CLASSIFICATION OF BREAST CANCER CELL LINES INTO SUBTYPES BASED ON GENETIC PROFILES

ANIRUDDHA VIKRAM PAWAR

Submitted to the faculty of Bioinformatics Graduate Program

in partial fulfillment of requirements for the degree

Master of Science in Bioinformatics

in the School of Informatics

Indiana University

MAY 2015

Accepted by the Faculty of Indiana University,

in partial fulfillment of the requirements for the degree of Master of Science

in Bioinformatics

_____

Dr. Lang Li, Ph.D., Chairperson

Master's Thesis

Committee

_____

Dr. Yunlong Liu, Ph.D.

_____

Dr. Xiaowen Liu, Ph.D.

Dedicated to my parents

ACKNOWLEDGMENTS

This thesis would not have materialized if it was not for the guidance and support from numerous people to whom I will always be indebted. Therefore I would like to thank everybody from bottom of my heart.

I am extremely grateful to my thesis advisor Dr. Lang Li for providing me with this opportunity and for his constant encouragement, excellent acumen, immense knowledge and constructive criticism. I would like to extend my gratitude to committee members Dr. Yunlong Liu and Dr. Xiaowen Liu for their valuable remarks. I sincerely thank Dr. Arindom Chakraborty and Dr. Lijun Chen for their guidance and helpful suggestions.

Special thanks to School of Informatics at Indiana University Purdue University Indianapolis for providing me an opportunity to learn and pursue an excellent career in bioinformatics.

I would like to thank Sumit Makshir for his timely inputs. I am also grateful to my friends and lab mates Akshay, Aditya, Chaitanya, Debolina, Guanglong, Hrishikesh, Michael, Mayuresh, Prasad, Swapnil and Swati for their unbiased comments and relentless assistance.

This acknowledgment would remain incomplete without expressing my gratitude towards my family. I sincerely thank my parents Mr. Vikram Pawar and Mrs. Anjali Pawar for their unceasing care, love and blessings. I would also like to thank my brother Uddhav for his support and encouragement. Finally, I would like to thank my fiancé, Neha who was always there cheering me up through the good and bad times.

ABSTRACT OF DISSERTATION

Title: Classification of breast cancer cell lines into subtypes based on genetic profile

Abstract: Today we know that there are several different types of breast cancer. Accurate identification breast cancer subtype is extremely important in treating this disease effectively. Consequently the process of invtro development of drugs to treat this disease should be naturally subtype specific. Until now several studies have identified multiple breast cancer cell lines and these cell lines have served as invaluable invitro tumor models. However very few of these cell lines are classified as per their subtypes. In this thesis an effort is made to classify 59 of such breast cancer cell lines using genetic profile comparison approach. This approach is based on comparing characteristic features such as copy number and gene expression of a given cell line to those observed from the tissue samples of different breast subtypes. The tissue data for this comparison comes from The Cancer Genome Atlas (TCGA) while cell line data is taken from Cancer Cell Line Encyclopedia (CCLE).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTODUCTION

1.1 Breast Cancer

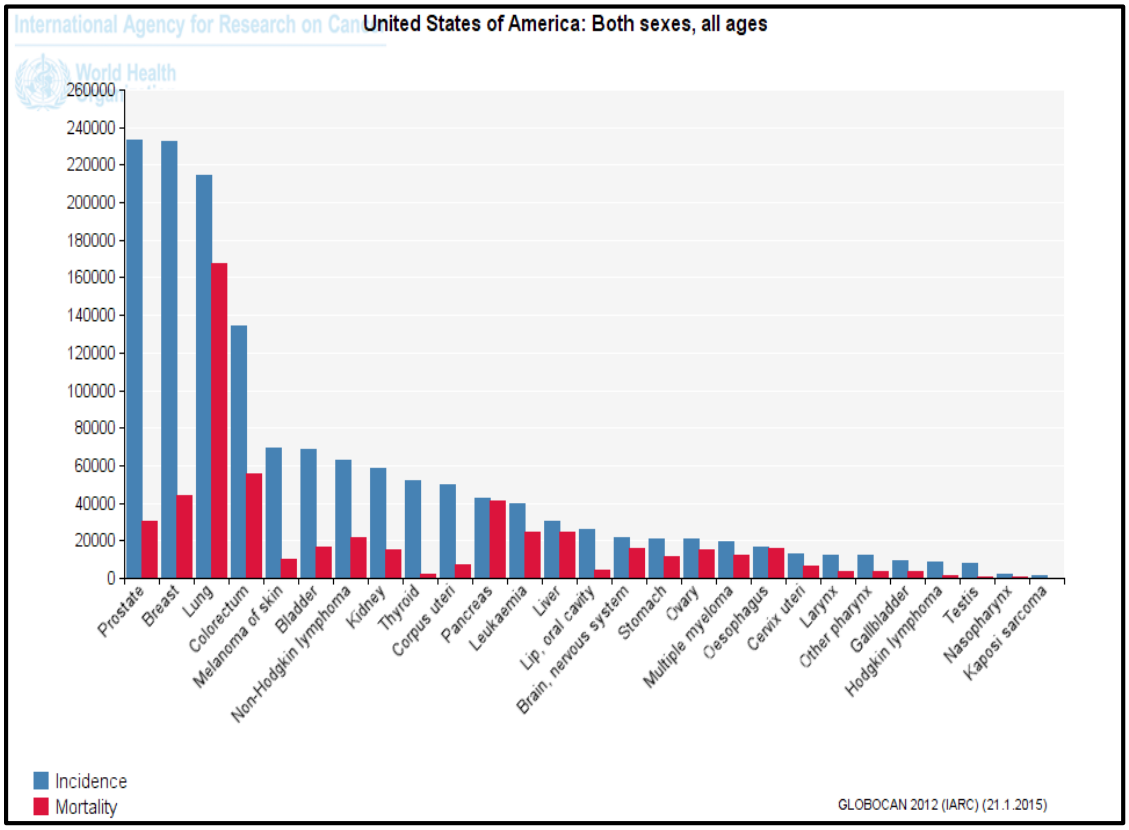
Cancer of cells of the breast is known as breast cancer. It is a condition characterized by uncontrolled growth of cells in breast tissue, which may lead in formation of a lump in breast. Actually breast cancer is a complex disease, in fact each of its subtype has its own clinical and pathological characteristic. Cancer originating from the lobules is called as lobular carcinoma, while cancer resulting from the ductal cells of breast is called as ductal carcinoma. Once the cancer spreads from its origin to other tissues it is termed as invasive carcinoma. Normally cancer limits to the tissue of origin and known as non-invasive form of carcinoma. Both ductal and lobular cancer can metastasize or spread to other parts of the body and may turn into invasive carcinomas. Invasive ductal carcinoma constitutes 80% of total invasive cancers while Invasive lobular carcinoma occurs around 10% to 15% of times. Other breast cancer types which are uncommon include inflammatory breast cancer, Phyllodes breast cancer, Paget disease of the nipple[1].

The risk of breast cancer is high among women and is one of the major cause of death, next to lung cancer [1]. In the US chances are that every one out of eight women will suffer from breast cancer [1] . In 2015 nearly 0.23 million new cases of breast cancer in women will be reported and around 40,000 women will lose their life to breast cancer [1]. According to American Cancer Society, the majority i.e. more than half of

the breast cancer cases belong to invasive ductal carcinoma, while breast cancer remains the most common condition affecting women [1]. Although breast cancer majorly occurs in females but there are rare cases of breast cancer in men. American Cancer Society states that there will be around 2000 new cases and 400 deaths amongst men due to breast cancer in the year 2015[1].

The risk of breast cancer is associated with multiple important factors. Age is one the most crucial factor[2]. The chances of breast cancer increase significantly with old age. American Cancer Society states among the first time detection of invasive ductal carcinoma cases, 33% of the women are 55 or above[1]. Gender is also a crucial factor, with significant risk in case of female rather than male. The risk is also associated with genetic factors or in other words it can be hereditary in nature. Out of all cases limited number of cases, that is 5% have hereditary mutations in breast cancer associated genes BRCA1, BRCA2 and TP53[3]. There is an increased chance of breast cancer when a direct relative suffers from it. Apart from age, gender, hereditary mutations other factors like menarche, menopausal age, obesity and ethnicity are found to be associated with breast cancer. Lifestyle habits like diet, smoking, exposure to radiation seem to have associated with occurrences of breast cancer.

Figure 1 Breast cancer incidence and mortality



Source: International Agency for Research on Cancer (IARC) and World Health Organization (WHO)[4]

Breast cancer can be treated. The percentage of survival has increased in present times as compared to few decades ago. This success can be contributed to the growth of knowledge about breast cancer as well as to the recent developments in medical science. Breast cancer can be treated either by surgery, chemotherapy or radiation. Studies have shown that combinatorial approach is more beneficial[5]. Treatment via surgery is literally the removal of cancer tissue from breast and for this purpose entire breast can be salvaged or infected part is dissected. In case of chemotherapy, drugs are administered to stop the growth of affected cells by blocking important

pathways involved in cancer progression. In some cases hormone blocking drugs or even monoclonal antibodies are administered depending upon the subtype of breast cancer. Radiation therapy involves bombardment of affected parts of breast with small amount of radiation which destroys cancer cells. Most of these methods have serious side effects including severe damage of neighboring healthy cells.

1.2 Breast Cancer Classification

As mentioned earlier breast cancer is heterogeneous in nature. It is therefore necessary to classify breast cancer into different types in order to capture the characteristic essence of individual elements of this heterogeneous disease. Each subclass of cancer are driven by characteristic features which may vary across the subclasses making breast cancer treatment difficult. There is no single treatment for all the subclasses. In fact each of the subclasses may have different treatment regimen consisting of two or more therapies or even combination therapy[5]. There are numerous method for classifying breast cancer and each of them is based on different principles. Some of the classification methods are stage, grade, histopathology and receptor status. All most all of these tests require study of cancer tissues. Generally procedure such as biopsy is used to obtain a small portion of breast cancer tissue from the patient (i.e. surgically removed) and is used for classification as well as further genetic analysis

Stage

This method of breast cancer classification is based on visual observation of cancer tissue obtained from patients. One of the most common method belonging to this category is TNM classification. This method was developed by Pierre Denoix and is maintained by Union of International Cancer Control[6]. This classification method is applicable to all forms of cancers and is not limited to breast cancer.

T – Denotes size of the tumor

N – Denotes the spread of the tumor to nearby lymph nodes

M – Denotes that cancer has undergone metastasis i.e. it has spread and infected other part(s) of the body

Further stages for breast cancer are denoted as follows

Stage 0 – Cancer is limited to tissue of origin

Stage 1, 2 & 3 – Cancer has grown in size and has affected lymph nodes

Stage 4 – Cancer has metastasized and spread to other part(s) of the body

Grade

Grade is a classification method based on differentiation of cancerous cells. A normal cell of breast is well differentiated and is designated to carry out a specific function. Like ductal cells are responsible for the formation of ducts. In cases of cancer these specialized cells begin losing their specialization/differentiation. Typically a pathologist would look at the biopsy sample and compare it with normal tissue to

determine dissimilarity or determine the amount of differentiation cancerous cells have lost. This classification method is applicable to all forms of cancers. The most used scale for grading is called as Elston-Ellis grade[6].

Grade 1 or low-grade – Small difference between normal and cancerous cells

Grade 2 or intermediate-grade – Cancerous cells are moderately differentiated

Grade 3 or high-grade – Cancerous cells completely lose their differentiation compared to normal cells

Histopathology

It is the direct observation of cancerous tissue obtained from biopsy. This is very common and foremost method where tissue samples are studied by a pathologist to determine the characteristic features. This method determines whether tissue exhibits in-situ or invasive behavior and other histologic features like inflammation, tubular and differentiation. Basic and important information like location of the tumor is also recorded. This method is also used to study other types of cancers.

Receptor Status

This is an important method for classification of breast cancer into subtypes based on the receptor information. Receptors are responsible in cell signaling. Hormones or other protein molecules bind receptors present on the surface of cell and bring about desired change in cell. Breast cancer cells may sometime show the presence of receptors, which in fact helps in deciding the course of treatment. Biopsy samples are always checked for presence of the receptors.

Currently there are three major receptors which are observed in breast cancer cells.

Estrogen receptor (ER)

Progesterone receptor (PR)

HER2/neu (HER2)

Estrogen binds to estrogen receptor (ER). Breast cancer cells which have estrogen receptor require estrogen for their growth and are denoted as ER+. Almost one third of invasive breast cancer cases are ER+[7]. Similarly progesterone binds to progesterone receptors present on breast cancer cells. This is denoted by PR+. In case of HER2+, cancer cells over express HER2/neu protein which plays an important role in growth of cancer cells.

Although the receptors are involved in growth of cancer cells, but they are quite important in treatment of cancer. Blocking of key hormones or proteins can stop the growth of cancer. Drugs have been designed to block hormones like estrogen and progesterone and thus help in treating ER+ and PR+ breast cancer cases. In case of HER/neu monoclonal antibodies are designed to bind to the protein and make it unavailable for cancer cells[8]. Due to the importance of receptors in cancer treatment new/other receptors are now being studied in detail.

For classification purpose breast cancer cells can have all three, few (one/two) or none of the receptors. And most commonly used classification method based on the receptor status makes use of all three receptors in classification.

Based on the receptors breast cancer can be classified as

1. Luminal A
2. Luminal B
3. HER2 positive
4. Basal or Triple negative

Luminal A subtype of breast cancer is characterized by presence of two receptor types. Estrogen receptors and progesterone receptors. This subtype does not show the presence of over expressed HER2/neu protein. To summarize Luminal A is ER+, PR+ and HER2/neu -.

Luminal B subtype of breast cancer is also characterized by presence of two receptors types. Estrogen receptors and progesterone receptors. However this subtype does show the presence of over expressed HER2/neu protein. To summarize Luminal B is ER+, PR+ and HER2/neu +.

Basically the luminal subtype is characterized by the presence of estrogen receptors alone[7].

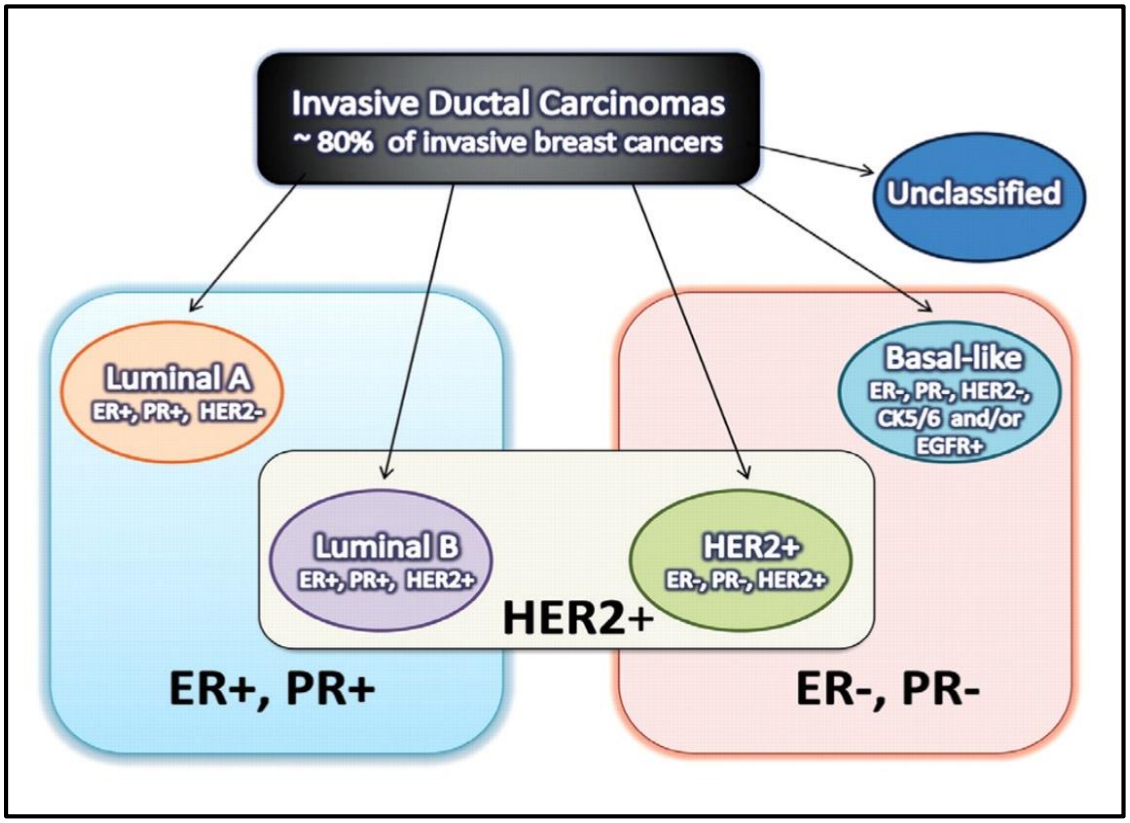
HER2/neu positive subtype of breast cancer is characterized by the absence of both estrogen receptors and progesterone receptors. But this subtype is characterized by over expression of HER2/neu protein. To summarize it is ER-, PR- and HER2/neu +.

Basal subtype is a special case. This subtype lacks estrogen receptors as well as progesterone receptors. Also cells do not show overexpression of HER2/neu protein. In fact it is also known as triple negative and this breast cancer subtype is denoted as TNBC (Triple Negative Breast Cancer). Most of the times the term Basal and TNBC are interchangeably used. This subtype is very difficult to treat as it does not have

important receptors. This subtype has poor prognosis and survival rate[7]. As there are drugs to block estrogen and progesterone and monoclonal antibodies for HER2/neu protein, these subtypes Luminal A, Luminal B and HER2/neu have comparatively better prognosis and survival rate[7].

In this study we will be using 4 subtype based breast cancer classification.

Figure 2 Breast cancer classification and its subtypes[7]



Source:[7]

1.3 Breast Cancer Cell lines

Cell lines are cell cultures which are grown in vitro in laboratory. In another words they are colonies of cells which are grown artificially under controlled conditions[9].

Cell lines have great pharmacological importance as they are used to study effects of various drugs. Almost all drugs are to be tested on cell lines in pre-clinical settings making cell lines inevitable part of drug development. Further cell line can be used to study cellular mechanisms and to determine new potential drug target sites[10].

Usually cancer cells are obtained from patient's tumor and cultivated into cell lines. Cell lines obtained specifically from breast cancer cells are known as breast cancer cell lines. These cell lines are very important aspect of breast cancer research. They provide continuous source of homogenous cancer cells[11]. BT-20 was the first breast cancer cell line to be established in 1958[12]. However development of cancer cell line is very tedious process with low success rate. This is because cells often get contaminated by the stromal cells. Even today the number of breast cancer cell lines available is below 100[11].

Following is a list of 59 breast cancer cell clines used in this thesis obtained from Cancer Cell Line encyclopedia[13](CCLE).

CHAPTER TWO

BACKGROUND

2.1 Research in Breast Cancer

Cancers are very intricate diseases. They abruptly start growing in the body. Various reasons are being cited for development of cancer but none of them have helped in eradicating cancer. It is therefore necessary to genetically study cancers. Because cancer cells arise from normal cells. Only after studying the genetic mechanism one can answer questions like what part of the genome is altered in case of cancer, what are the genes activated or overexpressed, which genes are inactivated or mutated and what sections of chromosomes are lost or gained.

A lot of new studies are now being published which make use of advance techniques like microarray, next generation sequencing and mass spectrometry etc. Breast cancer studies have revealed many insights. Scientific studies have published methods to classify breast cancer. Based on important genes various assays are now present to classify and predict breast cancer. Some of these assays are commercially available (PAM50®, MammaTyper®, MammaPrint®, Oncotype DX®, Endopredict®, Genomic Grade Index®)[14]. These methods use gene signatures to correctly estimate prognosis and thereby help in designing or choosing better therapy.

Most of the available gene signatures assume that breast cancer is not a single disease but a complex disease and try to classify it into different subtypes. This subtyping allows better treatment as these subtypes have different mechanism and thereby

different targets. Also these signature address the problem of prognosis and avoid overtreatment of patients by correctly identifying the type of breast cancer[14].

A brief information of commercially available methods.

PAM50 classifier

PAM50 classifier uses 50 genes as a gene signature to classify the breast cancer[14]. The reported subtypes are Luminal A, Luminal B, HER2 and Basal. Actually it is a trained classifier which utilizes feature of 50 genes to classify breast cancer into one of the four types. This technique is based on Nanostring nCounter technology[15].

The list of 50 genes[16].

Table 1 List of PAM50 genes[16].

ACTR3B	ANLN	BAG1	BCL2	BIRC5	BLVRA	CCNB1	CCNE1	CDC20	CDC6
CDH3	CENPF	CEP55	CXXC5	EGFR	ERBB2	ESR1	EXO1	FGFR4	FOXA1
FOXC1	GPR160	GRB7	KIF2C	KRT14	KRT17	KRT5	MAPT	MDM2	MELK
MIA	MKI67	MLPH	MMP11	MYBL2	MYC	NAT1	NDC80	NUF2	ORC6L
PGR	PHGDH	PTTG1	RRM2	SFRP1	SLC39A6	TMEM45B	TYMS	UBE2C	UBE2T

MammaTyper

MammaTyper is a kit which helps in classification of breast cancer into above mentioned four types on the basis of mRNA levels of four important genes. This technique quantitatively measures ER, PR, HER2 and Ki67 based on mRNA levels. The kit was developed by BioNTech AG[14]. It is also a classifier which uses information of the four genes generated from qRT-PCR to classify breast cancer. This method is good at distinction between Luminal A and Luminal B as it uses Ki67 level for identification of the subtype[14].

Mammaprint

This assay was developed by Agendia BV[14]. This method is based on gene signature of 70 identified genes. Mammaprint is predictive test which utilizes the information obtained from the 70 gene set to identify the risk of metastasis in breast cancer patient. Mammaprint is used to determine the chances of recurrence of breast cancer after 5 years and thus helps in determining the survival analysis. It classifies the patient into low risk and high risk of cancer reoccurrence[14].

Oncotype DX

This assay was developed by Genomic Health Inc. It is most widely used gene expression assay. This method is based on set of 21 genes. Most of these genes are related to ER, HER2 and proliferation and are measure of their characteristics. Oncotype DX is very good prognostic as well as predictive test. The result of such test provides risk assessment in treated breast cancer patient for its reoccurrence.

Oncotype DX classifies patient into three risk zones based on the 21 gene classifier, as high risk or moderate risk or low risk[14].

Endopredict

This test was developed by Sividon Diagnostics GmbH[14]. It is a prognostic test. The gene signature consist of 12 genes. However the prognosis is limited to ER + and HER2 – cases of breast cancer patients. Out of 12 genes, 4 genes are used as internal control while the 8 genes are used in assay to predict the risk of reoccurrence in case of treated breast cancer patients. The scoring system ranges from 0 to 15 and classifies patients as low risk or high risk[14].

Genomic Grade Index

This index was developed by Ipsogen S.A., Marseille. It is based on gene set which contains 97 genes. Genomic grade index is a prognostic test. The test is good for ER + breast cancer patient and also helps in grading breast cancer tumors. The result from the index gives risk of reoccurrence of cancer in next 5 years in patients who are completely cured. Generally it classifies patient based on 97 gene sets as high risk or a low risk patient[14].

Table 2 Table of gene signature based commercial assays[14].

Name of Assay	**Type of Test**	**Technology**	**Signature gene set**	**Result (Subclass/ risk)**
PAM50	Classifier, prognostic test	Nanostring nCounter	50 genes	4 class subtyping
MammaTyper	Classifier	qRT-PCR	4 genes	4 class subtyping
MammaPrint	Prognostic & predictive test	DNA microarrays	70 genes	low risk vs. high risk
Oncotype DX	Prognostic & predictive test	qRT-PCR	16 genes	high vs. moderate vs. low risk
Endopredict	Prognostic test	qRT-PCR	8 genes	low risk vs. high risk
Genomic Grade Index	Prognostic test	DNA microarrays	97 genes	low risk vs. high risk

Many breast cancer studies focus on one or two technologies. They themselves reveal lot of information by putting together the data and analyzing it. Studies have used mRNA expression profiling, DNA copy number analysis and also next generation sequencing techniques in combination [17-19].

The Cancer Genome Atlas is an organization with an aim to catalogue all the genetic mutations responsible for cancer[19]. TCGA has collected data about breast cancer using 6 different technologies. This has resulted in identification of many subtype specific mutations. Mutations accompanied with copy number alterations can reveal many drug targets. This breast cancer study cohort consisted of 825 patients[19]. Exome sequencing data is available for 510 tumors amongst 507 patients[19]. mRNA sequencing using Agilent platform was done for 547 samples , while copy number changes were detected using Affymetrix 6.0 single nucleotide polymorphism array (SNP- array)[19].

The study detected that certain mutations are associated with subclass of breast cancer. Nearly half of luminal A subtype were seen to have mutation in PIK3CA gene, luminal A was also dominated by MAP3K1 and MAP2K4 mutations[19]. The study reported that luminal A subtype was enriched in specific mutations in 3 genes (GATA3, PIK3CA, MAP3KI) while 29 % luminal B cases had TP53 and PIK3CA mutations. Samples of HER2 subtype of breast cancer show overexpression of HER2 and also had the characteristics TP53 (72%) and PIK3CA (39%) mutations. Samples belonging to basal class had characteristics TP53 mutations mainly in the version of non-sense and frame-shift mutations[19]. The study also compared and stated that basal subtype had lot of similarities with high-grade serous ovarian tumors and suggested that they may share drug targets on basis of commonalities.

TCGA also has a data portal where all the data is made available to researchers. Along with data from different technologies TCGA also have different types of data such raw

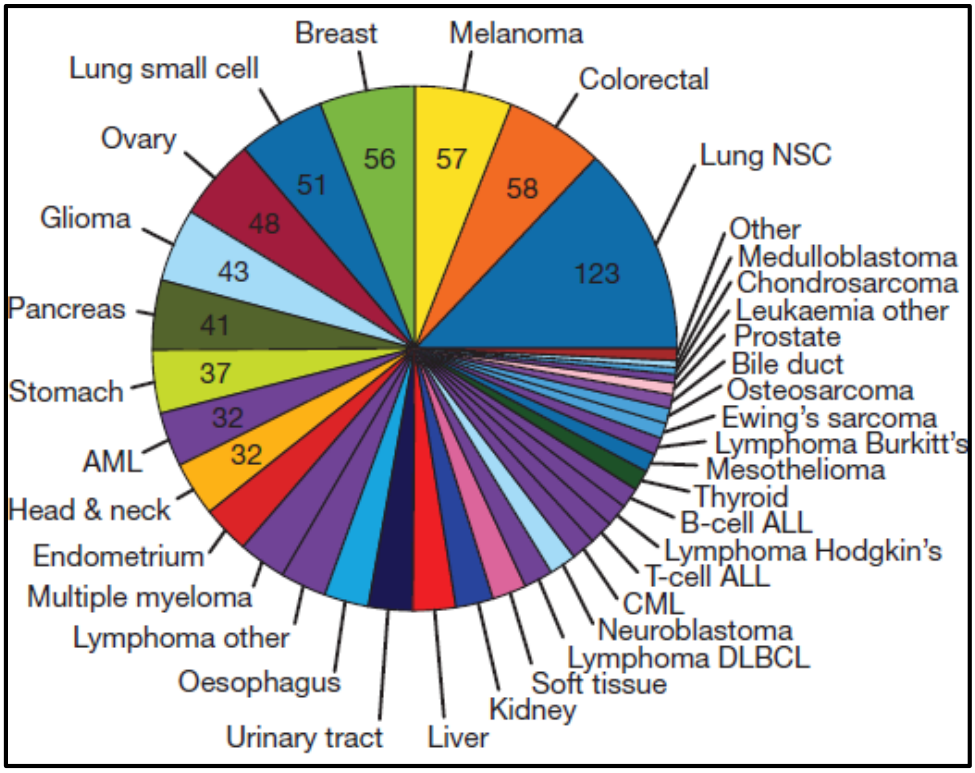
data to various forms of annotated data. Depending upon the annotation the data is assigned different levels. Level 3 is the highest annotated level and completely devoid of any patient information. Data which does not have any patient information is generally available for public to download freely. However access is controlled for different levels and access can be granted for research purposes after filing an application. Currently TCGA holds data for 34 types of cancer[19].

Knowledge gained about cancer from systematic research is interpreted for development of tumor biology and therapeutic opportunities. The outcomes must be verified and confirmed with the help of model system [13]. Cell lines provide to be a great model system. They are like the photographs of cells and actually reflect the genomic diversity of the cancers. Plus they are continuous source of homogenous cells. The results obtained using such cells are reproducible due to their continuous and homogenous nature. Further cell lines are easily available as an effort of commercial cell line banks like American Type Cell line collection (ATCC). These cell line banks maintain, cultivate and distribute numerous cell lines. As a result of which outcome obtained using cell line as model in one laboratory can be replicated at other laboratories.

Genomic profiles of cell lines must be studied. There are several cell line databases which have compiled together various omics aspect of cell lines. Cancer Cell Line Encyclopedia (CCLE) is one such database which has large collection of cell lines[13]. In fact it has 947 human cancer cell lines representing over 36 tumor types. CCLE holds gene expression, chromosomal copy number as well as massively parallel

sequencing data for the 947 cell lines of human origin. CCLE has also profiled 24 anticancer drugs against approximately 500 of these cell lines. There are 59 breast cancer cell lines in CCLE. CCLE has significantly helped cancer studies by annotating pre-clinical human cancer models. CCLE is maintained by the Broad institute and Novartis[13].

Figure 3 Pie chart showing different cell lines at CCLE[13].



Source: CCLE[13]

There are few other cell line databases. COSMIC (Catalogue of somatic mutations) is a database of somatic mutations. COSMIC cell line is part of the project. In cell line project all somatic mutations reported from literature survey are listed.

Simultaneously next generations sequencing methods are used to report somatic mutations. This database is developed by Wellcome Sanger Institute[20].

Another database called LINCS (Library of Integrated Network-based Cellular Signatures) exists with an aim to understand biological systems in healthy and diseased condition using networks. The outcome thereby can help in drug and biomarker development[21]. The ultimate goal is to completely understand cellular functions and their variance in normal and diseased conditions. A part of the data base called HMS LINCS deals with collection of information generated by the action of small experimental reagents on various model cell lines. As of March 2014 there were 1096 different cell lines. This database is maintained by Harvard[21].

Profiling Cancer Cell-line Sensitivities with Small Molecules (PCCSSM) is another database of cell lines developed by the Broad Institute[22]. PCCSSM maintains a record of information generated by the action small molecules like kinases and other small but potential drug compounds. The database has around 242 cell lines. PCCSSM does have DNA copy number data and mRNA gene expression data. PCCSSM is part of CTD2 (Cancer target Discovery and Development) database at Broad[22].

GDSC (Genomics of Drug Sensitivity in Cancer) is database developed by Sanger Institute[23]. The aim of this study is to predict the response of anticancer drug on the basis of molecular features of cancer. GDSC has a total of 138 compounds which are sequentially tested on 714 cell lines by using different drug concentration and recording various parameters like AUC (Area under the Curve), drug dosage and

IC_{50}[23]. GDSC also has gene expression values for cell lines determined using Affymetrix platform[23].

2.2 Thesis Statement

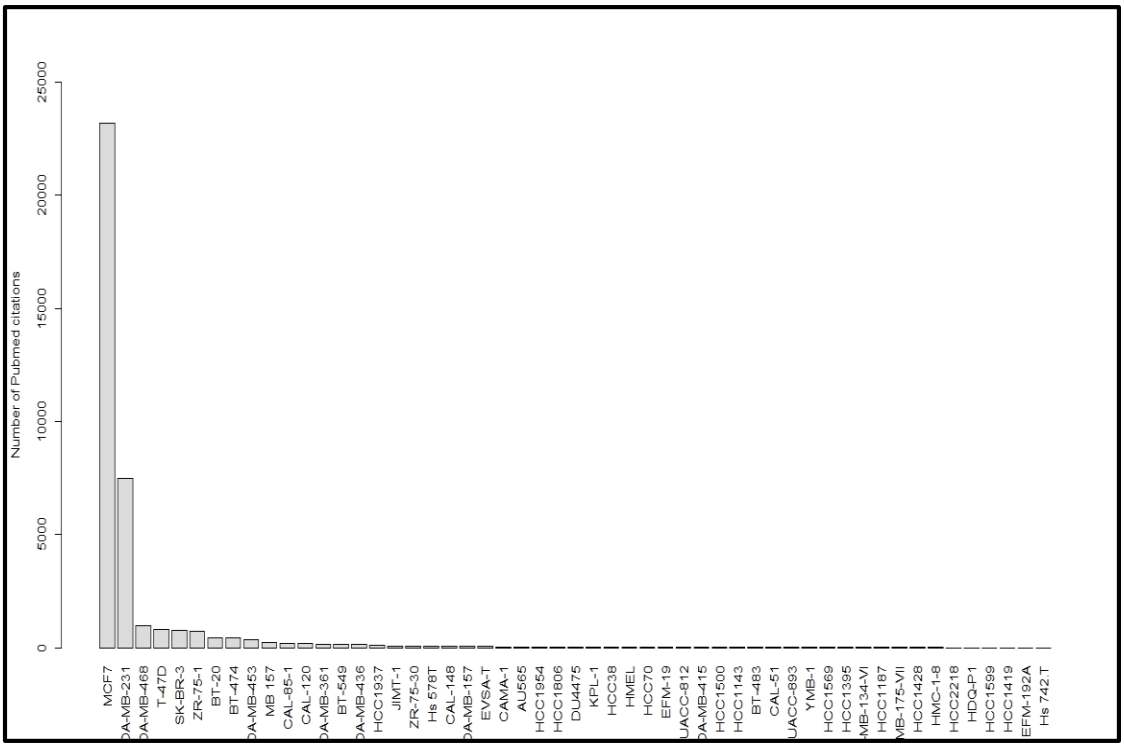
Breast cancer is heterogeneous in nature. It is very important to determine the subtype of breast cancer. As mentioned earlier that there are almost different course of treatment for luminal A, luminal B, HER2 and basal subtypes. The difference in treatment is due to the molecular features or genomic characteristics of each tumor subtype. A great deal of research is done on breast cancer and its subtype to identify changes in the molecular function and cause of these changes. The therapeutic targets revealed from these studies are being utilized for drug development.

Drug development is a very difficult process. There are almost around thousands of potential starting molecules. However very few potential molecules can make into the market[24]. One of the important step in drug development through which every molecule has to pass is determining effects of it on various models. Cell line are very important models in case of breast cancer.

It is very important to check the efficacy of potential molecule during drug development on the cell line in case of breast cancer. The molecule will perform better when the cell line shows presence of therapeutic target. Simultaneously the cell line should also have similar genomic characteristics as breast cancer subtype against which the molecule is designed. For this purpose it is necessary to compare the cell lines across tumor subtypes to determine the genetic similarity.

It is evident from PubMed citations that very few of the breast cancers have been studied in details. The most commonly studied breast cancer include MCF-7, MDA-MB-231, MDA-MB-468, SK-BR-3, T-47D and ZR-75-1 with each of them occurring in more than 500 PubMed citations.

Figure 4 PubMed citations of breast cancer cell lines



The goal of this thesis is to classify breast cancer cell line into tumor subtype based on the genomic profile comparison between tumor samples and breast cancer cell lines. There are very few breast cancer studies which compare tumor samples and cell lines with respect to genetic profile [25, 26]. This thesis compares tumor and cell lines of breast cancers using DNA copy number data from Affymetrix platform. An attempt is made to compare gene expression data (Affymetrix platform) between

tumors and cell line samples. Apart from that comparison is made between sections of genome altered in breast cancer tumors and cell lines. Mutational frequencies of important genes involved in breast cancer, obtained from mutation annotation format a product of downstream processing of massively parallel sequencing data, were compared between tumors and cell lines. All this data put together gives a fair idea about various cell lines and their genomic features along with possible subtypes of breast cancer they might more strongly associate.

This thesis has two fold objective.

1. Detect subtype information of the cell line with no prior information.

   There are few cell lines which do not have any prior subtype information. Literature survey of these cell line did not yield any information of their subtype.

2. Classify all cell lines into a subclass using genomic profile.

   Comparison of genomic features of tumor subtype with cell lines will help in classifying all cell lines into a particular tumor subtype.

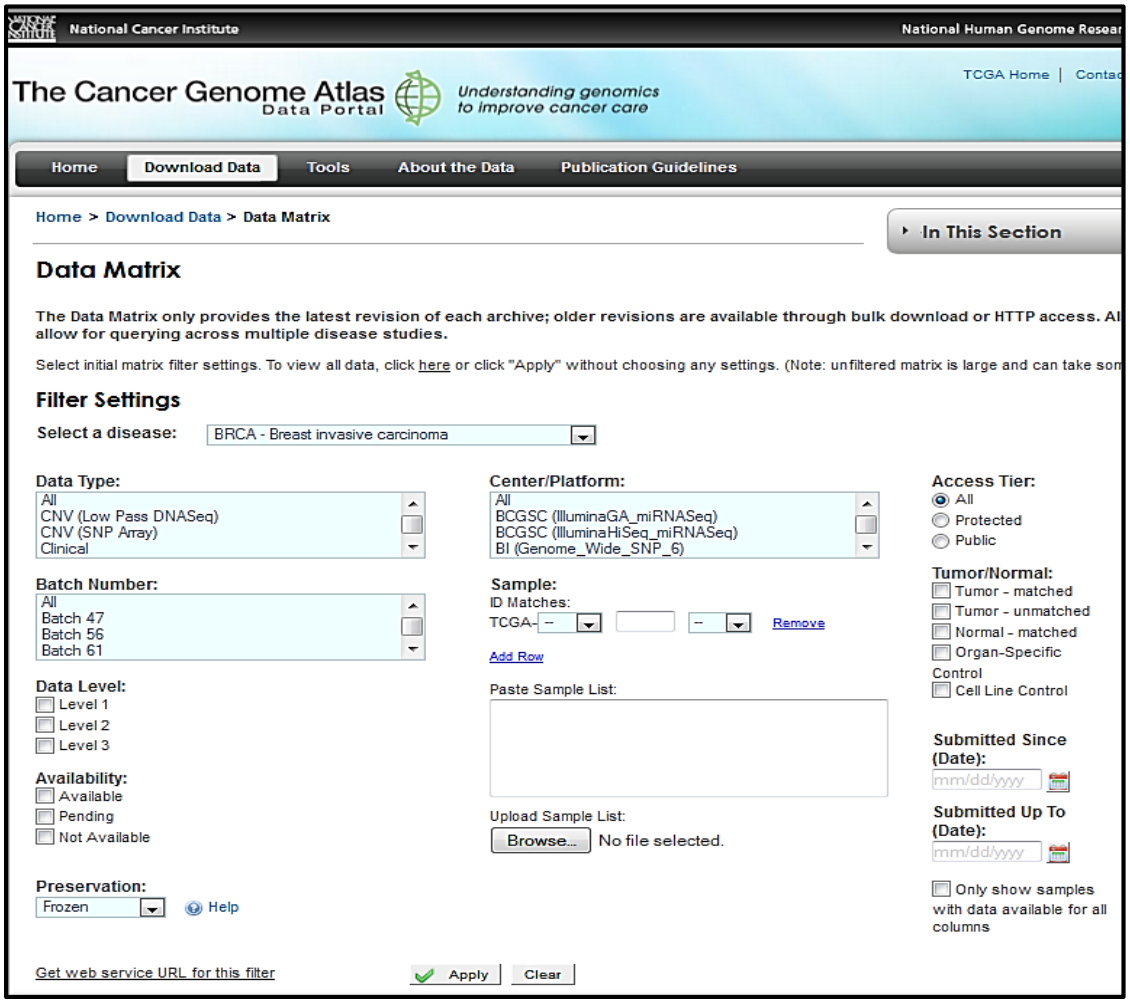
CHAPTER THREE

METHODOLOGY

3.1 Data Acquisition

All the data used in this thesis was downloaded from various data portals. These portals support unrestricted and free download of their data for research purposes. DNA copy number and mutation data for tumors was downloaded from TCGA data portal[19]. The total number of breast cancer samples present at TCGA was 1078 at the time of this thesis[19]. The DNA copy number data was available for 1049 samples. Mutation data was available for 992 samples. Not all samples from TCGA could be used in mutation related analysis. All the data downloaded from TCGA was level three data which was devoid of any patient information. All samples used in the analysis were tumor matched.

Figure 5 TCGA data portal[19].



Source: TCGA website

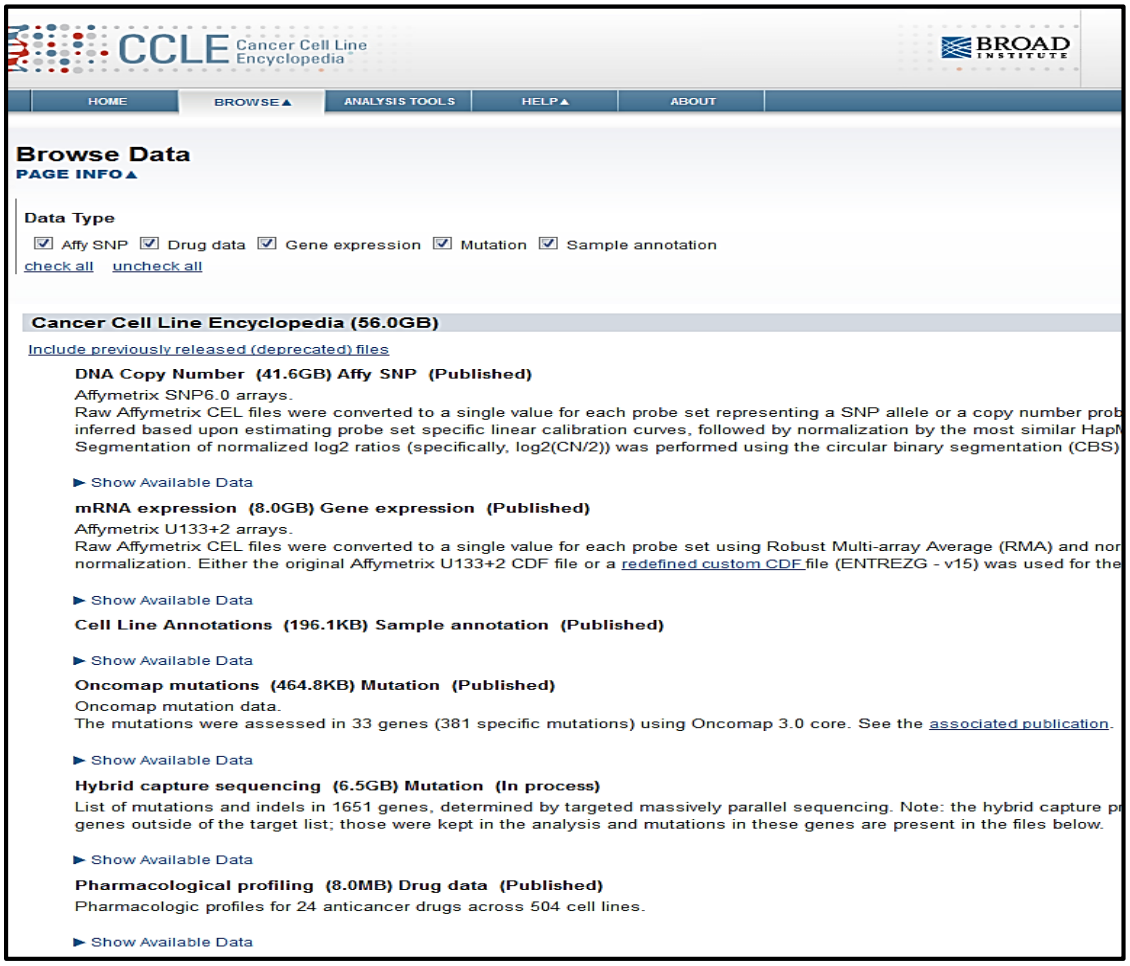
(https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter)

CCLE has an impressive collection of 947 human cancer cell lines[13]. The total number of breast cancer cell lines is 59. DNA copy number data, mutation data and mRNA expression data for the breast cancer cell lines was downloaded from CCLE data portal. DNA copy number data and mRNA expression data was available for 59

cell lines. Mutation data was available for 51 cell lines. CCLE data portal requires signing in for free download of the data[13].

The mRNA expression data for tumor samples could not be used from TCGA because of microarray platform used at TCGA was different than that present at CCLE. Because of this straightforward comparison of mRNA expression data from TCGA and CCLE was not possible. To compare the mRNA expression values of CCLE different datasets were downloaded from Gene Expression Omnibus (GEO) for tumor samples. GEO data set (GSE-41998) consisted of 279 tumor samples[27]. The data set also had the histopathology information which was used in classifying the tumor samples into subtypes. Both CCLE and GEO data sets use Affymetrix platform for mRNA expression.

Figure 6 CCLE data portal[13].



Source: CCLE website

(http://www.broadinstitute.org/ccle/data/browseData?conversationPropagation=begin)

Mutation analysis required some other essential files. 'WIG' format file which has the number of reads for each region was downloaded from CCLE data portal. In case of TCGA, bed files derived from 'WIG' files were available at for few samples at Synapse, a research sharing platform (https://www.synapse.org/#!Synapse:syn1695394).

3.2 Technology Background

DNA copy number

It is the measure of variation between genomes. One such example is difference between cancer and normal genome. The cancer genome may have certain regions of DNA which have either more than one copy or the region might be missing. In short it is a measure of structural variants present in genomes to be compared. One of the method to determine copy number variation is array based comparative genomic hybridization. In this method DNA from normal and cancer sample is obtained and fragmented, further each of the sample is labelled with different fluorescent dye. Dye labeled samples are made to hybridize with probes on array and washed to remove unbound DNA. Array is designed in specific manner such that it has probes representing specific region of the DNA. In case if the cancer sample has duplication it has more copy of DNA representing that particular region. Hence more of the cancer sample will bind to the probes and that particular region on array will glow with fluorescence of cancer labeled dye. In case the cancer sample has deletion no binding will occur and array region will emit fluorescence of normal labelled dye.

Affymetrix Genome –wide human SNP array 6.0 was used to detect the copy number variation in cancer samples at TCGA and CCLE. The result obtained from this array was processed and a downstream product called segmented file were used for further analysis.

Mutation

Mutation in cancer samples was detected using massively parallel sequencing method. In this sample DNA is broken down into very small fragments of around 100 bps. The fragments are further amplified. Then each of the fragment's sequence can be identified using variety of proprietary mechanisms. The sequence obtained is then mapped against normal reference sequence. With help of variant calling tools variants or mutations in cancer genome if any are detected. The result is processed and one of the downstream product called mutation annotation format (.maf) file which records all mutations present in the sample was used for mutation analysis.

mRNA expression data

Expression of mRNA in a sample genome can be detected using array. mRNA's are converted into complimentary DNA's (cDNA). These cDNA's are labeled with dye and are then hybridized with probes on array. Array has a collection of multiple copy probes representing various genes. The fluorescence form hybridized DNA and probes complex in measured optically and converted into numerical value depending upon the strength of fluorescence. These value in turn represent the quantity of mRNA in sample genome. Affymetrix platforms were used for mRNA expression at GEO and CCLE. GSE41998 used Affymetrix Human-Genome U133A 2.0 0 (GPL571) array with 22277 probes while CCLE used Affymetrix Human-Genome U133A Plus 2.0 (GPL570) with 54675 probes. Affymetrix GPL570 platform has all 22277 probes present on GPL 571 platform.

3.3 Data Analysis

3.3.1 Copy number data processing

3.3.1.1 FGA (Fraction Genome Altered)[28]

DNA copy number data was present in form of segmented file. Segmented files were available for TCGA tumor samples and CCLE cell line samples. Segmented files were used to calculate FGA. Fraction genome altered is a ratio. Here length of segments having value equal to or greater than a set threshold are added and are divided by the sum of length of all segments.

$$FGA = \left(\sum_{CNi>T} L(i) / \sum L(i) \right) \quad (1)[28]$$

The equation represents that sum lengths of all segments *(L(i))* whose copy number (CN) is above the set threshold (T) and divide by sum of lengths of all segments *(L(i))[*28]. In case of threshold it was set to 0.2 for tumor samples and 0.3 for CCLE cell line samples. The threshold values were based on the paper from where this concept of FGA was loaned[28]. The reason for lower threshold in case of tumor sample was due to contamination of tumor samples by other normal cells or also due to tumor heterogeneity. The cell lines were considered free of such aberration.

3.3.1.2 Correlation between TCGA and CCLE copy number data

The segmented files were converted into gene files where segments were replaced by gene/genes lying in corresponding segment region for both TCGA data and CCLE data.

This segment to gene conversion was made possible with the help of Bioconductor package called 'CNTools' [29]. CNTools helps in mapping gene back to a segment to which the gene belongs. These files with gene and copy number value were used for correlation. Clinical information file containing histopathological data of tumor samples was also available[19]. This information was used to classify tumor samples into subclass luminal A, luminal B, HER2 and TNBC. Only 628 samples could be classified using the receptor information (luminal A= 59, luminal B=419, HER2=30, basal=120). Each of the subtyped tumor class was grouped together in a file and correlated with all cell line samples. The correlation was computed in R using the function called 'cor'[30] . Thus all cell lines were correlated with four files generated from four tumor subtypes. Mean correlation was obtained for each cell line and tumor subtype and based on the highest correlation value cell line was classified into one of the four subtypes.

3.3.2 Mutational frequencies

Mutational frequencies were determined from data obtained from massively parallel sequencing technique. Mutational frequency can be described as a ratio of total number of mutations in the sample to total number of sample bases covered in sequencing. Actually it is the measure of total number of mutations detected out of total number bases sequenced. Mutations were limited to somatic mutations and functional mutations. Hence intronic, silent and other mutations were ignored and only exonic mutations were considered[28]. Mutations were reported as mutations per MB.

The number of mutations were obtained directly from .maf files for TCGA and CCLE. The number of bases for TCGA were detected from bed files. Bed file contains number of bases covered for each chromosome in form of start and end location. Subtracting end from start gives number of bases covered by the reads. All bases obtained for each sample were summed together to give us the number of bases covered for that sample. In case of CCLE, wig formats provides number of reads covering each base. So summing up all bases where there is read count will give the total number of bases.

3.3.3 Correlation between GEO and CCLE mRNA expression data

Raw files of microarray mRNA expression in form of 'CEL' files were downloaded from GEO and CCLE data portal. All the raw data was generated using Affymetrix platforms. Data to be compared between experiments had to be made comparable using methods such as back ground adjustments and nonspecific binding correction for which normalization algorithms were available. Affymetrix MAS5.0 normalization algorithm was used to normalize raw data. Clinical data was also available for GEO samples and based on receptor information GEO samples were divided into four subtypes luminal A, luminal B, HER2 and basal. There were 77 samples in luminal A group, 7 in luminal B, 16 in HER2 and 140 samples in basal group. All cell line samples were correlated with each of the four groups and correlation values were calculated in R[30]. Mean correlation value was calculated for each cell line and a group. Cell line was classified to a group with which it had highest correlation value.

3.3.4 Classifier

A classifier was developed to classify samples into luminal or basal based on the proportion of significantly mutated genes. The .maf files reported all mutated genes along with the types of mutation. For classifier TCGA samples were bifurcated into two classes. One class consisted of all luminal A and luminal B samples while the other class consisted only basal samples. The proportion of mutated genes for all samples were calculated for each class separately. The genes were sorted in decreasing order based on the difference in proportion. The genes with highest difference in proportion were deemed significant. These significant genes were further used to develop and train the classifier. Classifier was based on generalized linear model (GLM) under the assumption that there exist a linear relationship between the mutated genes. As the response variable was binary (luminal=1 and basal=0) logistic regression was used to detect the predictor variables (slope and intercept) from regression. These predicted variables were used to calculate the outcome of testing samples (In sample) as well as cell line samples (Out of sample). Majority of samples (70%) from TCGA were used to train the classifier and remaining (30%) served as testing samples. The classifier was optimized for number of variables starting from fifty variables. The optimized classifier was then used to classify the CCLE samples into luminal or basal.

3.3.5 PubMed citation analysis

The use of breast cancer cell lines in research can be best studied using PubMed citations. The number of citations can not only provide an estimate about cell lines being used in research but also suggests which cell line is most widely used.

PubMed search builder (http://www.pubmed.org) was used on January 21st 2015 to determine the number of citations for all breast cancer cell lines at CCLE. Various punctuation alternatives were used for each cell line. The number of citations cannot be considered final as there might be few false negative in the search.

Perl and Python were major scripting languages used in the analysis of data. R was used to perform statistical analysis and plotting graphs. Integrative genomic viewer (version 2.3) was used to create copy number profile plots[31].

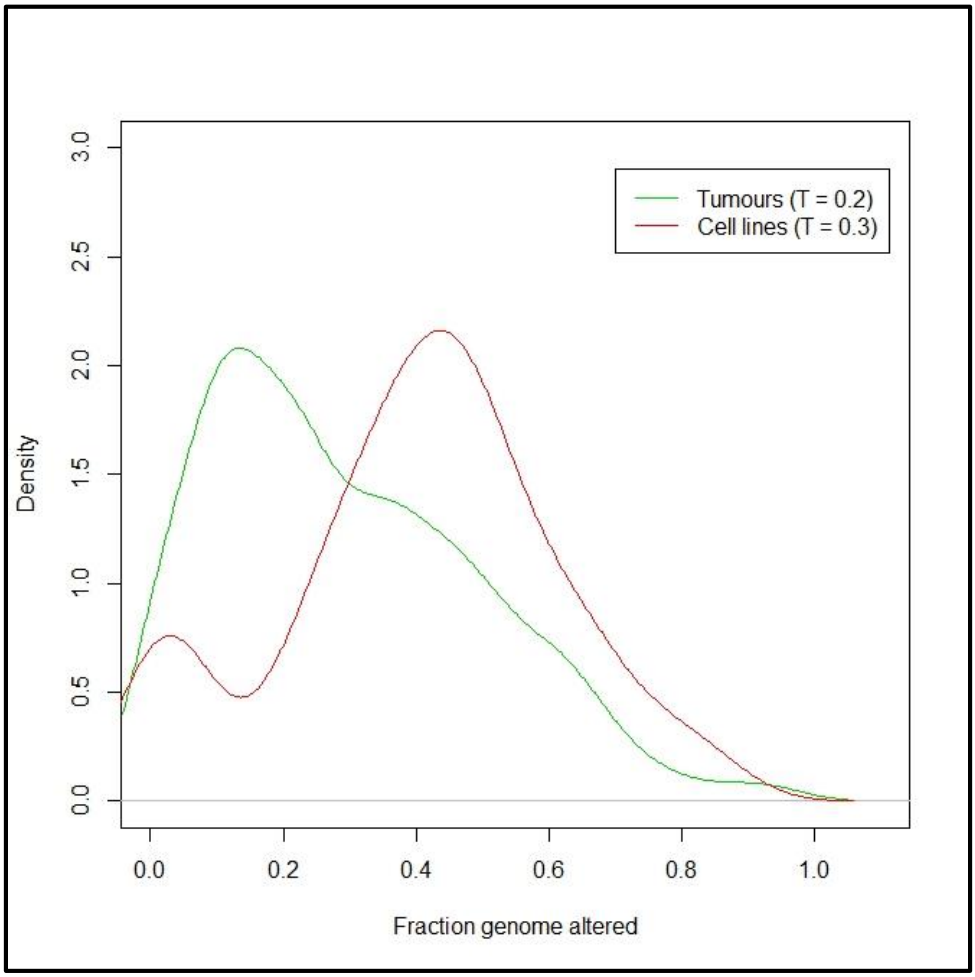
CHAPTER FOUR

RESULTS

TCGA paper found that luminal A, luminal B and HER2 samples show mutation in PIK3CA gene[19]. Luminal A was dominated with large number of samples with PIK3CA mutation along with mutation in other genes[19]. Luminal B had characteristic PIK3CA and TP53 mutation[19]. HER2 samples had characteristic PIK3CA and TP53 mutation along with HER2 gene amplification[19]. While TP53 mutation was present in Basal samples[19]

4.1 FGA

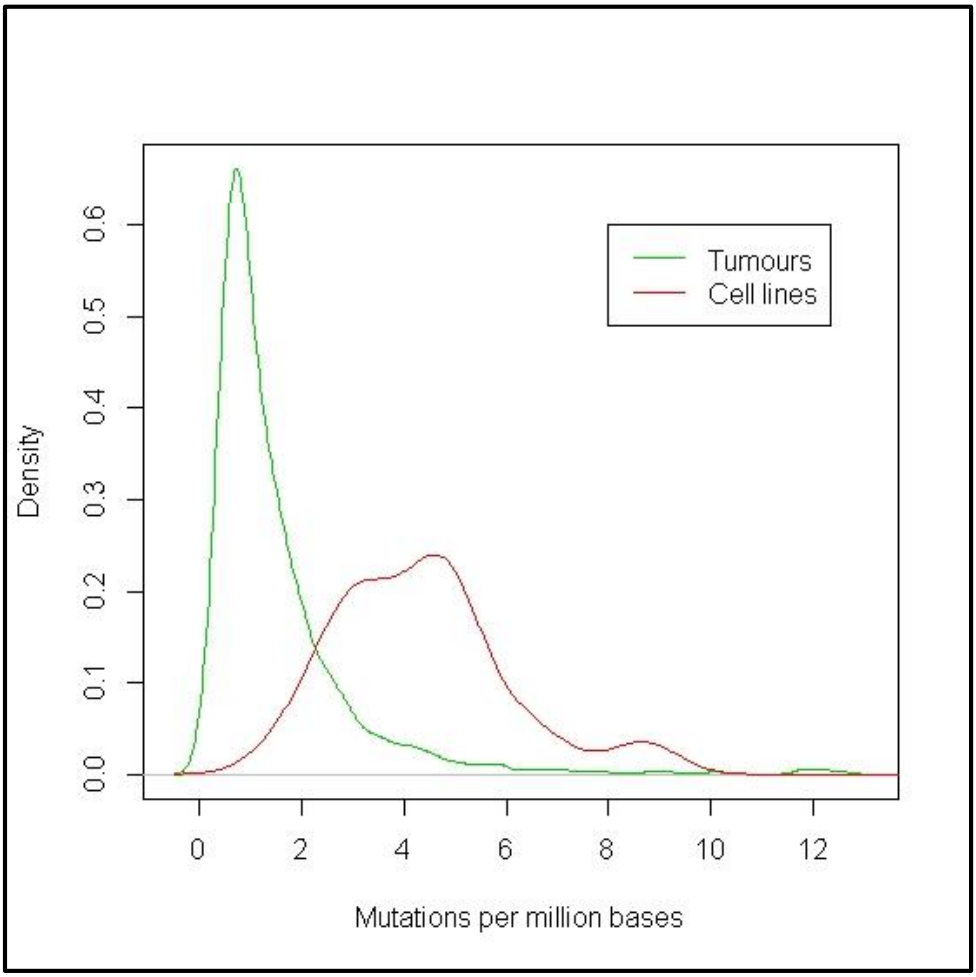
Figure 7 Fraction Genome Altered



It is evident from figure 7 that copy number alteration are significantly common in TCGA as well as CCLE. The median fraction of genome altered in TCGA is around 17% and 45% for CCLE. This indicates that cell line samples are subject to more copy number alterations. Nevertheless both datasets show wider distribution, this can be attributed to different subtypes present in TCGA as well as CCLE, as each of the subtype have different characteristic mutations and copy number profiles[28].

4.2 Mutation Frequencies

Figure 8 Mutation frequencies

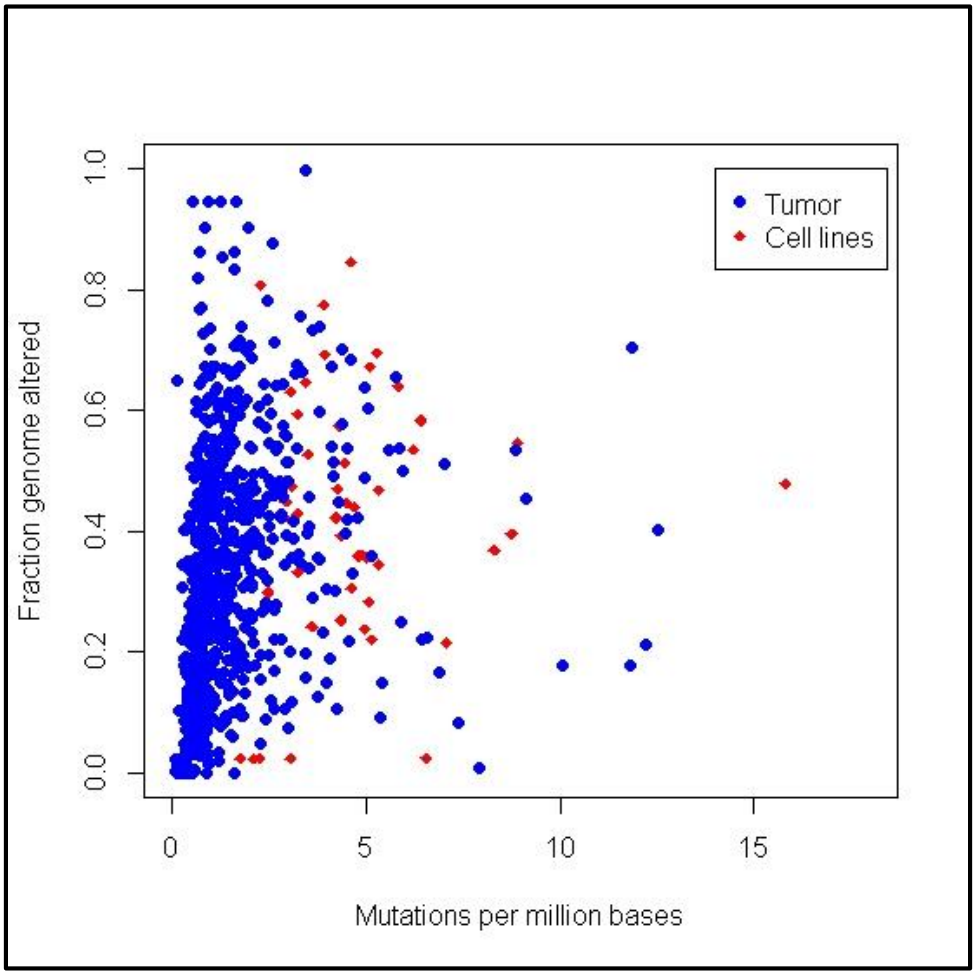


The above figure 8 illustrates the number of mutation per million bases for TCGA samples and CCLE samples. The median mutational frequency for TCGA is around 1 while for CCLE it is around 5. In case of CCLE the mutational frequency were available for only 1651 genes which the study considers important in cancer development, whereas TCGA captured all somatic mutations present in the genome. The reason for cell lines having more mutational frequencies is well defined in another study related

to ovarian cancer[28]. One of the reason cited is that cell lines being more pure capture more mutations than tumor samples which many times get contaminated with stromal cells. Another reason is cell lines may transform due to numerous passages over the period of time. Alternatively CCLE has reported all mutations including germline mutations which are excluded from TCGA tumor samples[28].

4.3 Hypermutation

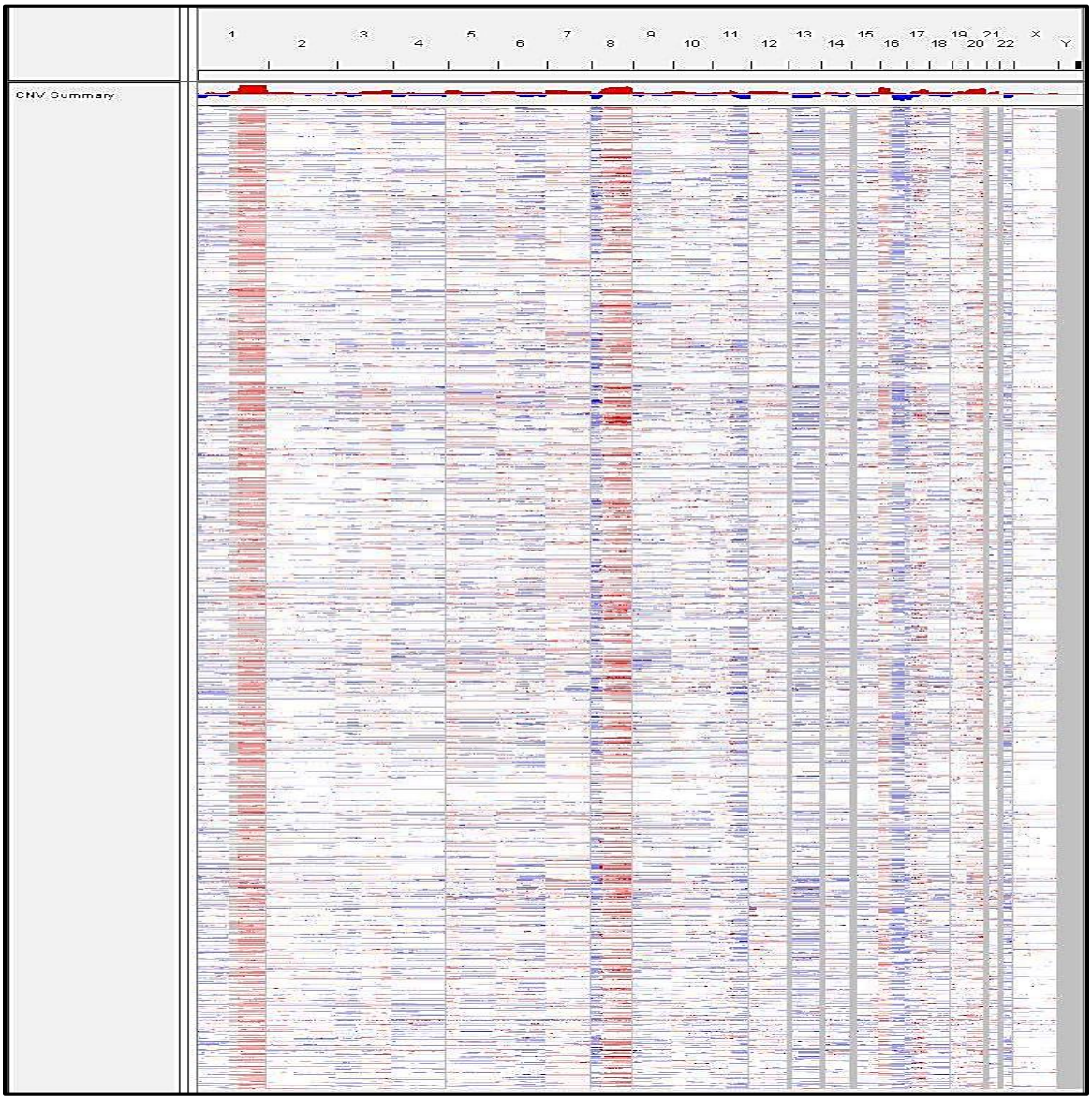
Figure 9 Comparison between Fraction Genome Altered and Mutational frequencies



From the figure 9 it can be seen that CCLE cell lines have a higher number of mutation compared to TCGA tumor samples. But both the data sets have similar number of copy number alterations. However one of the CCLE cell line (HCC1569) is reported with a very high mutational frequency, above 15 mutation per million bases. This clearly indicates that the cell line shows hypermutation.

4.4 Copy Number Profiles

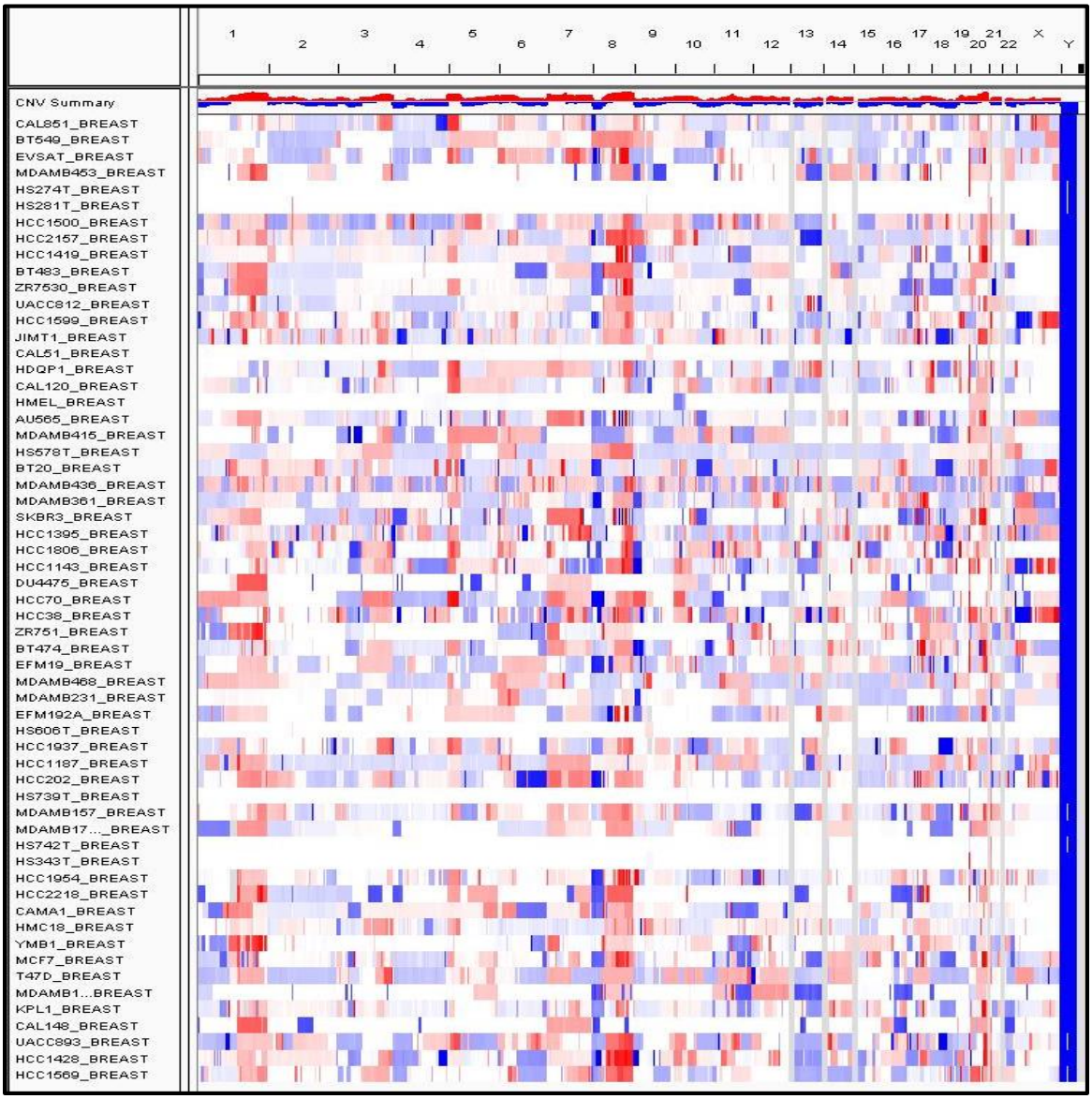
Figure 10 Copy number profiles for TCGA tumor samples



This figure provides copy number profiles for all tumor samples in TCGA (1049 samples).The copy numbers are shown for each chromosome. The red color indicates higher copy number value while the blue color indicates lower copy number value.

IGV helps to visualize all samples in a single frame and also supports zooming in to a particular chromosome or even a gene or a certain region of the chromosome[31].

Figure 11 Copy number profiles for CCLE cell line samples



This figure provides copy number profiles for all cell line samples in CCLE (59 samples).The copy numbers are shown for each chromosome. The red color indicates higher copy number value while the blue color indicates lower copy number value.

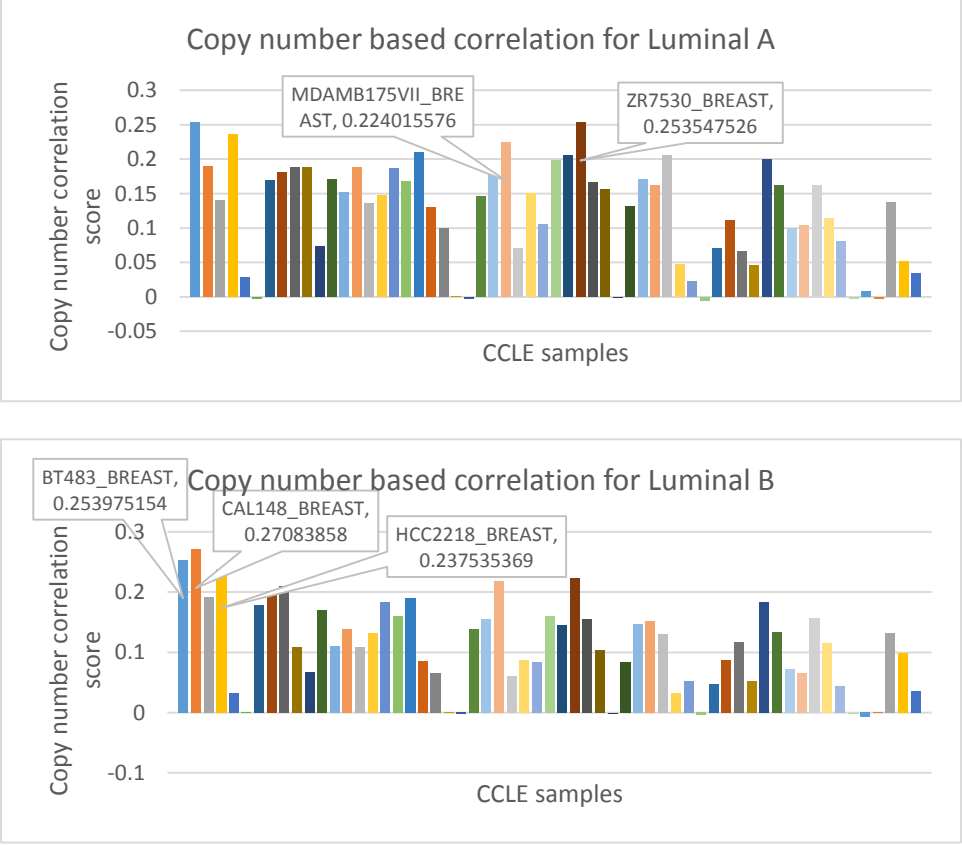
IGV helps to visualize all samples in a single frame and also supports zooming in to a particular chromosome or even a gene or a certain region of the chromosome[31].

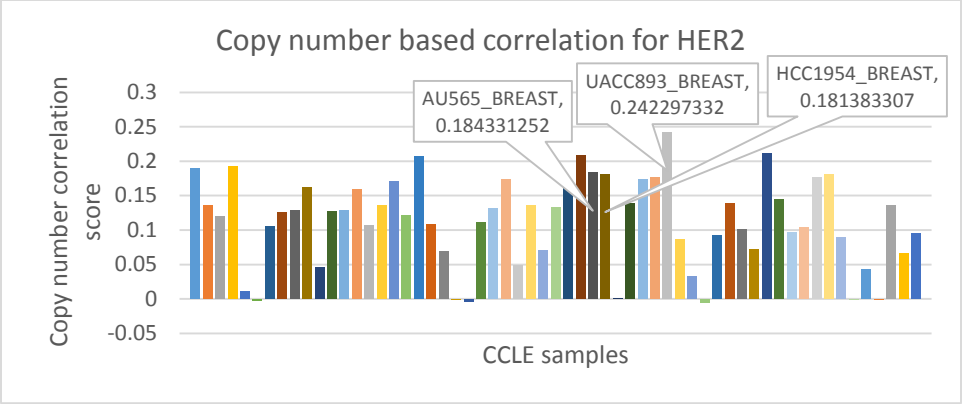
4.5 Correlation Score

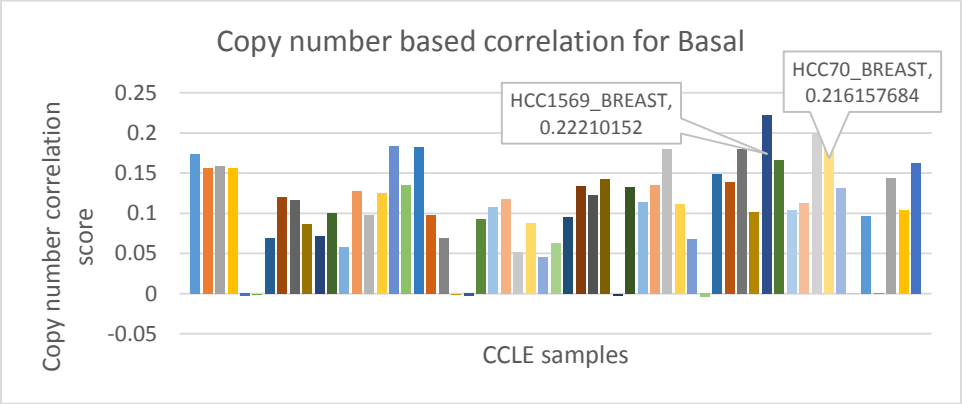
4.5.1 Correlation score for copy number profiles

Based on correlation scores obtained using Pearson correlation with TCGA tumor subtypes, each of the CCLE cell line was classified into a particular subtype. Out of 59 CCLE samples based on the correlation score 23 samples were classified as Luminal A, 9 as Luminal B, 7 as HER2 and 20 samples were classified as basal.

Figure 12 Copy number based correlation







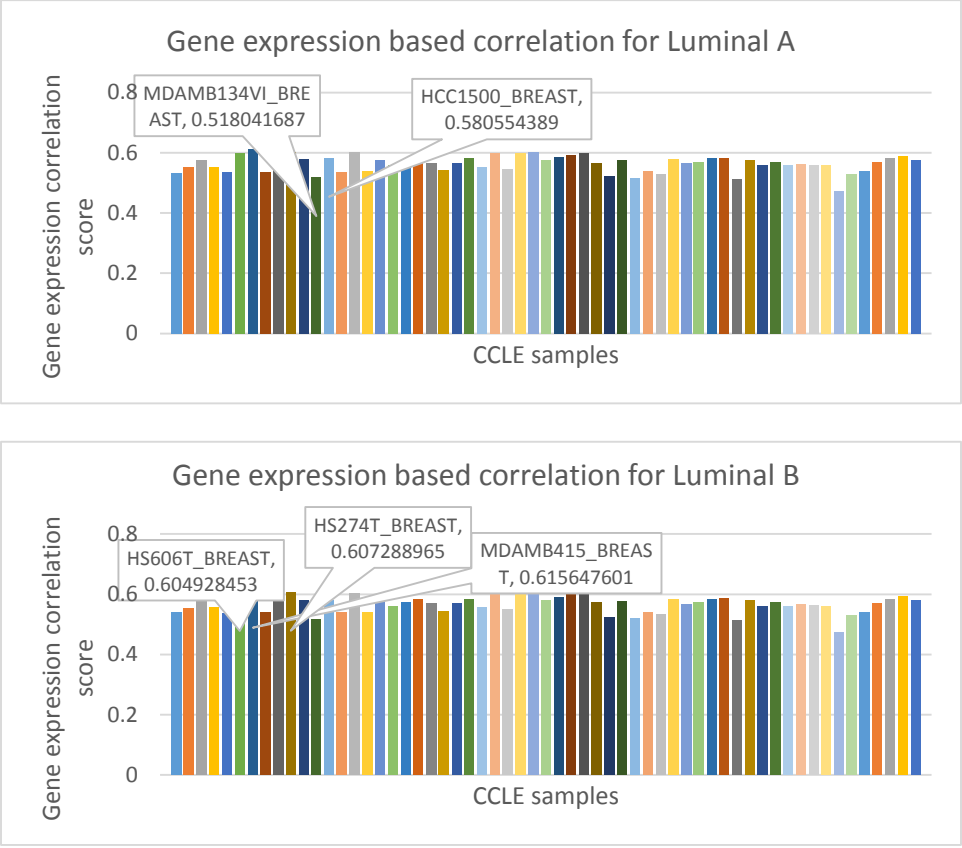
Above figures have 59 bar plots representing each of the 59 CCLE cell lines. For each subtype plot, few cell lines have been called out which have been classified into that particular subtype and have the highest correlation score.

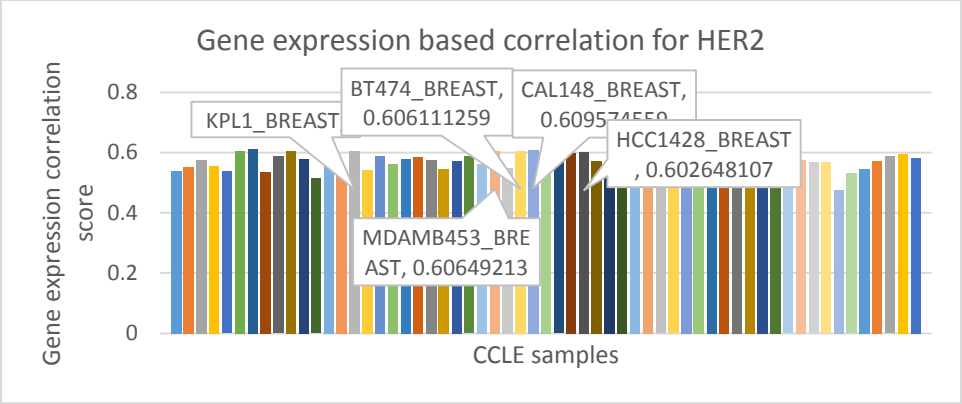
4.5.2 Correlation score for gene expression profiles

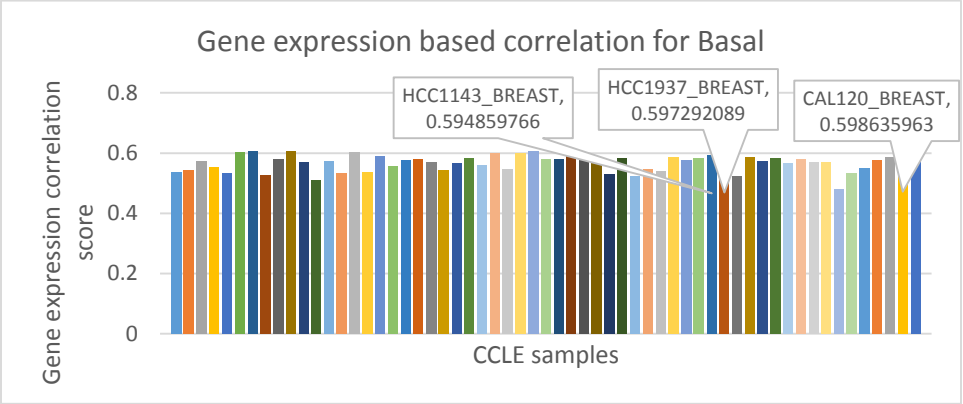
Correlation score was calculated for each CCLE cell line and TCGA tumor subtype based on Pearson correlation. Out of 59 CCLE samples based on gene expression correlation score 2 samples were classified as Luminal A, 11 as Luminal B, 21 as HER2 and 25 samples were classified as basal.

Figures below have 59 bar plots representing each of the 59 CCLE cell lines. For each subtype plot, few cell lines have been called out which have been classified into that particular subtype and have the highest correlation score.

Figure 13 Gene expression based correlation







4.6 Classifier

The GLM (logistic regression) based classifier was trained and tested on TCGA (in sample) data. It was found that the classifier had better accuracy when it used 50 variables as genes (significant genes based on proportion). These 50 variables were feed to step function, which uses stepwise AIC algorithm to generate a model. Sensitivity, specificity and AUC was determined for both training and test datasets. The classifier's training efficiency for predicting luminal samples based on 5 random trials (selection of samples) yielded sensitivity ranging from 81% to 86%, while specificity ranged from 79% to 90% and AUC ranged from 91% to 93%. In case of 5 testing trials sensitivity ranged from 78% to 89% while specificity ranged from 60%

to 90% and AUC ranged from 80% to 85%. The number of genes used as variables varied from 16 to 23.

However CCLE has mutation data for 1651 genes. And only 11 significant genes overlapped between TCGA and CCLE datasets. The model based on 11 genes as variables yielded following results. The classifier's training efficiency for predicting luminal samples yielded sensitivity of 82%, while specificity of 74%. And AUC obtained was 85%. In case of testing the sensitivity was 79% while specificity was 81% and AUC obtained was 87%.

4.7 Cell line Classification

Table 3 Classification of CCLE cell lines into four tumor subtypes

Cell line name	**PubMed Citations**	**Gene Expression based Classification**	**Copy number based Classification**	**Classifier based Classification**	**Classification from lit. survey**
AU565	41	HER2	HER2	Basal	HER2amp [32]
BT20	430	Basal	Basal	Luminal	Basal [32]
BT474	428	HER2	Luminal A	Luminal	HER2amp [32]

BT483	12	Luminal B	Luminal B	Luminal	Luminal
BT549	148	Basal	Basal	Luminal	Basal[32]
CAL120	181	Basal	Luminal A	Basal	Luminal B[33]
CAL148	57	HER2	Luminal B	Luminal	Luminal A[33]
CAL51	9	Basal	Basal	Luminal	
CAL851	183	Basal	Basal	Luminal	Basal[33]
CAMA1	41	HER2	Luminal A	Luminal	Luminal A
DU4475	28	Basal	Luminal B	Luminal	Basal
EFM19	18	HER2	Luminal A	Luminal	Luminal[34]
EFM192A	1	HER2	Luminal A	Luminal	Luminal B[34]
EVSAT	48	HER2	Basal	Luminal	
HCC1143	12	Basal	Luminal A	Basal	Basal[25, 35]
HCC1187	5	Basal	Basal	Luminal	Basal[32]
HCC1395	6	Basal	Basal	Luminal	Basal[32]
HCC1419	1	HER2	Luminal A	Luminal	HER2amp[32]
HCC1428	4	HER2	Luminal A	Luminal	Luminal[32]
HCC1500	13	Luminal A	Luminal A	Luminal	Luminal[32]

HCC1569	6	Basal	Basal	Luminal	HER2amp[32]
HCC1599	2	Basal	Basal	Luminal	Basal[32]
HCC1806	29	Basal	Basal	Luminal	Basal[32]
HCC1937	121	Basal	Basal	Basal	Basal[32]
HCC1954	40	HER2	HER2	NA	HER2amp[32]
HCC202	0	HER2	Luminal A	Luminal	HER2amp[32]
HCC2157	0	Basal	Basal	Luminal	Basal[32]
HCC2218	3	HER2	Luminal B	Luminal	HER2amp[32]
HCC38	23	Basal	Luminal A	Luminal	Basal[32]
HCC70	18	Basal	Basal	Luminal	Basal[32]
HDQP1	2	Basal	Basal	NA	
HMC18	4	Basal	Luminal A	Luminal	
HMEL	22	Basal	Luminal B	NA	
HS274T	0	Luminal B	Basal	NA	
HS281T	0	HER2	Luminal A	Luminal	
HS343T	0	HER2	HER2	Luminal	Basal[32]
HS578T	379	Basal	Basal	Luminal	Basal[25, 35]

HS606T	0	Luminal B	Basal	NA	
HS739T	0	Luminal B	Luminal B	Basal	
HS742T	1	Luminal B	Luminal A	Luminal	
JIMT1	62	Basal	HER2	Luminal	HER2amp[36]
KPL1	26	HER2	Luminal A	Luminal	Luminal[34]
MCF7	23201	HER2	Luminal A	Luminal	Luminal[32]
MDAMB134VI	5	Luminal A	Luminal B	Luminal	Luminal[32]
MDAMB157	50	Basal	Basal	NA	Basal[32]
MDAMB175VII	5	Luminal B	Luminal A	NA	Luminal[32]
MDAMB231	7507	Basal	Luminal A	Basal	Basal[32]
MDAMB361	155	HER2	Luminal A	Luminal	HER2amp[32]
MDAMB415	13	Luminal B	Luminal A	Luminal	Luminal[32]
MDAMB436	131	Basal	Basal	Luminal	Basal[32]
MDAMB453	350	HER2	HER2	Luminal	Basal[32]
MDAMB468	956	Basal	Basal	Luminal	Basal[32]
SKBR3	760	HER2	HER2	Basal	HER2amp[32]
T47D	795	Luminal B	Luminal A	Luminal	Luminal[32]

UACC812	15	Luminal B	Luminal A	Luminal	HER2amp[32]
UACC893	9	HER2	HER2	Luminal	HER2amp[32]
YMB1	8	HER2	Luminal B	NA	
ZR751	720	Luminal B	Luminal B	Luminal	Luminal[32]
ZR7530	59	Luminal B	Luminal A	Luminal	HER2amp[32]

FGA and mutational frequencies reveal information about TCGA and CCLE samples, but they alone are insufficient for classification of CCLE samples. The cell lines can be classified with more information obtained from correlation between copy number and gene expression datasets. These dataset comparisons in case of copy number account for amplification and deletion consistencies while gene expression account for characteristic expressions.

Based on correlation information cell lines are classified into one of the four breast cancer subtypes. The Classifier trained on proportion of mutation in significant genes further aids to confirm CCLE cell line classification.

The above table has classification for 59 CCLE cell lines based on correlation values obtained from copy number and gene expression datasets and this classification might be correctly identified and confirmed by classifier's outcome.

Cell line BT 483 is classified as luminal based on literature survey is confirmed as luminal B subtype by the correlation analysis. The classifier also identifies it to be luminal cell line. HCC1937 which is classified as basal cell line in literature is confirmed as basal by both the correlation results as well as by classifier's outcome. Most of the times correlation analysis correctly classifies cell lines into subtype found in literature. However the classifier's result are not always correct. This can be attributed to an error in classifier and therefore has a scope for improvement.

There are few cell lines where there is none or very less information available in literature. Further their subtypes are unknown and also histopathological information is unavailable. Few such cell lines were classified using correlation analysis. One such example is HDQP1 which was classified as Basal from correlation analysis. HS739T and HS742T were classified as luminal by correlation analysis. Enlightening facts about cell lines with no prior information was one of the important objective of this thesis. And on the basis of correlation few of them have been classified into a subtype. Thus unclassified cell lines were successfully classified in this study.

CHAPTER FIVE

DISCUSSION AND LIMTATIONS

5.1 Discussion

The goal of this thesis was to classify CCLE cell lines into one of the four breast cancer subtypes (luminal A, luminal B, HER2 and basal). This study has revealed valuable classification information about some cell lines which lacked earlier in depth studies. A careful study of tumor samples genomic parameters like copy number, gene expression and gene mutation data and comparison of it with cell line datasets has made possible cell lines classification based on a statistical measure known as correlation.

There has always been an ambiguity about subtype classification of breast cancer cell lines even if the number of cell lines available for breast cancer cell line is low. It is below 100. Further based on the literature survey very few of these cell lines are widely used. This gives rise the need of classifying cell lines which will help in providing new cell line models for various tumor subtype researches. This study tries to address these needs by classifying the cell lines.

Any cell line's use as model depends upon the question being asked. For example does cell line possess same gene alterations as the tumor or does cell line share same/similar type of mutations in gene/genes? Another question would be whether they share same genomic segment variation as in amplification or deletion? Also are gene expression pattern similar? These questions are very important during drug development as most studies target specific molecular mechanisms which should be

certainly present in model cell lines so as to confirm the pharmaceutical action of drugs. This study address similar problems by developing classification technique which uses all available genomic profiles to classify cell lines. Classification is done for each profile independently (copy number, gene expression and mutation). This ultimately provides measures of similarity of each cell line.

Cell line once subtyped and annotated to particular class will greatly help in development of personalized medicine. Annotated cell line matched to an individual patients will definitely help in selecting an established treatment regime. As there are various databases including CCLE which test various anticancer drugs on all available cell lines and record the activity. These studies are not limited to anticancer drugs. In fact all possible agents right from small molecules inhibitors, kinases to custom compounds are being tested as potential drugs.

5.2 Limitations

This thesis represents an effort for classification of cell lines. Concepts used in this thesis for classification are distinctive but they are not all inclusive. One of the main limitation is classifier, it is not completely accurate and is innate as it produces lot of incorrect results. One reason can be inadequate number of variables used in its creation. The error can also be attributed to CCLE data as only 1651 genes are available whereas more than 15000 genes are available at TCGA.

All the copy number analysis make use of segmented data. This segmented data is obtained from Affymertix SNP 6.0 platform. Even though the platform used at both the sources (TCGA and CCLE) are same there exist some artifacts in the data. Firstly

there exists some mechanical error occurred while conducting the experiments, many studies have also shown that even environmental conditions greatly affect the experiment outcome. However raw data obtained follow the same preprocessing and downstream processing at both sources which heavily normalize data to make it comparable across various other studies. Nevertheless these data normalization effort may not be sufficient enough to adjust for all errors and variations either internal or due to external factors. There should be additional steps involved where data for comparison is actually integrated and normalized together by using techniques like quantile normalization which can actually adjust for most of the variations and make the data more uniform.

Mutational data consist of all mutations present in all genes of an individual. This data is obtained using next generation sequencing technology. TCGA has both matched tumor data and normal data which results in identification of both somatic and germline mutations by comparing tumor and normal data. However segregation of observed mutations in to somatic and germline mutations is not possible for cell lines. As cell lines are often generated from tumor samples besides many of these were established decades ago at a time where normal tissue was not recorded along with the tumor tissue. Therefore it not possible to match tumor cell line with its normal tissue. Hence tumor mutation data consist of both germline as well as somatic mutation. Often this mutational data discrepancy yields inconclusive results between TCGA and CCLE samples.

However there are databases like 1000 genomes and HapMap which have catalogued generally occurring variants in various human races. These variants can be used as baseline against both the tumor and cell line data which ultimately take care of commonly occurring and germline mutations. This is based on the premise that if we are aware of the race of an individual from whom the cell line was obtained then common variants in that particular race might cover most of the germline mutations. Nonetheless this premise needs to be tried, tested and further verified.

Gene expression comparison between samples from different platform is not advisable. TCGA used Illumina platform for gene expression studies while CCLE used Affymetrix platform. In order to counter this problem a GSE study of breast cancer tumors was selected which was done on Affymetrix platform. Raw data was obtained from both sources. They were subjected to same pipeline and normalization techniques to make data comparable. However even for the same platform there are some chip artifacts which may affect the results. Also, as mentioned above human and environmental factors also affect the results to an extent. Method such as quantile normalization can adjust such artifacts by transforming original values to rank based values and thereby making them statistically comparable.

Correlation is measure of relationship between two variables. Correlation is a good method for determining similarity between cell lines and tumor subtypes. When gene expressions were used as variables correlation method worked fine. But in case of copy number data correlation observed was very low. A better method such as

regression may be used to correctly identify the relationship which may result in better classification.

The classification obtained using gene expression and copy number data was in agreement for most of the cell lines with their literature classifications. However in order to make an accurate classification a more advanced model system can be developed which can take in all the variables resulting in more congruent and better classification. This model can integrate gene expression values, copy number values, FGA and mutation data and assign a weight to each variable according to its significance which will lead into better prognosis.

The result obtained from correlation analysis might not be accurate as small number of cell lines were available. In order to improve the result it is necessary to have large number of cell lines along with multiple replicates.

Lot of issues exist with available data for tumors and cell lines. Mutational data is unavailable for all samples as some bed files of TCGA samples are missing. Also all samples at TCGA do not have genetic information for both genomic profiles of DNA copy number and mutation. Further platform used for gene expression at TCGA is different than CCLE which makes gene expression comparison difficult.

This study is limited to cell lines available at CCLE however there are other databases which have additional breast cancer cell lines. It can also be stated that this study is limited to only one type of cancer (breast cancer). However similar concepts can be applied to other form of cancers.

Some of the limitations can be addressed with additional data. Nevertheless this study can definitely be a base for future work.

CHAPTER SIX

REFERENCES

1. Society, A.C., *Cancer facts & figures*. 2008: The Society.
2. Reeder, J.G. and V.G. Vogel, *Breast cancer risk management.* Clinical breast cancer, 2007. **7**(11): p. 833-840.
3. Pasche, B., *Cancer genetics. Introduction.* Cancer Treat Res, 2010. **155**: p. xi-xii.
4. Cancer, I.A.f.R.o., *GLOBOCAN 2012: estimated cancer incidence, mortality and prevalence worldwide in 2012.* World Health Organization. http://globocan. iarc. fr/Pages/fact_sheets_cancer. aspx. Accessed on, 2014. **9**.
5. Saini, K., et al., *Role of the multidisciplinary team in breast cancer management: results from a large international survey involving 39 countries.* Annals of oncology, 2012. **23**(4): p. 853-859.
6. Scherpereel, A. and S. Dehette, *Nouvelle classification internationale TNM (7 e édition) du cancer pulmonaire: explications et implications pratiques.* Revue des Maladies Respiratoires, 2008. **25**: p. 66-68.
7. Sandhu, R., et al., *Microarray-based gene expression profiling for molecular classification of breast cancer and identification of new targets for therapy.* Lab Medicine, 2010. **41**(6): p. 364-372.
8. Romond, E.H., et al., *Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer.* New England Journal of Medicine, 2005. **353**(16): p. 1673-1684.
9. Sarntivijai, S., A. Ade, and B. Athey. *The Cell Line Ontology and its use in tagging cell line names in biomedical text*. in *AMIA... Annual Symposium proceedings/AMIA Symposium. AMIA Symposium*. 2006.
10. Allen, D.D., et al., *Cell lines as in vitro models for drug screening and toxicity studies.* Drug development and industrial pharmacy, 2005. **31**(8): p. 757-768.
11. Holliday, D.L. and V. Speirs, *Choosing the right cell line for breast cancer research.* Breast Cancer Res, 2011. **13**(4): p. 215.
12. Lasfargues, E.Y. and L. Ozzello, *Cultivation of Human Breast Carcinomas2.* 1958.
13. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-607.
14. Sinn, P., et al., *Multigene assays for classification, prognosis, and prediction in breast cancer: a critical review on the background and clinical utility.* Geburtshilfe und Frauenheilkunde, 2013. **73**(9): p. 932.
15. Reis, P.P., et al., *mRNA transcript quantification in archival samples using multiplexed, color-coded probes.* BMC biotechnology, 2011. **11**(1): p. 46.
16. Nielsen, T., et al., *Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens.* BMC cancer, 2014. **14**(1): p. 177.
17. Ding, L., et al., *Genome remodelling in a basal-like breast cancer metastasis and xenograft.* Nature, 2010. **464**(7291): p. 999-1005.

18. Shah, S.P., et al., *The clonal and mutational evolution spectrum of primary triple-negative breast cancers.* Nature, 2012. **486**(7403): p. 395-399.
19. Network, C.G.A., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
20. Bamford, S., et al., *The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website.* British journal of cancer, 2004. **91**(2): p. 355-358.
21. Sansone, S.-A., et al., *Toward interoperable bioscience data.* Nature genetics, 2012. **44**(2): p. 121-126.
22. Basu, A., et al., *An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules.* Cell, 2013. **154**(5): p. 1151-1161.
23. Yang, W., et al., *Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells.* Nucleic acids research, 2013. **41**(D1): p. D955-D961.
24. Stratman, H., *Bad Medicine: When Medical Research Goes Wrong.* Analog Science Fiction and Fact, September, 2010. **20**.
25. Kao, J., et al., *Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery.* PloS one, 2009. **4**(7): p. e6146.
26. Neve, R.M., et al., *A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes.* Cancer cell, 2006. **10**(6): p. 515-527.
27. Horak, C.E., et al., *Biomarker analysis of neoadjuvant doxorubicin/cyclophosphamide followed by ixabepilone or Paclitaxel in early-stage breast cancer.* Clinical Cancer Research, 2013. **19**(6): p. 1587-1595.
28. Domcke, S., et al., *Evaluating cell lines as tumour models by comparison of genomic profiles.* Nature communications, 2013. **4**.
29. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics.* Genome biology, 2004. **5**(10): p. R80.
30. Statistical Package, R., *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing, 2009.
31. Robinson, J.T., et al., *Integrative genomics viewer.* Nature biotechnology, 2011. **29**(1): p. 24-26.
32. Niepel, M., et al., *Analysis of growth factor signaling in genetically diverse breast cancer lines.* BMC biology, 2014. **12**(1): p. 20.
33. Lehmann, B.D., et al., *Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies.* The Journal of clinical investigation, 2011. **121**(7): p. 2750.
34. Finn, R.S., et al., *PD 0332991, a selective cyclin D kinase 4/6 inhibitor, preferentially inhibits proliferation of luminal estrogen receptor-positive human breast cancer cell lines in vitro.* Breast Cancer Res, 2009. **11**(5): p. R77.
35. Chavez, K.J., S.V. Garimella, and S. Lipkowitz, *Triple negative breast cancer cell lines: one tool in the search for better treatment of triple negative breast cancer.* Breast disease, 2010. **32**(1): p. 35-48.
36. Tanner, M., et al., *Characterization of a novel cell line established from a patient with Herceptin-resistant breast cancer.* Molecular cancer therapeutics, 2004. **3**(12): p. 1585-1592.

CHAPTER SEVEN

APPENDIX

Table 4 List of breast cancer cell lines from Cancer Cell Line encyclopedia[13].

Serial No.	Name of cell line
1	AU565
2	BT20
3	BT474
4	BT483
5	BT549
6	CAL120
7	CAL148
8	CAL51
9	CAL851
10	CAMA1
11	DU4475
12	EFM192A
13	EFM19
14	EVSAT
15	HCC1143
16	HCC1187

17	HCC1395
18	HCC1419
19	HCC1428
20	HCC1500
21	HCC1569
22	HCC1599
23	HCC1806
24	HCC1937
25	HCC1954
26	HCC202
27	HCC2157
28	HCC2218
29	HCC38
30	HCC70
31	HDQP1
32	HMC18
33	HMEL
34	HS274T
35	HS281T
36	HS343T
37	HS578T
38	HS606T

39	HS739T
40	HS742T
41	JIMT1
42	KPL1
43	MCF7
44	MDAMB134VI
45	MDAMB157
46	MDAMB175VII
47	MDAMB231
48	MDAMB361
49	MDAMB415
50	MDAMB436
51	MDAMB453
52	MDAMB468
53	SKBR3
54	T47D
55	UACC812
56	UACC893
57	YMB1
58	ZR751
59	ZR7530

Table 5 Copy number based correlation score and classification

Cell lines Names	Luminal A	Luminal B	HER2	Basal	Classification	Max. Score
AU565	0.166586	0.154257	0.184331	0.122502	HER2	0.184331
BT20	0.046715	0.031671	0.086495	0.110932	Basal	0.110932
BT474	0.188798	0.109514	0.162058	0.086534	Luminal A	0.188798
BT483	0.253735	0.253975	0.189633	0.174008	Luminal B	0.253975
BT549	0.022708	0.051819	0.031989	0.067934	Basal	0.067934
CAL120	0.072618	0.066888	0.046678	0.071756	Luminal A	0.072618
CAL148	0.189672	0.270839	0.135863	0.1557	Luminal B	0.270839
CAL51	-0.00547	-0.00421	-0.00496	-0.00409	Basal	-0.00409
CAL851	0.070278	0.047388	0.092087	0.148079	Basal	0.148079
CAMA1	0.170342	0.170115	0.126597	0.100618	Luminal A	0.170342
DU4475	0.140713	0.191345	0.11956	0.158038	Luminal B	0.191345
EFM19	0.187921	0.138244	0.158837	0.127828	Luminal A	0.187921
EFM192A	0.151507	0.110575	0.128764	0.058168	Luminal A	0.151507
EVSAT	0.110345	0.087168	0.138484	0.138799	Basal	0.138799
HCC1143	0.136037	0.109538	0.106757	0.097419	Luminal A	0.136037
HCC1187	0.065842	0.116121	0.100467	0.180329	Basal	0.180329
HCC1395	0.046239	0.051194	0.072145	0.101262	Basal	0.101262
HCC1419	0.147093	0.131807	0.135331	0.125562	Luminal A	0.147093
HCC1428	0.186741	0.183406	0.17101	0.183609	Luminal A	0.186741
HCC1500	0.167727	0.159965	0.121797	0.134641	Luminal A	0.167727
HCC1569	0.20042	0.183807	0.211373	0.222102	Basal	0.222102
HCC1599	0.161995	0.133576	0.144028	0.165898	Basal	0.165898
HCC1806	0.098709	0.071424	0.096514	0.104047	Basal	0.104047

HCC1937	0.103335	0.066144	0.103311	0.113123	Basal	0.113123
HCC1954	0.1567	0.103305	0.181383	0.142198	HER2	0.181383
HCC202	0.209316	0.190248	0.206212	0.18304	Luminal A	0.209316
HCC2157	0.161396	0.157346	0.176477	0.199652	Basal	0.199652
HCC2218	0.236173	0.237535	0.192882	0.156001	Luminal B	0.237535
HCC38	0.129315	0.08599	0.108364	0.097608	Luminal A	0.129315
HCC70	0.114064	0.114522	0.181384	0.216158	Basal	0.216158
HDQP1	0.081337	0.044456	0.088894	0.131949	Basal	0.131949
HMC18	0.098861	0.065408	0.069957	0.069847	Luminal A	0.098861
HMEL	0.028209	0.032638	0.01176	-0.00318	Luminal B	0.032638
HS274T	-0.00293	-0.00184	-0.00044	0.001242	Basal	0.001242
HS281T	0.000212	-0.00111	-5.3E-05	-0.00136	Luminal A	0.000212
HS343T	-0.00117	-0.00171	0.000114	-0.00243	HER2	0.000114
HS578T	0.007327	-0.00725	0.043033	0.095946	Basal	0.095946
HS606T	-0.00253	-0.00116	-0.00138	-0.00058	Basal	-0.00058
HS739T	-0.0026	-0.0002	-0.00242	-0.00074	Luminal B	-0.0002
HS742T	-0.00213	-0.00226	-0.00384	-0.00331	Luminal A	-0.00213
JIMT1	0.131891	0.083575	0.139405	0.132754	HER2	0.139405
KPL1	0.145989	0.139242	0.111829	0.092862	Luminal A	0.145989
MCF7	0.177755	0.155327	0.131767	0.107978	Luminal A	0.177755
MDAMB134VI	0.16867	0.177989	0.106024	0.068999	Luminal B	0.177989
MDAMB157	0.136728	0.132104	0.135413	0.143768	Basal	0.143768
MDAMB175VII	0.224016	0.21786	0.173201	0.117779	Luminal A	0.224016
MDAMB231	0.070701	0.060817	0.048531	0.051681	Luminal A	0.070701
MDAMB361	0.150759	0.087107	0.135484	0.087908	Luminal A	0.150759
MDAMB415	0.106	0.083785	0.070655	0.045863	Luminal A	0.106
MDAMB436	0.051987	0.099007	0.065526	0.10453	Basal	0.10453

MDAMB453	0.169822	0.14676	0.173551	0.113635	HER2	0.173551
MDAMB468	0.034743	0.034994	0.095231	0.162813	Basal	0.162813
SKBR3	0.161321	0.151328	0.176783	0.134858	HER2	0.176783
T47D	0.197928	0.160988	0.132668	0.062742	Luminal A	0.197928
UACC812	0.205642	0.145025	0.169032	0.095471	Luminal A	0.205642
UACC893	0.205827	0.129598	0.242297	0.179517	HER2	0.242297
YMB1	0.180665	0.201045	0.126246	0.120499	Luminal B	0.201045
ZR751	0.188406	0.209278	0.128404	0.116821	Luminal B	0.209278
ZR7530	0.253548	0.222583	0.207872	0.134416	Luminal A	0.253548

Table 6 Gene expression based correlation score and classification

Cell lines Names	Luminal A	Luminal B	HER2	Basal	Classification	Max. Score
AU565	0.570026	0.57432	0.579763	0.576286	HER2	0.579763
BT20	0.564632	0.567867	0.575154	0.575742	Basal	0.575742
BT474	0.599571	0.602351	0.606111	0.599208	HER2	0.606111
BT483	0.589095	0.589134	0.587257	0.579759	Luminal B	0.589134
BT549	0.559006	0.562753	0.566721	0.570646	Basal	0.570646
CAL120	0.590019	0.592141	0.594513	0.598636	Basal	0.598636
CAL148	0.602213	0.60476	0.609575	0.604806	HER2	0.609575
CAL51	0.580683	0.582498	0.587331	0.588143	Basal	0.588143
CAL851	0.557406	0.560935	0.567473	0.571492	Basal	0.571492
CAMA1	0.55942	0.560788	0.561223	0.558354	HER2	0.561223
DU4475	0.517213	0.518507	0.522048	0.523279	Basal	0.523279
EFM19	0.546366	0.548151	0.54819	0.545028	HER2	0.54819
EFM192A	0.580619	0.583738	0.587441	0.581366	HER2	0.587441
EVSAT	0.553026	0.555579	0.561578	0.559659	HER2	0.561578
HCC1143	0.581157	0.58385	0.58917	0.59486	Basal	0.59486
HCC1187	0.559603	0.561519	0.568442	0.572434	Basal	0.572434
HCC1395	0.57537	0.578106	0.582776	0.587424	Basal	0.587424
HCC1419	0.586698	0.589336	0.590237	0.58094	HER2	0.590237
HCC1428	0.599791	0.602025	0.602648	0.600679	HER2	0.602648
HCC1500	0.580554	0.58022	0.577432	0.573	Luminal A	0.580554
HCC1569	0.511839	0.513867	0.518865	0.524124	Basal	0.524124
HCC1599	0.53831	0.540238	0.54575	0.550942	Basal	0.550942
HCC1806	0.52837	0.530124	0.530812	0.535139	Basal	0.535139

HCC1937	0.582408	0.585304	0.591811	0.597292	Basal	0.597292
HCC1954	0.576913	0.580767	0.589402	0.589169	HER2	0.589402
HCC202	0.565565	0.569084	0.573789	0.566555	HER2	0.573789
HCC2157	0.521166	0.522985	0.528185	0.53164	Basal	0.53164
HCC2218	0.538542	0.540795	0.542325	0.53553	HER2	0.542325
HCC38	0.574749	0.576395	0.582056	0.585008	Basal	0.585008
HCC70	0.570171	0.573132	0.57965	0.584247	Basal	0.584247
HDQP1	0.56835	0.570672	0.573425	0.577578	Basal	0.577578
HMC18	0.530087	0.532197	0.534592	0.539375	Basal	0.539375
HMEL	0.473079	0.47269	0.474992	0.480046	Basal	0.480046
HS274T	0.601471	0.607289	0.604908	0.605178	Luminal B	0.607289
HS281T	0.592256	0.598214	0.598233	0.597762	HER2	0.598233
HS343T	0.566836	0.572646	0.573177	0.573118	HER2	0.573177
HS578T	0.577653	0.583453	0.584346	0.586865	Basal	0.586865
HS606T	0.599365	0.604928	0.604132	0.60359	Luminal B	0.604928
HS739T	0.534029	0.538806	0.537864	0.535947	Luminal B	0.538806
HS742T	0.550857	0.555965	0.554277	0.55295	Luminal B	0.555965
JIMT1	0.575458	0.578295	0.583062	0.586071	Basal	0.586071
KPL1	0.602228	0.603976	0.606325	0.602303	HER2	0.606325
MCF7	0.54375	0.544845	0.544922	0.542028	HER2	0.544922
MDAMB134VI	0.518042	0.51605	0.514159	0.510426	Luminal A	0.518042
MDAMB157	0.557329	0.56068	0.563801	0.566379	Basal	0.566379
MDAMB175VII	0.55053	0.552245	0.550921	0.542686	Luminal B	0.552245
MDAMB231	0.539184	0.54147	0.544249	0.547399	Basal	0.547399
MDAMB361	0.580423	0.582531	0.586684	0.58096	HER2	0.586684
MDAMB415	0.6126	0.615648	0.613167	0.605896	Luminal B	0.615648
MDAMB436	0.568883	0.572186	0.577076	0.581384	Basal	0.581384

MDAMB453	0.598087	0.600993	0.606492	0.59937	HER2	0.606492
MDAMB468	0.563058	0.565885	0.574286	0.578314	Basal	0.578314
SKBR3	0.567165	0.570952	0.574411	0.570755	HER2	0.574411
T47D	0.576485	0.577867	0.576657	0.572871	Luminal B	0.577867
UACC812	0.536555	0.539204	0.535373	0.525882	Luminal B	0.539204
UACC893	0.574544	0.579545	0.586481	0.579644	HER2	0.586481
YMB1	0.536735	0.538636	0.539808	0.533984	HER2	0.539808
ZR751	0.537049	0.538381	0.537582	0.532485	Luminal B	0.538381
ZR7530	0.577092	0.579765	0.578823	0.568499	Luminal B	0.579765

Table 7 CCLE cell line FGA values

Cell line Name	FGA	Cell line Name	FGA	Cell line Name	FGA
CAL851	0.454959	HS578T	0.448292	HCC202	0.640264
BT549	0.422156	BT20	0.344444	HS739T	0.022813
EVSAT	0.583262	MDAMB436	0.845678	MDAMB157	0.436991
MDAMB453	0.3686	MDAMB361	0.546427	MDAMB175VII	0.278391
HS274T	0.022146	SKBR3	0.574749	HS742T	0.022504
HS281T	0.023455	HCC1395	0.670912	HS343T	0.022843
HCC1500	0.774086	HCC1806	0.428133	HCC1954	0.355743
HCC2157	0.51379	HCC1143	0.470245	HCC2218	0.392192
HCC1419	0.24137	DU4475	0.252385	CAMA1	0.283514
BT483	0.361104	HCC70	0.5935	HMC18	0.220312
ZR7530	0.236611	HCC38	0.645516	YMB1	0.366913
UACC812	0.446035	ZR751	0.353006	MCF7	0.525897
HCC1599	0.440108	BT474	0.395631	T47D	0.806502
JIMT1	0.484	EFM19	0.629806	MDAMB134VI	0.298061
CAL51	0.024347	MDAMB468	0.694528	KPL1	0.332333
HDQP1	0.47564	MDAMB231	0.305308	CAL148	0.214785
CAL120	0.422619	EFM192A	0.5461	UACC893	0.692182
HMEL	0.057846	HS606T	0.023226	HCC1428	0.467515
AU565	0.358645	HCC1937	0.487248	HCC1569	0.478483
MDAMB415	0.535105	HCC1187	0.472756		

Table 8 CCLE cell line mutation frequencies

Cell line Name	No. of mutation	No. of base pairs	Mutation frequency	Cell line Name	No. of mutation	No. of base pairs	Mutation frequency
AU565	68	14139745	4.809139	HCC2218	60	13738586	4.367262
BT20	73	13674648	5.338346	HCC38	45	13065569	3.444167
BT474	118	13453355	8.771046	HCC70	48	14834814	3.235632
BT483	58	11819716	4.907055	HMC18	66	12825948	5.145818
BT549	57	13476716	4.229517	HS281T	31	13729498	2.257912
CAL120	36	12436514	2.894702	HS343T	29	13775110	2.105246
CAL148	84	11855608	7.085255	HS578T	40	13580708	2.945355
CAL51	86	13111361	6.559197	HS739T	45	14655824	3.070452
CAL851	28	9761310	2.868467	HS742T	27	15306548	1.763951
CAMA1	67	13179068	5.083819	JIMT1	35	11903158	2.940396
DU4475	51	11691745	4.362052	KPL1	43	13206920	3.255869
EFM19	43	13999889	3.071453	MCF7	34	9688929	3.50916
EFM192A	33	12968331	2.544661	MDAMB134VI	35	14115109	2.479612
EVSAT	85	13226558	6.426464	MDAMB231	61	13146670	4.639958
HCC1143	59	13833410	4.265037	MDAMB361	126	14146129	8.90703
HCC1187	41	13231558	3.098652	MDAMB415	63	10110941	6.230874
HCC1395	69	13523808	5.102113	MDAMB436	54	11704866	4.613466
HCC1419	41	11350180	3.612278	MDAMB453	106	12747117	8.315606
HCC1428	55	10331529	5.32351	MDAMB468	59	11164856	5.284439
HCC1500	45	11496602	3.9142	SKBR3	63	14593061	4.31712
HCC1569	180	11367730	15.8343	T47D	34	14832063	2.292331
HCC1599	50	10634068	4.70187	UACC812	64	14222633	4.49987
HCC1806	44	13544676	3.248509	UACC893	46	11672003	3.941055
HCC1954	65	12945787	5.020938	ZR751	20	14639370	1.366179
HCC202	73	12485249	5.8469	ZR7530	71	14270318	4.975362
HCC2157	58	13012549	4.457236				

Aniruddha Pawar

532 Drake Street, Indianapolis, IN 46202
Phone (317) 559 5454, Email: apawar@iupui.edu

OBJECTIVE

Seeking a full time opportunity in Bioinformatics where my skills can be utilized to make data driven decisions to solve biological problems and deliver optimized solutions

QUALIFICATION HIGHLIGHTS

- Proficient in microarray, NGS and large multi-omics data analysis
- Strong proficiency in Perl, Python and R , familiarity with SQL, PHP
- Experience with UNIX clusters – PBS torque
- In-depth knowledge of computational and as well as biological concepts
- Experience of working efficiently on multiple collaborative projects
- Good communication skills and a team player

WORK EXPERIENCE

Data analyst – (CCBB), IU School of Medicine

October 2012 – Present

Model cell-lines for testing anti-cancer drugs

- Identify the cancer cell-lines most similar to breast cancer tissues in patients
- Characterize critical features such as copy number, gene expression and gene mutation in tissue samples and cancer cell-lines and then measure the similarities between these features
- Compute measures of similarities such as correlations between cancer cell-lines and cancerous tissues from patients

Next Generation Sequencing – Identification of excess variants from whole-exome sequencing

- Investigated the genetic variants in two traits of Alzheimer disease using NGS data
- Annotated variants using ANNOVAR from sequencing data of the participants
- Identified significant variants using SKAT package in R

Relationship between age/gender and expression of 'pharmaco-genes'

- Detected and reported relationships between several pharmaco-genes and age, gender which indicated that age and gender should be taken into account while determining the dosage of certain drugs
- Analyzed microarray data from a liver cohort study to detect these associations
- Processed the raw microarray dataset to restructure and normalize it

Analyst – Symphony Teleca Corporation, India
June 2011 to July 2012

- Worked in a team to build a unified support system for various open source bioinformatics tools
- Explored various open source and licensed biological software tools used for tasks such as genome annotation, format conversion, and variant analysis
- Prepared educational course ware documents to facilitate use and application of Bioinformatics software

Graduate research assistant – National Chemical Laboratory (NCL) Pune, India
(Council of Scientific & Industrial Research, CSIR Lab) February 2011 – May 2011

- Screened and partially purified extracellular nuclease enzyme from *Streptomyces aureofaciens* (NCIM-2614)

EDUCATION

- **Master of Science in Bioinformatics, GPA – 3.95/4.00**
  Indiana University, School of Informatics, Indianapolis – May 2015

- **Master of Science in Biotechnology**
  University of Pune, Pune, INDIA – May 2011

- **Bachelor of Science in Biotechnology**
  University of Pune, Pune, INDIA – May 2009

TECHNICAL SKILLS

- **Programming languages:** Perl, Python, R, SQL, PHP, HTML
- **Databases:** MySQL
- **Tools:** SVN, GitHub
- **Microarray analysis:** Normalization, MeV
- **NGS tools:** BWA, Bowtie, SAM tools, VCF tools, GATK, ANNOVAR
- **Public Databases:**1000 Genomes, TCGA, NCBI, KEGG, PDB, Gene Expression Omnibus (GEO), UCSC genome browser, Cancer Cell Line Encyclopedia (CCLE)
- **Statistical knowledge:** Descriptive statistics, hypothesis testing, regression, ANOVA, clustering
- **Operating Systems:** Windows, Unix/Linux and Macintosh
