Linux clusters and Cloud Computing various services.

In near future, the applications can be made and developed using the Hadoop technology and its various modules with more efficient algorithm in analysing the genomic datasets of Bioinformatics and Biotechnological experiments.

## MODEL DEVELOPMENT:

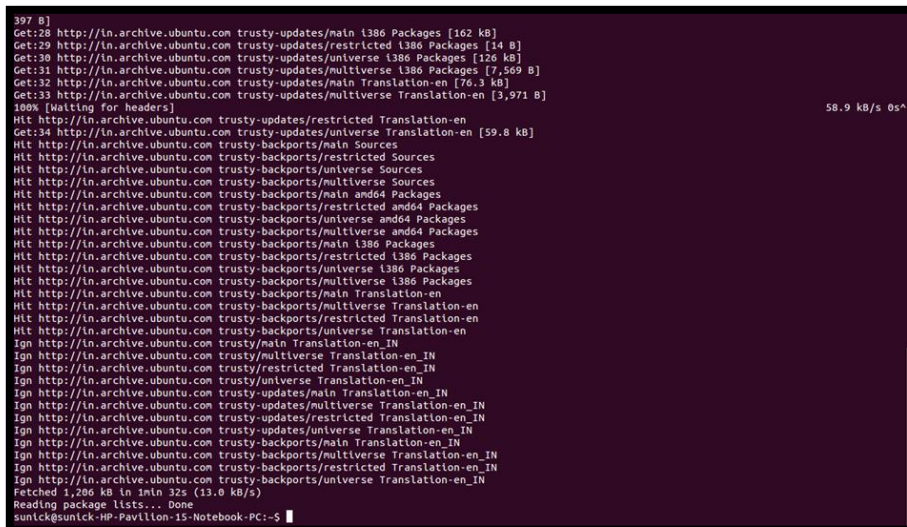
The accuracy of the model was checked based on the given data.

Processes involved:

### 1. Hadoop Installation:

```
user@ubuntu:~$ sudo apt-get update
```

```
user@ubuntu:~$ sudo apt-get install sun-java6-jdk
```

A screenshot of a terminal window with a dark background and light-colored text. The terminal shows the output of the command 'sudo apt-get update'. It lists various package repositories being updated, including 'main', 'restricted', 'universe', and 'multiverse' for both 'main' and 'restricted' architectures. It also shows the download of package lists and the calculation of package sizes. The output ends with 'Reading package lists... Done' and a prompt for the next command.

```
397 B]
Get:28 http://in.archive.ubuntu.com trusty-updates/main i386 Packages [162 kB]
Get:29 http://in.archive.ubuntu.com trusty-updates/restricted i386 Packages [14 B]
Get:30 http://in.archive.ubuntu.com trusty-updates/universe i386 Packages [126 kB]
Get:31 http://in.archive.ubuntu.com trusty-updates/multiverse i386 Packages [7,569 B]
Get:32 http://in.archive.ubuntu.com trusty-updates/main Translation-en [76.3 kB]
Get:33 http://in.archive.ubuntu.com trusty-updates/multiverse Translation-en [3,971 B]
100% [Waiting for headers]
Hit http://in.archive.ubuntu.com trusty-updates/restricted Translation-en
Get:34 http://in.archive.ubuntu.com trusty-updates/universe Translation-en [59.8 kB]
Hit http://in.archive.ubuntu.com trusty-backports/main Sources
Hit http://in.archive.ubuntu.com trusty-backports/restricted Sources
Hit http://in.archive.ubuntu.com trusty-backports/universe Sources
Hit http://in.archive.ubuntu.com trusty-backports/multiverse Sources
Hit http://in.archive.ubuntu.com trusty-backports/main amd64 Packages
Hit http://in.archive.ubuntu.com trusty-backports/restricted amd64 Packages
Hit http://in.archive.ubuntu.com trusty-backports/universe amd64 Packages
Hit http://in.archive.ubuntu.com trusty-backports/multiverse amd64 Packages
Hit http://in.archive.ubuntu.com trusty-backports/main i386 Packages
Hit http://in.archive.ubuntu.com trusty-backports/restricted i386 Packages
Hit http://in.archive.ubuntu.com trusty-backports/universe i386 Packages
Hit http://in.archive.ubuntu.com trusty-backports/multiverse i386 Packages
Hit http://in.archive.ubuntu.com trusty-backports/main Translation-en
Hit http://in.archive.ubuntu.com trusty-backports/restricted Translation-en
Hit http://in.archive.ubuntu.com trusty-backports/universe Translation-en
Ign http://in.archive.ubuntu.com trusty/main Translation-en_IN
Ign http://in.archive.ubuntu.com trusty/multiverse Translation-en_IN
Ign http://in.archive.ubuntu.com trusty/restricted Translation-en_IN
Ign http://in.archive.ubuntu.com trusty/universe Translation-en_IN
Ign http://in.archive.ubuntu.com trusty-updates/main Translation-en_IN
Ign http://in.archive.ubuntu.com trusty-updates/restricted Translation-en_IN
Ign http://in.archive.ubuntu.com trusty-updates/universe Translation-en_IN
Ign http://in.archive.ubuntu.com trusty-backports/main Translation-en_IN
Ign http://in.archive.ubuntu.com trusty-backports/restricted Translation-en_IN
Ign http://in.archive.ubuntu.com trusty-backports/universe Translation-en_IN
Fetched 1,206 kB in 1m1s 32s (13.0 kB/s)
Reading package lists... Done
sunick@sunick-HP-Pavilion-15-Notebook-PC:~$
```

Figure 3.2 : Hadoop Installation in process

Adding a dedicated Hadoop system user

```
user@ubuntu:~$ sudo addgroup hadoop_group
```

```
user@ubuntu:~$ sudo adduser --ingroup hadoop_group hduser1
```

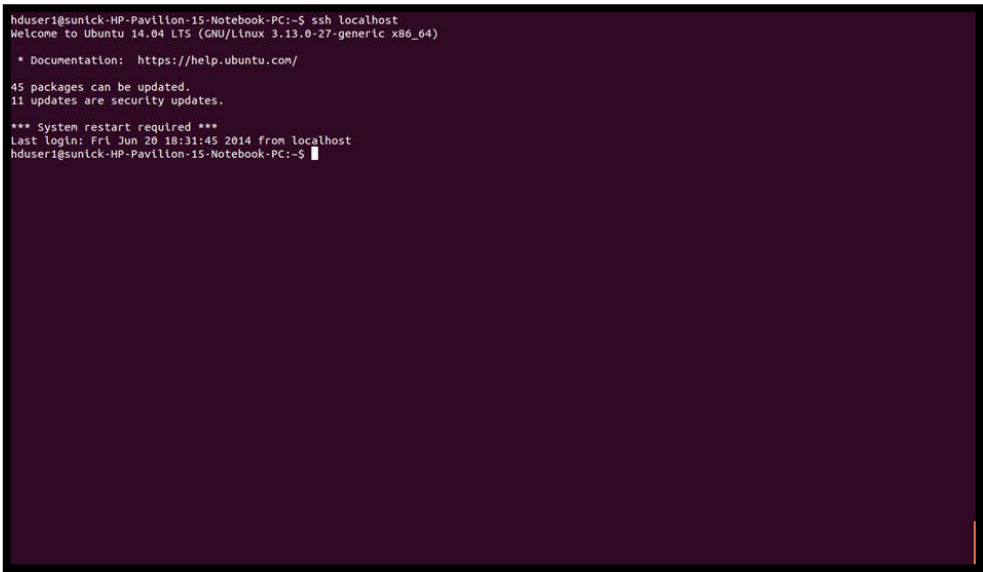
```
user@ubuntu:~$ sudo adduser hduser1 sudo
```

We have to generate an SSH key for the hduser user.

```
user@ubuntu:~$ su - hduser1
```

```
hduser1@ubuntu:~$ ssh-keygen -t rsa -P
```

```
hduser@ubuntu:~$ ssh localhost
```

A terminal window with a dark purple background. The text shows an SSH session for user 'hduser1' on a machine named 'sunick-HP-Pavilion-15-Notebook-PC'. It displays the Ubuntu 14.04 LTS welcome message, system updates (45 packages can be updated, 11 security updates), and a system restart requirement. The last login was on Fri Jun 20 18:31:45 2014 from localhost. The prompt is 'hduser1@sunick-HP-Pavilion-15-Notebook-PC:~\$' with a cursor.

```
hduser1@sunick-HP-Pavilion-15-Notebook-PC:~$ ssh localhost
Welcome to Ubuntu 14.04 LTS (GNU/Linux 3.13.0-27-generic x86_64)

 * Documentation:  https://help.ubuntu.com/

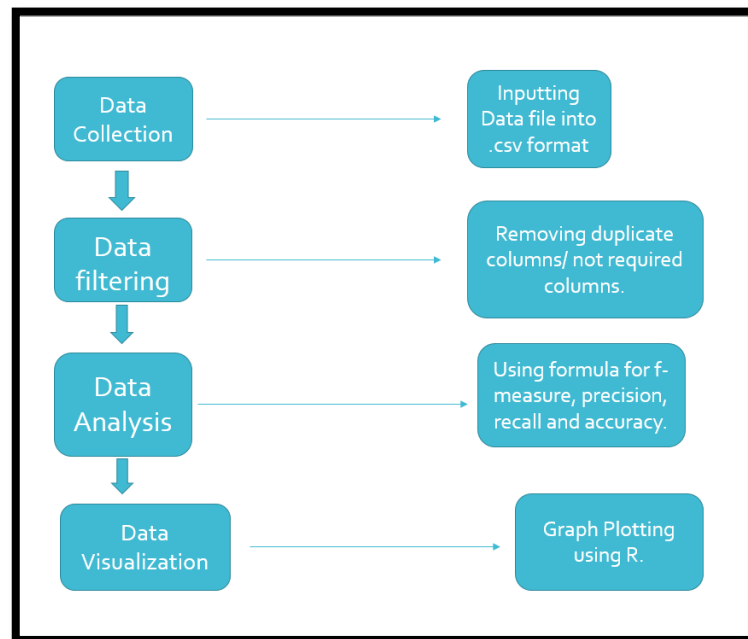
45 packages can be updated.
11 updates are security updates.

*** System restart required ***
Last login: Fri Jun 20 18:31:45 2014 from localhost
hduser1@sunick-HP-Pavilion-15-Notebook-PC:~$
```

Figure 3.3 : Installation of Hadoop



## 2. WORKFLOW OF THE MODEL DEVELOPMENT



Flowchart: Workflow of the task

### 1. Data Collection:

The data was collected.

It was converted into csv file format -> breastcancer.csv

Then the csv file was input into HDFS file system.

### 2. Data Filtering:

The data was fetched from the HDFS system.

The duplicate columns were removed and the redundant data was deleted using Map Reduce technology.

### 3. Data Analysis:

The datasets were compared.

The model accuracy was checked on the basis of :

F-measure: it is the measure of the test's accuracy. It is the weighted harmonic mean of precision and recall.

Recall: it is defined as the sensitivity.

Precision: It is the measure of how close your data points are to the true mean. First we need to find the average, then the variance, then standard deviation, then standard error of the mean is calculated.

Accuracy: It is the difference between the accepted value and experimental value. And then it is divided by the accepted value.

### 4. Data Visualization:

The graph was plotted in R using the function of boxplot in R.

RESULT:

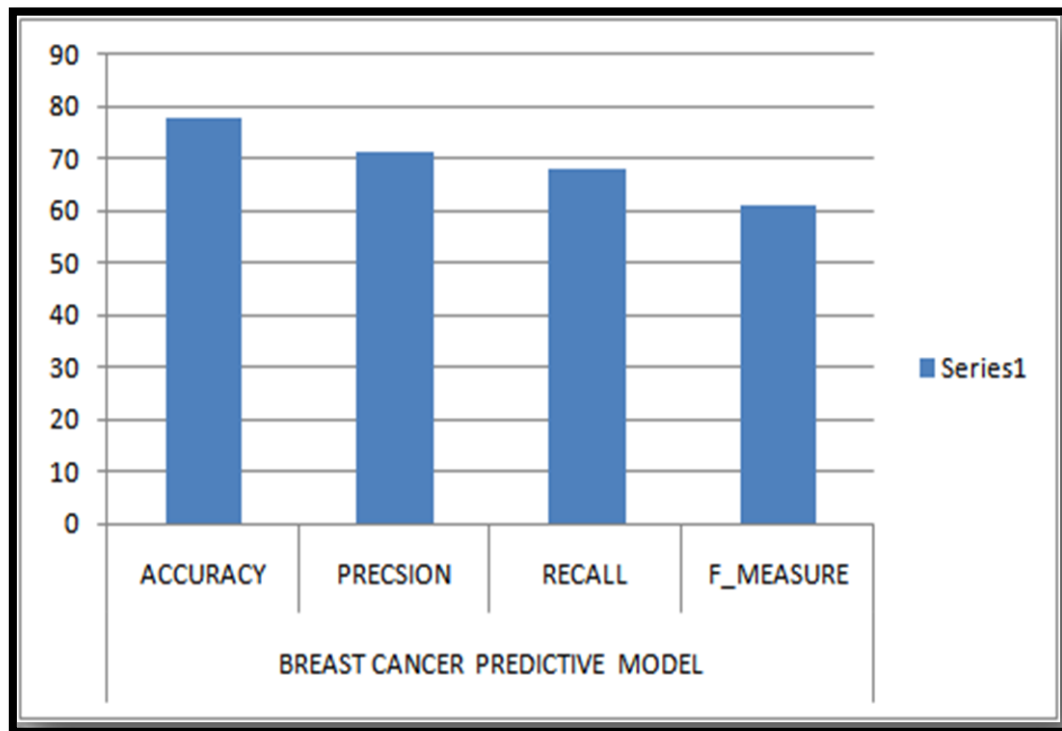


Figure: 3.4 Model Development using Hadoop & R.

- The accuracy came out to be: 78.092
- The precision of the model was: 71.117
- The recall was calculated as: 68.091
- The f-measure was: 61.096

## GEO2R:

GEO2R is an interactive web tool that allows users to compare two or more groups of Samples in a GEO Series in order to identify genes that are differentially expressed across experimental conditions. Results are presented as a table of genes ordered by significance.

GEO2R performs comparisons on original submitter-supplied processed data tables using the GEOquery and limma R packages from the Bioconductor project. The GEOquery R package parses GEO data into R data structures that can be used by other R packages. Thus, GEO2R provides a simple interface that allows users to perform R statistical analysis without command line expertise.

GEO2R does not rely on curated DataSets and interrogates the original Series Matrix data file directly. This allows a greater proportion of GEO data to be analyzed in a timely manner.

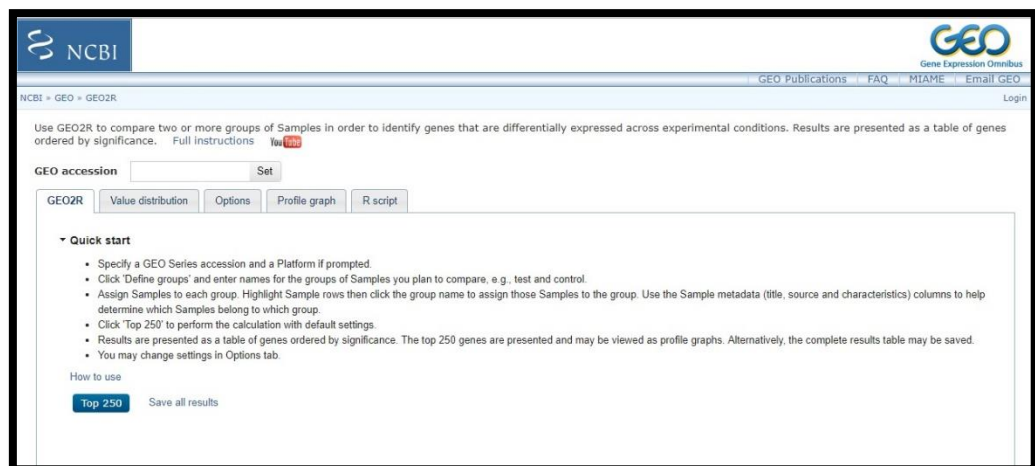


Figure 4.1 NCBI GEO homepage [1]

## WORKFLOW:

### 1. Data collection:

The data was collected from GEO2R database. There were 2 sets of data collected:

#### SET 1:

1. The overexpression of miR-203 on mesenchymal cells of BC cell line. The study of the cell line was done to see the overexpression of miR-203 in metastasis of BC.

The reference series was GSE50679. The series consisted of 6 samples from which 3 were controlled and the rest 3 were miR-203 affected.

- Then the samples were fed into the GEO2R tool.

The screenshot displays the GEO2R web interface. At the top, the 'GEO accession' field contains 'GSE50697' and the 'Set' button is visible. The title of the page is 'Gene expression of SUM159 breast cancer cell line expressing microRNA--203'. Below this, the 'Samples' section is active, showing a table with 6 samples. A 'Define groups' dropdown menu is open, showing 'Enter a group name:' and a 'List' button. The table has columns for Group, Accession, Source name, Cell line, Treatment, and Tissue. The first three samples are control, and the last three are treated with pBabe puro miR-203. Below the table, there are buttons for 'GEO2R', 'Value distribution', 'Options', 'Profile graph', and 'R script'. At the bottom, there is a 'View' button and an 'Export' link.

Group	Accession	Source name	Cell line	Treatment	Tissue
-	GSM1226581	SUM159	SUM159	control	Claudin-low breast cancer
-	GSM1226582	SUM159 Control rep2	SUM159	control	Claudin-low breast cancer
-	GSM1226583	SUM159 Control rep3	SUM159	control	Claudin-low breast cancer
-	GSM1226584	SUM159 miR-203 rep1	SUM159	pBabe puro miR-203	Claudin-low breast cancer
-	GSM1226585	SUM159 miR-203 rep2	SUM159	pBabe puro miR-203	Claudin-low breast cancer
-	GSM1226586	SUM159 miR-203 rep3	SUM159	pBabe puro miR-203	Claudin-low breast cancer

Figure 4.2 Feeding the data into the GEO2R

- Then the division was done into 2 categories of samples: the control and the miR-203 affected.

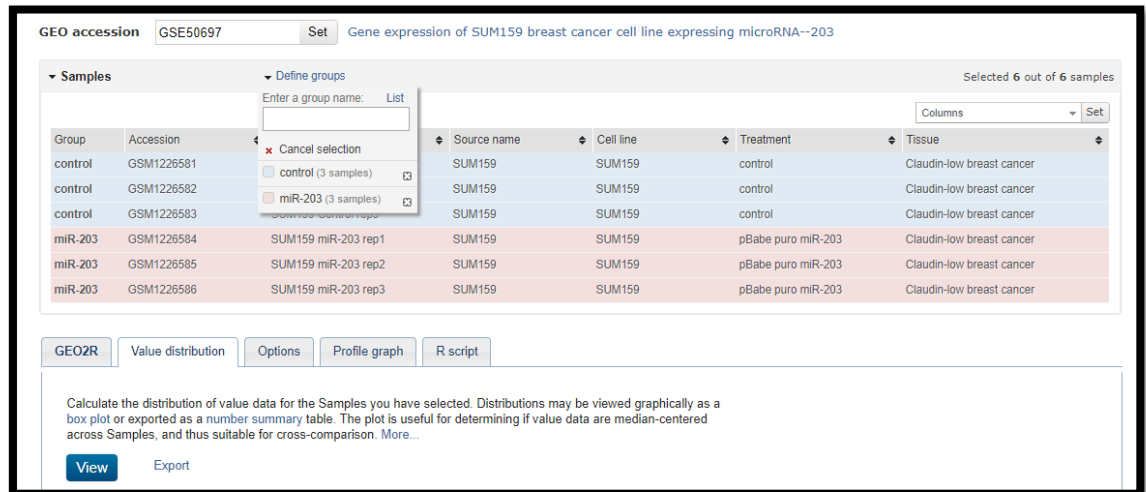


Figure 4.3 Creating groups in GEO2R

- Then the run option was done and the top 250 genes are displayed after the process is finished.
- The gene IDs are shown along with their logFC value from which they are further differentiated.

Log-transformation has been applied to the data. You can change this in the Options tab.

Recalculate if you changed any options. Save all results Select columns

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
▶ 231905_at	0.111	0.0000202	22.39	3.751	2.946	C20orf96	chromosome 20 open...
▶ 228946_at	0.119	0.0000566	-18.37	3.392	-2.79	INTU	inturned planar cell p...
▶ 209719_x_at	0.119	0.00000836	17.04	3.229	2.197	SERPINB3	serpin family B memb...
▶ 218499_at	0.119	0.0000119	-15.92	3.067	-2.597	STK26	serine/threonine prote...
▶ 222920_s_at	0.119	0.0000169	-14.87	2.894	-4.08	TESPA1	thymocyte expressed....
▶ 211906_s_at	0.119	0.00001878	14.57	2.839	2.131	SERPINB4	serpin family B memb...
▶ 1569981_at	0.119	0.00001997	14.4	2.807	3.875		
▶ 208891_at	0.119	0.00003037	13.27	2.574	1.691	DUSP6	dual specificity phosph...
▶ 1556773_at	0.119	0.00003424	12.96	2.503	4.039		
▶ 209946_at	0.119	0.0000348	12.92	2.494	1.677	VEGFC	vascular endothelial g...
▶ 210413_x_at	0.119	0.00003581	12.85	2.477	2.262	SERPINB4//SERPINB3	serpin family B memb...
▶ 202949_s_at	0.119	0.00003606	12.83	2.473	1.604	FHL2	four and a half LIM do...
▶ 1553333_at	0.119	0.00003629	12.82	2.469	3.022	MAB21L3	mat-21 like 3
▶ 1557883_s_at	0.119	0.0000375	-12.74	2.449	-2.019		
▶ 214183_s_at	0.119	0.00003813	12.69	2.439	2.879	TKTL1	transketolase like 1
▶ 1565917_at	0.119	0.00003827	-12.69	2.437	-4.039		
▶ 206172_at	0.119	0.00003866	12.66	2.431	1.954	IL13RA2	interleukin 13 recepto...
▶ 215695_s_at	0.119	0.00003907	-12.63	2.424	-1.848	GYG2	glycogenin 2
▶ 201721_s_at	0.124	0.00004453	-12.31	2.344	-2.253	LAPTM5	lysosomal protein tra...
▶ 213831_at	0.124	0.00004686	12.19	2.312	2.508	LOC100509457//HL...	HLA class II histocom...
▶ 226189_at	0.124	0.00005081	12	2.261	1.795	ITGB8	integrin subunit beta 8

Figure 4.4 Top results of the query

- Then the gene list was sorted in excel sheet according to the logFC value. The logFC value determines the expression levels of the genes in the sample. The comparison is done between the control and miR-203 samples.

	A	B	C	D	E	F	G	H	I
1	ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title	
2	232618_at	0.137	0.00026065	-8.67	1.079	-4.428	TXLNGY	taxilin gamma pseudogene, Y-linked	
3	1553602_at	0.142	0.00031614	-8.34	0.922	-4.283	MUC1L	mucin like 1	
4	219947_at	0.124	0.00008382	-10.87	1.929	-4.081	CLEC4A	C-type lectin domain family 4 member A	
5	222920_s_at	0.119	0.0000169	-14.87	2.894	-4.08	TESPA1	thymocyte expressed, positive selection associated 1	
6	215468_at	0.163	0.00071169	-7.07	0.227	-4.062	LOC647070	nascent polypeptide-associated complex subunit alpha, muscle-s	
7	1565917_at	0.119	0.00003827	-12.69	2.437	-4.039			
8	238169_at	0.124	0.00013849	-9.84	1.569	-3.736			
9	1563513_at	0.15	0.00044416	-7.78	0.638	-3.523	SYTL4	synaptotagmin like 4	
10	1570584_at	0.15	0.00042064	-7.87	0.684	-3.479	MPZL3	myelin protein zero like 3	
11	1561644_x_at	0.124	0.00011658	-10.19	1.695	-3.442			
12	231424_at	0.138	0.00029155	-8.48	0.989	-3.39	SLC5A12	solute carrier family 5 member 12	
13	230869_at	0.163	0.00067716	-7.14	0.271	-3.376	FAM155A	family with sequence similarity 155 member A	
14	1564039_at	0.124	0.00009164	-10.68	1.867	-3.365	ZSCAN23	zinc finger and SCAN domain containing 23	
15	1566762_at	0.163	0.00072435	-7.04	0.211	-3.335			
16	205888_s_at	0.163	0.00072372	-7.04	0.212	-3.253	JAKMIP2	janus kinase and microtubule interacting protein 2	
17	1570207_at	0.16	0.00061245	-7.29	0.36	-3.251	FRRS1	ferric chelate reductase 1	
18	1567273_at	0.15	0.0004607	-7.73	0.606	-3.242	OR2K2	olfactory receptor family 2 subfamily K member 2	
19	236430_at	0.126	0.00019093	-9.23	1.325	-3.236	TMED6	transmembrane p24 trafficking protein 6	
20	237750_at	0.124	0.00012823	-10	1.626	-3.157	XPNPPEP3	X-prolyl aminopeptidase 3	
21	205826_at	0.15	0.00037232	-8.07	0.787	-3.142	MYOM2	myomesin 2	
22	215294_s_at	0.124	0.00007772	-11.04	1.981	-3.115	SMARCA1	SWI/SNF related, matrix associated, actin dependent regulator of	
23	213219_at	0.155	0.000521	-7.53	0.501	-3.103	ADCY2	adenylate cyclase 2	

Figure 4.5 Sorted value on the basis of logFC value.

- Then the list was divided into :  
Top 10 over-expressed genes.

1	ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
2	244829_at	0.157	0.00057547	7.38	0.414	3.181	LINC00518	
3	211756_at	0.124	0.00013229	9.93	1.603	3.214	PTHLH	long intergenic non-protein coding RNA 518
4	1557267_s_at	0.15	0.00041786	7.88	0.689	3.481	GEN1	parathyroid hormone like hormone
5	1553183_at	0.126	0.00017467	9.4	1.394	3.7	UMODL1	GEN1, Holliday junction 5' flap endonuclease
6	1562579_at	0.163	0.00070687	7.08	0.233	3.724		uromodulin like 1
7	210355_at	0.143	0.00032064	8.32	0.911	3.751	PTHLH	
8	1569981_at	0.119	0.00001997	14.4	2.807	3.875		parathyroid hormone like hormone
9	1556773_at	0.119	0.00003424	12.96	2.503	4.039		
10	220030_at	0.163	0.00073575	7.02	0.197	5.27	STYK1	serine/threonine/tyrosine kinase 1
11								

Figure 4.6 Division of the result on the basis of their expression level

Top 10 under-expressed genes.

1	ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
2	232618_at	0.137	0.00026065	-8.67	1.079	-4.428	TXLNGY	taxillin gamma pseudogene, Y-linked
3	1553602_at	0.142	0.00031614	-8.34	0.922	-4.283	MUCL1	mucin like 1
4	219947_at	0.124	0.00008382	-10.87	1.929	-4.081	CLEC4A	C-type lectin domain family 4 member A
5	222920_s_at	0.119	0.0000169	-14.87	2.894	-4.08	TESPA1	thymocyte expressed, positive selection associated 1
6	215468_at	0.163	0.00071169	-7.07	0.227	-4.062	LOC647070	nascent polypeptide-associated complex subunit alpha, muscle-specific form-like
7	1565917_at	0.119	0.00003827	-12.69	2.437	-4.039		
8	238169_at	0.124	0.00013849	-9.84	1.569	-3.736		
9	1563513_at	0.15	0.00044416	-7.78	0.638	-3.523	SYTL4	synaptotagmin like 4
10	1570584_at	0.15	0.00042064	-7.87	0.684	-3.479	MPZL3	myelin protein zero like 3
11								
12								
13								

Figure 4.7 Division of the result on the basis of their expression level

Then with the help of NCBI, GeneCards, their symbol ID and Gene ID and chromosomal location were found:



ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	GENE ID	Gene.title	Location
244829_at	0.157	0.00057547	7.38	0.414	3.181	LINC00518	221718	long intergenic non-protein coding RNA 518	6p24.3
211756_at	0.124	0.00013229	9.93	1.603	3.214	PTHLH	5744	long intergenic non-protein coding RNA 518	12p11.22
1557267_s_at	0.15	0.00041786	7.88	0.689	3.481	GEN1	348654	parathyroid hormone like hormone	2p24.2
1553183_at	0.126	0.00017467	9.4	1.394	3.7	UMODL1	89766	GEN1, Holliday junction 5' flap endonuclease	21q22.3
1562579_at	0.163	0.00070687	7.08	0.233	3.724			Transcribed locus, weakly similar to NP_078841.2 hypothetical protein FLJ14166	
210355_at	0.143	0.00032064	8.32	0.911	3.751	PTHLH	5744	long intergenic non-protein coding RNA 518	12p11.22
1556773_at	0.119	0.00003424	12.96	2.503	4.039	PTHLH			
220030_at	0.163	0.00073575	7.02	0.197	5.27	STYK1	55359	serine/threonine/tyrosine kinase 1	12p13.2

Figure 4.7 Updation of the list

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	GENE ID	Gene.title	chromosomal location
232618_at	0.137	0.00026065	-8.67	1.079	-4.428	TXLNGY	246126	taxilin gamma pseudogene, Y-linked	Yq11.222-q11.223
1553602_at	0.142	0.00031614	-8.34	0.922	-4.283	MUC1L	118430	mucin like 1	12q13.2
219947_at	0.124	0.00008382	-10.87	1.929	-4.081	CLEC4A	50856	C-type lectin domain family 4 member A	12p13.31
222920_s_at	0.119	0.0000169	-14.87	2.894	-4.08	TESPA1	9840	thymocyte expressed, positive selection associated 1	12q13.2
215468_at	0.163	0.00071169	-7.07	0.227	-4.062	LOC647070	647070	nascent polypeptide-associated complex subunit alpha, muscle-specific form-like	1q25.3
1565917_at	0.119	0.00003827	-12.69	2.437	-4.039				
1563513_at	0.15	0.00044416	-7.78	0.638	-3.523	SYTL4	94121	synaptotagmin like 4	Xq22.1
1570584_at	0.15	0.00042064	-7.87	0.684	-3.479	MPZL3	196264	myelin protein zero like 3	11q23.3
230869_at	0.163	0.00067716	-7.14	0.271	-3.376	FAM155A	728215	family with sequence similarity 155 member A	13q33.3

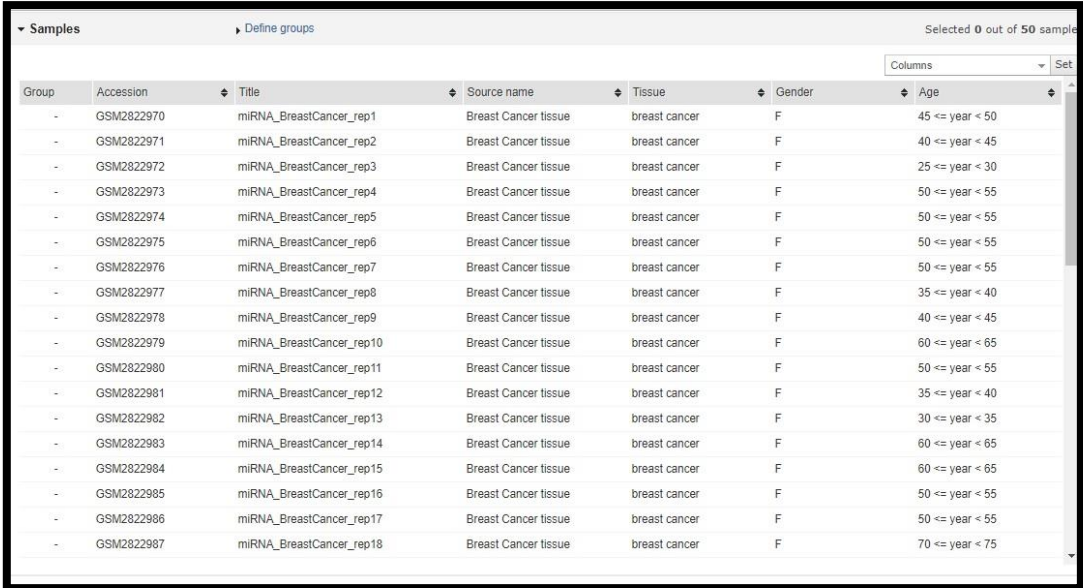
Figure 4.7 Updation of the list

## DATA SET 2:

The second set of data was collected from NCBI GEO database. The dataset was derived from Cancer Institute Hospital of Japanese Foundation for Cancer Research. The dataset has 50 samples of affected BC patients. The samples were tissues affected from BC in females.

The dataset was differentiated on the basis of age groups. There were 50 samples out of which 3 were defined as 'unknown'. Different age groups were ranging from 25 years old to 110 years old. The majority of the samples were under the age group of 45-55 years old.

The data was loaded into the GEO2R tool.



Group	Accession	Title	Source name	Tissue	Gender	Age
-	GSM2822970	miRNA_BreastCancer_rep1	Breast Cancer tissue	breast cancer	F	45 <= year < 50
-	GSM2822971	miRNA_BreastCancer_rep2	Breast Cancer tissue	breast cancer	F	40 <= year < 45
-	GSM2822972	miRNA_BreastCancer_rep3	Breast Cancer tissue	breast cancer	F	25 <= year < 30
-	GSM2822973	miRNA_BreastCancer_rep4	Breast Cancer tissue	breast cancer	F	50 <= year < 55
-	GSM2822974	miRNA_BreastCancer_rep5	Breast Cancer tissue	breast cancer	F	50 <= year < 55
-	GSM2822975	miRNA_BreastCancer_rep6	Breast Cancer tissue	breast cancer	F	50 <= year < 55
-	GSM2822976	miRNA_BreastCancer_rep7	Breast Cancer tissue	breast cancer	F	50 <= year < 55
-	GSM2822977	miRNA_BreastCancer_rep8	Breast Cancer tissue	breast cancer	F	35 <= year < 40
-	GSM2822978	miRNA_BreastCancer_rep9	Breast Cancer tissue	breast cancer	F	40 <= year < 45
-	GSM2822979	miRNA_BreastCancer_rep10	Breast Cancer tissue	breast cancer	F	60 <= year < 65
-	GSM2822980	miRNA_BreastCancer_rep11	Breast Cancer tissue	breast cancer	F	50 <= year < 55
-	GSM2822981	miRNA_BreastCancer_rep12	Breast Cancer tissue	breast cancer	F	35 <= year < 40
-	GSM2822982	miRNA_BreastCancer_rep13	Breast Cancer tissue	breast cancer	F	30 <= year < 35
-	GSM2822983	miRNA_BreastCancer_rep14	Breast Cancer tissue	breast cancer	F	60 <= year < 65
-	GSM2822984	miRNA_BreastCancer_rep15	Breast Cancer tissue	breast cancer	F	60 <= year < 65
-	GSM2822985	miRNA_BreastCancer_rep16	Breast Cancer tissue	breast cancer	F	50 <= year < 55
-	GSM2822986	miRNA_BreastCancer_rep17	Breast Cancer tissue	breast cancer	F	50 <= year < 55
-	GSM2822987	miRNA_BreastCancer_rep18	Breast Cancer tissue	breast cancer	F	70 <= year < 75

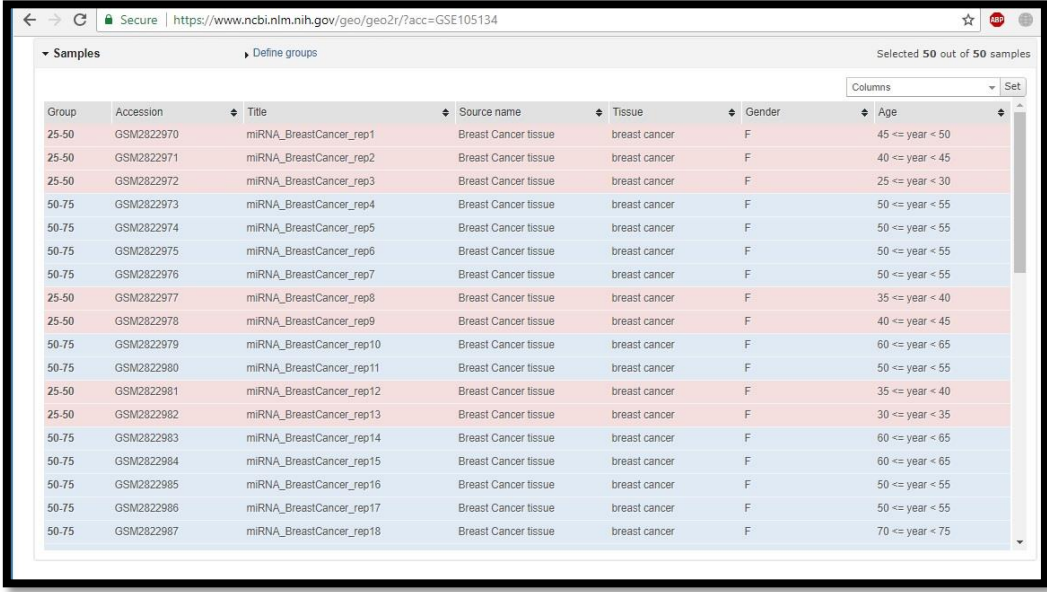
Figure 4.8 Second dataset loaded into the GEO2R

The groups were defined on the basis of 3 age groups:

- (25-50)
- (50-75)
- (75-110)

On the basis of the samples, the number of samples which were under these age groups were:

- (25-50) – 16 samples + 1 unknown age group sample
- (50-75) – 29 samples + 1 unknown age group sample
- (75-110) – 2 samples + 1 unknown age group sample



The screenshot shows the NCBI GEO dataset GSE105134. The table displays 18 samples, each with a Group, Accession, Title, Source name, Tissue, Gender, and Age. The samples are categorized into three age groups: 25-50, 50-75, and 75-110. The table is sorted by Age, and the selected 50 out of 50 samples are displayed.

Group	Accession	Title	Source name	Tissue	Gender	Age
25-50	GSM2822970	miRNA_BreastCancer_rep1	Breast Cancer tissue	breast cancer	F	45 <= year < 50
25-50	GSM2822971	miRNA_BreastCancer_rep2	Breast Cancer tissue	breast cancer	F	40 <= year < 45
25-50	GSM2822972	miRNA_BreastCancer_rep3	Breast Cancer tissue	breast cancer	F	25 <= year < 30
50-75	GSM2822973	miRNA_BreastCancer_rep4	Breast Cancer tissue	breast cancer	F	50 <= year < 55
50-75	GSM2822974	miRNA_BreastCancer_rep5	Breast Cancer tissue	breast cancer	F	50 <= year < 55
50-75	GSM2822975	miRNA_BreastCancer_rep6	Breast Cancer tissue	breast cancer	F	50 <= year < 55
50-75	GSM2822976	miRNA_BreastCancer_rep7	Breast Cancer tissue	breast cancer	F	50 <= year < 55
25-50	GSM2822977	miRNA_BreastCancer_rep8	Breast Cancer tissue	breast cancer	F	35 <= year < 40
25-50	GSM2822978	miRNA_BreastCancer_rep9	Breast Cancer tissue	breast cancer	F	40 <= year < 45
50-75	GSM2822979	miRNA_BreastCancer_rep10	Breast Cancer tissue	breast cancer	F	60 <= year < 65
50-75	GSM2822980	miRNA_BreastCancer_rep11	Breast Cancer tissue	breast cancer	F	50 <= year < 55
25-50	GSM2822981	miRNA_BreastCancer_rep12	Breast Cancer tissue	breast cancer	F	35 <= year < 40
25-50	GSM2822982	miRNA_BreastCancer_rep13	Breast Cancer tissue	breast cancer	F	30 <= year < 35
50-75	GSM2822983	miRNA_BreastCancer_rep14	Breast Cancer tissue	breast cancer	F	60 <= year < 65
50-75	GSM2822984	miRNA_BreastCancer_rep15	Breast Cancer tissue	breast cancer	F	60 <= year < 65
50-75	GSM2822985	miRNA_BreastCancer_rep16	Breast Cancer tissue	breast cancer	F	50 <= year < 55
50-75	GSM2822986	miRNA_BreastCancer_rep17	Breast Cancer tissue	breast cancer	F	50 <= year < 55
50-75	GSM2822987	miRNA_BreastCancer_rep18	Breast Cancer tissue	breast cancer	F	70 <= year < 75

Figure 4.9 Division of the samples into group.

Then the run option was selected and the top genes are displayed. The genes are sorted on the basis of F value. The genes with the highest F-value were listed first. The highest F-value was of the has-miRNA 1305.

Based on the F value top 20 results were extracted out:

1	ID	adj.P.Val	P.Value	F	miRNA_ID	SPOT_ID
2	hsa-miR-1305	0.0192	0.0000406	12.329	hsa-miR-1305	
3	kshv-miR-K12-10b	0.2502	0.0010558	7.828	kshv-miR-K12-10b	
4	hsa-miR-92a-1*	0.362	0.0022909	6.838	hsa-miR-92a-1*	
5	hsa-miR-100	0.4577	0.0038626	6.186	hsa-miR-100	NA
6	hsa-miR-99b	0.5341	0.0068376	5.489	hsa-miR-99b	
7	hsa-miR-146a	0.5341	0.0099765	5.035	hsa-miR-146a	
8	hsa-miR-210	0.5341	0.0133406	4.691	hsa-miR-210	
9	hsa-miR-550*	0.5341	0.0151596	4.541	hsa-miR-550*	
10	hsa-miR-30c-2*	0.5341	0.0163149	4.455	hsa-miR-30c-2*	
11	hsa-miR-7-1*	0.5341	0.0167557	4.424	hsa-miR-7-1*	
12	hsa-miR-301b	0.5341	0.0179427	4.344	hsa-miR-301b	
13	hsa-miR-219-5p	0.5341	0.0180636	4.336	hsa-miR-219-5p	
14	hsa-miR-30c	0.5341	0.0183269	4.319	hsa-miR-30c	
15	hsa-miR-532-5p	0.5341	0.0200452	4.215	hsa-miR-532-5p	
16	hsa-miR-20a*	0.5341	0.0210649	4.158	hsa-miR-20a*	
17	hsa-miR-484	0.5341	0.0216187	4.128	hsa-miR-484	
18	hsa-miR-200c*	0.5341	0.0218467	4.115	hsa-miR-200c*	
19	hsa-miR-23a	0.5341	0.0231273	4.05	hsa-miR-23a	
20	hsa-miR-892b	0.5341	0.0240059	4.007	hsa-miR-892b	
21						
22						

Figure 4.10 sorted table based on the F value

Further information was collected from NCBI and MiRNA database:

	A	B	C	D	E	F	G	H
1	ID	adj.P.Val	P.Value	F	miRNA_ID	SPOT_ID	Gene ID	SYMBOL
2	hsa-miR-1305	0.0192	0.0000406	12.329	hsa-miR-1305		100302270	MIR1305
3	hsa-miR-92a-1*	0.362	0.0022909	6.838	hsa-miR-92a-1*		407048	MIR92A1
4	hsa-miR-100	0.4577	0.0038626	6.186	hsa-miR-100	NA	406892	MIR100
5	hsa-miR-99b	0.5341	0.0068376	5.489	hsa-miR-99b		407056	MIR99B
6	hsa-miR-146a	0.5341	0.0099765	5.035	hsa-miR-146a		406938	MIR146A
7	hsa-miR-210	0.5341	0.0133406	4.691	hsa-miR-210		406992	MIR210
8	hsa-miR-550*	0.5341	0.0151596	4.541	hsa-miR-550*		693134	MIR550A2
9	hsa-miR-30c-2*	0.5341	0.0163149	4.455	hsa-miR-30c-2*		407032	MIR30C2
10	hsa-miR-7-1*	0.5341	0.0167557	4.424	hsa-miR-7-1*		407043	MIR7-1
11	hsa-miR-301b	0.5341	0.0179427	4.344	hsa-miR-301b		100126318	MIR301B
12	hsa-miR-219-5p	0.5341	0.0180636	4.336	hsa-miR-219-5p		407002	MIR219A1
13	hsa-miR-30c	0.5341	0.0183269	4.319	hsa-miR-30c		407031	MIR30C1
14	hsa-miR-532-5p	0.5341	0.0200452	4.215	hsa-miR-532-5p		693124	MIR532
15	hsa-miR-20a*	0.5341	0.0210649	4.158	hsa-miR-20a*		406982	MIR20A
16	hsa-miR-484	0.5341	0.0216187	4.128	hsa-miR-484		619553	MIR484
17	hsa-miR-200c*	0.5341	0.0218467	4.115	hsa-miR-200c*		406985	MIR200C
18	hsa-miR-23a	0.5341	0.0231273	4.05	hsa-miR-23a		407010	MIR23A
19	hsa-miR-892b	0.5341	0.0240059	4.007	hsa-miR-892b		100126307	MIR892B

Figure 4.11 Updation of the table with information from various sources

Then further info was collected for the miRNA values greater than 5 which includes chromosomal location, their links to NCBI site and mirbase site :

1	ID	adj.P.Val	P.Value	F	miRNA_ID	SPOT_ID	Gene ID	SYMBOL	chromosomal location	info link	info ncbi
2	hsa-miR-1305	0.0192	0.0000406	12.329	hsa-miR-1305		100302270	MIR1305	4q34.3	<a href="http://mirbase.org/miR-1305">http://mirbase.org/miR-1305</a>	which
3	hsa-miR-92a-1*	0.362	0.0022909	6.838	hsa-miR-92a-1*		407048	MIR92A1	13q31.3	<a href="http://mirbase.org/cgi-bin/mirna_e">http://mirbase.org/cgi-bin/mirna_e</a>	In the presence of
4	hsa-miR-100	0.4577	0.0038626	6.186	hsa-miR-100	NA	406892	MIR100	11q24.1	<a href="http://mirbase.org/cgi-bin/mirna_e">http://mirbase.org/cgi-bin/mirna_e</a>	
5	hsa-miR-99b	0.5341	0.0068376	5.489	hsa-miR-99b		407056	MIR99B	19q13.41	<a href="http://mirbase.org/miR-99b">http://mirbase.org/miR-99b</a>	was exp
6	hsa-miR-146a	0.5341	0.0099765	5.035	hsa-miR-146a		406938	MIR146A	5q33.3	<a href="http://mirbase.org/cgi-bin/mirna_e">http://mirbase.org/cgi-bin/mirna_e</a>	
7	hsa-miR-210	0.5341	0.0133406	4.691	hsa-miR-210		406992	MIR210	11p15.5	<a href="http://mirbase.org/cgi-bin/mirna_e">http://mirbase.org/cgi-bin/mirna_e</a>	It has been descr
8	hsa-miR-550*	0.5341	0.0151596	4.541	hsa-miR-550*		693134	MIR550A2			
9	hsa-miR-30c-2*	0.5341	0.0163149	4.455	hsa-miR-30c-2*		407032	MIR30C2			
10	hsa-miR-7-1*	0.5341	0.0167557	4.424	hsa-miR-7-1*		407043	MIR7-1			
11	hsa-miR-301b	0.5341	0.0179427	4.344	hsa-miR-301b		100126318	MIR301B			

Figure 4.12 F value greater than 5 of the results were extracted out.

## DISCUSSION

### Over Expressed top genes:

#### PTHLH:

PTHLH gene is found to be expressed in many of the genes which play an important role in cancers like breast cancer and colon cancer. It was seen in the study that this gene modulates the cell cycle progression. Cell migration also increased when PTHLH was over-expressed. It was also found in the study that it was a positive prognostic indicator in women with lung cancer also.

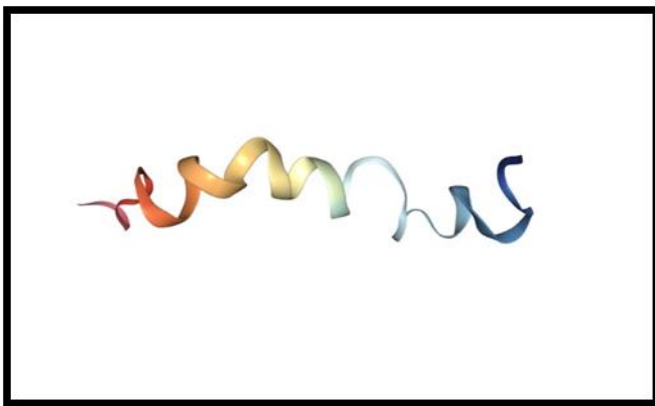


Figure 5.1 PTHLH structure

#### GEN1:

This gene is involved in DNA double strand break. It is also involved in maintaining centrosome integrity. The study showed that GEN1 is involved in differentiation, proliferation of the mammary gland epithelial cells.

The results in the study showed that the damage in the GEN1 may lead to the development of breast cancer in humans.

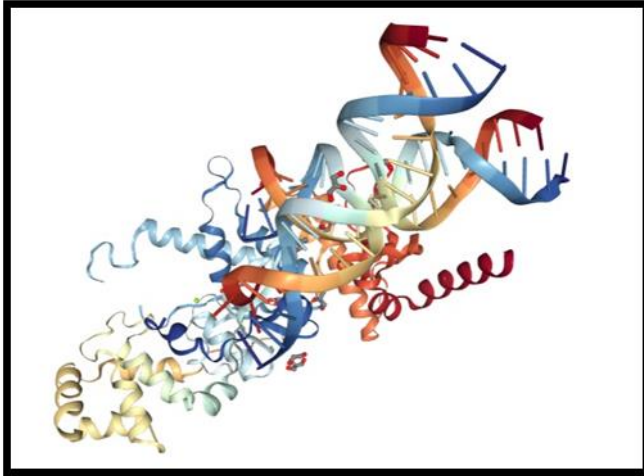


Figure 5.2 GEN1 structure

PTHLH(5744):

It regulates epithelial-mesenchymal interactions during the formation of the mammary glands and teeth. The over-expression or under-expression of this gene is found to contribute in the development of cancer. Knocking down the expression of this gene inhibited the formation of tumours in mice.

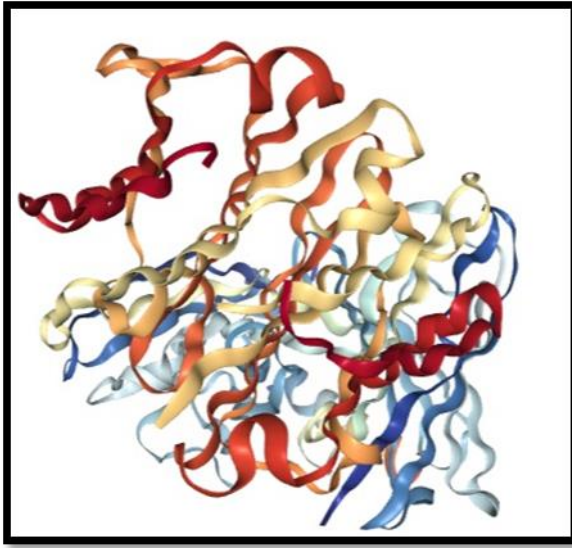


Figure 5.3 PTHLH (5744) structure



## UNDER EXPRESSED TOP GENES

### CLEC4A:

This gene functions in cell signalling and cell adhesion. This gene has shown expression in many of the parts of the body tissues like bone marrow, appendix, spleen, lung and many more. It encodes a member of C-type lectin. This gene plays a role in inflammatory and immune response. It plays a role in modulating dendritic cell modulation and maturation.

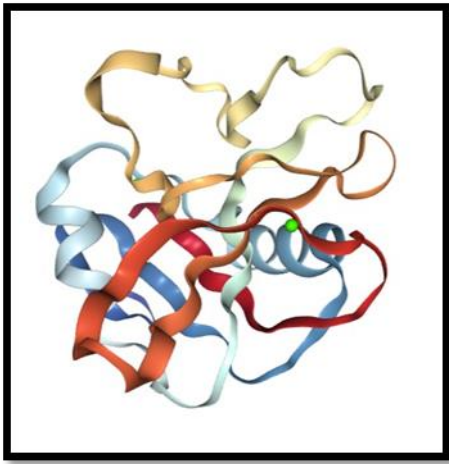


Figure 5.4 CLEC4A structure

### MPZL3:

It is a protein coding gene. It mediates homophilic cell-cell adhesion. This gene has shown broad expressions in skin, oesophagus, thyroid, breast, stomach, lung and many other organs as well. The under-expression of this gene has shown severe skin abnormalities and hair loss as well.

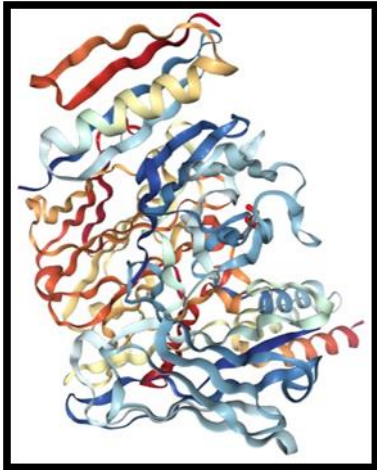


Figure 5.5 MPZL3 structure

#### FAM155A:

The major bias expression of this gene is shown in brain apart from other tissues as well like testis prostate and bladder.

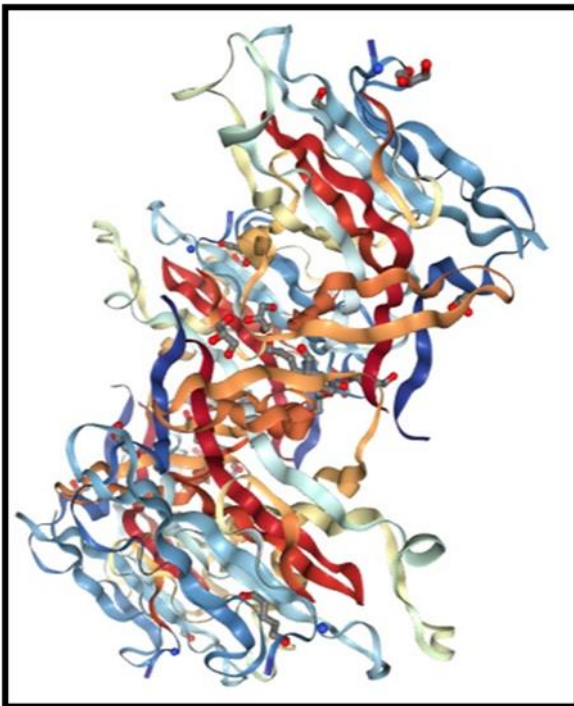


Figure 5.6 FAM155A structure

Top miRNAs:

hsa-miR-1305:

This miRNA's overexpression is responsible for induced differentiation of pluripotent stem cells, increased cell differentiation during the cell cycle. The down regulation of this miRNA 1305 resulted in a better cell life.

In a study, a genome wide association analyses for smokers and non-smokers in Indian and Bangladesh was performed in which miRNA 1305 was the 3rd most affecting gene which further led to cancer symptoms in people :

MIR1305||TENM3    4q35.1 rs11724903    G    0.09    0.13     $2.90 \times 10^{-7}$

hsa-miR-92a-1:

miRNA 92 has been found to be a major cancer related in many of the cancers types. Its upregulation was found in many of the cancer cell lines. Its high expression is correlated with clinical stage.

Patients having high expression of miRNA 92a had poorer survival rates. The overexpression of miRNA 92a has shown increased cell cycle rate and also apoptosis of U2OS cells.

In a separate study, it was found that this has-miR-92a acts as an onco-miRNA and contributed to the progression and invasion of cancer, specifically cervical and breast cancer. The serum extracted had high amount of has-miR-92a in the advanced stages of cancer than the stage I and stage II cancer.

Hence upregulation of hsa-miR-92a was associated with progression of cancer.

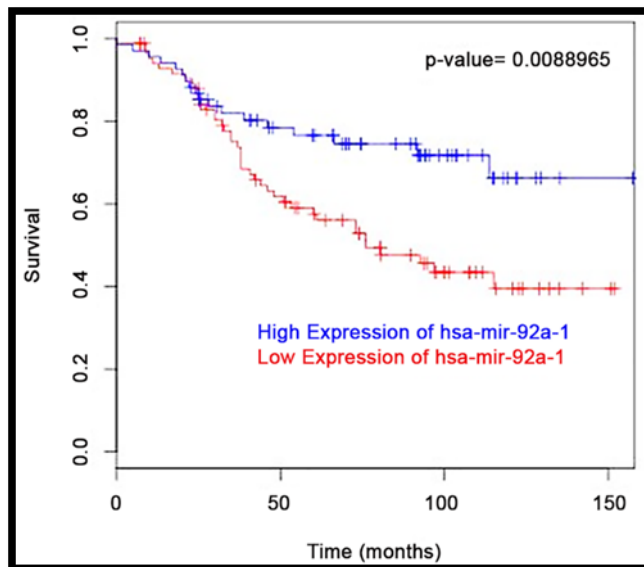


Figure 5.7 Expression studies of has-mir-92a-1 [13]

#### hsa-miR-100:

It has shown to be a precursor for ovarian and breast cancer progression in women. The over expression of this miRNA has shown increased and fast cell cycle.

In another study, in an OS specimen, it was found that the samples contained lower levels of miR-100. Lower levels of miR-100 were associated with poor prognosis of the OS patients.

#### hsa-miR-99b:

Over expression or under-expression of this miRNA results in proliferation of cell cycle. Elevated levels of this miRNA in patients showed a poor overall survival rate in hepatocellular carcinoma.

“Overexpression or knockdown of miR-99b expression increased or inhibited, respectively, the metastasis of HCC cells in vitro.

These findings suggest that a high level of miR-99b expression is an independent prognostic factor and correlates with poor survival of patients with HCC. Therefore, inhibition of miR-99b expression may serve as a therapeutic approach for inhibiting the metastatic phenotype of HCC.”

hsa-miR-146a:

“The microRNA-146a and microRNA-146b (miR-146a/b) when expressed in the highly metastatic human breast cancer cell line MDA-MB-231 function to negatively regulate NF- $\kappa$ B activity.”

Also, in most of the studies the down regulated levels of hsa-miR-146a and also hsa-miR-146b were found to be an important or potential biomarkers in young women suffering from breast cancer.

## CHAPTER 6

### CONCLUSION:

The BC prediction model was developed. The dataset was provided by Kaggle. The training and the testing of the model was done. The predictive model was developed and various measures were obtained which tells the accuracy, f-measures, recall and precision for the dataset. The graph was then plotted in R.

Analysis of two datasets was carried out which consisted of miRNA 203 affected and controlled dataset and the other dataset was of Japanese patients suffering from BC of different age groups. The comparative analysis was done through the microarray data analysis tool: GEO2R. Certain genes were found to be of much importance than the others which signified a major role in the progression of BC.

Some of the potential biomarkers found were

## REFERENCES

[1]"GEO2R - GEO - NCBI", Ncbi.nlm.nih.gov, 2018. [Online]. Available: <https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE105134>. [Accessed: 19- Feb- 2018].

[2]G. Database, "GeneCards - Human Genes | Gene Database | Gene Search", Genecards.org, 2018. [Online]. Available: <https://www.genecards.org/>. [Accessed: 19-May- 2018].

[3]"mahmoudparsian/data-algorithms-book", GitHub, 2018. [Online]. Available: <https://github.com/mahmoudparsian/data-algorithms-book/tree/master/src/main/java/org/dataalgorithms/machinelearning/logistic/cancer>. [Accessed: 9- April- 2018].

[4]R. Bank, "RCSB PDB: Homepage", Rcsb.org, 2018. [Online]. Available: <https://www.rcsb.org/>. [Accessed: 19- March- 2018].

[5]M. Salmans, F. Zhao and B. Andersen, "The estrogen-regulated anterior gradient 2 (AGR2) protein in breast cancer: a potential drug target and biomarker", Breast Cancer Research, vol. 15, no. 2, 2013.

[6]R. Sandhu, J. Rein, M. D’Arcy, J. Herschkowitz, K. Hoadley and M. Troester, "Overexpression of miR-146a in basal-like breast cancer cells confers enhanced tumorigenic potential in association with altered p53 status", 2018. .

[7]A. Corter, R. Broom, D. Porter, V. Harvey and M. Findlay, "Predicting nonadherence to adjuvant endocrine therapy in women with early stage breast cancer", *Psycho-Oncology*, 2018.

[8]Q. Kong, Z. Tang, F. Xiang, J. Jiang, H. Yue, R. Wu and X. Kang, "Diagnostic Value of Serum hsa-mir-92a in Patients with Cervical Cancer", *Clinical Laboratory*, vol. 63, no. 022017, 2017.

[9]J. YANG, X. LIU, X. YUAN and Z. WANG, "miR-99b promotes metastasis of hepatocellular carcinoma through inhibition of claudin 11 expression and may serve as a prognostic marker", *Oncology Reports*, vol. 34, no. 3, pp. 1415-1423, 2015.

[10]"Apache Hadoop 2.7.4 – Hadoop: Setting up a Single Node Cluster.", *Hadoop.apache.org*, 2018. [Online]. Available: <https://hadoop.apache.org/docs/r2.7.4/hadoop-project-dist/hadoop-common/SingleCluster.html>. [Accessed: 19- May- 2018].



[11]"HADOOP INSTALLATION — Installation and Configuration 1.0.1 documentation", Doctuts.readthedocs.io, 2018. [Online]. Available: <http://doctuts.readthedocs.io/en/latest/hadoop.html>. [Accessed: 19- May- 2018].

[12]K. Boras-Granic and J. Wysolmerski, "PTHrP and breast cancer: more than hypercalcemia and bone metastases", *Breast Cancer Research*, vol. 14, no. 2, 2012.

[13]M. Lukamowicz-Rajska, C. Mittmann, M. Prummer, Q. Zhong, J. Bedke, J. Hennenlotter, A. Stenzl, A. Mischo, S. Bihr, M. Schmidinger, U. Vogl, I. Blume, C. Karlo, P. Schraml and H. Moch, "MiR-99b-5p expression and response to tyrosine kinase inhibitor treatment in clear cell renal cell carcinoma patients", *Oncotarget*, vol. 7, no. 48, 2016.

[14]"GLOBOCAN Cancer Fact Sheets: Breast cancer", *Globocan.iarc.fr*, 2018. [Online]. Available: <http://globocan.iarc.fr/old/FactSheets/cancers/breast-new.asp>. [Accessed: 19- May- 2018].

[15]A. Zedan, T. Hansen, J. Assenholt, M. Pleckaitis, J. Madsen and P. Oster, "microRNA expression in tumour tissue and plasma in patients with newly diagnosed metastatic prostate cancer", *Tumor Biology*, vol. 40, no. 5, p. 101042831877586, 2018.

[16]F. Jerry R. Balentine, "Breast Cancer Causes, Types, Symptoms, Signs, Stages, Treatment", MedicineNet, 2018. [Online]. Available: [https://www.medicinenet.com/breast\\_cancer\\_facts\\_stages/article.htm](https://www.medicinenet.com/breast_cancer_facts_stages/article.htm). [Accessed: 19-May- 2018].

[17]M. Christina Chun, "Breast cancer: Symptoms, risk factors, and treatment", Medical News Today, 2018. [Online]. Available: <https://www.medicalnewstoday.com/articles/37136.php>. [Accessed: 19- May- 2018].

[18]B. cancer, "Breast cancer: MedlinePlus Medical Encyclopedia", Medlineplus.gov, 2018. [Online]. Available: <https://medlineplus.gov/ency/article/000913.htm>. [Accessed: 06- Nov- 2017].

[19]L. Greeshma and G. Pradeepini, "Big Data Analytics with Apache Hadoop MapReduce Framework", Indian Journal of Science and Technology, vol. 9, no. 26, 2016.

[20]"Apache Hadoop 2.9.1 – HDFS High Availability Using the Quorum Journal Manager", Hadoop.apache.org, 2018. [Online]. Available: <https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HDFSHighAvailabilityWithQJM.html>. [Accessed: 03- Mar- 2018].

[21]"Bioinformatics tools for Transcription analysis - OMICtools", omictools, 2018. [Online]. Available: <https://omictools.com/gene-expression-analysis-category>. [Accessed: 19- May- 2018].

[22]R. Taylor, "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics", BMC Bioinformatics, vol. 11, no. 12, p. S1, 2010.

[23]R. Taylor, "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics", BMC Bioinformatics, vol. 11, no. 12, p. S1, 2010.

[24]"RefSeq: NCBI Reference Sequence Database", Ncbi.nlm.nih.gov, 2018. [Online]. Available: <https://www.ncbi.nlm.nih.gov/refseq/>. [Accessed: 19- May- 2018].

[25]"NCBI's genome browser - Genome Data Viewer", Ncbi.nlm.nih.gov, 2018.  
[Online]. Available: <https://www.ncbi.nlm.nih.gov/genome/gdv/>. [Accessed: 19- May- 2018].

[26]L. Sun, Y. Zhang, Z. Pan, B. Li, M. Sun and X. Zhang, "Expression and Localization of GEN1 in Mouse Mammary Epithelial Cells", Journal of Biochemical and Molecular Toxicology, vol. 28, no. 10, pp. 450-455, 2014.

[27]F. Lerebours, G. Cizeron-Clairac, A. Susini, S. Vacher, E. Mouret-Fourme, C. Belichard, E. Brain, J. Alberini, F. Spyrtos, R. Lidereau and I. Bieche, "miRNA expression profiling of inflammatory breast cancer identifies a 5-miRNA signature predictive of breast tumor aggressiveness", International Journal of Cancer, vol. 133, no. 7, pp. 1614-1623, 2013.

[28]E. Elghoroury, H. ElDine, S. Kamel, A. Abdelrahman, A. Mohammed, M. Kamel and M. Ibrahim, "Evaluation of miRNA-21 and miRNA Let-7 as Prognostic Markers in Patients With Breast Cancer", Clinical Breast Cancer, 2017.

[29]A. Sahlabadi, R. Chandren Muniyandi, M. Sahlabadi and H. Golshanbafghy, "Framework for Parallel Preprocessing of Microarray Data Using Hadoop", 2018. .

[30]M. Kallio, J. Tuimala, T. Hupponen, P. Klemelä, M. Gentile, I. Scheinin, M. Koski, J. Käki and E. Korpelainen, "Chipster: user-friendly analysis software for microarray and other high-throughput data", BMC Genomics, vol. 12, no. 1, 2011.

[31]R. Hamam, A. Ali, K. Alsaleh, M. Kassem, M. Alfayez, A. Aldahmash and N. Alajez, "microRNA expression profiling on individual breast cancer patients identifies novel panel of circulating microRNA for early detection", Scientific Reports, vol. 6, no. 1, 2016.

[32]"Female Breast Cancer - Cancer Stat Facts", Seer.cancer.gov, 2018. [Online]. Available: <https://seer.cancer.gov/statfacts/html/breast.html>. [Accessed: 19- May- 2018].

[33]F. Wahid, A. Shehzad, T. Khan and Y. Kim, "MicroRNAs: Synthesis, mechanism, function, and recent clinical trials", 2018.

[34]Intellipaat.com, 2018. [Online]. Available: <https://intellipaat.com/tutorial/big-data-and-hadoop-tutorial/the-hadoop-module-high-level-architecture/>. [Accessed: 19- May- 2018].