Functional Analysis of Differentially Expressed Genes Associated with the Conditions of Polycystic Ovary Syndrome and Endometrial Cancer

By

SONIA MUNAWAR

Fall 2017-MS BI-2 00000203601

Supervised by

Dr. Rehan Zafar Paracha

RESEARCH CENTRE FOR MODELING & SIMULATION

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

SEPTEMBER 2020

Functional analysis of differentially expressed genes associated with the conditions of Polycystic Ovary Syndrome and Endometrial Cancer

SONIA MUNAWAR

Research Centre for Modeling & Simulation

A thesis submitted to the National University of Sciences & Technology in partial fulfillment of the requirement for the degree of Master of Science in Computational Science & Engineering

SEPTEMBER 2020

DEDICATION

This research is dedicated to my beloved parents, who have supported me and are a source of constant motivation.

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

______________ ________________________

Date SONIA MUNAWAR

Acknowledgement

Praises to The Creator and The Benevolent for without Him there is no purpose to life. Asking for aide from Almighty Allah are my loving parents and their prayers which formed a platform for me to stand upon. Many salutations of gratitude are directed towards Dr. Rehan Zafar Paracha, a knowledgeable supervisor and a kind-hearted human-being guided me all the way through to the end and the thesis would not be at the current standard as it is now. I am indebted to my GEC members, Dr. Ishrat Jabeen and Dr. Zamir Hussain, for their motivation and constructive criticism so as to prepare me for future endeavors. Many appreciations for the supportive friends Ayesha Shabir,Maham Hamid,Sadaf Saleem, Misha.Waheed, Hassan Tahir Shah, Muhammad Hasnain, Ahtisham fazeel abbasi, Arslan Ali, Samawia Rizwan, Sadia bano,Sidra Adil, and fellows, the professional faculty members, kind staff and humble students of RCMS, NUST.

Abstract

Polycystic ovary syndrome is a metabolic and endocrinal disorder common in women of reproductive age and has led to infertility prevailing across the world. The prevalence of PCOS is 33% worldwide. PCOS can also lead to serious conditions such as anxiety, depression, Insulin resistance, obesity, metabolic disorders (dyslipidemia), cardiovascular complications and endometrial cancer. Linkage of PCOS with endometrial cancer has been proposed by many studies. Beginning in the lining of uterus (endometrium), the most common gynecological malignancy is endometrial cancer. PCOS is a major risk factor associated with endometrial cancer. So, there is a dire need to analyze these diseases at molecular level to propose potential biomarkers and therapeutic targets. Microarray and RNA-Seq data analysis have been performed for analyzing the molecular patterns of the diseases. Different datasets of PCOS and EC have been analyzed in order to compare expression profiles of transcripts. Subsequently, pathway analysis of all datasets provides the evidences for a common pathway which has been affected in both diseases. This pathway is known as "Focal adhesion". This pathway is further quantitatively modeled by using the expression values of differentially expressed genes. High sensitivity level of ACTG1 and CycD have been observed towards cell mobility and cell proliferation conditions, respectively. Change in expression of these genes can lead towards endometrial cancer and can be used as important therapeutic targets.

Contents

List of Figures

List of Tables

Chapter 1 INTRODUCTION

1.1 POLYCYSTIC OVARY SYNDROME:

Polycystic ovary syndrome (PCOS) is a common disorder associated with hormonal imbalance in women of reproductive age mainly due to excess in the levels of androgen (a male reproductive hormone). This hormonal imbalance affects menstrual period and results in infertility.(Ndefo, Eaton and Green, 2013).This syndrome mainly affects ovaries and dysregulate the production of sex hormones. Abnormal hormonal activities also result in the production of cysts in ovaries. Ovarian cysts are sacs filled with fluid in ovaries commonly formed during ovulation. It is characterized by increased number of oocytes of poor quality. Ovulation occurs when eggs are released by ovary each month. In women with PCOS syndrome, ovarian cysts are formed due to anovulation. Due to the reason, follicles keep growing and form multiple cysts (Lee and Rausch, 2012). Change in metabolic and hormonal activities increase the risk of insulin resistance, obesity, pancreatic b-cell dysfunction, hirsutism, cardiovascular disease and some types of cancer. Women with PCOS usually have increased inflammation levels in their body because of overweight and higher androgen level (Escobar-Morreale, 2018).

1.2 PREVELANCE OF PCOS:

This is a second most common cause of infertility. The prevalence of PCOS is 33% worldwide. It was estimated that 40% of women who attended infertility clinics have PCOS. It is affecting 5-10% of women of reproductive age. The highest reported prevalence of PCOS was 52% among South Asian women (George and Malini, 2014). A study suggested the lowest ratio of

PCOS in Chinese as well as black women and an increasing prevalence rate was reported in Caucasians and middle eastern groups. Variation in prevalence of PCOS is due to the difference in diagnostic criteria in different ethnic groups (Lee and Rausch). Overall, prevalence rate of PCOS cannot be reported because most of the cases remain undiagnosed due to less awareness. This issue is still more challenging because of large undiagnosed population. According to NIH diagnostic criteria, prevalence of PCOS among 6-7% population is documented in United States, United Kingdom, Spain, Greece, Australia as well as Mexico (Wolf *et al.*, 2018). In Pakistan, PCOS is the most pervasive gynecological disorder with 55.41% of cases (Sidra *et al.*, 2019) .

1.3 PATHOGENESIS OF PCOS:

Complete understanding of pathophysiology of PCOS is lacking. Due to heterogeneity of this syndrome it remains difficult to find out the exact mechanism. There are multiple pathophysiologic mechanisms and defects like disturbance in androgen production, hyperinsulinemia, obesity, genetics and environmental factors related to it (Rosenfield and Ehrmann, 2016).

Main feature that is associated with PCOS is hyper androgenism. Many studies suggested that hyperinsulinemia and obesity play role in its pathophysiology. Inhibition of lipolysis and abundance of GLUT4 (Glucose transporter type 4) is stimulated by insulin but in case of PCOS both are reduced and leads towards metabolic disturbance. In case of obesity, level of SHBG (Sex hormone-binding globulin) decreases which increase the availability of testosterone and as a result causes hyperandrogenimia.Obese PCOS women have higher rate of glucose intolerance therefore obesity is aggravating environmental factor (Barber *et al.*, 2019).

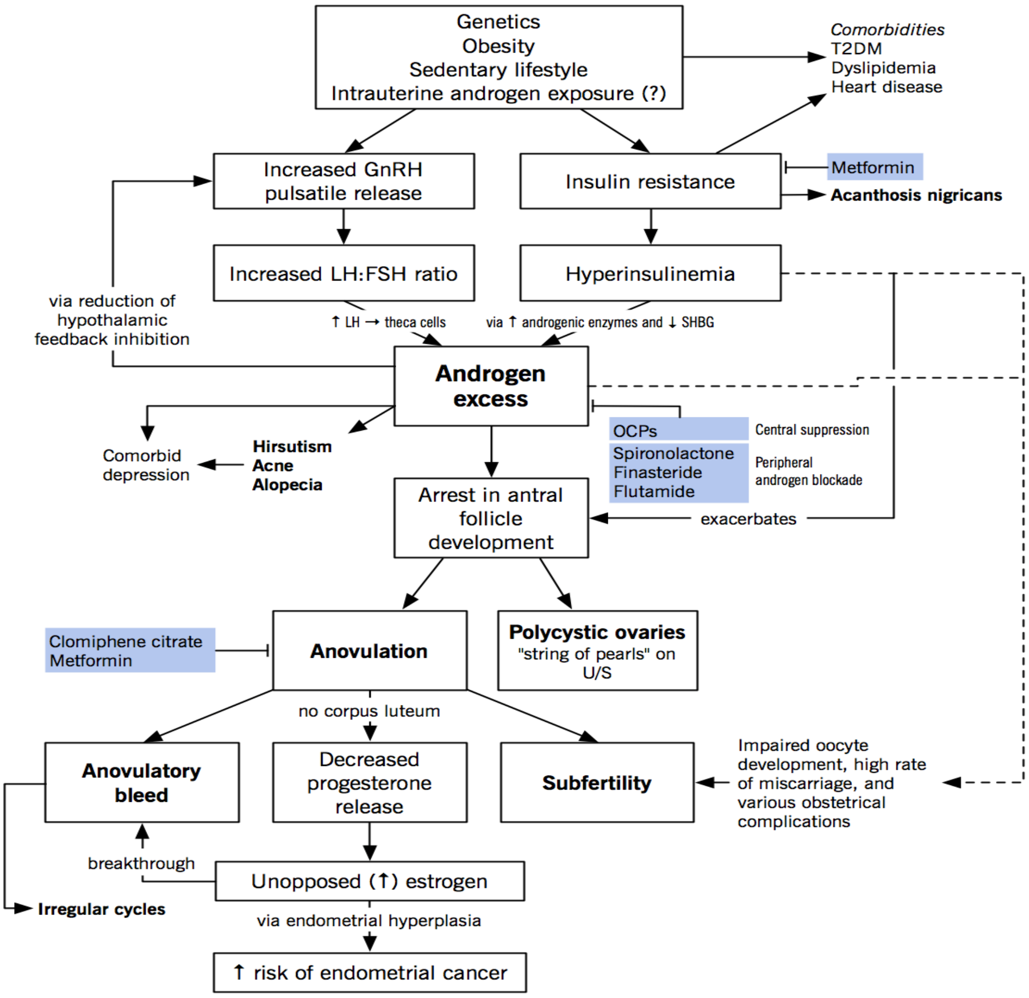

Figure 1.1: Pathophysiology of PCOS. Figure is adapted from (Goodarzi et al., 2011)

1.3.1 Treatment of the Disease

To control the hyperandrogenism and insulin resistance different insulin reducing agents are used like Thiazolidinediones (TZDs) and Metformin. Metformin shows active result even in women without hyperinsulinemia. It decrease androgen production by decreasing glucose production and direct action on the ovaries (Physician, 2016). Another insulin sensitizing agent TZDs trigger genes that encode insulin action, gene transcription and normal FFA metabolism in androgen secreting cells. In this way, TZDs bring into being effective for both hyperandrogenism and insulin resistance. In short, all insulin sensitizing drugs decrease free fatty acids (FFA) level by improving insulin sensitivity in adipocyte and improves hyperandrogenimia (Singh and Rai, 2019)

1.4 ENDOMETRIAL CANCER

Endometrial cancer (EC) is utmost occurring reproductive disorder worldwide specially in advanced countries. It initiates in the endometrium of the uterus. Uterus is main hormone-responsive secondary sex organ of female reproductive system. Once the egg left the ovary and get fertilized it is implanted in uterus where fetus develops during pregnancy. Uterus thick wall has three layers, among them, innermost layer is known as endometrium from where EC embarks on. It is made up of glandular cells and cause secretions. EC is also known as uterine cancer. It often results in abnormal vaginal bleeding even after menopause, painful urination, pelvic pain, enlarged uterus and bleeding between periods (Tomao *et al.,* 2016). The exact mechanism of EC is unidentified but it is proposed that hormonal imbalance or genetic mutation can lead towards it. Higher estrogen level without progesterone production causes the thickness of endometrium. Cells that make up the lining crowd together become abnormal and lead towards cancer. Other risk factors associated with this disease includes poly cystic ovaries, diabetes mellitus, obesity, more years of menstruation, older age, never having been pregnant, and white race (Crandall *et al.*, 2018). PCOS and EC share common risk factors like obesity, hormonal imbalance and insulin resistance so there is a possibility of strong linkage between both diseases (Mravec and Tibensky, 2020).

1.4.1 Subtypes of Endometrial cancer

On the basis of clinical outcomes and histology, EC is divided in to further two subtypes (Burke *et al.*, 2014).

- Type I tumors (Estrogen dependent)
- Type II tumors (Estrogen independent)

1.4.1.1 Type I tumors

Type I tumors are proceeded by overgrowth of cells in the endometrium (hyperplasia). It mostly covers endometriosis adenocarcinomas and associated with disturbances in the estrogen level. This type is usually not very destructive and do not spread to other tissues swiftly (Mäenpää, 2020).

1.4.1.2 Type II tumors

Type II tumors are predominantly serous carcinomas. This is not independent of estrogen level and is more aggressive. It tends to spread to other tissues quickly even to the outside of the uterus (Mäenpää, 2020). Therefore, it is suggested that both subtypes have distinct etiologies.

1.4.2 Prevalence

EC is the sixth most occurring cancer in women. The highest incidence rates of EC were seen in United States and in in the Slovakia. Rates are lowest in middle-income countries such as India, Pakistan, South Africa and highest in North America and Europe(Parsons *et al.*, 2018).. Rate is increasing day by day. In 2018, an estimated 382,069 women are identified with endometrial cancer worldwide (Ghoubara, Sundar and Ewies, 2019).

Pakistan has also observed fewer cases of EC as compared to other gynecologic cancers. The risk of endometrial carcinoma rises with age from 1% at the age of 50 years to approximately 25% at the age of 80 years in Pakistan (Ghoubara, Sundar and Ewies, 2019). In contrast to USA, Pakistan observes less cases of EC.

1.4.3 Pathophysiology

Disturbances in the Estrogen level is important in the etiology of EC. Estrogen metabolism, including the expression pattern of aromatase and regulation of 17β-hydroxysteroid dehydrogenase type 2 (responsible for the inactivation of estradiol to estrogen) is altered in women with endometriosis. Aromatase is and enzyme that catalyzes the conversion of steroids to estrogen (Dumesic *et al.*, 2015).

Modification in the (estrogen receptor) ER-*α* gene, the *CYP19* gene encoding aromatase, and several other genes are associated with the endometriosis. Other frequently altered genes which are reported in type I carcinomas are PTEN, ß-catenin and K-Ras. In comparison to type I, the typical type II serous carcinoma is estrogen independent and seems to be driven by TP53 mutations (Ghoubara, Sundar and Ewies, 2019). Mutations in p16, cyclin E and CTNNB1 have also been observed in Type II serous carcinoma. Further research of these will lead to better understandings of the pathophysiology and etiology of endometrial cancer.

PCOS and EC share common risk factors like obesity, hormonal imbalance and insulin resistance. It is proposed that women with PCOS have endocrine abnormalities, hormonal imbalance and it play vital role when it comes to cancer. PCOS women have increased levels of estrogen and abnormally low levels of progesterone and eventually if PCOS is not treated well it can lead towards cancer (Holm *et al.*, 2012). Also during ovulation uterine wall shed but in case of PCOS women does not ovulate regularly and lining can build up to increase risk of EC. These are multifactorial and most debatable reproductive endocrinal disorders in the young females. Female with PCOS and EC cannot conceive and could have various disorders so there is a strong need to understand this disorder at molecular level (Mravec and Tibensky, 2020).

Comparison and identification of differentially expressed genes in PCOS and EC is important for the identification of risk factors. It can also inform about differences of events occurring in these diseases which are common and can be treated with a single therapeutic agent. Moreover, this study will help to identify a linkage between PCOS and EC by understanding common pathways among both conditions. Different genetic mutation can act as biomarkers and provide major therapeutic targets for these diseases. So, there is a need to find the novel biomarkers and potential therapeutic targets that can assist in poly cystic ovaries and endometrial cancer therapies in near future. Due to the advancements of next generation sequencing many technologies have been developed for this purpose and the most significant ones are Microarray and RNA-Seq analysis.

1.5 TECHNIQUES FOR TRANSCRIPTOME ANALYSIS

Transcriptome analysis is very important to understand underlying biology of many diseases and different physiological activities. In past few decades, techniques like microarray and NGS (Next Generation Sequencing) have revolutionized the field of genomics(Levy and Myers, 2016). Using these techniques, one can analyze whole genome effectively and genetic causes of many diseases can be explored quickly. NGS includes different techniques (Illumina, Roche, Ion torrent sequencing) of DNA and RNA sequencing (Grada and Weinbrecht, 2013). Focus of this study is on microarray and RNA-Seq techniques. These two techniques are explained in detail below.

1.5.1 Microarray Analysis

This technique is one of the recent advances of sequencing techniques and it provides assistance in research of cancer and various other diseases. Microarray is a DNA or a glass chip with thousands of different known DNA fragments organized in the form of rows and columns in such a manner that each probe can be identified through its location on array. In this technique, unknown fragments from mRNA molecules are converted to complementary DNA (cDNA) fragments by reverse transcription. They are then labeled by using fluorescent markers. These cDNA fragments are then subjected to react with complementary probes on chip. After binding to all the complementary sequences on chip, the remaining fragments not bound to probes are washed away. The labeled targets are then identified by their fluorescence by passing a laser beam. This fluorescence emission pattern is then recorded in the form of intensities in order to identify the RNAs from the sample. It also assists in biomarker identification for a particular disease by comparison of the normal versus diseased states. (Villaseñor-Park and Ortega-Loayza, 2013).

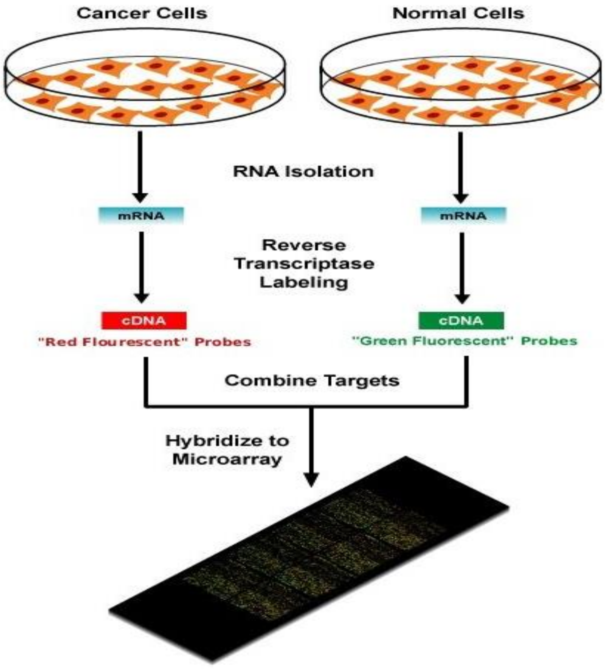

Figure 1.2: Overview of microarray technique. In the first step, mRNA is extracted from both control and experimental samples. These mRNAs are reverse transcribed to cDNA and are fluorescently labeled. Finally, these cDNAs are allowed to hybridize with probes present on microarray chips and these chips are scanned. Figure is adapted from (Redon, Fitzgerald and Carter, 2009).

1.5.2 RNA-Sequencing

RNA-Sequencing (RNA-Seq) has replaced microarrays in last few years (Kukurba and Montgomery, 2015).RNA-Seq has several advantages over microarrays such as larger dynamic range, novel transcript discovery that is not possible in microarrays as results are limited to used probes (Nagalakshmi, Waern and Snyder, 2010).

RNAs from biological samples are extracted in the form of small fragments which are reverse transcribed into cDNA sequences. These sequences are amplified using PCR and sequenced using sequencing machines. Output of sequencing machines is in the form of small sequences called reads. These reads are then preprocessed to remove reads with low quality and any biases

(Pareek, Smoczynski and Tretyn, 2011). Once the reads are preprocessed, these are aligned to reference genome. Each read represent a specific genomic feature. Number of reads mapping to a specific position of a reference genome represent the total count of that specific genomic feature in biological sample The number of reads aligned to a particular site of genome are then counted and the gene expression is quantified, accordingly (Hrdlickova, Toloue and Tian, 2017).

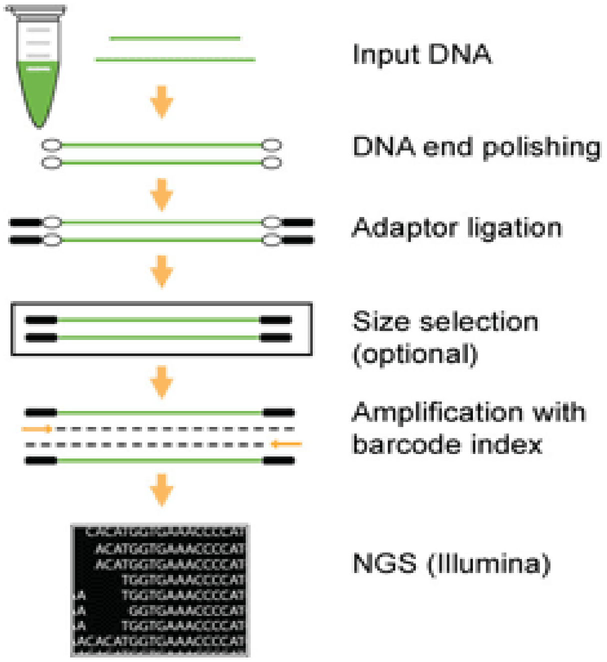

Figure 1.2: RNA-Sequencing technique. Figure shows step by step methodology of RNA-Seq technique. First of all mRNAs are extracted from RNAs. In the next step poly, A tails are removed and mRNAs are converted to cDNA. This cDNA are then sequenced and reads are obtained. Each read represents a specific mRNA. These reads are allowed to map with human reference genome. Finally read counts are obtained. Figure is adapted from (Kumar *et al.*, 2012).

1.6 PATHWAY ANALYSIS

Pathway analysis plays an important role in understanding the holistic molecular behavior of physiological and pathological systems. The aim of pathway analysis is to analyze the high throughput sequencing data, interacting genes and expression of genes in a particular pathway

(García-Campos, Espinal-Enríquez and Hernández-Lemus, 2015). There are different databases like Kyoto Encyclopedia of Genes and Genomes (KEGG) and DAVID that can reveal the association of input genes with different physiological and pathological singling pathways.

1.7 SYSTEMS BIOLOGY

Systems biology is an inter-disciplinary field which uses computational and mathematical modeling approaches to study complex biological and molecular networks (Smith, 2008).Biological and molecular networks involve, metabolic, genetic and signaling networks. The study of complex biological networks is very important to understand the role of each component in disease. Use of these approaches improve our understanding of complex systems. This approach is used to build model of differentially expressed genes involved in disease (Krallinger and Tendulkar, 2006).

1.8 APPLICATIONS OF TRANSCRIPTOMIC ANALYSIS

The downstream transcriptomic analysis encompasses various applications for achieving further knowledge regarding genes in normal and diseased states (García-Campos, Espinal-Enríquez and Hernández-Lemus, 2015). Some of applications covered in this thesis includes

- Differential Gene Expression Analysis
- Pathway analysis.

1.9 PROBLEM STATEMENT

It has been studied that women with PCOS have three to five times more chances of having endometrial cancer. Both share common risk factors like obesity, hormonal imbalance, insulin resistance, infertility. Most studies have investigated the link between infertility and EC. So, here we are performing analysis for the identification of differentially expressed genes related to PCOS and EC.

1.10 OBJECTIVES

The main objective is the identification of differentially expressed genes and analysis of common and uncommon gene patterns among different disorders associated with PCOS and EC in order to:

- Identify potential biomarkers through Microarray analysis and RNA-Seq.
- Identify the possible affected signaling events and related pathways by using systems biology approach.
- Compare the regulatory events involved in the pathophysiology of both diseases.

Chapter 2 LITERATURE REVIEW

A review of literature has been performed to understand different perspectives of the diseases. It contains two main sections. First section illustrates the brief description of PCOS and endometrial cancer along with their molecular causes. Second section is comprised of review of microarray and RNA-Seq studies in context of these disease.

Polycystic ovarian syndrome (PCOS) is a disorder of women of reproductive age mainly due to disturbance in hormonal level which further disturb menstrual cycle and leads towards infertility. This disease is prevailing across the world. Due to disturbance in hormonal level, cysts in ovaries are formed. Ovarian cysts are sacs filled with fluid in ovaries that appear like "string of pearls". Most studies have discussed that higher androgen level is the major cause of disorder. (Giallauria *et al.*, 2009).

The etiology of PCOS is still uncertain. However, several studies have suggested that insulin resistance and beta-cell-dysfunction plays an important role in the pathogenesis of the syndrome. Among PCOS women risk of diabetes seems to be approximately 5 to 10 fold higher than normal. Different risk factors such as a positive family history, obesity and hyperandrogenism may contribute to increase the diabetes risk in PCOS (Diamanti-Kandarakis, Kandarakis and Legro, 2006).

There are many abnormalities at molecular level contributing to the disease. It was hypothesized that associated hirsutism is due to hyperandrogenism, which was either of ovarian or adrenal origin. In PCOS, there is a defect in the biosynthesis of estrogens within the cystic follicles, which, further leads towards hirsutism (Jayasena and Franks, 2014).

Evidences suggest that insulin signaling pathways play critical role in its pathophysiology. Many other defects that might be due to environmental factor or genetic abnormality or both are also observed. Therefore, it is difficult to propose the exact mechanism of PCOS (Tomao *et al.*, 2016).

Change in metabolic and hormonal activities may increase their risk of insulin resistance, pancreatic b-cell dysfunction, hirsutism, obesity, cardiovascular disease and some types of cancer. Insulin resistance is the peculiar feature of PCOS and seen in approximately 50–70% of affected women (Escobar-Morreale, 2018).

PCOS is considered as heredity trait that might result from genetic mutation under the effect of environmental factors. Candidate genes for PCOS are associated with metabolic and androgenic pathways. Genes encoding inflammatory cytokines have been identified as target genes for PCOS, as these are also linked with cardiovascular risk, insulin resistance and obesity (Roldán, San Millán and Escobar-Morreale, 2004).

Dysfunction of granulosa cells may be etiologically important in the pathogenesis of PCOS. Comparative analysis identified candidate genes involved in MAPK/ERK signaling pathways that may influence the function of granulosa cells in PCOS. (Kaur *et al.*, 2012). It was later confirmed that mitogen-activated protein kinase 4 and phospho-ERK1/2 were decreased in PCOS granulosa cells (Chen-Wei Lan et al 2015).

Another study also suggested that genes of cumulus cells may act as biomarkers because many identified DEGs (ANGPTL1, SERP4, TNIK, LHCGR, SOCS3 and GRIN2A) were involved in different pathways like metabolism, focal adhesion, calcium signaling, Wnt signaling and type 2 diabetes. After RT-PCR analysis LHCGR, SOCS3 and TNIK were proposed as putative biomarkers in PCOS (Liu *et al.*, 2016).

Several microarray and RNA-Seq studies have also been carried out to understand the etiology of PCOS. In past few decades, microarray technology has revolutionized the field of genomics. Microarrays allow looking into molecular biology and evaluation of biological processes. Microarray studies have been implemented in order to analyze genetic alterations (Villaseñor-

Park and Ortega-Loayza, 2013). With the advancement in NGS techniques, RNA-Seq has surpassed the microarrays in transcriptome analysis. RNA-Seq has the ability of novel transcripts identification (Qian *et al.*, 2014).

Bioinformatics approaches gave an insight to the genetic aspects of PCOS and revealed that several genes such as DENND1A, CYP17A1, CYP11A1, CYP19, HSD17B2, STAR, HSD17B1 and FSHB are responsible for the disease causing events (Mariano *et al.*, 2017).

Recent developments in sequencing enables researchers to identify rare genetic variants contributing to PCOS as well as to map the genetic variants (Zhang *et al.*, 2019).

Endometrial cancer (EC) is the most widely recognized gynecological threat among ladies worldwide with 287,000 new cases and 74,000 deaths every year (Burke *et al.*, 2014). EC has been divided into two types, (1) less vigorous type I and (2) extremely aggressive type II. Due to different histopathology, clinical behavior and underlying molecular profiling. Around 75% ECs are type I and are estrogen dependent adenocarcinomas with endometrioid morphology. They are generally analyzed at initial stages. On the other hand, type II ECs are estrogen independent with myometrial invasion, extra uterine spread and serous histology. A potential method for the treatment is to target EC cells by hindering key signaling pathways that are important for tumor growth (Dong *et al.*, 2013).

The use of microarray, whole genome sequencing and other techniques of NGS enlightened the path towards better understanding of molecular basis of disease. We can investigate the gene expression in a solitary run or in an efficient way. Such techniques gives a huge platform to scientific community to work on essential aspects of development of life and to investigate the hereditary issues responsible for disease (Zhang *et al.*, 2018).

Studies suggested that EC develops as a result of a multistep process of tumor suppressor gene inactivation and oncogene activation. It was proposed that type I EC is characterized by mutation of KRAS2, PTEN, and type II malignancies frequently contain mutations of Her-2/neu and TP53 (Risinger *et al.*, 2003).

High degree of somatic mutations were identified in SPOP, FBXW7, ARID1A, CHD4, ABCC9 and MAP3K4. Serous tumors had mutation in chromatin-redesigning gene and ubiquitin ligase complex gene. Mutational interruption in these processes is probably the deadliest type of endometrial malignant growth. Mutation in PI3K and WNT signaling pathways were also identified, recommending therapeutic targets for treatment of EC (Li *et al.*, 2014).

Worldwide, gene expression analysis is perceived as a powerful method for analyzing the transcriptional profiles. The most significantly upregulated estrogen responsive genes (KIAA1324, TFF3, MLPH) and genes that involved in estrogen-related processes (FOXA2, ESR1, PGR) were identified. Two other upregulated genes identified in EC have previously been reported are TFF310 and CEACAM110. These genes are involved in cell-adhesion pathways, extra-cellular matrix processes and have been linked to other cancer types. Results from this study contribute to understand and identify the molecular mechanisms of endometrial cancer (O'Mara, Zhao and Spurdle, 2016).

In addition, it was proved by microarray studies that JQ1 (thienotriazolodiazepine) adjusted PTEN (Phosphatase and tensin) and its downstream PI3K/AKT signaling targets. JQ1 is a bromodomain inhibitor that upregulate expression of PTEN gene, block the PI3K/AKT signaling pathway and silent tumor development. Studies recommended that focusing on JQ1 may fill in as a novel helpful target in endometrial diseases (Qiu *et al.*, 2016).

Molecular anomalies in PIK3R1, AKT2 and FOXO1 that could lead to the activation of the IL-7 signaling pathway has not been previously linked with EC was also reported (Suhaimi, Ab Mutalib and Jamal, 2016). PIK3CA amplification is a solid prognostic marker and a potential marker for typeII endometrial cancer (Holst *et al.*, 2019).

The prognosis for EC is more in women with PCOS because their endometrial tumors tend to exhibit a great degree of differentiation than in women without PCOS. This was further investigated that prevalence of EC in women with PCOS was 37% (Pillay *et al.*, 2006). Many studies have investigated the link between EC and anovulatory infertility or with endometrial hyperplasia but not with PCOS. Later, it was also suggested that women with PCOS may be at increased risk of developing EC (Barry, Azizia and Hardiman, 2014).

It was analyzed that expression of IGFBP1, PTEN, IGF1 was upregulated in endometrium of women with PCOS and EC. These findings can also help out to find biomarkers which can predict that women with PCOS will go on to develop endometrial cancer or not (Shafiee *et al.*, 2016).

The exact mechanism of the association between PCOS and EC is however unclear. A recently published review in a meta-analysis proposed that women with PCOS were almost three to five times more likely to develop EC. However, due to lower amount of solid evidence further research will be required to confirm genetic linkage among both conditions. A deeper understanding of intracellular pathways and gene networks will allow the identification of therapeutic targets as well as novel biomarkers. (Mravec and Tibensky, 2020).

Chapter 3 METHODOLOGY

The main goal of this research is to find potential biomarkers and therapeutic targets by focusing on differential expression analysis of genes in poly cystic ovary syndrome and endometrial cancer using Microarray and RNA-Seq techniques. Publicly available datasets from different regions of the world have been selected to perform analysis. After finding differentially expressed genes and their comparative analysis common pathways are identified. Common pathway consists of enriched genes. Pathway analysis further clarifies that how these genes are involved in different physiological and pathological conditions. This further helps to determine potential biomarkers and therapeutic targets against the disease. The overall methodology of the study shown in Figure 3.1

3.1 DATASET SELECTION

The datasets selected for this study are from two different platforms i.e. Microarray and RNA-Seq. These datasets are retrieved from two publicly available repositories such as Array Express and NCBI- GEO (Clough and Barrett, 2016). GEO is a public repository that contains microarray and various high throughput sequencing genomic datasets. It is an international public repository with a user-friendly interface to easily locate simple submission procedures and download genomic data of interest (Agarwala *et al.*, 2016). Array express is also a public repository which accepts functional genomics data generated on microarray and high-throughput sequencing (HTS) platforms such as Illumina, Solid and 454.(Athar *et al.*, 2019) All data can be downloaded easily and to extract meaningful biological information further analysis is done by specific Bioconductor packages.

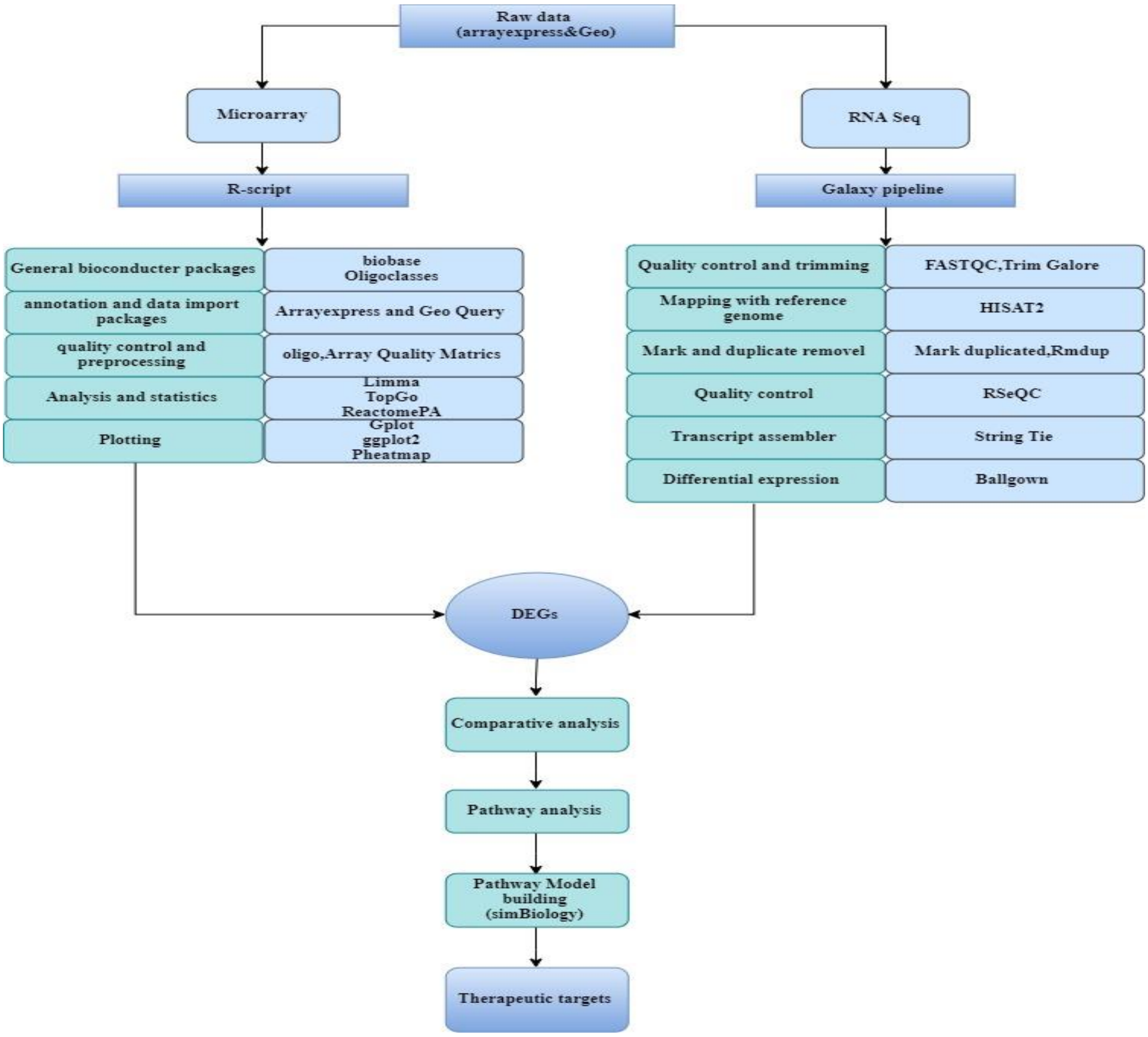

Figure 3.1: Methodology Workflow. Publicly available datasets from different regions of the world have been selected to perform micro array and RNA-Seq analysis. After finding differentially expressed genes and their comparative analysis common pathways are identified. Pathway analysis further clarifies that how genes are involved in different physiological and pathological conditions. This further helps to determine potential biomarkers and therapeutic targets against the disease.

3.2 MICROARRAY ANALYSIS

To analyze the expression of multiple genes microarray analysis is performed. It provides the significant information about the differential expression of genes by analyzing the intensities.

Different packages, tools and methods can be used for better understanding of such type of analysis. Data selection is further followed by pre-processing, quality control, normalization and differential expression analysis by using Bioconductor packages.

Bioconductor is an open source R programming language. It provides different packages and tools for the analysis of genomic data. It also provides a wide platform for Microarray and RNA-Seq analysis. The selected datasets are of different platforms, so there are different packages used for particular platforms (Kauffmann and Huber, 2010). Datasets for microarray analysis have been shown below

Table 3.1: Microarray datasets of PCOS and EC

Sr. No.	Accession No.	No. of Samples	Platform	Country	Disease Type
1	E-GEOD-34526	10	Affymetrix	India	PCOS
2	E-GEOD-6798	29	Affymetrix	Denmark	PCOS
3	E-GEOD-5090	17	Affymetrix	Spain	PCOS
4	E-GEOD-5850	12	Affymetrix	USA	PCOS
5	E-GEOD-63678	12	Affymetrix	USA	EC
6	E-MTAB-2532	176	Agilent	Norway	EC

3.2.1 Quality Control and Pre-processing

After importing raw data quality control is performed. To evaluate quality of a raw data, to detect outliers, to make microarray data more reproducible because data collected from microarray experiments can be biased, noisy and incomplete. It starts with the visual assessment of the scanned microarray images to make sure that there are no problems in data. To perform quality control oligo package from Bioconductor packages is used. Different plots are produced to check the quality of data by using different packages like ggplot2 (Kauffmann and Huber, 2010). These quality analysis and plots are discussed in detail below:

3.2.2 Principal component analysis (PCA)

It is a mathematical data compression or dimensionality reduction method used to reduce large set of correlated variables into a smaller uncorrelated variable without losing information. It is also applicable on expression data in which gene expression measurements are the observations and experimental conditions are the variables in order to check the variability. It provides information regarding the similarity between individual samples through cluster formation (Sharov, Dudekula and Ko, 2005). Here, PCA is performed on both raw and processed data to check the pattern of classification. The data is composed of two phenotypes i.e. disease vs normal.

3.2.3 Box plot

Box plots are used to check genomic variations between samples. All the samples were analyzed to have same median expression values after normalization. In box plot whiskers show extreme values and box depicts the main body of data. Each box represents an individual sample. Variability is checked by length of whiskers and size of the box. Same median estimations of all boxes show the good nature of data (Spitzer *et al.*, 2014). To check the nature of raw and processed data boxplot has been made in this research.

3.2.4 Relative Log Expression (RLE)

RLE is also a quality control method before calibrating and evaluating the data. RLE Plots are used for visualization of unwanted variation that can be highly problematic. Their detection is often crucial in high dimensional data. It was basically formulated for measurement of expression levels of many thousands of genes simultaneously. It is also used to reveal unwanted variation in many other kinds of high dimensional data, where these variations can be problematic and may not of biological interest Raw sequence file is given as input. RLE analyze every spot in array by computing the median log2 intensity of every transcript (Gandolfo and Speed, 2018).

3.2.5 Robust Multi-Array Average (RMA)

RMA algorithm allows the user to perform background correction, quantile normalization and summarization in one single step by using oligo package (Kim, Hwang and Zhang, 2014). Oligo is a package to analyze oligonucleotide arrays at probe-level. It is based on bioconductor principles of efficiency and reproducibility help in accessing, visualization, preprocessing and normalization of data.

3.2.5.1 Background adjustment

After quality assessment next step is background adjustment to adjust observed intensities. This is important because a proportion of the measured probe intensities are due to chemicals on the glass, noise in the optical detection system and non-specific hybridization. So, we would like to measure and adjust observed intensities to give correct measurements of hybridization.

3.2.5.2 Normalization

Is used to correct systematic differences between genes or arrays. The objective is to adjust the gene expression values of all genes so that the ones that are not differentially expressed have similar values across the array. For this purpose quantile normalization is used. It is applied by replacing intensity expression values to the average values and recommended after trimming.

3.2.5.3 Summarization

After normalization, summarization is necessary to be done because, transcripts are represented by multiple probes, that is multiple locations on the array. For each gene, the background-adjusted and normalized intensities of all probes need to be summarized into one quantity that estimates an amount proportional to the amount of RNA transcript. In this study all above steps are performed increase the quality of the datasets.

3.2.6 Heat Map

In order to analyze the clustering of samples according to given phenotype and to calculate sample to sample distance heatmap analysis is performed by using manhattan distance. Manhattan distance between the two points is the sum of the differences of their corresponding components (Deu-Pons, Schroeder and Lopez-Bigas, 2014). Pheatmap package in R is used to draw clustered heatmap by controlling dimensions and parameters according to given feature.

3.2.7 Differential Expression Analysis

One common goal is to rank all the genes on a chip in order to get differential expression. This analysis is performed by comparing the normal versus disease samples by applying paired t-test (Loughborough University, 2009). This statistic follows normal distribution. It further affirms this difference by checking its significance through p-value. The value around which it is considered to be significant is less than decision rule is based on p-value < 0.05.

For controlling false discovery rates, multiple testing was applied using Benjamini-Hochberg method. It helps to avoid type 1 errors alpha and adjusts the FDR. The results have been summarized in the form of table which includes ($\log_2$) fold changes, standard errors, t-statistics, p-values and adj. p-value. The decision rule for selecting differentially expressed genes was based on both p-value and adjusted p-value less than 0.05. To find differentially expressed genes between two phenotypic condition Limma package is used.

3.2.8 Limma package

Limma is a readily available package in R (Ritchie *et al.*, 2015). that allows many functions including differential expression for data arising from microarray experiments. It allows comparison of many target genes and miRNAs concurrently. It mainly fits a linear model to input data. Methods like empirical Bayes are also applied allowing the analysis for experiments with small number of arrays. Complete procedure requires two matrices which are design matrix and contrast matrix.

- Design matrix: tells whether samples are assign to each array. The row represents the array of experiment and column represents to coefficient of samples
- Contrast matrix: shows the comparison of conditions of samples.

For visualization of DE genes, volcano plots are illustrated. It is a plot of p-value versus log FC (Li, 2012). Significant and differentially expressed genes are visualized using different threshold values.

3.3 RNA-SEQUENCING

RNA-Seq analysis is performed in order to evaluate novel DEGs. RNA-Seq analysis include quality control, preprocessing, mapping to human reference genome (hg38) and obtaining read counts for each sample. Novel raw sequence reads in the form of FASTQ files are produced by RNA-Seq experiment. Further, quality control analysis using FASTQC tool is performed and the data files with errors were send for preprocessing to Trim galore. After preprocessing, reads are mapped with reference genome hg38 using HISet2. The mapped files are then subjected to StringTie tool with hg38 annotation file for obtaining read counts. Moreover, normalization, differential expression analysis and pathway analysis are also performed.

Two datasets were used for RNA-Seq. These datasets were obtained through different platforms like Illumina and Solexa etc. The detail of datasets has been shown in **Error! Reference source not found.**

3.3.1 Galaxy pipeline

Galaxy is a web portal that allows intensive genomic analysis (Giardine et al., 2005). It is an open source platform that provides users data management and contain many tools required for bioinformatics analysis. An automated pipeline in galaxy platform was made to perform quality control, preprocessing, mapping to human reference genome and obtaining feature counts. Tools used were FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), Trim galore (Bolger et al., 2014), HIset2 (Langmead, 2010), and StringTie. This pipeline provides users with extensive data management, data storage up to 250 GB, efficient processing of multiple files simultaneously and reproducibility of results. Pipeline takes accession ids of samples and human reference genome annotation file as input. As an output, it gives one file at each step which was subjected to the next tool for further processing. It was built to access four samples at a time and process through the tools periodically. This pipeline was made to process four input files at a time. All methods applied by the pipeline along with tools and thresholds are discussed below.

3.3.2 Data Retrieval

For data retrieval paired end raw sample files is downloaded using NCBI SRA toolkit (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software) present in Galaxy which fetches the relevant sample files when the accession number is provided. The Sequence Read Archive (SRA) is a public repository which provides the researchers with biological sequence data and enables them to make new advancements and discoveries by making comparisons of different datasets. It stores the raw sequencing data from different high-throughput sequencing platforms along with their sequencing information (Wheeler et al., 2007). The data files retrieved were in FASTQ format.

Table 3.1: Datasets of RNA-Sequencing

Sr. No.	Accession No.	No. of Samples	Platform	Region	Disease type
1	GSE84958	30	Illumina HiSeq 2500	Birmingham	PCOS
2	GSE56087	10	Illumina HiSeq 2000	China	EC

3.3.3 Quality control

Once all FASTQ files are retrieved next step is to check quality of reads generated. For this purpose, Fast QC is used.

3.3.3.1 FASTQC

In order to check, there is no sequencing error or biases in data FASTQC apply various quality checks. FASTQC has graphical interface to show different quality measures for the input file which can spot technical problems in the data. The main functions of FASTQC are

- Importing data from files of different formats like BAM, SAM or FASTQ
- Identifying problems quickly and providing an immediate overview of problems
- Generating results in the form of summary graphs and tables for quick assessment of data
- Generate an HTML based report of results

Raw reads may contain duplicates, adapter sequences of low quality. Preprocessing is a key step in RNA-Seq analysis as it removes noise from data. If the sequences have adapter contents than there is a need to remove these for better quality of the sequences for further analysis.

3.3.3.2 Trim Galore

The sequences with poor quality were further improved by using Trim Galore tool. It is a preprocessing adaptor trimmer tool which use raw data files as input to remove adaptor content, filter the sequences with low mean quality score, with many ambiguous (N) bases and too short sequences (Krueger, 2016). Again quality check is performed on trim galore files using FASTQC tool.

3.3.4 Mapping to the reference genome

After quality assessment reads are subjected for mapping with the reference genome hg38. Alignment with reference genome is performed by using HISAT2.

3.3.4.1 HISAT2

Is a fast and sensitive spliced alignment tool for mapping NGS reads. It is much faster than other alignment tools like 50 times more than Top Hat2. It is based on Burrows –wheeler transform algorithm. It accepts files in FASTQ and FASTA format (Kim, Langmead and Salzberg, 2015).

3.3.5 RSeQC

To comprehensively evaluate different aspects and quality of RNA-Seq data before analysis again quality control is performed on HISET2 files by using RSeQC tool (Wang, Wang and Li, 2012).

3.3.6 Mark duplicate

To locate duplicate molecule, Mark duplicate examine aligned reads in supplied SAM and BAM files. Duplicates may arise during sample preparation or during sequencing via sequencer machine and must be removed to get worthy result. This tool works by comparing sequences in the 5 prime positions of both reads. It produces a metrics file indicating the numbers of duplicates for both single- and paired-end reads.

3.3.7 RmDup

Duplicates that have been marked is removed by using RmDup. It retains the reads having good quality score (Ebbert *et al.,* 2016).

3.3.8 StringTie

Is a highly efficient and fast assembler of RNA –seq alignment into potential transcripts. It uses a novel network flow algorithm as well as an optional *de novo* assembly step to assemble and quantitate full-length transcripts representing multiple splice variants for each gene locus.

It takes the input of BAM file that came from HISAT2.The reference annotation file in form of GTF/GFF3 is also required. As an output, it gives transcript assembled file and gene abundance file in the form of GTF. In order to identify differentially expressed genes StringTie's output is processed by specialized software like Ballgown (Pertea *et al.*, 2016).

3.3.9 Ballgown

Tools for statistical analysis of assembled transcriptomes, including flexible differential expression analysis, visualization of transcript structures, and matching of assembled transcripts to annotation. Ballgown is a program that can be used to visualize the transcript assembly on a gene-by-gene basis, extract abundance estimates for exons, introns, transcripts or genes, and perform linear model–based differential expression analyses (Frazee *et al.*, 2014). The fragment per kilo base of transcript per million mapped reads (FPKM) values are used to compare gene expression level.

3.3.10 Enhanced Volcano plot

Volcano plot give overview of differentially expressed genes, we create a volcano plot, which is commonly used to summarize the results of a differential expression analysis in a single figure (Blighe, 2019)It arranges genes according to the dimension of both biological and statistical significance.

3.4 COMPARATIVE ANALYSIS

Comparative analysis is performed on the basis of differentially expressed genes to find out common and uncommon genes it is performed through Bioinformatics and Evolutionary Genomics tools which use gene name as input file to provide Venn diagram that describes the comparison between datasets

3.5 PATHWAY ANALYSIS

After finding differentially expressed genes and co-expressed genes next step is to find underlying functions of these genes. Pathway analysis is a powerful tool to interpret complex genomic data because genes do not work alone, but in an intricate network of interactions. This is essentially a multi-dimensional representation of the data (García-Campos, Espinal-Enríquez and Hernández-Lemus, 2015). There are different databases like Kyoto Encyclopedia of Genes and Genomes (KEGG, DAVID that reveal the association of input genes with different physiological and disease related pathways here, we are using DAVID for pathway analysis.

3.5.1 DAVID (**Database for Annotation, Visualization, and Integrated Discovery)**

is an integrated biological knowledge base and analytic tools aimed at systematically extracting biological meaning and analyze the function and pathway enrichment for DEGs from large gene/protein lists in the modules experiments (Huang *et al.,* 2007)

3.6 SYSTEMS BIOLOGY

Systems biology is an inter disciplinary field which uses computational and mathematical modeling approaches to study complex biological and molecular networks (Babu et al., 2006). Biological and molecular networks involve, metabolic, genetic and signaling networks.in each network interaction depends upon the molecular procedures like complex development, transcriptional regulation, chemical reaction and cellular response to stimuli. It provides insights into modeling and analysis of quantitative and qualitative nature of complex networks. Systems biology approaches can help scientist in determining the overall behavior and functioning of a system (Wingreen and Botstien).

3.6.1 SimBiology

is an application of MATLAB that is used to model complex networks (Huang *et al.*, 2007). The model is built in SimBiology by using entity/species which represent DNA, RNA, complexes etc. The entities are connected through reaction and such reaction rate can be calculated. We can perform different tasks i.e. simulate the data to check the behavior of entities with respect to time and sensitivity analysis to determine the influence of each component on overall dynamics of pathway.

Chapter 4 RESULTS

The goal of this study is to identify the potential therapeutic targets of PCOS and endometrial cancer through microarray and RNA-Seq expression profiling. Datasets of different areas and platforms have been used in order to identify differentially expressed genes. The presence of genes in different metabolic and signaling pathways associated with PCOS and EC have been checked through pathway analysis using David. Comparative analysis on the basis of identified pathways gives a common pathway among all datasets. A systems level model is designed by joining entities to understand the important processes of cell mobility and proliferation. By applying different approaches of systems biology, the dynamic behavior of a model is determined and changes in concentration of entities affecting a signaling pathway are identified with respect to time. Such information will be helpful to find potential targets associated with diseases.

4.1 MICROARRAY ANALYSIS

is performed for analyzing differentially expressed genes involved in PCOS and endometrial cancer. This process starts with quality control and normalization in order to avoid any false positive results and biasness A published pipeline based on R language is used to perform microarray analysis. For quality control oligo package is used in order to check the quality of raw data. After estimation of quality of raw data, normalization is performed. Quantile normalization method is used to normalize the data in order to avoid any biasness. RMA (robust multi array average) method is used for background adjustment, summarization and normalization of microarray data. Then again quality assessment on calibrated data is performed and clustering heatmap is produced to see how well the samples cluster. After that intensity based filtering is performed by using the *Limma* (a package used for the differential expression analysis). Before fitting linear model, we annotate transcript cluster by adding feature data to it. At the end to analyze the differentially expressed genes between normal and disease samples linear model is fit which uses appropriate design and contrast matrices and fits model to each gene separately. For visualization of the differentially expressed genes, a volcano plot is created to summarize the results of a differential expression analysis in a single figure. Result of each dataset is explained briefly below.

4.1.1 Poly Cystic Ovary Syndrome (PCOS)

4.1.1.1 Dataset *1*

First dataset of PCOS with accession no E-GEOD-34526 is obtained from Array Express. This data set contains 3 samples of normal and 7 samples of disease. After importation of related raw data, its nature is analyzed for possible anomalies that may occur due to errors while performing microarray analysis.

Quality control is performed on raw data to check normalization. PCA, box plot, RLE plot and heatmap is generated. PCA plot shows the expression of data in the form of cluster. Gene expression values are the observations and experimental conditions represented as variables. Here, it indicates the condition of PCOS and normal samples.

PCA of raw data depicts that there is a need to normalize the data for better understanding of cluster analysis among normal and disease samples because normal condition is not differentiated from disease condition. PCA plot of raw data is shown in Figure 4.1. Box plot of raw data describes that medians of all samples does not lie at the same level and same in case of RLE. Therefore, normalization is to be performed on raw data.

PCA of normalize data depicts that samples of normal is differentiated from disease sample. Different color palate is used to characterize normal condition and disease condition. It also represents the separate cluster for normal and disease samples. X-axis represents PC1 and Y-axis represents PC2. According to PCA, PC1 shows more variations of 43.6% as compared to PC2 which is 26% as shown in Figure 4.1

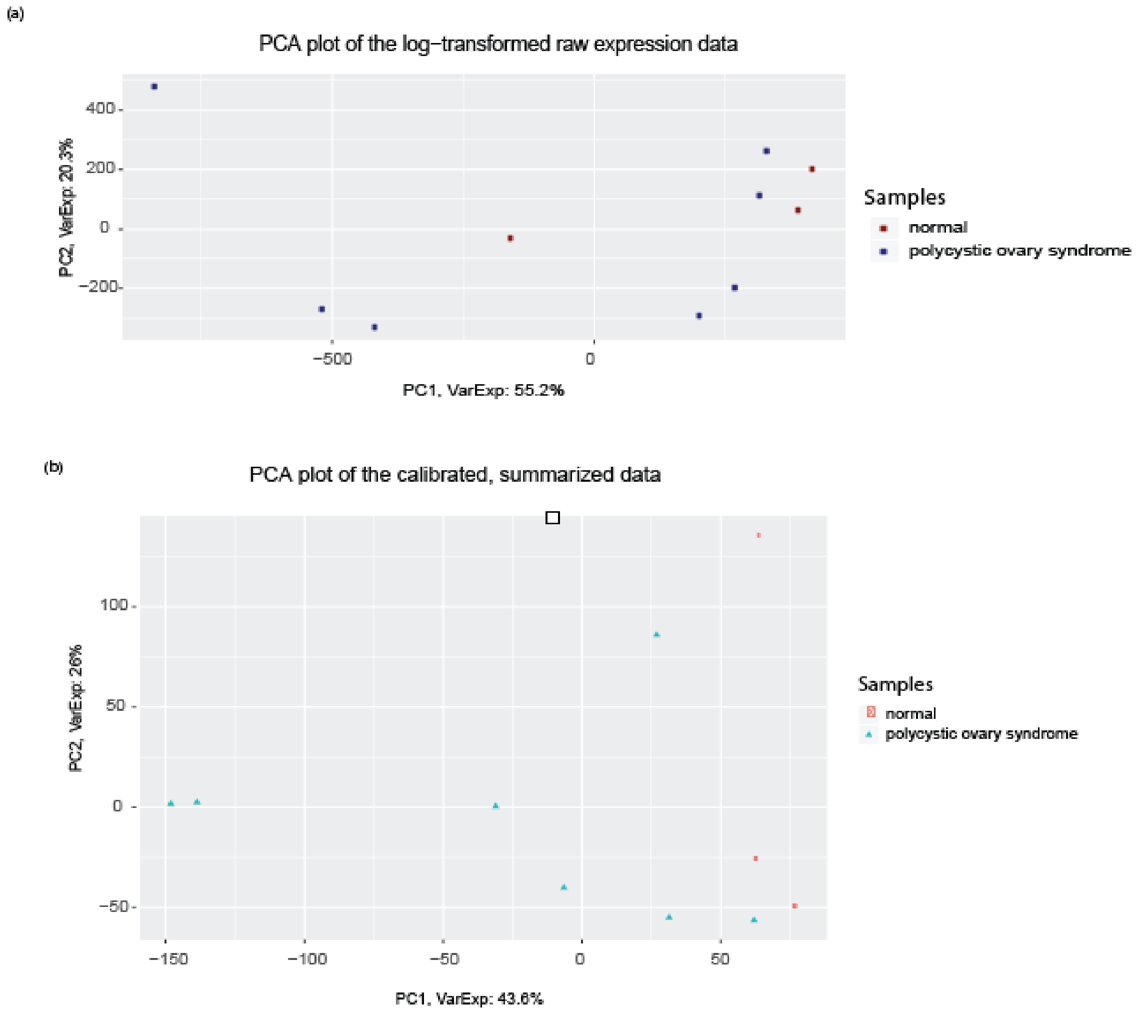

Figure 4.1: (a) PCA of Raw data, (b) PCA of Normal data

According to Figure 4.2 Box plot of quantile normalized data describes the expression of genes whereas X-axis represents samples and Y- axis shows intensity values. Box plot shows that all samples follow the same median distribution and lie at same point which is “4”. Black line represents the median value.

RLE (Relative log expression) is another graphical representation of quality control which is used to calculate the median log2 intensity values of every genes across all arrays. According to RLE of normalized data medians of all the samples lie approximately at “0”., X-axis shows samples and Y-axis represents log2 expression deviation values. It describes the expression values according to “0” as how many boxes lie above and below it as shown in Figure 4.2

Heat map is generated after annotation to describes the distances and clustering among samples. In heatmap light color tiles represents the larger distance and dark color shows small distance among samples. It describes Manhattan distance among samples. Large distance among samples shows low correlation and small distance shows high correlation Diagonal tiles show in gray color represents that there is no distance of samples with each other. According to heat map, one of the normal sample (GSM850527) shows larger distance as compared to other sample of normal conditions as shown in Figure 4.2.

Volcano plot is used to visualize the differential expression of genes. In volcano plot X-axis represents the $\log_2$FC and Y- axis shows $-\text{Log}_{10}$ p values. According to volcano plot, genes towards positive value shows upregulation and genes towards negative values shows downregulation of genes. Red color dots represent that these genes are significant and differentially expressed while blue color dots show only significant genes. Green color dots represent the genes that are differentially expressed but not significant and gray color dots show non-significant genes. DEGs were visualized using threshold of fold change 2 and p-values <0.05 as shown in Figure 4.2. Top 20 DEGs of PCOS are shown in Table 4.1

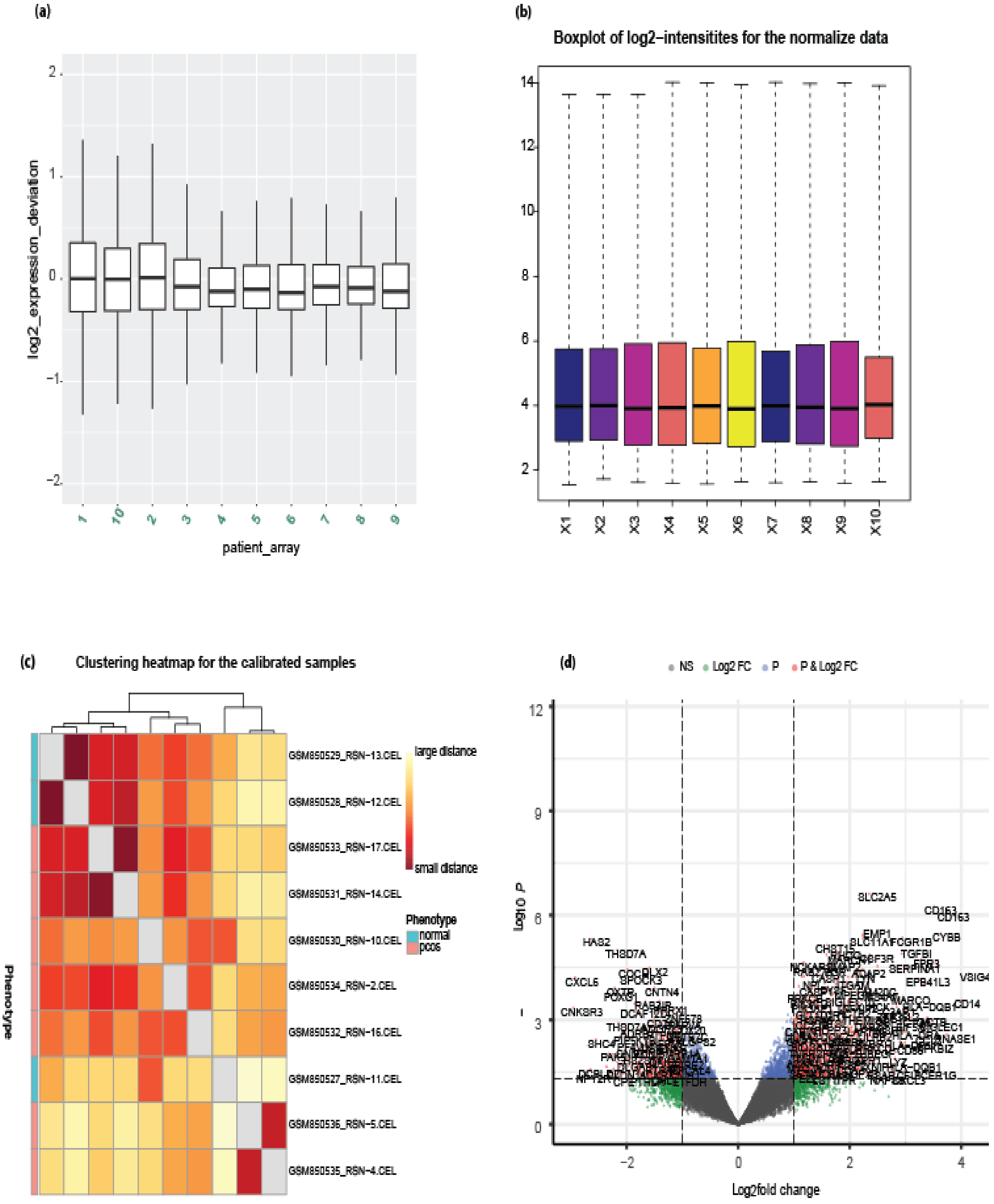

Figure 4.2: (a) RLE of normalized data, (b) Box plot of normalized data, (c) Heatmap, (d) Enhanced Volcano plot

Table 4.1: Top 20 DEGs of PCOS 1

Sr. No.	Symbol	$\log_2$ FC	P-Value
1	SLC2A5	2.347232119	2.44E-07
2	CD163	3.509026439	5.66E-07
3	CD163	3.735628621	8.44E-07
4	EMP1	2.389023268	2.77E-06
5	CYBB	3.628343535	3.39E-06
6	HLA-DQB1	2.228316901	3.63E-06
7	FCGR1B	2.943861979	4.47E-06
8	SLC11A1	2.236843611	4.54E-06
9	HAS2	-2.64165624	4.81E-06
10	MAFB	2.864050652	6.01E-06
11	CHST15	1.591082298	7.28E-06
12	THSD7A	-2.169116016	1.02E-05
13	TGFBI	3.078025017	1.04E-05
14	RHOQ	1.811728103	1.14E-05
15	SORL1	1.788960308	1.19E-05
16	FAR2	1.824238324	1.23E-05
17	SLAMF8	1.950919337	1.24E-05
18	MAFB	2.931706747	1.26E-05
19	PADI4	1.706064227	1.32E-05
20	EMB	1.580833205	1.34E-05

4.1.1.2 Dataset 2

Second dataset of PCOS having accession no E-GEOD-6798 is obtained from Array Express. It contains total 29 samples including 13 of normal and 16 of disease samples. Quality control is performed on raw data. PCA, box plot and RLE of raw data depicts that there is a need to normalize the data for better understanding of nature of data.

PCA of normalize data depicts that samples of normal is not well differentiated from disease sample and there are not much variations among them but much improved than PCA of raw data. It also represents that clustering among both samples is not well define. According to PCA, PC1 shows more variations of 32.6% as compared to PC2 which is 11.4% as shown in Table 4.3.

Box plot of normalized data shows that all samples follow the same median distribution and lie at same point which is approximately around “5.7” and we can interpret that our data has been normalized as shown in Table 4.3.

RLE of normalized data depicts that medians of log2 expression deviation values of all the samples lie approximately around “0” as shown in Table 4.3.

Heat map describes the clustering and distance among samples. light color stripes represent the larger distance and dark color shows small distance among samples. gray color along diagonal represents no distance among samples. According to heatmap shown in figure, it is observed that sample do not cluster strongly confirming the impression from the PCA plot as shown in Table 4.3.

To visualize and summarize the results of the differential expression of genes **Volcano plot** is created. X-axis represents the $\log_2$FC and Y- axis shows $-\log_{10}$ p-values. Red color dots represent those genes that are significant and differentially expressed while blue color dots show only significant genes. Green color dots represent the genes that are differentially expressed but not significant and gray color dots show non-significant genes. DEGs were visualized using threshold of fold change 0.5 as shown in Table 4.3.

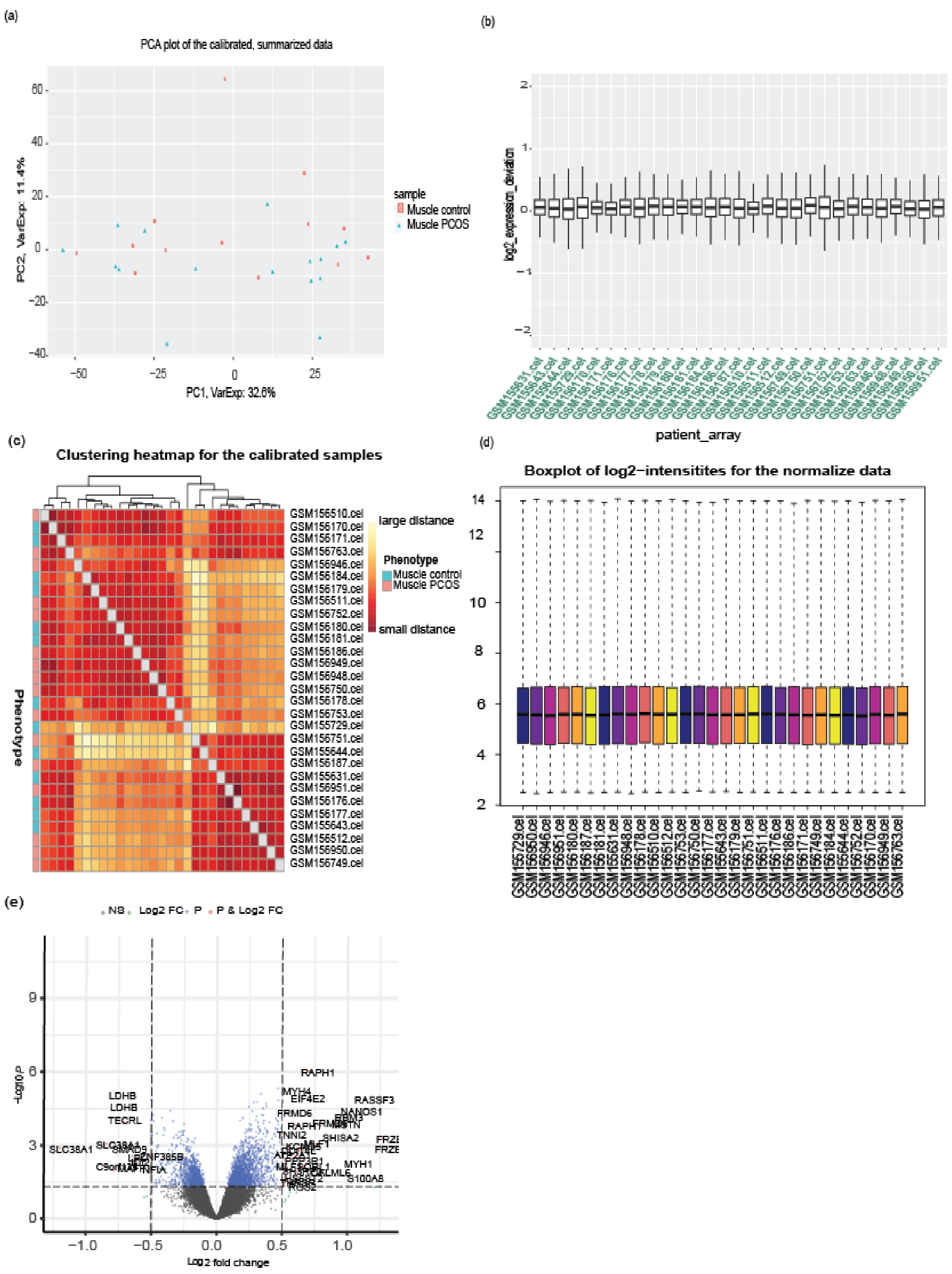

Figure 4.3 : (a) PCA plot of normalized data, (b)RLE plot of normalized data, (c) Heatmap, (d) Boxplot, (e) Enhanced Volcano

Top 20 DEGs are shown in Table 4.2

Table 4.2: Top 20 DEGs of PCOS 2

Sr. No.	Symbol	$\log_2$ FC	P-Value
1	RAPH1	0.7095638	9.64E-07
2	MYH4	0.5611468	5.81E-06
3	LDHB	-0.7718302	7.79E-06
4	EIF4E2	0.632862	1.15E-05
5	RASSF3	1.1297739	1.24E-05
6	LDHB	-0.761657	2.39E-05
7	NANOS1	1.0278984	3.94E-05
8	FRMD6	0.5284531	4.36E-05
9	RBM3	0.9585737	6.09E-05
10	TECRL	-0.7670667	8.87E-05
11	LPL	-0.7063329	0.00010782
12	FRMD6	0.8002285	0.0001183
13	RASSF3	0.92207	0.00012613
14	MSTN	0.9362215	0.00012857
15	RAPH1	0.6079741	0.00015283
16	TNNI2	0.5174287	0.00034889
17	TMEM182	0.5957761	0.000401
18	SHISA2	0.8785306	0.00041291
19	SH2D1B	0.8738964	0.00043128
20	FRZB	1.2708399	0.00052757

4.1.1.3 **Dataset *3***

Third dataset of PCOS contains total 17 samples including 8 of normal and 9 of disease samples. Quality check of raw data depicts that there is a need to normalize the data to avoid biasness.

PCA of calibrated and normalized data shows that there are not many variations among normal and disease samples but clustering is better than PCA of raw data. In plot PC1 shows more variations of 40% as compared to PC2 which is 13.1% as shown in Figure 4.4.

Box plot of normalized data shows that all samples follow the same median distribution and lie at same point which is approximately around “2.4” and we can say that our data has been normalized as shown in Figure 4.4.

RLE plot shows that shape and medians of log2 expression deviation values of all the samples are same and lie approximately around “0”and these samples can be used for cluster analysis later as shown in Figure 4.4.

Heat map describes the distances and clustering among samples. In heat map light color tiles represents the larger distance and dark color shows small distance among samples. Diagonal tiles show in gray color represents that there is no distance of samples with each other. According to heat map, one of the normal sample (GSM850527) shows larger distance as compared to other sample of normal conditions as shown in Figure 4.4.

Summary of results of DEGs is shown in **Volcano plot**. Red color dots represent significant and genes that are differentially expressed which is analyzed using threshold of 0.5 as shown in Figure 4.4. and top 20 DEGs are shown in Table 4.3

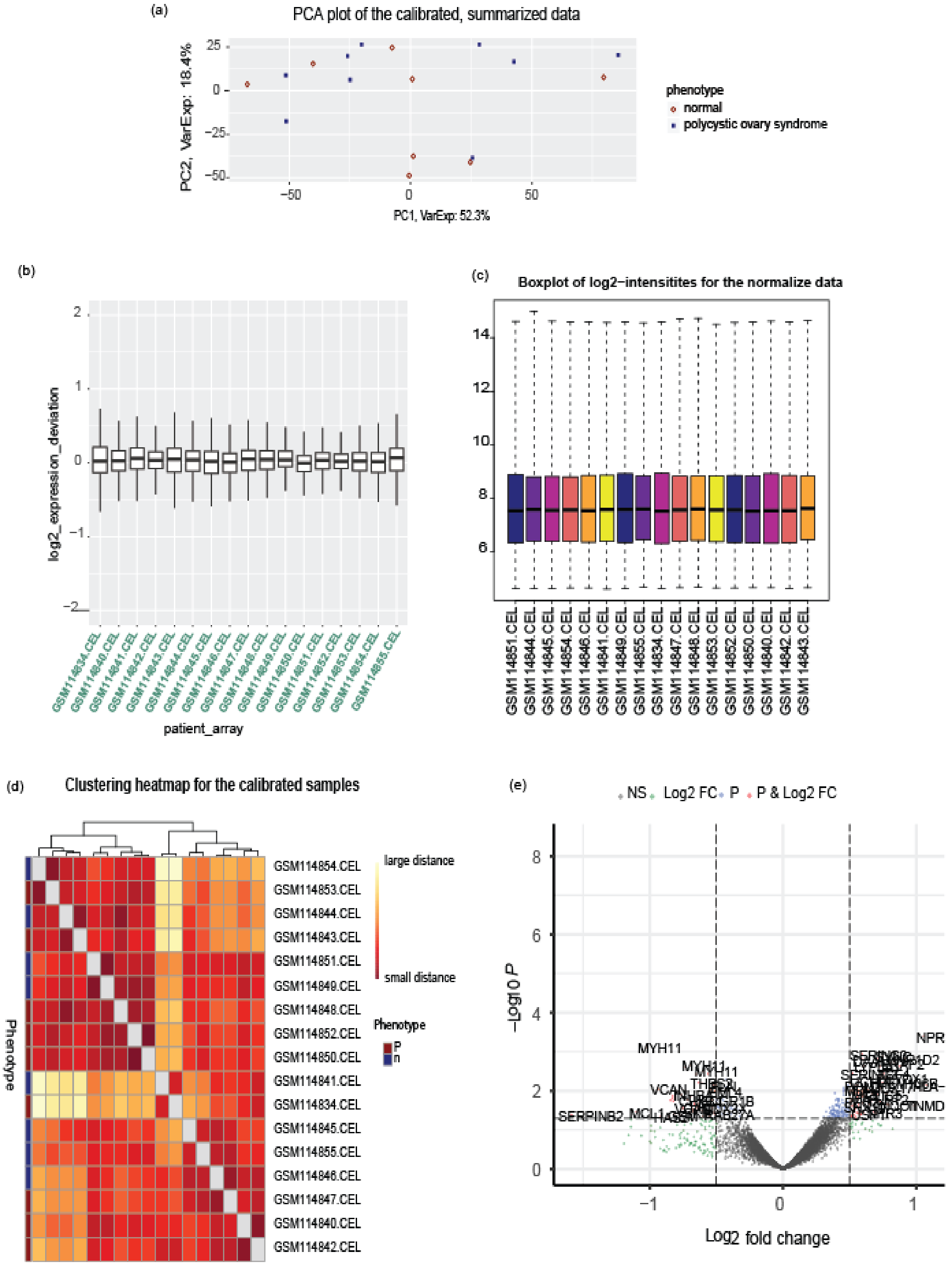

Figure 4.4: (a) PCA plot of normalized data, (b)RLE plot of normalized data, (c) Heatmap, (d) Boxplot, (e) Enhanced Volcano

Table 4.3: Top 20 DEGs of PCOS 3

Sr. No.	Symbol	log FC	p-value
1	NPR3	1.057046	0.000395
2	MYH11	-1.023	0.000773
3	SERINC3	0.591718	0.001126
4	SYNC	0.748062	0.001334
5	NR1D2	0.914403	0.001516
6	BBX	0.710622	0.001587
7	DAAM1	0.594814	0.001601
8	KDM5A	0.608339	0.001617
9	PKN2	0.835201	0.001779
10	ERAP2	0.803183	0.00198
11	MYH11	-0.69173	0.002215
12	CTSZ	0.72489	0.002367
13	TGFBR3	0.880228	0.002375
14	LYPLA1	0.604413	0.002406
15	PLN	-0.73169	0.002523
16	DNAJB14	0.58432	0.003061
17	MYH11	-0.59747	0.00318
18	SERINC3	0.521114	0.003565
19	TCF4	0.755709	0.003657
20	HNRNPH1	-0.56025	0.003688

4.1.1.4 Dataset *4*

Fourth dataset of PCOS consist of total 12 samples including 6 of normal and 6 of disease samples. After quality control it was analyzed that there is need of normalization.

After seeing **PCA** of normalize data we can interpret that samples of normal is differentiated from disease sample both have separate cluster. Blue color represents disease and orange color shows normal sample which is shown in figure. On X-axis PC1 and on Y-axis PC2 is plotted. According to plot, PC1 shows more variation of 15.6% as compared to PC2 which is 13.8% as shown in Figure 4.5.

Box plot of normalized data depicts that all samples follow the same median distribution and lie at same point which is approximately around “2.5” which shows that data is normalized as shown in Figure 4.5.

RLE of normalized data depicts that medians of log2 expression deviation values of all the samples lie approximately around “0” as shown in Figure 4.5.

Heat map shows that there is a larger distance among samples so clustering is not proper among them. Blue color represent disease and orange shows normal samples. Two samples of disease GSM136527, GSM136560 and one sample of normal condition GSM136526 is showing larger distance as compared to other samples as shown in Figure 4.5.

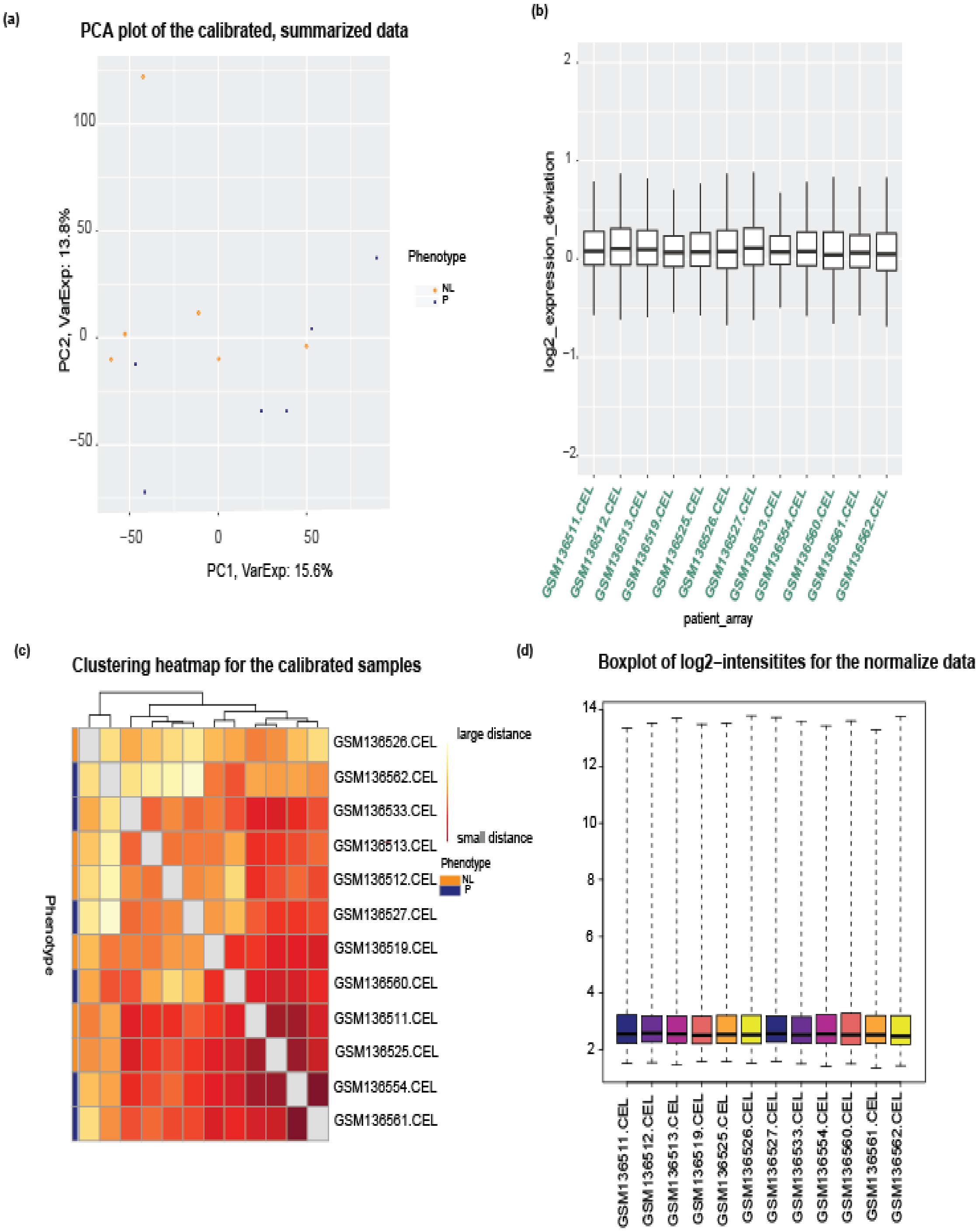

Figure 4.5: (a) PCA plot of normalized data, (b)RLE plot of normalized data, (c) Heatmap, (d) Boxplot

Significant and differentially expressed genes are visualized using threshold of” 2” in **volcano plot** as shown in Figure 4.6

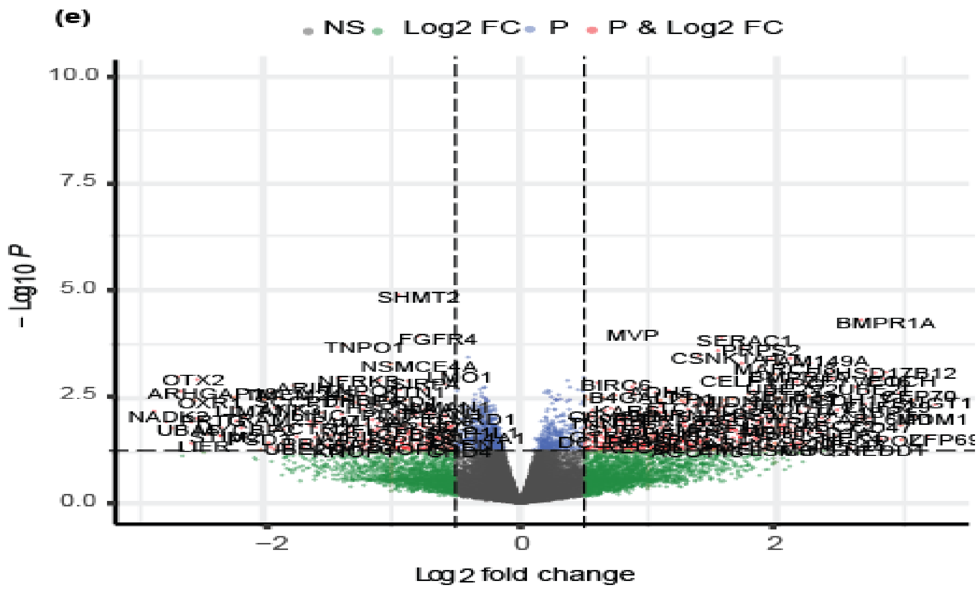

Figure 4.6: (e) Enhanced Volcano

Top 20 DEGs are shown in Table 4.4

Table 4.4: Top 20 DEGs of PCOS 4

Sr. No.	Symbol	log FC	p-value
1	SHMT2	-0.94707	1.27E-05
2	BMPR1A	2.659511	4.87E-05
3	MVP	0.781649	9.77E-05
4	FGFR4	-0.79139	0.000117
5	SERAC1	1.569803	0.000124
6	TNPO1	-1.3658	0.000186
7	PRPS2	1.726461	0.000216
8	NR2F2	1.544974	0.000248
9	CSNK1A1	1.395226	0.000329
10	FAM149A	2.107995	0.000375
11	ARHGEF7	1.956188	0.000388
12	CNBP	2.152597	0.000456
13	RNF10	2.256027	0.000483
14	FGFR1OP2	2.001186	0.000484
15	TRIM68	1.72983	0.000487
16	NSMCE4A	-1.00993	0.000501
17	8-Mar	1.868997	0.000593
18	HSD17B12	2.696975	0.000719
19	KIAA1671	2.434818	0.000737
20	LARP4	2.852178	0.000783

4.1.2 Endometrial Cancer

4.1.2.1 Dataset *1*

First dataset of endometrial cancer attained from Array Express. This data set contains 10 sample of normal and 7 samples of disease. Quality control is performed on raw data to generate PCA, box plot, RLE plot and heatmap. PCA of raw data depicts that there is a need to normalize the data for better understanding of cluster analysis among normal and disease samples because normal condition is not differentiated from disease condition. Box plot of raw data describes that medians of all samples not lie at the same level. RLE of raw data also represents that medians of all samples not lie at point “0”. Therefore, normalization is need to perform on raw data.

PCA of normalize data depicts that samples of normal is differentiated from disease sample. Different color palate is used to characterize normal condition and disease condition. X-axis represents PC1 and Y-axis represents PC2. According to PCA, PC1 shows more variations of 33.2% as compared to PC2 which is 22.3% as shown in Figure 4.7.

RLE (relative log expression) is another graphical representation of quality control which is used to calculate the median log2 intensity values of every genes across all arrays. According to RLE of normalized data, medians of all the samples lie approximately at around “1.5” but two samples shows a little variation which can be considered as outlier as shown in Figure 4.7.

Heat map defines sample to sample distance and clustering among them. light color tiles represent the larger distance and dark color shows small distance among samples. Here, blue color phenotype shows disease and pink color represent samples of normal condition. Gray color tiles across the diagonal represents that there is no distance among samples. According to heat map shown in figure two of the disease sample (GSM1555094, GSM1555091) shows larger distance as compared to other disease sample as shown in Figure 4.7.

Box plot of quantile normalized data shows that all samples follow the same median distribution and lie at same point which is “6.5” which shows data has been normalized as shown in Figure 4.7.

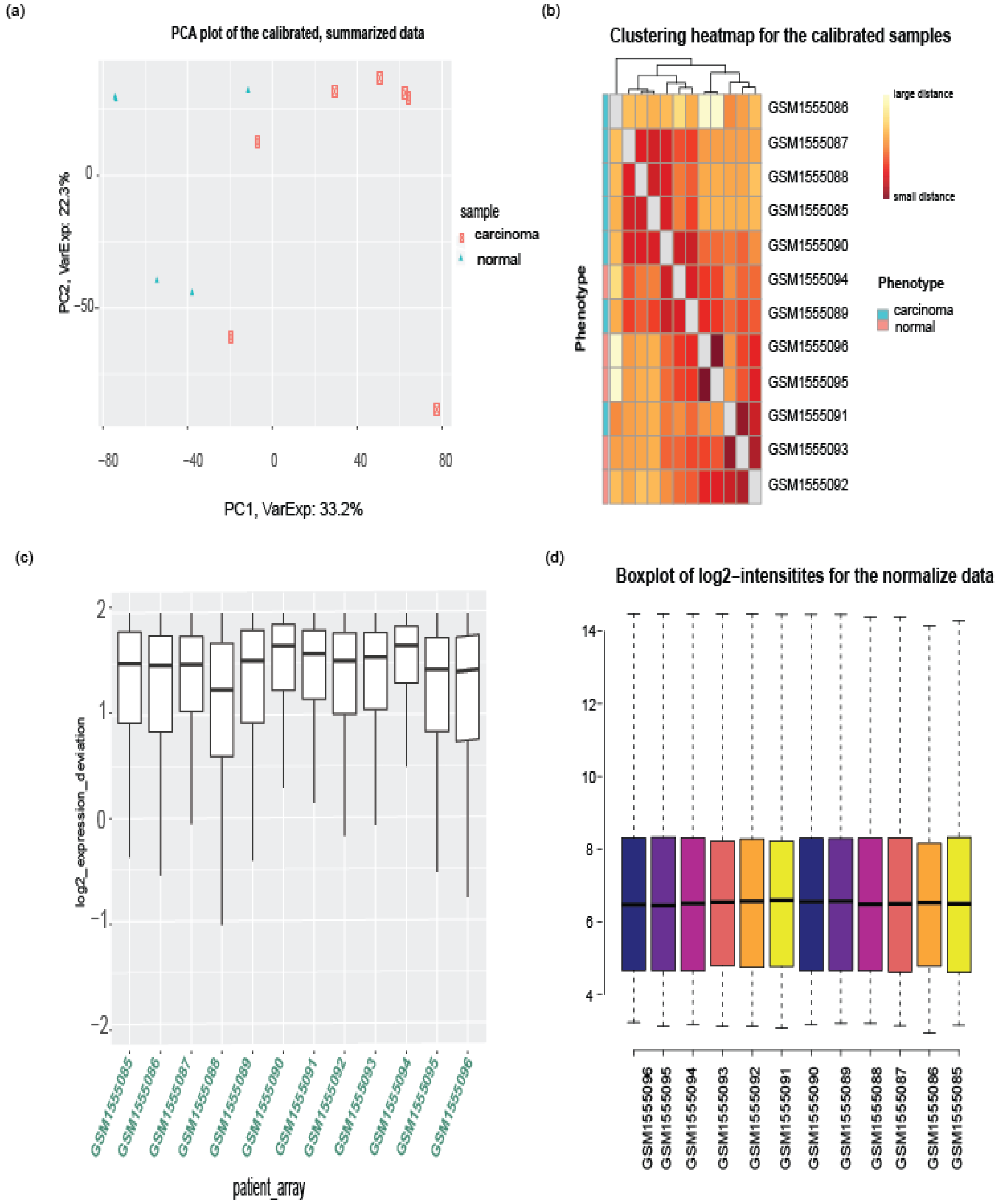

Figure 4.7: (a) PCA plot of normalized data, (b)RLE plot of normalized data, (c) Heatmap, (d) Boxplot

To visualize and summarize the results of the differential expression of genes **Volcano plot** is created. Genes towards positive value shows upregulation and genes towards negative values shows downregulation of genes. Significant and differentially expressed genes are visualized using threshold of “1” as shown in Figure 4.8.

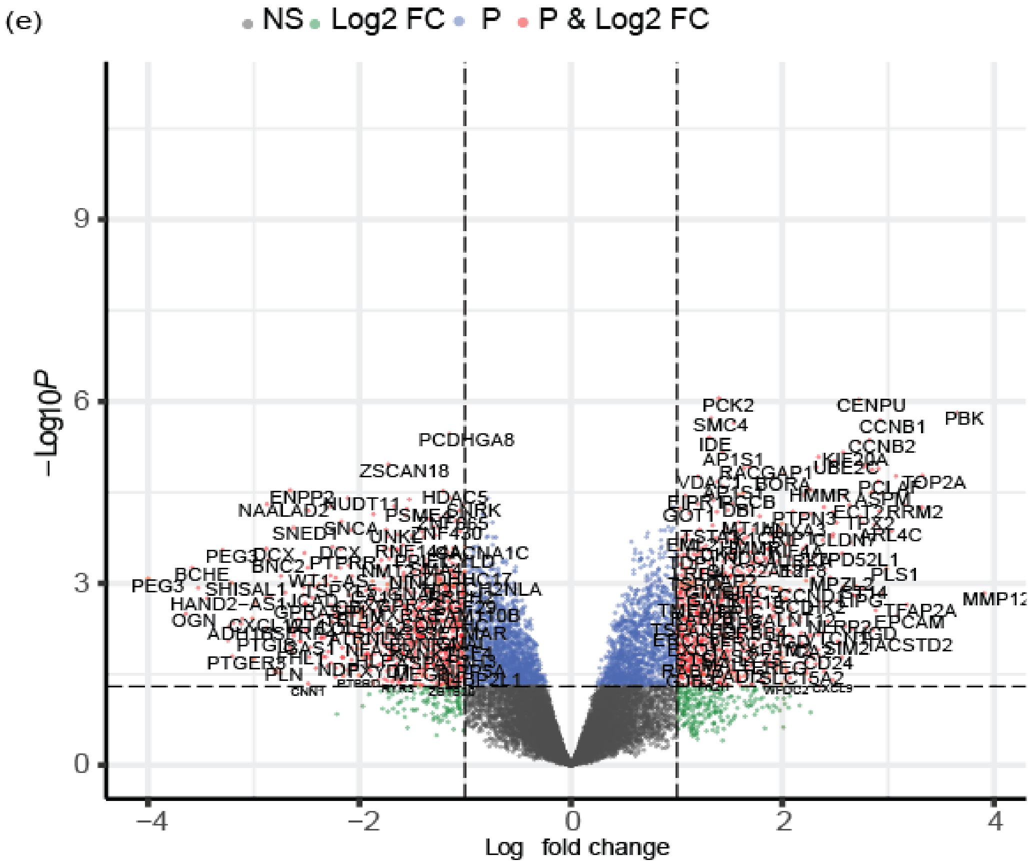

Figure 4.8: (e) Enhanced Volcano

Top 20 DEGs are shown in Table 4.5

Table 4.5: Top 20 DEGs of EC 1

Sr. No.	Symbol	Log FC	p-value
1	PCK2	1.397524111	8.82E-07
2	CENPU	2.724203176	9.27E-07
3	PBK	3.653914248	1.52E-06
4	SMC4	1.322774847	1.86E-06
5	CCNB1	2.927956801	2.04E-06
6	IDE	1.538315391	2.24E-06
7	PCDHGA8	-1.151049643	3.45E-06
8	IDE	1.307197554	3.92E-06
9	CCNB2	2.823835889	4.35E-06
10	ZWINT	2.725083233	4.74E-06
11	KIF20A	2.575803767	6.90E-06
12	AP1S1	1.431212616	6.97E-06
13	KIF11	2.339732151	8.17E-06
14	UBE2C	2.489716958	9.56E-06
15	TOP2A	2.720557061	1.01E-05
16	PTTG1	2.788224727	1.07E-05
17	ZSCAN18	-1.726012744	1.09E-05
18	RACGAP1	1.686288959	1.17E-05
19	MELK	2.7691245	1.19E-05
20	CENPF	2.320010251	1.21E-05

4.1.2.2 Dataset 2

Second dataset of endometrial cancer comprises 198 total samples of normal and 7 samples of disease. Quality control is performed on raw data to check the authenticity of data. PCA, box plot and RLE of raw data depicts that there is a need to normalize the data for better understanding of nature of data.

PCA of normalize data depicts that samples of normal is not well differentiated from disease sample and there are not much variations among them. It also represents that clustering among both samples is not well define. According to PCA, PC1 shows more variations of 16.1% as compared to PC2 which is 7.8% as shown in Figure 4.9.

RLE of normalized data depicts that medians of log2 expression deviation values of all the samples lie approximately around "0" as shown in Figure 4.9.

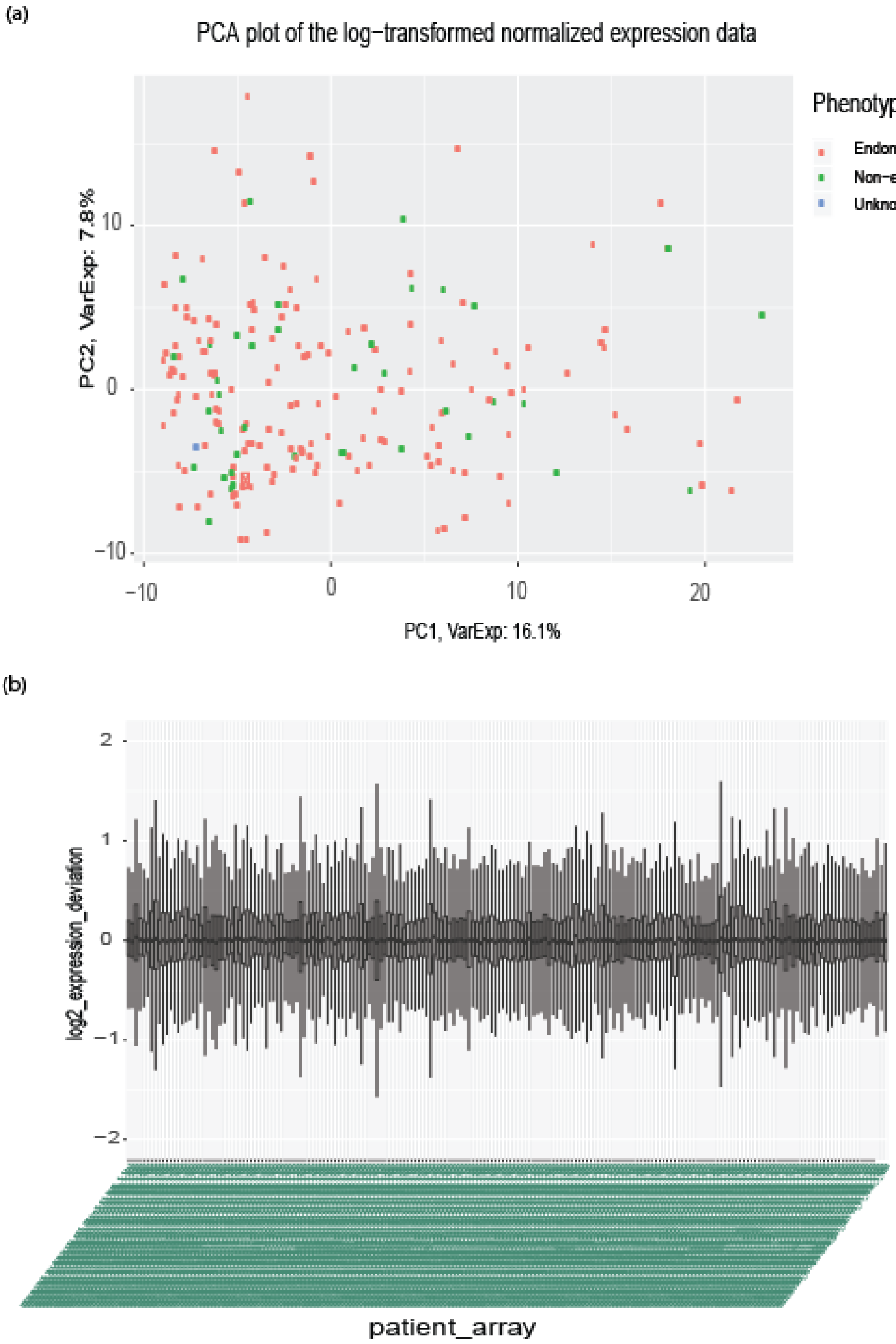

Figure 4.9: (a) PCA plot of normalized data, (b) RLE of normalized data

Here, **Box plot** of normalized data shows that all samples follow the same median distribution and lie at same point which is approximately “8” and we can interpret that our data has been normalized as shown in Figure 4.10.

Heat map describes the clustering and distance among samples According to heatmap shown in figure, it is observed that sample do not cluster and have larger distance among them confirming the impression from the PCA plot as shown in Figure 4.10.

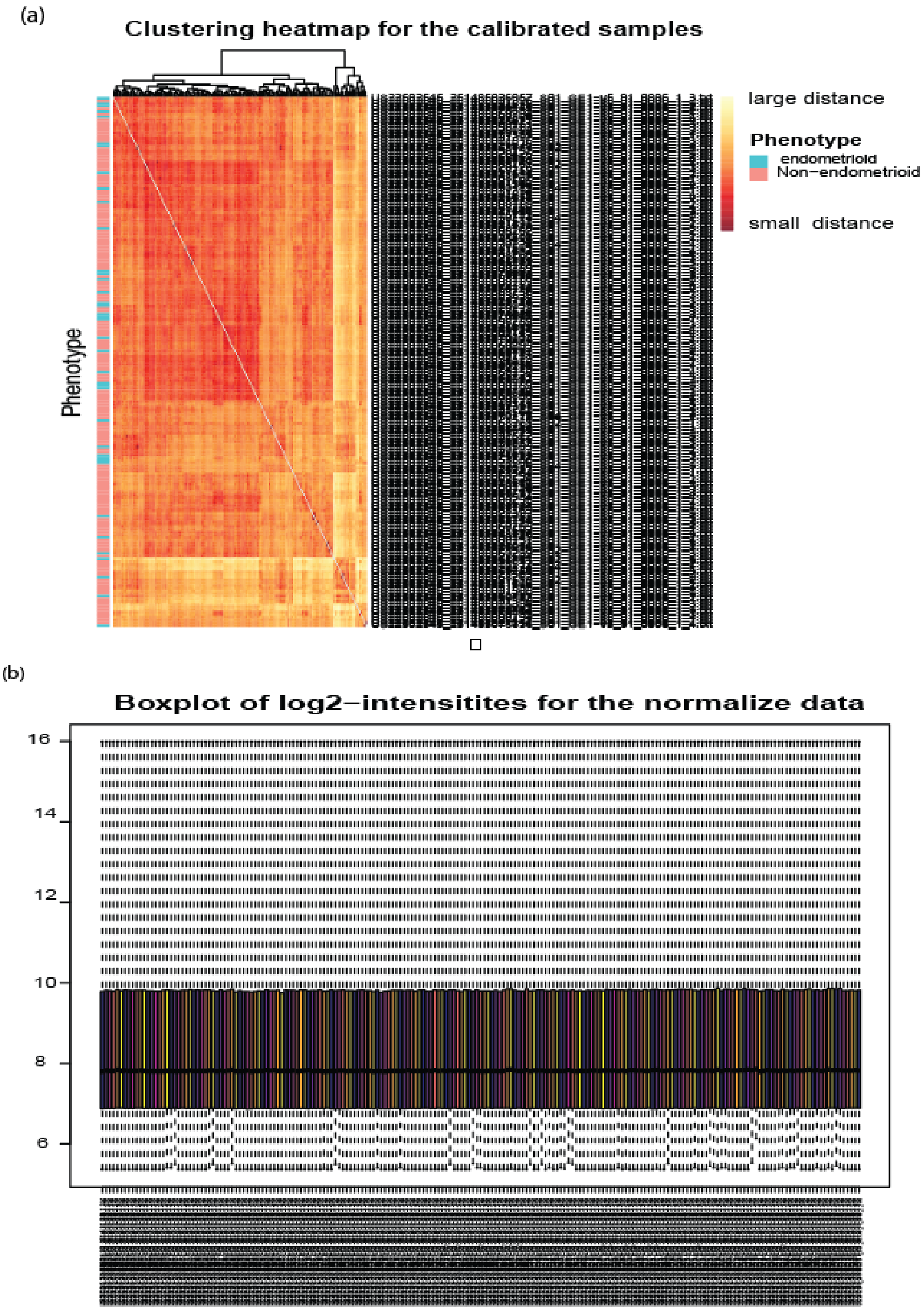

Figure 4.10: (a) Heatmap, (b) Boxplot

To visualize and summarize the results of the differential expression of genes **Volcano plot** is created. DEGs are visualized using threshold of fold change 0.5 as shown in Figure 4.11.

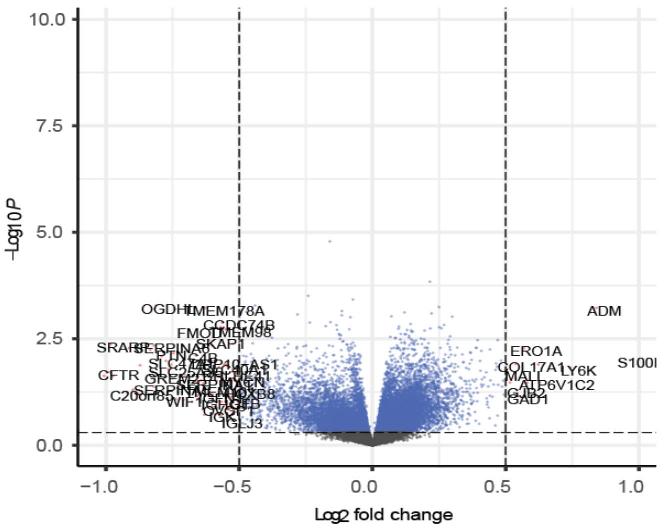

Figure 4.11: Enhanced Volcano

Top 20 DEGs are shown in Table 4.6

Table 4.6: Top 20 DEGs of EC 2

Sr. No.	Symbol	log FC	p-value
1	OGDHL	-0.81258	0.000525
2	ADM	0.840509	0.000595
3	TMEM178A	-0.62819	0.000604
4	SEMA3D	-0.74064	0.000634
5	CCDC74B	-0.56346	0.001314
6	CCDC74B	-0.57006	0.001671
7	TMEM98	-0.5503	0.001903
8	FMOD	-0.68783	0.001952
9	SKAP1	-0.61268	0.003466
10	ZNF516	-0.52477	0.004134
11	SRARP	-0.98153	0.004244
12	KCNQ1	-0.60252	0.004543
13	SERPINA6	-0.81973	0.004574
14	TSPAN8	-0.91312	0.005197
15	ERO1A	0.568664	0.005281
16	EDN3	-0.83356	0.005727
17	PTN	-0.78132	0.006558
18	HPGD	-0.74637	0.006864
19	C4B	-0.66209	0.007738
20	PTN	-0.79159	0.008096

4.2 RNA-SEQUENCING

To quantify the expression level of genes RNA-Seq analysis has become a main research tool. Many studies do not determine the genuine quality of RNA-Seq . data which can lead to misinterpretation of results here, we proposed a simple method evaluate the RNA-Seq data with mapped RNA-Seq reads.

RNA-Seq analysis include quality control, preprocessing, mapping to human reference genome (hg38) and obtaining read counts for each sample, assembling the short reads through StringTie and further for differential expression analysis by using Ballgown which is based on R.

4.2.1 Results of Poly Cystic Ovary Syndrome (PCOS)

Quality control of raw reads is performed via fastQC which point out problems in two main areas. Problems were found in per base sequence quality and sequence content at start of bases. These problems were fixed using trim galore. Trim galore checks the quality of each base and removes the bases with low average quality. Per base sequence content error occurs usually when difference between proportion of bases (A and T, G and C) at any position is greater than 20. This type of error is mostly seen at the starting of reads due to adapter ligation. Trim galore adjusts this problem by trimming bases from the start of reads.

All preprocessed reads were aligned to human reference genome (hg38) through Hiset2. After alignment the quality of aligned reads is checked through RSeQC. mark duplicate is used to mark the duplicated sequences which are further removed by Rmdup, then assembled the short reads through StringTie. After mapping and again quality check the differential expression analysis is performed using ballgown in R. different graphs are obtained through Ballgown.

PCA of calibrated and normalized data shows that there are not much variations among normal and disease samples and clustering is not proper. Different color palate is used to characterize normal condition and disease condition. In plot PC1 shows more variations of 81.9% as compared to PC2 which is 8.8% as shown in Figure 4.12.

Heat map defines sample to sample distance and clustering among them. light color stripes represent the larger distance and dark color shows small distance among samples. gray color along diagonal represents no distance among samples. According to heatmap, it is observed that sample do not cluster strongly confirming the impression from the PCA plot as shown in Figure 4.12.

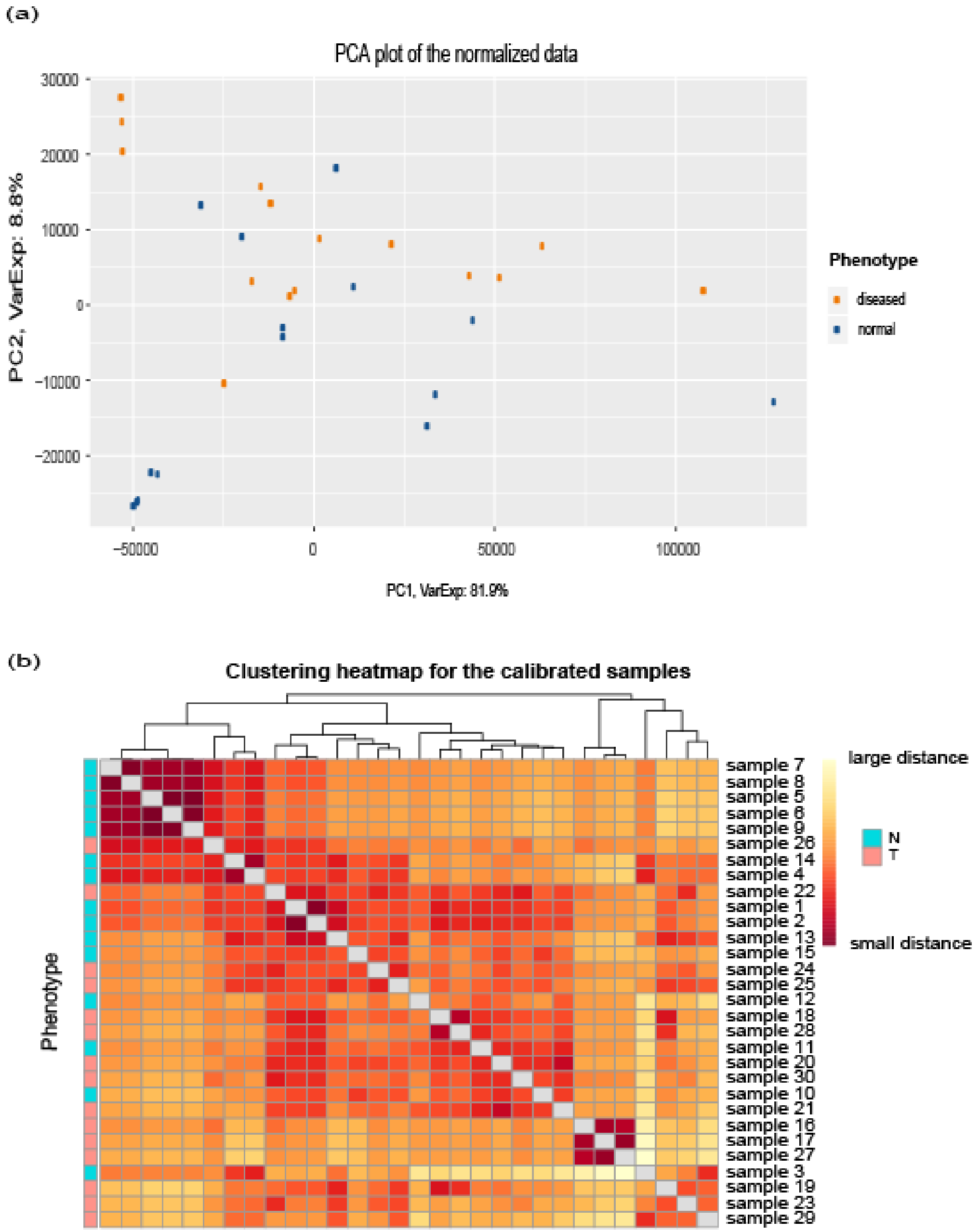

Figure 4.12: (a) PCA of normalized data, (b) Heatmap

Volcano plot is used to visualize the differential expression of genes According to volcano plot, genes towards positive value shows upregulation and genes towards negative values shows downregulation of genes. Red color dots represent that these genes are significant and differentially expressed while blue color dots show only significant genes. DEGs were visualized using threshold of fold change 1 as shown in Figure 4.13.

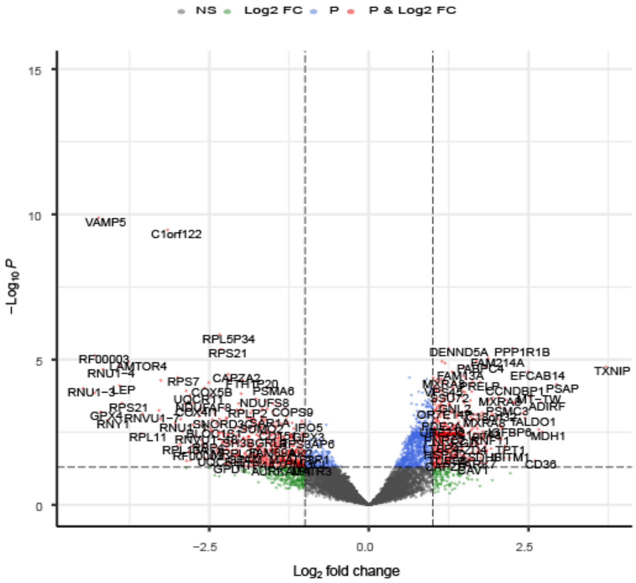

Figure 4.13: Enhanced Volcano

Top 20 DEGs are shown in Table 4.7

Table 4.7: Top 20 DEGs of PCOS

Sr. No.	Symbol	log FC	p-value
1	A4GALT	1.899251652	0.003635
2	AAMP	1.655550156	0.004576
3	AARSD1	1.688042544	0.001132
4	ABHD11	1.834476166	0.003902
5	AC004792.1	1.815440692	0.004959
6	AC005726.3	2.135744915	0.000384
7	AC006116.7	1.648870068	0.003606
8	AC006238.1	1.716650818	0.004574
9	AC007663.3	1.995945588	0.002781
10	AC008626.2	1.654356824	0.002862
11	AC009244.1	3.182155362	0.002013
12	AC010507.2	2.649056781	0.003464
13	AC010547.1	2.194295933	0.003991
14	AC011468.5	1.789314975	0.000356
15	AC011816.2	2.040255796	0.001441
16	AC016727.1	1.699969835	0.003246
27	AC017100.2	1.764360873	0.003546
18	AC027229.1	2.009845432	0.003734
19	AC060766.4	1.999268323	0.000424
20	AC073332.1	1.724050239	0.004065

4.2.2 Results of Endometrial Cancer

Quality control of raw reads is performed via fastQC which indicate problems in our data. These problems are fixed using trim galore. to adapter ligation. All preprocessed reads were aligned to human reference genome (hg38) through Hiset2 which is alignment program for mapping NGS reads and based on Burrows-wheeler transformation. RSeQC again check the quality of mapped reads. mark duplicate marked the duplicated sequences which is removed by Rmdup and assembled through StringTie. differential expression analysis is performed using Ballgown in R. Different graphs obtained through Ballgown are as following:

PCA plot is generated using the ggplot2 package in R. characterization of RNA-Seq data is determined using the gene expression (FPKM). The result shows that data of some of samples are dissimilar to other samples with in the same group (normal or disease) which indicate heterogeneous transcriptomes due to different cell population for .example two of disease samples is located far from other disease sample cluster in the gene expression PCA as shown in Figure 4.14.

Heatmap is generated by using gene expression file which depicts that clustering is not proper among both normal and disease conditions which might be due to very few number of differential expressed genes as shown in Figure 4.14.

According to heatmap, blue color represents disease condition and pink color represents normal condition. Diagonal tiles show that there is no distance of samples with each other.

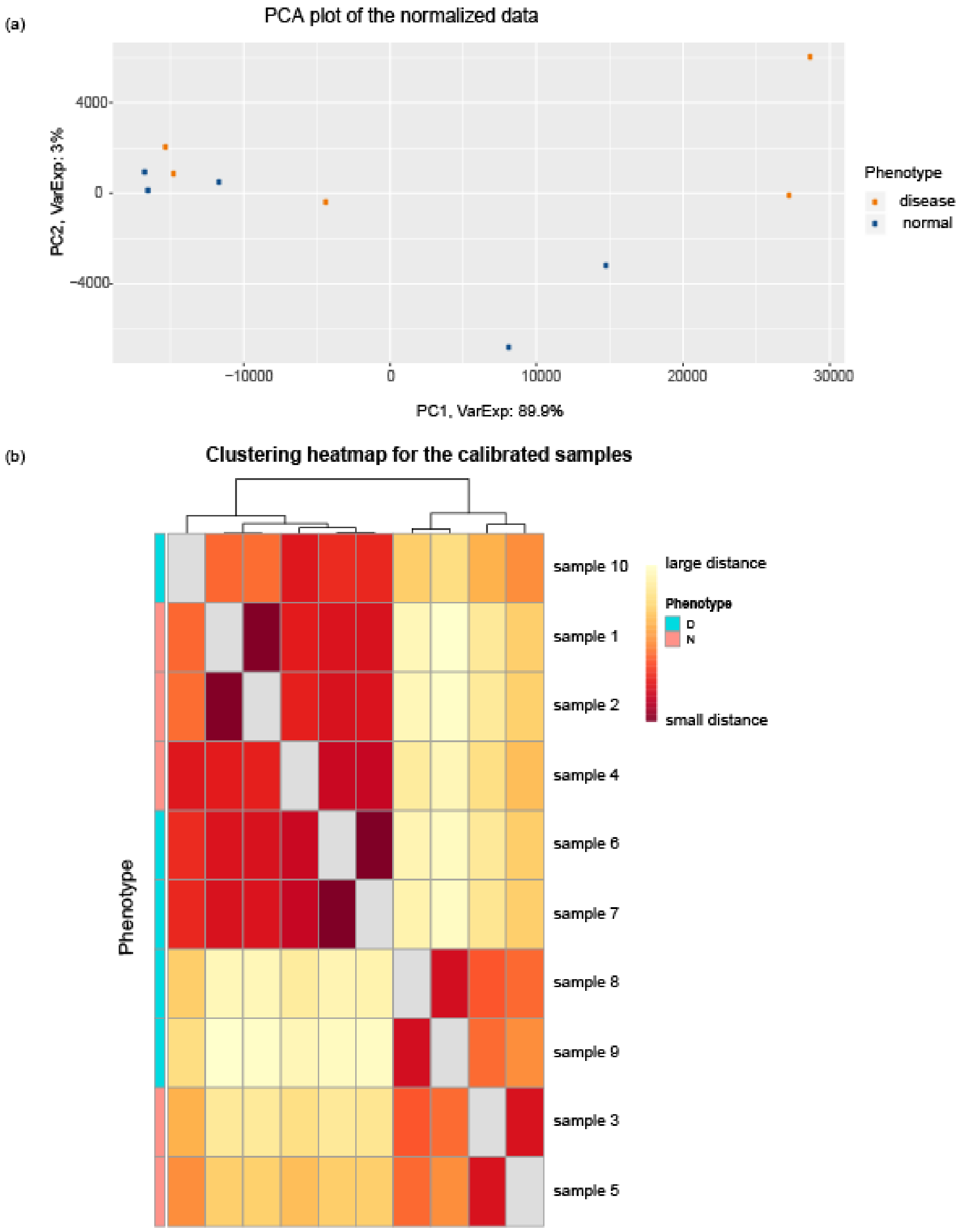

Figure 4.14: (a) PCA of normalized data, (b) Heatmap

Summary of results of DEGs is shown in **Volcano plot**. Red color dots represent significant and genes that are differentially expressed which is analyzed using threshold of 1 as shown in Figure 4.15

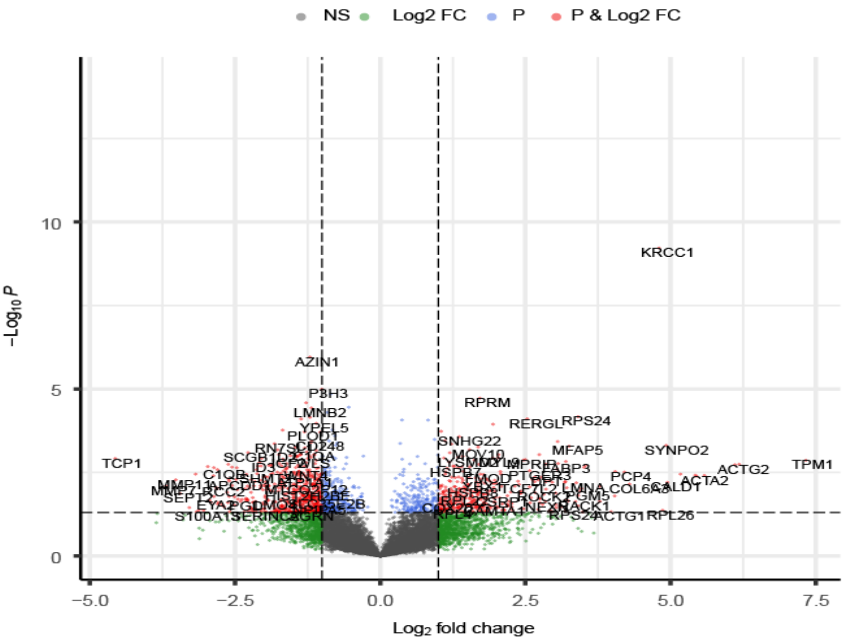

Figure 4.15: Enhanced Volcano

Top 20 DEGs are shown in Table 4.8

Table 4.8: Top 20 DEGs of EC

Sr. No.	Symbol	Log FC	p-value
1	TMEM88	2.118786	0.012055
2	NAA38	2.56104	0.009941
3	RPL26	29.10501	0.043011
4	RPLP1P11	2.093705	0.042592
5	MPRIP	5.532379	0.001346
6	TVP23B	3.403707	0.037356
7	EIF1	4.648828	0.02306
8	LSM12	2.073857	0.008363
9	AC005180.1	2.918101	0.0319
10	CDK5RAP3	2.893403	0.004585
11	LUC7L3	2.425833	0.002063
12	NT5C	3.174033	0.010235
13	ACTG1	15.69431	0.047318
14	THOC1	2.457515	0.025952
15	THOC1	2.39932	0.04037
16	COLEC12	6.149283	0.005218
17	PPP4R1	2.978643	0.024225
18	GATA6-AS1	2.153568	0.032757
19	AC027449.1	3.191513	0.012776
20	OSBPL1A	3.919042	0.005432

4.3 COMPARATIVE ANALYSIS OF DIFFERENTIALLY EXPRESSED GENES

One of the main purposes of this study is to identify similarities and differences between different datasets. In case of microarray analysis comparison of differentially expressed genes of PCOS gave three common genes but no common gene with RNA-Seq data set of PCOS. In case of Endometrial cancer 10 common genes were found in case of microarray analysis and 1 common gene in case of comparison with RNA-Seq dataset.

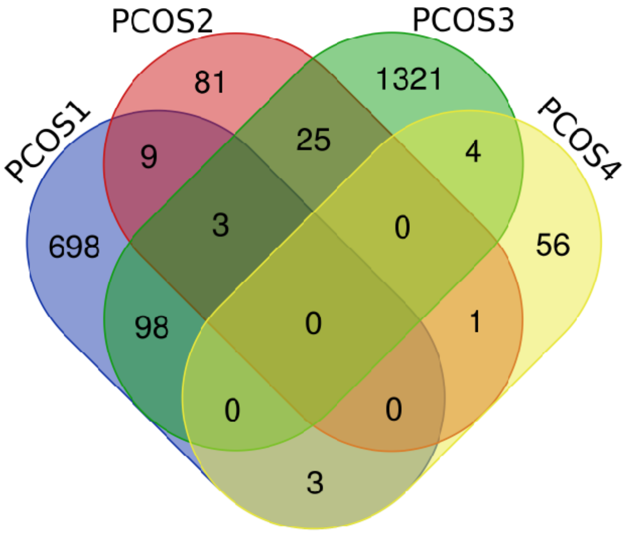

Figure 4.16: Comparative analysis of PCOSs datasets

Comparison among RNA-Seq datasets of both PCOS and EC gave 27 common genes. comparative analysis of all datasets shows that no common genes are present in all datasets (microarray, RNA-Seq). Therefore, further pathway analysis is performed on all datasets to find a common pathway.

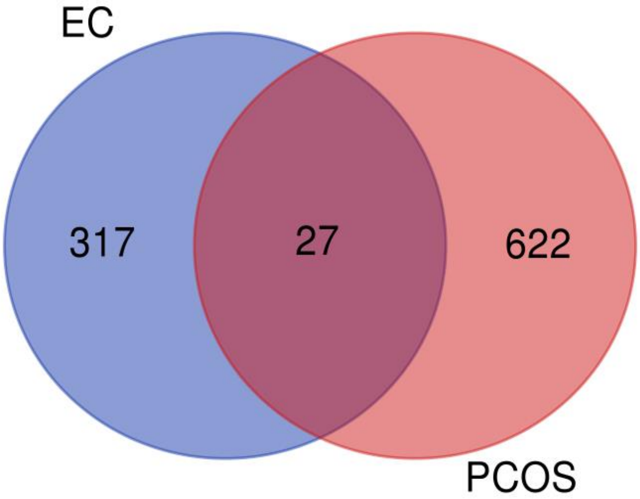

Figure 4.17: Comparative analysis of PCOS and EC

4.4 PATHWAY ANALYSIS

Pathway analysis is performed through David for all datasets of microarray and RNA-Seq datasets in order to identify metabolic and signaling pathways associated with PCOS and EC. After comparison of identified pathways, it has been seen that Focal adhesion is the common pathway among all the datasets having greater number of differentially expressed genes. A pathway has been extracted from this pathway and used for model building according to some important entities to further analyze quantitatively through systems biology approach.

Top pathways of all datasets have been shown in the following tables.

Table 4.9: Pathway analysis of PCOS 1(M.A)

Pathways	No. of Genes
Ubiquitin mediated proteolysis	15
Protein processing in endoplasmic reticulum	16
RNA degradation	9
Endocytosis	18

Table 4.10: Pathway analysis of PCOS 2(M.A)

Pathways	No. of Genes
Focal adhesion	30
MAPK signaling pathway	35
Type I diabetes mellitus	16
Cell adhesion molecules	32

Table 4.11: Pathway analysis of PCOS 3(M.A)

Pathways	No. of Genes
Jak-STAT signaling pathway	5
Amoebiasis	4
PI3K-Akt signaling pathway	7
Viral carcinogenesis	6

Table 4.12: Pathway analysis of PCOS 4(M.A)

Pathways	No. of Genes
cGMP-PKG signaling pathway	6
Calcium signaling pathway	6
Glucagon signaling pathway	4
Oxytocin signaling pathway	4

In Endometrial Cancer (EC):

Table 4.13: Pathway analysis of EC 1(M.A)

Pathways	No. of Genes
Focal adhesion	31
Cell cycle	27
p53 signaling pathway	19
Oocyte meiosis	22

Table 4.14: Pathway analysis of EC 2(M.A)

Pathways	No. of Genes
Staphylococcus aureus infection	6
Complement and coagulation cascades	6
Rheumatoid arthritis	5
Wnt signaling pathway	4

Pathway analysis of RNA-Seq datasets

Table 4.15: Pathway analysis of PCOS(RNA-Seq)

Pathways	No. of Genes
Focal adhesion	13
Ribosome	15
Vascular smooth muscle contraction	14
Oxytocin signaling pathway	13

Table 4.16: Pathway analysis of EC(RNA-Seq)

Pathways	No. of Genes
Focal adhesion	15
Carbon metabolism	15
Pathways in cancer	28
Biosynthesis of antibiotics	17

4.5 SYSTEMS BIOLOGY

For quantitative based analysis system biology approach is used. Focal adhesion pathway is selected through literature review in order to generate model. FPKM average expression values of samples for each phenotype separately retrieved from RNA-Seq datasets are used for simulation and sensitivity analysis. A pathway is designed by using simbiology which is an application of MATLAB, has been shown in Figure 4.18 to further analyze by using quantitative modeling approach of systems biology..

Pathway model indicates the inhibition and activation pattern of genes. Model shows that FN activates the complex which further activates many entities that move towards actin polymerization, P13K-Akt signaling pathway and MAPK signaling pathway and lead the path towards cell mobility and cell proliferation.

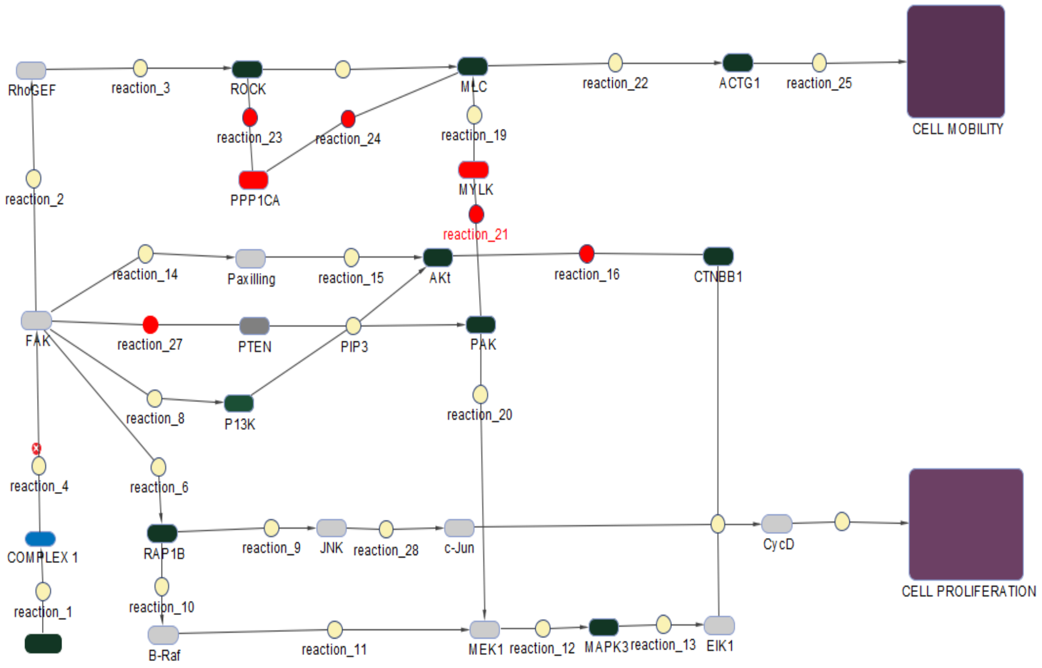

Figure 4.18: Pathway model. Green color represents the differentially expresed genes, Red color shows the inhibitor, Blue color shows the complex of two genes and Purple color shows the conditions

Simulation of models shows the dynamic behavior of entities. Graph shows increase and decrease concentration of cell mobility and cell proliferation with respect to time. x-axis shows time and y-axis shows concentration of entities

Normal state Figure 4.19 depicts that in normal condition, concentration of proliferation is up to 1.5 and concentration of cell mobility lies approximately at 8.5 with respect to time.

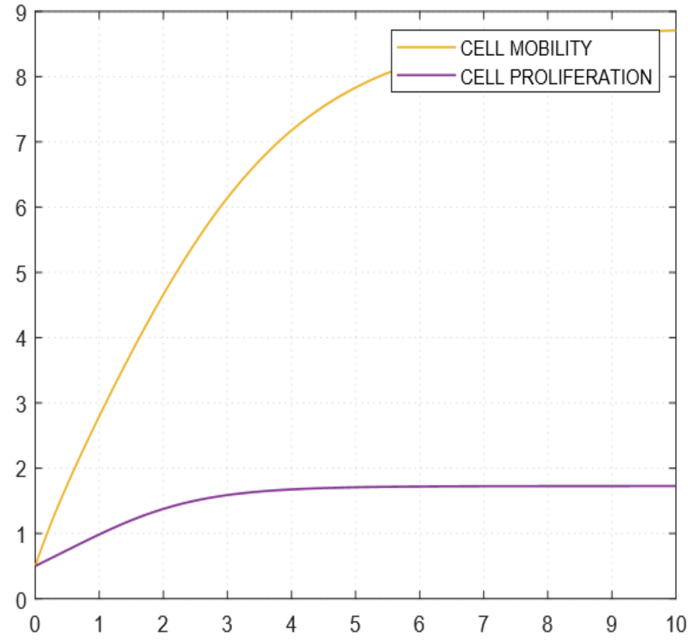

Figure 4.19: Simulation model for normal conditions

Disease state Figure 4.20, represents that in disease condition concentration of proliferation and cell mobility changes with respect to time as compare to normal state.

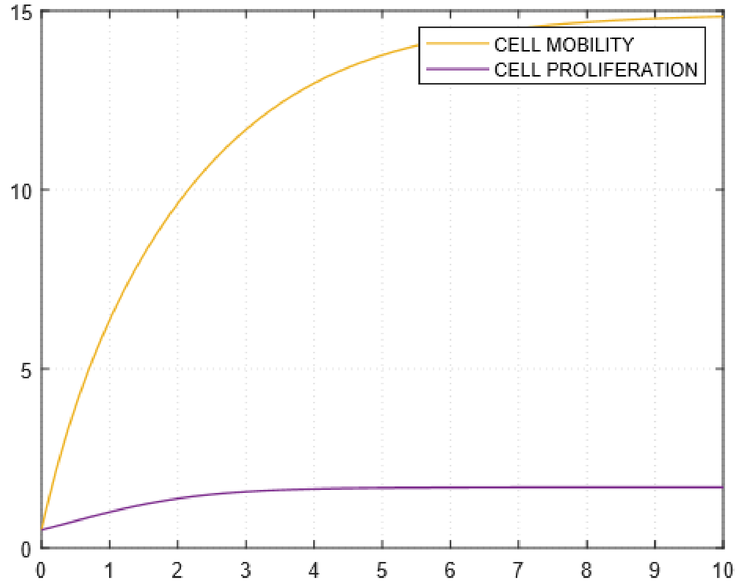

Figure 4.20: Simulation model for disease conditions

Sensitivity Analysis

In sensitivity analysis one can check the influence of entities on overall pathway. It can be seen that sensitivity of ACTG1, PPICA, ROCK and MLC has increased in cell mobility shown in Figure 4.21.

In case of proliferation CycD show high sensitivity level Figure 4.21 (Below). The x-axis represents entities and y-axis represents sensitivity of desired condition (cell proliferation).

ACTG1, PPICA, ROCK and MLC are the genes that show differential expression in our analysis. Approximately in all datasets these genes are up-regulated except ROCK which is down regulated. They show high sensitivity level according to the designed network. So, they may predict as important therapeutic targets.

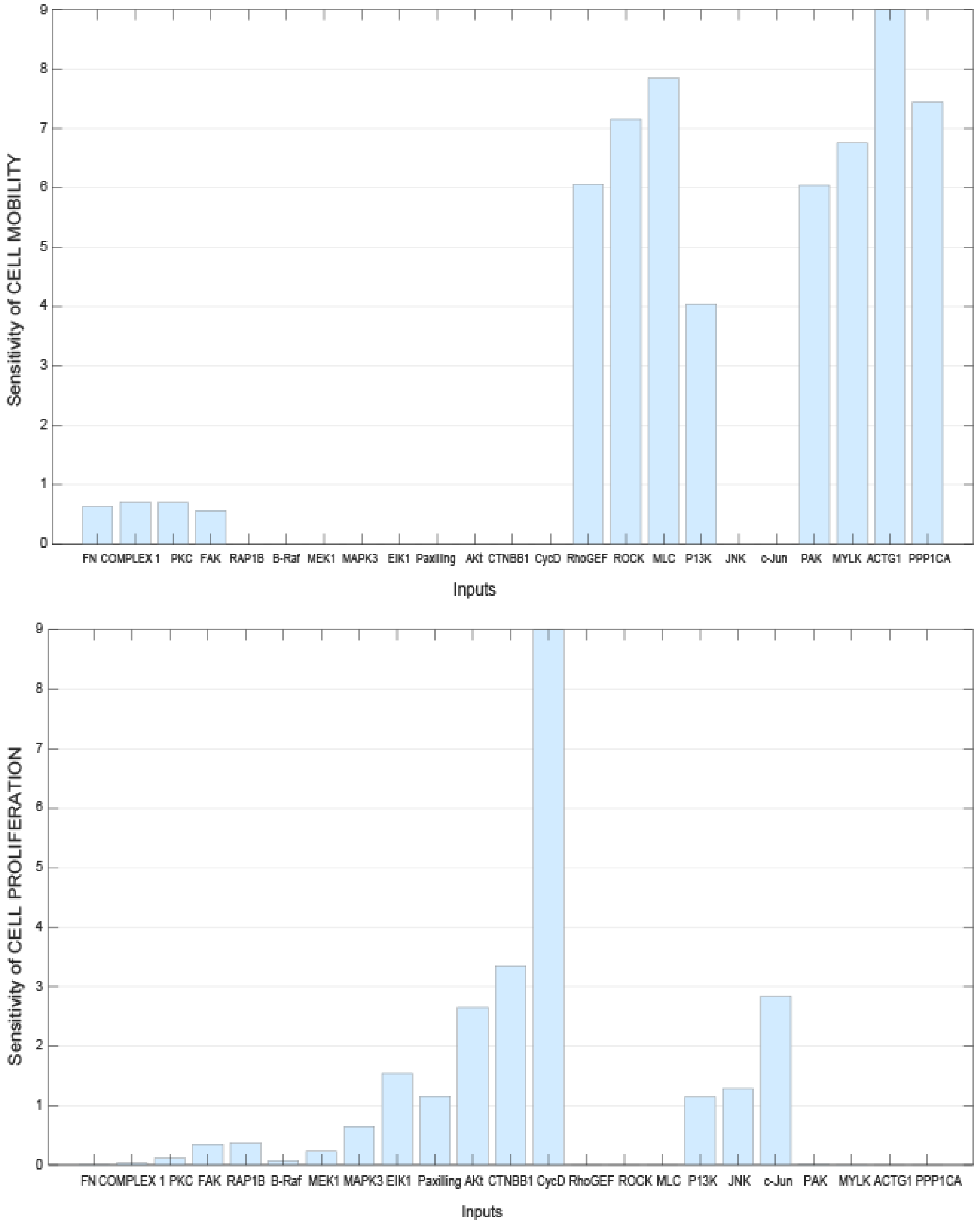

Figure 4.21: (Above) Sensitivity plot of cell mobility, (Below) Sensitivity plot of cell proliferation

Chapter 5 DISCUSSION

Poly cystic ovary syndrome is most common endocrine disorder affecting 5-11% of reproductive age and can lead towards other diseases. Different risk factors are involved which includes diabetes, obesity, Changes in the balance of female hormones in the body, metabolic disorder, anxiety and sometime heredity that can play significant role in describing the exact mechanism. Endometrial cancer is one of most common gynecological cancer associated with poly cystic ovaries. The complexity of diseases is not properly understood although many metabolic and signaling pathways were reported which elaborates the mechanism and pathogenicity of the disease to some extent. But there are some ambiguities in understanding the molecular basis of the disease associated with their causative agents. agents.

Expression profiling of genes gain importance after development in high throughput sequencing techniques as it clarifies the path for researchers to find potential therapeutic targets of the disease. These techniques make it very easy to comprehend the pathogenicity of the disease so, there is a need to completely understand the pathogenicity of the disease and to identify biomarkers and potential therapeutic targets. Different studies reported many genes as biomarkers but they didn't use multiple data of different regions and also did not use integrated sequencing analysis techniques and systems biology approaches to predict the biomarkers and therapeutic targets. This study uses different sequence analysis techniques to find similarities and differences between different expression profiles, analysis of pathways and quantitative modeling approach for the identification of therapeutic targets.

Different datasets have been used for microarray and RNA-Seq that is of different regions and platforms. Information in these datasets are further processed to retrieve some meaningful biological information. These datasets are composed of different number of samples. The purpose of using these datasets of different regions is to check any kind of genetic variations that may be different from region to region. Although microarray and RNA-Seq have already performed on selected datasets at experimental level but they did not compare their results of differential expressions with other datasets of the same disease. We performed the whole analysis again by using some computational and statistical techniques i.e. the use of different quality checks, algorithms for normalization and clustering to statistically visualize and analyze the differentially expressed genes. We have generated different plots to check the distribution and variation of samples. Box plot shows the distribution of the data between upper quartiles, medians, lower quartiles, minimum and maximum values and also give information about outliers if exist. In different dataset PCA shows that some of the diseased samples come nearer to normal ones. From statistical point of view these samples may considered as outliers but as we already discussed that the data is secondary we are not familiar about any kind of technical errors and environmental factors which affects the whole experiment. So we could not remove the samples as they may contain some worthy genes that are very important to analyze. We got differentially expressed genes after performing microarray and NGS analysis and to get a common gene among all datasets comparative analysis was performed. Comparison of DEGs of all datasets gave no common gene among both PCOS and EC so, no gene came as a potential therapeutic target and we could not rely only on comparative analysis.

Pathway enrichment analysis help to find important metabolic and signaling pathways in which these are enriched. Pathways of all datasets have been identified. Comparative analysis on the basis of pathways has been performed which give a common pathway i.e. focal adhesion. Many of the important genes in this pathway have been associated with the disease.

Makker et al. (Makker *et al.*, 2012) analyzed that PI3K/AKT pathway has been activated in PCOS patient. PI3K-Akt and MAPK are two important cell signaling pathways that are activated by steroid hormones and growth factors leading to cellular events including gene expression, cell proliferation and survival. These pathways are considered as an attractive target for the development of novel anticancer molecules, and selective inhibitors specifically targeting different components of these cascades have been developed.

Morgensztern *et. al.* (Morgensztern and McLeod, 2005) demonstrated the importance of dual role of mTOR and AKT and analyze that over expression of mTOR/AKT signaling pathway enhance the chances of tumor growth. some genetic alternation in these pathways are studied. In this study Regulation of actin cytoskeleton pathway is also observed which play role in actin polymerization and cell mobility. Actually Malignant growth cell metastasis is a multi-step process including invasion into encompassing tissue, intravasation, travel in the blood or lymph, extravasation, and development at another site. Huge numbers of these means require cell motility, which is driven by cycles of actin polymerization, cell adhesion and acto-myosin contraction. These procedures have been contemplated in malignant growth cells (Olson and Sahai, 2009)

Importance of WNT/Beta-catenin signaling pathway is also reported. There are some important protein ligands and regulatory proteins like GSK3B and Beta catenin which play crucial role in growth and development. Mutations in Beta-catenin has been reported in cancers (Matsumoto *et al.*, 2015).

Simulation of the designed network shows high concentration of cell mobility and with respect to time in diseased state. Sensitivity analysis was performed on generated model which propose potential biomarkers that could lead towards cancer. Two main process were targeted in analysis cell mobility and cell proliferation. ACTG1, PP1CA, MLC and ROCK Observed to be significant genes in case of cell mobility that could serve as biomarker that lead towards cancer. In case of cell proliferation CycD gene was more sensitive that could also act as potential biomarker. Therefore, to minimize the susceptibility of these genes should be studied and observed in future.

Chapter 6 CONCLUSION & FUTURE PERSPECTIVES

This study incorporates identification of therapeutic targets focuses by utilizing high throughput sequencing procedures. Microarray and RNA-Seq based study are performed by utilizing distinctive datasets of PCOS and Endometrial cancer. Differentially expressed genes have been distinguished. After comparative analysis, no common gene is found. Differentially expressed genes are further used to perform pathway analysis. Pathway enrichment analysis give us significant metabolic and signaling pathways related with the disease. Focal adhesion is recognized as regular pathway that incorporate number of significant entities connected towards the significant procedures, for example, cell mobility and proliferation. A pathway is structured which is made out of some significant entities that are differentially expressed in our study. This pathway also includes Wnt signaling pathway, actin polymerization, P13K-AKT signaling pathway and MAPK signaling pathway. Quantitative modeling approach is used to investigate the impact of DEGs in general pathway. This model is generated by taking FPKM average expression values of DEGs in normal and disease samples. Simulation model expounds the dynamic conduct of model and tell the significance of cell mobility which increases during disease condition. Sensitivity analysis of pathway shows the general effect of entities on it which was performed and shows higher sensitivity of ACTG1, MLC and ROCK. In addition, ACTG1 and MLC are up-regulated in our analysis but ROCK is down regulated so, these genes can be used as noteworthy potential biomarker against PCOS and change in expression of these can lead towards endometrial carcinoma.

In future, we can check the expression level of differentially expressed genes by using the different techniques of wet lab for more validation. Further study on remaining pathways by using different approaches of systems biology can provide us more important targets associated with PCOS and how it is linked with Endometrial carcinoma. We can also identify and design inhibitors against these targets as well.

References

1. Agarwala, R. *et al.* (2016) 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Research*. doi: 10.1093/nar/gkv1290.
2. Athar, A. *et al.* (2019) 'ArrayExpress update - From bulk to single-cell expression data', *Nucleic Acids Research*. doi: 10.1093/nar/gky964.
3. Barber, T. M. *et al.* (2019) 'Obesity and Polycystic Ovary Syndrome: Implications for Pathogenesis and Novel Management Strategies', *Clinical Medicine Insights: Reproductive Health,* 13, p. 117955811987404. doi: 10.1177/1179558119874042.
4. Barry, J. A., Azizia, M. M. and Hardiman, P. J. (2014) 'Risk of endometrial, ovarian and breast cancer in women with polycystic ovary syndrome: A systematic review and meta-analysis', *Human Reproduction Update*. doi: 10.1093/humupd/dmu012.
5. Blighe, K. (2019) 'Publication-ready volcano plots with enhanced colouring and labeling', *R-Package*.
6. Burke, W. M. *et al.* (2014) 'Endometrial cancer: A review and current management strategies: Part i', *Gynecologic Oncology*. doi: 10.1016/j.ygyno.2014.05.018.
7. Clough, E. and Barrett, T. (2016) 'The Gene Expression Omnibus database', in *Methods in Molecular Biology*. doi: 10.1007/978-1-4939-3578-9_5.
8. Crandall, C. J. *et al.* (2018) 'Breast cancer, endometrial cancer, and cardiovascular events in participants who used vaginal estrogen in the Women's Health Initiative Observational Study', *Menopause*. doi: 10.1097/GME.0000000000000956.
9. Deu-Pons, J., Schroeder, M. P. and Lopez-Bigas, N. (2014) 'JHeatmap: An interactive heatmap viewer for the web', *Bioinformatics*. doi: 10.1093/bioinformatics/btu094.
10. Diamanti-Kandarakis, E., Kandarakis, H. and Legro, R. S. (2006) 'The role of genes and environment in the etiology of PCOS', *Endocrine*. doi: 10.1385/ENDO:30:1:19.

11. Dong, P. *et al.* (2013) ‘Emerging therapeutic biomarkers in endometrial cancer’, *BioMed Research International*, 2013(Figure 1). doi: 10.1155/2013/130362.
12. Dumesic, D. A. *et al.* (2015) ‘Scientific statement on the diagnostic criteria, epidemiology, pathophysiology, and molecular genetics of polycystic ovary syndrome’, *Endocrine Reviews*. doi: 10.1210/er.2015-1018.
13. Ebbert, M. T. W. *et al.* (2016) ‘Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches’, *BMC Bioinformatics*. doi: 10.1186/s12859-016-1097-3.
14. Escobar-Morreale, H. F. (2018) ‘Polycystic ovary syndrome: Definition, aetiology, diagnosis and treatment’, *Nature Reviews Endocrinology*. Nature Publishing Group, pp. 270–284. doi: 10.1038/nrendo.2018.24.
15. Frazee, A. *et al.* (2014) ‘Flexible analysis of transcriptome assemblies with Ballgown’, *bioRxiv*. doi: 10.1101/003665.
16. Gandolfo, L. C. and Speed, T. P. (2018) ‘RLE plots: Visualizing unwanted variation in high dimensional data’, *PLoS ONE*. Public Library of Science, 13(2). doi: 10.1371/journal.pone.0191629.
17. García-Campos, M. A., Espinal-Enríquez, J. and Hernández-Lemus, E. (2015) ‘Pathway analysis: State of the art’, *Frontiers in Physiology*. Frontiers Research Foundation. doi: 10.3389/fphys.2015.00383.
18. George, K. R. O. Y. and Malini, N. A. (2014) ‘the Prevalence and Etiology of Polycystic Ovarian Syndrome (Pcos) As a Cause of Female Infertility in Central Travancore’, 9(1), pp. 1–6.
19. Ghoubara, A., Sundar, S. and Ewies, A. A. A. (2019) ‘Black women with postmenopausal bleeding have lower prevalence of endometrial cancer than other ethnic groups’, *Climacteric*. Taylor and Francis Ltd, 22(6), pp. 632–636. doi:

10.1080/13697137.2019.1606794.

20. Giallauria, F. *et al.* (2009) 'Androgens in polycystic ovary syndrome: The role of exercise and diet', *Seminars in Reproductive Medicine*. doi: 10.1055/s-0029-1225258.
21. Goodarzi, M. O. *et al.* (2011) 'Polycystic ovary syndrome: Etiology, pathogenesis and diagnosis', *Nature Reviews Endocrinology*. doi: 10.1038/nrendo.2010.217.
22. Grada, A. and Weinbrecht, K. (2013) 'Next-generation sequencing: Methodology and application', *Journal of Investigative Dermatology*. doi: 10.1038/jid.2013.248.
23. Holm, N. S. L. *et al.* (2012) 'The prevalence of endometrial hyperplasia and endometrial cancer in women with polycystic ovary syndrome or hyperandrogenism', *Acta Obstetricia et Gynecologica Scandinavica*. doi: 10.1111/j.1600-0412.2012.01458.x.
24. Holst, F. *et al.* (2019) 'PIK3CA amplification associates with aggressive phenotype but not markers of AKT-mTOR signaling in endometrial carcinoma', *Clinical Cancer Research*, 25(1), pp. 334–345. doi: 10.1158/1078-0432.CCR-18-0452.
25. Hrdlickova, R., Toloue, M. and Tian, B. (2017) 'RNA-Seq methods for transcriptome analysis', *Wiley Interdisciplinary Reviews: RNA*. doi: 10.1002/wrna.1364.
26. Huang, D. W. *et al.* (2007) 'The DAVID Gene Functional Classification Tool: A novel biological module-centric algorithm to functionally analyze large gene lists', *Genome Biology*, 8(9). doi: 10.1186/gb-2007-8-9-r183.
27. Jayasena, C. N. and Franks, S. (2014) 'The management of patients with polycystic ovary syndrome', *Nature Reviews Endocrinology*. doi: 10.1038/nrendo.2014.102.
28. Kauffmann, A. and Huber, W. (2010) 'Microarray data quality control improves the detection of differentially expressed genes', *Genomics*. doi: 10.1016/j.ygeno.2010.01.003.
29. Kaur, S. *et al.* (2012) 'Differential Gene Expression in Granulosa Cells from Polycystic

Ovary Syndrome Patients with and without Insulin Resistance : Identification of Susceptibility Gene Sets through Network Analysis', 97(October), pp. 2016–2021. doi: 10.1210/jc.2011-3441.

30. Kim, C. S., Hwang, S. and Zhang, S. D. (2014) 'RMA with quantile normalization mixes biological signals between different sample groups in microarray data analysis', in *Proceedings - 2014 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2014*. Institute of Electrical and Electronics Engineers Inc., pp. 139–143. doi: 10.1109/BIBM.2014.6999142.
31. Kim, D., Langmead, B. and Salzberg, S. L. (2015) 'Hisat2', *Nature methods*. doi: 10.1038/nmeth.3317.
32. Krallinger, M. and Tendulkar, A. V (2006) 'Text Mining Tools in Biology', *Cancer Research*.
33. Krueger, F. (2016) *Trim Galore, Babraham Bioinformatics*. doi: http://www.bioinformatics.babraham.ac.uk/projects/trim galore/.
34. Kukurba, K. R. and Montgomery, S. B. (2015) 'RNA sequencing and analysis', *Cold Spring Harbor Protocols*. doi: 10.1101/pdb.top084970.
35. Kumar, R. *et al.* (2012) 'A high-throughput method for Illumina RNA-Seq library preparation', *Frontiers in Plant Science*. doi: 10.3389/fpls.2012.00202.
36. Lee, T. T. and Rausch, M. E. (2012) 'Polycystic ovarian syndrome: Role of imaging in diagnosis', *Radiographics,* 32(6), pp. 1643–1657. doi: 10.1148/rg.326125503.
37. Lee, T. T. and Rausch, M. E. (no date) 'THE REPRODUCTIVE YEARS'. doi: 10.1148/rg.326125503.
38. Levy, S. E. and Myers, R. M. (2016) 'Advancements in Next-Generation Sequencing', *Annual Review of Genomics and Human Genetics*. doi: 10.1146/annurev-genom-083115-022413.

39. Li, N. *et al.* (2014) 'Identification of chimeric TSNAX-DISC1 resulting from intergenic splicing in endometrial carcinoma through high-throughput RNA sequencing', *Carcinogenesis,* 35(12), pp. 2687–2697. doi: 10.1093/carcin/bgu201.

40. Li, W. (2012) 'Volcano plots in analyzing differential expressions with mRNA microarrays', *Journal of Bioinformatics and Computational Biology*. doi: 10.1142/S0219720012310038.

41. Liu, Q. *et al.* (2016) 'Single-cell analysis of differences in transcriptomic profiles of oocytes and cumulus cells at GV, MI, MII stages from PCOS patients', *Scientific Reports*. doi: 10.1038/srep39638.

42. Loughborough University (2009) 'Statistical Analysis 3 : Paired t-test', *Discovering statistics*.

43. Mäenpää, J. (2020) 'Epidemiology, Risk Factors, and Prevention for Endometrial Cancer', in *Management of Endometrial Cancer*. Springer International Publishing, pp. 61–67. doi: 10.1007/978-3-319-64513-1_5.

44. Makker, A. *et al.* (2012) 'PI3K-Akt-mTOR and MAPK signaling pathways in polycystic ovarian syndrome, uterine leiomyomas and endometriosis: An update', *Gynecological Endocrinology*. doi: 10.3109/09513590.2011.583955.

45. Mariano, D. C. B. *et al.* (2017) 'A guide to performing systematic literature reviews in bioinformatics'. Available at: http://arxiv.org/abs/1707.05813 (Accessed: 17 January 2020).

46. Matsumoto, T. *et al.* (2015) 'Distinct β-catenin and PIK3CA mutation profiles in endometriosis-associated ovarian endometrioid and clear cell carcinomas', *American Journal of Clinical Pathology*. doi: 10.1309/AJCPZ5T2POOFMQVN.

47. Morgensztern, D. and McLeod, H. L. (2005) 'PI3K/Akt/mTOR pathway as a target for cancer therapy', *Anti-Cancer Drugs*. doi: 10.1097/01.cad.0000173476.67239.3b.

48. Mravec, B. and Tibensky, M. (2020) 'Increased cancer risk in polycystic ovary syndrome: An (un)sympathetic connection?', *Medical Hypotheses*. doi: 10.1016/j.mehy.2019.109437.

49. Nagalakshmi, U., Waern, K. and Snyder, M. (2010) 'RNA-seq: A method for comprehensive transcriptome analysis', *Current Protocols in Molecular Biology*. doi: 10.1002/0471142727.mb0411s89.

50. Ndefo, U. A., Eaton, A. and Green, M. R. (2013) 'Polycystic Ovary Syndrome A Review of Treatment Options With a Focus on Pharmacological Approaches', 38(6).

51. O'Mara, T. A., Zhao, M. and Spurdle, A. B. (2016) 'Meta-analysis of gene expression studies in endometrial cancer identifies gene expression profiles associated with aggressive disease and patient outcome', *Scientific Reports*. Nature Publishing Group, 6(October), pp. 1–9. doi: 10.1038/srep36677.

52. Olson, M. F. and Sahai, E. (2009) 'The actin cytoskeleton in cancer cell motility', *Clinical and Experimental Metastasis*, pp. 273–287. doi: 10.1007/s10585-008-9174-2.

53. Pareek, C. S., Smoczynski, R. and Tretyn, A. (2011) 'Sequencing technologies and genome sequencing', *Journal of Applied Genetics*. doi: 10.1007/s13353-011-0057-x.

54. Parsons, L. H. P. *et al.* (2018) 'The prevalence of occult endometrial cancer in women undergoing hysterectomy for benign indications', *European Journal of Obstetrics and Gynecology and Reproductive Biology*. doi: 10.1016/j.ejogrb.2018.02.017.

55. Pertea, M. *et al.* (2016) 'Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown', *Nature Protocols*. doi: 10.1038/nprot.2016.095.

56. Physician, A. F. (2016) 'Diagnosis and Treatment of Polycystic Ovary Syndrome'.

57. Pillay, O. C. *et al.* (2006) 'The association between polycystic ovaries and endometrial cancer', *Human Reproduction*, 21(4), pp. 924–929. doi: 10.1093/humrep/dei420.

58. Qian, X. *et al.* (2014) 'RNA-seq technology and its application in fish transcriptomics',

OMICS A Journal of Integrative Biology. doi: 10.1089/omi.2013.0110.

59. Qiu, H. *et al.* (2016) 'JQ1 suppresses tumor growth via PTEN/PI3K/AKT pathway in endometrial cancer', *Oncotarget*, 7(41), pp. 66809–66821. doi: 10.18632/oncotarget.11631.
60. Redon, R., Fitzgerald, T. and Carter, N. P. (2009) 'Comparative genomic hybridization: DNA labeling, hybridization and detection.', *Methods in molecular biology (Clifton, N.J.)*. doi: 10.1007/978-1-59745-538-1_17.
61. Risinger, J. I. *et al.* (2003) 'Microarray analysis reveals distinct gene expression profiles among different histologic types of endometrial cancer', *Cancer Research*, 63(1), pp. 6–11.
62. Ritchie, M. E. *et al.* (2015) 'Limma powers differential expression analyses for RNA-sequencing and microarray studies', *Nucleic Acids Research*. doi: 10.1093/nar/gkv007.
63. Roldán, B., San Millán, J. L. and Escobar-Morreale, H. F. (2004) 'Genetic Basis of Metabolic Abnormalities in Polycystic Ovary Syndrome: Implications for Therapy', *American Journal of PharmacoGenomics*. doi: 10.2165/00129785-200404020-00004.
64. Rosenfield, R. L. and Ehrmann, D. A. (2016) 'The Pathogenesis of Polycystic Ovary Syndrome (PCOS): The hypothesis of PCOS as functional ovarian hyperandrogenism revisited', *Endocrine Reviews*, 37(5), pp. 467–520. doi: 10.1210/er.2015-1104.
65. Shafiee, M. N. *et al.* (2016) 'Up-regulation of genes involved in the insulin signalling pathway (IGF1, PTEN and IGFBP1) in the endometrium may link polycystic ovarian syndrome and endometrial cancer', *Molecular and Cellular Endocrinology*, 424, pp. 94–101. doi: 10.1016/j.mce.2016.01.019.
66. Sharov, A. A., Dudekula, D. B. and Ko, M. S. H. (2005) 'A web-based tool for principal component and significance analysis of microarray data', *Bioinformatics*. doi: 10.1093/bioinformatics/bti343.

67. Sidra, S. *et al.* (2019) 'Evaluation of clinical manifestations, health risks, and quality of life among women with polycystic ovary syndrome', *PLoS ONE*. doi: 10.1371/journal.pone.0223329.
68. Singh, P. and Rai, S. N. (2019) 'Factors affecting obesity and its treatment', *Obesity Medicine*, 16(June). doi: 10.1016/j.obmed.2019.100140.
69. Smith, T. F. (2008) 'Computational biology: Its challenges past, present, and future', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-540-78839-3_1.
70. Spitzer, M. *et al.* (2014) 'BoxPlotR: A web tool for generation of box plots', *Nature Methods*, pp. 121–122. doi: 10.1038/nmeth.2811.
71. Suhaimi, S. S., Ab Mutalib, N. S. and Jamal, R. (2016) 'Understanding molecular landscape of endometrial cancer through next generation sequencing: What we have learned so far?', *Frontiers in Pharmacology*, 7(NOV), pp. 1–7. doi: 10.3389/fphar.2016.00409.
72. Tomao, F. *et al.* (2016) 'Special issues in fertility preservation for gynecologic malignancies', *Critical Reviews in Oncology/Hematology*. doi: 10.1016/j.critrevonc.2015.08.024.
73. Villaseñor-Park, J. and Ortega-Loayza, A. G. (2013) 'Microarray technique, analysis, and applications in dermatology', *Journal of Investigative Dermatology*. doi: 10.1038/jid.2013.64.
74. Wang, L., Wang, S. and Li, W. (2012) 'RSeQC: Quality control of RNA-seq experiments', *Bioinformatics*. doi: 10.1093/bioinformatics/bts356.
75. Wolf, W. M. *et al.* (2018) 'Geographical Prevalence of Polycystic Ovary Syndrome as Determined by Region and Race / Ethnicity', pp. 1–13. doi: 10.3390/ijerph15112589.
76. Zhang, Y. *et al.* (2018) 'SequencEnG: an Interactive Knowledge Base of Sequencing

Techniques', *Sequenceng: An interactive knowledge base of sequencing techniques*. doi: 10.1101/319079.

77. Zhang, Y. *et al.* (2019) 'Sequenceng: An interactive knowledge base of sequencing techniques', *Bioinformatics*. doi: 10.1093/bioinformatics/bty794.
78. Agarwala, R. *et al.* (2016) 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Research*. doi: 10.1093/nar/gkv1290.
79. Athar, A. *et al.* (2019) 'ArrayExpress update - From bulk to single-cell expression data', *Nucleic Acids Research*. doi: 10.1093/nar/gky964.
80. Barber, T. M. *et al.* (2019) 'Obesity and Polycystic Ovary Syndrome: Implications for Pathogenesis and Novel Management Strategies', *Clinical Medicine Insights: Reproductive Health*, 13, p. 117955811987404. doi: 10.1177/1179558119874042.
81. Barry, J. A., Azizia, M. M. and Hardiman, P. J. (2014) 'Risk of endometrial, ovarian and breast cancer in women with polycystic ovary syndrome: A systematic review and meta-analysis', *Human Reproduction Update*. doi: 10.1093/humupd/dmu012.
82. Blighe, K. (2019) 'Publication-ready volcano plots with enhanced colouring and labeling', *R-Package*.
83. Burke, W. M. *et al.* (2014) 'Endometrial cancer: A review and current management strategies: Part i', *Gynecologic Oncology*. doi: 10.1016/j.ygyno.2014.05.018.
84. Clough, E. and Barrett, T. (2016) 'The Gene Expression Omnibus database', in *Methods in Molecular Biology*. doi: 10.1007/978-1-4939-3578-9_5.
85. Crandall, C. J. *et al.* (2018) 'Breast cancer, endometrial cancer, and cardiovascular events in participants who used vaginal estrogen in the Women's Health Initiative Observational Study', *Menopause*. doi: 10.1097/GME.0000000000000956.
86. Deu-Pons, J., Schroeder, M. P. and Lopez-Bigas, N. (2014) 'JHeatmap: An interactive heatmap viewer for the web', *Bioinformatics*. doi: 10.1093/bioinformatics/btu094.

87. Diamanti-Kandarakis, E., Kandarakis, H. and Legro, R. S. (2006) ‘The role of genes and environment in the etiology of PCOS’, *Endocrine*. doi: 10.1385/ENDO:30:1:19.

88. Dong, P. *et al.* (2013) ‘Emerging therapeutic biomarkers in endometrial cancer’, *BioMed Research International*, 2013(Figure 1). doi: 10.1155/2013/130362.

89. Dumesic, D. A. *et al.* (2015) ‘Scientific statement on the diagnostic criteria, epidemiology, pathophysiology, and molecular genetics of polycystic ovary syndrome’, *Endocrine Reviews*. doi: 10.1210/er.2015-1018.

90. Ebbert, M. T. W. *et al.* (2016) ‘Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches’, *BMC Bioinformatics*. doi: 10.1186/s12859-016-1097-3.

91. Escobar-Morreale, H. F. (2018) ‘Polycystic ovary syndrome: Definition, aetiology, diagnosis and treatment’, *Nature Reviews Endocrinology*. Nature Publishing Group, pp. 270–284. doi: 10.1038/nrendo.2018.24.

92. Frazee, A. *et al.* (2014) ‘Flexible analysis of transcriptome assemblies with Ballgown’, *bioRxiv*. doi: 10.1101/003665.

93. Gandolfo, L. C. and Speed, T. P. (2018) ‘RLE plots: Visualizing unwanted variation in high dimensional data’, *PLoS ONE*. Public Library of Science, 13(2). doi: 10.1371/journal.pone.0191629.

94. García-Campos, M. A., Espinal-Enríquez, J. and Hernández-Lemus, E. (2015) ‘Pathway analysis: State of the art’, *Frontiers in Physiology*. Frontiers Research Foundation. doi: 10.3389/fphys.2015.00383.

95. George, K. R. O. Y. and Malini, N. A. (2014) ‘the Prevalence and Etiology of Polycystic Ovarian Syndrome (Pcos) As a Cause of Female Infertility in Central Travancore’, 9(1), pp. 1–6.

96. Ghoubara, A., Sundar, S. and Ewies, A. A. A. (2019) ‘Black women with

postmenopausal bleeding have lower prevalence of endometrial cancer than other ethnic groups', *Climacteric*. Taylor and Francis Ltd, 22(6), pp. 632–636. doi: 10.1080/13697137.2019.1606794.

97. Giallauria, F. *et al.* (2009) 'Androgens in polycystic ovary syndrome: The role of exercise and diet', *Seminars in Reproductive Medicine*. doi: 10.1055/s-0029-1225258.
98. Goodarzi, M. O. *et al.* (2011) 'Polycystic ovary syndrome: Etiology, pathogenesis and diagnosis', *Nature Reviews Endocrinology*. doi: 10.1038/nrendo.2010.217.
99. Grada, A. and Weinbrecht, K. (2013) 'Next-generation sequencing: Methodology and application', *Journal of Investigative Dermatology*. doi: 10.1038/jid.2013.248
100. Holst, F. *et al.* (2019) 'PIK3CA amplification associates with aggressive phenotype but not markers of AKT-mTOR signaling in endometrial carcinoma', *Clinical Cancer Research*, 25(1), pp. 334–345. doi: 10.1158/1078-0432.CCR-18-0452.
101. Hrdlickova, R., Toloue, M. and Tian, B. (2017) 'RNA-Seq methods for transcriptome analysis', *Wiley Interdisciplinary Reviews: RNA*. doi: 10.1002/wrna.1364.
102. Huang, D. W. *et al.* (2007) 'The DAVID Gene Functional Classification Tool: A novel biological module-centric algorithm to functionally analyze large gene lists', *Genome Biology*, 8(9). doi: 10.1186/gb-2007-8-9-r183.
103. Jayasena, C. N. and Franks, S. (2014) 'The management of patients with polycystic ovary syndrome', *Nature Reviews Endocrinology*. doi: 10.1038/nrendo.2014.102.
104. Kauffmann, A. and Huber, W. (2010) 'Microarray data quality control improves the detection of differentially expressed genes', *Genomics*. doi: 10.1016/j.ygeno.2010.01.003.
105. Kaur, S. *et al.* (2012) 'Differential Gene Expression in Granulosa Cells from Polycystic Ovary Syndrome Patients with and without Insulin Resistance : Identification of Susceptibility Gene Sets through Network Analysis', 97(October), pp.

2016–2021. doi: 10.1210/jc.2011-3441.

106. Kim, C. S., Hwang, S. and Zhang, S. D. (2014) 'RMA with quantile normalization mixes biological signals between different sample groups in microarray data analysis', in *Proceedings - 2014 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2014*. Institute of Electrical and Electronics Engineers Inc., pp. 139–143. doi: 10.1109/BIBM.2014.6999142.

107. Kim, D., Langmead, B. and Salzberg, S. L. (2015) 'Hisat2', *Nature methods*. doi: 10.1038/nmeth.3317.

108. Krallinger, M. and Tendulkar, A. V (2006) 'Text Mining Tools in Biology', *Cancer Research*.

109. Krueger, F. (2016) *Trim Galore, Babraham Bioinformatics*. doi: http://www.bioinformatics.babraham.ac.uk/projects/trim galore/.

110. Kukurba, K. R. and Montgomery, S. B. (2015) 'RNA sequencing and analysis', *Cold Spring Harbor Protocols*. doi: 10.1101/pdb.top084970.

111. Kumar, R. *et al.* (2012) 'A high-throughput method for Illumina RNA-Seq library preparation', *Frontiers in Plant Science*. doi: 10.3389/fpls.2012.00202.

112. Lee, T. T. and Rausch, M. E. (2012) 'Polycystic ovarian syndrome: Role of imaging in diagnosis', *Radiographics*, 32(6), pp. 1643–1657. doi: 10.1148/rg.326125503.

113. Lee, T. T. and Rausch, M. E. (no date) 'THE REPRODUCTIVE YEARS'. doi: 10.1148/rg.326125503.

114. Levy, S. E. and Myers, R. M. (2016) 'Advancements in Next-Generation Sequencing', *Annual Review of Genomics and Human Genetics*. doi: 10.1146/annurev-genom-083115-022413.

115. Li, N. *et al.* (2014) 'Identification of chimeric TSNAX-DISC1 resulting from intergenic splicing in endometrial carcinoma through high-throughput RNA

sequencing', *Carcinogenesis*, 35(12), pp. 2687–2697. doi: 10.1093/carcin/bgu201.

116. Li, W. (2012) 'Volcano plots in analyzing differential expressions with mRNA microarrays', *Journal of Bioinformatics and Computational Biology*. doi: 10.1142/S0219720012310038.

117. Liu, Q. *et al.* (2016) 'Single-cell analysis of differences in transcriptomic profiles of oocytes and cumulus cells at GV, MI, MII stages from PCOS patients', *Scientific Reports*. doi: 10.1038/srep39638.

118. Loughborough University (2009) 'Statistical Analysis 3 : Paired t-test', *Discovering statistics*.

119. Mäenpää, J. (2020) 'Epidemiology, Risk Factors, and Prevention for Endometrial Cancer', in *Management of Endometrial Cancer*. Springer International Publishing, pp. 61–67. doi: 10.1007/978-3-319-64513-1_5.

120. Makker, A. *et al.* (2012) 'PI3K-Akt-mTOR and MAPK signaling pathways in polycystic ovarian syndrome, uterine leiomyomas and endometriosis: An update', *Gynecological Endocrinology*. doi: 10.3109/09513590.2011.583955.

121. Mariano, D. C. B. *et al.* (2017) 'A guide to performing systematic literature reviews in bioinformatics'. Available at: http://arxiv.org/abs/1707.05813 (Accessed: 17 January 2020).

122. Matsumoto, T. *et al.* (2015) 'Distinct β-catenin and PIK3CA mutation profiles in endometriosis-associated ovarian endometrioid and clear cell carcinomas', *American Journal of Clinical Pathology*. doi: 10.1309/AJCPZ5T2POOFMQVN.

123. Morgensztern, D. and McLeod, H. L. (2005) 'PI3K/Akt/mTOR pathway as a target for cancer therapy', *Anti-Cancer Drugs*. doi: 10.1097/01.cad.0000173476.67239.3b.

124. Mravec, B. and Tibensky, M. (2020) 'Increased cancer risk in polycystic ovary syndrome: An (un)sympathetic connection?', *Medical Hypotheses*. doi:

10.1016/j.mehy.2019.109437.

125. Nagalakshmi, U., Waern, K. and Snyder, M. (2010) ‘RNA-seq: A method for comprehensive transcriptome analysis’, *Current Protocols in Molecular Biology*. doi: 10.1002/0471142727.mb0411s89.

126. Ndefo, U. A., Eaton, A. and Green, M. R. (2013) ‘Polycystic Ovary Syndrome A Review of Treatment Options With a Focus on Pharmacological Approaches’, 38(6).

127. O’Mara, T. A., Zhao, M. and Spurdle, A. B. (2016) ‘Meta-analysis of gene expression studies in endometrial cancer identifies gene expression profiles associated with aggressive disease and patient outcome’, *Scientific Reports*. Nature Publishing Group, 6(October), pp. 1–9. doi: 10.1038/srep36677.

128. Olson, M. F. and Sahai, E. (2009) ‘The actin cytoskeleton in cancer cell motility’, *Clinical and Experimental Metastasis*, pp. 273–287. doi: 10.1007/s10585-008-9174-2.

129. Pareek, C. S., Smoczynski, R. and Tretyn, A. (2011) ‘Sequencing technologies and genome sequencing’, *Journal of Applied Genetics*. doi: 10.1007/s13353-011-0057-x.

130. Parsons, L. H. P. *et al.* (2018) ‘The prevalence of occult endometrial cancer in women undergoing hysterectomy for benign indications’, *European Journal of Obstetrics and Gynecology and Reproductive Biology*. doi: 10.1016/j.ejogrb.2018.02.017.

131. Pertea, M. *et al.* (2016) ‘Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown’, *Nature Protocols*. doi: 10.1038/nprot.2016.095.

132. Physician, A. F. (2016) ‘Diagnosis and Treatment of Polycystic Ovary Syndrome’.

133. Pillay, O. C. *et al.* (2006) ‘The association between polycystic ovaries and endometrial cancer’, *Human Reproduction*, 21(4), pp. 924–929. doi: 10.1093/humrep/dei420.

134. Qian, X. *et al.* (2014) ‘RNA-seq technology and its application in fish transcriptomics’, *OMICS A Journal of Integrative Biology*. doi: 10.1089/omi.2013.0110.

135. Qiu, H. *et al.* (2016) 'JQ1 suppresses tumor growth via PTEN/PI3K/AKT pathway in endometrial cancer', *Oncotarget*, 7(41), pp. 66809–66821. doi: 10.18632/oncotarget.11631.

136. Redon, R., Fitzgerald, T. and Carter, N. P. (2009) 'Comparative genomic hybridization: DNA labeling, hybridization and detection.', *Methods in molecular biology (Clifton, N.J.)*. doi: 10.1007/978-1-59745-538-1_17.

137. Risinger, J. I. *et al.* (2003) 'Microarray analysis reveals distinct gene expression profiles among different histologic types of endometrial cancer', *Cancer Research*, 63(1), pp. 6–11.

138. Ritchie, M. E. *et al.* (2015) 'Limma powers differential expression analyses for RNA-sequencing and microarray studies', *Nucleic Acids Research*. doi: 10.1093/nar/gkv007.

139. Roldán, B., San Millán, J. L. and Escobar-Morreale, H. F. (2004) 'Genetic Basis of Metabolic Abnormalities in Polycystic Ovary Syndrome: Implications for Therapy', *American Journal of PharmacoGenomics*. doi: 10.2165/00129785-200404020-00004.

140. Rosenfield, R. L. and Ehrmann, D. A. (2016) 'The Pathogenesis of Polycystic Ovary Syndrome (PCOS): The hypothesis of PCOS as functional ovarian hyperandrogenism revisited', *Endocrine Reviews*, 37(5), pp. 467–520. doi: 10.1210/er.2015-1104.

141. Shafiee, M. N. *et al.* (2016) 'Up-regulation of genes involved in the insulin signalling pathway (IGF1, PTEN and IGFBP1) in the endometrium may link polycystic ovarian syndrome and endometrial cancer', *Molecular and Cellular Endocrinology*, 424, pp. 94–101. doi: 10.1016/j.mce.2016.01.019.

142. Sharov, A. A., Dudekula, D. B. and Ko, M. S. H. (2005) 'A web-based tool for principal component and significance analysis of microarray data', *Bioinformatics*. doi: 10.1093/bioinformatics/bti343.

143. Sidra, S. *et al.* (2019) 'Evaluation of clinical manifestations, health risks, and quality

of life among women with polycystic ovary syndrome', *PLoS ONE*. doi: 10.1371/journal.pone.0223329.

144. Singh, P. and Rai, S. N. (2019) 'Factors affecting obesity and its treatment', *Obesity Medicine*, 16(June). doi: 10.1016/j.obmed.2019.100140.
145. Smith, T. F. (2008) 'Computational biology: Its challenges past, present, and future', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-540-78839-3_1.
146. Spitzer, M. *et al.* (2014) 'BoxPlotR: A web tool for generation of box plots', *Nature Methods*, pp. 121–122. doi: 10.1038/nmeth.2811.
147. Suhaimi, S. S., Ab Mutalib, N. S. and Jamal, R. (2016) 'Understanding molecular landscape of endometrial cancer through next generation sequencing: What we have learned so far?', *Frontiers in Pharmacology*, 7(NOV), pp. 1–7. doi: 10.3389/fphar.2016.00409.
148. Tomao, F. *et al.* (2016) 'Special issues in fertility preservation for gynecologic malignancies', *Critical Reviews in Oncology/Hematology*. doi: 10.1016/j.critrevonc.2015.08.024.
149. Villaseñor-Park, J. and Ortega-Loayza, A. G. (2013) 'Microarray technique, analysis, and applications in dermatology', *Journal of Investigative Dermatology*. doi: 10.1038/jid.2013.64.
150. Wang, L., Wang, S. and Li, W. (2012) 'RSeQC: Quality control of RNA-seq experiments', *Bioinformatics*. doi: 10.1093/bioinformatics/bts356.
151. Wolf, W. M. *et al.* (2018) 'Geographical Prevalence of Polycystic Ovary Syndrome as Determined by Region and Race / Ethnicity', pp. 1–13. doi: 10.3390/ijerph15112589.
152. Zhang, Y. *et al.* (2018) 'SequencEnG: an Interactive Knowledge Base of Sequencing

Techniques', *Sequenceng: An interactive knowledge base of sequencing techniques*. doi: 10.1101/319079.

153. Zhang, Y. *et al.* (2019) 'Sequenceng: An interactive knowledge base of sequencing techniques', *Bioinformatics*. doi: 10.1093/bioinformatics/bty794.

Appendix A

Source code for Microarray (Differential expression analysis)

```
if (!require("BiocManager"))

install.packages("BiocManager")

BiocManager::install("maEndToEnd", version = "3.8")

install.packages("devtools")

library(devtools)

devtools::install_github("r-lib/remotes")

library(remotes)

packageVersion("remotes") # has to be 1.1.1.9000 or later

# To install packages in R 3.5.2

BiocManager::install(c( "checkmate"))

library(BiocGenerics)

remotes::install_github("b-klaus/maEndToEnd", ref="master")

#General Bioconductor packages

install.packages("affyio")

library(Biobase)

library(oligoClasses)

#Annotation and data import packages

library(ArrayExpress)

library(pd.hugene.1.0.st.v1)

library(hugene10sttranscriptcluster.db)

#Quality control and pre-processing packages

library(oligo)

library(arrayQualityMetrics)

#Analysis and statistics packages
```

```
library(limma)
library(topGO)
library(ReactomePA)
library(clusterProfiler)
#Plotting and color options packages
library(gplots)
library(ggplot2)
library(geneplotter)
library(RColorBrewer)
library(pheatmap)
#Formatting/documentation packages
#library(rmarkdown)
#library(BiocStyle)
library(dplyr)
library(tidyr)
#Helpers:
library(stringr)
library(matrixStats)
library(genefilter)
library(openxlsx)
#Downloading the raw data from ArrayExpress
raw_data_dir <- "C:/Micro Array 1 GSE 5850/rawDataMAWorkdown"
if (!dir.exists(raw_data_dir)) {
  dir.create(raw_data_dir)
}
#setting working directory
Tutorial_MicArrays_3 <- "C:/MicArrays_1"
if(!dir.exists(Tutorial_MicArrays_3)){
  dir.create(Tutorial_MicArrays_3)
}
```

```
setwd(Tutorial_MicArrays_3)
raw_data_dir <- file.path(getwd(), "rawDataMAWorkdown")
if(!dir.exists(raw_data_dir)){
  dir.create(raw_data_dir)
}
anno_AE <- getAE("E-GEOD-5850", path=raw_data_dir, type="raw")
setwd("C:/Micro Array 1 GSE 5850")
SDRF <- read.delim("E-GEOD-5850.sdrf.txt")
rownames(SDRF) <- SDRF$Array.Data.File
SDRF <- AnnotatedDataFrame(SDRF)
SDRF
pData(SDRF)
sdrf_location <- file. path(raw_data_dir,"E-GEOD-5850.sdrf.txt")
#by changing directory
SDRF <- read.csv(sdrf_location)
write.csv(SDRF,"Sdrf.csv")
SDRF
raw_data <- oligo::read.celfiles(filenames = file.path(raw_data_dir,
SDRF$Array.Data.File),
verbose = FALSE, phenoData = SDRF)
write.csv(raw_data,"raw_data.csv")
stopifnot(validObject(raw_data))
# Columns has been changed from sdrf
head(Biobase::pData(raw_data))
pData(raw_data)
#TROUBLESHOOT:
#- There must be two columns needed to be filtered for pDATA
Biobase::pData(raw_data) <- Biobase::pData(raw_data)[,c("Scan.Name","Array.Data.File")]
head(Biobase::pData(raw_data))
# Quality control
```

```
Biobase::exprs(raw_data)[1:12, 1:12]
exp_raw <- log2(Biobase::exprs(raw_data))
PCA_raw <- prcomp(t(exp_raw), scale. = FALSE)
percentVar <- round(100*PCA_raw$sdev^2/sum(PCA_raw$sdev^2),1)
sd_ratio <- sqrt(percentVar[2] / percentVar[1])
# scan name and array data file has been added
dataGG <- data.frame(PC1 = PCA_raw$x[,1], PC2 = PCA_raw$x[,2],
Disease = pData(raw_data)$Scan.Name,
Individual = pData(raw_data)$Array.Data.File)
# to save plots in pdf format and define manual values and color
pdf("PCAplot.pdf " ,width=10,height=10,paper='special')
ggplot(dataGG, aes(PC1, PC2)) +
  geom_point(aes(shape = Individual, colour = Disease)) +
  ggtitle("PCA plot of the log-transformed raw expression data") +
  xlab(paste0("PC1, VarExp: ", percentVar[1], "%")) +
  ylab(paste0("PC2, VarExp: ", percentVar[2], "%")) +
  theme(plot.title = element_text(hjust = 0.5))+
  coord_fixed(ratio = sd_ratio) +
  scale_shape_manual(values = c(4,15)) +
  scale_color_manual(values = c("darkorange", "darkblue"))
dev.off()
# box plot
pdf("boxplot.pdf " ,width=10,height=10,paper='special')
oligo::boxplot(raw_data, target = "core",
               main = "Boxplot of log2-intensitites for the raw data")
dev.off()
arrayQualityMetrics(expressionset = raw_data,
outdir =raw_data_dir,
force = TRUE, do.logtransform = TRUE,
intgroup = c("Scan.Name"))
```

```
#Background adjustment, calibration, summarization and annotation
head(ls("package:hugene10sttranscriptcluster.db"))
palmieri_eset <- oligo::rma(raw_data, normalize = FALSE)
# plotting RLE
pdf("RLE.pdf " ,width=10,height=10,paper='special')
row_medians_assayData <-
  Biobase::rowMedians(as.matrix(Biobase::exprs(palmieri_eset)))
RLE_data <- sweep(Biobase::exprs(palmieri_eset), 1, row_medians_assayData)
wd3RLE_data <- as.data.frame(RLE_data)
RLE_data_gathered <-
tidyr::gather(RLE_data, patient_array, log2_expression_deviation)
ggplot2::ggplot(RLE_data_gathered, aes(patient_array,
log2_expression_deviation)) +
 geom_boxplot(outlier.shape = NA) +
 ylim(c(-2, 2)) +
 theme(axis.text.x = element_text(colour = "aquamarine4",
angle = 60, size = 6.5, hjust = 1 ,
face = "bold"))dev.off()
#quality assessment of data
pdf("PCA.pdf " ,width=10,height=10,paper='special')
palmieri_eset_norm <- oligo::rma(raw_data)
#PCA analysis
exp_palmieri <- Biobase::exprs(palmieri_eset_norm)
write.csv(exp_palmieri,"exp palmieri..normalized data.csv")
PCA <- prcomp(t(exp_palmieri), scale = FALSE)

percentVar <- round(100*PCA$sdev^2/sum(PCA$sdev^2),1)
sd_ratio <- sqrt(percentVar[2] / percentVar[1])
dataGG <- data.frame(PC1 = PCA$x[,1], PC2 = PCA$x[,2],
Disease =
Biobase::pData(palmieri_eset_norm)$Scan.Name)
ggplot(dataGG, aes(PC1, PC2)) +
  geom_point(aes(shape = Disease, colour = Disease)) +
  ggtitle("PCA plot of the calibrated, summarized data") +
  xlab(paste0("PC1, VarExp: ", percentVar[1], "%")) +
  ylab(paste0("PC2, VarExp: ", percentVar[2], "%")) +
  theme(plot.title = element_text(hjust = 0.5)) +
  coord_fixed(ratio = sd_ratio) +
  scale_shape_manual(values = c(5,15)) +
  scale_color_manual(values = c("darkorange","darkblue"))
dev.off()
#Heatmap clustering analysis
pData(palmieri_eset_norm)
disease_names <- ifelse(str_detect(pData
(palmieri_eset_norm)$Scan.Name,
"NL"), "NL", "P")
annotation_for_heatmap <-
  data.frame(  Disease = disease_names)
row.names(annotation_for_heatmap) <- row.names(pData(palmieri_eset_norm))
#Distances
dists <- as.matrix(dist(t(exp_palmieri), method = "manhattan"))
rownames(dists) <- row.names(pData(palmieri_eset_norm))
hmcol <- rev(colorRampPalette(RColorBrewer::brewer.pal(9, "YlOrRd"))(255))
colnames(dists) <- NULL
diag(dists) <- NA
ann_colors <- list(
  Disease = c(NL = "darkorange", P = "darkblue")
)
pdf("heatmap.pdf " ,width=10,height=10,paper='special')
pheatmap(dists, col = (hmcol),
annotation_row = annotation_for_heatmap,
annotation_colors = ann_colors,
legend = TRUE,
treeheight_row = 0,
legend_breaks = c(min(dists, na.rm = TRUE),
max(dists, na.rm = TRUE)),
legend_labels = (c("small distance", "large distance")),
main = "Clustering heatmap for the calibrated samples")
dev.off()
#Filtering based on intensity
pdf("historam.pdf " ,width=10,height=10,paper='special')
palmieri_medians <- rowMedians(Biobase::exprs(palmieri_eset_norm))
hist_res <- hist(palmieri_medians, 100, col = "cornsilk1", freq = FALSE,
main = "Histogram of the median intensities",
border = "antiquewhite4",
xlab = "Median intensities")
dev.off()
# Histogram of the median intensities per gene with manual intensity filtering threshold (red line)
pdf("historam.pdf " ,width=10,height=10,paper='special')
man_threshold <- 2
hist_res <- hist(palmieri_medians, 100, col = "cornsilk", freq = FALSE,
main = "Histogram of the median intensities",
border = "antiquewhite4",
xlab = "Median intensities")
dev.off()
#transcript less than threshhold are excluded
no_of_samples <-
  table(paste0(pData(palmieri_eset_norm)$Scan.Name,"_"))
no_of_samples
abline(v = man_threshold, col = "coral4", lwd = 2)
```

```
#sample cuttoff
samples_cutoff <- min(no_of_samples)
idx_man_threshold <- apply(Biobase::exprs(palmieri_eset_norm), 1,
function(x){
sum(x > man_threshold) >= samples_cutoff})
table(idx_man_threshold)
palmieri_manfiltered <- subset(palmieri_eset_norm, idx_man_threshold)
# annotation of transcription cluster
annotation(raw_data)
BiocManager::install(c("pd.hg.u133.plus.2","hgu133plus2.db"))
library(hgu133plus2.db)
anno_palmieri <- AnnotationDbi::select(hgu133plus2.db,
keys = (featureNames(palmieri_manfiltered)),
columns = c("SYMBOL", "GENENAME"),
keytype = "PROBEID")
anno_palmieri <- subset(anno_palmieri, !is.na(SYMBOL))
#Removing multiple mappings
anno_grouped <- group_by(anno_palmieri, PROBEID)
anno_summarized <-
  dplyr::summarize(anno_grouped, no_of_matches = n_distinct(SYMBOL))
head(anno_summarized)
anno_filtered <- filter(anno_summarized, no_of_matches > 1)
head(anno_filtered)
probe_stats <- anno_filtered
nrow(probe_stats)
#exclude duplicate
ids_to_exlude <- (featureNames(palmieri_manfiltered) %in% probe_stats$PROBEID)
table(ids_to_exlude)
palmieri_final <- subset(palmieri_manfiltered, !ids_to_exlude)
validObject(palmieri_final)
```

```
head(anno_palmieri)
# recall fdata
fData(palmieri_final)$PROBEID <- rownames(fData(palmieri_final))
fData(palmieri_final) <- left_join(fData(palmieri_final), anno_palmieri)
# restore rownames after left_join
rownames(fData(palmieri_final)) <- fData(palmieri_final)$PROBEID
validObject(palmieri_final)
#A linear model for the data
individual <-
as.character(Biobase::pData(palmieri_final)$Scan.Name)
disease <-
str_replace_all(Biobase::pData(palmieri_final)$Scan.Name,
" ", "_")
disease <-
  ifelse(str_detect(Biobase::pData(palmieri_final)$Scan.Name,
"P"), "D", "N")
i_PCOS <- individual[disease == "PCOS"]
design_palmieri_PCOS <- model.matrix(~ 0 + disease)
colnames(design_palmieri_PCOS)[1:2] <- c("D", "N")
rownames(design_palmieri_PCOS) <- individual
write.xlsx(design_palmieri_PCOS,file="designmatrix.xlsx")
contrast_matrix_PCOS <- makeContrasts(D-N, levels = design_palmieri_PCOS)
write.xlsx(contrast_matrix_PCOS,file="contrast_matrix_PCOS.xlsx")
palmieri_fit_PCOS <- eBayes(contrasts.fit(lmFit(palmieri_final,
design = design_palmieri_PCOS),
contrast_matrix_PCOS))
table_PCOS <- topTable(palmieri_fit_PCOS, number = Inf)
write.csv(table_PCOS,"table_PCOS.csv")
head(table_PCOS)
# Histogram of the p-values for disease
```

```
pdf("histogram.pdf " ,width=10,height=10,paper='special')
hist(table_PCOS$P.Value, col = brewer.pal(3, name = "Set2")[2],
main = "NL vs P", xlab = "p-values")
dev.off()
#subset
tail(subset(table_PCOS, P.Value < 0.05))
total_genenumber_PCOS <- length(subset(table_PCOS, P.Value < 0.05)$SYMBOL)
#volcano plot
write.csv(table_PCOS,file="DEGenes.csv")
library(EnhancedVolcano)
de =read.csv("DEGenes.csv" , header = TRUE, sep=",")
de_U <- de[!(is.na(de$SYMBOL) | de$SYMBOL==""),]
#threshold value
pval <- 0.05
fc<- 0.5
pdf("volcanofinal.pdf" , width=10, height=10, paper='special')
EnhancedVolcano(de_U,
lab = de_U$SYMBOL,
x = "logFC",
y = "P.Value",
pCutoff = pval,
FCcutoff = fc,
title = "volcano plot")
dev.off()
remp <- de_U[de_U$P.Value < pval,]
remFU <- remp[remp$logFC < -fc | remp$logFC > fc,]
dim(remFU)
write.csv(remFU,"final.csv")
```

Appendix B

Source code for RNA-Seq (Differential expression analysis)

```
library(ballgown)

library(RSkittleBrewer)

library(genefilter)

library(plyr)

library(devtools)

library(ggplot2)

data_dir<-("E:/StringTie file pcos 1")

getwd()

pheno_data = read.csv("pheno.csv")

bg = ballgown(dataDir= "source", samplePattern='sample', meas='all', pData=pheno_data)

#save(bg, file='bg.rda')

#structure(bg)$exon

bg

#Filtering low abundance genes

bg_filt = subset(bg,"rowVars(texpr(bg)) >1",genomesubset=TRUE)

#loading gene names

bg_table = texpr(bg_filt, 'all')

#bg_gene_names = unique(bg_table[, 1:10])

#Pull the gene_expression data frame from the ballgown object

gene_expression = as.data.frame(gexpr(bg_filt))

transcript_expression = as.data.frame(texpr(bg_filt))

head(transcript_expression)

row.names(transcript_expression)

write.csv(bg_gene_names, "bg_gene_names.csv")

write.csv(bg_table, "bg_table.csv")
write.csv(results_genes, "results_genes.csv")
# transcripts analysis
results_transcripts = stattest(bg_filt,
feature="transcript",covariate="Phenotype",adjustvars = NULL,
getFC=TRUE, meas="FPKM")
results_genes = stattest(bg_filt, feature="gene",
covariate="Phenotype", adjustvars = NULL, getFC=TRUE,
meas="FPKM")
results_transcriptsmer =
data.frame(geneNames=ballgown::geneNames(bg_filt),
geneIDs=ballgown::geneIDs(bg_filt), transcriptNames=ballgown::transcriptNames(bg_filt),
results_transcripts)
results_transcripts = arrange(results_transcripts,pval)
results_genes = arrange(results_genes,pval)
write.csv(bg_gene_names, "bg genes name.csv")
write.csv(results_transcripts, "transcript2_results.csv",
row.names=FALSE)
write.csv(results_genes, "gene_results.csv",
row.names=FALSE)
tra <- subset(results_transcripts,results_transcripts$pval<0.05)
gen <- subset(results_genes,results_genes$pval<0.05)
write.csv(tra, "filtered transcripts.csv")
write.csv(gen, "filtered genes.csv")
#######Box plot########
tropical=
c("cyan","cyan3","cyan4","lightcoral","salmon","cadetblue1","burlywood","darkseagreen","darkslategray"

,"gray73","gray29","lightgray","ivory","mediumorchid1","mediumorchid4","midnightblue","powderblue","skyblue",

"hotpink","limegreen","yellow","darkorange","dodgerblue","blue","purple","brown","red","seagreen","pink","yellow")
```

```
palette(tropical)
fpkm = texpr(bg_filt,meas="FPKM")
fpkm = log2(fpkm+1)
write.csv(fpkm, "fpkm")
library(viridis)
#oligo::boxplot(fpkm,col=viridis_pal(pheno_data$phenotype)(n=6),las=2, cex.axis= 0.4,
             main = "log2(FPKM+1)")
#boxplot (fpkm,col=as.numeric(pheno_data$phenotype),las=2,ylab='log2(FPKM+1)')
pdf("boxplotNorm.pdf " ,width=10,height=10,paper='special')
boxplot(gene_expression, col=rainbow(28),
        las=2, ylab="log2(FPKM)",
        main="Distribution of FPKMs for all 6 samples")
boxplot(log2(gene_expression+1), col=rainbow(6),
        las=2, ylab="log2(FPKM)",
        main="log2(FPKM+1)")
dev.off()
transcript_gene_table = indexes(bg_filt)$t2g
head(transcript_gene_table)
#Each row of data represents a transcript
. Many of these transcripts represent the same gene. Determine the numbers of transcripts and unique genes
length(row.names(transcript_gene_table))
length(unique(transcript_gene_table[,"g_id"]))
counts=table(transcript_gene_table[,"g_id"])
c_one = length(which(counts == 1))
c_more_than_one = length(which(counts > 1))
c_max = max(counts)
hist(counts, breaks=50, col="bisque4", xlab="Transcripts per gene", main="Distribution of transcript count per gene")
legend_text = c(paste("Genes with one transcript =", c_one), paste("Genes with more than one transcript =", c_more_than_one
            ), paste("Max transcripts for single gene = ", c_max))
```

```
legend("topright", legend_text, lty=NULL)

#Plot #2 - the distribution of transcript sizes as a histogram

full_table <- texpr(bg_filt , 'all')

hist(full_table$length, breaks=50, xlab="Transcript length (bp)", main

        ="Distribution of transcript lengths",

         col="steelblue")

##########################################################

data_colors=(c("cyan","cyan3","cyan4","lightcoral","salmon","cadetblue1","burlywood","darkseagreen","darkslategray",


"gray73","gray29","lightgray","ivory","mediumorchid1","mediumorchid4","midnightblue","powderblue","skyblue",

"green","limegreen","yellow","darkorange","dodgerblue","blue","purple","brown","red","seagreen","pink","yellow"))

min_nonzero=1

#Set the columns for finding FPKM and create shorter names for figures

data_columns=c(1:30)

short_names=c("sample 1","sample 10","sample 11","sample 12","sample 13","sample 14",

        "sample 15","sample 16","sample 17","sample 18","sample 19","sample 2","sample 20","sample 21"

        ,"sample 22","sample 23","sample 24","sample 25","sample 26","sample 27","sample 28","sample 29","sample 3","sample 30","sample 4","sample 5","sample 6","sample 7","sample 8","sample 9")

#Plot #3 - View the range of values and general distribution of FPKM values for all libraries Create boxplots for this purpose

#Display on a log2 scale and add the minimum non-zero value to avoid log2(0)

boxplot(log2(transcript_expression [,data_columns]+min_nonzero),

 col=data_colors, names=short_names, las=2, ylab="log2(FPKM)", main="

Distribution of FPKMs ")

####################################################

colors = colorRampPalette(c("blue", "blue", "#007FFF", "cyan","#7FFF7F", "yellow", "#FF7F00", "red", "#7F0000"))

#smoothScatter(x=log2(x+min_nonzero), xlab="FPKM (SRR218_N, Replicate 1)", ylab="FPKM (SRR219_N, Replicate 2)", main="Comparison of expression values for a pair of replicates", colramp=colors, nbin=200)
```

```
#Compare the correlation 'distance' between all replicates

transcript_expression[,"sum"]=apply(transcript_expression[,data_columns], 1, sum)

#Identify the genes with a grand sum FPKM of at least 5 - we will filter out the genes with very low expression across the board

i = which(transcript_expression[,"sum"] > 5)

#Calculate the correlation between all pairs of data

r=cor(transcript_expression[i,data_columns], use="pairwise.complete.obs",

 method="pearson")

#Plot #8 - Convert correlation to 'distance', and use 'multi-dimensional scaling' to display the relative differences between libraries

d=1-r

data_columns=c(1:30)

mds=cmdscale(d, k=2, eig=TRUE)

par(mfrow=c(1,1))

plot(mds$points, type="n", xlab="", ylab="", main="MDS distance plot ", xlim=c(-0.25,0.25), ylim=c(-0.25,0.25))

points(mds$points[,1], mds$points[,2], col="grey", cex=2, pch=16)

text(mds$points[,1], mds$points[,2], short_names, col=data_colors)

sig=which(results_transcripts$pval<0.05)

results_transcriptsmer[,"de"] =

 log2(results_transcriptsmer[,"fc"])

hist(results_transcripts[sig,"de"], breaks=50,

 col="seagreen", xlab="log2(Fold change Normal-Diseased", main="

            Distribution of differential expression values")

abline(v=-2, col="black", lwd=2, lty=2)

abline(v=2, col="black", lwd=2, lty=2)

legend("topleft", "Fold-change > 4", lwd=2, lty=2)

#enhanced volcano###########################

library(EnhancedVolcano)

results_transcriptsmer

EnhancedVolcano(results_transcriptsmer,

lab = results_transcriptsmer$geneNames,
x = "de",
y = "pval",
pCutoff = 0.05,
FCcutoff =0.5,
title = "RNA-dataset")
trans <- subset(tra, pval < 0.05)
trans <- subset(tra, fc < -0.5 | fc > 0.5)
write.csv(tra,"results_transcriptsmert with cutoffvolcano.csv")
  #volcano
volcano_names <- ifelse(abs(tra$coefficients)>=1,
tra$id, NA)
volcanoplot(tra, coef = 1L, style = "p-value", highlight = 100,
names = volcano_names,
            xlab = "Log2 Fold Change", ylab = NULL, pch=16, cex=0.35)
#################################PCA####################################
pca_data=prcomp(t(gene_expression))
percentVar <- round(100*pca_data$sdev^2/sum(pca_data$sdev^2),1)
percentVar
sd_ratio <- sqrt(percentVar[2] / percentVar[1])
sd_ratio
sd_ratio = 1.5
dataGG <- data.frame(PC1 = pca_data$x[,1], PC2 =pca_data$x[,2],
             Phenotype = pheno_data$Phenotype)
dataGG
#pdf("PCA plot of the log-transformed raw expression data.pdf",width=7,height=5,paper='special')
pdf("pcaNorm.pdf " ,width=10,height=10,paper='special')
ggplot(dataGG, aes(PC1, PC2)) +
  geom_point(aes(colour = Phenotype)) +
  ggtitle("PCA plot of the normalized data") +
  xlab(paste0("PC1, VarExp: ", percentVar[1], "%")) +
  ylab(paste0("PC2, VarExp: ", percentVar[2], "%")) +
  theme(plot.title = element_text(hjust = 0.5))+
  coord_fixed(ratio = sd_ratio) +
  scale_shape_manual(values = c(4,15)) +
  scale_color_manual(values = c("darkorange2", "dodgerblue4"))
dev.off()
```