

**Meta-analysis of genomic and
expression data
in endometrial cancer**

M.A.I. Perera



Meta–analysis of genomic and expression data in endometrial cancer

**M.A.I.Perera
Index No: 14001136**

Supervisor: Mrs. Rupika Wijesinghe

December 2018

Submitted in partial fulfillment of the requirements of the
B.Sc in Computer Science Final Year Project (SCS4124)



Declaration

I certify that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, be made available for photocopying and for interlibrary loans, and for the title and abstract to be made available to outside organizations.

Candidate Name: M.A.I. Perera

.....

Signature of Candidate

Date:

This is to certify that this dissertation is based on the work of Ms. M.A.I.Perera under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Supervisor's Name: Mrs. Rupika Wijesinghe

.....

Signature of Supervisor

Date:

Co- Supervisor's Name: Dr. Ruwan Weerasinghe

.....

Signature of Supervisor

Date:

Abstract

Endometrial carcinoma is the commonest gynaecological cancer in the world, with huge hormonal influence, as it occurs in the inner lining of the uterus where the cell proliferation, regeneration, and functions are maintained through hormonal influence. But there is a lack of researches done on the analysis of the genetic level hormonal influence of endometrial cancer. Mostly endometrial carcinoma is diagnosed when the carcinoma is confined to the uterus and also current therapeutics fail to treat late-stage disease.

This study was conducted to identify meaningful subtypes in endometrial carcinoma and to identify the hormonal influence in each subtype. Study will help in improving the efficiency of the treatments and reducing the toxicity of the treatments by identifying clues to find target therapeutics through the analysis of genetic level hormonal influence. Gene expression data obtained from cBioportal has been analysed in this research using unsupervised learning techniques since gene expression data illustrate the biological processes happening inside the cells. Partitioning around medoids, K-means and Hierarchical clustering techniques used in initial clustering and hierarchical clustering technique has been used with different distance and linkage measures, to identify meaningful clusters.

After the cluster identification, cluster validation has been conducted according to internal measures like Silhouette, Dunn index and relative measures to identify optimal number of clusters. External validations like comparing the classes with clinical variables and visual analysis of the classes using heatmaps has also been conducted. Identified clusters are analysed to find the hormonal influence inside each cluster and stage-specific hormonal influence also analysed using feature filtering methods like Between Sum of Square / Within Sum of Square.

After heatmap filtering it was identified three cluster analysis results have meaningful clusters, out of them identified the cluster analysis with 15 clusters from hormonal gene dataset having significant genes in each cluster.

Preface

The results in this study rely upon data provided by Pan Cancer study through cBioPortal. The analysis of the data is entirely my own work which I carried out with the help of R language, R packages, Panther as tools.

Acknowledgement

I take this moment to convey my appreciation and gratitude towards my supervisor Senior Lecture Mrs. Rupika Wijesinghe, UCSC and co supervisor Senior Lecturer Dr. Ruwan Weerasinghe, UCSC for their commendable heading, observing, constance support, proper guidance, helpful advices, sparing their valuable time from the very early stage of the research, motivation not only for the research work but also for the life throughout this task. The help and direction given by them time to time might convey us long course in the trip of life on which we are going to set out.

I would like to express my gratitude for bioinformatics team; Bhagya Madhubhanie Senadheera, Upul Anuradha, Kulani Sumanasekera, Dinithi Rajapaksha for the help they given me to make this study success from the beginning.

I owe a considerable amount to my family and friends who gave me support and encouragement to make this success.

Finally, I would like to thank all the people whose names are not appeared, but their untiring effort was crucial to make this study success.

Table of Contents

Declaration	iii
Abstract	iv
Preface	v
Acknowledgement	vi
Table of Contents	vii
List of Figures	x
List of Tables	xii
List of Acronyms	xiii
Chapter 1 - Introduction	1
1.1 Background to the Research	1
1.2 Research Problem and Research Questions	2
1.3 Justification for the research	3
1.4 Methodology	3
1.5 Outline of the Dissertation	5
1.7 Delimitations of Scope.....	5
1.8 Conclusion	6
Chapter 2 - Literature Review	7
2.1 Introduction.....	7
2.2 Theories.....	7
2.3 Related works.....	16
2.4 Summary	18
Chapter 3 - Design	20
3.1 Introduction.....	20

3.2 Research Methodology	20
3.3 Design	20
3.3.1 <i>Collecting Data</i>	22
3.3.2 <i>Identifying Endometrial cancer subtype</i>	22
3.3.3 <i>Identifying Hormonal Influence</i>	24
3.3.4 <i>Identifying Stage Specific Genes</i>	24
3.4 Tools used for analysing	25
3.4.1 <i>R, R studio, R markdown</i>	25
3.4.2 <i>Panther</i>	25
3.4.3 <i>NCBI</i>	25
3.5 Conclusion	26
Chapter 4 - Implementation.....	27
4.1 Introduction.....	27
4.2 Collecting Data	27
4.3 Pre-process the data	29
4.3.1 <i>Basic analysis and handling missing values</i>	29
4.3.2 <i>Feature extraction</i>	31
4.4 Clustering.....	32
4.4.1 <i>Assessing Cluster Tendency</i>	32
4.4.2 <i>Identifying Optimal number of K</i>	32
4.4.2 <i>Clustering</i>	32
4.5 Validating.....	33
4.5.1 <i>Internal Validation</i>	33
4.5.2 <i>Relative Validation</i>	33
4.5.3 <i>External Validation</i>	33
4.6 Analysing	33
Chapter 5 - Results and Evaluation	35

5.1 Data collecting and pre-processing	35
5.1.1 <i>Selecting data set</i>	35
5.1.2 <i>Handling Missing values</i>	38
5.1.3 <i>Extracting Features</i>	39
5.2 Identifying Endometrial Cancer Subtypes	40
5.2.1 <i>Analysing Cluster Tendency</i>	40
5.2.2 <i>Identifying optimal number of K</i>	40
5.2.3 <i>Clustering</i>	42
5.2.4 <i>Validating</i>	46
5.3 Identifying Hormonal influence in subtypes and stages	51
Chapter 6 - Conclusions	58
6.1 Introduction.....	58
6.2 Conclusions about research questions (aims/objectives)	58
6.3 Conclusions about research problem	59
6.4 Limitations	60
6.5 Implications for further research.....	60
References	61
Appendix A: Diagrams	64
Appendix B: Selected Hormonal gene Details.....	67

List of Figures

Figure 3.1: Mind map of Highlevel Research design	21
Figure 4.1: Import and arrange data frame for analysis	27
Figure 4.2: In detail Research Conceptual Model	28
Figure 4.3: Basic analysis done for expression data	29
Figure 4.4: Hard to interpret missing data patterns due to high dimensionality	30
Figure 4.5: Hard to interpret missing data patterns due to high dimensionality	30
Figure 4.6: Missing patterns using VIM package	31
Figure 4.7: BSS/WSS function	34
Figure 5.1 Missing value pattern of expression data	38
Figure 5.2: Eliminate the variables having high missingness	39
Figure 5.3: Ordered dissimilarity matrix of genes selected with high variance	40
Figure 5.4: Optimal number of clusters genes having high variance	41
Figure 5.5: Optimal number of clusters hormonal genes	42
Figure 5.6: Comparison between Hierarchical. K-means and PAM	44
Figure 5.7: Spearman distance and complete linkage hierarchical clustering applied	45
Figure 5.8: Internal validity of correlational distance of PAM, Kmeans, Hierarchical clustering	46
Figure 5.9: Internal validity of Euclidean distance of PAM, K-means, Hierarchical clustering	47
Figure 5.10: Correlational plot of different distance measures and linkage measures	48
Figure 5.11: Two Heatmaps of highly varied genes split under 3 and 4 clusters identified in spearman distance and complete linkage in highly varied gene dataset.	49
Figure 5.12: comparison between classes identified using spearman distance and complete linkage	50
Figure 5.13: comparison between classes identified using spearman distance and complete linkage.	50
Figure 5.14: Heatmap with 15 clusters in hormonal gene dataset	50

Figure 5.15: Discriminative 15 hormonal genes highly expressed only at one cluster at a time 57

Figure 5.16: Pathway analysis of 15 discriminative hormonal genes of 15 clusters identified using hormonal gene dataset 57

List of Tables

Table 5.1: Basic data analysis on the uterus cancer dataset	35
Table 5.2 Basic data analysis on the Endometrial cancer dataset	36
Table 5.3 Basic data analysis on Uterine Corpus Endometrial Carcinoma (TCGA, PanCancer Atlas 2018)	37
Table 5.4: Basic exploration on the datafiles available in Uterine Corpus Endometrial Carcinoma (TCGA, PanCancer Atlas 2018) dataset	37
Table 5.5 Stages specific discriminative gene set from all gene set	51
Table 5.6: Stages specific discriminative hormones in all gene set	54
Table 5.7: 15 discriminative hormonal genes of Cluster 15 in hormonal gene dataset	57

List of Acronyms

A	-	Adenine
C	-	Cytosine
CNA	-	Copy Number Alteration
CNV	-	Copy Number Variation
DCC	-	Data Coordination
DNA	-	Deoxyribonucleic Acid
EC	-	Endometrial Cancer
G	-	Guanine
GDAC	-	Genome Data Analysis Centre
HGNC	-	HUGO Gene Nomenclature Committee
HUGO	-	The Human Genome Organisation
NCBI	-	National Centre for Biotechnology Information
PANTHER	-	Protein Analysis Through Evolutionary Relationships
PCA	-	Principle Component Analysis
SDAE	-	Stacked denoising autoencoder
SNP	-	Single Nucleotide Polymorphism
T	-	Thymine
TCGA	-	The Cancer Genome Atlas
UCEC	-	Uterine Corpus Endometrial Carcinoma
VAT	-	Visual Assessing Technique
VIM	-	Visualization and Imputation of Missing Values

Chapter 1 - Introduction

1.1 Background to the Research

In human body, there are different kinds of cells. For example, brain cells, muscle cells, blood cells and etc. Most of the cells, have a nucleus in the middle. In this nucleus, there are 46 chromosomes and DNA (Deoxyribonucleic Acid) strands are bundled up inside these chromosomes. If these DNA strands pull out 20,000+ genes can be identified and if we look more closely, we could see all these genes in the sense DNA. DNA strand has been built with basically 4 nucleotides. Those are Adenine(A), Thymine(T), Cytosine(C), and Guanine(G). A is connected with T and C is connected with G in making base pairs. Aforementioned DNA strand's fundamental building blocks are these two types of base pairs and there are 3 billion base pairs in human DNA. This DNA structure contains information to make proteins, and proteins determine the traits and characteristics of an organism. Basically, in the human body every cell contains same DNA, that is maintained by DNA replication where, cell divide to make more cells.

Cell cycle regulation is essential for healthy cell growth. It is controlled by genes and environmental factors. Cancer cells occur due to alterations happened in the genes which regulate the cell cycle in the sense of mutations that happened to those genes. Mutations are the changes in nitrogen bases in DNA sequence. Some of these mutations are good for organisms and some are bad, but most of the mutations are neutral. There are two kinds of mutations. Those are point mutations and frameshift mutations. In point mutations, a single base is changed. those also called as single nucleotide variations (SNV). This will cause to change only one amino acid. In frameshift mutations, there can be insertion or deletions of at least one nucleotide in the DNA sequence which will affect all amino acids after the mutation when synthesizing proteins. Copy number alteration (CNA) is a kind of frameshift mutation. As mentioned above every cell in human body has the same exact set of DNAs, but all cells are not the same. There are blood cells, nerve cells, cardiac cells and so on. The process of the cell becoming specialized or cell differentiation is controlled by the way

genes are expressed. In a cell all genes are not expressed. Some may be expressed and some may not. This is like a light bulb. When the switch is turned on the bulb will emit light as the output, same way expression controls act as the switch if they present the gene will be expressed and output proteins. There can be internal factors like hormones and external factors like radiation or alcohol which may influence this process. This study will analyse expression data to identify expression patterns in the tumor samples collected from endometrial cancer (EC) patients.

EC is a cancer that arises from the endometrium (the lining of the uterus/womb). It is the commonest gynaecological cancer in the world [1] and has an alarmingly increasing incidence related to longevity and obesity. Estimates suggest, by 2025, an incidence increase of EC in the range of 50%–100% will occur, relative to the observed incidence in 2005 [1]. Most of the women with Endometrial carcinoma are postmenopausal whose age is averaged to 60 years[2], and due to other commonly coexisting medical problems, such as obesity, hypertension and diabetes not suitable for standard surgical [3] and majority of these gynaecological surgery for cancer experience serious surgical complications. The 5-year survival rates for advanced EC are 23%, worse than for other common gynaecological cancers, and similar to that of ovarian cancer. Cancer statistics centre has estimated that there can be 63,230 cases in 2018 in the world [4].

1.2 Research Problem and Research Questions

1.2.1 Research Problem

Estimated number of EC cases in 2018 according to the cancer statistic center is 63,230 [4]. Although EC is the most common gynaecological cancer, there is a knowledge gap in the genetic level hormonal influence of EC [5] which can help to identify effective and less toxic therapeutics by finding subtypes which has significant hormonal gene impact by identifying subtypes of EC related to hormonal genes.

1.2.2 Research Questions

- What are the subtypes in endometrial cancer?
- What are subtype specific gene expressions of endometrial cancer?
- What are the hormonal gene expression profiles for each identified subtype of EC?
- What are the stage specific gene expression profiles in EC?

1.3 Justification for the research

It could be **maximized efficacy in treatments and minimized toxicity** target specific therapies to genetically different tumor types. Primarily cancers classify using morphological appearance of the tumor which has many limitations. Although a tumor has similar histopathological appearance it can have different clinical causes and different responses to therapies [6]. Improvements in cancer classification will lead to advance cancer treatments. Earlier EC has been broadly classified into two groups as endometrioid which are linked to estrogen excess, obesity, hormone-receptor positivity and serous tumors which are common in older, non-obese women. Early-stage endometrioid cancers are often treated with adjuvant radiotherapy, where serous tumors are treated with chemotherapy, similar to advanced stage cancers of either histological subtype. As such, proper subtype classification is a necessary for select appropriate therapies. Identifying the hormonal gene expression profiles in cancer subtype will help to identify better target therapeutics

In order to **improve survival rate** in cancer, identifying stage specific hormonal genes and predicting the stage of the cancer will be helpful.

1.4 Methodology

In the methodology of this study expression data collected of endometrial carcinoma from cBioPortal [7] to find the similarities between samples to identify classes of EC with the intention of helping to **find clues for target therapeutics**. Then find hormonal genes and genes having high coefficient of variance expression profiles inside the subtypes identified. To find clues to **identify target therapeutics for stages**

and identify the stage of the patient, analyze stage specific data inside previously identified subtypes and whole data set.

In order to **find similar samples(clusters)** in EC need to pre-process the data gathered from cBioPortal. In pre-processing first handle the missing values. Rather than imputing the missing values which can be adversely affect the underlying patterns in data, remove missing values after identifying the missing value patterns. There are 20,000+ number of genes which considered as variables and there are only 500+ samples which create curse of dimensionality, it is necessary to extract features. Two methods are used in this study. Extract features using variance thresholds which remove features considering coefficient of variance. To make sure biological value is considered, and as the main objective of this study is to analyze hormonal gene impact, using the hormonal gene group identified by the literature. Then using the extracted features cluster tendency has been analysed using Cluster tendency assessing visual method (VAT) [8] and Hopkin's statistics. Using Gap statistics, Silhouette and Elbow method the number of appropriate clusters can be evaluated.

Then using PAM, K-means and hierarchical clustering techniques used to cluster the samples. After identified the classes to validate clusters used internal, external and relative validation measures.

After identify the meaningful classes **find subtype specific gene expressions** for each subtype by comparing the correlation and variation of expression levels of genes among clusters can identify significant genes have different patterns among the classes identified.

To find **hormonal genomic profiles in identified subtypes** using the hormonal genes identified through literature, generate the genomic profiles considering the variation of expression levels in each subtype.

To **find stage specific gene expressions** in each subtype identified the analyze samples in each subtype separately to identify significant changes among stages within subtypes and also if a subtype contains whole set of samples with one particular stage then analyze whole dataset for identify significant changes among stages to help stage specific target therapeutics.

1.5 Outline of the Dissertation

Chapter 01 - It has an explanation of background of the study, research problem, objectives of the study and the significant of the study.

Chapter 02 - This chapter mainly emphasizes on biological theories and related works carried out in the relevant theoretical findings of the problem in this study investigate

Chapter 03 - This chapter explains the approach of the research conducted

Chapter 04 - This chapter is based on implementations of the analysis carried out in the research

Chapter 05 - Finally this presents the Results and evaluation of the results obtained in the research

Chapter 06 - Findings for each research question or hypothesis are summarized explained within the context of this and prior research examined

1.7 Delimitations of Scope

- There are two kinds of uterus cancers which were endometrial carcinoma and uterine carcinosarcoma. ECs are more frequently occur (80%) than the carcinosarcoma (2%-4%) and the available data set is relatively smaller than the EC in carcinosarcoma, which will be discussed in Chapter – 5 section and also the lack of knowledge of hormonal influence in molecular level research gaps was related to EC. Considering those facts Endometrial carcinoma has been chosen as the cancer type will be analysed in this study.
- In this study Expression data in Endometrial carcinoma will be analysed as they are available in cBioPortal and they give genetic level information about cancers.
- There were three datasets available under endometrial carcinoma, but after the basic analysis about the datasets, it was found that the most of the samples used in those three samples were the same. Hence one dataset from a study of Uterine Corpus Endometrial Carcinoma (TCGA, PanCancer Atlas) [9] which have 529 samples was selected as it was the most recently updated one and it was a part of a big project done by TCGA named as pan cancer atlas which make the data set more reliable.

- The data set is having samples from different countries and different ethnic groups.

1.8 Conclusion

This chapter gives a brief introduction to this dissertation. It has introduced the research problem, research questions, hypothesis and key deliverables key resources and brief description about the work flow, the dissertation was outlined, and the limitations were given. On these foundations, the dissertation can proceed with a detailed description of the research.

Chapter 2 - Literature Review

2.1 Introduction

Literature review is presented in order to provide a general idea about the key aspects and theories about the data analysing process of biological data and the related works. This chapter will explain theoretical models and key concepts will be used to build the logic in this research study.

2.2 Theories

Theories section consist of the theoretical information gained through the literature about biological information of ECs and phases of analysing biological data such as data gathering, pre-processing, reducing the dimensionality, transforming, analysing, and interpreting biological data to identify the subtypes of ECs.

2.2.1 Biological Data & Bioinformatics

DNA (Deoxyribonucleic Acid) is the genetic material in every organism, structure of this is same in all organisms. It contains information to make protein and protein will determine the characteristics of organisms. Each of cell in a person consist with exact same set of DNA.

There are two type of tumors (clumps formed by cancer cells, but not all cancers form solid tumors),

- **Benign** - Not harmful, not cancerous, can be removed, remain clustered
- **Malignant** - Harmful, Cancerous, Metastasized or breakaway can form more tumors

Cancers can happen due to the damage happen to cells, which come from mutations. Mutations are the changes in the sequence of DNA (changes in nitrogen bases in DNA). Mutations can happen after somebody is born or passed down from parents. But not all mutations harm organisms. Some are good, some are extremely bad and

most of them are neutral. Basically, there are two kind of mutations, which are point mutations and frameshift mutation.

Protein synthesis is a process of making protein using instructions encoded in DNA (genes). This happen in two steps which are

- **Transcription** - Convert DNA into messenger RNA
 1. DNA unwound at a gene
 2. RNA polymerase (enzyme) copies DNA strand creating a complementary mRNA strand
 3. mRNA moves from nucleus to cytoplasm and bind with ribosome to begin translation
- **Translation** - Using information in messenger RNA make protein which can be understood by the body like enzymes, hormones.
 1. mRNA attached to start codon (triplet sequence of those new nucleotide bases)
 2. transfer RNA (tRNA) carries amino acid specific to the codon on the mRNA
 3. Amino acids are bound to create a polypeptide, which will continue till stop codon is reached.

Although all the cells consist with exact same set of DNAs, not all genes in DNA is expressed every time. So, the behaviour of a cell can be identified using the mutation data, copy number data and expression data.

New advancements in technology allow biologist to generate a humongous amount of data from measurements of genomic sequence to images of physiological structure. There are different levels of information like mRNA expression, transcription factor binding, protein expression and etc. Now the technology exists to probe each of these levels in high throughput. These technological advancements lead to two key questions to address. What could be measured? And how could be analysed collected data to discover underlying biology? The fundamental problem which will be addressed in this study is the second one. How could be analyse collected data to discover underlying biology of ECs?

This study will use Pan cancer dataset (Uterine Corpus Endometrial Carcinoma (TCGA, PanCancer Atlas)) [9].

2.2.2 Endometrial Cancer

Endometrium is the lining of uterus, which consist with two layers. Single layer of columnar epithelium resting on stroma, a layer of connective tissue that varies in thickness according to the hormonal influences. Breast and ovarian cancers, obesity, late menopause due to estragon excess, diabetes, hypertension, unopposed oestrogen, polycystic ovarian syndrome considered as risk factors [10].

Earlier EC has been broadly classified into two groups as

- Endometrioid Adenocarcinoma

linked to oestrogen excess, obesity, hormone-receptor positivity. Early-stage endometrioid cancers are often treated with adjuvant radiotherapy while advanced stage ECs treated with chemotherapy. Postmenopausal bleeding considers as clinical sign.

- Serous Carcinoma/ Clear Cell

common in older, non-obese women and have worse outcome, no precursor, more aggressive. Serous tumours are treated with chemotherapy in both advanced and early stages. Bloating, pelvic pressure, bowel dysfunction is considered as clinical signs.

This classification is based on the model proposed by Bokhman in 1983, according to the oestrogen dependent and oestrogen independent [11] [12] are shown that the current pathological classification and grading system of high-grade endometrial carcinomas is limited in both reproducibility and prognostic ability. Therefore, to select appropriate adjuvant therapies it is necessary to have proper subtype classification in molecular level [13].

EC is the commonest gynaecological cancer has an alarmingly increasing incidence related to longevity and obesity which arises from the lining of uterus. Estimates suggest, by 2025, an incidence increase of EC in the range of 50%–100% will occur, relative to the observed incidence in 2005 [1]. Most of the women with Endometrial carcinoma are postmenopausal whose age is averaged to 60 years[2], and not suitable for standard surgical treatment because of other commonly coexisting medical problems, such as obesity, hypertension and diabetes [3] and the majority of these gynaecological surgeries for cancer experience serious surgical complications. The 5-year survival rates for advanced EC are 23%, worse than for other common gynaecological cancers, and similar to that of ovarian cancer. Cancer statistics centre has estimated that there can be 63,230 cases in 2018 [4].

Human endometrium is the primary recipient organ for ovarian steroid hormonal signal and is intricately responsive to these hormones. The three main classical ovarian steroid hormones, estragon, progesterone and androgens regulate normal human endometrial cell proliferation, regeneration and function therefore are implicated in endometrial carcinogenesis directly or via influencing other hormones and metabolic pathways [5].

This hormonal control of endometrium starts from the hypothalamus which stimulate pituitary gland by releasing Gonadotropin-releasing hormone (GnRH). The pituitary gland releases Follicle-Stimulating hormone (FSH) which circulate through blood to ovaries and stimulate group of follicles to grow from primary to secondary and produce oestrogen hormone which stimulate the growth of endometrium. When oestrogen hormone increases hypothalamus will increase GnRH and induce production of Luteinizing hormone (LH) which triggers ovulation and release egg from ovaries where the swept (ruptured follicle) become corpus luteum, which secretes progesterone hormone. The Progesterone hormone stimulate uterine development, in the absence of pregnancy corpus luteum atrophies and reduce Progesterone hormone which will break down the endometrium.

All types of ECs may share common etiological factors, including their response to/stimulation by oestrogen and other ovarian steroid hormones. Although many reviews previously describe the role of oestrogen and progesterone in endometrial carcinogenesis, they have largely disregarded the involvement of androgens and other hormones like gonadotropin-releasing hormone (GnRH) and luteinizing hormones (LH) in EC. This [5] study is a review article done to consolidate the current knowledge by carrying out PubMed (Medline) and Ovid searches systematically for publications from November 2000 until November 2015 of the involvement of the three main endogenous ovarian hormones (estragon, progesterone and androgens) as well as the other hormones in endometrial carcinogenesis, to identify important avenues for future research.

2.2.3 Machine learning

In the article 'big data bioinformatics' [14][15] concentrate on analysing the big data gathered using supervised and unsupervised machine learning techniques. In supervised learning approach, labelled data has been analysed and produce an inferred

function, which can be used for classifying samples to improve diagnosis. In unsupervised method does not need labelled data (without predefined classes). This will discover hidden patterns within the data which can be used to identify subtypes [10]. In simplest words supervise learning will discover the relationships between data and predefined labels(output) and unsupervised will discover the relationships between data itself (uncover the underlying organizational structure of data). There is another method called semi-supervised method, which combined both labelled and unlabelled data. But this method is not yet widely applied in biology. Unsupervised learning to detect clusters (identifying subtypes) and supervised learning to classify new samples (improve diagnosing) can be used in this study.

2.2.4 Imputing missing values

Imputing missing values in genomic data is difficult due to the high dimensionality of data. The main occurrence of missing values in gene expression data are dust or scratches on the slides, hybridization failures, corruptions, image noise and insufficient resolution. These missing values effect machine learning techniques like dimensionality reduction, clustering as they use the patterns in the data itself without using any output labels. Simply removing the observations containing missing values and replacing missing values with average/mean are basic missing value imputation techniques which have a disadvantage of ignoring the correlation among genes.

In general, there are two types of information can be used.

- Correlation structure between entries in the data matrix
- Domain knowledge of the data or the process that generate the data

Missing value imputation techniques can be broadly categorized into 3 sections, according to the information used to impute the missing values [16].

- Global approach

Use the correlation information derived from the entire data matrix. Less accurate when the genes exhibit dominant local similarities. Singular value decomposition-based imputation and Bayesian principle component analysis (PCA) are some techniques used in global approach. This method is preferred when the data is homogeneous.

- Local approach

Use the correlation information derived from the local similarity structure (subset of genes which show high correlation with the gene showing missing value). K nearest neighbour based, Sequential K nearest neighbour, Iterative K nearest neighbour, Gaussian mixture clustering based, Single linear regression, Multiple linear regression, Sequential multiple linear regression, Iterated multiple linear regression, Least square regression with principal components, Linear and non-linear regression with Bayesian gene selection, Linear regression with multiple parallel imputations, CMVE with automatic determination of number of reference genes, AR modelling with least square regression can be considered as imputation techniques using local approach. This approach is preferred when the data is heterogenous

- Knowledge assisted approach

In this approach knowledge or external information integrate to the process. This method will obviously improve the accuracy than the two purely data driven approaches discussed above. Especially for the small number of samples and high missing rate.

In [17][18] analysed the impact of the missing value imputation techniques in classification and clustering. Analysis done on hierarchical clustering with both complete and average linkage, k-medoids. Imputation techniques make more impact on clustering has been revealed by this study.

2.2.5 Feature reduction

Another problem face in genetic data analysing is the dimensionality, the number of features is very high while the number of data samples(observations) very low. In [19] to reduce the dimensionality of features they eliminated the genes with minimum variation across the samples. In [17] they reduce dimensionality in three steps, first remove genes which have missing value rate more than 10% across the gene. Then select genes with expression level differing by **l-fold** in at least **c** samples. 'l' and 'c' selected to save at least 10% of original features. Then discard genes which are 10% largest and smallest values according to the mean of gene expression across samples. In [20] study they have discretized the expression data to 0 and 1 using Gaussian distribution based on the non-cancer control samples only from the same tissue of origin. They have done pre-processing for genes with low expression variance in normal cells, i.e., standard deviation of expression smaller than 0.2, they used 3-fold

change to determine whether the genes were differentially expressed in tumor cells and also expression changes due to CNA were also masked; as such changes are not regulated by the cellular signalling system, but are due to genomic alterations. Then they have removed performed to remove genes with low Bernoulli variance because of their general lack of information and removed any genes that were highly correlated with a specific cancer type or tissue type, by removing all genes with a Pearson correlation coefficient, with respect to cancer or tissue type labels, greater than 0.85.

2.2.6 Transformation

Data transformation could be carried out to reduce the error and increase the efficiency of analysis. This can be broadly categorised into three main sections such as standardizing, scaling, ranking [21].

- **Standardizing** - Z score is an example for standardizing approach. In this approach transform features will have zero mean and unit variance, which means mean and variance is constant across all features. but this approach adversely affects by outliers.
- **Scaling** - In this type of transformation minimum and maximum value on the gene will be transformed to (0,1) respectively, if there are non-negative values and (-1,+1) respectively if there are negative values. This will also be adversely affected by outliers. Mean and variance is not constant across all features like in standardizing method.
- **Ranking** - This approach will transform the values of gene to ranking. This approach is more robust to outliers than the above two approaches.

2.2.7 Assessing cluster tendency

Clustering tendency can be used to check whether the clustering suitable for the dataset, two kinds of cluster tendency assessing technique discussed.

- **Cluster tendency assessing testing visual method (VAT)** - Compute the dissimilarity (DM) matrix between the objects in the data set using the Euclidean distance measure. Reorder the DM so that similar objects are close to one another. This process creates an ordered dissimilarity matrix (ODM). The ODM is displayed as an ordered dissimilarity image (ODI), which is the visual output of VAT.

- **Hopkins statistic** - used to assess the clustering tendency of a dataset by, measuring the probability that a given dataset is generated by a uniform data distribution. In other words, it tests the spatial randomness of the data.

Let D be a real dataset. The Hopkins statistic can be calculated as follow:

Sample uniformly n points (p_1, \dots, p_n) from D . For each point $p_i \in D$, find its nearest neighbour p_j ; then compute the distance between p_i and p_j and denote it as $x_i = \text{dist}(p_i, p_j)$. Generate a simulated dataset ($\text{random}D$) drawn from a random uniform distribution with n points (q_1, \dots, q_n) and the same variation as the original real dataset D . For each point $q_i \in \text{random}D$, find its nearest neighbour q_j in D ; then compute the distance between q_i and q_j and denote it $y_i = \text{dist}(q_i, q_j)$. Calculate the Hopkins statistic (H) as the mean nearest neighbour distance in the random dataset divided by the sum of the mean nearest neighbour distances in the real and across the simulated dataset.

The formula is defined as follow:

$$H = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i + \sum_{i=1}^n y_i}$$

Equation 1: Hopkins statistic

A value of H about 0.5 means that $\sum y_i$ and $\sum x_i$ are close to each other, and thus the data D is uniformly distributed. If the value of Hopkins statistic is close to zero, then we can reject the null hypothesis and conclude that the dataset D is significantly a clusterable data.

2.2.8 Identifying number of clusters

Identifying number of classes beforehand is a big challenge many researchers face in analysing genomic data using clustering. There are statistical methods to estimate the appropriate number of groups such as gap statistics [22]. In this study, they propose a method names ‘gap statistic’ for estimating the number of clusters in a dataset. The goal of this study is to provide a statistical procedure to formalize the heuristic that the location of an ‘elbow’ indicates the appropriate number of clusters. This method can be applied to virtually any clustering method. This has been implemented in cluster package of R.

2.2.9 Clustering

Although unsupervised method is the best way to address the main goal of this study which is to discover molecular subtypes in endometrial carcinoma, there are several challenges exist in this method. The main challenge which has been identified by [15], was that unsupervised learning algorithm tends to discover a dominant recurrent feature in the data, that may lead to susceptible to confounding factors as these are the features that explain most of the variability in the data. If datasets from two different studies have combined, such study, platform or batch effects can confound unsupervised analyses, hence these kinds of algorithms better to apply within a single homogenous data set.

Another challenge in many clustering algorithms used in analysing genomic data is that they divide the data into a predefined number of groups which has been discussed in the 2.2.8 Identifying number of clusters.

Partitioning clustering methods use pre-defined number of groups(k) and subdivide data into k number of classes. K-means, K-Medoids and CLARA algorithms are partitioning clustering methods. These clustering algorithms have the challenge of need pre-defined number of classes which is mentioned above in [15]. Agglomerative hierarchical clustering methods do not require pre-defined number of classes. The results of hierarchical clustering are a tree-based representation of data which is also called as dendrogram.

When doing the analysis there are two approaches to integrating data. Such as, Deep Integrative approach which combines many measurements of different levels such as mRNA, miRNA, transcription factor binding, this will let us understand how regulatory systems work together and broad integrative analyses over many distinct datasets, this leads to the discovery of general principles that are maintained across distinct conditions.[15].

In this study, it was targeted to conduct deep and broad integrative studies, but after the basic analysis done on datasets, it was revealed that all three studies done on endometrial cancers has been conducted on a single data set where the samples Ids were same. This research will be conducted only on deep integrative analysis approach. In this ‘Principles and methods of integrative genomic analyses in cancer’ [23] article, they have done a review of this methodology.

Many of the biological studies have used classical clustering methods such as hierarchical clustering. The reason for this issue is availability of implementation in many standard microarrays data analysing software and the ease of use. In [17] they have conducted a comparative study on 7 different clustering methods (hierarchical clustering with single, complete and average linkage, k-means, mixture of multivariate Gaussians, spectral clustering and a nearest neighbour-based method) and 4 proximity measures (Pearson's correlation coefficient, cosine, Spearman's correlation coefficient and Euclidean distance). According to their evaluation finite mixture of Gaussians and k-means, exhibited the best performance. But it is not that much reliable as this study had conducted their analysis on benchmark data set which consist of different cancer types, they evaluated their clustering considering the cancer types and subtype clusters. Most of the subtypes and types of cancers are changing according to the molecular level analysis later on.

The Clustering techniques used in the related works are Self Organizing Maps [6], Supercluster (an integrative clustering algorithm)[13], Hierarchical clustering[13], [17], [19], K-means [13], [17].

2.3 Related works

There are basically two challenges in cancer classification and cancer diagnosis using genomic data, which are high dimensionality of genomic data and availability of only a few hundred samples for a given tumor which leads to overfitting.

The study [6] demonstrate the feasibility of cancer classification based solely on gene expression data analysing and suggest a general way to discover and predict cancer classes for other types of cancers, independently to the previous biological knowledge, basically they evaluate class prediction and class discovery.

Class prediction-

Using neighbourhood analysis find genes whose expression pattern were strongly correlated with the class distinction to be predicted. Calculate the weighted votes of informative genes and find the winning class using predetermined threshold. Check the validity of the class predictors using cross-validation on the initial data set and assesses its accuracy on an independent set of samples

Class discovery -

Cluster tumors using self-organizing map. Predefining the number of classes in the data set when using SOM algorithm, can be assume as this was conducted in 1999 due to the lack of computational power. To check the putative classes are true structure create class predictor based on those classes and assume should perform well if the classes are true structures.

Earlier EC has been broadly classified into two groups as endometrioid and serous tumors, which was discussed in ‘2.2.2 Endometrial Cancer’. Proper subtype classification is crucial for selecting appropriate adjuvant therapy. According to “Integrated genomic characterization of endometrial carcinoma” [13] article, study has classified ECs into four categories: POLE ultra-mutated, microsatellite instability hypermutated, copy-number low, and copy-number high using integrated genomic, transcriptomic and proteomic characterization of biospecimens which were obtained from 373 patients. Integrated cross-platform analyses were performed using MEMo, iCluster and PARADIGM and also used their own integrative clustering algorithm called “supercluster” to derive overall subtypes based on sample cluster memberships across all the data type [13]. The studies [24] and [25] are two review papers which discuss the importance of applying TCGA molecular approach done on [13] to clinical practices.

In [21] study they have performed whole exome sequencing of 98 tumor biopsies with complex atypical hyperplasia, primary tumors and paired abdominopelvic metastases. Then analysed somatic mutations and copy number mutations of different biopsies of same individual to reconstruct phylogenetic relationships and identify putative cancer drivers across site of disease. But in this study, it was unable to identify examples of tumor self-seeding, which have been observed in human prostate cancers and breast cancer models and also couldn’t identify recurrent metastasis-specific driver mutations.

In [19] study used agglomerative hierarchical clustering to identify molecular level subtypes of breast cancers. Analysis was conducted in two phases after calculated the expression values. In first phase compare the expression values with the standard prognostic variables like tumor size, grade, menopausal status etc, then in the second phase done clustering. Eliminate genes with minimum variance across specimens to reduce curse of dimensionality. Then used hierarchical clustering with single linkage and average linkage with Euclidean distance and one minus Pearson correlation

distance as proximity measures. Out of these measures most distinctive clusters obtained by compact linkage with one minus Pearson correlation distance. Kaplan Meier estimation, cox-proportional hazard regression used as survival comparison among clusters identified.

2.4 Summary

Most of the biological studies have used classical clustering methods such as hierarchical clustering and K-means in clustering expression data. In [17] they have conducted a comparative study on 7 different clustering methods as hierarchical clustering with single, complete and average linkage, k-means, mixture of multivariate Gaussians, spectral clustering and a nearest neighbour-based method and 4 proximity measures as Pearson's correlation coefficient, cosine, Spearman's correlation coefficient and Euclidean distance. According to their evaluation finite mixture of Gaussians and k-means, exhibited the best performance. But it is not that much reliable as this study had conducted their analysis on benchmark data set which consist of different cancer types, they evaluated their clustering considering the cancer types and subtype clusters. Most of the subtypes and types of cancers are changing according to the molecular level analysis later on.

There are basically two challenges in cancer classification and cancer diagnosis using genomic data, which are high dimensionality of genomic data and availability of only a few hundred samples for a given tumor which leads to overfitting. This [6] study demonstrate the feasibility of cancer classification based solely on gene expression data analysing and suggest a general way to discover and predict cancer classes for other types of cancers, independently to the previous biological knowledge, basically they evaluate class prediction and class discovery.

Class prediction using neighbourhood analysis find genes whose expression pattern were strongly correlated with the class distinction to be predicted. Calculate the weighted votes of informative genes and find the winning class using predetermined threshold. Check the validity of the class predictors using cross-validation on the initial data set and assesses its accuracy on an independent set of samples. Class discovery by clustering tumors using self-organizing map. Predefining the number of classes in the data set when using SOM algorithm, can be assume as this was conducted in 1999 due

to the lack of computational power. To check the putative classes are true structure create class predictor based on those classes and assume should perform well if the classes are true structures.

Earlier EC has been broadly classified into two groups as endometrioid and serous tumors. Proper subtype classification is crucial for selecting appropriate adjuvant therapy. Reference [13] has classified endometrial cancers into four categories: POLE ultra-mutated, microsatellite instability hypermutated, copy-number low, and copy-number high using integrated genomic, transcriptomic and proteomic characterization of biospecimens which were obtained from 373 patients. The studies [24][25] are two review papers which discuss the importance of applying TCGA molecular approach done on [13] to clinical practices. According to the study conducted on “Hormones and endometrial carcinogenesis” [5] there is a knowledge gap in genetic level influence of UCEC although endometrium is a hormonal receptive organ.

In [21] study they have performed whole exome sequencing of 98 tumor biopsies with complex atypical hyperplasia, primary tumors and paired abdominopelvic metastases. Then analysed somatic mutations and copy number mutations of different biopsies of same individual to reconstruct phylogenetic relationships and identify putative cancer drivers across site of disease. But in this study, it was unable to identify examples of tumor self-seeding, which have been observed in human prostate cancers and breast cancer models and also couldn't identify recurrent metastasis-specific driver mutations.

In [19] study used agglomerative hierarchical clustering to identify molecular level subtypes of breast cancers. Analysis was conducted in two phases after calculated the expression values. In first phase compare the expression values with the standard prognostic variables like tumor size, grade, menopausal status etc, then in the second phase done clustering. Eliminate genes with minimum variance across specimens to reduce curse of dimensionality. Then used hierarchical clustering with single linkage and average linkage with Euclidean distance and one minus Pearson correlation distance as proximity measures. Out of these measures most distinctive clusters obtained by compact linkage with one minus Pearson correlation distance. Kaplan Meier estimation, cox-proportional hazard regression used as survival comparison among clusters identified.

Chapter 3 - Design

3.1 Introduction

This chapter describes the design and the methodology of the proposed solution to the research problem.

3.2 Research Methodology

This study has basically three main objectives, which are to find more precise subtypes in ECs and then find the genetic level hormonal influence in each of those sub types and find the stages specific gene expression levels. In order to achieve these goals, exploratory type of research was carried out. There are many studies which have been carried out to analyse cancer related data. They have applied different machine learning algorithms to identify subtypes by clustering samples and conduct gene enrichment analysis by clustering the genes. But according to [5] study, there is a gap in the hormonal influence of endometrial carcinoma and failed to treat late-stage endometrial carcinoma. In order to fulfil aforementioned goals, this study focus to analyse using a proper methodology. Following design describe the most appropriate method selected by analysing information from the literature review.

3.3 Design

When considering the research methodology, it can be divided into basically 4 phases as shown in the Figure 3.1, as:

- Data collecting
- Identifying EC subtype
- Identifying Hormonal influence
- Identifying stages specific gene expressions

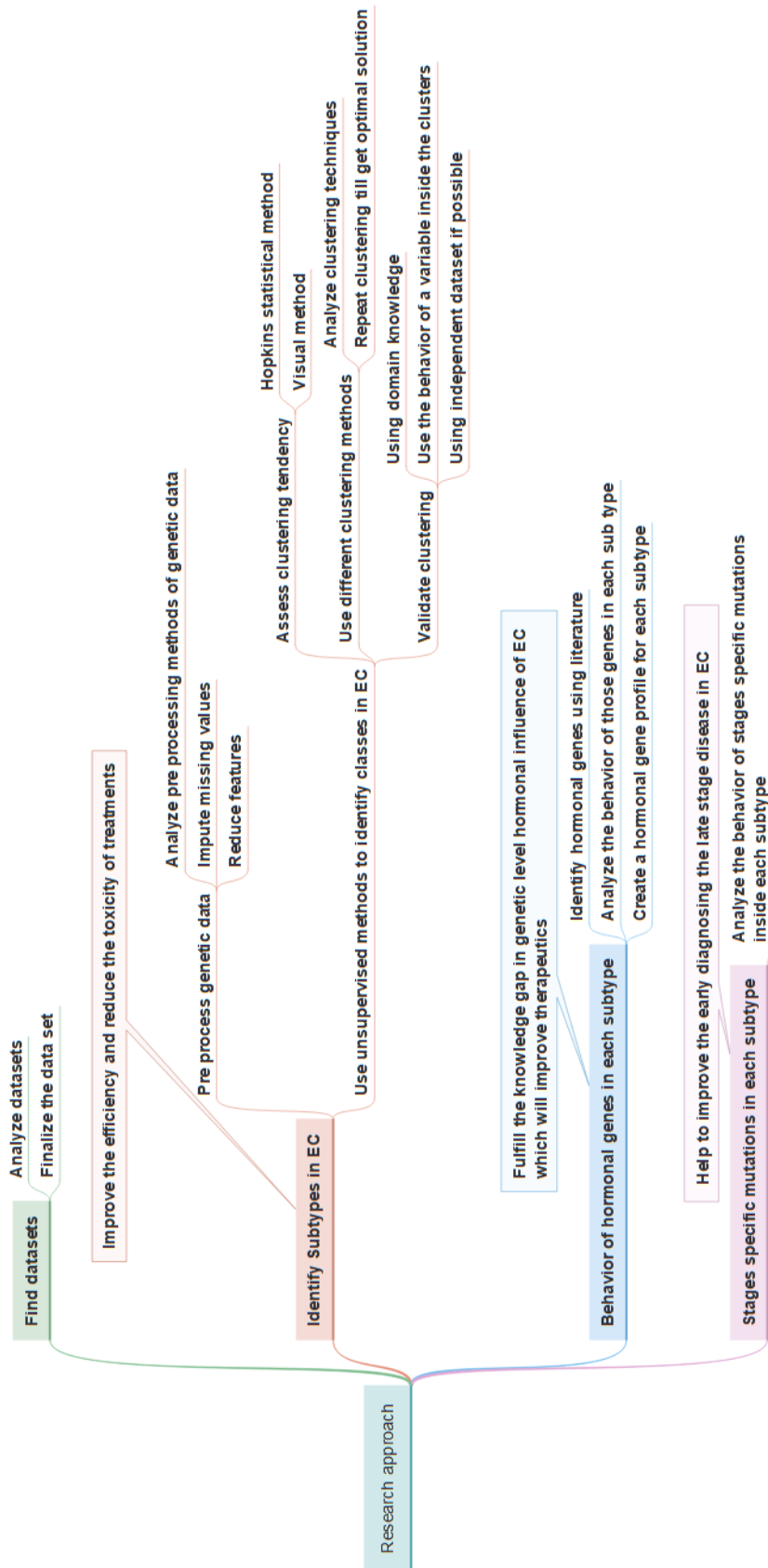


Figure 3.1: Mind map of Highlevel Research design

3.3.1 Collecting Data

Used cBioPortal to get the genetic data of EC. cBioPortal is an open platform for exploring multidimensional Cancer Genomics Data, integrated with the cancer genome atlas project (TCGA), all data go through Data Coordination Centre (DCC) and then reviewed by GDAC Broad which is available at cBioPortal. Currently provide access to more than 5000 tumor samples of 31 cancer studies[7].

Multidimensional genomic data types available in this data portal are Expression data, mutation data and copy number variation (CNV) data which has been described in ‘1.1 Background to the Research’ and ‘2.2.1 Biological Data & Bioinformatics’ sections.

Analyse datasets to check whether the same cancer samples repeated in the studies which will be used, and what kind of data types are available in each study. Outcomes of these basic analysis will be discussed in the ‘Chapter 5 – Results and Evaluation’ section. Fundamentally there are two different approaches [8] to conduct analysis of genomic data as:

- Deep - Integrative approach which combines many measurements of different levels such as mRNA, miRNA, transcription factor binding, this approach helps to understand how regulatory systems work together
- Broad - integrative analyses over many distinct datasets, this leads to the discovery of general principles that are maintained across distinct conditions.

This study couldn’t conduct broad integrative method in spite of the fact that the all three data sets containing almost same data samples which will be discussed in ‘Chapter 5’.

3.3.2 Identifying Endometrial cancer subtype

After selecting the datasets, EC subtype identifying phase is the next phase which will be the deepest part of this study. In this phase, analyse different unsupervised learning methods as the data set is unlabelled and the objective of unsupervised learning method is to discover hidden patterns within the data to find more precise subtypes.

In order to find similar samples(clusters) in EC, need to pre-process the data gathered from cBioPortal. In order to handle missing values, Missing value pattern was created using the VIM package [26] (Visualization and Imputation of Missing Values) and noted that there are some genes with 100% and 70% missing values across patients.

Due to this high amount of missingness across samples, those genes were eliminated from the dataset rather than imputing the missing value. After eliminating the genes with missing values there were 17,398 genes remained for analysis.

The dataset downloaded from cBioPortal was already z-score normalized using zero copy number variation samples (unaltered samples in a gene) across genes which remove the expression levels of unaltered samples.

As the dataset contained 17000+ variables(genes) and 500+ observations (samples) it created curse of dimensionality. In order to identify most meaningful subtypes of UCEC, 1000 genes were selected with a high coefficient of the variance. To preserve the biological value 170 hormonal genes were selected from NCBI (National Centre for Biotechnology Information) gene database and from that 25 hormonal genes that are related to UCEC were selected using literature [5], [21], [27]–[29].

Assess the cluster tendency of the datasets using Hopkins.

Optimal number of clusters was estimated using Elbow method and Silhouette methods which use internal measures such as within sum of square of clusters and silhouette width for K-means, Hierarchical and PAM clustering.

Hierarchical clustering and K-means clustering were used for cluster analysis as most of the background studies used them. As K-means clustering is sensitive to outliers, PAM was used to avoid that limitation. According to the literature, correlational distances which identify the clusters with the same profile regardless of the magnitude are better than Euclidean and Manhattan for expression data clustering. Spearman Correlation distance mitigates the effect of sensitiveness to outliers of Pearson Correlational distance. For further analysis we used Spearman correlation distance which measure the correlation and coefficient between the rank of two variables. Ward and Complete linkage measures give compact clusters rather than Single and Average linkage measures, as Ward linkage try to link two clusters having minimum within sum of square and Complete linkage try to link the clusters having the minimum of the maximum distance between clusters.

Basically, validation of the clusters carried out in three different approaches [30], such as internal validation, relative validation, and external validation.

A main requirement of good clusters is high similarity of objects in same cluster and high distinction between objects in different clusters which can be measured using internal measures like compactness and separation. Compactness and separation

measure the closeness of objects within clusters and how well the clusters separated from other clusters.

$$\text{Internal Validity} = \alpha (\text{separation} / \text{connectivity})$$

In this study internal validation was measured using Dunn index, Silhouette width, and connectivity. Dunn index is a ratio between smallest distance between objects in two clusters to the largest distance between objects in one cluster (intra cluster distance) which should be maximised. If a cluster is good then the diameter of the cluster should be small and distance between clusters should be large. Silhouette width measures the average distances between clusters which also should be larger if the observations are well clustered. Connectivity measures to what extent the items are placed in the same cluster as their nearest neighbours in the data space which should be minimized.

Relative validation refers to the analysis of different k values and different clustering measures. It is used in identifying optimal number of clusters.

External validation conducted visually using heatmaps[31] along with clinical variables.

3.3.3 Identifying Hormonal Influence

After choosing the meaningful clusters, in the next phase the behaviour of hormonal genes in each subtype will be observed, which will help to improve the therapeutics. First identify hormonal genes by referring literature. Then analyse the behaviour of those genes inside each subtype identified in the previous phase. Then create hormonal gene profile for each subtype which will be the summary of the behaviour of the hormonal genes in each subtype. This can be done by the BSS/WSS feature filtering method [32] by considering the discrimination power of the features among classes which will be further discussed in ‘Chapter 4 and 5’

3.3.4 Identifying Stage Specific Genes

The last phase which is identifying stage specifically expressed genes in each subtype of EC identified in phase 2 can be conducted parallelly with the previous phase (identifying behaviour of hormonal genes in each subtype). This can be done by the BSS/WSS[32] method mentioned above phase.

3.4 Tools used for analysing

Software and other tools used in this study will be discussed in this section

3.4.1 R, R studio, R markdown

For implementation used R tool which is open source, cross platform which runs on many operating systems. R is the most comprehensive statistical analysis package. It has many freely available bioinformatics libraries. As it is open source any person can use it and change it. New packages can be introduced easily.

And for the organizational purposes and to create reproducible research used R markdown(notebook) facility.

Use R studio as the integrated development environment (IDE) to run R programming language.

3.4.2 Panther

Panther used to analyse genes identified as feature for the clustering analysis and the genes identified as significant among clusters after clustering classes obtained (gene enrichment analysis). The PANTHER (Protein Analysis Through Evolutionary Relationships)[33] is part of Gene Ontology Phylogenetic Annotation Project. Version used was Version 14.0 (released 2018-12-03) contains 15512 protein families, divided into 104496 functionally distinct protein subfamilies. Downloads from this site may be Subject to the terms of the GNU General Public License Version 2, June 1991. 'Copyright © Paul Thomas' copyright statement should place on the top of each file.

3.4.3 NCBI

NCBI (National Center for Biotechnology Information) is a website developed by NCBI, contain series of databases which contain comprehensive biological information. In this study NCBI gene database has been used to validate and extract hormonal genes identified using literature.

3.5 Conclusion

This chapter gives an introduction to research design created using the information gained through the literature review, the tools used in this study with the ethical background of using them. On the foundation of research design, the dissertation can proceed with a detailed description of the research design implementation.

Chapter 4 - Implementation

4.1 Introduction

This chapter describes the implementation of the proposed solution to the research problem. Overall in detail research design and implementation was interpreted in Figure 4.2 below.

4.2 Collecting Data

After selected the Cancer type, Data set, Data types which will discuss further in 'Chapter – 5', importing data to the programming environment was different from data type to type.

Read the tab separated expression data file by selecting the particular file through browsing as shown in Figure 4.1. It makes the code more independent in creating reproducible research and creating package from the analysis.

```
14
15 #Import data from raw cbioportal data file
16 '{r}'
17 #.txt file: Read tab separated values
18 original_file <- read.delim(file.choose(),sep = '\t')
19 data_file <- as.data.frame(t(original_file[,c(-1,-2)]))
20
21 #take the unique nams of genes
22 gene_type_hugo <- original_file$COMMON
23 gene_type_entrez <- original_file$GENE_ID
24
25 #assign column names
26 colnames(data_file) <- gene_type_hugo
27
28 ...
```

Figure 4.1: Import and arrange data frame for analysis

Row data set contains Rows – genes and Columns – patients. In the raw data file, there were two gene naming conventions used as Hugo (Gene identifier according to HGNC) and Entrez (gene identifier according to NCBI), took both the names to variables and store for further analysis and create data frame with one naming convention to proceed.

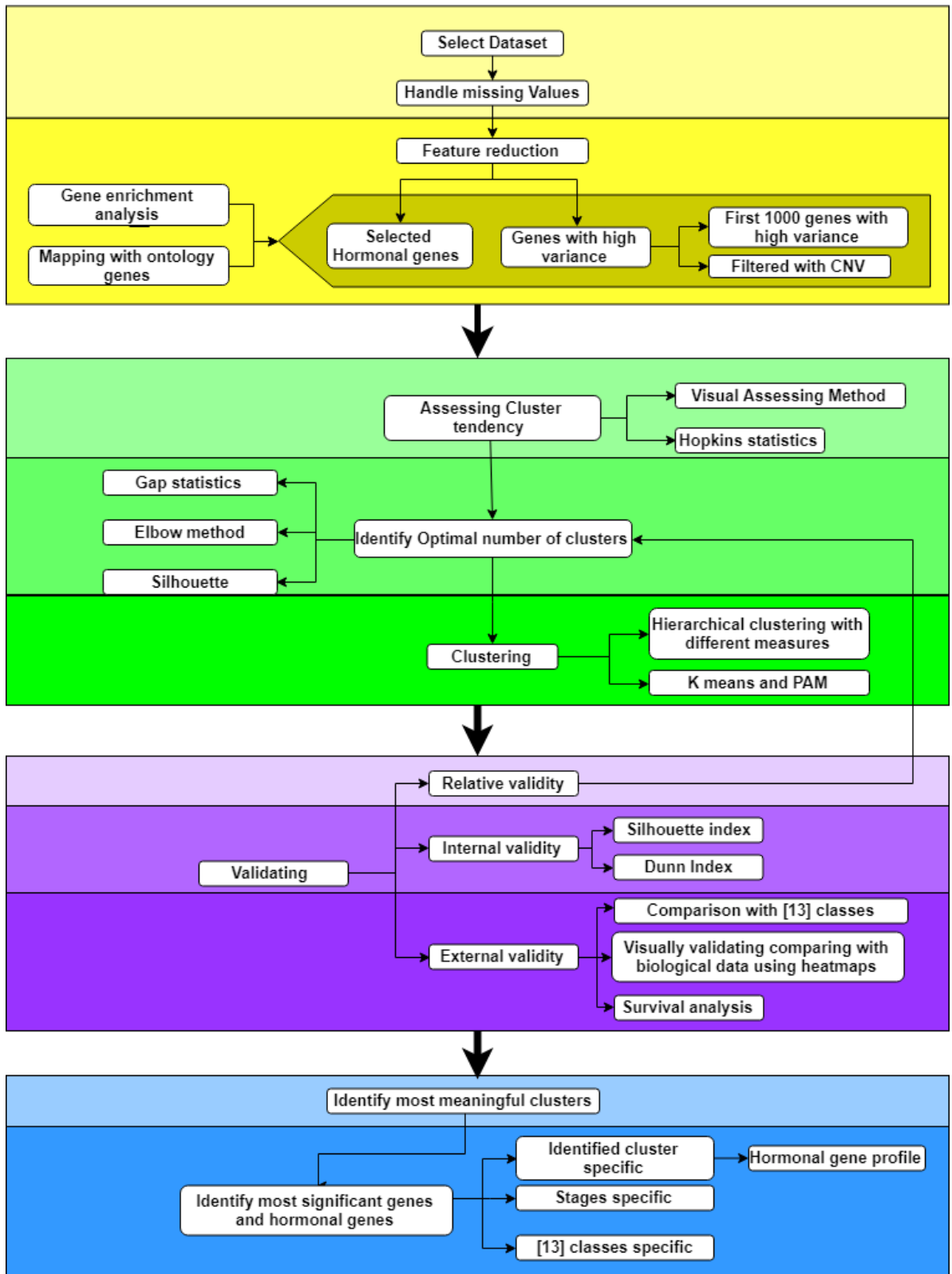


Figure 4.2: In detail Research Conceptual Model

4.3 Pre-process the data

4.3.1 Basic analysis and handling missing values

Basic analysis conducted to get an idea about the data going to be deal with, in here check the dimensionality, class of dataset, number of missing values and pattern of missing data as shown in Figure 4.3.

```
Basic analyse
Do this in order to take basic idea of data set
##find some graphical representation to visualize data

```{r,comment=""}
library(ggplot2)

class(data_file) #class of dataset
nrow(data_file) # number of rows
ncol(data_file) # number of columns
dim(data_file) # number of rows and columns
#summary(data_file) # descriptive statistics

#ggplot(data_file,aes(x=Samples,y=Genes, color=num)) + geom_point(size = 4) # scatterplot
#ggplot(temp_o, aes(Genes, fill=factor(num))) + geom_bar() # stacked histogr
...

Identify the missing values

```{r,comment=""}
library(mice)
library(VIM)
library(Hmisc)
temp_missing_patterns <- md.pattern(data_file) # (mice) displays all the missing values, NA for missing values

mice_plot <- aggr(data_file, col=c("green","red"),
  numbers=TRUE, sortvars=TRUE,
  labels=names(data_file), cex.axis=.7,
  gap=3, ylab=c("Missing data","Pattern")) # (VIM) display graphically missing values
```

Figure 4.3: Basic analysis done for expression data

But there was a challenge, in expression data there were many missing values, can't eliminate genes with missing values as the missingness can occur due to the experimental error and the particular gene may be significant gene in identifying subtype of EC and also it is hard to impute the perfect values as well. If the imputed value is not the perfect one it will lead to wrong output as in the cluster analysis, need to analyse the underline structure of the data. Another issue is to graphically view the patterns of missingness is really hard to interpret due to high dimensionality as shown in Figure 4.4 and Figure 4.5. Therefore, display the missing values using vim library as shown in the latter part of Figure 4.3, which interpret the missing patterns in a more understandable way as shown in Figure 4.6.

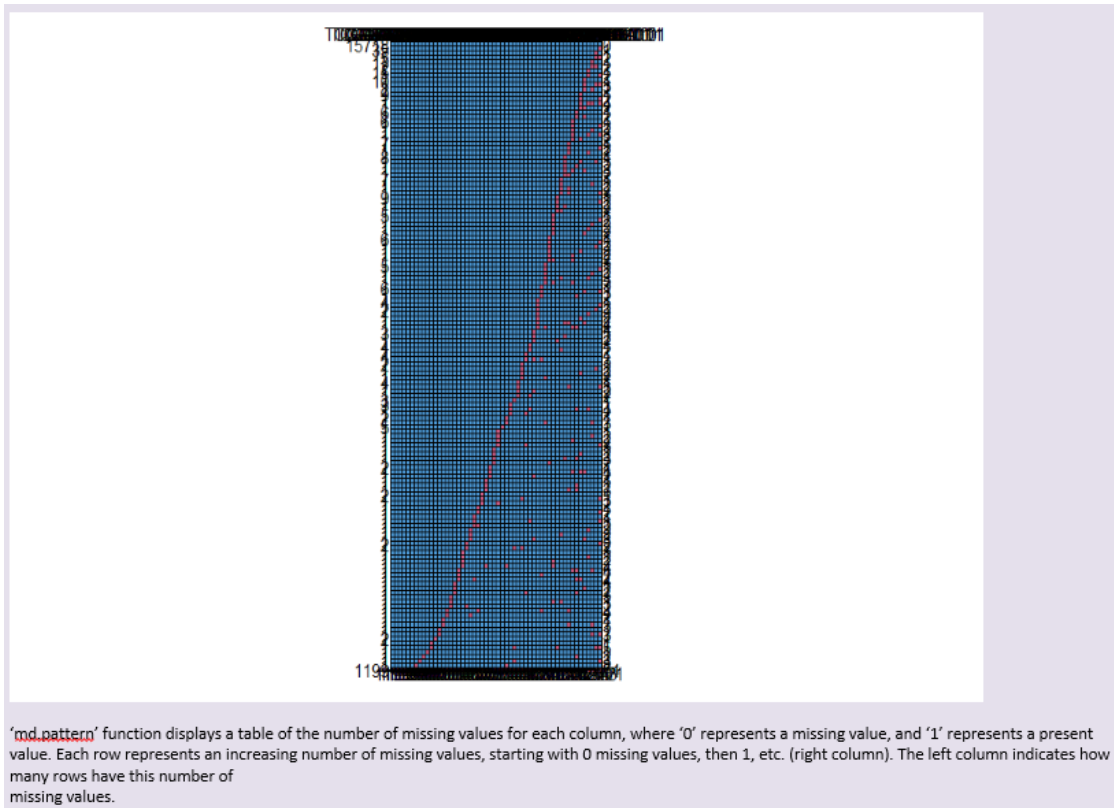


Figure 4.4: Hard to interpret missing data patterns due to high dimensionality

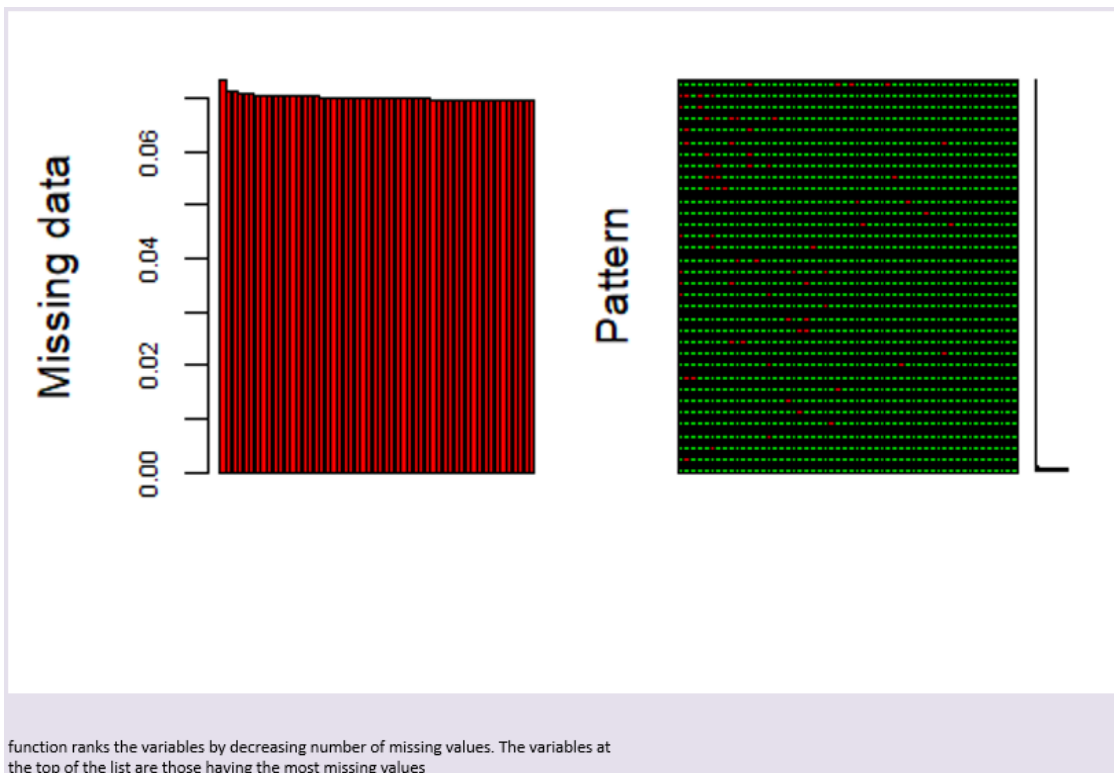


Figure 4.5: Hard to interpret missing data patterns due to high dimensionality

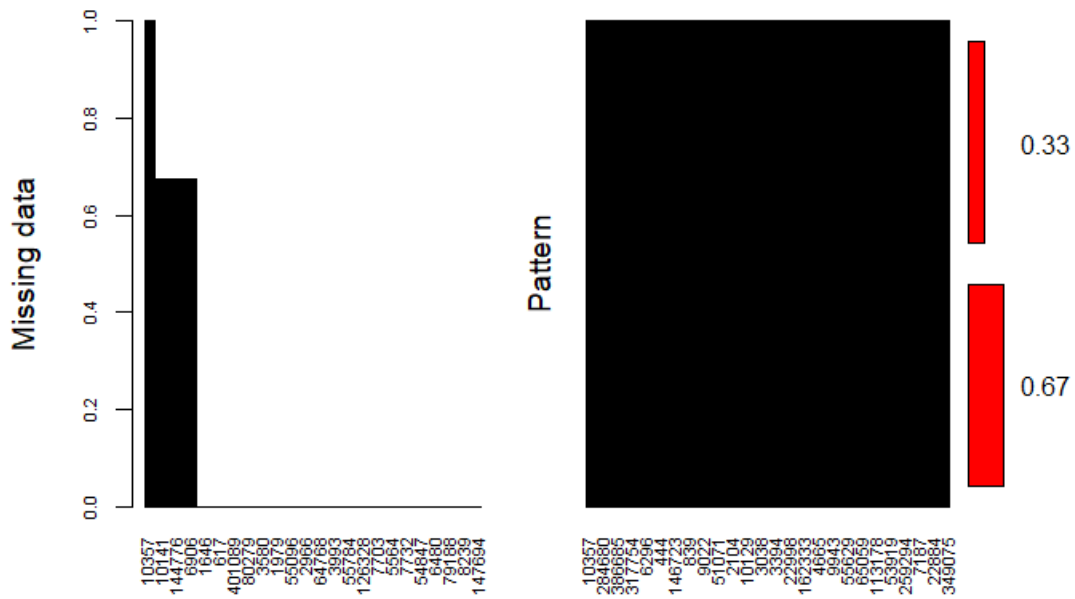


Figure 4.6: Missing patterns using VIM package

4.3.2 Feature extraction

Different features affect clusters differently. Some are important for clusters while others may hinder the clustering task. An efficient way of handling it is by selecting a subset of important, relevant and non-redundant features discovering similar or better cluster structures in the data than those obtained by using the whole set of features.

It helps in finding clusters efficiently understanding the data better and reducing data size for efficient storage collection and processing. As clustering is done on unsupervised data without class information traditional feature selection algorithms for classification do not work as there is no information (class labels) that can guide the search for relevant features. In terms of classification accuracy and runtime the best option for datasets with 6000 or less features are univariate spectral feature selection method (USFSM), while for datasets with more than 6000 features Variance is the best option [34].

As discussed in the ‘Chapter 3’ to preserve biological information and to reduce the dimensionality this study used 2 feature set as 1000 Genes with high variance among classes as they contribute to discriminate the classes and second feature set was hormonal genes selected using literature and NCBI gene data base. Details of the selected hormonal genes will be shown in the ‘Appendix B’. First identify the 170+

hormonal genes from NCBI gene database and extract 30 genes from that using the literature related to hormone encoding genes and hormone receptive genes.

4.4 Clustering

In his section implementation of the clustering analysis will be discussed.

4.4.1 Assessing Cluster Tendency

First assess feasibility of the clustering as clustering methods sometimes return clusters even the data doesn't contain clusters. To evaluate the cluster tendency of the data sets used visual method where visualise the ordered dissimilarity matrix. In this study it assesses dissimilarity matrices of Euclidean, Manhattan, spearman correlation and Pearson correlation. Hopkins statistical method has been also used to test the spatial randomness of the data.

4.4.2 Identifying Optimal number of K

For identifying optimal number of clusters used visualizing the dendrograms created using the hierarchical clustering which doesn't required predefined number of clusters as partition clustering and direct methods optimizing a measure of perfect clustering such as average silhouette measure (Silhouette method) and within cluster sum of square (Elbow method). As the methods mention above are subjective to the distance measures used, and the criteria used to partition, analyse different distance measures and algorithms in terms of identifying the optimal number of k. Calculate optimal number of k for Hierarchical clustering, K-means and PAM.

4.4.2 Clustering

In order to identify the clusters in expression data used K-means, PAM and Hierarchical clustering with different distance measures and linkage measures. Through comparing the algorithms, performance and different limitations which will be discussed in the next Chapter, the study proceeds further with hierarchical clustering.

4.5 Validating

Basically, validation of the clusters carried out in three different approaches, such as internal validation, relative validation and external validation.

4.5.1 Internal Validation

This measures the goodness of the clusters using the internal measures such as compactness, connectivity and separation. Generally, this used separation/connectivity. In this study to internally validate the clusters used 3 indexes as Dunn index, Silhouette width and connectivity.

4.5.2 Relative Validation

Relative validation refers to analysing different k values and different clustering measures which has been already done in this analysis.

4.5.3 External Validation

This validation phase refers to the validation conducted using the external information such as clinical variables, a ground truth of the cancers or the existing classes. To conduct external validation in this study used two separate approaches as visual method and statistical method. For the visual method use heatmaps[31] to interpret the relationship between the classes identified and the clinical variables and behaviour of the gene expressions.

Second approach was to conduct a comparison between the existing subtypes identified in the [13] study and subtypes identified in this study using corrected rand index.

4.6 Analysing

To identify the discriminative genes (significant genes) among stages and identified subtype used bss/wss function which rank the gene with their discriminative power.

```

...{r}
bssWssFast <- function (X, givenClassArr, numClass=2)
# between squares / within square feature selection
{
  classVec <- matrix(0, numClass, length(givenClassArr))
  for (k in 1:numClass) {
    temp <- rep(0, length(givenClassArr))
    temp[givenClassArr == (k - 1)] <- 1
    classVec[k, ] <- temp
  }
  classMeanArr <- rep(0, numClass)
  ratio <- rep(0, ncol(X))
  for (j in 1:ncol(X)) {
    overallMean <- sum(X[, j]) / length(X[, j])
    for (k in 1:numClass) {
      classMeanArr[k] <-
        sum(classVec[k, ] * X[, j]) / sum(classVec[k, ])
    }
    classMeanVec <- classMeanArr[givenClassArr + 1]
    bss <- sum((classMeanVec - overallMean)^2)
    wss <- sum((X[, j] - classMeanVec)^2)
    ratio[j] <- bss/wss
  }
  sort(ratio, decreasing = TRUE, index = TRUE)
}
...

```

Figure 4.7: BSS/WSS function

Chapter 5 - Results and Evaluation

5.1 Data collecting and pre-processing

5.1.1 Selecting data set

As discussed in above chapters data for this study collected through cBioPortal. There are two types of uterus cancers, Endometrial carcinoma, and Carcinosarcoma. Basic data analysis had been conducted on the datasets retrieved on the aforementioned uterus cancers which are illustrated in the Table 5.1 below. ECs are more frequently occur (80%) than the Carcinosarcoma (2%-4%).

Table 5.1: Basic data analysis on the uterus cancer dataset

Cancer study Name	Uterine Carcinosarcoma (Johns Hopkins University, Nat Commun 2014)	Uterine Carcinosarcoma (TCGA, PanCancer Atlas)	Uterine Carcinosarcoma (TCGA, Provisional)	Uterine Corpus Endometrial Carcinoma (TCGA, Nature 2013)	Uterine Corpus Endometrial Carcinoma (TCGA, PanCancer Atlas)	Uterine Corpus Endometrial Carcinoma (TCGA, Provisional)
Samples	22	57	57	373	529	548
Patients	22	57	57	373	529	548
Number of Genes	19150	20532	17814	17242	20472	17814
Sequenced	22	57	57	248	517	248
CNA	0	56	56	363	523	539
RNA Seq	0	0	57	333	0	177
Uterine Endometrioid Carcinoma	22	57	57	307	399	410
Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma	-	-	-	53	109	115
Endometrial Carcinoma (mixed)	-	-	-	13	21	24
data_clinical_patient	Y	Y	Y	Y	Y	Y
data_clinical_sample	Y	Y	Y	Y	Y	Y
data_CNA	-	Y	Y	Y	Y	Y
data_expression_median	-			Y		Y
data_gistic_genes_amp	-		Y	Y		Y
data_gistic_genes_del	-		Y	Y		Y
data_linear_CNA	-		Y	Y		Y
data_methylation_hm27	-			Y		Y
data_methylation_hm450	-		Y			
data_mRNA_median_Zscores	-			Y		Y
data_mutations_extended	Y		Y	Y		Y
data_mutations_mskcc	-		Y			Y
data_mutsig	-			Y		Y
data_RNA_Seq_v2_expression_median	-		Y	Y		Y
data_RNA_Seq_v2_mRNA_median_Zscores	-		Y	Y		Y
data_rppa	-		Y	Y		Y
data_rppa_Zscores	-		Y			Y
data_subtypes	-			Y		
			628 (1995-2013)		549 (1995-2013)	

According to the Table 5.1, available Carcinosarcoma data set is relatively smaller than the Endometrial carcinoma dataset, and also the knowledge gap in hormonal influence in EC identified from [5] were related to EC. Cancer type which will be analysed in this study was selected as Endometrial Carcinoma.

Then conducted a basic analysis of Endometrial carcinoma datasets, there were 3 studies conducted on EC:

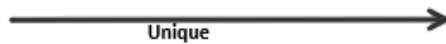
- Uterine Corpus Endometrial Carcinoma (TCGA, Nature 2013)
- Uterine Corpus Endometrial Carcinoma (TCGA, PanCancer Atlas 2018)
- Uterine Corpus Endometrial Carcinoma (TCGA, Provisional)

Although there were 1400+ samples in those three studies separately. After the basic analysis, it was revealed that those three studies were conducted on the same dataset, as shown in the Table 5.2 below.

Table 5.2 Basic data analysis on the Endometrial cancer dataset

Unique with->	TCGA	PAN	PUB	Total
TCGA	548	9	175	548
PAN	0	539	223	539
PUB	0	13	373	373

There were only 548 unique samples available



It was concluded to use the Uterine Corpus Endometrial Carcinoma (TCGA, PanCancer Atlas 2018)[9] dataset for the further analysis as it has been updated recently and contain 500+ samples.

Then conducted a basic analysis of the Uterine Corpus Endometrial Carcinoma (TCGA, PanCancer Atlas 2018) dataset as shown in the Table 5.3 below. This dataset contains three types of genomic data. Such as CAN data, Expression data, Mutations.

Table 5.3 Basic data analysis on Uterine Corpus Endometrial Carcinoma (TCGA, PanCancer Atlas 2018)

Cancer study ID	uces_tcga_pan_can_atlas_2018
Cancer study Name	Uterine Corpus Endometrial Carcinoma (TCGA, PanCancer Atlas)
Samples	529
Patients	529
Number of Genes	20472
Sequenced	517
CNA	523
RNA Seq	0
Uterine Endometrioid Carcinoma	399
Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma	109
Endometrial Carcinoma (mixed)	21
data_clinical	Y
data_CNA	Y
data_log2CNA	Y
data_expression	Y
data_expression_Zscores	Y
data_expression_merged	Y
data_expression_merged_Zscores	Y
data_fusions	Y
data_mutations_extended	Y
	529 (1998-2009)

The basic discovery carried out on the data files available in the Uterine Corpus Endometrial Carcinoma (TCGA, PanCancer Atlas 2018) dataset is shown in the Table 5.4 below:

Table 5.4: Basic exploration on the datafiles available in Uterine Corpus Endometrial Carcinoma (TCGA, PanCancer Atlas 2018) dataset

data_filename	genetic_alteration_type	datatype	profile_description
data_CNA	COPY_NUMBER_ALTERATION	DISCRETE	Putative copy-number from GISTIC 2.0. Values: -2 = homozygous deletion; -1 = hemizygous deletion; 0 = neutral / no change; 1 = gain; 2 = high level amplification.
data_log2CNA	COPY_NUMBER_ALTERATION	LOG-VALUE	Log2 copy-number values for each gene (from Affymetrix SNP6).
data_expression	MRNA_EXPRESSION	CONTINUOUS	mRNA Expression from Illumina HiSeq_RNASeqV2: https://www.synapse.org/#!Synapse:syn4874822.6 .
data_expression_Zscores	MRNA_EXPRESSION	Z-SCORE	mRNA z-Scores (U133 microarray only) compared to the expression distribution of each gene tumors that are diploid for this gene.
data_expression_merged	MRNA_EXPRESSION	CONTINUOUS	mRNA Expression from Illumina HiSeq Batch Normalized: https://www.synapse.org/#!Synapse:syn4976369 (mRNA Expression Batch Normalized/Merged from Illumina HiSeq_RNASeqV2 syn4976369)
data_expression_merged_Zscores	MRNA_EXPRESSION	Z-SCORE	mRNA Expression Zscores from Illumina HiSeq Batch Normalized: https://www.synapse.org/#!Synapse:syn4976369 (mRNA Expression Zscores Batch Normalized/Merged from Illumina HiSeq_RNASeqV2 syn4976369)
data_fusions	FUSION	FUSION	Fusions
data_mutations_extended	MUTATION_EXTENDED	MAF	Mutation data from exome sequencing data that were compiled using handpicked individual MAFs taken from the individual tumor type AWG lists.

5.1.2 Handling Missing values

After gain understanding about the data types and studies through these evaluations did basic analysis about the data types inside the dataset selected (Expression data of Pan cancer dataset) and understand the missing value patterns in order to impute missing values as shown in Figure 5.1. Graph in the left side shows the missing patterns of variables(genes), such as there are few genes with all the expression values missing and a few with approximately 70% missing values. In the right-side graph shows missing patterns in observations (samples), such as there were two types of groups in samples as, 33% of samples and 67% of samples according to the missing genes discussed earlier.

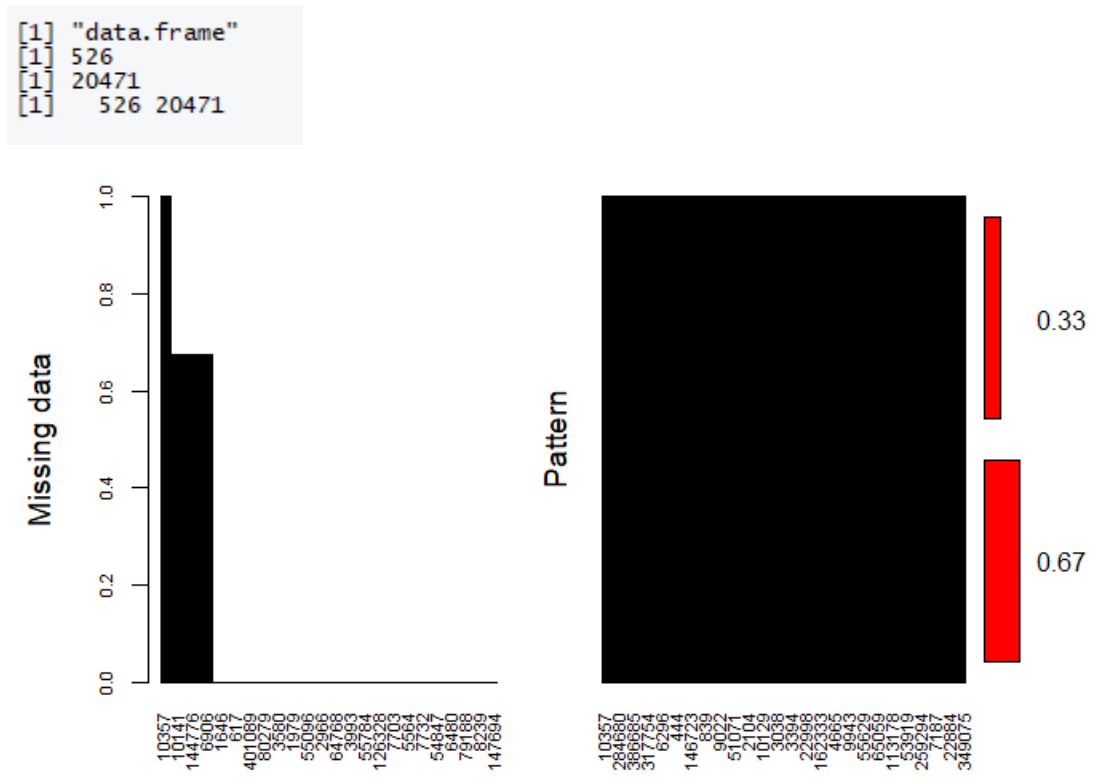


Figure 5.1 Missing value pattern of expression data

As the limitations in missing value imputations discussed in the previous chapters it is hard to impute the missing values under this situation. Imputation can affect the underlying patterns of data as the missingness of variables(genes) were 100% and 70% which are very high values. Although discussed many missing value imputation algorithms in literature review, concluded to eliminate the missing variables considering the related works [19] having more than 60% of missingness which will be

shown in Figure 5.2 below without any missing values. After removing the missing genes there were 17398 genes.

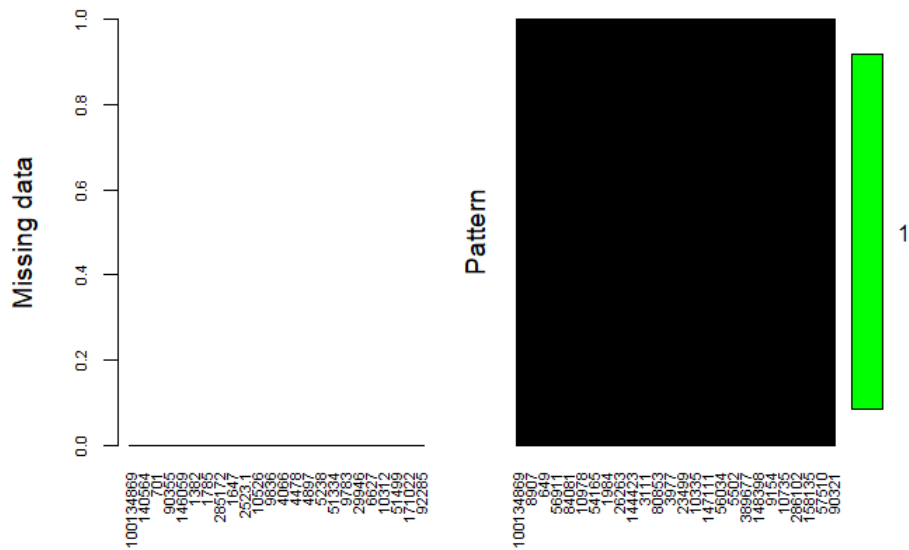


Figure 5.2: Eliminate the variables having high missingness

5.1.3 Extracting Features

After handled the missing values next identify the most important features out of the data set as this contains 20000+ variables(genes) and 500+ observations(samples) which create dimensionality curse.

As this study has to use unsupervised feature selection techniques due to unavailability of class labels, feature selection algorithms for classification do not work and has to use unsupervised feature selection methods such as variance, laplacian score, univariate spectral feature selection method, unsupervised discriminative feature selection method, SVD-entropy and etc. In terms of classification accuracy and runtime the best option for datasets with more than 6000 features is Variance [34].

First 1000 genes with high variance, selected considering coefficient of the variance Second feature set was hormonal genes selected using literature [5], [21], [27]–[29] and NCBI gene data base. More details about selected hormonal genes will be shown in the in ‘Appendix B’.

5.2 Identifying Endometrial Cancer Subtypes

5.2.1 Analysing Cluster Tendency

Before applying the clustering, techniques assess the feasibility of clustering in the data set using visual assessing technique and statistical techniques as discussed in ‘Chapter – 2 and Chapter - 4’. Clustering tendency was analysed for the both feature sets and for the visual assessing method used magnitude based and correlational based distance measures to create the ordered dissimilarity matrix.

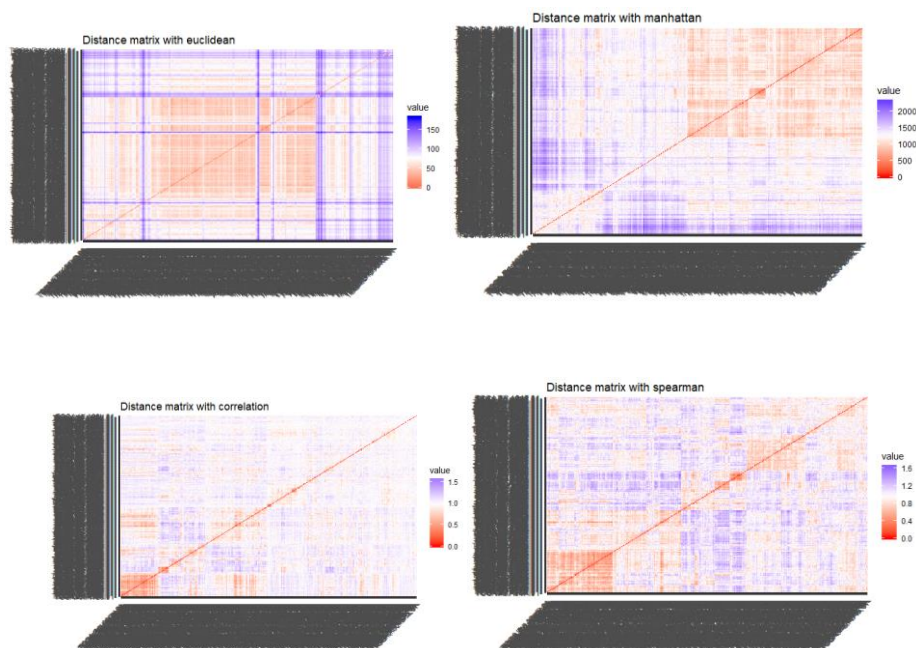


Figure 5.3: Ordered dissimilarity matrix of genes selected with high variance

According to Hopkins statistics Feature set with genes having high variance show 0.0985 value and Feature set with hormonal genes shows 0.2431 value, which will reject the null hypothesis and conclude that in both scenarios data set is significantly cluster able.

5.2.2 Identifying optimal number of K

As mentioned in ‘4.4.2 Identifying optimal number of K’ conduct a comparative analysis of identifying the optimal number of K as methods of identifying the optimal number of clusters are subjective to the distance measures and partitioning criteria.

Calculate the optimal number of k for Hierarchical clustering, Kmeans, PAM and use the numbers obtained for the further analysis. Results are shown in the Figure 5.4 and Figure 5.5 below.

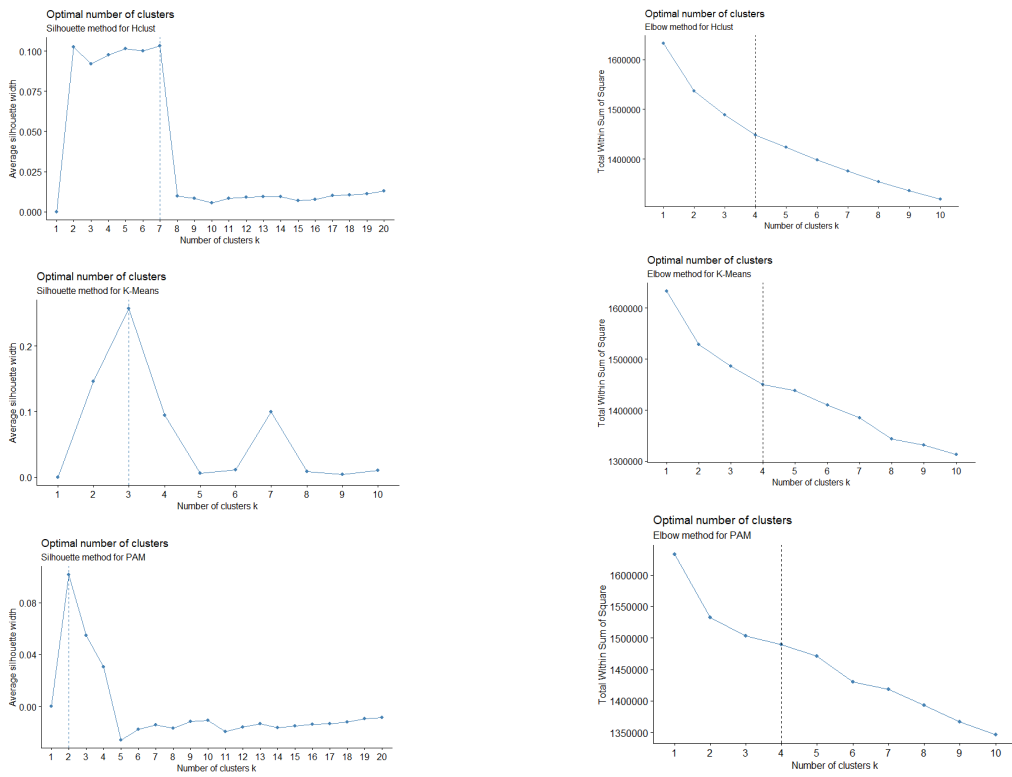
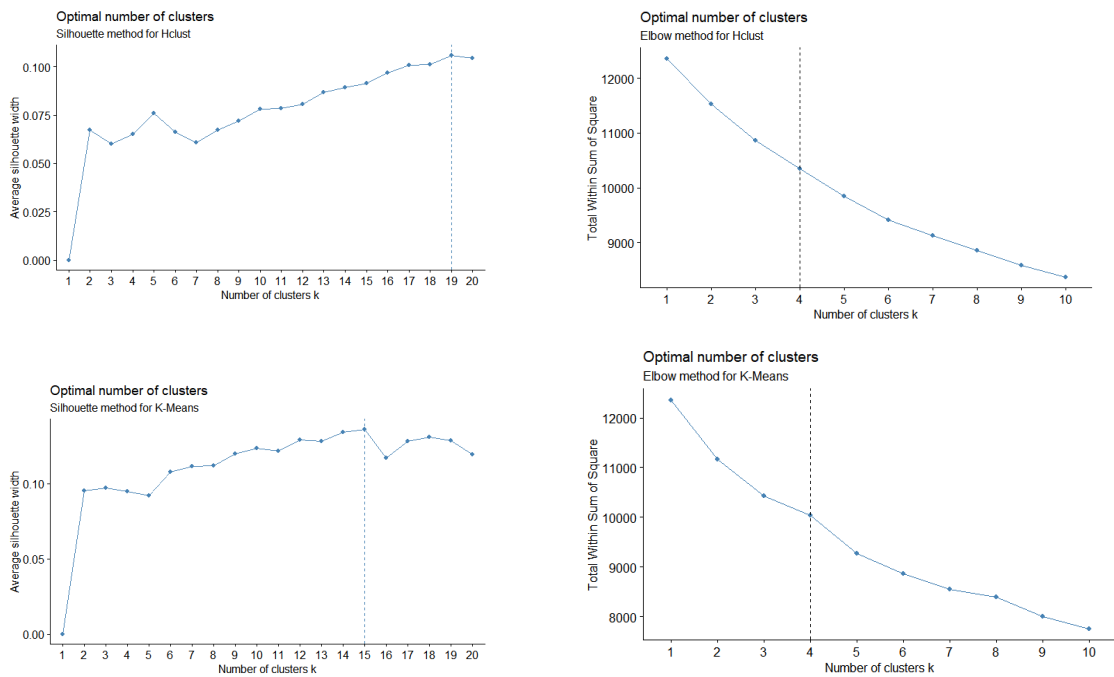


Figure 5.4: Optimal number of clusters genes having high variance



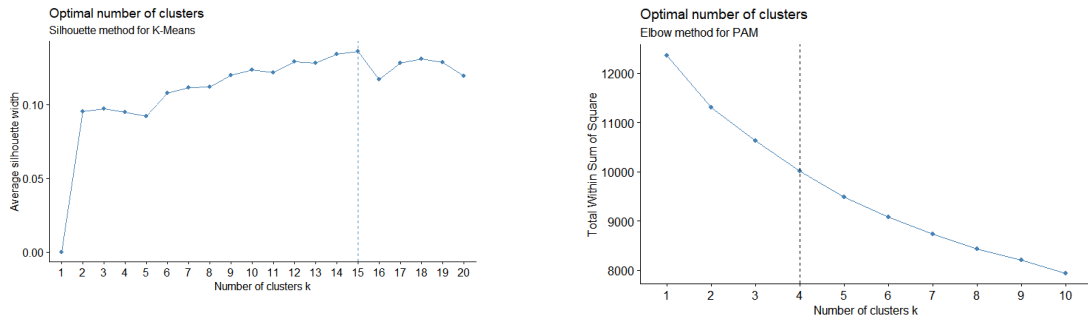


Figure 5.5: Optimal number of clusters hormonal genes

For genes with high variance used k as 3,4,7,11 and for hormonal gene set 2,3,4,15 considering the algorithms and visualizing the dendrograms.

5.2.3 Clustering

As mentioned in the previous chapter to identify clusters in samples of EC, used unsupervised learning techniques as the data is unlabelled. Considering the literature review as many cancer related researches has conducted on hierarchical and k-means clustering algorithms, used both algorithms for comparison. K means algorithm was sensitive to outliers. To overcome that outlier sensitivity, try PAM clustering algorithm. But as shown in the Figure 5.7, PAM also have the weakness of sensitive to the outliers same as kmeans and according to the internal validity which will be discussed in the next section PAM shows poor internal validity and stability compared to hierarchical clustering. Considering those limitations and as hierarchical clustering is more informative at the same time can be customised using different distance measures and linkage measures, this study was proceeding with hierarchical clustering methods.

According to the Figure 5.7 it can be seen that internal validity measure of average silhouette width won't give optimal validation as it measures global clustering characteristics only.

K-means

	cluster <fctr>	size <int>	ave.sil.width <dbl>
1	1	526	0.89
2	2	1	0.00

2 rows

	cluster <fctr>	size <int>	ave.sil.width <dbl>
1	1	505	0.57
2	2	21	0.29
3	3	1	0.00

3 rows

	cluster <fctr>	size <int>	ave.sil.width <dbl>
1	1	4	-0.11
2	2	1	0.00
3	3	21	0.29
4	4	501	0.58

Hclust

	cluster <fctr>	size <int>	ave.sil.width <dbl>
1	1	321	0.19
2	2	206	0.22

2 rows

3

	cluster <fctr>	size <int>	ave.sil.width <dbl>
1	1	217	0.22
2	2	206	0.15
3	3	104	0.09

3 rows

	cluster <fctr>	size <int>	ave.sil.width <dbl>
1	1	217	0.20
2	2	98	0.10
3	3	104	0.06
4	4	108	0.09

PAM			
	cluster <fctr>	size <int>	ave.sil.width <dbl>
1	1	526	0.89
2	2	1	0.00
2 rows			
	cluster <fctr>	size <int>	ave.sil.width <dbl>
1	1	515	0.65
2	2	11	0.49
3	3	1	0.00
3 rows			
	cluster <fctr>	size <int>	ave.sil.width <dbl>
1	1	514	0.65
2	2	11	0.49
3	3	1	0.00
4	4	1	0.00

Figure 5.6: Comparison between Hierarchical, K-means and PAM

According to the literature review [35] correlational distance are more suitable for clustering gene expression data than magnitude based distances. Basically, most popular correlational distances are Spearman correlation which measure the correlation coefficient of rank of two variables and Pearson correlation measures the degree of linear relationship. Spearman correlation was used for further analysis as Pearson distance is more sensitive to outliers. According to visual assessing technique used Euclidian distance for comparison and further evaluation purposes. Cluster the both datasets with Spearman distance measure and complete and Ward linkage measures. Results are shown in the Figure 5.7 below. To visualise how classes, scatter around use first two principle components to display multidimensional data set.

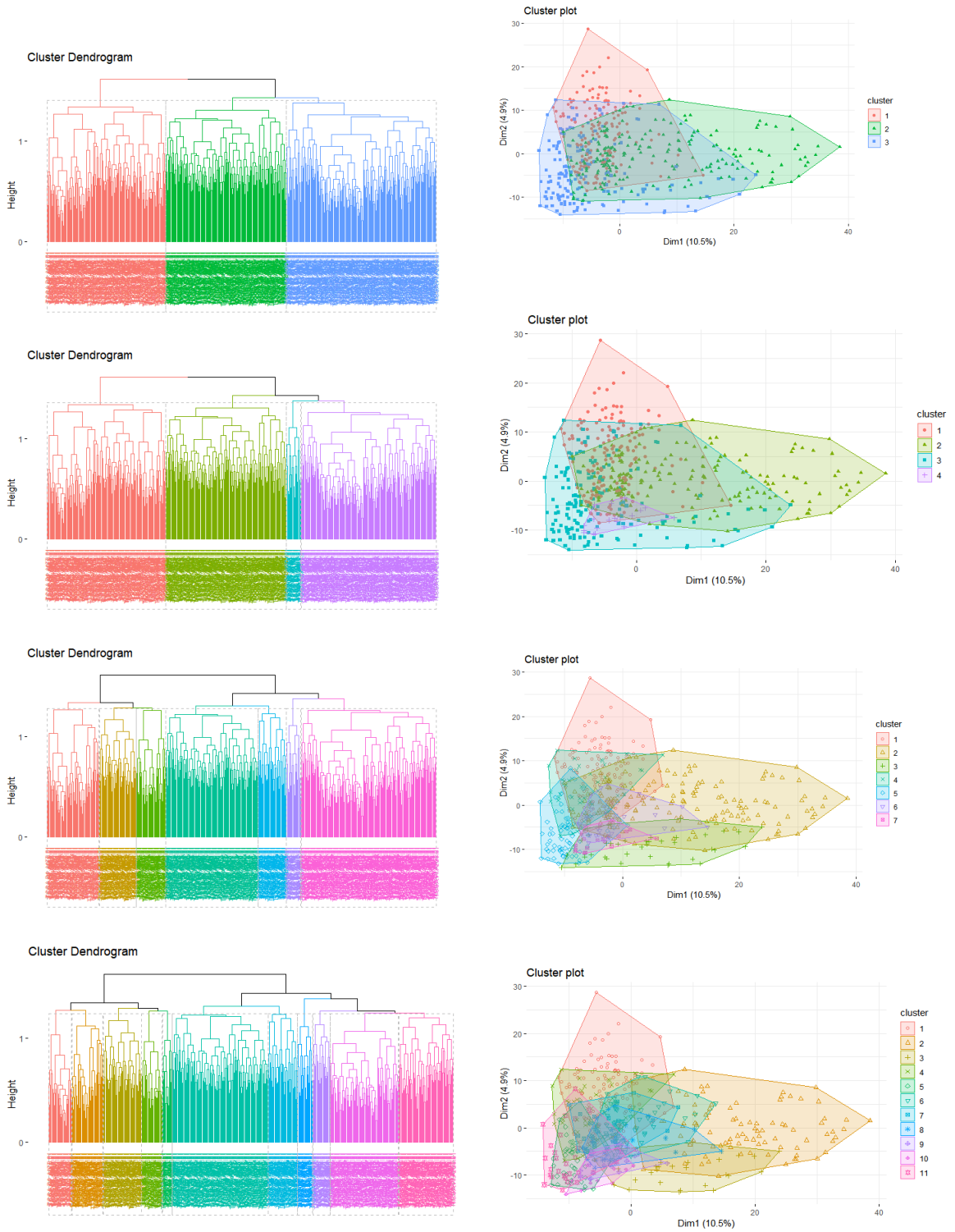


Figure 5.7: Spearman distance and complete linkage hierarchical clustering applied

5.2.4 Validating

Up to this stage clusters couldn't be eliminated as the basic rule of clustering is to identify meaningful clusters, not just clusters. As such three phases of validation carried out.

5.2.4.1 Internal validation

As discussed in the previous chapter internal validation measures the goodness of clusters using internal factors like connectivity and separation without considering external factors. To internally validate the clusters identified this study, used Dunn index, silhouette width and connectivity indexes as shown in Figure 5.8 and Figure 5.9 below.

Clustering Methods:		Validation Measures:										
hierarchical pam kmeans		2	3	4	5	6	7	8	9	10	11	
Cluster sizes:		2	3	4	5	6	7	8	9	10	11	
hierarchical	Connectivity	210.9413	232.7532	379.2917	416.6048	510.2706	552.5421	572.6143	579.4968	586.1230	608.0603	
	Dunn	0.2754	0.2797	0.2484	0.2535	0.2604	0.2614	0.2652	0.2691	0.2694	0.2756	
	Silhouette	0.0750	0.0548	0.0256	0.0214	0.0247	0.0291	0.0389	0.0388	0.0360	0.0332	
pam	Connectivity	260.5567	360.4246	462.8536	487.8417	552.3813	554.7361	579.4115	592.9794	581.8333	600.6516	
	Dunn	0.1582	0.1351	0.1390	0.1439	0.1390	0.1390	0.1390	0.1479	0.1479	0.0516	
	Silhouette	0.0893	0.0786	0.0738	0.0765	0.0749	0.0778	0.0779	0.0786	0.0833	0.0847	
kmeans	Connectivity	166.3290	261.7321	361.2202	391.9032	428.2913	457.0389	498.8333	488.8290	569.7639	587.4496	
	Dunn	0.0407	0.0407	0.0459	0.0479	0.0479	0.0499	0.0493	0.1430	0.1404	0.1477	
	Silhouette	0.0655	0.0558	0.0756	0.0785	0.0769	0.0831	0.0799	0.0812	0.0475	0.0582	

	Score	Method	Clusters
	<dbl>	<fctr>	<fctr>
Connectivity	166.3290	kmeans	2
Dunn	0.2797	hierarchical	3
Silhouette	0.0893	pam	2

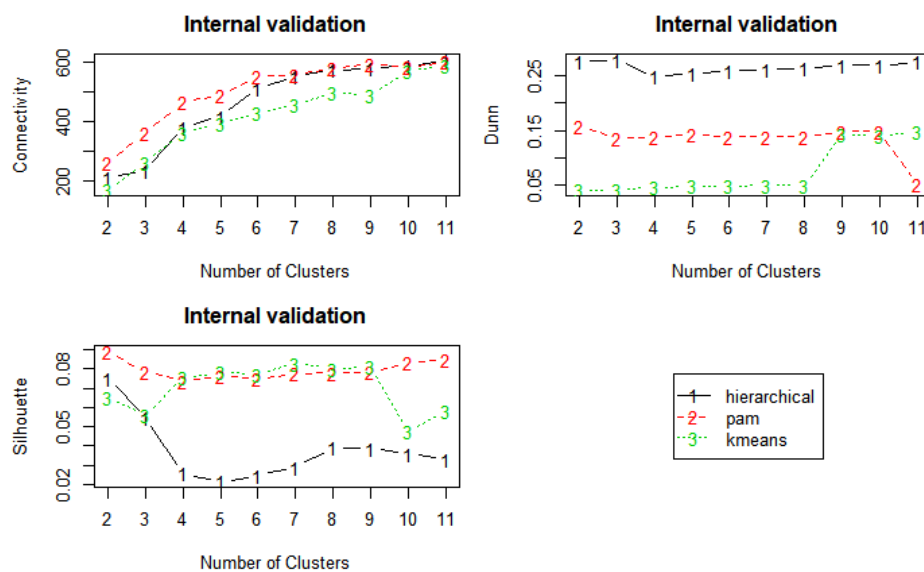


Figure 5.8: Internal validity of correlational distance of PAM, Kmeans, Hierarchical clustering

Clustering Methods:		Validation Measures:									
hierarchical pam kmeans		2	3	4	5	6	7	8	9	10	11
Cluster sizes:		2 3 4 5 6 7 8 9 10 11									
hierarchical	Connectivity	4.8579	7.7869	20.8341	25.6210	25.6210	28.5500	32.4079	32.4079	40.9810	43.9099
	Dunn	0.3972	0.4022	0.4034	0.4053	0.4121	0.4215	0.4295	0.4396	0.4441	0.4534
	Silhouette	0.3584	0.3590	0.3582	0.3568	0.3544	0.3551	0.3479	0.3464	0.3378	0.3383
pam	Connectivity	93.7433	381.1917	493.3671	560.4726	555.1583	556.3897	554.1460	554.5429	550.8175	596.5310
	Dunn	0.2079	0.1633	0.1398	0.1420	0.1420	0.1536	0.1536	0.1536	0.1536	0.1536
	Silhouette	0.1015	0.0545	0.0304	-0.0257	-0.0176	-0.0142	-0.0169	-0.0114	-0.0108	-0.0192
kmeans	Connectivity	4.8579	7.7869	15.4052	20.1921	23.1210	26.0500	29.9079	32.8369	41.7873	44.7163
	Dunn	0.3972	0.4022	0.3969	0.3980	0.4019	0.4053	0.4154	0.4327	0.4416	0.4509
	Silhouette	0.3584	0.3590	0.3261	0.3280	0.3287	0.3294	0.3300	0.3308	0.3073	0.3079

	Score	Method	Clusters
	<dbl>	<fctr>	<fctr>
Connectivity	4.8579	hierarchical	2
Dunn	0.4534	hierarchical	11
Silhouette	0.3590	hierarchical	3

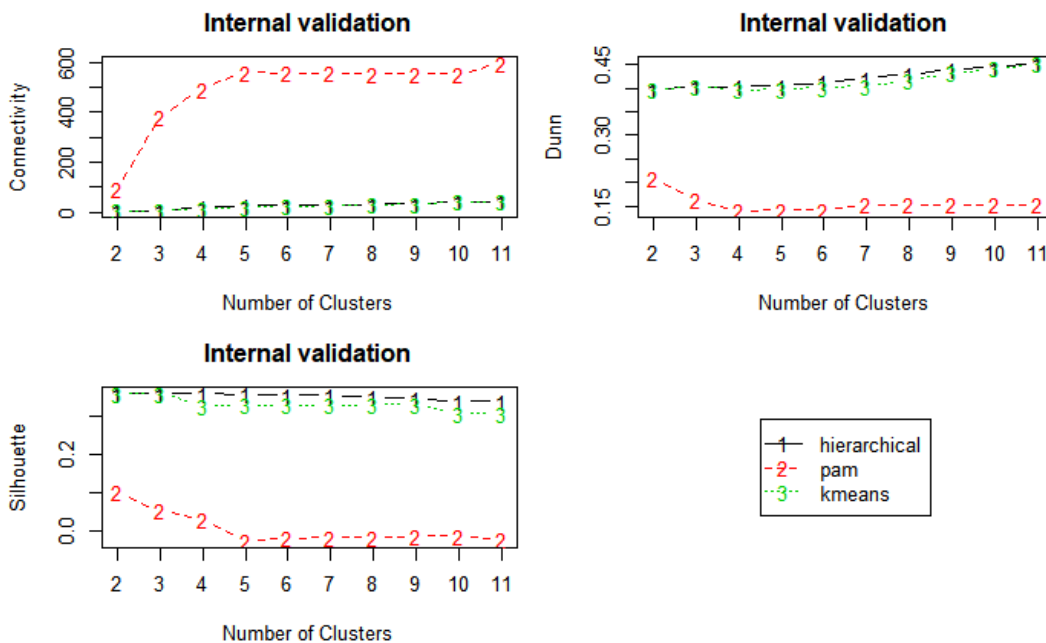


Figure 5.9: Internal validity of Euclidean distance of PAM, K-means, Hierarchical clustering

According to these internal measures, although kmeans and Hierarchical clustering act differently in correlational distances, act similarly in Euclidean distance. For correlational measures hierarchical clustering was the best method out of three, which confirms the choosing hierarchical clustering over other two as most suitable distance measure for expression data is correlational data. These images are belonging to the data analysis of genes with high variance among the samples.

5.4.2.2 Relative Validation

As mentioned in previous chapter relative validation refers to validation done using different number of clusters and different clustering measure

According to this internal validation it can be assume that 2,3,11 are good k values for hierarchical clustering under correlational distance and according to identifying the optimal number of k we have identified that k should be 3,4,7,11. Number of 2 clusters can be identified for any data set. And also 2 subtypes in EC won't be meaningful. Proceed the analysing process with number of clusters as 3,4,7,11. In the same way considering the figures attached in the appendix A, (Figure A.1, Figure A.2, Figure A.3) internal validity, cluster stability and the silhouette and elbow method, most appropriate number of clusters for hormonal gene set is 3,4,10,11

According to the literature it is better to use correlational distance for expression data analysing. Compared the correlation of combinations of distances and linkage measure in hierarchical clustering as shown in the Figure 5.10.

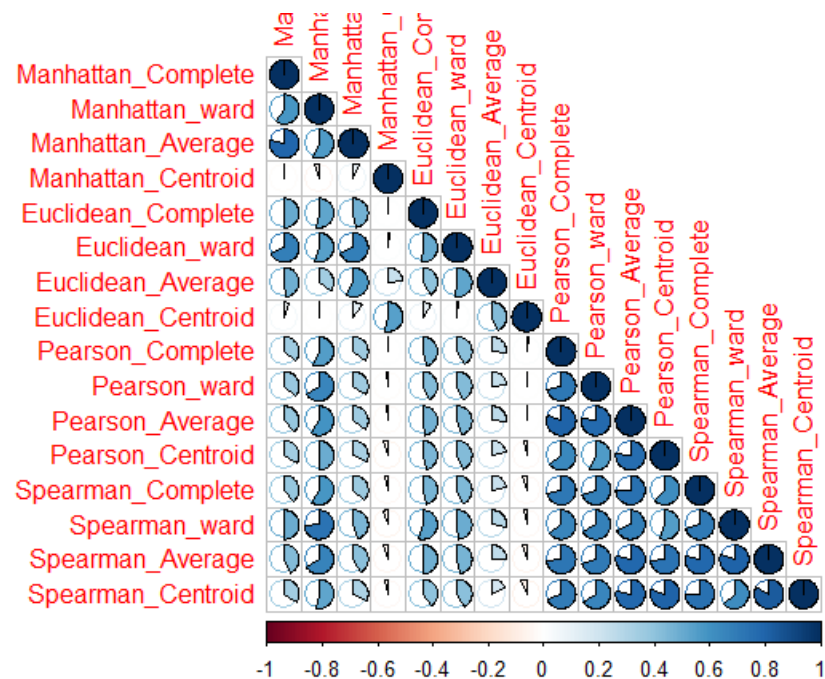


Figure 5.10: Correlational plot of different distance measures and linkage measures

According to the correlational plot it can be seen that correlational distances are having more correlation than the magnitude base distances.

5.2.2.3 External cluster validation

As discussed in the previous chapter, used two approaches to validate the clusters using external information. In visual method used heatmaps to evaluate the relationships of clinical data to classes and gene expression data to classes shown in the Figure 5.11 below.

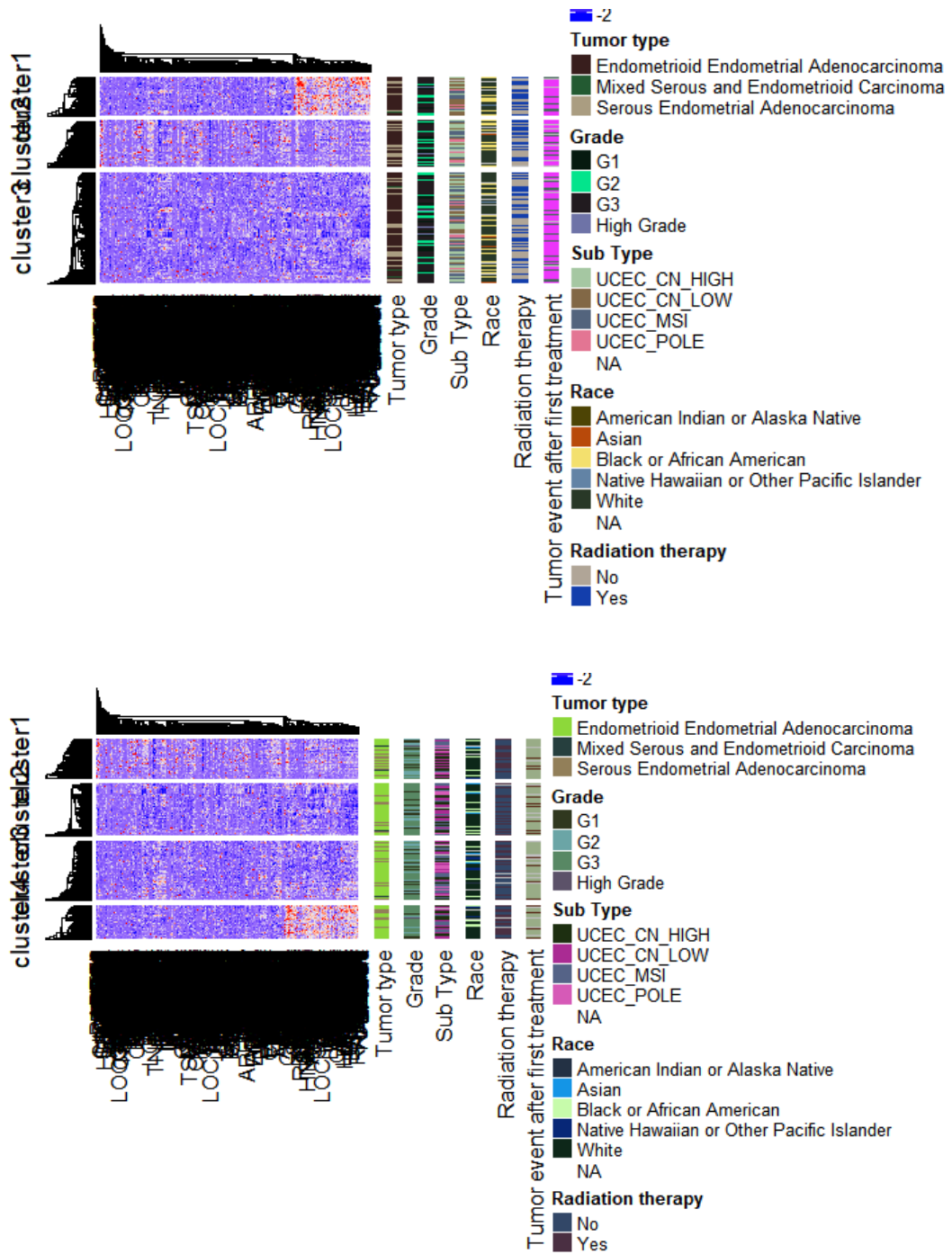


Figure 5.11: Two Heatmaps of highly varied genes split under 3 and 4 clusters identified in spearman distance and complete linkage in highly varied gene dataset.

Due to high dimensionality and heterogeneity it was really hard to identify most meaningful classes. There is no relative similarity between clinical variables considered in this study. But it can be seen that there is a pattern in expression data which is related with the classes, although it can be take k=4 cluster for further analysis it is hard to tell whether this is the most meaningful subtypes could be identified.

Then conducted a comparison between currently existing subtypes of EC's as in [13] using corrected rand index. Following Figure 5.12 and Figure 5.13 shows the comparison between the classes identified using spearman distance, complete linkage and ward linkages. Which shows disagreement. The existing subtypes have identified basically considering CNV. Expression data and copy number data can be different due to various reasons, as discussed in the literature review CNV is a frame shift mutation. Sometimes this frameshift wont effect the functionality of a gene which was basically represent by expression data and due to epigenetic reasons also copy number and expression data can be different.

	UCEC_CN_HIGH	UCEC_CN_LOW	UCEC_MSI	UCEC_POLE
1	113	108	108	34
2	50	39	40	15
NAs introduced by coercion[1] -0.0007628906				

Figure 5.12: comparison between classes identified using spearman distance and complete linkage

	UCEC_CN_HIGH	UCEC_CN_LOW	UCEC_MSI	UCEC_POLE
1	75	58	57	19
2	88	89	91	30
NAs introduced by coercion[1] -0.001386385				

Figure 5.13: comparison between classes identified using spearman distance and complete linkage.

In the hormonal gene data set it can be seen that there are 15 clusters according to the Figure 5.15. Selected k = 15 and conducted further analysis to identify the significant hormonal genes in the clusters identified.

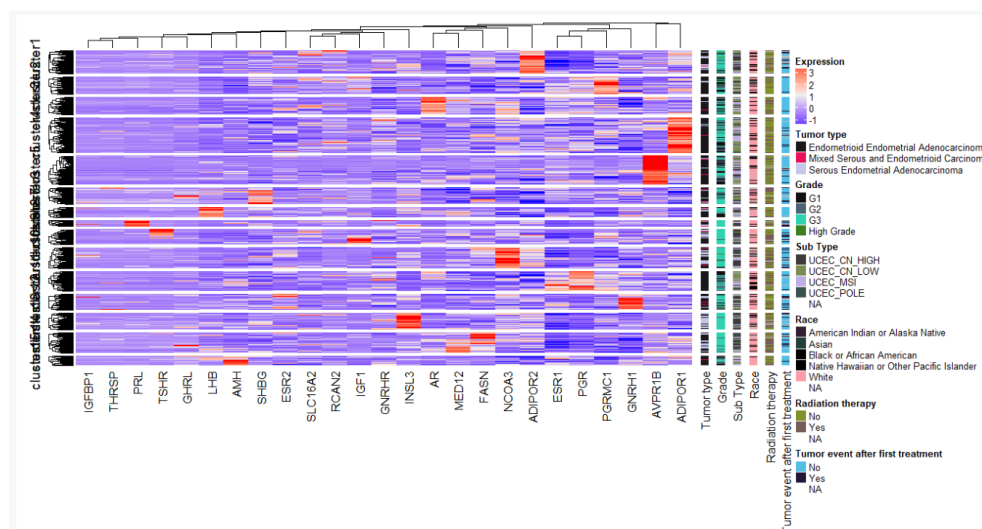


Figure 5.14: Heatmap with 15 clusters in hormonal gene dataset

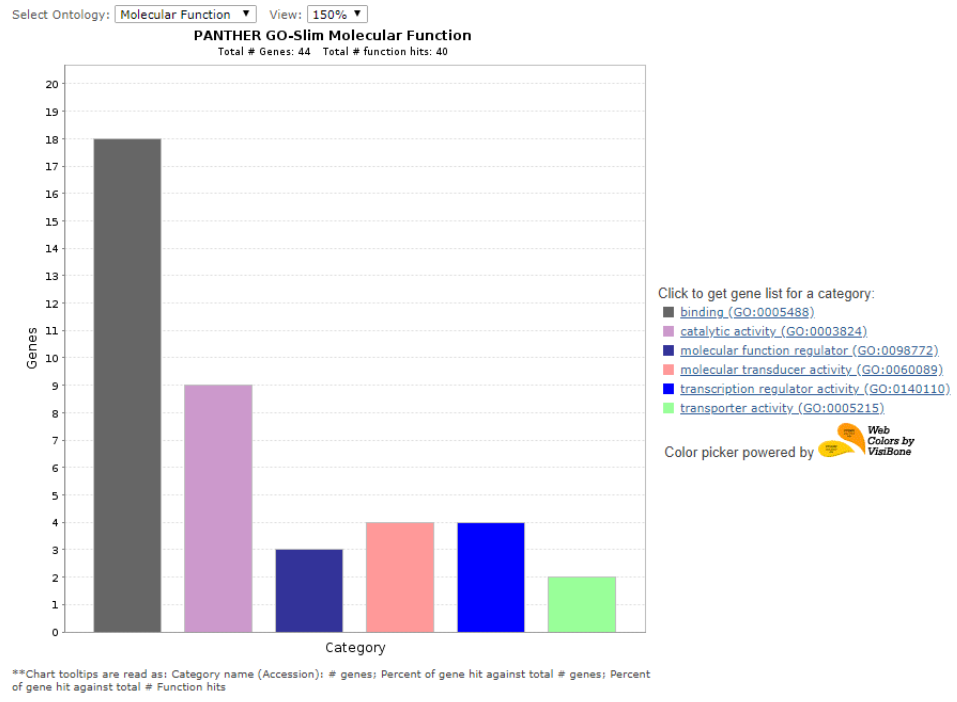
5.3 Identifying Hormonal influence in subtypes and stages

Identifying the discriminative genes in classes and stages were conducted through feature filtering method BSS/WSS which rank features according to their discriminative power and then did a gene enrichment analysis using panther tool. The discriminative features identified and the details about the genes are shown in the following tables.

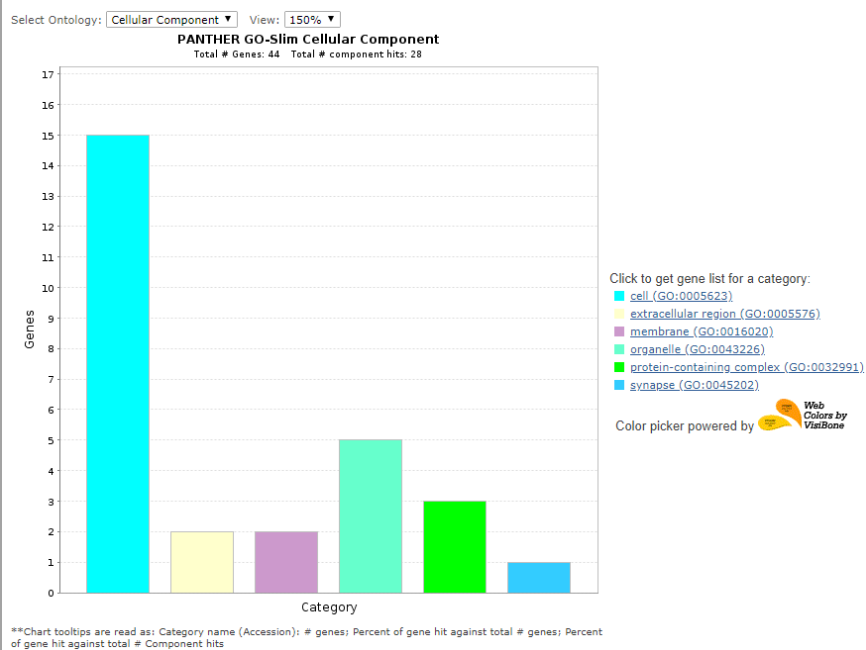
Table 5.5 Stages specific discriminative gene set from all gene set

	From all genes
Genes	P2RX2, MUC2, HIST1H3D, RGS7, VRTN, IRX2, TDRD9, CYP4F35P, PNMT, RGMB, WFIKKN2, FSTL4, MOV10L1, CYP1A1, STAP1, CCAR2, HIST1H4H, TAC3, PCAT18, NKX2.2, TM9SF4, DYM, TRIM44, HIST1H2AD, HIST1H2BF, HIST1H2BG, TNNT3, RECQL, HIST1H2AM, NHLH1, WNK4, TSPY26P, PAOX, APP, CHRM1, MAGIX, TEX14, EPB41, LINC00461, QSER1, EIF4G2, SLC9A3R2, FLRT1, SMARCA5, POFUT1, SGF29, HIST1H2BJ, USP38, SKOR1, CENPI
Protein class	<p>Select Ontology: Protein Class View: 150%</p> <p>PANTHER Protein Class Total # Genes: 44 Total # protein class hits: 31</p> <p>Click to get gene list for a category:</p> <ul style="list-style-type: none"> calcium-binding protein (PC000660) chaperone (PC00072) cytoskeletal protein (PC00085) enzyme modulator (PC00095) hydrolase (PC00121) nucleic acid binding (PC00171) oxidoreductase (PC00176) receptor (PC00197) transcription factor (PC00218) transferase (PC00220) transporter (PC00227) <p>Color picker powered by Web Colors by VistaZone</p> <p>**Chart tooltips are read as: Category name (Accession): # genes; Percent of gene hit against total # genes; Percent of gene hit against total # Protein Class hits</p>

Molecular function



Cellular component

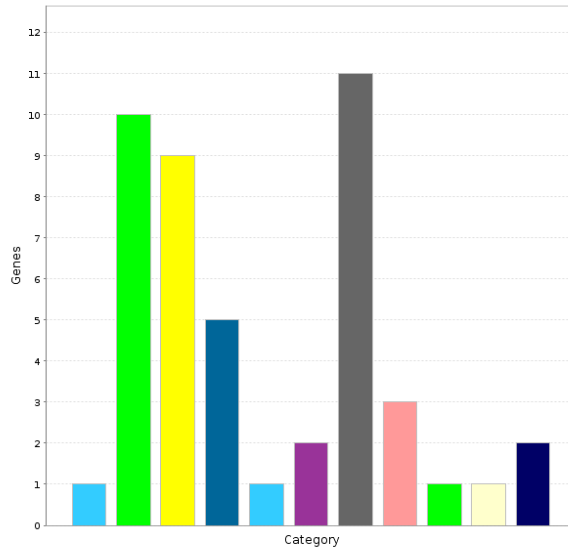


Biological Process

Select Ontology: Biological Process View: 150%


PANTHER GO-Slim Biological Process

Total # Genes: 44 Total # process hits: 46



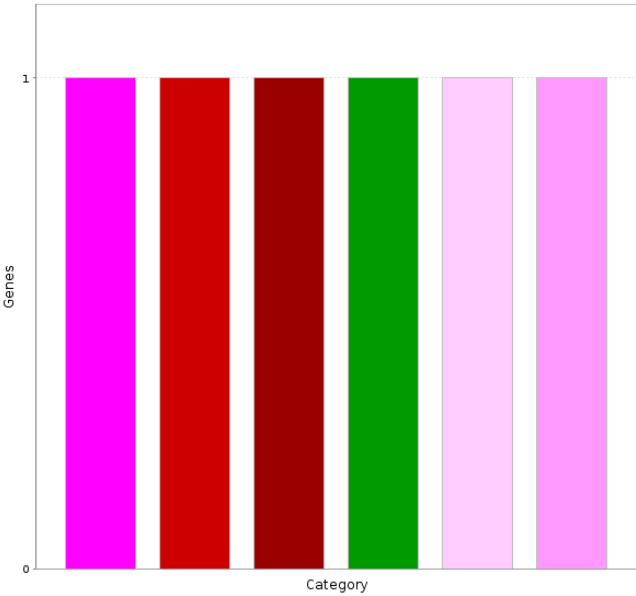

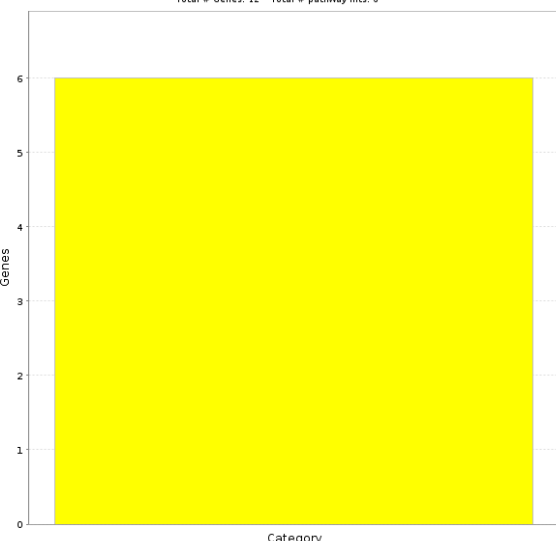

Click to get gene list for a category:

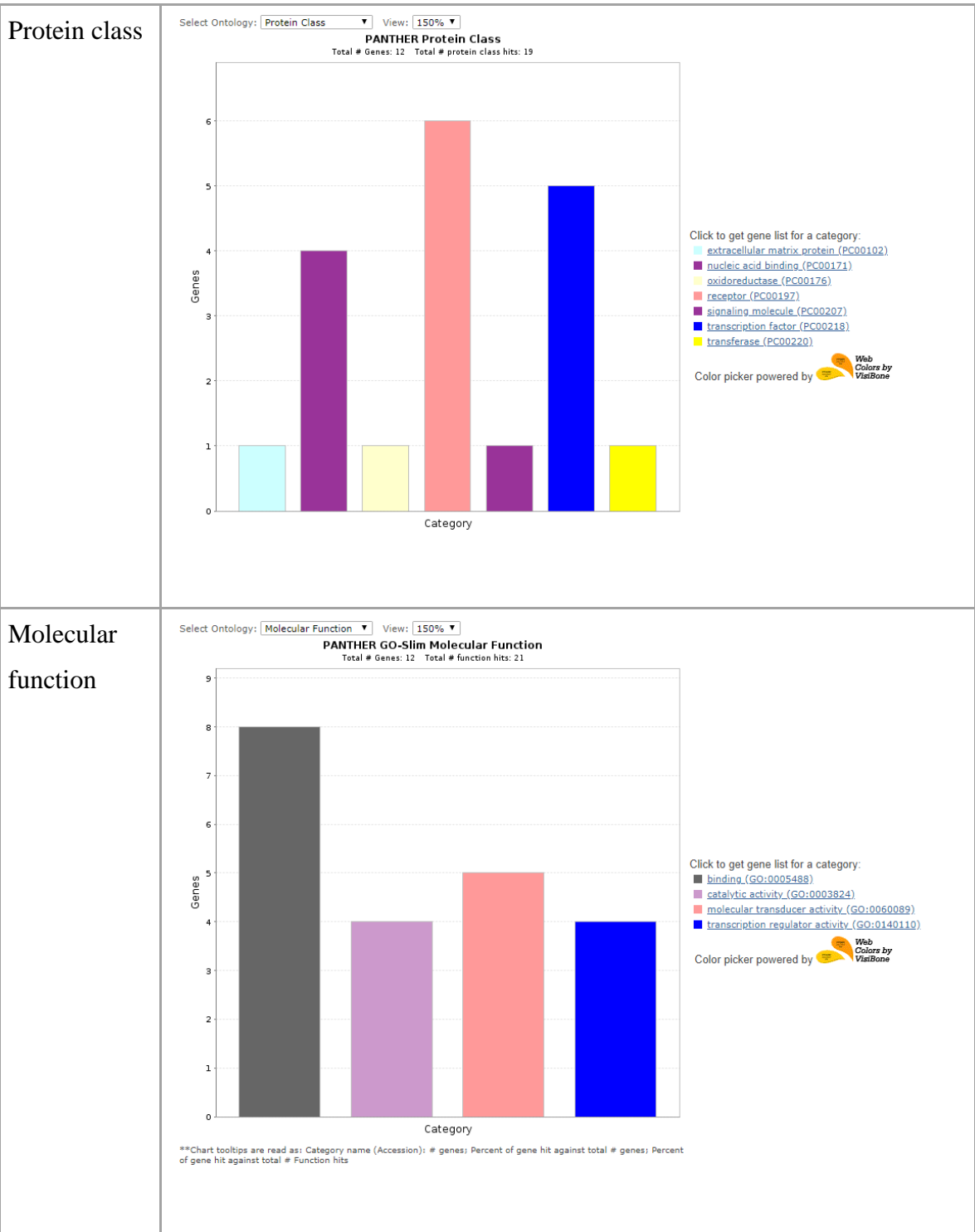
- [biological adhesion \(GO:0022610\)](#)
- [biological regulation \(GO:0065007\)](#)
- [cellular component organization or biogenesis \(GO:0071840\)](#)
- [cellular process \(GO:0009987\)](#)
- [developmental process \(GO:0032502\)](#)
- [localization \(GO:0051179\)](#)
- [metabolic process \(GO:0008152\)](#)
- [multicellular organismal process \(GO:0032501\)](#)
- [reproduction \(GO:0000003\)](#)
- [response to stimulus \(GO:0050896\)](#)
- [signaling \(GO:0023052\)](#)

Color picker powered by 

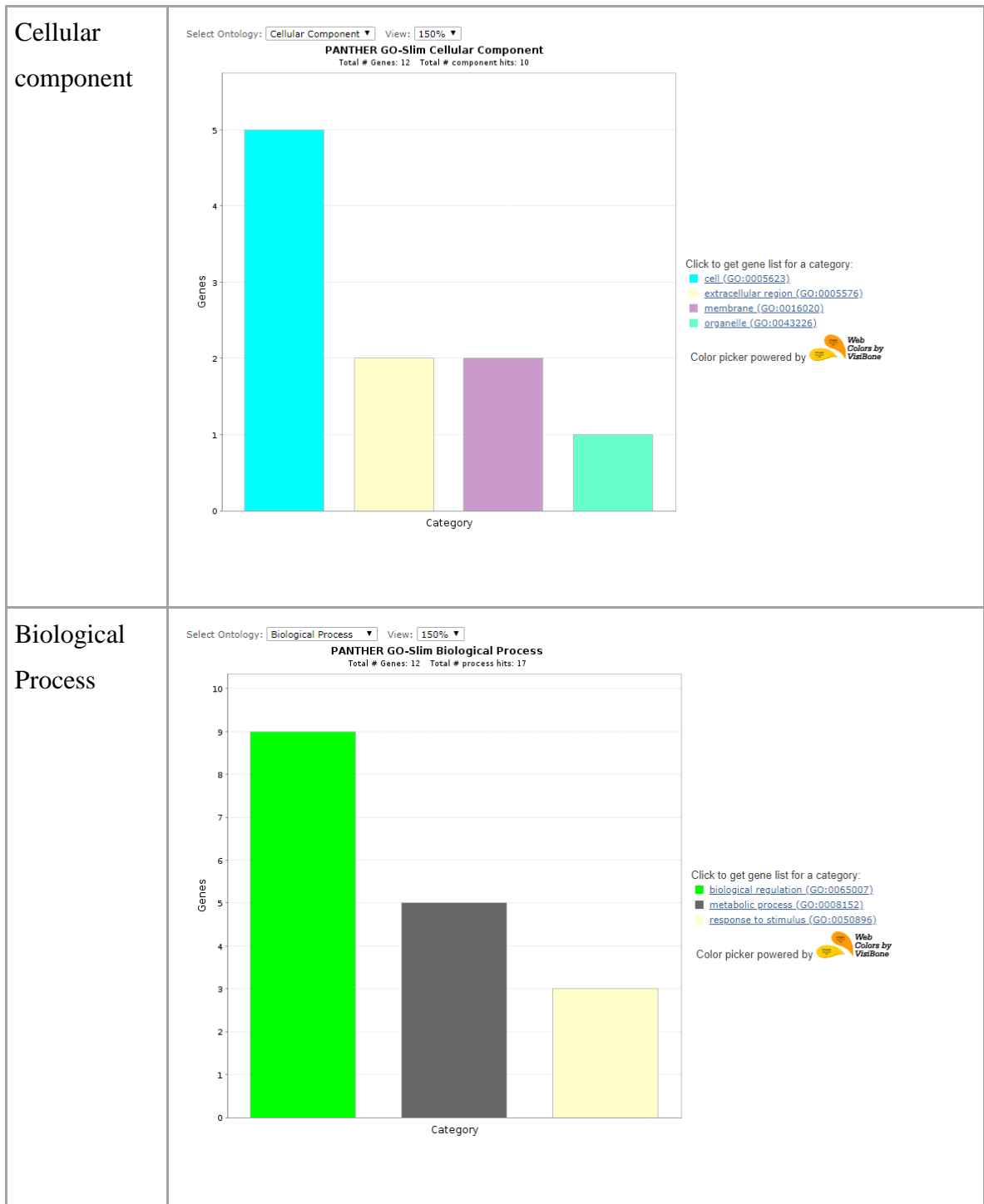
**Chart tooltips are read as: Category name (Accession): # genes; Percent of gene hit against total # genes; Percent of gene hit against total # Process hits

Table 5.6: Stages specific discriminative hormones in all gene set

	Stages specific discriminative hormonal genes
Genes	PGR,ESR1,FASN,ADIPOR2,INSL3,TSHR,AVPR1B,NCOA3,PRL,AR
Pathway	<p>Select Ontology: <input type="text" value="Pathway"/> View: <input type="text" value="150%"/></p> <p>PANTHER Pathway Level 1: Gonadotropin-releasing hormone receptor pathway (P06664) Total # Genes: 0 Total # pathway hits: 6</p>  <p>Click to get gene list for a category:</p> <ul style="list-style-type: none"> ■ AR (P06774) ■ AdipoR1/R2 (P06706) ■ Ncoa3 (P06719) ■ PR (P06708) ■ Pbx1 (P06729) ■ Prolactin (P06789) <p>Color picker powered by </p> <p>**Chart tooltips are read as: Category name (Accession): # genes; Percent of gene hit against total # genes; Percent of gene hit against total # Pathway hits</p> <p>Select Ontology: <input type="text" value="Pathway"/> View: <input type="text" value="150%"/></p> <p>PANTHER Pathway Total # Genes: 12 Total # pathway hits: 6</p>  <p>Click to get gene list for a category:</p> <ul style="list-style-type: none"> ■ Gonadotropin-releasing hormone receptor pathway (P06664) <p>Color picker powered by </p> <p>**Chart tooltips are read as: Category name (Accession): # genes; Percent of gene hit against total # genes; Percent of gene hit against total # Pathway hits</p>



**Chart tooltips are read as: Category name (Accession): # genes; Percent of gene hit against total # genes; Percent of gene hit against total # Function hits



It could be identified 15 discriminative genes which are in the hormonal gene dataset shown in the Table 5.7 and Figure 5.15.

Conducted a gene enrichment analysis using panther tool for the discriminative 15 genes in Table 5.7 as shown in Figure 5.16. which shows these significant genes are connected with gonadotropin releasing hormone pathway and Insulin/ IGF pathway.

Table 5.7: 15 discriminative hormonal genes of Cluster 15 in hormonal gene dataset

Gene Symbol	Description
ADIPOR1	adiponectin receptor 1
ADIPOR2	adiponectin receptor 2
AMH	anti-Mullerian hormone
AR	androgen receptor
AVPR1B	arginine vasopressin receptor 1B
FASN	fatty acid synthase
IGF1	insulin like growth factor 1
INSL3	insulin like 3
MED12	mediator complex subunit 12
NCOA3	nuclear receptor coactivator 3
PGR	progesterone receptor
PGRMC1	progesterone receptor membrane component 1
SHBG	sex hormone binding globulin
TSHR	thyroid stimulating hormone receptor
GNRH1	gonadotropin releasing hormone 1

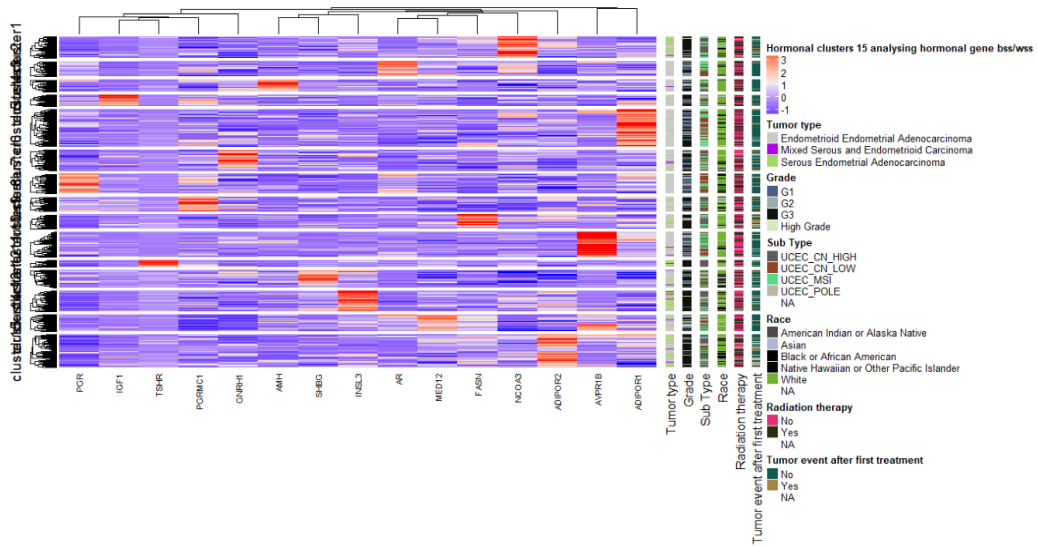


Figure 5.15: Discriminative 15 hormonal genes highly expressed only at one cluster at a time

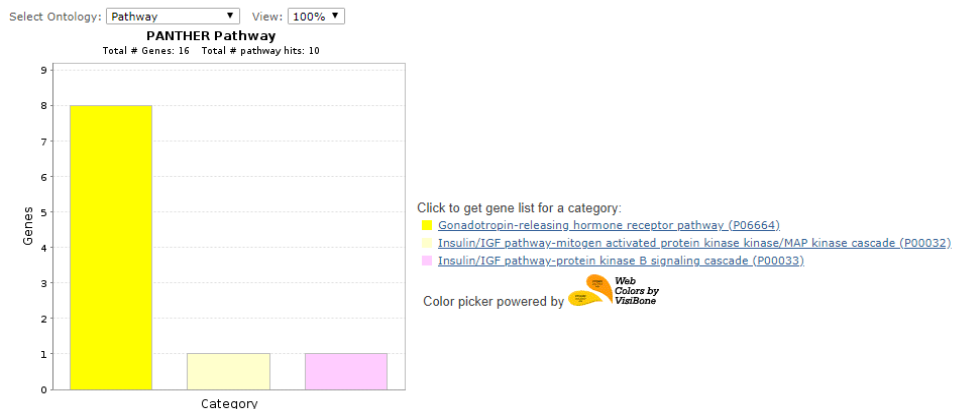


Figure 5.16: Pathway analysis of 15 discriminative hormonal genes of 15 clusters identified using hormonal gene dataset

Chapter 6 - Conclusions

6.1 Introduction

This chapter includes a review of the research with the objectives and the results obtained. Limitations of the current work identified during the research and directions for future researches.

6.2 Conclusions about research questions (aims/objectives)

Main objective of this study was to identify meaningful subtypes of EC, which was obtained using the knowledge of prior researches and data analytical techniques. Due to the high dimensionality and heterogeneity of data, clustering is hard in genomic data. Then identifying the hormonal influence in the subtypes and stages were analysed using panther gene enrichment analysing tool mentioned in the Chapter 3.

Identified a set of k values using Elbow and Silhouette methods and validated them using internal validation measures calculated using clValid R package as shown in Figure 5.8 and Figure 5.9. The optimal number of k values identified and used for the further analysis are 3, 4, 7 and 11 for high coefficient of variance genes dataset and 2, 3, 4 and 15 for hormonal genes dataset.

From the internal validation measures in Figure 5.8. and Figure 5.9. it can be seen that PAM has low internal validity than Hierarchical clustering and k means. Further analysis was done using K-means and Hierarchical clustering with Spearman distance, Ward and Complete linkages.

Visually validated the clusters using heat maps obtained for the set of K values in hierarchical clustering and k-means and identified Cluster 3, Cluster 4 of highly varied genes dataset and Cluster 15 of hormonal gene dataset as shown in Figure 5.11 and Figure 5.14 respectively as meaningful clusters having good discrimination of expression levels between clusters. Although tried to align clusters with clinical data like tumor type, Grade, Subtypes identified in [5] study, Race, Radiation therapy, tumor event after first treatment a better distinction among clusters was not shown.

Analysis was conducted in two phases. In first phase the genes were analysed within the clusters(subtypes) identified from clustering and in the second phase genes were analysed within the grades of UCEC.

Most discriminative genes among clusters identified using external validation were identified using BSS/WSS [36]. First calculated the rank and sorted them in descending order. First 50 genes were selected out of whole gene set (17,000+) and first 15 genes out of the hormonal gene set (25) were selected from the ranked list. And validated the results using heatmaps by plotting heatmaps of the samples with the genes identified as variables. Next gene enrichment analysis was conducted using the PANTHER tool.

From the heatmaps of the 3 clusters it could be seen that in Cluster 15 identified using hormonal gene dataset has discriminative hormonal genes which has highly expressed in one cluster at a time as shown in Figure 5.15 above.

According to the gene enrichment analysis of those 15 hormonal genes as shown in the Figure 5.16 above, these genes have 8 hits in Gonadotropin-releasing hormone receptor pathway and 2 hits in Insulin/IGF pathway. Details of these 15 genes can be seen in Table 5.7 above.

Then the most discriminative genes among grades were identified using BSS/WSS [36]. First calculated the rank and sorted them in descending order and selected first 50 genes out of whole gene set (17,000+) and first 15 genes out of hormonal gene set (25). And validated the results using heatmaps by plotting heatmaps of the samples with the genes identified as variables. Then conducted the gene enrichment using PANTHER, but identified grades discriminative genes were not clearly shown better discrimination among grades.

6.3 Conclusions about research problem

Research problem was to analyse gene expression data in order to identify meaningful classes which will improve the efficiency of the treatments while reducing the toxicity. According to the knowledge gap of genetic level hormonal influence identified using [5] study can be addressed by identifying the hormonal gene behaviour inside the subtypes and stages.

15 clusters output obtained from hormonal gene dataset, very well discriminated the hormonal genes as they are highly expressed in one cluster at a time (Figure 5.15). Considering that and heatmaps, it was clear that better subtypes can be identified through the cluster analysis results with 15 clusters identified with hormonal gene dataset using K-means clustering. Significant genes in 15 clusters are connected with gonadotropin releasing hormone pathway and Insulin/ IGF pathway (Figure 5.16).

6.4 Limitations

High dimensionality of the data is a huge limitation that came across when interpreting results and identifying the important features.

Less amount of availability of samples also became a limitation which stop applying deep learning techniques to automate the feature selection technique

Difficulty face in validating the clustering results is a limitation, as although the clusters can be identified and validated through statistical methods, meaningfulness of the classes are difficult to identify.

6.5 Implications for further research

As PanCancer project integrating all the cancer types and identifying underlying genetic level patterns using deep learning techniques like SDAE will improve the dimensionality reduction. Despite of the perfect algorithms, dataset plays a major role in data analysing can improve the feature selection techniques. Try to improve measures used in clustering. Improving the feature filtering methods to identify most important genomic features to help in target therapeutics. Identify integrative clustering methods to cluster considering all types of genomic data.

References

- [1] K. Lindemann, A. Eskild, L. J. Vatten, and F. Bray, “Endometrial cancer incidence trends in Norway during 1953-2007 and predictions for 2008-2027,” *Int. J. Cancer*, vol. 127, no. 11, pp. 2661–2668, 2010.
- [2] N. Colombo *et al.*, “Endometrial cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up,” *Ann. Oncol.*, vol. 24 Suppl 6, pp. vi33-8, 2013.
- [3] R. Iyer *et al.*, “Predictors of complications in gynaecological oncological surgery: A prospective multicentre study (UKGOSOC - UK gynaecological oncology surgical outcomes and complications),” *Br. J. Cancer*, vol. 112, no. 3, pp. 475–484, 2015.
- [4] *Survival by stage of endometrial cancer*. American Cancer Society, 2014.
- [5] A. Kamal *et al.*, “Hormones and endometrial carcinogenesis,” *Horm. Mol. Biol. Clin. Investig.*, vol. 25, no. 2, pp. 129–148, 2016.
- [6] T. R. Golub *et al.*, “Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring,” *Science (80-.)*, vol. 286, no. 5439, pp. 531–527, 1999.
- [7] J. Gao *et al.*, “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal,” *Sci. Signal.*, vol. 6, no. 269, 2013.
- [8] J. C. Bezdek and R. J. Hathaway, “VAT: a tool for visual assessment of (cluster) tendency,” *Proc. 2002 Int. Jt. Conf. Neural Networks. IJCNN’02 (Cat. No.02CH37290)*, pp. 2225–2230.
- [9] K. Chang *et al.*, “The Cancer Genome Atlas Pan-Cancer analysis project,” *Nat. Genet.*, vol. 45, no. 10, pp. 1113–1120, 2013.
- [10] F. S. Liu, “Molecular carcinogenesis of endometrial cancer,” *Taiwan. J. Obstet. Gynecol.*, vol. 46, no. 1, pp. 26–32, 2007.
- [11] M. K. McConechy *et al.*, “Use of mutation profiles to refine the classification of endometrial carcinomas,” *J. Pathol.*, vol. 228, no. 1, pp. 20–30, 2012.
- [12] R. A. Soslow, *Endometrial carcinomas with ambiguous features*, vol. 27. 2010.
- [13] G. Getz *et al.*, “Integrated genomic characterization of endometrial carcinoma,”

Nature, vol. 497, no. 7447, pp. 67–73, 2013.

- [14] D. C. Cornelius and B. Lamarca, “TH17- and IL-17- mediated autoantibodies and placental oxidative stress play a role in the pathophysiology of pre-eclampsia,” *Minerva Ginecol.*, vol. 66, no. 3, pp. 243–249, 2014.
- [15] C. S. Greene, J. Tan, M. Ung, J. H. Moore, and C. Cheng, “Big data bioinformatics.,” *J. Cell. Physiol.*, vol. 229, no. 12, pp. 1896–1900, Dec. 2014.
- [16] A. C. Liew, N. F. Law, and H. Yan, “Missing value imputation for gene expression data: Computational techniques to recover missing data from available information,” *Brief. Bioinform.*, vol. 12, no. 5, pp. 498–513, 2011.
- [17] M. C. P. de Souto, I. G. Costa, D. S. A. de Araujo, T. B. Ludermir, and A. Schliep, “Clustering cancer gene expression data: A comparative study,” *BMC Bioinformatics*, vol. 9, pp. 1–14, 2008.
- [18] M. C. P. de Souto, P. A. Jaskowiak, and I. G. Costa, “Impact of missing data imputation methods on gene expression clustering and classification,” *BMC Bioinformatics*, vol. 16, no. 1, p. 64, Feb. 2015.
- [19] C. Sotiriou *et al.*, “Breast cancer classification and prognosis based on gene expression profiles from a population-based study,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 18, pp. 10393–10398, 2003.
- [20] J. D. Young, C. Cai, and X. Lu, “Unsupervised deep learning reveals prognostically relevant subtypes of glioblastoma,” *BMC Bioinformatics*, vol. 18, no. Suppl 11, 2017.
- [21] W. J. Gibson *et al.*, “The genomic landscape and evolution of endometrial carcinoma progression and abdominopelvic metastasis.,” *Nat. Genet.*, vol. 48, no. 8, pp. 848–855, Aug. 2016.
- [22] R. Tibshirani, G. Walther, and T. Hastie, “Estimating the number of clusters in a data set via the gap statistic,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 63, no. 2, pp. 411–423, 2001.
- [23] V. N. Kristensen, O. C. Lingjærde, H. G. Russnes, H. K. M. Vollan, A. Frigessi, and A. L. Børresen-Dale, “Principles and methods of integrative genomic analyses in cancer,” *Nat. Rev. Cancer*, vol. 14, no. 5, pp. 299–313, 2014.
- [24] J. M. Piulats *et al.*, “Molecular approaches for classifying endometrial carcinoma,” *Gynecol. Oncol.*, vol. 145, no. 1, pp. 200–207, 2017.
- [25] M. Le Gallo and D. W. Bell, “The emerging genomic landscape of endometrial

- cancer,” *Clin. Chem.*, vol. 60, no. 1, pp. 98–110, 2014.
- [26] M. Templ, A. Alfons, and P. Filzmoser, “Exploring incomplete data using visualization techniques,” *Adv. Data Anal. Classif.*, vol. 6, no. 1, pp. 29–47, 2012.
- [27] M. Dasgupta, “The Estrogen and Progesterone Receptors in Endometrial Carcinoma - An Update,” *Endocrinol. Int. J.*, vol. 5, no. 2, 2017.
- [28] D. A. Koutoukidis, M. T. Knobf, and A. Lanceley, “Obesity, diet, physical activity, and health-related quality of life in endometrial cancer survivors.,” *Nutr. Rev.*, vol. 73, no. 6, pp. 399–408, Jun. 2015.
- [29] M. E. Sherman, “Theories of endometrial carcinogenesis: A multidisciplinary approach,” *Mod. Pathol.*, vol. 13, no. 3, pp. 295–308, 2000.
- [30] O. Granichin, “Cluster validation,” *Intell. Syst. Ref. Libr.*, vol. 67, pp. 163–228, 2015.
- [31] Z. Gu, R. Eils, and M. Schlesner, “Complex heatmaps reveal patterns and correlations in multidimensional genomic data,” *Bioinformatics*, vol. 32, no. 18, pp. 2847–2849, 2016.
- [32] S. Dudoit, J. Fridlyand, and T. P. Speed, “Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data,” *J. Am. Stat. Assoc.*, vol. 97, no. 457, pp. 77–87, 2002.
- [33] H. Mi, A. Muruganujan, and P. D. Thomas, “PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees,” *Nucleic Acids Res.*, vol. 41, no. D1, pp. 377–386, 2013.
- [34] S. Solorio-Fernández, J. Ariel Carrasco-Ochoa, and J. F. Martínez-Trinidad, “Ranking Based Unsupervised Feature Selection Methods: An Empirical Comparative Study in High Dimensional Datasets: 17th Mexican International Conference on Artificial Intelligence, MICAI 2018, Guadalajara, Mexico, October 22–27, 2018, Proceedings, Part I,” 2018, pp. 205–218.
- [35] A. Kassambara, “Practical Guide To Cluster Analysis in R (preview),” pp. 1–38, 2015.
- [36] T. N. Buhtoiarova, C. A. Brenner, and M. Singh, “Role of current and emerging biomarkers in resolving persistent clinical dilemmas,” *Am. J. Clin. Pathol.*, vol. 145, no. 1, pp. 8–21, 2016.

Appendix A: Diagrams

Clustering Methods:
hierarchical pam kmeans

Cluster sizes:
2 3 4 5 6 7 8 9 10 11

Validation Measures:

	2	3	4	5	6	7	8
hierarchical	196.9579	286.3722	298.6250	335.7349	367.1226	374.1496	401.7603
Connectivity	420.3377	451.7242	457.2718				
Dunn	0.0556	0.0585	0.0599	0.0605	0.0636	0.0661	0.0664
Silhouette	0.0973	0.0839	0.0871	0.0939	0.0986	0.1075	0.1071
pam	221.6333	296.5504	372.7071	413.1377	450.3286	446.0643	466.3825
Connectivity	475.7853	497.0313	467.5171				
Dunn	0.0428	0.0596	0.0462	0.0383	0.0374	0.0403	0.0410
Silhouette	0.0430	0.0430	0.1730	0.1728	0.1685	0.1915	0.1953
kmeans	182.6401	261.1226	342.7226	355.6349	364.1163	372.9171	373.6615
Connectivity	393.8948	414.6075	427.9659				
Dunn	0.0414	0.0414	0.0414	0.0422	0.0380	0.0395	0.0400
Silhouette	0.2116	0.2298	0.2321	0.1904	0.1769	0.1773	0.1779

	Score	Method	Clusters
	<dbl>	<fctr>	<fctr>
Connectivity	182.6401	kmeans	2
Dunn	0.0676	hierarchical	11
Silhouette	0.2321	kmeans	11

3 rows

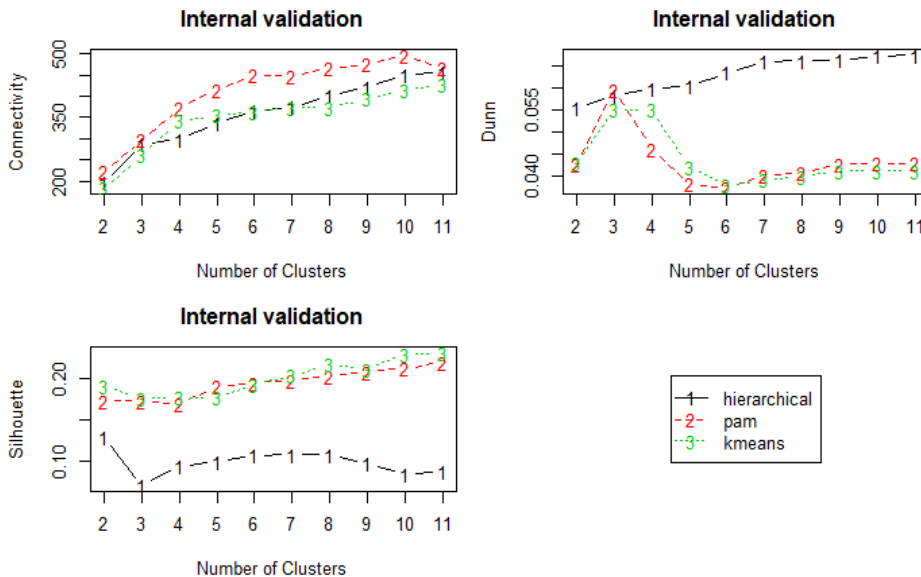
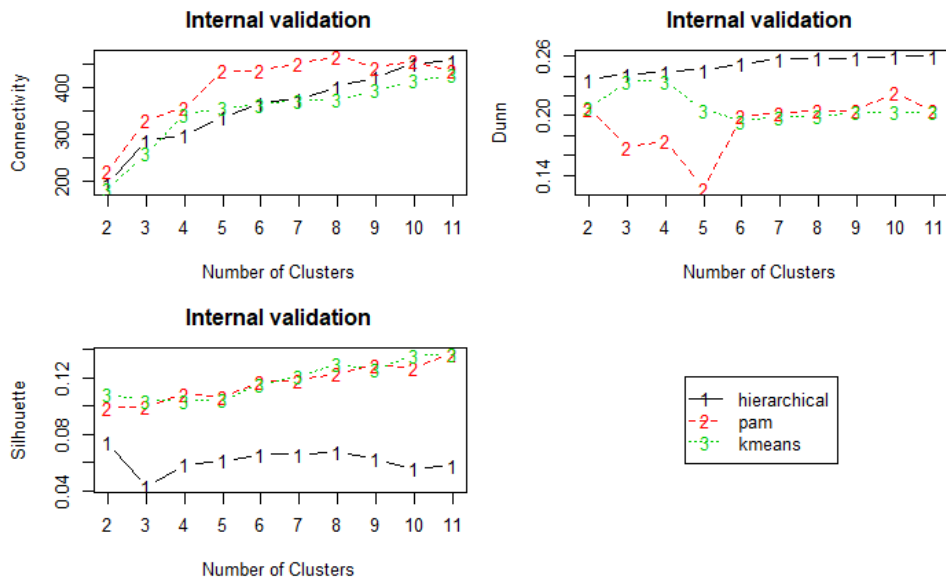


Figure A.1: Internal validity of correlational distance of PAM, Kmeans, Hierarchical clustering of hormonal gene set



Clustering Methods:
hierarchical pam kmeans

Cluster sizes:
2 3 4 5 6 7 8 9 10 11

Validation Measures:

	2	3	4	5	6	7	8	9	10	11
hierarchical Connectivity	196.9579	286.3722	298.6250	335.7349	367.1226	374.1496	401.7603	420.3377	451.7242	457.2718
hierarchical Dunn	0.0556	0.0585	0.0599	0.0605	0.0636	0.0661	0.0664	0.0664	0.0673	0.0676
hierarchical Silhouette	0.1294	0.0695	0.0939	0.0986	0.1069	0.1075	0.1071	0.0973	0.0839	0.0871
pam Connectivity	221.6333	296.5504	372.7071	413.1377	450.3286	446.0643	466.3825	475.7853	497.0313	467.5171
pam Dunn	0.0428	0.0596	0.0462	0.0383	0.0374	0.0403	0.0410	0.0428	0.0430	0.0430
pam Silhouette	0.1730	0.1728	0.1685	0.1915	0.1953	0.1985	0.2038	0.2088	0.2124	0.2205
kmeans Connectivity	182.6401	261.1226	342.7226	355.6349	364.1163	372.9171	373.6615	393.8948	414.6075	427.9659
kmeans Dunn	0.0431	0.0552	0.0552	0.0422	0.0380	0.0395	0.0400	0.0414	0.0414	0.0414
kmeans Silhouette	0.1904	0.1769	0.1773	0.1779	0.1933	0.2039	0.2178	0.2116	0.2298	0.2321

	Score <dbl>	Method <fctr>	Clusters <fctr>
Connectivity	182.6401	kmeans	2
Dunn	0.2599	hierarchical	11
Silhouette	0.1371	pam	11

3 rows

Figure A.2: Internal validity of Euclidean distance of PAM, Kmeans, Hierarchical clustering of hormonal gene set

```

Clustering Methods:
  hierarchical pam kmeans

Cluster sizes:
  2 4 5 6 9 10

Validation Measures:
           2      4      5      6      9      10

hierarchical APN  0.3765 0.5304 0.5624 0.5843 0.6016 0.6136
              AD  0.9636 0.9107 0.8956 0.8758 0.8291 0.8212
              ADM 1.3444 1.8057 1.9366 2.0072 2.1680 2.2121
              FOM 0.9218 0.9116 0.9099 0.9060 0.8994 0.8970
pam          APN  0.1290 0.2915 0.2739 0.3580 0.2758 0.3214
              AD  0.9011 0.8262 0.7752 0.7645 0.6860 0.6838
              ADM 0.4911 1.0528 1.0024 1.2920 1.0537 1.2677
              FOM 0.9226 0.9066 0.9055 0.8997 0.8895 0.8895
kmeans      APN  0.0548 0.3623 0.4619 0.2957 0.3980 0.3822
              AD  0.8771 0.8402 0.8264 0.7417 0.7084 0.6819
              ADM 0.2252 1.5812 1.9011 1.1925 1.5935 1.5143
              FOM 0.9082 0.9046 0.9001 0.8977 0.8875 0.8860

Optimal Scores:

```

	Score <dbl>	Method <fctr>	Clusters <fctr>
APN	0.0548	kmeans	2
AD	0.6819	kmeans	10
ADM	0.2252	kmeans	2
FOM	0.8860	kmeans	10

Figure A.3: Stability of clusters identified using PAM, Kmeans, Hierarchical clustering of hormonal gene set

Appendix B: Selected Hormonal gene Details

GeneID	Org_name	Symbol	description	other_designations
51094	Homo sapiens	ADIPOR1	adiponectin receptor 1	adiponectin receptor protein 1 progestin and adipoQ receptor family member 1 progestin and adipoQ receptor family member I
79602	Homo sapiens	ADIPOR2	adiponectin receptor 2	adiponectin receptor protein 2 progestin and adipoQ receptor family member 2 progestin and adipoQ receptor family member II
268	Homo sapiens	AMH	anti-Mullerian hormone	muellerian-inhibiting factor Mullerian inhibiting factor Mullerian inhibiting substance anti-Muellerian hormone muellerian-inhibiting substance
367	Homo sapiens	AR	androgen receptor	androgen receptor dihydrotestosterone receptor nuclear receptor subfamily 3 group C member 4
553	Homo sapiens	AVPR1B	arginine vasopressin receptor 1B	vasopressin V1b receptor AVPR V1b AVPR V3 antidiuretic hormone receptor 1B arginine vasopressin receptor 3 pituitary vasopressin receptor 3 vasopressin V3 receptor
2099	Homo sapiens	ESR1	estrogen receptor 1	estrogen receptor E2 receptor alpha ER-alpha estradiol receptor estrogen nuclear receptor alpha estrogen receptor alpha E1-E2-1-2 estrogen receptor alpha E1-N2-E2-1-2 nuclear receptor subfamily 3 group A member 1 oestrogen receptor alpha
2100	Homo sapiens	ESR2	estrogen receptor 2	estrogen receptor beta estrogen receptor beta 4 estrogen receptor beta splice variant, ERbeta2delta7 estrogen receptor beta splice variant, ERbeta4delta7 estrogen receptor beta splice variant, ERbeta6 estrogen receptor beta splice variant, ERbeta6delta7 estrogen receptor beta splice variant, ERbeta7 estrogen receptor beta splice variant, ERbeta7delta7 estrogen receptor beta

				splice variant, ERbetaEx. 4L estrogen receptor beta splice variant, ERbetaEx. 6L nuclear receptor subfamily 3 group A member 2 oestrogen receptor beta
2194	Homo sapiens	FASN	fatty acid synthase	fatty acid synthase short chain dehydrogenase/reductase family 27X, member 1
51738	Homo sapiens	GHRL	ghrelin and obestatin prepropeptide	appetite-regulating hormone In2c-preproghrelin ghrelin, growth hormone secretagogue receptor ligand ghrelin/obestatin preprohormone ghrelin/obestatin prepropeptide growth hormone-releasing peptide motilin-related peptide prepro-appetite regulatory hormone preproghrelin
3479	Homo sapiens	IGF1	insulin like growth factor 1	insulin-like growth factor I insulin-like growth factor 1 (somatomedin C) insulin-like growth factor IB mechano growth factor somatomedin-C
3484	Homo sapiens	IGFBP1	insulin like growth factor binding protein 1	insulin-like growth factor-binding protein 1 IBP-1 IGF-binding protein 1 IGFBP-1 alpha-pregnancy-associated endometrial globulin amniotic fluid binding protein binding protein-25 binding protein-26 binding protein-28 growth hormone independent-binding protein placental protein 12
3640	Homo sapiens	INSL3	insulin like 3	insulin-like 3 insulin-like 3 (Leydig cell) leydig insulin -like hormone leydig insulin-like peptide prepro-INSL3 relaxin-like factor b
9968	Homo sapiens	MED12	mediator complex subunit 12	mediator of RNA polymerase II transcription subunit 12 CAG repeat protein 45 OPA-containing protein activator-recruited cofactor 240 kDa component human opposite paired mediator of RNA polymerase II transcription, subunit 12 homolog putative mediator subunit 12 thyroid hormone receptor-associated protein complex 230 kDa component thyroid hormone receptor-associated protein, 230 kDa subunit trinucleotide repeat containing 11 (THR-associated protein, 230 kDa subunit) trinucleotide repeat-containing gene 11 protein
8202	Homo	NCOA3	nuclear receptor	nuclear receptor coactivator 3 CBP-

	sapiens		coactivator 3	interacting protein amplified in breast cancer 1 protein class E basic helix-loop-helix protein 42 receptor-associated coactivator 3 steroid receptor coactivator protein 3 thyroid hormone receptor activator molecule 1
5241	Homo sapiens	PGR	progesterone receptor	progesterone receptor nuclear receptor subfamily 3 group C member 3
10857	Homo sapiens	PGRMC1	progesterone receptor membrane component 1	membrane-associated progesterone receptor component 1 progesterone binding protein
5617	Homo sapiens	PRL	prolactin	prolactin decidual prolactin growth hormone A1
10231	Homo sapiens	RCAN2	regulator of calcineurin 2	calcipressin-2 Down syndrome candidate region 1-like 1 Down syndrome critical region gene 1-like 1 myocyte-enriched calcineurin-interacting protein 2 thyroid hormone-responsive (skin fibroblasts) thyroid hormone-responsive protein ZAKI-4
6462	Homo sapiens	SHBG	sex hormone binding globulin	sex hormone-binding globulin sex steroid-binding protein testis-specific androgen-binding protein testosterone-binding beta-globulin testosterone-estradiol-binding globulin testosterone-estrogen-binding globulin
6567	Homo sapiens	SLC16A2	solute carrier family 16 member 2	monocarboxylate transporter 8 X-linked PEST-containing transporter monocarboxylate transporter 7 solute carrier family 16, member 2 (thyroid hormone transporter)
7069	Homo sapiens	THRSP	thyroid hormone responsive	thyroid hormone-inducible hepatic protein SPOT14 homolog lipogenic protein 1 spot 14 protein thyroid hormone responsive (SPOT14 homolog, rat) thyroid hormone responsive SPOT14
7253	Homo sapiens	TSHR	thyroid stimulating hormone receptor	thyrotropin receptor TSH receptor seven transmembrane helix receptor thyrotropin receptor-I, hTSHR-I
2798	Homo sapiens	GNRHR	gonadotropin releasing hormone receptor	gonadotropin-releasing hormone receptor GnRH receptor GnRH-R gonadotropin-releasing hormone (type 1) receptor 1 leutinizing hormone releasing hormone receptor leutinizing-releasing hormone receptor luliberin receptor type I GnRH receptor

2796	Homo sapiens	GNRH1	gonadotropin releasing hormone 1	progonadoliberin-1 GnRH-associated peptide 1 gonadotropin-releasing hormone 1 (luteinizing-releasing hormone) leuteinizing-releasing hormone luliberin I prolactin release-inhibiting factor
2492	Homo sapiens	FSHR	follicle stimulating hormone receptor	follicle-stimulating hormone receptor FSH receptor follitropin receptor
2488	Homo sapiens	FSHB	follicle stimulating hormone subunit beta	follitropin subunit beta FSH-B FSH-beta follicle stimulating hormone beta subunit follicle stimulating hormone, beta polypeptide follitropin, beta chain
3972	Homo sapiens	LHB	luteinizing hormone beta polypeptide	lutropin subunit beta interstitial cell stimulating hormone, beta chain luteinizing hormone beta subunit lutropin beta chain
3973	Homo sapiens	LHCGR	luteinizing hormone/choriogonadotropin receptor	lutropin-choriogonadotropic hormone receptor hypergonadotropic hypogonadism lutropin/choriogonadotropin receptor