META-ANALYSIS OF GENE EXRESSION DATA FOR IDENTIFICATION OF COMMON ALTERED GENES IN BREAST AND OVARIAN CANCERS

Abhishikta Hazra

Delhi Technological University, Delhi, India

Abstract

A technology that has been used extensively for analyzing breast and ovarian cancer malignancies is microarray technology. The combined investigation for breast and ovarian cancers across multiple gene expression studies are not well reported. In preliminary phase of study, data analysis was performed by combining gene expression profiles of eight different published microarray studies based on breast and ovarian cancers. Breast and ovarian cancers' genetic makeup are very similar and heritable mutations in the tumor suppressor genes BRCA1 and BRCA2 incline individuals to breast and ovarian cancers. Microarray data of both cancers was screened and downloaded from NCBI GEO. The raw data files were extracted and only low, high and normal grade tumor samples were included for the meta-analysis. After combining all the eight microarray data, normalization, and pre-processing, differential gene expression analysis (DEGs) were carried out. The statistical test that was used for identifying DEGs was one way ANOVA and this was followed by clustering the top DEGs. The clustering analysis explored the common genes and sample expression pattern including co-expressed gene sets across two types of cancers (breast and ovarian). This meta-analysis unified eight results of previous gene expression studies in breast and ovarian cancers. This analysis was performed using two different softwares viz. Robina and Genespring. The combined microarray data analysis result revealed the connection between common expressed genes in breast and ovarian cancers. It was found that the common DEGs and subsequently the co-expressed genes have strong enrichment from cell proliferation, ER signaling, actin cytoskeleton and Mitogen Activated Protein Kinase (MAPK) pathways. The research was continued by further pathway analysis of DEGs and co-expressed genes which then explored the common molecular basis, signatures and potential important regulatory pathways in these two cancer developments. The common up-regulated genes deduced after performing all the steps were *IRF5*, *IKZF2* and *CCNL1* and the common down-regulated genes included *ATF3*, *HMGA1* and *NRIP3*. The identified common altered genes in breast and ovarian expression data, which can serve as potential biomarkers, were validated using *in silico* method.

Chapter 1

Introduction

Breast cancer ranks second both in terms of being the most common diagnosed cancers after non-melanoma skin cancer and being the leading cause of cancer deaths after lung cancer in women (Perou *et al*., 2000). In women, ovarian cancer is the fifth most lethal cancer (Soegard *et al*., 2009). An obvious fact is both breast and ovarian cancers are heterogeneous diseases and the characteristics that define them includes biological subtypes of the tumor, age of outset, clinical course that it follows and it's response to treatment. While studying the epidemiology of these two cancers, it was established that family history plays an important role in the onset of both these cancers. The pattern in which autosomal dominant cancer susceptibility is inherited can be noticed in the development of breast and ovarian cancers as in some families, several family members are diagnosed with breast and/or ovarian cancers. Further, women with family history of these cancers are more susceptible to developing breast and ovarian cancers. All these are clear indications that genetic factors contribute to the risk of developing breast and ovarian cancers. At present, *BRCA1* and *BRCA2* are the most widely used biomarker genes for the prognostication and early detection of breast and ovarian cancers. If disease causing mutations are found in the *BRCA1* and *BRCA2* genes then that person is more likely to develop breast and/or ovarian cancers. Several tumor growths may arise from the ovary but the most predominant ovarian malignancy is found in the epithelial cells of the ovary, which is further one of the most frequently diagnosed gynaecological malignant growth (Cho, WSC. 2007). Approximately, five to ten percent of ovarian cancers are caused when there is a family history of this cancer and the pattern in which it is inherited can be classified into three groups, viz. ovarian and breast cancers together or ovarian and colon cancers together and ovarian cancer alone (Hanahan *et al*., 2000).

Gene expression studies (microarray studies) were developed more than a decade ago and since then the studies to analyze the variations in mRNA transcripts in disordered tissues are usually microarray based (Cho, WSC. 2007). Substantial microarray data have already been deposited into several international repositories and it includes Array Express and Gene Expression Omnibus (GEO) (Srinivas *et al.*, 2001).

Meta-analysis of openly accessible microarray data has been made use of, to establish shared characteristics in same cancer subtypes, for instance, lung vs. lung, liver vs. liver, breast vs. breast (Jemal *et al*., 2011). It is probable that the proteins that are translated from the mRNA transcripts are present in a differential manner in the disordered tissue, which might subsequently get secreted in the blood and then be detected. Wong *et al*. (2001) first integrated Oncomine cancer gene expression data and then several gene ontology annotations were used to filter the data viz., "extracellular", "extracellular matrix" and "extracellular space", which made them the first to predict serum protein biomarkers (Wong *et al*. 2001).

It was established from a multitude studies that a bunch of gene's expression levels can be employed as a "molecular fingerprint" in order to classify diverse types of tumor (Esquela *et al.,* 2006). Even within the same cancer type, the subtypes can be compared in order to find the common biomarkers. Meta-analyzing the microarray data of the same subtype of breast and ovarian cancers individually revealed common breast and ovarian cancer biomarkers. Till date no common biomarker meta-analysis study has been reported by combining breast and ovarian cancers gene expression data from publicly available data (NCBI GEO and Array Express). Integrating microarray data from multiple subtypes of breast and ovarian cancers in order to increase the sample size will be a promising approach for the identification of more robust common altered genes in breast and ovarian cancers. This method will further increase the data accuracy by assisting in pooling huge amount of raw data. It is possible that while a given gene may be not declared significantly to be differentially expressed by any one lab, the combination of results across labs in an integrative analysis can provide sufficient evidence to declare significant differential expression.

In the current study, eight publicly available gene expression datasets (four for each cancer) from different laboratories, representing breast and ovarian cancers on the same Affymetrix microarray platform U133 array were combined. The purpose was to obtain more powerful statistical results than the single dataset and identify common altered genes and their expression pattern in epithelial breast and ovarian cancer tissues. Robina and Genespring were used separately in order to identify the common altered genes in breast and ovarian cancers. Around ten common altered genes were identified using both these softwares individually. Finally the results of both these softwares were integrated in order to find the common co-expressed genes which are of more statistical significance and these can be predicted as probable biomarkers more accurately. The selection of these two methods was based on the fact that both have same normalization, that is robust multiarray averaging (RMA), for quality checks and the main analysis.

Chapter 2

Literature Review

2.1 Cancer

More than 11 million people are diagnosed with cancer every year. It is estimated that there will be 16 million new cases every year by 2020 (Cho, WSC. 2007). Cancer is a cluster of diseases involving alterations in the status and expression of multiple genes that confer a survival advantage and undiminished proliferative potential to somatic or germinal cells (Hanahan *et al.*, 2000). Alterations primarily in three main classes of genes viz., (proto) oncogenes, tumour suppressor genes and DNA repair genes collectively contribute to the development of cancer genotype and phenotype that resists the natural and inherent death mechanism(s) embedded in cells (apoptosis and like processes), coupled with dysregulation of cell proliferation events.

2.2 Proto-oncogenes

Proto-oncogenes are genes that normally help cells grow. When a proto-oncogene mutates or there are too many copies of it, it becomes a "bad" gene that can become permanently turned on or activated when it is not supposed to be. When this happens, the cell grows out of control, which can lead to cancer. This bad gene is called an oncogene (Bahcall, O. 2013).

3.3 Tumor suppressor genes

Tumor suppressor genes are normal genes that slow down cell division, repair DNA mistakes, help in apoptosis or programmed cell death. When tumor suppressor genes don't work properly, cells can grow out of control, which can lead to cancer (Berger *et al.*, 2011).

3.4 DNA repair

DNA in most cells is regularly damaged by endogenous and exogenous mutagens. Unrepaired damage can result in apoptosis or may lead to unregulated cell growth and cancer. If DNA damage is recognized by cell machinery, several responses may occur to prevent replication in the presence of genetic errors. At the cellular level, checkpoints can be activated to arrest the cell-cycle, transcription can be up-regulated to compensate for the damage, or the cell can apoptose (Vispe *et al.*, 2000). Alternatively, the damage can be repaired at the DNA level enabling the cell to replicate as planned. Complex pathways involving numerous molecules have evolved to perform such repair. Because of the importance of maintaining genomic integrity in the general and specialized functions of cells as well as in the prevention of carcinogenesis, genes coding for DNA repair molecules have been proposed as candidate cancer-susceptibility genes (Cairns *et al.*, 1982; Knudson *et al.*, 1989; Shields *et al.*, 1991).

3.5 Biomarkers

Technologies to recognize and understand the signatures of normal cells and how these become cancerous, promises to provide important insights into the aetiology of cancer that can be useful for early detection, diagnosis and treatment of cancers. Advancement in such technologies has instigated renewed interest in developing new biomarkers. Biomarkers of cancer could include a broad range of biochemical entities, such as nucleic acids, proteins, sugars, lipids, and small metabolites, cytogenetic and cytokinetic parameters as well as whole tumour cells found in the body fluid. Biomarkers are therefore invaluable tools for cancer detection, diagnosis, patient prognosis and treatment selection (Ludwig *et al.,* 2005). These can also be used to localize the tumor and determine its stage, subtype, and response to therapy (Bayli *et al.,* 2006).

Genetics, genomics, proteomics, many non invasive imaging techniques etc., allow measurement of several biomarkers. Currently, there is a greater understanding of the disease pathways, the protein targets and the pharmacologic consequences of drug administration. Therefore, application of biomarkers in the clinical practice is likely to result in advanced knowledge leading to a better understanding of the disease process that will facilitate development of more effective and disease specific drugs with minimal undesired systemic toxicity (Egger *et al.,* 2004). Establishment of biomarkers requires a comprehensive understanding of the molecular mechanisms and cellular processes underlying the initiation of cancer, especially focusing on how small changes in only a few regulatory genes or proteins can disrupt a variety of cellular functions. A major challenge in cancer diagnosis is to establish the exact relationship between cancer biomarkers and the clinical pathology, as well as, to be able to non-invasively detect tumors at an early stage. Similarly, identification of subtle changes in the genomics and proteomics status specific to malignant transformation will allow molecular targets to be used for developing therapeutics (Sawyers, CL. 2008).

3.6 Breast cancer

Worldwide, breast cancer accounts for 22.9% of all cancers (excluding non-melanoma skin cancers) in women. In 2008, breast cancer caused 458,503 deaths worldwide (13.7% of cancer deaths in women). Breast cancer is more than 100 times more common in women than in men, although men tend to have poorer outcomes due to delays in diagnosis (Srinivas *et al.,* 2001).

Breast cancer is a heterogeneous disease, comprising multiple entities associated with distinctive grading, histological and biological features, clinical presentations and behaviours and responses to therapy (Farley *et al*., 2008). Grading focuses on the appearance of the breast cancer cells compared to the appearance of normal breast cancer cells (Ikpatt *et al.,* 2005). Normal cells in breast become differentiated, taking specific shapes and forms that reflect their function as part of that organ. But, cancerous breast cells loose that differentiation. In cancerous condition, the cells that would normally line up in an orderly way to make up the milk ducts become disorganized. Further, cell division becomes uncontrolled and cell nuclei become less uniform. Pathologists describe cells as well

differentiated (low-grade), moderately differentiated (intermediate-grade), and poorly differentiated (high-grade) as the cells progressively lose the features seen in normal breast cells. Poorly differentiated cancerous cells have worst prognosis (Farley *et al*., 2008).

Breast cancer being a heterogeneous disease comprises of various types of neo-plasms, which involves different profile changes in both mRNA and micro-RNA (miRNA) expression. Extensive studies on mRNA expression in breast tumor have yielded some very interesting findings, some of which have been validated and used in clinic. It's a proven fact that *BRCA1* mRNA expression plays a major role as a marker of time to progression and overall survival in sporadic breast cancers treated with chemotherapy (Margeli *et al*., 2010). Recent miRNA research advances showed great potential for the development of novel biomarkers and therapeutic targets. It has been demonstrated that miRNA expression is frequently deregulated in breast cancer, which warrants further in-depth investigation to decipher their precise regulatory role in tumorigenesis. Several studies were directed towards the regulatory mechanism of miRNA, expression level of miRNA in tumorous state, and their potential use as breast cancer biomarkers for early disease diagnosis (Filipowicz *et al.*, 2008).

3.7 Ovarian cancer

Most (more than 90%) ovarian cancers are classified as "epithelial" and are believed to arise from the surface epithelium of the ovary. Upon histological evaluation, most ovarian cancers are found to be epithelial in nature and are collectively referred to as ovarian epithelial cancers (OEC). The most common OEC subtypes include, in decreasing order of frequency, serous adenocarcinomas, followed by endometrioid, and smaller subsets of mucinous, clear cell, transitional, and undifferentiated carcinomas (Pradhan *et al.*, 2010).

However, some evidence suggests that the fallopian tube could also be the source of some ovarian cancers. Since the ovaries and tubes are closely related to each other, it is thought that these fallopian cancer cells can mimic ovarian cancer. Other types may arise from the egg cells (germ cell tumor) or supporting cells. Ovarian cancers are included in the category gynecologic cancer (Soegard *et al.*, 2009)

In the United States, invasive ovarian cancer is the 5th most deadly malignancy in females, accounting for an estimated 13,850 deaths in 2010 (Ahmad, S. 2011). The risk of dying from ovarian cancer depends on staging and varies greatly. Ovarian cancer patients diagnosed at the localized stage exhibit a 5 year survival rate of 94%. This rate is 73% when diagnosed at the regional stage following local dissemination and drops to 28% when a patient is diagnosed at the distant stage with metastasis to organs outside the pelvis. Overall, the combined 5 year survival rate for all ovarian cancer patients is an unmanageable 46% (Abbott, KL. 2010).

Ninety percent of human cancers, however, are epithelial in origin and display marked aneuploidy, multiple gene amplifications and deletions, and genetic instability, making resulting downstream effects difficult to study with traditional methods. Recent technologies, like microarray technology corroborates beneficial for such analysis (Gray *et al.*, 2000).

3.8 Microarray technology

Gene expression studies in human cancer can identify genetic markers of malignant transformation. Traditionally, such studies were limited to examining a few genes at a time. However, different methods are now available for large-scale gene expression analysis. For example, microarray technology is used to find out the expression of large number of genes simultaneously (Ponder, BA. 2001).

Microarray methods were initially developed to study differential gene expression using complex populations of RNA. Refinement of these methods now permits the analysis of copy number imbalances and gene amplification of DNA (Jain *et. al.*, 2003).

3.9 Affymetrix GeneChip array

Affymetrix, Inc. is an American company that manufactures DNA microarrays; it is based in United States. Affymetrix makes quartz chips for analysis of DNA microarrays called GeneChip arrays. Affymetrix's GeneChip arrays assist researchers in quickly scanning for the presence of particular genes in a biological sample. Within this area, Affymetrix is focused on oligonucleotide microarrays. These microarrays are used to determine which genes exist in a sample by detecting specific pieces of mRNA. A single chip can be used only once to analyze thousands of genes in one assay (Quackenbush, J. 2001).

The GeneChip Human Genome U133 Plus 2 array is a single array representing 14,500 well-characterized human genes that can be used to explore human biology and disease processes. The salient features of Human Genome U133 Plus 2 array includes coverage of well-substantiated genes in the transcribed human genome on a single array, analytical ability of the expression level of 18,400 transcripts and variants, including 14,500 well-characterized human genes, being comprised of more than 22,000 probe sets and 500,000 distinct oligonucleotide features (Tusher *et. al.*, 2001).

3.10 Data analysis

There is an exponential growth in the numbers of microarray-based studies identifying new genes or molecular pathways involved in tumor classification, cancer progression, or patient outcome. We are now in “postgenomic era”, during which the diagnostic, prognostic, and treatment response biomarker genes identified by microarray screening are about to be cross-examined to provide personalized management of patients (Rousseau *et al.,* 2012).

Gene expression studies pose many challenges for data organization, storage and analysis (Quackenbush, J. 2001). Present technology allows for the evaluation of nearly the entire genome from a single biologic sample. Databases are required for efficient storage and retrieval of this information, but most biomedical laboratories are not set up to handle this type of data (Ermolaeva *et al.,* 1998). Furthermore, there are no standards for the design and implementation of expression databases. These limitations presently make it difficult to compare datasets generated in different laboratories. To date, the computational analysis of gene expression data has centered on two approaches. One is unsupervised learning or

clustering and the other one is supervised learning. Unsupervised learning involves the aggregation of a diverse collection of data into clusters based on different features in a data set. For example, one could divide a group of people into clusters based on any combination of eye color, waist size or height. Similarly, one can gather data about the various expressed genes in a collection of tumor samples and then cluster the samples as best as possible into groups based on the similarity of their aggregate expression profiles.

One could cluster genes across all samples, to identify genes that share similar patterns of expression in varying biologic contexts. Such approaches have the advantage of being unbiased and allow for the identification of structure in a complex data set without making any a priori assumptions. However, because many different relationships are possible in a complex data set, the predominant structure uncovered by clustering may not necessarily reflect clinical or biologic distinctions of interest (Jain *et al.,* 2003).

The other approach, supervised learning, on the other hand incorporates the knowledge of class label information to make distinctions of interest. A training data set is used to select those features that best make a distinction. These features are then applied to an independent test data set to validate the ability of selected features to make that distinction. For example, one could select a subset of expressed genes that are best able to distinguish between two cancer types and build a computational model that uses these selected genes to sort an independent, unlabelled collection of those tumor types into the two groups of interest. However, supervised learning is dependent on accurate sample labels, which can be an issue given the limitations of histopathologic cancer diagnosis (Tusher *et al.,* 2001).

Sometimes, results from unsupervised and supervised learning on a single data set can overlap, but this does not have to be the case. An important issue with either analytic approach is that of statistical significance of observed correlations. A typical microarray experiment yields expression data for thousands of genes from a relatively small number of samples, and gene-class correlations, therefore, can be revealed by chance alone. This issue can be addressed by collecting more samples for each class studied, but this is often difficult with clinical cancer samples (Tusher *et al.,* 2001; Jain *et al.,* 2003).

Another approach is to perform exploratory data analysis on an initial data set and apply findings to an independent test set. Findings confirmed in this fashion are less likely a result of chance. Permutation testing, which involves randomly permuting class labels and determining gene-class correlations, has also been used to determine statistical significance. Observed gene-class correlations that are stronger than those seen in permuted data are considered statistically significant (Ermolaeva *et al.,* 1998).

Clinicians will be able to use microarrays during early clinical trials to confirm the mechanisms of action of drugs and to assess drug sensitivity and toxicity. Coupled with more conventional biochemical analysis such as Immunohistochemistry (IHC) and Enzyme Linked Immunosorbent Assay (ELISA), microarrays will be used for diagnostic and prognostic purposes. Kim *et al.* (2011) published an example of such a potential “bench to bedside” translation (Kim *et al.,* 2011). The osteopontin gene, which encodes a calcium binding

glycophosphoprotein, had been identified by cDNA microarray analysis as being up-regulated in ovarian cancer. Kim *et al*. (2011) showed that screening of plasma samples from ovarian cancer patients revealed that osteopontin protein concentrations in plasma was significantly higher in a majority of samples with ovarian cancer compared with normal controls. This study demonstrated the potential value of cDNA microarray analysis in identifying biomarker genes in cancer and the feasibility of subsequently testing these genes at the protein level by conventional biochemical assays (Quackenbush, J. 2001). The technology is becoming increasingly user friendly, automated and cost effective too with the advent of freely available potent softwares like Robina (Gyorffy *et al*., 2009).

Chapter 3
Methodology

3.1 Tools used in the analysis

3.1.1 NCBI GEO

GEO is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community. The GEO DataSets database stores original submitter-supplied records (Series, Samples and Platforms) as well as curated DataSets. The GEO DataSets database stores original submitter-supplied records (Series, Samples and Platforms) as well as curated DataSets. (Barrett *et al.*, 2009)

3.1.2 Robina tool

Robina represents an easy to use graphical interface for microarray (Affymetrix GeneChip, other single channel (e.g. Agilent) and two colour) analysis functions from R/BioConductor. Here, we have used Robina for three main purposes viz. quality assessment of our data, normalization of our microarray data and detection of differentially expressed genes. (Anders *et al.*, 2010)

3.1.3 Genespring

Agilent's Genespring provides powerful, accessible statistical tools for intuitive data analysis and visualization. Designed specifically for the needs of biologists, Genespring offers an interactive environment that promotes investigation and enables understanding of Transcriptomics, Metabolomics, Proteomics and NGS data within a biological context. Genespring allows us to quickly and reliably identify targets of interest that are both statistically and biologically meaningful (Buscaglia *et. al.*, 2011).

3.1.4 Affymetrix

NetAffx™ Analysis Center. The NetAffx™ Analysis Center enabled us to correlate the GeneChip array results with array design and annotation information (Quackenbush 2001).

3.1.5 Comparative toxicogenomics database

CTD promotes understanding about the effects of environmental chemicals on human health by integrating data from curated scientific literature to describe chemical interactions with genes and proteins, and associations between diseases and chemicals, and diseases and genes/proteins (Li *et al*., 2009).

3.1.6 DAVID database

DAVID bioinformatics resources consist of an integrated biological knowledgebase and analytic tools aimed at systematically extracting biological meaning from large gene/protein lists. DAVID, is a high-throughput and integrated data-mining environment, to analyze gene lists derived from high-throughput genomic experiments. The procedure first requires uploading a gene list containing any number of common gene identifiers followed by analysis using one or more text and pathway-mining tools such as gene functional classification, functional annotation chart or clustering and functional annotation table. By following this protocol, investigators are able to gain an in-depth understanding of the biological themes in lists of genes that are enriched in genome-scale studies (Huang *et. al*., 2008).

3.1.7 PANTHER database

PANTHER (Protein Analysis Through Evolutionary Relationships) classification system was designed to classify proteins (and their genes) in order to facilitate high-throughput analysis. Proteins have been classified according to: Family and subfamily: families are groups of evolutionarily related proteins; subfamilies are related proteins that also have the same function; Molecular function: the function of the protein by itself or with directly interacting proteins at a biochemical level, e.g. a protein kinase; Biological process: the function of the protein in the context of a larger network of proteins that interact to accomplish a process at the level of the cell or organism, e.g. mitosis; Pathway: similar to biological process, but a pathway also explicitly specifies the relationships between the interacting molecules (Bateman *et. al*., 2002).

3.1.8 GOBO web interface

GOBO is a user-friendly online tool that allows rapid assessment of gene expression levels, identification of co-expressed genes and association with outcome for single genes, gene sets or gene signatures in an 1881-sample breast cancer data set. Moreover, GOBO offers the possibility of investigation of gene expression levels in breast cancer subgroups and breast cancer cell lines for gene sets, as well as creation of potential metagenes based on iterative correlation analysis to a prototype gene (Gyorffy *et. al*., 2009). The web interface of GOBO allows precompiled data sets to be queried by the three main applications of GOBO: Gene Set Analysis (GSA), Co-expressed Genes (CG), and Sample Prediction (SP). Currently, the precompiled data sets consist of gene expression data and annotation data for a pooled 1881-sample breast tumor set and 51 previously reported breast cancer cell lines. The 881-sample breast tumor set comprises 11 public data sets analyzed using Affymetrix U133A arrays and processed. GSA is further divided into outcome analysis in breast tumors (GSA-Tumor) and expression patterns in breast cancer cell lines (GSA-Cell line). In both GSA applications the input is either a single gene or probe identifier, or a set of gene/probe identifiers (referred to as a gene set hereinafter). CG allows identification of coexpressed genes by provision of a single gene identifier in both the breast tumor data set and the panel of breast cancer cell lines. SP allows users to investigate the association of their classifiers (in certain predefined forms) with outcome in the 1881-sample breast cancer set (Karn *et. al.*, 2010).

3.2 Gene expression data from NCBI GEO

The raw gene expression data on Affymetrix platform for eight studies was screened and downloaded from the journal articles and Gene Expression Omnibus (GEO). These studies were selected on the basis of similar cancer grade, platform, array and large sample size. The measures in comparing and combining the two different cancer types expression data were similarity of experiment, sample source (epithelial cancer cells were considered), cancer grade or stages, number of sub groups in each experiment and total number of samples present in the group.

3.3 Data pre-processing and normalization

The entire original downloaded microarray data (CEL files) for all the experiments were pre-processed using RMA (robust multi array averaging) algorithm. Firstly background adjustment was performed, followed by normalization of data and finally, a linear model was fitted to the corrected and normalized probe intensities. These were the three steps followed in RMA. The Robina and Genespring tools were used for data normalization, visualization and analysis. An experimental design was fed into Robina in order to compare the data and specify the direction of comparison (Figure 3.1).

Figure 3.1: Experiment design fed in Robina

Next, filtering of raw data was carried out by excluding the probes whose intensities were less than twenty percentile and probes whose coefficient of variation (CV) was less than fifty percent were selected for analysis.

3.4 Identification of differentially expressed genes (DEGs)

One way ANOVA analysis was performed for differential gene expression analysis by matching breast and ovarian cancer groups with normal samples. A threshold value of 0.05 for P value and 1.5 for fold change were considered during differentially expressed genes analysis using Robina and Genespring tools. That is, unpaired t-test was used for this gene expression analysis. NetAffx analysis center of Affymetrix was used to correlate array information with annotation to identify ‘gene symbol’ and ‘gene title’ among others. In following step, the top DEGs (highly up-regulated and highly down-regulated) were screened on the basis of fold change and processed for the clustering and pathway analysis. Top twenty DEGs were considered for the following steps.

3.5 Identification of co-expressed and co-regulatory sets of genes

In order to reveal the expression and co-regulation of genes, the screened top twenty genes from the DEGs analysis were processed for the cluster analysis. Unsupervised hierarchical clustering (HCL) was the preferred clustering method. Separate clustering was done for up-regulated genes of breast and ovarian cancers and down-regulated genes of breast and ovarian cancers. Comparative Toxicogenomics Database (CTD) was used to generate a comparative data of the co-expressed genes of both these differently regulated cancer groups.

3.6 Identification of common altered genes, pathways and functional annotations

The Venn diagram analysis was performed for mining the co-expressed genes in both breast and ovarian cancer groups. Some unique gene expression signatures in breast and ovarian cancers were also identified from the Venn diagram analysis. The results from Robina and Genespring were assessed to identify the common co-expressed genes between breast and ovarian cancers. The common altered genes' Affymetrix probe IDs were submitted to David database for both pathway analysis and functional classification analysis and Panther online analysis server for functional classification analysis.

3.7 *In silico* validation of common altered genes in both breast and ovarian cancers

The common altered gene symbols and IDs in both breast and ovarian cancers were mapped in GOBO online web server for *in silico* validation and prediction of screened biomarker genes. The common identified altered genes were given as input and the result included only those which were showing strong enrichment from breast cancer expression data. *In silico* validation of the altered genes was possible only for breast cancer as a validating web server like GOBO is still missing for ovarian tumor data set.

The workflow shown in the next page was used in this study (Figure 3.2).

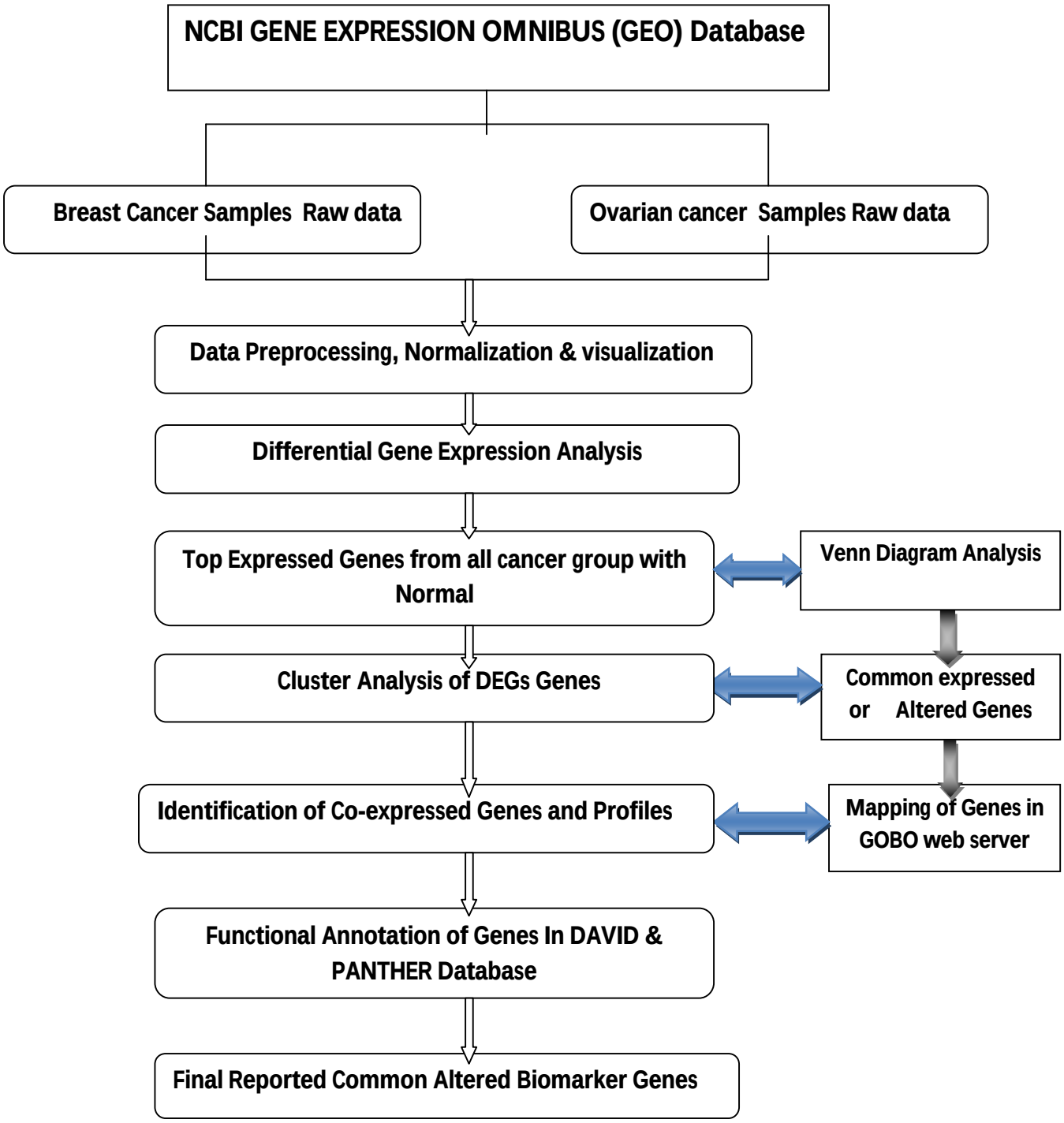


Figure 3.2: Workflow of study

Chapter 4

Results

4.1 Gene expression data from NCBI GEO

Similarity of experiment, sample source (epithelial cancerous cells), experiment groups or stages, number of sub groups in each experiment and total number of samples present in the groups are the conditions that were checked while selecting and downloading raw data from NCBI GEO and finally four pairs of similar type experiment groups were identified (Table 4.1). The normal samples for both these cancers were also downloaded along with the different experiments from the NCBI GEO.

Table 4.1: Overview of datasets retrieved and used in analysis

NCBI GEO ID	Array type	No of Samples	Authors	Source	Data Type
GSE52262	HG-U133_Pl us_2	27	Liu S *et al.*, 2013	Epithelial cells	Breast Cancer
GSE52327	HG-U133_Pl us_2	16	Conley *et al.*, 2012	Tissue	Breast Cancer
GSE31192	HG-U133_Pl us_2	33	Harvell DM, Kim J, O'Brien J, Tan AC *et al.*, 2013	Epithelial Cells	Breast Cancer
GSE27018	HG-U133_Plus_2	17	Luciani MG, Seok J, Sayeed A, Champion S *et al.*, 2011	Epithelial Cells	Breast Cancer
GSE29220	HG-U133_Plus_2	22	Lee Y, Kim J, Zhou H, Wong DT, 2011	Tissue	Ovarian Cancer
GSE18680	HG-U133_Pl us_2	12	Kulbe H	Tissue	Ovarian Cancer
GSE15578	HG-U133_Pl us_2	17	Pejovic T,2009	Epithelial cells	Ovarian cancer
GSE14001	HG-U133_Pl us_2	23	Tung CS, Mok SC, *et al*	Surface Epithelia	Ovarian cancer

4.2 Robina

4.2.1 Data preprocessing and normalization

Box plots of the unnormalized expression values on each chip give a global overview of the

signal intensity distributions. Preferably, comparable distribution of all the chips is desired even before performing normalization. Plotting smoothed histograms of the (log2) signal intensity of all perfect match (PM) probes is another way of visualizing the distribution of signal intensities (Figure 4.1, 4.2). In order to set the median standard error of each probeset to one, NUSE plots standardize the probe level models for each probeset across all chips. It visualizes the standard error distribution of each individual chip. Chips with consistently increased standard errors are bound to be of low quality and the experiment proceeds keeping those aside.

The logarithmic expressions of each probeset on every chip are compared to the median expression of the probes in order to compute the Relative Logarithmic Expression (RLE). The median RLE value should be zero if it is to be assumed that majority of genes, under a particular given treatment are not differentially expressed.When there is a deviation from the zero mark or when it is noticed that the box plot of RLE is having an increased spread for individual arrays then it is concluded that those arrays are of low quality. Low quality chips with strong outlier behaviour are indicated by red ellipses (Figure 4.3, 4.4).

Next is the RNA degradation plot, where the probes are ordered from 5’ to 3’ direction. In general, it is found that RNA degradation is more dynamic at the 5’ end so, correspondingly, probes closer to this end have low signal intensities. If the slopes of individual chips are deviating from the median by more than 10% then Robina issues a warning (Figure 4.5).

Thus, the identified two chips of low quality showing strong outlier behaviour were removed from study and we continued with the other chips (Table 4.2).

Table 4.2: Removed chips after quality check

EXPERIMENT/PATIENT	CONTROL/HEALTHY
GSM722641	GSM722652

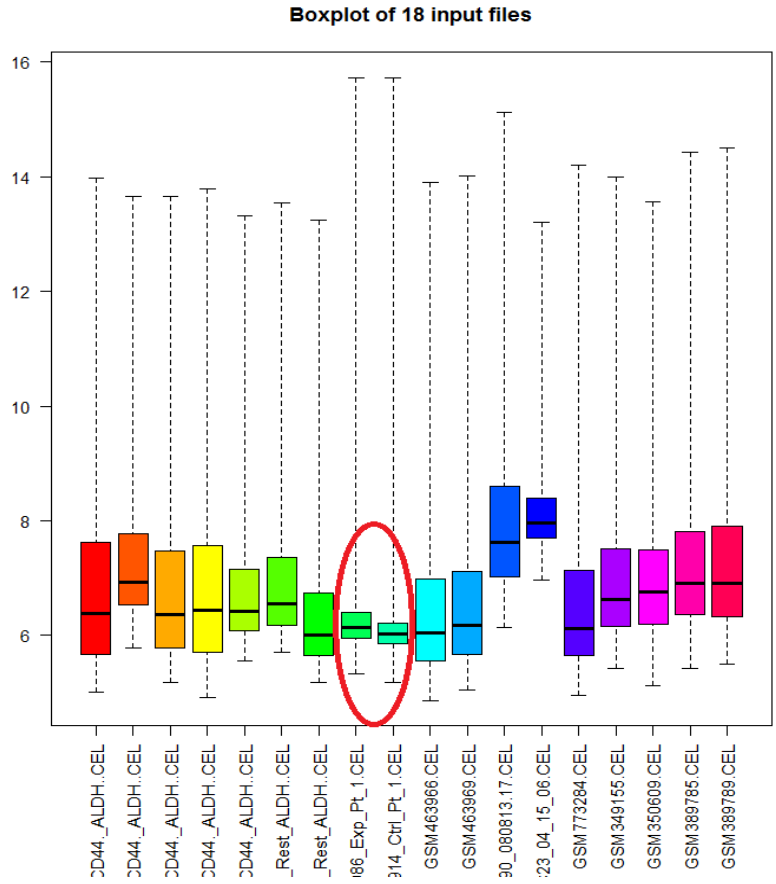


Figure 4.1: Boxplots from Robina

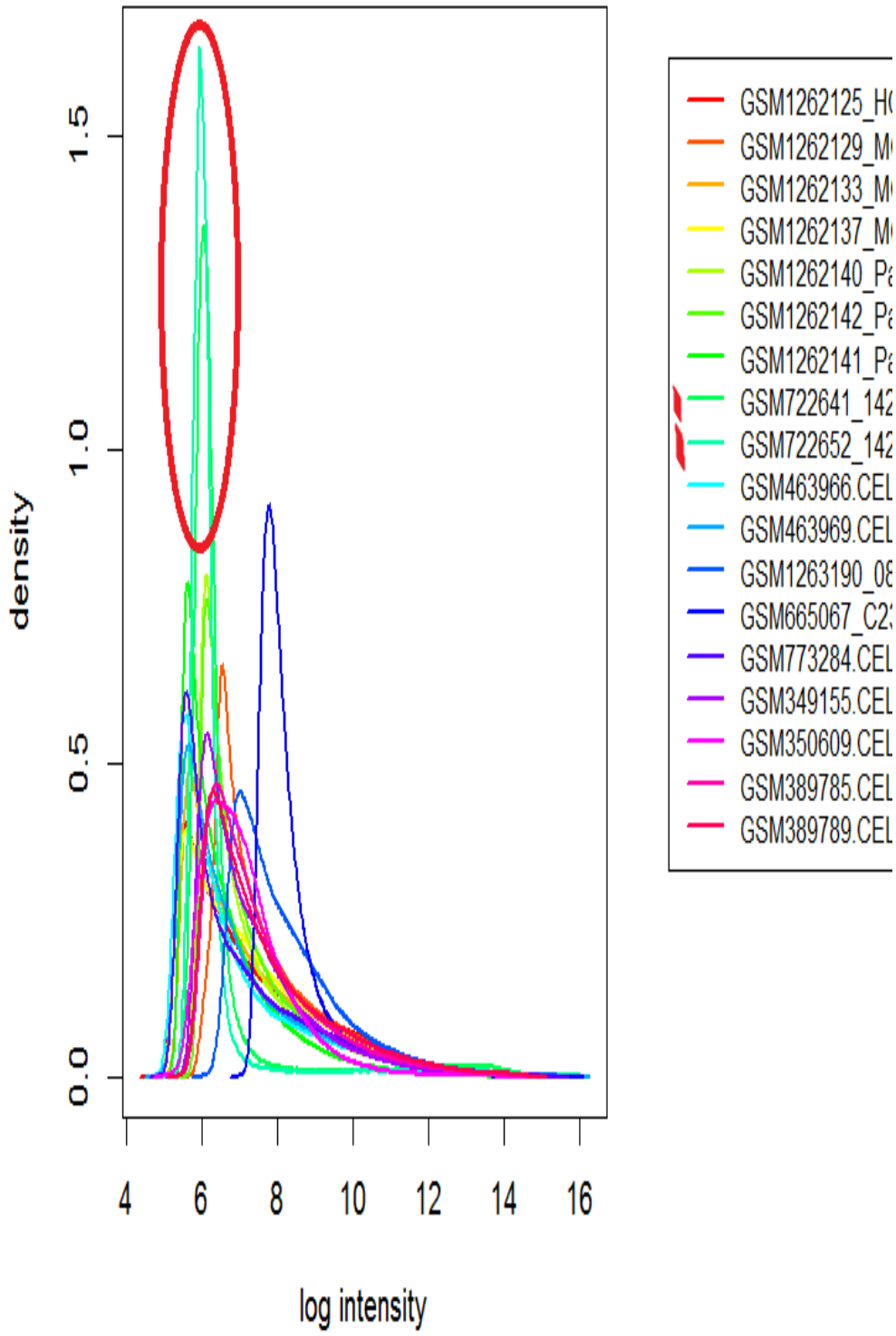


Figure 4.2: Smoothed histograms of the signal intensities of all perfect match probes

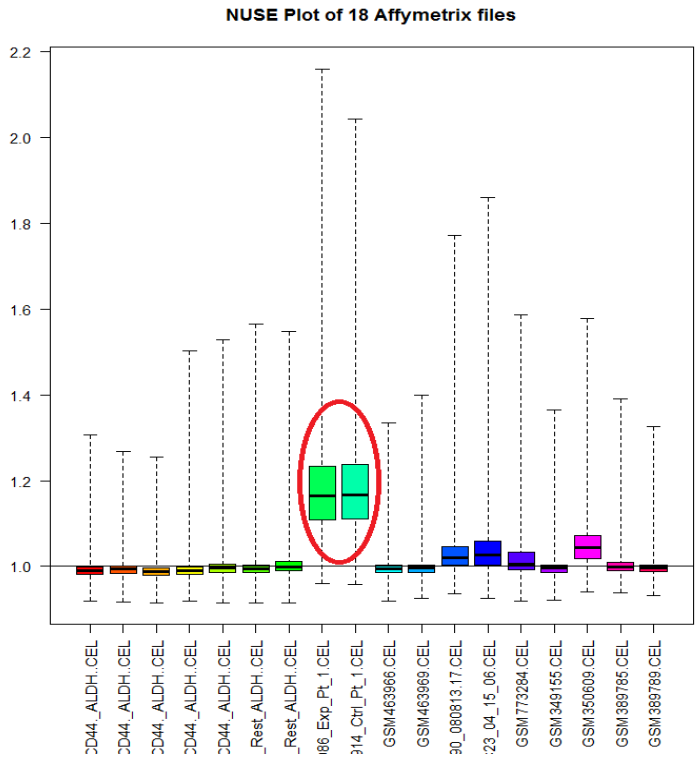


Figure 4.3: Normalized un-scaled standard error plots

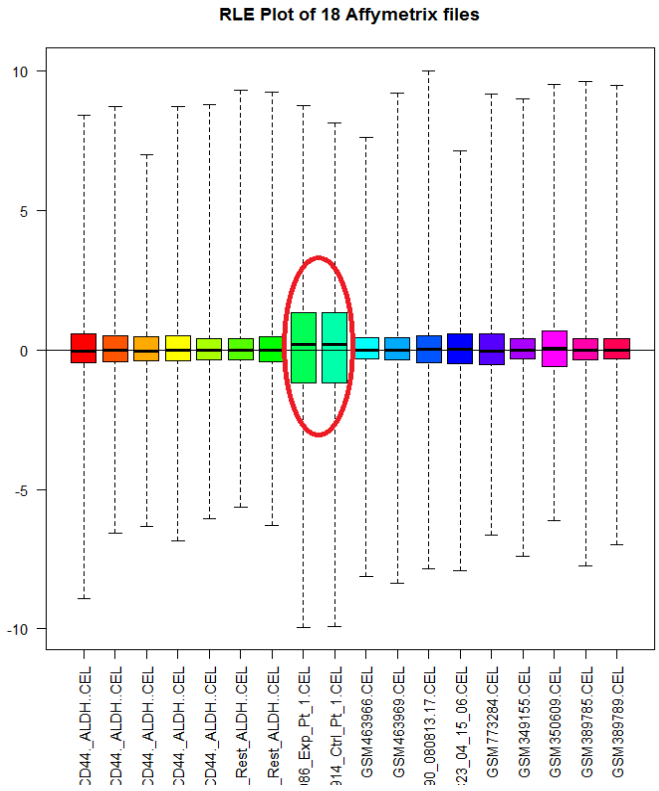


Figure 4.4: Relative Logarithmic Expression

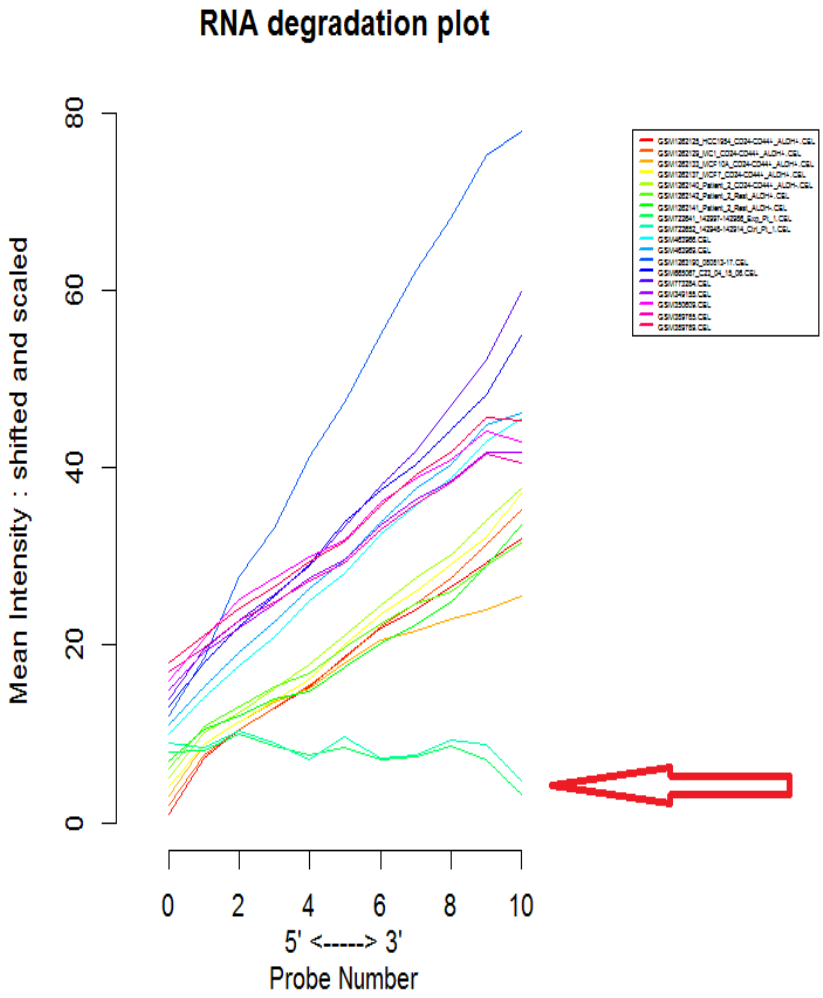


Figure 4.5: RNA degradation plot

4.2.2 Identification of differentially expressed genes (DEGs)

Annotation information was retrieved using NetAffx analysis centre of Affymetrix and top ten up-regulated and down-regulated genes for following four different sets were analyzed and categorized-

Breast Cancer mapped with Breast Normal (BC_BN)
Breast Cancer mapped with Ovarian Cancer
Breast Cancer mapped with Ovarian Normal
Ovarian Cancer mapped with Ovarian Normal (OC_ON)

For further analysis we needed only mapped data of 'Breast Cancer vs. Breast Normal' and 'Ovarian Cancer vs. Ovarian Normal', so the table of only these two comparisons were

considered. The 20 DEGs represented by these probes are enlisted in tables (Table 4.3, 4.4, 4.5, 4.6).

Table 4.3: Up-regulated breast cancer genes compared to breast normal genes from Robina

logFC	P.Value	ID	Gene Symbol	Gene Title	Representative Public Id
5.560972	0.000122	201291_s_at	TOP2A	topoisomerase (DNA)	AU159942
4.968754	0.000268	1557094_at	LOC100996760	uncharacterized LOC1	BC029890
4.880495	0.00022	218542_at	CEP55	centrosomal protein 5	NM_018131
4.746589	2.38E-05	212022_s_at	MKI67	marker of proliferatic	BF001806
4.709012	2.15E-05	203418_at	CCNA2	cyclin A2	NM_001237
4.463984	1.25E-05	202095_s_at	BIRC5	baculoviral IAP repea	NM_001168
4.35521	3.20E-07	1555826_at	BIRC5 /// EPR-	baculoviral IAP repea	BQ021146
4.200034	4.47E-05	218355_at	KIF4A	kinesin family memb	NM_012310
3.936971	5.49E-05	215509_s_at	BUB1	BUB1 mitotic checkpo	AL137654
3.855997	6.06E-05	231534_at			
3.783167	3.90E-05	229490_s_at			
3.679598	0.000249	216228_s_at	WDHD1	WD repeat and HMG-l	AK001538
3.588418	5.70E-05	205469_s_at	IRF5	interferon regulatory	AI028035
3.555828	0.000269	1552682_a_at			
3.509179	0.000107	209709_s_at			
3.2345	0.000267	231929_at	IKZF2	IKAROS family zinc fir	AI458439
3.186033	0.000104	229492_at			
3.175463	0.000155	1555827_at	CCNL1	cyclin L1	AY034790
3.14236	0.000142	222380_s_at	PDCD6	programmed cell dea	AI907083

Table 4.4: Down-regulated breast cancer genes compared to breast normal genes from Robina

-3.74191	0.000233	221992_at	NPIPB15	nuclear pore complex	AI925734
-3.88275	7.93E-05	209242_at			
-4.03309	3.79E-05	211819_s_at			
-4.05855	8.74E-05	207113_s_at			
-4.08195	9.21E-05	210457_x_at	HMGA1	high mobility group A	AF176039
-4.15545	0.000282	230233_at			
-4.28245	8.21E-05	204213_at			
-4.32458	0.000131	222900_at	NRIP3	nuclear receptor inte	AJ400877
-4.4753	0.000109	227742_at			
-4.50724	3.41E-06	205030_at	FABP7	fatty acid binding pro	NM_001446
-4.70227	0.000114	220133_at	ODAM	odontogenic, amelob	NM_017855
-4.73123	2.09E-07	211302_s_at	PDE4B	phosphodiesterase 4	L20966
-4.73385	1.74E-05	202672_s_at	ATF3	activating transcriptic	NM_001674
-4.81775	0.000104	206509_at	PIP	prolactin-induced pro	NM_002652
-4.91922	0.000103	210413_x_at	SERPINB3 /// S	serpin peptidase inhi	U19557
-5.01027	5.22E-05	209842_at	SOX10	SRY (sex determining	AI367319
-5.24087	2.82E-05	203708_at	PDE4B	phosphodiesterase 4	NM_002600
-5.52384	2.63E-05	203665_at	HMOX1	heme oxygenase (de	NM_002133
-5.6322	7.97E-05	205916_at	S100A7	S100 calcium binding	NM_002963
-5.74102	2.58E-05	228245_s_at	LOC100509445	uncharacterized LOC1	AW594320
-6.13025	8.86E-05	206378_at	SCGB2A2	secretoglobin, family	NM_002411

Table 4.5: Up-regulated ovarian cancer genes compared to ovarian normal genes from Robina

logFC	P.Value	ID	Gene Symbol	Gene Title	Representative Pub
2.38585	0.001201	207542_s_at	AQP1	aquaporin 1	NM_000385
2.206819	0.00736	210619_s_at	HYAL1	hyaluronoglu	AF173154
2.158729	0.005803	1569555_at	GDA	guanine dea	BC012859
2.082519	0.00404	229797_at	MCOLN3	mucolipin 3	AI636080
2.056253	0.000553	224179_s_at	MIOX	myo-inositol	AF230095
2.041843	0.008646	227394_at	NCAM1	neural cell a	W94001
1.970595	0.002712	220332_at	CLDN16	claudin 16	NM_006580
1.91088	0.012877	234723_x_at	---	---	AK024881
1.864484	0.002707	236717_at	FAM179A	family with s	AI632621
1.815494	0.012086	232046_at	KIAA1217	KIAA1217	AU148164
1.806412	0.001814	209755_at	NMNAT2	nicotinamide	AF288395
1.664403	0.005341	1556029_s_at	NMNAT2	nicotinamide	H90656
1.648862	0.00963	231929_at	IKZF2	IKAROS fami	AI458439
1.601447	0.007862	1552395_at			
1.598335	0.010261	1555827_at	CCNL1	cyclin L1	AY034790
1.552637	0.008786	204729_s_at			
1.546504	0.005654	239907_at			
1.530273	0.009699	205469_s_at	IRF5	interferon re	NM_002200
1.520127	0.011722	1557669_at			

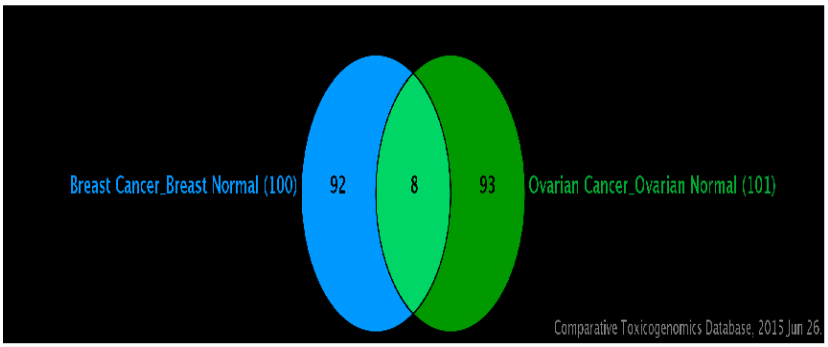
Table 4.6: Down-regulated ovarian cancer genes compared to ovarian normal genes from Robina

-2.0919	0.0115	213899_at			
-2.12668	0.001304	220327_at			
-2.14827	0.00761	214954_at			
-2.2234	0.001429	213362_at	PTPRD	protein tyros	N73931
-2.22758	0.002741	205517_at	GATA4	GATA bindin	AV700724
-2.23604	0.005772	202920_at	ANK2	ankyrin 2, ne	BF726212
-2.33957	0.00053	222900_at	NRIP3	nuclear rece	AJ400877
-2.41506	0.00437	33767_at	HHLA1	HERV-H LTR-	AU148706
-2.41583	0.008085	210457_x_at	HMGA1	high mobility	AF176039
-2.46114	0.004228	214841_at	CNIH3	cornichon fa	AF070524
-2.51472	0.007544	242277_at			
-2.56126	0.008528	234304_s_at	IPO11 /// IPO1	importin 11 /	AL162083
-2.74044	0.009836	211340_s_at	MCAM /// MIR	melanoma c	M28882
-2.91192	0.001597	209087_x_at	MCAM	melanoma c	AF089868
-2.95982	4.18E-05	202672_s_at	ATF3	activating tra	NM_001674
-3.07096	2.07E-05	229160_at	MUM1L1	melanoma a	AI967987
-3.28701	0.003959	205347_s_at	TMSB15A /// T	thymosin be	NM_021992
-3.93675	0.001985	227705_at	TCEAL7	transcription	BF591534
-4.15203	0.007862	218469_at	GREM1	gremlin 1, D/	NM_013372
-4.31605	0.012297	218468_s_at	GREM1	gremlin 1, D/	AF154054

4.2.3 Identification of co-expressed and co-regulatory sets of genes

Comparative Toxicogenomics Database generated a Venn diagram of the co-expressed genes among the different experimental designs specifically between Breast_Cancer vs. Breast_Normal and Ovarian_Cancer vs. Ovarian_Normal (Figure 4.6).

Figure 4.6: Co-expressed genes of Breast cancer and Ovarian cancer from Robina studuy using CTD



⊞ Items only in Breast Cancer_Breast Normal (92)

⊞ Items only in Ovarian Cancer_Ovarian Normal (93)

⊟ Items common to Breast Cancer_Breast Normal and Ovarian Cancer_Ovarian Normal (8)

1. 1555827_at
2. 202672_s_at
3. 205469_s_at
4. 210457_x_at
5. 221992_at
6. 222380_s_at
7. 222900_at
8. 231929_at

⊞ All items in Breast Cancer_Breast Normal (100)

4.3 Genespring

4.3.1 Data pre-processing and normalization

The intensity value of each sample was normalized and the distribution of such normalized values was represented in box-whisker plot. The experiment was carried forward with those normalized set of values (Figure 4.7).

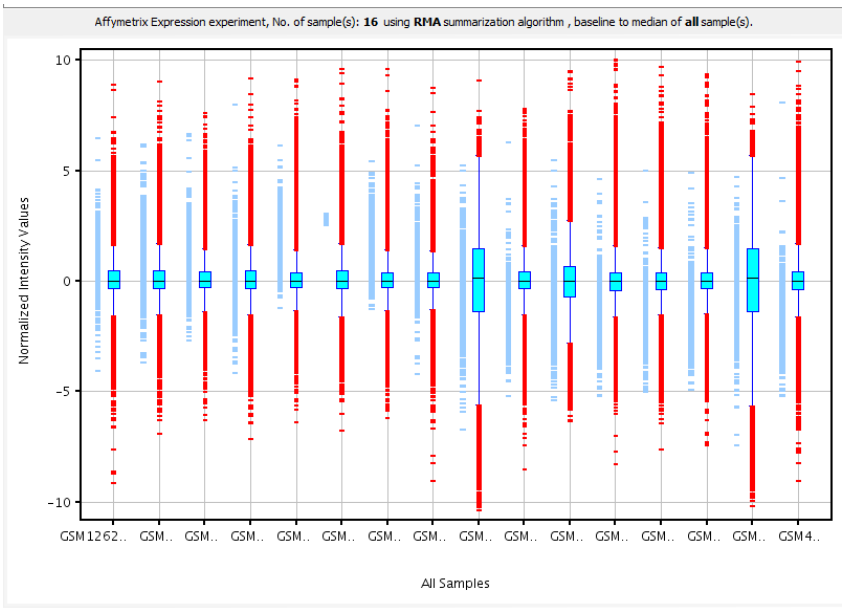


Figure 4.7: Normalized data using Genespring

The grouping structure of the experiment is defined by the experimental parameters. The parameters inserted were: Breast_Cancer; Breast_Normal; Ovarian_Cancer and Ovarian_Normal (Figure 4.8).

Samples	New Parameter
GSM1262125_HCC1954_CD24-CD44+_ALDH+.CEL	Breast_Cancer
GSM1262129_MC1_CD24-CD44+_ALDH+.CEL	Breast_Cancer
GSM1262133_MCF10A_CD24-CD44+_ALDH+.CEL	Breast_Cancer
GSM1262137_MCF7_CD24-CD44+_ALDH+.CEL	Breast_Cancer
GSM1262140_Patient_2_CD24-CD44+_ALDH-.CEL	Breast_Normal
GSM1262141_Patient_2_Rest_ALDH-.CEL	Breast_Normal
GSM1262142_Patient_2_Rest_ALDH+.CEL	Breast_Normal
GSM1262126_HCC1954_Rest_ALDH-.CEL	Breast_Normal
GSM722641_142997-142986_Exp_Pt_1.CEL	Ovarian_Cancer
GSM463966.CEL	Ovarian_Cancer
GSM350609.CEL	Ovarian_Cancer
GSM389785.CEL	Ovarian_Cancer
GSM389789.CEL	Ovarian_Normal
GSM349155.CEL	Ovarian_Normal
GSM722652_142948-142914_Ctrl_Pt_1.CEL	Ovarian_Normal
GSM463969.CEL	Ovarian_Normal

Figure 4.8: Parameters used in Genespring

Sample quality was assessed by examining the values in PCA plot (Figure 4.9). All the samples complied with the quality assessment steps so none was removed from the experiment as two of the samples showing outlier behaviour in Robina were not included in Genespring at all.

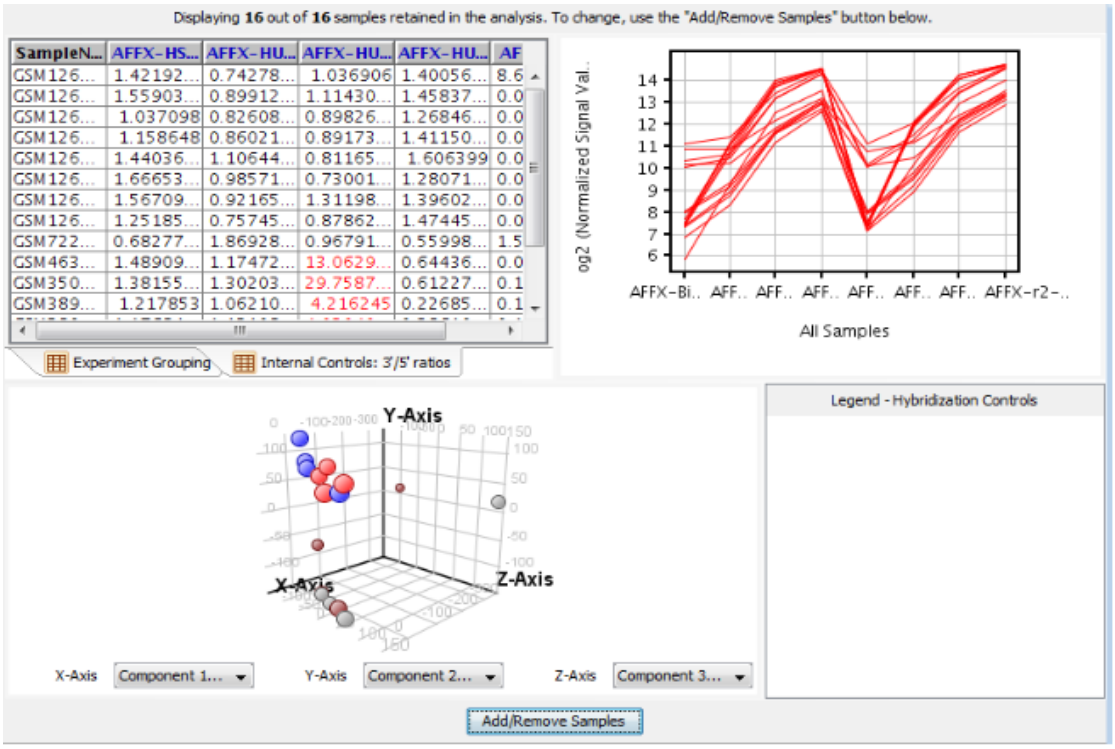


Figure 4.9: Quality control on samples in Genespring

In order to filter the probe sets, entities were filtered based on their signal intensity values (Figure 4.10). Probe sets which had values between 20.0 and 100 percentile were kept and the rest were filtered.

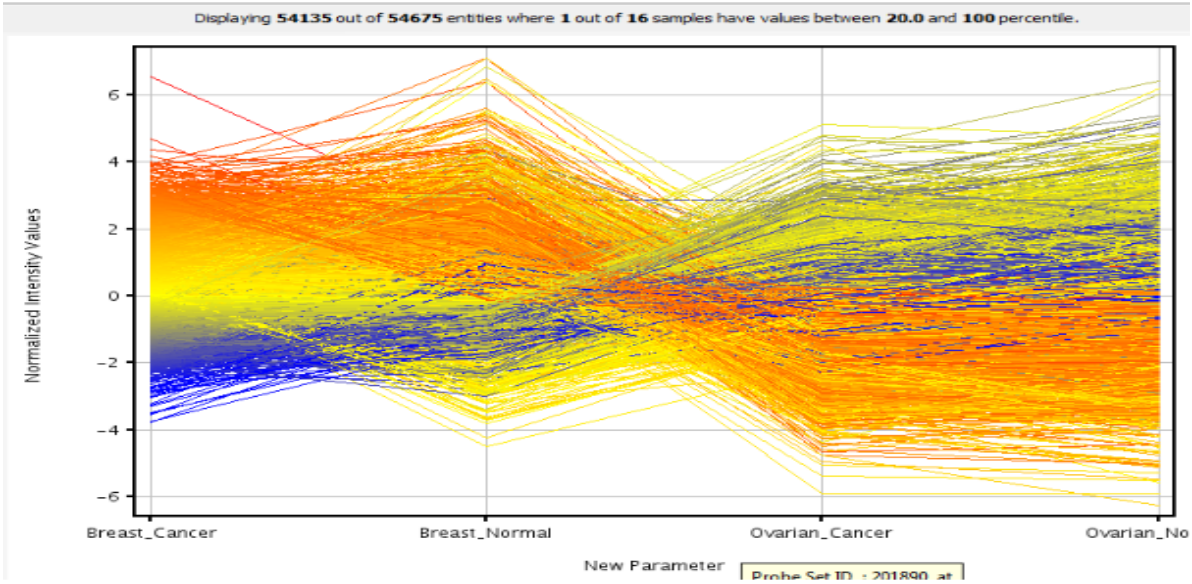


Figure 4.10: Filter probesets in Genespring

4.3.2 Identification of differentially expressed genes (DEGs)

Entities were filtered based on their corrected p-values calculated from statistical analysis (Figure 4.11). Statistical significant p-value cut off 0.05 was taken in this step.

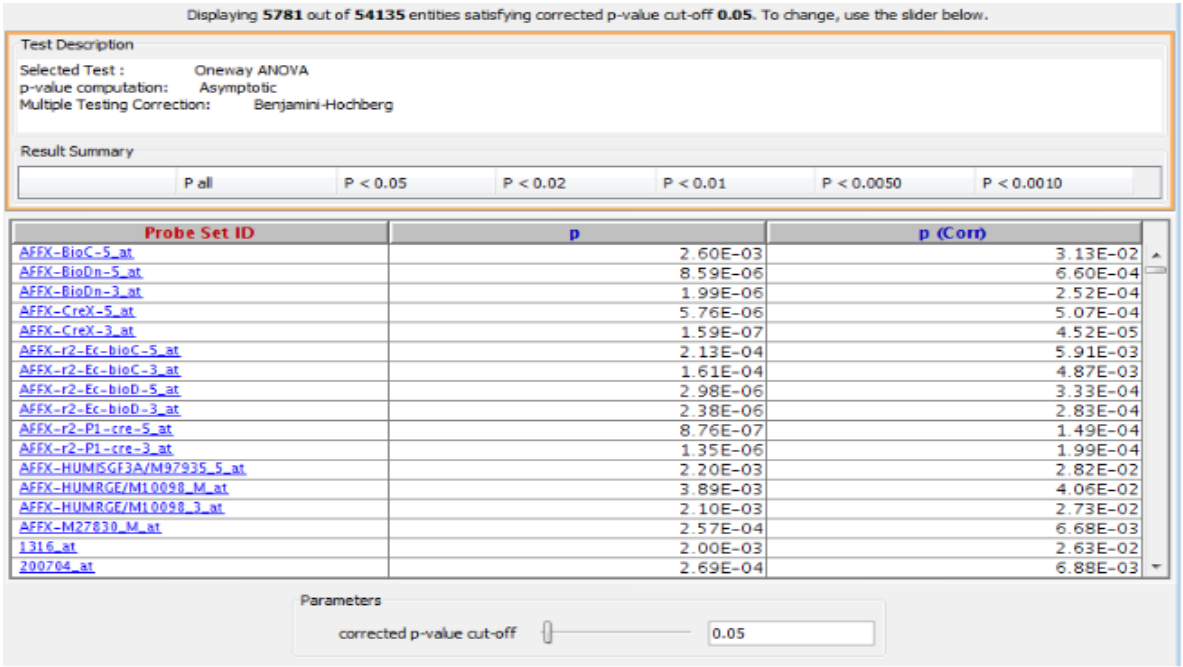


Figure 4.11: Significance analysis in Genespring

A fold change threshold of two was considered in at least one condition pair to select entities for the next step (Figure 4.12).

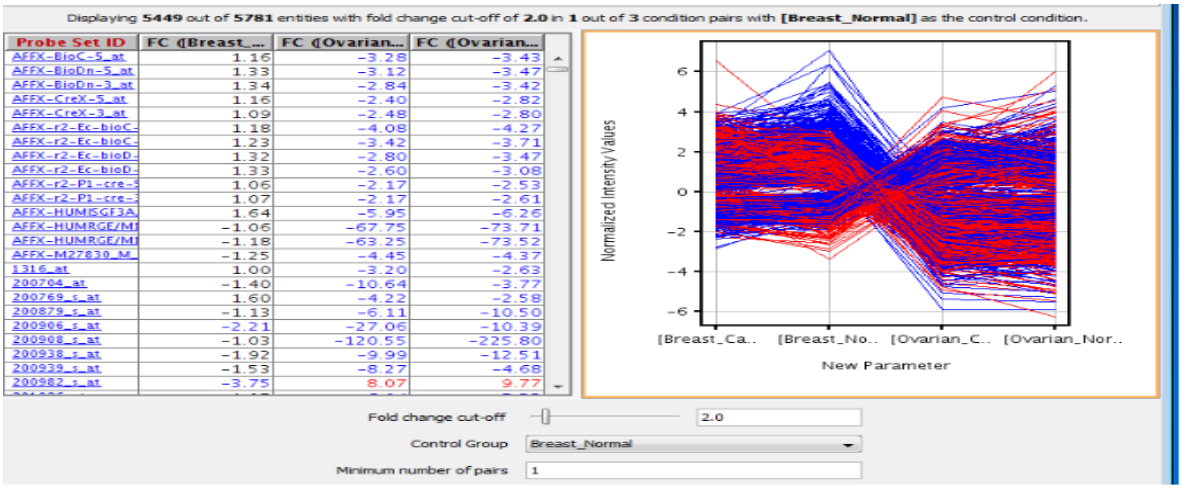


Figure 4.12: Fold change in Genespring

Gene ontology was chosen next in order to unravel the characteristics of genes in three different categories. First, the biological process of genes was studies. Molecular function and cellular component of the genes were revealed too. The saved entity lists contained entities corresponding to the p- value cut-off of 0.05 (Figure 4.13).

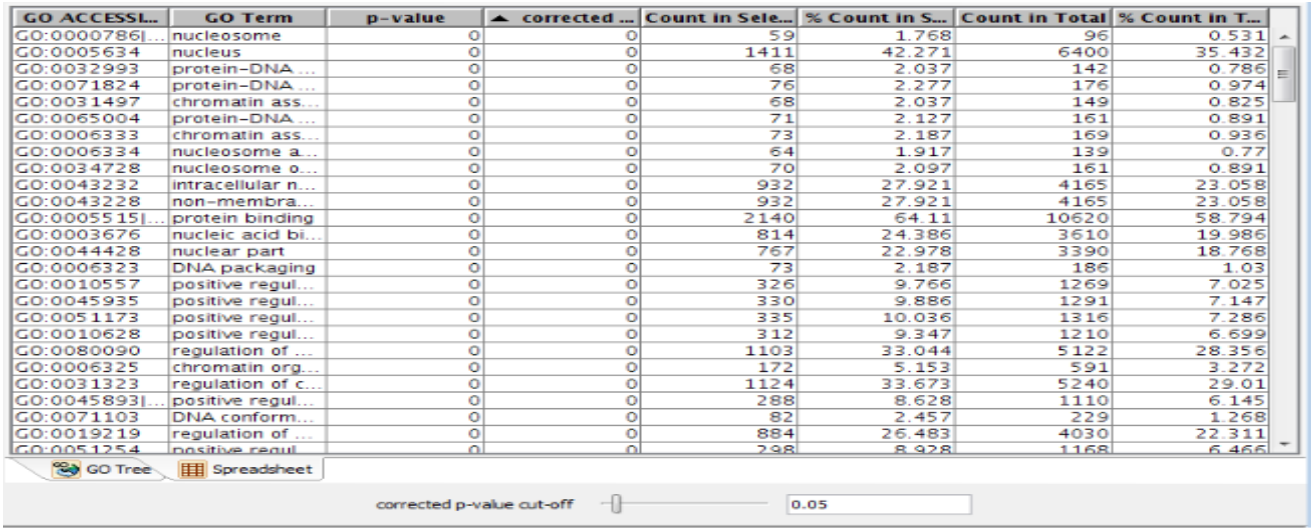


Figure 4.13: GO analysis in Genespring

4.3.3 Identification of co-expressed and co-regulatory sets of genes

Two different experiments were performed, in one breast normal was taken as control and the rest of the groups (viz., breast cancer, ovarian cancer and ovarian normal) were compared with it and in another experiment ovarian normal was taken as control. The up-regulated and down-regulated genes were grouped based on the Fold change values. Next, Venn diagrams were drawn, one for up-regulated 'Breast_Cancer mapped with Breast_Normal' and up-regulated 'Ovarian_Cancer mapped with Ovarian_Normal' (Figure 4.14, 4.15). The co-expressed genes were obtained in tabular format (Table 4.7). Venn diagram was also drawn for down-regulated 'Breast_Cancer mapped with Breast_Normal' and down-regulated 'Ovarian_Cancer mapped with Ovarian_Normal' (Figure 4.16, 4.17). Co-expressed genes were tabulated (Table 4.8). Fold change cut off in case of Genespring was 1.0.

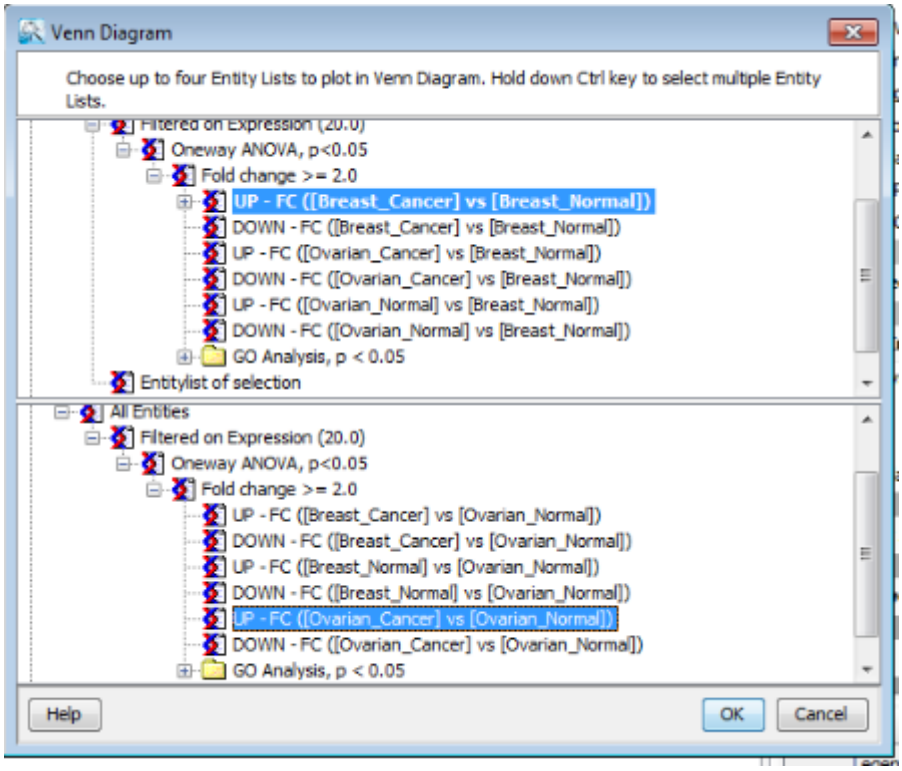


Figure 4.14: Venn diagram parameters for up-regulated genes in Genespring

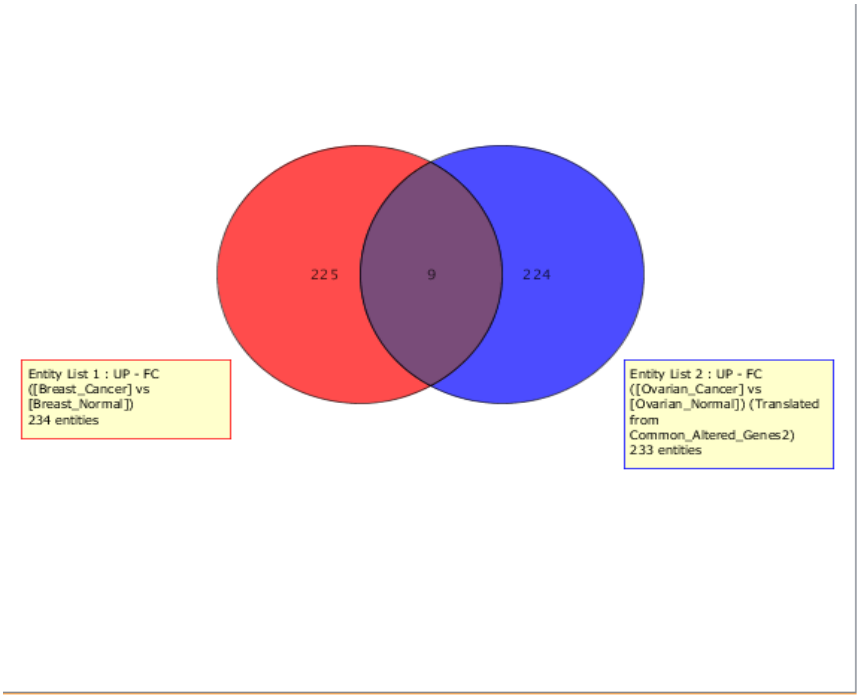


Figure 4.15: Venn diagram output for up-regulated genes in Genespring

Table 4.7: Co-expressed up-regulated genes between breast cancer and ovarian Cancer from Genespring

Probe Set ID	Log FC ([Breast_Can	Log FC ([Ovarian	Gene Symbol	Entrez Ge	Gene Ontology Biological Process
205469_s_at	1.0073632	1.2465585	IRF5	3663	0006351 // transcription, DNA-depen
209173_at	1.5897613	1.3001348	AGR2	10551	0070254 // mucus secretion // inferre
209446_s_at	1.532597	1.2938672			
60815_at	1.7942345	1.1246194	POLR2J4	84820	
231211_s_at	2.0449913	1.0823724	YIF1B	90522	
231929_at	1.087458	1.0462956	IKZF2	22807	0006351 // transcription, DNA-depen
1555827_at	1.1349396	1.2977515	CCNL1	57018	0000079 // regulation of cyclin-deper
1558154_at	1.5593636	1.2733743			
1555827_at	1.1349396	1.2977515	CCNL1	57018	0000079 // regulation of cyclin-deper

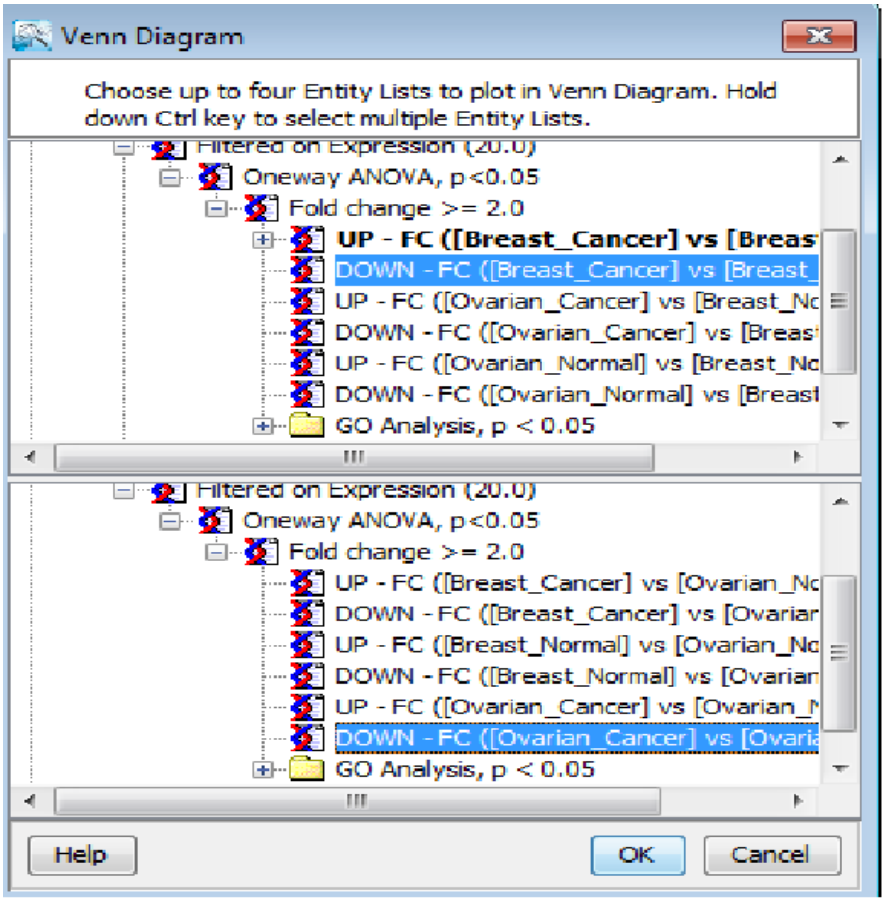


Figure 4.16: Venn diagram parameters for down-regulated genes in Genespring

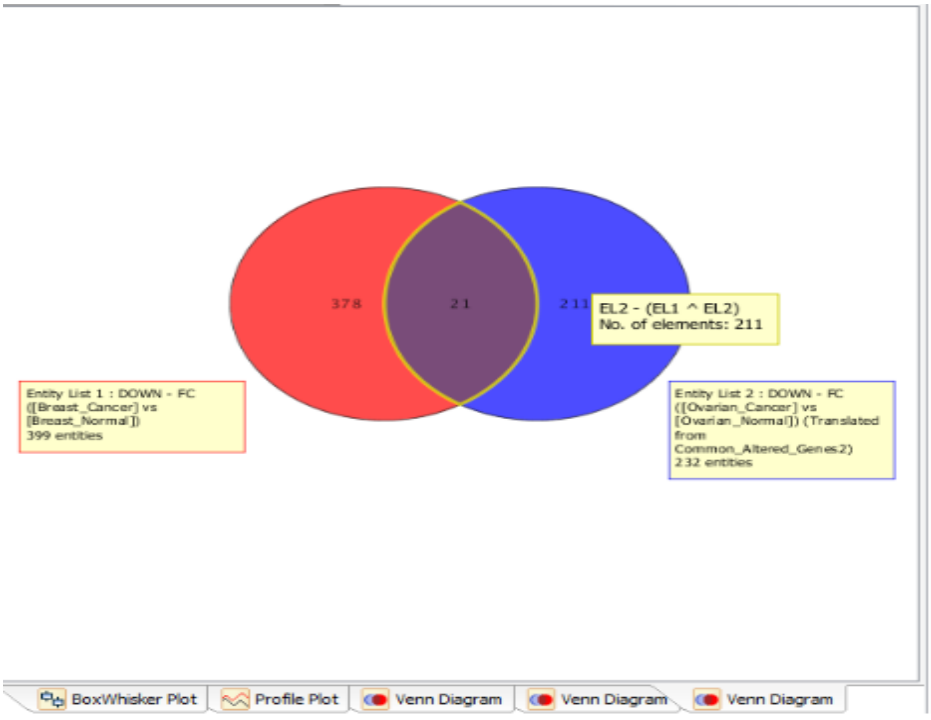


Figure 4.17: Venn diagram output for down-regulated genes in Genespring

Table 4.8: Co-expressed down-regulated genes between breast cancer and ovarian cancer from Genespring

Probe Set	Log FC ([B	Log FC ([O	Gene Sym	Entrez Gene	Gene Ontology Biological Process
200906_s_	-1.14393	-1.38062	PALLD	23022	0007010 // cytoskeleton organizatio
202672_s_	-1.55899	-1.84247	ATF3	467	0006094 // gluconeogenesis // infer
204400_at	-3.49492	-1.9062	EFS	10278	0007155 // cell adhesion // inferred
204560_at	-1.17778	-2.0133	FKBP5	2289	0000413 // protein peptidyl-prolyl is
205379_at	-2.69232	-1.68385	CBR3	874	0008152 // metabolic process // infe
209230_s_	-1.74556	-1.23956	NUPR1	26471	0002526 // acute inflammatory resp
209615_s_	-1.39969	-1.34978	PAK1	5058	0000165 // MAPK cascade // inferre
210180_s_	-1.35846	-1.50405	TRA2B	6434	0000302 // response to reactive oxy
210457_x_	-1.07408	-1.25531	HMGA1	3159	0006268 // DNA unwinding involved
210458_s_	-1.22354	-1.3808	TANK	10010	0007165 // signal transduction // tra
219557_s_	-1.01636	-1.92921	NRIP3	56675	0006508 // proteolysis // inferred fr
222161_at	-2.11741	-1.17873	NAALAD2	10003	0006508 // proteolysis // non-tracea
222528_s_	-1.72506	-1.30228	SLC25A37	51312	0006810 // transport // inferred fror
222900_at	-1.31917	-1.49579	NRIP3	56675	0006508 // proteolysis // inferred fr
227554_at	-1.34788	-2.40222	MAGI2-AS	100505881	
228527_s_	-1.92205	-1.23765	SLC25A37	51312	0044281 // small molecule metaboli
230790_x_	-1.37745	-1.04585			
231274_s_	-1.45256	-1.28668			
231411_at	-1.2446	-1.40736	LHFP	10186	
235267_at	-1.42096	-1.38113	MAGI2-AS	100505881	
236600_at	-1.25716	-1.05529	SPG20	23111	0000910 // cytokinesis // inferred fr

4.4 Identification of common altered genes, functional annotation and pathways

After finding the co-expressed and co-regulatory sets of genes using Robina and Genespring, both the results were compared to find the common altered genes in both the analysis (Table 4.9).

In the process, David database and Panther Database were employed to find the functional classification of the genes individually for both up-regulated and down-regulated genes. The 'binding' functional group was studied to identify the common-altered genes (Figure 4.18, 4.19, 4.20, 4.21).

Pathway enrichment analysis was studied using David database. For the 5 up-regulated gene probes of potential breast and ovarian cancer biomarkers that were deduced using Genespring, one KEGG pathway was found in enrichment analysis:"Toll-like receptor ovarian cancer biomarkers, that were deduced using Genespring, another KEGG pathway

was found in enrichment analysis: “Mitogen-activated protein kinase pathway” (Table 4.10). The number of genes involved in each was 3 (Table 4.11).

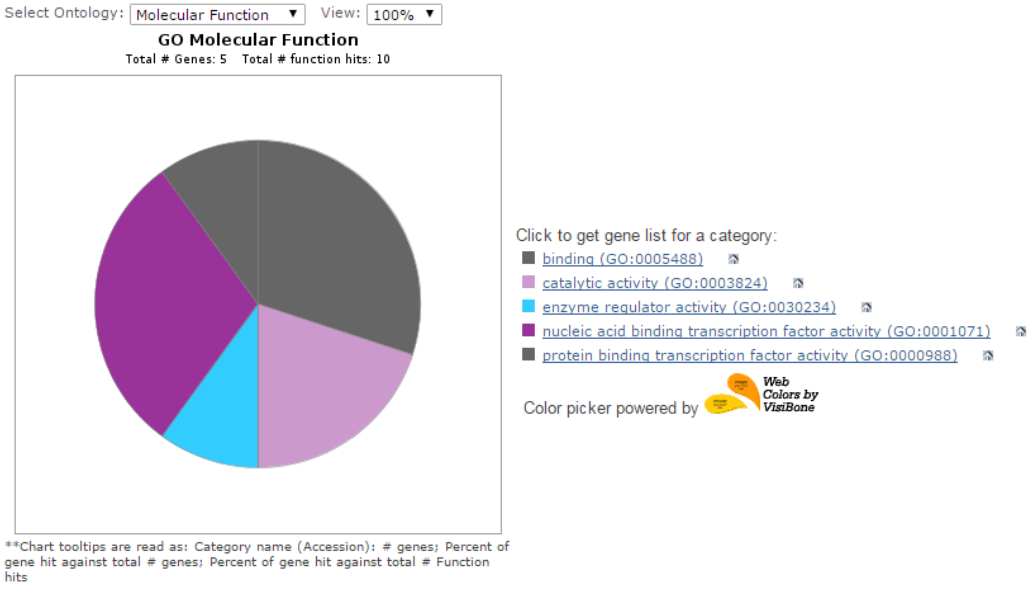


Figure 4.18: Functional classification of up-regulated genes viewed in pie-chart in Panther to identify common altered genes

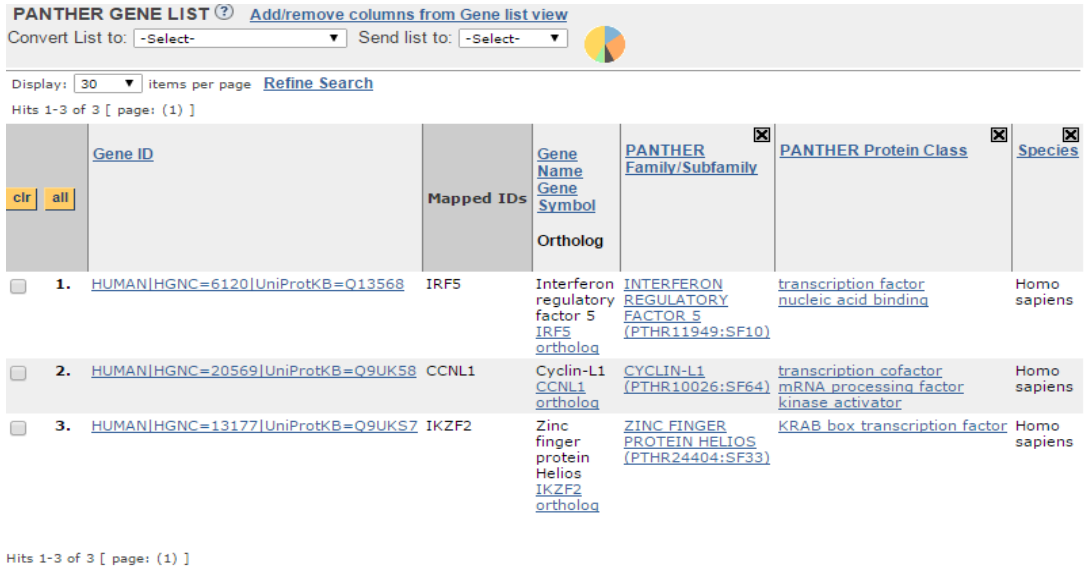


Figure 4.19: Panther up-regulated gene list of “binding” functional annotation

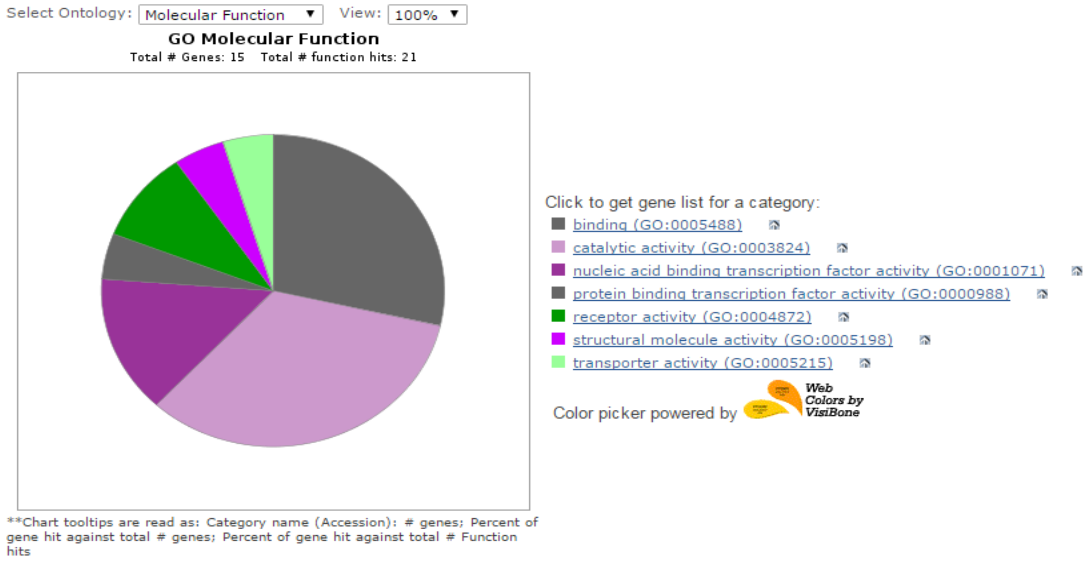
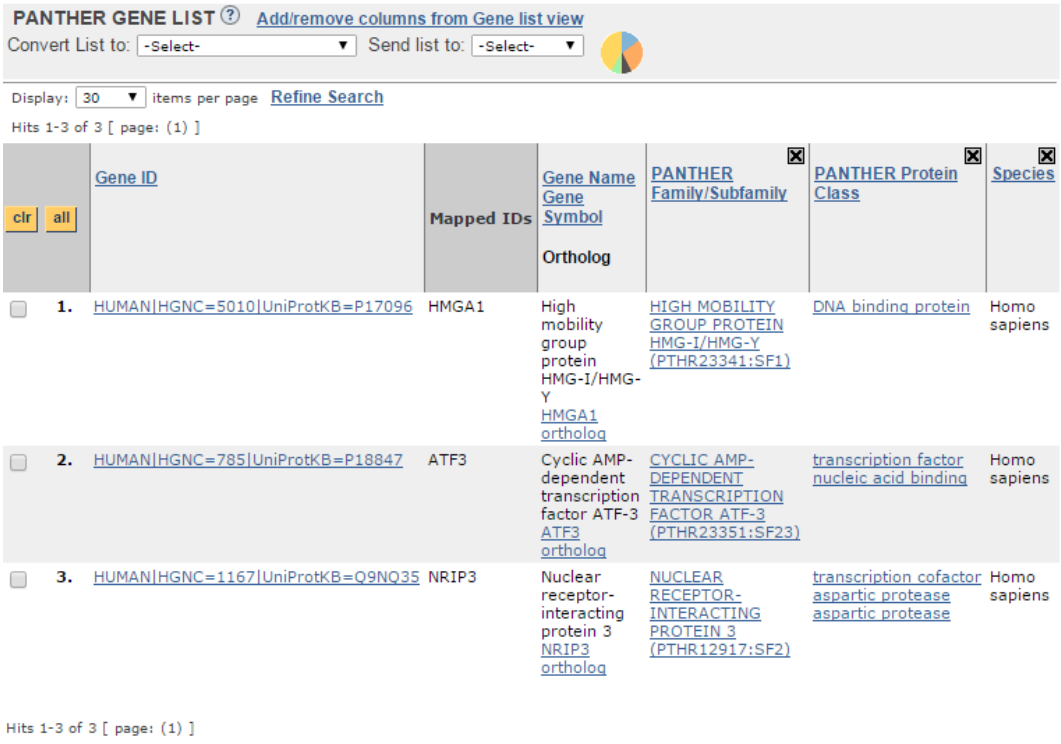


Figure 4.20: Functional classification of down-regulated genes viewed in pie-chart in Panther to identify common altered genes



PANTHER GENE LIST Add/remove columns from Gene list view

Convert List to: -Select- Send list to: -Select-

Display: 30 items per page Refine Search

Hits 1-3 of 3 [page: (1)]

clr all		Gene ID	Mapped IDs	Gene Name Gene Symbol Ortholog	PANTHER Family/Subfamily	PANTHER Protein Class	Species
☐	1.	HUMAN|HGNC=5010|UniProtKB=P17096	HMGA1	High mobility group protein HMG-I/HMG-Y HMGA1 ortholog	HIGH MOBILITY GROUP PROTEIN HMG-I/HMG-Y (PTHR23341:SF1)	DNA binding protein	Homo sapiens
☐	2.	HUMAN|HGNC=785|UniProtKB=P18847	ATF3	Cyclic AMP-dependent transcription factor ATF-3 ATF3 ortholog	CYCLIC AMP-DEPENDENT TRANSCRIPTION FACTOR ATF-3 (PTHR23351:SF23)	transcription factor nucleic acid binding	Homo sapiens
☐	3.	HUMAN|HGNC=1167|UniProtKB=Q9NQ35	NRIP3	Nuclear receptor-interacting protein 3 NRIP3 ortholog	NUCLEAR RECEPTOR-INTERACTING PROTEIN 3 (PTHR12917:SF2)	transcription cofactor aspartic protease aspartic protease	Homo sapiens

Hits 1-3 of 3 [page: (1)]

Figure 4.21: Panther down-regulated gene list of "binding" functional annotation

Table 4.9 Pathway analysis of the up-regulated genes between breast cancer and ovarian cancer using David

#	Gene	Species	Kappa
1	hypothetical LOC494760	Xenopus laevis	1.0
2	interferon regulatory factor 5	Danio rerio	0.9411729182173478
3	interferon regulatory factor 5	Gallus gallus	0.9142803715820024
4	Interferon regulatory factor 5	Salmo salar	0.9090856062340688
5	interferon regulatory factor 5	Bos taurus	0.849989412715865
6	interferon regulatory factor 5	Mus musculus	0.8292560016214627
7	**interferon regulatory factor 5**	**Homo sapiens**	**0.8095098415165741**
8	interferon regulatory factor 5	Xenopus silurana	0.7142726196392726
9	cyclin L1	Danio rerio	0.2631081267438044
10	IKAROS family zinc finger 2	Mus musculus	0.2534496151415429
11	**IKAROS family zinc finger 2**	**Homo sapiens**	**0.2534496151415429**
12	**cyclin L1**	**Homo sapiens**	**0.21733017091838225**
13	cyclin L1	Rattus norvegicus	0.21270362152867422
14	cyclin L1	Mus musculus	0.20826983575362146
15	IKAROS family zinc finger 2	Rattus norvegicus	0.18598578159603424
16	IKAROS family zinc finger 2	Gallus gallus	0.15377515336074796

Table 4.10: Pathway analysis of the down-regulated genes between breast cancer and ovarian cancer using David

#	Gene	Species	Kappa
1	p21 protein (Cdc42/Rac) -activated kinase 1	Homo sapiens	1.0
2	p21 protein (Cdc42/Rac) -activated kinase 1	Rattus norvegicus	0.5302384392329444
3	nuclear Receptor Interacting protein 3	Homo sapiens	0.44225350432604216
4	p21 protein (Cdc42/Rac) -activated kinase 1	Bos taurus	0.39978509381613203
5	activating transcription factor 3	Homo sapiens	0.38399026838766287
6	p21/Cdc42/Rac1-activated kinase 1	Danio rerio	0.38399026838766287
7	activating transcription factor 3	Xenopus laevis	0.38399026838766287
8	high Mobility group AT-hook 1	Homo sapiens	0.38399026838766287

Table 4.11: Pathway analysis comparison of the DEGs

Regulation	Category	Terms	Genes	Count	Percentage
UP-Regulated	KEGG_Pathway	Toll-like Receptor Pathway	*IRF5, CCNL1, IKZF2*	3	18.5
Down-Regulated	KEGG_Pathway	MAPK Pathway	*NRIP3, HMGA1, ATF3*	3	10

A complete comparison of the fold-change values of the co-expressed genes using both the tools has been presented in a tabular format (Table 4.12).

Table 4.12 Comparison of Fold-Change Values of the co-expressed genes using both the tools

ID	Gene Symbol	Gene Title	Log Fc Value in Robina		Regulation in Robina		Log Fc Value in Genespring		Regulation in Genespring	
			BC_BN	OC_ON			BC_BN	OC_ON		
1555827_at	*CCNL1*	cyclin L1	1.75463	1.598335	UP	UP	1.134939 6	1.297751 5	UP	UP
231929_at	*IKZF2*	IKAROS family zinc finger 2 (Helios)	3.2345	1.64882	UP	UP	1.087458	1.046295 6	UP	UP
205469_s_at	*IRF5*	interferon regulatory factor 5	3.588418	1.530273	UP	UP	1.007363 2	1.246558 5	UP	UP
202672_s_at	*ATF3*	activating transcription factor 3	-4.73385	-2.95982	DOWN	DOWN	-1.55899	-1.84247	DOWN	DOWN
222900_at	*NRIP3*	nuclear receptor interacting protein 3	-4.32458	-2.33957	DOWN	DOWN	-1.31917	-1.49579	DOWN	DOWN
210457_x_at	*HMGA1*	high mobility group AT-hook 1	-4.08195	-2.41583	DOWN	DOWN	-1.07408	-1.25531	DOWN	DOWN

4.5 Online validation of common altered genes in both breast and ovarian cancers

GOBO online tool was used to study the association with outcome for gene sets in a sample breast cancer datasets (Figure 4.22). The six probes, deduced from the above experiments,

matched with the breast cancer datasets present in GOBO server.

```
## Parameters ##
Nbr of groups:
2
Data set selection:
all
Input type:
GeneSymbol
Censoring (Years):
10
endPoint:
DMFS
Cut definition:
complete
## End Parameters ##
Gene/Probe ID    matched
CCNL1    TRUE
IKZF2    TRUE
IRF5     TRUE
ATF3     TRUE
NRIP3    TRUE
HMGA1    TRUE
```

Figure 4.22: *In silico* validation result from GOBO web server for breast cancer altered genes

Chapter 5
Conclusion

The experiment started with 167 samples of breast and ovarian cancer malignancies and after consequent literature survey, eighteen high grade samples of breast and ovarian cancers, normal breast and ovarian samples were put for a thorough study in Robina and Genespring.

Stringent quality checks were applied, removing probes whose intensities were less than twenty percentile and accepting probes whose coefficient of variation (CV) was less than fifty percent.

The DEGs analysis was carried out using Robina and Genespring by applying strict threshold values of 0.05 for P value and 1.5 for fold change. The highly expressed DEGs list showed a definite pattern of gene expression wherein the roles of few genes were pre-established in breast and ovarian cancers, few novel genes not reported in the pertaining literature, were found to be present and few common altered genes in breast and ovarian malignancies were reported. The top DEGs list contained *TOP2A, CCNA2, BIRC5, BUB1*, and *KIF4A*, which were already known to play roles in breast cancer malignancy. Further, the list contained *FAM179A, CBX5, HHLA1, and GATA4,* which were known to be involved with ovarian cancer.

IKZF2, NRIP3, which were never reported to be associated with breast cancer were found in the list of common altered genes. *IKZF2, CCNL1, NRIP3* of the list were never reported to be associated with ovarian cancer previously. These can be potential biomarkers for breast and ovarian cancers respectively, which can only be confirmed after further stringent benchmarking.

The genes that were found to be up-regulated in both breast and ovarian cancers included *IKZF2, CCNL1* and *IRF3*, and the down-regulated ones included *NRIP3, HMGA1* and *ATF3*. These are the six common altered genes in breast and ovarian cancers that were identified by meta-analysis of the raw data of both these cancers.

Nonetheless identifying biomarkers among these requires serious benchmarking and wet lab studies.

Chapter 6
Discussion and Future Perspective

The CEL files across eight different laboratories were associated and the same normalization method was applied in order to pre-process the data. The reason for assembling data from different laboratories was to increase the number of overall samples for consideration.

Data accuracy increased with this approach of integrative analysis and one of the benefits included more easy detection of any significant differential expression. It was very likely that a particular gene may not have been showing noteworthy differential expression in a particular laboratory's data but the same gene could have been showing differential expression in another laboratory. The present work supports this standpoint.

Among the common altered genes between breast and ovarian cancers, studies already showed the expression pattern of *CCNL1, IRF5, HMGA1* and *ATF3* to be linked to breast cancer and *IRF5, HMGA1* and *ATF3* to be associated with ovarian cancer. The gene, which was commonly up-regulated in both breast and ovarian cancers and was reported separately to be associated with both of these cancers, was *IRF5*. Equivalently, in this study, the combined data revealed the up-regulation of *IRF5* in both the cancers. In the same way, commonly down-regulated genes included *HMGA1* and *ATF3* and were individually reported. Correspondingly, in this study, the combination of the data showed the down-regulation of these two genes in both of these cancers.

Further, a couple of reported common altered genes in breast and ovarian cancers, *IKZF2* and *NRIP3*, were never reported previously to have any sort of link with either breast cancer or ovarian cancer. The present study confirmed the role of these novel genes as potential biomarkers in both these cancers.

Literature survey showed a fine balance between Cyclin L1 (*CCNL1*) and tissue inhibitor of matrix metalloproteinase-1 (*TIMP1*) contributing to the development of breast cancer cells. *In vitro* experiments showed a stimulatory effect of *TIMP1* and *CCNL1* on growth of MDA-MB-231 cells, a particular breast-cancer cell line. Co-expression or co-repression of these two genes did not affect cell growth. But then again, over-expression of *CCNL1* and *TIMP1* individually induced overexpression of each other. These data demonstrated a fine balance between *CCNL1* and *TIMP1*, which might contribute to breast cancer development. (Peng *et al.,* 2011). So it is very likely that up-regulation of *CCNL1* has a profound role in breast cancer development.

IRF5 has established role in regulation of cell motility, invasive action and in maintenance of equilibrium inside cell. Pimenta *et al.* (2015) hypothesized that IRF5 may not be transcription driven as its expression is predominantly found in mammary epithelial cells of human (Pimenta *et al.,* 2015).

Pegoraro *et al.* (2013), reported that the *HMGA1* protein has a role in breast cancer cells but the exact mechanism of action was not explained (Pegoraro *et al.,* 2013).

ATF3 gene is shown to be induced by many signals which include some of those involved in cancerous cells and further, it codes a member transcription factor linked to mammalian action (Norman EB 2013). There are evidences suggesting that *ATF3* is involved in apoptosis of cells thus, ceasing any tumor formation. Particularly, it is an established fact that tumor suppression is not possible in colorectal cancer on removal of *ATF3* gene (Huang *et al.,* 2008). Further, the metastasis of ovarian cancer cells increased and cell death was possible in ovarian cancer cells due to *ATF3* up-regulation (Huang *et al.,* 2008).

In the process, David database and Panther database were used to find the functional classification of the genes individually for up-regulated and down-regulated genes. It was noted that all the common altered genes between breast and ovarian cancer that were deduced by comparing the result of Robina and Genespring viz., *IKZF2, CCNL1, IRF5, NRIP3, HMGA1* and *ATF3* belonged to the same protein class, namely "binding class".

The pathway analysis of the 5 probes of up-regulated genes, which were deduced from Genespring, in David found one significant pathway: "toll-like receptor signalling pathway". A significant number of studies have reported the function of TLRs in invasive action and metastasis of cancerous growths. Further, TLRs put up a resistance to apoptosis, whose mechanism is not completely defined. However, few studies suggested the involvement of TLRs in tumorogenesis. Pro-inflammatory factors are produced on TLR signalling activation which causes "immune evasion" through "cytotoxic lymphocyte attack" (Huang *et al.,* 2008).

So, the association of the up-regulated differentially expressed genes of breast and ovarian cancers with TLR signalling pathway further supports the up-regulatory activity of these genes.

The pathway analysis of the 15 probes of down-regulated genes, which were deduced from Genespring, in David found one significant pathway: "Mitogen-activated protein kinase pathway". MAPKs have dual activity in cancerous cells and it is difficult to reach at a consensus about the role of MAPKs in tumor cells. But the strength of activation of this pathway decides whether it is harmful for the cells or not. Nonetheless, the pathway is very much involved with operational activity of tumor cells (Dhillon *et al.,* 2007).

The common identified altered genes, viz., *IKZF2, CCNL1, IRF5, NRIP3, HMGA1,* and *ATF3,* were given as input in GOBO and the output included all six, which showed strong enrichment from breast cancer expression data. This was an *in silico* validation step. Further benchmarking and wet-lab validation will confirm the function of these genes as biomarkers.

The application of microarray technology to breast and ovarian cancers has provided a molecular basis for grade of tumors. DEGs identified pathways have provided new realm to the potential treatment of these cancers. This knowledge will help in precise diagnosis and judicious treatment for breast and ovarian cancers.

The accessibility of publicly available microarray databases has made the whole process of microarray data analysis cost effective and has expunged the practical limitations associated

with the undersized biological samples. The best part is increase in the diversity of data due to consideration of samples from a range of research groups (Farley *et al.*, 2008).

Instead of combining data from different platforms belonging to the same research group, combining data on same microarray platform from a variety of research groups decrease the chance of any technical glitch (Dennis *et al*., 2003). In this work, the datasets from Affymetrix platform for breast and ovarian cancers were chosen for meta-analysis and integrative analysis by combining both cancer data sets including normal, high and low grade ovarian cancer profiles. Further, the use of different softwares for the microarray data analysis increased the degree of data accuracy and the genes were predicted more securely. Integrative microarray analysis is a competent way of finding biomarker genes and in future, such data integration studies implicates great potential for cancer treatment.

Comprehensively, usage of integrative analysis for both breast and ovarian cancers evaluated the global gene expression data. The combined analysis also describes those genes which shares common expression pattern in breast and ovarian cancers including different subtypes of cancer group. The identified common altered genes in both breast and ovarian cancer can be taken as potential and prognostic biomarker genes for the early detection of cancer after further standardization and subsequent wet-lab validation.

References

Abbott, KL. (2010). Identification of candidate biomarkers with cancer-specific glycosylation in the tissue and serum of endometrioid ovarian cancer patients by glycoproteomic analysis. Proteomics. ***10***, 470-481.

Ahmad, S. (2011). Advances in ovarian cancer screening: health and medicine for women: a multidisciplinary, evidence-based review of mid-life health concerns. Yale Journal of Biology and Medicine. ***84***, 47-49.

Anders, S; Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology. ***11(10)***, R106.

Bahcall, O. (2013). Common variation and heritability estimates for breast, ovarian and prostate cancers. Nature Genetics. ***35***, 23-25.

Barrett, T; Troup, DB; Wilhite, SE; Ledoux, P; Rudnev, D; Evangelista, C; Kim, IF; Soboleva, A; Tomashevsky, M; Marshall, KA. (2009). NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Research. ***37***, D885–D890.

Bateman, A; Birney, E; Cerruti, L; Durbin, R; Etwiller, L; Eddy, SR; Griffiths-Jones, S; Howe, KL; Marshall, M; Sonnhammer, EL. (2002). The Pfam protein families database. Nucleic Acids Research. ***30***, 276–280.

Bayli, SB; Ohm, JE. (2006). Epigenetic gene silencing in cancer - mechanism for early oncogenic pathway addiction. Nature Reviews Cancer. ***6***, 107-17.

Berger, AH; Pandolfi, PP. (2011). Cancer Susceptibility Syndromes. Principles and Practice of Oncology. ***8***, 161–172.

Buscaglia, LE; Li, Y. (2011). Apoptosis and the target genes of microRNA-21. Chinese Journal of Cancer. ***30(6)***, 371-380.

Cairns, J. (1982). Aging and cancer as genetic phenomena. National Cancer Institute. ***60***, 237–239.

Cho, WSC. (2007). Contribution of oncoproteomics to cancer biomarker discovery. Molecular Cancer. ***6***, 25.

Conley, SJ; Gheordunescu, E; Kakarala, P; Newman, B; Korkaya, H; Heath, AN; Clouthier, SG; Wicha, MS. (2012). Antiangiogenic agents increase breast cancer stem cells via the generation of tumor hypoxia. Proceedings of the National Academy of Sciences. ***109***, 2784–2789.

Dennis, G; Sherman, BT; Hosack, DA. (2003). DAVID: Database for annotation, visualization, and integrated discovery. Genome Biology. ***4(5)***, 3.

Dhillon, AS; Hagan, S; Rath, O; Kolch, W. (2007). MAP kinase signalling pathways in cancer. Nature Genetics. ***26***, 3279–3290.

Egger, G; Liang, G; Aparicio, A; Jones, PA. (2004). Epigenetics in human disease and prospects for epigenetic therapy. Nature. ***429***, 457-63.

Eisen, MB; Spellman, PT; Brown, PO.(2000). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences. ***95(25)***, 14863-14868.

Ermolaeva, O; Rastogi, M; Pruitt, KD. (1998). Data management and analysis for gene expression arrays. Nature Genetics. ***20***, 19-23.

Esquela, KA; Slack, FJ. (2006). Oncomirs - microRNAs with a role in cancer. Nature Reviews Cancer. ***6(4)***, 259-69.

Farley, J; Ozbun, L; Birrer, MJ. (2008). Genomic analysis of epithelial ovarian cancer. Cell Research. ***18(5)***, 538-548.

Ferlay, J; Shin, HR; Bray, F. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. International Journal of Cancer. ***127(12)***, 2893–2917.

Filipowicz, W; Bhattacharyya, SN; Sonenberg, N. (2008). Mecha-nisms of post-transcriptional regulation by microRNAs. Nature Reviews Genetics. ***9(2)***, 102-114.

Gray, JW; Collins, C. (2000). Genome changes and gene expression in human solid tumors. Carcinogenesis. ***21***, 443-452.

Gyorffy, B; Lanczky, A; Eklund, AC; Denkert, C; Budczies, J. (2009). An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. Breast Cancer Research and Treatment. ***123***, 725–731.

Harvell, DM; Kim, J; O'Brien, J; Tan, AC. (2013). Genomic signatures of pregnancy-associated breast cancer epithelia and stroma and their regulation by estrogens and progesterone. Hormones and Cancer. ***4(3)***, 40-53.

Hanahan, D; Weinberg, RA. (2000). The hallmarks of cancer. Cell. ***100***, 57-70.

Huang, B; Zhao, J; Unkeless, JC; Feng, ZH; Xiong, H. (2008). TLR signaling by tumor and immune cells: a double-edged sword. Oncogene. ***27***, 218–224.

Huang, DW; Sherman, BT; Lempicki, RA. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols. ***4***, 44 – 57.

Huang, X; Li, X; Guo, B. (2008). KLF6 induces apoptosis in prostate cancer cells through up-regulation of ATF3. Journal of Biological Chemistry. ***283***, 29795–29801.

Jain, N; Thatte, J; Braciale, T; Ley, K; O'Connell, M; Lee, JK. (2003). Local pooled error test for Identifying Differentially Expressed genes with a small number of replicated microarrays. Oxford Journals Bioinformatics. ***19***, 1945-1951.

Jemal, A; Bray, F; Center, MM; Ferlay, J; Ward, E; Forman, D. (2011). Global cancer statistics. CA: a cancer journal for clinicians. ***61(2)***, 69-90.

Karn, T; Metzler, D; Ruckhaberle, E; Hanker, L; Gatje, R. (2010). Data-driven derivation of cutoffs from a pool of 3,030 Affymetrix arrays to stratify distinct clinical types of breast cancer. Breast Cancer Research and Treatment. ***120***, 567–579.

Kim, JH; Skates, SJ; Uede, T; Wong,KK; Schorge, JO; Feltmate, CM. (2002). Osteopontin as a potential diagnostic biomarker forovarian cancer. The Journal of the American Medical Association. ***287***, 1671–9.

Knudson, AG. (1989) The genetic predisposition to cancer. Birth defects original article series. ***25***, 15–27.

Li, H; Liang, S. (2009). Local network topology in human protein interaction data predicts functional association. Public Library of Sciences One. ***4(7)***, e6410.

Liu, S; Cong, Y; Wang, D; Sun, Y. (2014). Breast cancer stem cells transition between epithelial and mesenchymal states reflective of their normal counterparts. Stem Cell Reports. ***2(1)***, 78-91.

Luciani, MG; Seok, J; Sayeed, A; Champion, S. (2011). Distinctive responsiveness to stromal signaling accompanies histologic grade programming of cancer cells. Public Library of Sciences One. ***6(5)***, e20016.

Ludwig, JA; John, N. (2005). Weinstein biomarkers in cancer staging, prognosis and treatment selection. Nature Reviews Cancer. ***5***, 845-56.

Marc, L; Anthony, MB; Axel, Nagel; Alisdair, RF; John, EL; Mark, S; Biorn, U. (2012). *R*obina: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Research. ***40***, W622-W627.

Margeli, M; Cirauqui, B; Castella, E; Tapia, G; Costa, C; Gimenez, CA; Barnadas, A; Sanchez, RM; Benlloch, S; Taron, M; Rosell, R. (2010). The prognostic value of BRCA1 mRNA expression levels following neoadjuvant chemotherapy in breast cancer. Public Library of Science. ***5(3)***, e9499.

Norman, EB. (2013). ATF3 rSNPs, transcriptional factor binding sites and human etiology. Hormones and Cancer. ***3***, 253-261.

Ohta, S; Shimekake, Y; Nagata, K. (1997). Molecular cloning and characterization of a transcription factor for the C-type natriuretic peptide gene promoter. European Journal of Biochemistry. ***242 (3)***, 460–6.

Pegoraro, S; Ros, G; Piazza, S; Sommaggio, R; Ciani, Y; Rosato, A; Sgarra, R; Del, SG; Manfioletti, G. (2013). HMGA1 promotes metastatic processes in basal-like breast cancer regulating EMT and stemness. Hormones and Cancer. ***4(8)***, 1293-308.

Peng, L; Yanjiao, M; Ai-guo, W; Pengtao, G; Jianhua, L; Ju, Y; Hongsheng, O; Xichen, Z. (2011). A fine balance between CCNL1 and TIMP1 contributes to the development of breast cancer cells. Biochemical and biophysical research communications. ***409(2)***, 344-9.

Perou, CM; Sorlie, T; Eisen, MB. (2000). Molecular portraits of human breast tumours. Nature. ***406 (6797)***, 747-52.

Pimenta, EM; Barnes, BJ. (2015). A conserved region within interferon regulatory factor 5 controls breast cancer cell migration through a cytoplasmic and transcription-independent mechanism. Molecular Cancer. ***14***, 32.

Ponder, BA. (2001). Cancer genetics. Nature. ***411***, 336-341.

Pradhan, M; Pal, T. (2010). Gross genomic alterations and gene expression profiles of high grade serous carcinoma of the ovary with and without BRCA1 inactivation. BMC Cancer. ***10***, 493.

Quackenbush, J. (2001). Computational analysis of microarray data. Nature Reviews Genetics. ***2***, 418-427.

Rousseau, A; Badoual, C. (2012). Head and Neck: Squamous cell carcinoma: an overview. Atlas of Genetics and Cytogenetics in Oncology and Haematology. ***2***, 145-155.

Sawyers, CL. (2008). The cancer biomarker problem. Nature. ***452***, 548-52.

Shields, PG; Harris, CC. (1991). Molecular epidemiology and the genetics of environmental cancer. JAMA. ***266***, 681–687.

Soegaard, M; Frederiksen, K; Jensen, A; Høgdall, E; Høgdall, C; Blaakaer, J; Ramus, SJ; Gayther, S A; Kjaer, SK. (2009). Risk of ovarian cancer in women with first-degree relatives with cancer. Acta Obstetricia at Gynecologica Scandinavica. ***88 (4)***, 449–456.

Srinivas, PR; Kramer, BS; Srivastava S. (2001). Trends in biomarker research for cancer detection. The Lancet Oncology. ***2***, 698-704.

Syed , V; Mukherjee, K; Lyons-Weiler, J; Lau, KM; Mashima, T; Tsuruo, T; Ho, SM. (2005). Identification of ATF-3, caveolin-1, DLC-1, and NM23-H2 as putative antitumorigenic, progesterone-regulated genes for ovarian cancer cells by gene profiling. Oncogene. ***24***, 1774–1787.

Tung, CS; Mok, SC; Tsang, YT; Zu, Z. (2009). PAX2 expression in low malignant potential ovarian tumors and low-grade ovarian serous carcinomas. Molecular Pathology. ***22(9)***, 1243-1250.

Tusher, VG; Tibshirani, R; Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences. ***98***, 5116-5121.

Vispe, S; Yung, T; Ritchot, M; Serizawa, J; Satoh, MS. (2000). A cellular defense pathway regulating transcription through poly(ADP-ribosyl)ation in response to DNA damage. Proceedings of the National Academy of Sciences. ***97***, 9886–9891.

Wang, E; Lichtenfels, R; Bukur, J. (2004). Ontogeny and oncogenesis balance the transcriptional profile of renal cell cancer. Cancer Research. ***64***, 7279–87.

Wong, KK; Cheng, RS; Mok, SC. (2001). Identification of differentially expressed genes from ovarian cancer cells by MICROMAX cDNA microarray system. Biotechniques. ***30***, 670–5.

Yasuhoro, H; Akiko, O; Hisao, S; Shingo, H; Takahiro, Y; Takuya, Koie; Chikara, O. (2013). Gene Expression Changes in Venous Segment of Overflow Arteriovenous Fistula. International journal of Nephrology. ***2013***, 16-24.

Appendices

#1 TOP 100 DEGs when Breast Cancer data is compared to Breast Normal

logFC	P.Value	ID	Gene Symbol	Gene Title	Representative Public Id
5.560972	0.000122	201291_s_at	TOP2A	topoisomerase (DNA)	AU159942
4.968754	0.000268	1557094_at	LOC100996760	uncharacterized LOC1	BC029890
4.880495	0.00022	218542_at	CEP55	centrosomal protein 5	NM_018131
4.746589	2.38E-05	212022_s_at	MKI67	marker of proliferatic	BF001806
4.709012	2.15E-05	203418_at	CCNA2	cyclin A2	NM_001237
4.463984	1.25E-05	202095_s_at	BIRC5	baculoviral IAP repea	NM_001168
4.35521	3.20E-07	1555826_at	BIRC5 /// EPR-	baculoviral IAP repea	BQ021146
4.200034	4.47E-05	218355_at	KIF4A	kinesin family memb	NM_012310
3.936971	5.49E-05	215509_s_at	BUB1	BUB1 mitotic checkpo	AL137654
3.855997	6.06E-05	231534_at			
3.783167	3.90E-05	229490_s_at			
3.679598	0.000249	216228_s_at	WDHD1	WD repeat and HMG-	AK001538
3.588418	5.70E-05	205469_s_at	IRF5	interferon regulatory	AI028035
3.555828	0.000269	1552682_a_at			
3.509179	0.000107	209709_s_at			
3.2345	0.000267	231929_at	IKZF2	IKAROS family zinc fir	AI458439
3.186033	0.000104	229492_at			
3.175463	0.000155	1555827_at	CCNL1	cyclin L1	AY034790
3.14236	0.000142	222380_s_at	PDCD6	programmed cell dea	AI907083
3.102367	0.000116	1555310_a_at			
3.047999	0.000105	232238_at			
3.025457	0.00027	213873_at			
3.009803	0.00026	219691_at			
2.963853	3.07E-05	1557029_at			
2.905201	0.000154	219461_at			
2.688539	0.000151	236267_at			
2.684126	2.26E-05	215507_x_at	---	---	AL049985
2.677544	1.00E-04	227316_at	---	---	AF116715
2.645413	1.43E-05	206558_at			
2.595265	6.35E-05	225079_at			
2.59231	7.06E-06	205046_at	CENPE	centromere protein E	NM_001813
2.547156	0.000205	235846_at			
2.52455	8.03E-05	232740_at	MCM3AP-AS1	MCM3AP antisense R	BC002458
2.492939	7.97E-05	1559174_at			
2.470123	0.000181	1557129_a_at	FAM111B	family with sequence	AA960844
2.415391	0.000143	220885_s_at			
2.39817	3.83E-05	1552680_a_at	CASC5	cancer susceptibility	NM_020380
2.29475	4.15E-05	241713_s_at			
2.200942	0.000181	229538_s_at	IQGAP3	IQ motif containing G	AW271106
2.110325	6.60E-05	239265_at			
2.058794	0.000206	214474_at			
-1.57241	0.00011	230579_at			
-1.72197	0.00023	230937_at			
-1.89782	0.000159	215156_at			
-1.94501	3.51E-05	205013_s_at			
-1.95036	0.000218	233469_at			
-2.0089	0.000175	240383_at			
-2.02358	0.000129	220373_at			
-2.03682	7.43E-05	237718_at			
-2.04511	0.000181	1558711_at			
-2.09833	0.000183	226824_at			
-2.11512	0.000275	244544_at			
-2.11714	0.000275	229438_at			
-2.15834	0.000202	1562657_a_at			
-2.18983	0.000163	242904_x_at	RP11-489E7.4	---	AI351653
-2.295	0.000181	235557_at			
-2.40619	0.00016	242773_at			
-2.42052	0.00027	219480_at			
-2.52107	4.32E-05	228557_at			
-2.53195	0.000218	233520_s_at			
-2.64968	0.00014	1564027_a_at			
-2.68074	9.51E-05	1558586_at			
-2.72096	5.11E-06	231013_at			
-2.75201	3.94E-05	239845_at			
-2.83349	4.49E-05	235272_at			
-2.87925	1.98E-05	230477_at			
-2.8997	0.00014	234052_at			
-3.01717	6.72E-05	228888_at			
-3.05685	1.21E-05	230068_s_at			
-3.11845	3.88E-05	220528_at			
-3.19141	8.54E-05	211105_s_at			
-3.24741	5.91E-06	205040_at			
-3.35109	9.80E-05	230318_at			
-3.37232	0.000259	1553785_at			
-3.44992	9.55E-05	206628_at			
-3.48877	4.18E-06	209843_s_at			
-3.58641	7.36E-05	242913_at			
-3.69515	0.000216	209720_s_at			
-3.71826	0.000121	222513_s_at			
-3.74191	0.000233	221992_at	NPIPB15	nuclear pore complex	AI925734
-3.88275	7.93E-05	209242_at			
-4.03309	3.79E-05	211819_s_at			
-4.05855	8.74E-05	207113_s_at			
-4.08195	9.21E-05	210457_x_at	HMGA1	high mobility group A	AF176039
-4.15545	0.000282	230233_at			
-4.28245	8.21E-05	204213_at			
-4.32458	0.000131	222900_at	NRIP3	nuclear receptor inte	AJ400877
-4.4753	0.000109	227742_at			
-4.50724	3.41E-06	205030_at	FABP7	fatty acid binding pro	NM_001446
-4.70227	0.000114	220133_at	ODAM	odontogenic, amelob	NM_017855
-4.73123	2.09E-07	211302_s_at	PDE4B	phosphodiesterase 4	L20966
-4.73385	1.74E-05	202672_s_at	ATF3	activating transcriptic	NM_001674
-4.81775	0.000104	206509_at	PIP	prolactin-induced pro	NM_002652
-4.91922	0.000103	210413_x_at	SERPINB3 /// S	serpin peptidase inhi	U19557
-5.01027	5.22E-05	209842_at	SOX10	SRY (sex determining	AI367319
-5.24087	2.82E-05	203708_at	PDE4B	phosphodiesterase 4	NM_002600
-5.52384	2.63E-05	203665_at	HMOX1	heme oxygenase (de	NM_002133
-5.6322	7.97E-05	205916_at	S100A7	S100 calcium binding	NM_002963
-5.74102	2.58E-05	228245_s_at	LOC100509445	uncharacterized LOC1	AW594320
-6.13025	8.86E-05	206378_at	SCGB2A2	secretoglobin, family	NM_002411

#2 TOP 100 DEGs when Ovarian Cancer data is compared to Ovarian Normal

logFC	P.Value	ID	Gene Symbol	Gene Title	Representative Pub
2.38585	0.001201	207542_s_at	AQP1	aquaporin 1	NM_000385
2.206819	0.00736	210619_s_at	HYAL1	hyaluronogl	AF173154
2.158729	0.005803	1569555_at	GDA	guanine dea	BC012859
2.082519	0.00404	229797_at	MCOLN3	mucolipin 3	AI636080
2.056253	0.000553	224179_s_at	MIOX	myo-inositol	AF230095
2.041843	0.008646	227394_at	NCAM1	neural cell a	W94001
1.970595	0.002712	220332_at	CLDN16	claudin 16	NM_006580
1.91088	0.012877	234723_x_at	---	---	AK024881
1.864484	0.002707	236717_at	FAM179A	family with s	AI632621
1.815494	0.012086	232046_at	KIAA1217	KIAA1217	AU148164
1.806412	0.001814	209755_at	NMNAT2	nicotinamide	AF288395
1.664403	0.005341	1556029_s_at	NMNAT2	nicotinamide	H90656
1.648862	0.00963	231929_at	IKZF2	IKAROS fami	AI458439
1.601447	0.007862	1552395_at			
1.598335	0.010261	1555827_at	CCNL1	cyclin L1	AY034790
1.552637	0.008786	204729_s_at			
1.546504	0.005654	239907_at			
1.530273	0.009699	205469_s_at	IRF5	interferon re	NM_002200
1.520127	0.011722	1557669_at			
1.500774	0.005174	1552719_at			
1.499399	0.006829	214421_x_at			
1.48463	0.001563	1553264_a_at			
1.477921	0.0105	215515_at			
1.429067	0.007577	222380_s_at	PDCD6	programmec	AI907083
1.426704	0.005815	240555_at			
1.396656	0.00368	208213_s_at			
1.362415	0.004724	207352_s_at			
1.340936	0.00765	233333_x_at			
1.322993	0.00533	1561910_at			
1.298273	0.005787	216025_x_at			
1.263328	0.01065	233953_at	GUCA1C	guanylate cy	AF110003
1.239832	0.005185	205163_at			
1.223125	0.007721	229645_at			
1.200062	0.004683	216661_x_at			
1.180744	0.011497	238222_at	GKN2	gastrokine 2	AI821357
1.155955	0.010044	1558855_at			
1.147164	0.005664	221374_at			
1.12787	0.01042	237253_at	IGSF11-AS1	IGSF11 antis	AA789243
1.11778	0.009229	1552568_at			
1.083468	0.0118	211328_x_at			
1.072177	0.009118	226402_at			
1.06603	0.010565	214485_at			
1.060391	0.006872	236987_at			
-1.02105	0.010882	240849_at			
-1.16459	0.008534	229749_at			
-1.24356	0.011445	224169_at			
-1.25998	0.012054	238818_at			
-1.36327	0.004911	204412_s_at			
-1.38334	0.011245	243110_x_at			
-1.38863	0.012519	221992_at	NPIPB15	nuclear pore	AI925734
-1.40839	0.006424	238865_at			
-1.46476	0.007626	241360_at			
-1.48146	0.012855	214913_at			
-1.49064	0.008012	209883_at			
-1.5225	0.003427	239410_at			
-1.53174	0.00987	227554_at	MAGI2-AS3	MAGI2 antis	AU145805
-1.56107	0.012682	218297_at			
-1.57804	0.000505	227971_at			
-1.59439	0.001262	239376_at			
-1.61027	0.005127	237034_at			
-1.63933	0.002146	226085_at	CBX5	chromobox h	AA181060
-1.65745	0.01018	238418_at			
-1.69449	0.011726	228333_at			
-1.70573	0.002443	217525_at			
-1.70952	0.008428	219949_at			
-1.72077	0.010086	227444_at	ARMCX4	armadillo re	AW519141
-1.7275	0.006174	230121_at			
-1.73039	0.012636	204940_at			
-1.7389	0.005495	220171_x_at			
-1.79774	0.001455	224508_at			
-1.81293	0.004123	208368_s_at			
-1.93575	0.007032	213788_s_at			
-1.95782	0.006338	209789_at			
-1.96305	0.009752	209679_s_at			
-1.9776	0.011992	219557_s_at			
-2.02574	0.011779	221107_at			
-2.03351	0.001195	219315_s_at			
-2.04489	0.000445	231938_at			
-2.05531	0.006992	238443_at			
-2.05986	0.007344	209840_s_at			
-2.0919	0.0115	213899_at			
-2.12668	0.001304	220327_at			
-2.14827	0.00761	214954_at			
-2.2234	0.001429	213362_at	PTPRD	protein tyro	N73931
-2.22758	0.002741	205517_at	GATA4	GATA bindin	AV700724
-2.23604	0.005772	202920_at	ANK2	ankyrin 2, ne	BF726212
-2.33957	0.00053	222900_at	NRIP3	nuclear rece	AJ400877
-2.41506	0.00437	33767_at	HHLA1	HERV-H LTR-	AU148706
-2.41583	0.008085	210457_x_at	HMGA1	high mobilit	AF176039
-2.46114	0.004228	214841_at	CNIH3	cornichon fa	AF070524
-2.51472	0.007544	242277_at			
-2.56126	0.008528	234304_s_at	IPO11 /// IPO1	importin 11 /	AL162083
-2.74044	0.009836	211340_s_at	MCAM /// MIR	melanoma c	M28882
-2.91192	0.001597	209087_x_at	MCAM	melanoma c	AF089868
-2.95982	4.18E-05	202672_s_at	ATF3	activating tr	NM_001674
-3.07096	2.07E-05	229160_at	MUM1L1	melanoma a	AI967987
-3.28701	0.003959	205347_s_at	TMSB15A /// T	thymosin be	NM_021992
-3.93675	0.001985	227705_at	TCEAL7	transcription	BF591534
-4.15203	0.007862	218469_at	GREM1	gremlin 1, D	NM_013372
-4.31605	0.012297	218468_s_at	GREM1	gremlin 1, D	AF154054