Gene expression analysis in cancer microarray datasets, investigating the role of an Embryonic Stem Cell Factor in prognosis

Deepankar Chakroborty
Master's Thesis
Master's Degree Programme in Bioinformatics
Department of Information Technology
University of Turku
September 2014

UNIVERSITY OF TURKU

Department of Information Technology

CHAKROBORTY, DEEPANKAR : Gene expression analysis in cancer microarray datasets, investigating the role of an Embryonic Stem Cell Factor in prognosis

Master's thesis 70 p.

Master's Degree Programme in Bioinformatics

September 2014

Cancer is a condition that is demanding more research with new cases being reported each year. In this thesis the aim was to investigate the behaviour of a embryonic stem cell factor and its partners in various types of cancers. The embryonic stem cell factor under study in this thesis is responsible for the maintenance of pleuripotency in stem cells and its interaction partners maintain the self-renewal ability of the embryonic stem cells. With the discovery of cancer stem cells and detections of stem cell like gene signatures from cancers, it becomes important to address the issue to identify the responsible genes. The embryonic stem cell factor of our interest when knocked down in cell line studies showed downregulation of stem cell pleuripotency factors therefore, we believe it may be playing a key role in cancer tissues where it is expressed.

We use gene expression analysis of microarray data of cancer patient samples along with the available survival information to test whether the gene and its partners have any effect on survival. We use correlation measures to establish that partners of the embryonic stem cell factor of our interest might be co-expressed in patient samples. In particular, we were able to identify colon cancer and seminoma samples that express our gene of interest at high levels. We used T-test and ROTS (Reproducibility Optimized Test Statistic) on these datasets to detect which genes are differentially expressed.

The project also presents a different approach to microarray data analysis where the focus is not on the disease or condition but a set of genes are central theme of the study, and the research is done to find the cancer or datasets where the gene set is perturbed. This is desired under situations similar to the premise of this project, that if an embryonic stem cell factor is expressed in adult tissues it is a sign of problems.

The project suggests that the embryonic stem cell factor under question alone cannot be held responsible for poor survival of the cancer patients, instead it seems that it is a pro-survival factor after all. But further analyses are being done in this area to uncover more information and also to find factors that can explain the poor survival of the samples for the cancer datasets under study.

Keywords

Cancer, Stem Cells, Embryonic Stem Cell Factor, Gene Expression Analysis, Differential expression Analysis,ROTS, Survival analysis, Prognosis, Microarray dataset, Co-Expression, Gene-centric Microarray study.

Contents

Acknowledgement

This thesis would not have been possible without several people, towards whom I would like to express my gratitude. The first person to thank would be my supervisor, Laura Elo, Ph.D, who is Adjunct Professor in BioMathematics at the University of Turku. She is a great mentor and guided me throughout the project, her suggestions and questions on results made me think more about solving the problems.

I would also like to thank Professor Riitta Lahesmaa, for giving me this wonderful opportunity to work on this project at Turku Center for Biotechnology. I thank her for financially as well as morally supporting this project. Her encouragement and valuable insights into the mechanism of cancer and stem cells were very essential in guiding me through the project and helped a lot in interpreting the result. I would also like to thank Sigrid Juselius Foundation and Cancer Foundation for funding this project and showing faith in our research.

I would also like to thank Emani Maheswara Reddy, Ph.D., who is a post doctoral research fellow of Academy of Finland, for his constant support and motivation for the project. I would also like to thank the staff and personnel at Turku Center for Biotechnology, for making my working experience pleasant.

I would thank Professor Tapio Salakoski and Eija Nordlund for encouraging my research. I would also thank all my teachers here at University of Turku, who taught me the skills and imparted me with the knowledge necessary for carrying out this research work.

My parents' motivation and encouragement cannot go unnoticed in my eyes, as they have been on this journey with me and have inspired me to keep working hard regardless of the failures or successes, because "*The show must go on*".

Last but not the least, I would thank my colleagues at Computational Biomedicine group, at Turku Center for biotechnology, for all their suggestions during group meetings and discussions, which helped this project take shape.

Deepankar Chakroborty

Turku, 02. September. 2014

List of Abbreviations.

CSC	Cancer Stem Cells
DNA	Deoxyribonucleic acid
ESC-1	Embryonic Stem Cell factor 1 (pseudonym for our gene)
ESCP- (N)	Pseudonyms for partners of ESC-1 e.g. ESCP-1, ESCP-2, …, ESCP-18
GEO	Gene Expression Omnibus
hESC	Human Embryonic Stem Cells
KRAS	V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
MIAME	Minimum Information About a Microarray Experiment
mRNA	messenger RNA
NDUFA1	NADH dehydrogenase 1
PCA	Principal Component Analysis
POL2RA	RNA polymerase II A
PSAT1	Phosphoserine aminotransferase 1
RMA	Robust Multiarray Average
RNA	Ribonucleic acid
RNI	RMA Normalized Intensity
ROTS	Reproducibility Optimized Test Statistic
SCID	Severe Combined Immuno Deficient
TCGA	The Cancer Genome Atlas
UPC	Universal exPression Code
UV	Ultra Violet

1. *Introduction*

Cancer affects millions of people globally, [1] and thousands of new cases are reported in Finland each year [2]. However with more research being conducted on cancer more knowledge is being revealed and more questions are raised. In the last decade of the 20th century Cancer Stem Cells were identified for the first time [3] and then in the first decade of the 21st century much research has been done characterizing them and in identifying them in various cancer tissues [4]–[6]. Embryonic stem cell transcription factors are shown to be highly expressed in cancer samples [7], [8]. In this study we investigate the role of a novel embryonic stem cell factor, which we refer as ESC-1, in different cancers. More information about ESC-1 and its properties is presented in section 2.4.

Gene expression analysis was performed on several cancer datasets (description of datasets presented in section 4.1) spanning over 5500 samples. The gene expression of ESC-1 along with its partners was studied to see how they behave in different cancers. This study was necessary because not much information is available about ESC-1 in cancer samples. It was important to look into ESC-1 because it is an embryonic stem cell factor and should not be expressed in normal healthy adult tissue. It was established that there is a positive correlation between ESC-1 and its interaction partners in cancer samples in section 5.3. Cancer samples that were expressing ESC-1 at high levels were identified (section 5.7). Differentially expressed genes between the samples expressing ESC-1 at high levels and the samples that were not expressing ESC-1 were also identified (section 5.9). There was overlap among the differentially expressed genes between the different colon cancer datasets that we used in this project, more about this in section 5.9. Survival information was analyzed from cancer datasets and contrary to our initial hypothesis, which was that ESC-1 will have a negative effect on survival of patients, there are indications from four colon cancer datasets that ESC-1 might be a positive prognostic factor, in section 5.10. However, further research investigations are required to understand the role of ESC-1 in cancer.

This thesis is a part of a project aimed at characterizing ESC-1in a better way and to study its properties and investigation of its role in cancer. The research work in this project was conducted under the guidance of Dr. Laura Elo Computational Biomedicine group at Turku Centre for Biotechnology. The research work in this project is a part of collaboration between laboratories of Prof. Riitta Lahesmaa and Dr. Laura Elo. The funding for this project was received from Sigrid Juselius Foundation and Cancer Foundation.

Section 2) Background

2.1 Cancer

2.1.1 Hallmarks of Cancer

2.2 Stem Cells

2.2.1 Properties of Stem Cells

2.3 Cancer Stem Cells

2.4 ESC-1 and other pluripotency factors

2.5 Gene expression quantification

2.5.1 Affymetrix platform for microarray

2.5.2 Why perform gene expression analysis?

2.1 *Cancer*

In today's world, everyone is aware of the devastating impact cancer has on our society. World Cancer Research Fund states that in 2012 globally there were 14.1 million reported cases of cancer [9]. Finland has 15,000 new reported cases of cancer every year [2]. Cancer is a biological condition where a group of cells, which all living beings are made of, loose control over their processes and start to behave abnormally [10]. As they are altered self-cells* our immune system is also helpless most of the time. Under normal circumstances events from cell birth to its death are well regulated by the signals received from growth factors, growth inhibitors and apoptotic factors. The important mechanisms in the cell that are impaired when the cell becomes cancerous are cell division, DNA repair, apoptosis (programmed cell death) and cell signalling [11]. However, most often we find that there is a combination of pathways and mechanisms that are disrupted in a cancer cell and not just one, which makes it difficult to target. On top of this there are several components in the pathways and finding out the key factors that are not behaving properly is a large task at hand. Along with this the cancer cells take the route of the circulatory system and lodge themselves at different sites, away from their place of origin. Now they cause problems not on just local but on a systemic level, a condition defined by the term 'metastasis' [11].

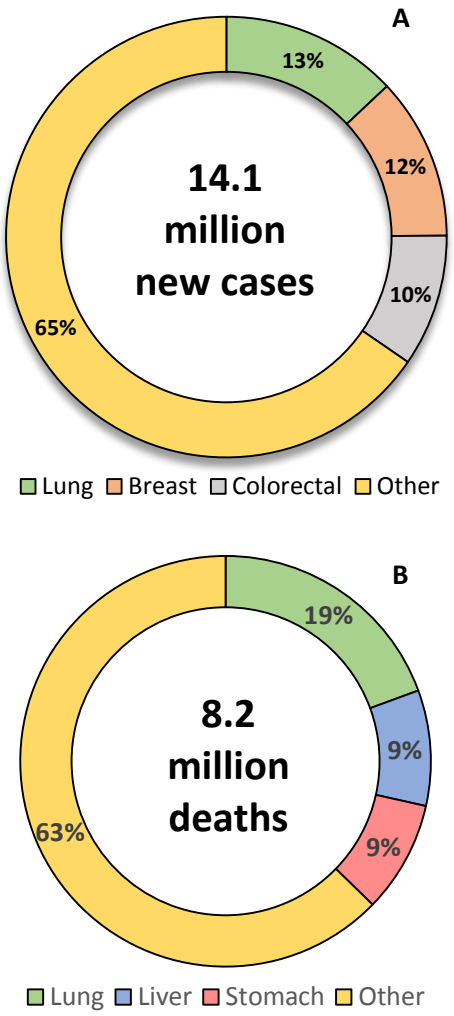


Fig 2.1 A) Pie chart showing the breakdown of 14.1 million new incidences[9] of cancer worldwide in 2012. **B)** Pie chart showing the breakdown of 8.2 million deaths[1] due to cancer worldwide in 2012.

The primary cause of these deviations from the normal life trajectory of the cells (from birth → fulfil function → orchestrated death) is genetic damage to the genes that regulate the important processes. The target genes that are affected by this genetic damage are classified into two groups. The first group of genes are proto-oncogenes, which are when activated due

* Self-cells are cells of our body that our immune system is trained to recognize so that it does not initiate an immunological response once it encounters them.

to mutations become oncogenic (i.e. cancer causing). These genes promote growth of the tissue where they are activated or promote production of a product that promotes growth [11]. The other group is tumour suppressor genes, which are the genes that keep a check on abnormal growth [11]. Most genes in this group are regulators and effectors of cell cycle, apoptosis and DNA damage repair. Cancer causing mutations are acquired during the lifespan of cells due to the influence of carcinogens, like UV light, some chemical agents like acetaldehyde, which is commonly introduced in the human system by smoking tobacco. As these mutations mostly occur in somatic cells*[12] instead of germ cells#[12], thankfully it is not passed on to the next generation. However, there are certain mutations that are carried on through the germ cells, which increase the probability of acquiring cancer. For instance, mutated BRCA1 and BRCA2 genes elevate the risk of developing breast cancer up to 80% and the risk of ovarian cancer up to 40% in a woman's lifetime [13].

2.1.1 *Hallmarks of Cancer*

There is a constant presence of mutagens in the environment, most importantly in the form of Ultra Violet light from the Sun. However despite this fact, we do not observe such high rates of mutations (eventually causing cancer) because our body has certain mechanisms to repair these damages and to keep the cells with irreparable DNA damage from dividing [12]. However, there are certain traits that help a cancer breach through the lines of defence of an organism. These help them to start growing as a tumour and cause more damage over time, and becoming life threatening in most cases. The development of cancer is characterized by certain "Hallmarks of Cancer" [14] which represent the procurement of a capability by the cancer cell that marks a successful breach in one of the anti-cancer mechanisms. A brief explanation is presented below:

1. **Self sufficiency in growth signals:** Under normal circumstances a cell requires signal from external factors that bind to the cellular receptors and give a signal for mitogenesis (i.e. initiation of mitosis), but several cancer tissues gain independence from these growth signals by either over expressing the cell surface receptors [15], or

---

Germ Cells – Reproductive cells (also known as gametes e.g. ovum, sperm), undergo meiosis, progeny is haploid, precursor may be haploid or diploid, genetic recombination occurs thus the progeny is not identical to its precursor cells [12], [14].

* Somatic Cells – Non reproductive cells, undergo mitosis, ploidy same as parent cell, progeny identical to parent cell in genetic content (no recombination occurs in most cases with some exceptions) [12], [14].

by a mutation that keeps the cell surface receptor active irrespective of the binding of the growth signal, which is also known as ligand independent signalling [14].

2. **Insensitivity to anti-growth signals:** Cancer cells become insensitive to factors that keep a check on cell growth and division, e.g. mutation in the pRB (Retinoblastoma protein) and other components in its pathway make the cell insensitive to several anti-growth signals. [14]
3. **Evading apoptosis:** Apoptosis or Programmed Cell Death is an orchestrated destruction of a cell once it receives the signal about its fate. There are several conditions that trigger this in a normal and healthy cell. Apoptosis can be initiated by extra cellular factors, like deficiency of IL-3 survival factor [16], or by some intracellular signals such as signal after detection of DNA damage and signalling imbalance caused by an oncogene [17] to name a few. However, mutations in the apoptotic genes or the receptor molecules for death or survival signals, confers on the cancer cells the ability to evade apoptosis. [14]
4. **Limitless potential for replication (i.e. limitless cell division)**: The above three capabilities facilitate the cancer cells to divide and over time generate macroscopic tumours. In contrast to cancerous cells, normal human cells cannot divide infinitely, which is governed by Hayflick limit, which showed that cells in a culture have a finite replicative potential of about 60-70 doublings [18]. This is primarily due to a phenomenon called 'telomere shortening', which is basically the loss of ends or terminals of chromosomes called telomeres (which are made up of short hexanucleotide repeats, in vertebrates it is TTAGGG), every time a cell divides. It is inevitable and over time when too much information is lost, the cell loses its capability to divide further on [18]. But there is an enzyme called telomerase, which is not expressed under normal circumstances but when expressed in tumours it repairs the telomeres by restoring the fragments lost during cell division. 85% to 90% of the malignant cells use the activation of telomerase enzyme for the maintenance of telomeres [19]. Role of telomerase in immortalizing cells in-vitro was demonstrated in 1998 [20], [21].
5. **Tissue invasion and metastasis:** 90% of human cancer deaths are caused by metastasized tumour [22]. Some cancer cells escape from their location and invade the neighbouring tissue to create a new tumour. They also take the route of the circulatory system to transport the cancer cells to distant body parts and create new tumours over time [14].

6. **Sustained angiogenesis:** To ensure a proper supply of nutrients and oxygen to the tumour once it grows to macroscopic scale, the cancer tissue initiates angiogenesis (i.e. formation of blood vessels). It is necessary for the tumour to develop means of blood supply in order to grow bigger [23].

Cancer Biology is an active field of research, ranging from studies to diagnose and characterize cancer through image analysis [24], to in-silico simulation studies [25], [26], to studies investigating use of natural compounds to compound cancer [27], to studies with novel methods of treatment e.g. using oncolytic viruses as cancer therapy [28].

2.2 *Stem Cells*

Stem cells are undifferentiated cells that have the capability to give rise to progeny that can differentiate into different cell types [11], [29]. They are found in most multicellular organisms. There are two types of stem cells Adult stem cells, found in various tissues (e.g. bone marrow has hematopoietic stem cells) and Embryonic stem cells, which are isolated from the inner cell mass of a blastocyst [29]–[31]. Embryonic Stem Cells are pluripotent [10], i.e. they are capable of giving rise to all cell types of an adult body, if they are provided with the right set of signals to do so. In the recent years (Nobel Prize in Medicine, 2012) there have been studies to induce pluripotency in adult somatic cells [32]. Adult somatic cells do not have properties of the stem cells but introduction of certain transcription factors can convert these cells to stem cells [32]–[34].

2.2.1 Properties of *Stem Cells*

Stem cells are well characterized and are recognized to possess the following properties:

State of differentiation: Stem cells are in an undifferentiated state[29], i.e. they have not yet specialized to carry out a particular function. Despite having the same genetic content, the stem cells are not able to perform the functions that the other somatic cells are capable of[11]. For instance, Neural Stem Cells (NSCs)[35] and neurons both are found in adult mammalian brain, but the difference between them is that NSCs are undifferentiated and hence cannot transmit nerve impulses while their counterparts neurons are differentiated to carry out

nerve impulses.

Self-Renewal: Stem cells have the ability to go through several cycles of cell division while maintaining the capacity to divide further. The stem cells do this while maintaining their undifferentiated state [11].

Potency: Stem cells have the ability of a stem cell to differentiate into multiple cell types based on the signals it receives [32].

Homing: Homing is the migration of cells to their tissue of origin, Hematopoietic Stem Cells undergo this process to reach Bone marrow [36], [37]. It is noteworthy to mention that all stem cells do not perform homing.

Plasticity: Adult stem cells derived from one tissue can be reprogrammed to differentiate into cell types from different germ layers (they are the three primary germ layers in the mammalian embryo: ectoderm, mesoderm, and endoderm [30]). This is known as stem cell transdifferentiation or stem cell plasticity [38]. For instance, Neural Stem Cells from brain (which are derived from ectoderm), can be reprogrammed to differentiate into other cells derived from either of the three germ layers [35].

2.3 *Cancer Stem Cells*

Some of these properties of stem cells are similar to the properties of cancer cell. In a condition like cancer when the cells lose their sensitivity to regulatory elements and resort to various means to sustain themselves in the body, it is not unlikely that they start adopting some properties of stem cells. This is observed in cases where the cancer relapses even after the tumour resection and chemotherapy. This is due to the presence of certain cells in the cancer tissue that have stem cell like properties, primarily the property of Self-Renewal and the ability to give rise to all cells found in that particular cancer. These cells, though very few in number, are called Cancer Stem Cells (CSC) [39]. CSCs are, therefore, capable of giving rise to the whole tumour by themselves [39]. This is what is suspected to take place when there is a tumour relapse after surgery and drug therapy. Both CSCs and Stem cells are known to be stress resistant and have the ability to survive as well as perform their functions in challenging conditions [40], [41]. CSCs are a recent concept in Cancer Biology as they were first cited in 1990s and then, only after a decade there has been a significant focus on research pertaining to CSCs [5] and their properties.

The first Cancer Stem Cells were suspected and identified in the case of leukaemia (a hematological malignancy) [3]. The study found that a small volume of leukaemia cells were able to generate colonies in "Colony Forming Assay" and later on it was proved that not all leukaemia cells but only certain cells are capable of doing this. The study showed that when the cells from Human Acute Myeloid Leukaemia(AML) when transplanted in SCID mice (Severe Combined Immuno-Deficiency) [42], (in these mice the adaptive immune system is compromised), the transplanted human AML cells initiated AML in the mice[3]. Later similar observations were made in colon cancer and colon cancer stem cells[4] were identified using a similar procedure of transplanting human colon cancer initiating cells to SCID mice [4]. The study also suggested that CSCs need to be targeted for cancer therapy to be effective [4]. CSCs have been shown to provide drug resistance to tumours like in lung cancer [43], pancreatic cancer and breast cancer [44].

Following these discoveries, cancer stem cells were identified in other tissues like colon [45] and also in-silico analyses were done which identified stem-cell like gene expression [7] signature in malignancies, e.g. ovarian cancer [46]. Studies have also been conducted to predict the prognosis of the patients in ovarian cancer [46].

2.4 *ESC-1 and other Pluripotency factors*

ESC-1 is a pseudonym for our gene of interest as the findings are unpublished. This acronym will be used throughout this thesis. The aim of this study was to study gene expression of ESC-1 and its partners in various cancer patients. The following findings of collaborators were the motivation for pursuing this study:

a. ESC-1 is an embryonic stem cell marker and is highly expressed in human embryonic stem cells (hESCs).

b. ESC-1 is downregulated rapidly in response to differentiation, i.e. it is expressed when cell is in an undifferentiated state.

c. ESC-1 is required for maintenance of the self-renewal capacity of the hESCs.

d. ESC-1 depletion resulted in downregulation of key markers of undifferentiated embryonic stem cells.

e. ESC-1 was highly expressed in Seminoma (Tcam2 cell line) and Embryonal Carcinoma (2012Ep cell line).

f. ESC-1 is an RNA binding protein and also binds to RNA Helicase A.

The hypothesis was that if an embryonic stem cell factor is being expressed in an adult tissue, it is not a good sign as it is not supposed to be expressed there.

2.5 *Gene Expression Quantification*

In this project we analyse the gene expression of patients across several cancer types. We chose to investigate the gene expression because the varying gene expression in different tissues is due to the varying combination of genes that are on and off in a particular tissue. It is because of this that our skin cells are morphologically and functionally so much different than the neurons in our brain, or how the muscle cells of our stomach function in ways highly contrasting to the way muscle cells in our heart function.

The identification of differential gene expression, i.e. genes having a different expression level as compared to normal tissues, is among the very first methods a bioinformatician can use to judge that is there something unusual going on in the samples under study. The gene expression data is available abundantly on various public portals.

2.5.1 *Affymetrix Platform for microarray*

Microarray is a chip with a 2-dimensional array of microscopic dots where the DNA, RNA or Protein binding probes are present. In each of the cells or individual compartments which are called “spots”, lies a chromogenic probe or an antibody that detects its target and emits a signal. This signal is detected when the microarray chip is scanned by a microarray reader which subsequently generates an image file [47]. The raw intensities from each of the spots representing a gene is retrieved by processing the image file.

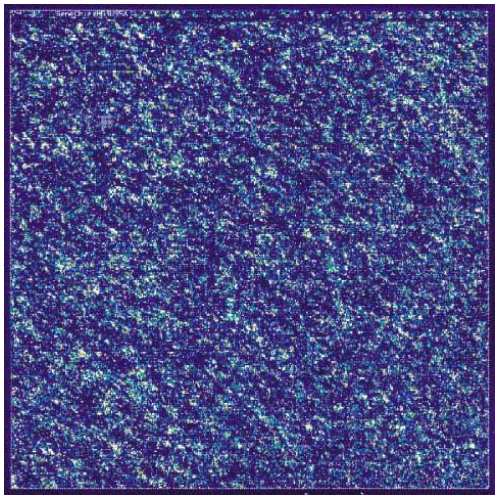


Fig 2. – A) – Microarray image file[76], showing output from an Affymetrix GeneChip, which is shown on the right **(B)** Affymetrix GeneChip Human Genome U133 Plus 2.0.[77]

Affymetrix GeneChip microarrays have probes chemically synthesized on the chip itself. Each probe corresponds to a specific region of its target gene's RNA or cDNA[47]. These probes are 25-30 bases long oligonucleotides. They are synthesized by a process called photolithography [47], more information can be found on the website of Affymetrix. For each gene there are 2 probes, one Perfect Match (PM) to the target gene and one Mismatch (MM) which serves as a control for hybridization specificity [47].

2.5.2 *Why perform gene expression analysis ?*

The basic idea behind gene expression analysis is *to identify the genes that are statistically significantly differentially expressed between two or more groups under consideration*. Differential expression implies the difference in gene expression levels between two groups of samples. The gene expression in the same tissue type varies not only from individual to individual but it varies within the same individual based on different sampling time. This is due to the fact that gene expression is not a static in nature, instead it is a dynamic phenomenon. So there are several differences that arise due to the basic biology of the individual and they might have nothing to do with the actual disease or condition we want to study. To resolve this ambiguity and to judge which genes are actually differentially expressed and are not just a consequence of natural variation we use statistics.

With the help of statistical tests we determine how likely is it to have a gene expression profile like our Gene X. If it is unlikely that we have a gene expression profile like that, we call that result statistically significant. We can use several statistical tests to do this, but most common is Student's T-test (described in section 4.6.1). Followed by this, conventionally we look at the resulting genes, and look for their annotations to identify in which biological processes they are involved and try to look for the set(s) of genes that work together, or genes from a pathway that are enriched in that particular cancer or disease under study.

Section 3) Hypothesis & Workflow

3.1 Hypotheses

3.2 Workflow

3.1 *Hypothesis*

Hypothesis:

- It is possible to differentiate between cancer and non cancer samples using ESC-1 alone.
- ESC -1 explains the survival of patient.

Alternate Hypothesis:

- It is not possible to differentiate between cancer and non cancer samples using ESC-1 alone.
- ESC -1 alone does not explain the survival of patient.

3.2 *Workflow*

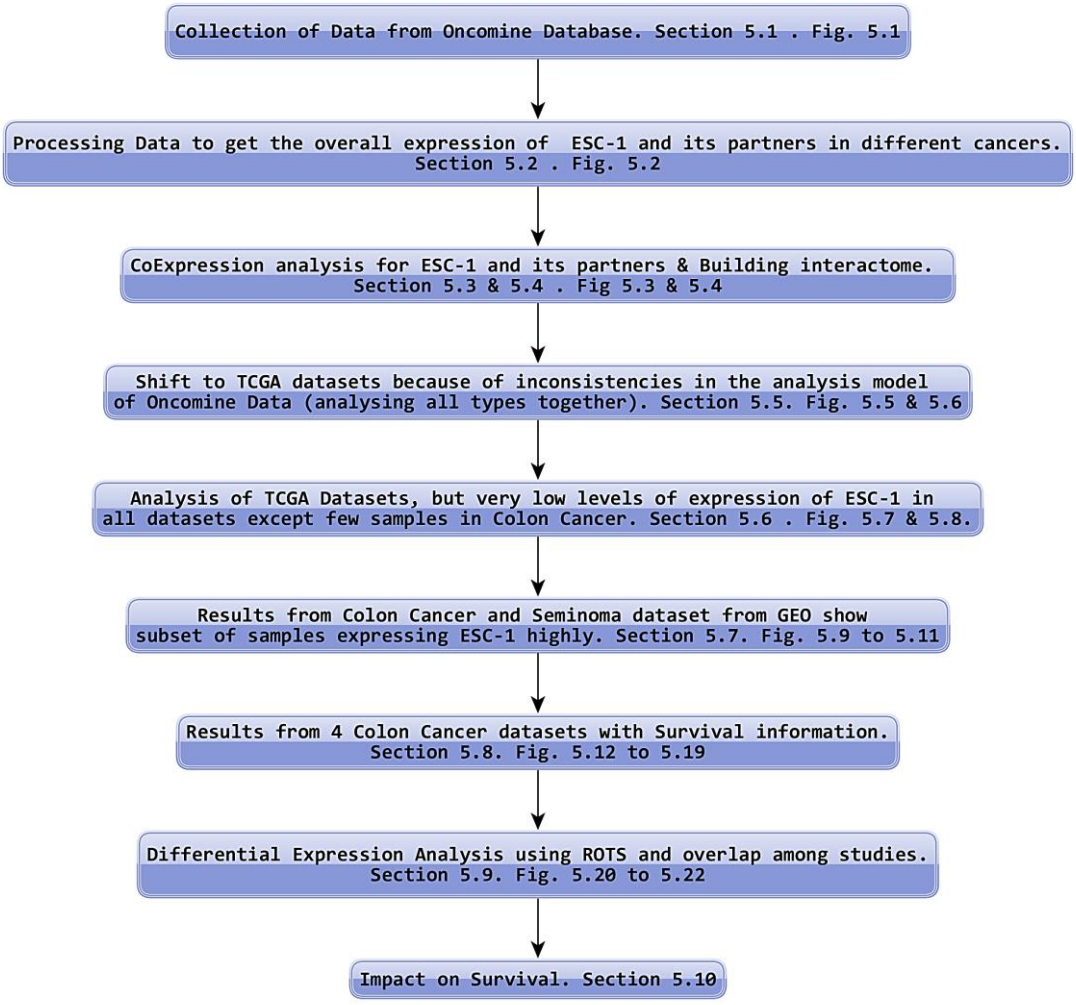


Fig. 3.1 – Flowchart showing the workflow of this project while mentioning the sections and images that correspond to that part.

This project was different from conventional gene expression data analysis projects in two main aspects. Primarily, in conventional gene expression analysis projects there is a disease of interest and we are looking for genes that are differentially expressed (between the healthy and diseased samples). Then we draw the conclusion that these genes might be involved in the pathogenesis or development of that condition/disease. Contrary to that in this project there were genes of interest and the key interest was to find cancers in which they are perturbed. Secondly, the sample size for this study was much larger as compared to a conventional microarray study. In this study over 5500 cancer samples were analysed.

A flowchart representing the workflow of this project is presented as Fig. 3.1 and a flowchart highlighting the differences between the conventional microarray analysis and this project is shown as Fig. 3.2.

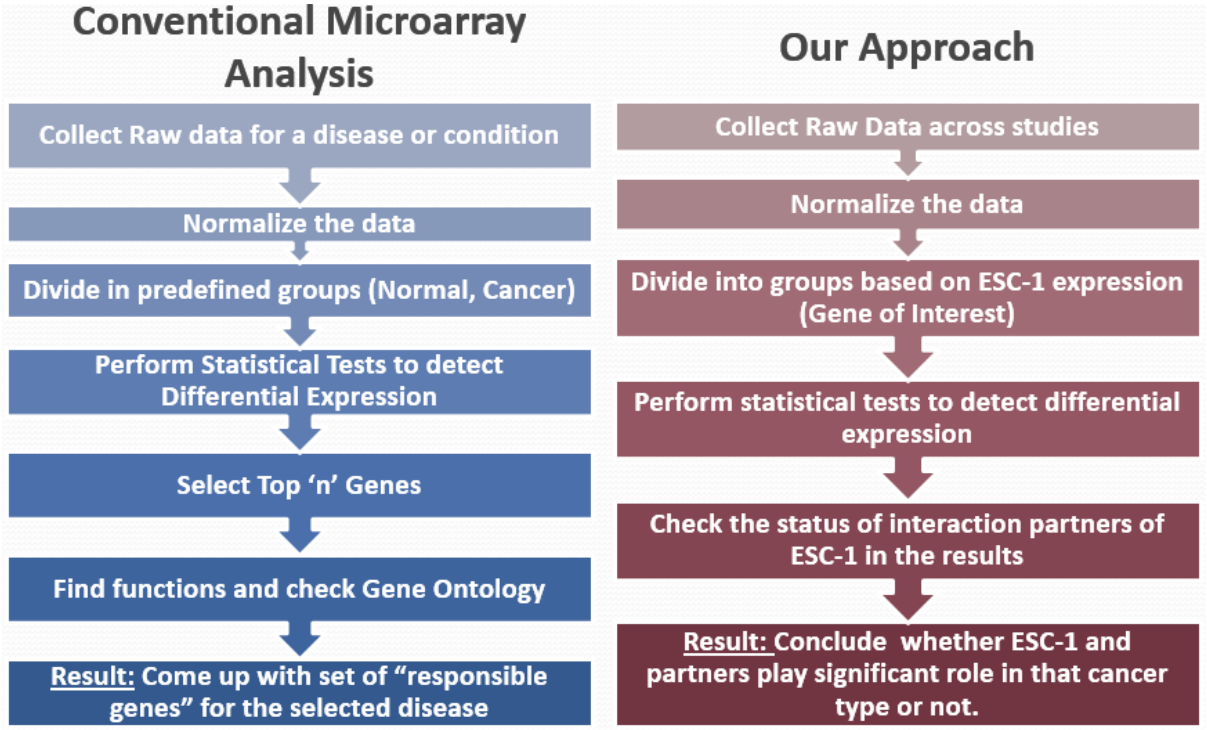


Fig. 3.2 – Flowchart showing the contrast between conventional microarray data analysis pipeline and our approach.

Section 4) Resources and Methods

4.1 *Datasets*

4.1.1 *TCGA*

The first source of datasets used was TCGA [48] (The Cancer Genome Atlas), as they had raw as well as normalized data for a large number of samples along with clinical information (e.g. Tumour stage, Tumour size and most important of all survival information). They also host RNA sequencing, SNP, Methylation, Protein Expression data along with Exome Sequence. But the prime interest in this study was in gene expression data for various cancer types.

4.1.2 *Datasets from TCGA*

Datasets that were used from TCGA in this study are enumerated in the Table IV.I below.

Dataset ID	Cancer Type	Number of Probes	Number of Samples	Tumour & Normal count	Platform
TCGA COAD	Colon Adenocarcinoma[49]	90797	174	Tumour 19 : Normal 155	Agilent 244K Custom Gene Expression G4502A-07-3
TCGA LUAD	Lung Adenocarcinoma[50]	90797	32	Tumour 32 : Normal 0	Agilent 244K Custom Gene Expression G4502A-07-3
TCGA LUSC	Lung Small Cell Carcinoma [51]	90797	153	Tumour 153 : Normal 0	Agilent 244K Custom Gene Expression G4502A-07-3
TCGA OV	Ovarian serous cystadenocarcinoma[52]	22277	586	Tumour 586 : Normal 8*	Affymetrix HT Human Genome U133 Array Plate Set
TCGA OV	Ovarian serous cystadenocarcinoma[52]	90797	588	Tumour 588 : Normal 8*	Agilent 244K Custom Gene Expression G4502A-07-3
TCGA GBM	Brain Glioblastoma [53]	22277	519	Tumour 519 : Normal 0	Affymetrix HT Human Genome U133 Array Plate Set
TCGA GBM	Brain Glioblastoma [53]	90797	89	Tumour 89 : Normal 0	Agilent 244K Custom Gene Expression G4502A-07-1
TCGA GBM	Brain Glioblastoma [53]	90797	473	Tumour 473 : Normal 0	Agilent 244K Custom Gene Expression G4502A-07-2

Table IV.I – Table listing the datasets from TCGA with the details like sample count, probe count.

4.1.3 *GEO*

Gene Expression Omnibus [54], [55], generally referred by its abbreviation GEO, is a database for Gene expression data hosted by the National Centre for Biotechnology Information at the United States of America. GEO provides access to raw data (or normalized at some occasions depends on what authors have provided). However the additional clinical information about the samples is rarely provided by the authors at GEO, which is partly due to the fact that GEO requires the data to comply to MIAME (Minimum Information About a Microarray Experiment) standards, which do not have the clinical information as mandatory. Therefore most authors choose to leave it out. GEO is one of the most widely used portal for the retrieval and submission of Microarray datasets[54]. There are datasets from several experiment types like mRNA Expression, miRNA and siRNA studies from several organisms and conditions.

4.1.4 *Datasets from GEO*

Datasets that were used in this thesis from GEO are listed below in Table IV.II. Nicknames are assigned to the datasets for this thesis. This was done specifically for the Colon cancer datasets, as there were 5 of them.

Dataset ID	Cancer Type	Number of Probes	Number of Samples	Tumour & Normal count	Platform	Dataset Nickname in Thesis
GSE3218 [56]	Seminoma	22283	107	Tumour 101 : Normal 6	Affymetrix HG U133	Seminoma
GSE8671 [57]	Colon Adenoma	54675	64	Tumour 32 : Normal 32	Affymetrix HG U133 Plus 2.0	Colon1
GSE2109 [58]	Expression Project for Oncology	54675	2158	Tumour 2158	Affymetrix HG U133 Plus 2.0	ExPO
GSE14333 [59]	Colon Cancer	54675	290	Tumour 290	Affymetrix HG U133 Plus 2.0	Colon2
GSE17536 [60]	Colon Cancer	54675	177	Tumour 177	Affymetrix HG U133 Plus 2.0	Colon3
GSE33113 [61]	Colon Cancer	54675	96	Tumour 90 : Normal 6	Affymetrix HG U133 Plus 2.0	Colon4
GSE17537 [60]	Colon Cancer	54675	57	Tumour 57	Affymetrix HG U133 Plus 2.0	Colon5

Table IV.II – Table listing the datasets from GEO with the details like sample count, probe count and the nickname used for the datasets later on in the thesis.

4.2 *Normalization*

Microarray experiments are susceptible to variation due to the fact that there are several sources that may affect the quantification of gene expression. Normalization is the process which tries to compensate for such variation. In other words it removes, or rather attempts to remove the variations that can be attributed to factors other than biological variation itself [62]. It is also worth mentioning that the various factors affect different experiments to different extents as all of the runs are independent events. Thus, it becomes even more important to normalize the data so that the different microarrays are comparable to each other. Normalization also attempts to bridge the gap between different platforms [62]. However, there are separate methods for implementing those. As the datasets in this work were analysed independent of each other and at the results were compared at the end, it is not relevant to discuss them in-depth here.

Some sources or variation (other than biological phenomenon) that might affect the microarray data are dye bias (not required on 1-channel array e.g. Affymetrix), scanner malfunction, batch effect and array design. Also another very significant issue is the variation introduced by the experimenters themselves.

4.2.1 *RMA Normalization*

Among the assortment of normalization methods, RMA (Robust Multiarray Average) was chosen in this study as the method for normalizing the arrays from Affymetrix platform as it performs global background correction as well as does quantile normalization across-array [63]. RMA also fits a linear model per probe set which removes probe-specific affinities. It makes the distribution of the expression values across the arrays identical, and thereby making them more comparable to one another [63].

Normalization of the microarray expression data from the Affymetrix platform was performed using the “affy” package [64], [65] from Bioconductor. In case of the TCGA datasets, the normalized expression values were downloaded directly from the TCGA website[48]. This was done keeping in mind that it would be easier for other researchers to replicate the results. Otherwise there are a lot of small things like the version of the normalization software, the parameters etc. that can influence the results.

The “affy” package [65] in Bioconductor provides functions to read and analyse the

raw microarray data files, having the extension .CEL. The basic process involves reading the .CEL files stored in a directory using the `ReadAffy()` function, followed by normalization using the `rma()` function. There is also another option to use the `justRMA()` function which does not require using the `ReadAffy()` function instead, it can be invoked directly once the current working directory is set to the location where the .CEL files are stored. The function `justRMA()` was used because it is an implementation of RMA method in C programming language, which is fast as well as uses lesser primary memory (RAM) as compared to using the `ReadAffy()` → `rma()` [65].

```
# Installing affy package
source("http://bioconductor.org/biocLite.R")
biocLite("affy")
library(affy)

# Normalizing the expression values
# Also reads the .CEL files from the current working directory
result<-justRMA()
dat2<-exprs(result)
```

4.2.2 *UPC Normalization*

UPC (Universal exPression Code) is a normalization method which uses a mixture model to estimate the activation status of a Gene in a sample [66]. The mixture model that UPC uses is made of two components 1) Background noise and 2) Background noise + Biological signal [61]. UPC also corrects for platform-specific background noise. As UPC scores are derived from the information within a sample theoretically, it should perform equally well if ran on the whole dataset all together or on individual samples or even on batches of samples for that matter. These claims made in the publication seemed to hold well during the analyses. Therefore, UPC was chosen as another approach to normalize the data.

UPC produces standardized scores or UPC values which are on a 0 to 1 continuous scale, where the lower value indicates that it is likely to be in the background while a higher value indicates that the gene is transcriptionally active [66]. UPC score represents the probability that a gene is expressed in that sample, and in the publication they classify UPC > 0.5 as active and UPC ≤ 0.5 as inactive[66]. Using UPC has an added advantage that samples from different profiling techniques can be compared e.g. RNA-seq and Microarray. Additionally, it computes the background noise on a per-sample basis, thereby making it

possible to account for sample-specific bias[66].

It is noteworthy to shed some light on the contrast between RMA and UPC scores. As evident from the density distribution of RMA scores (Fig. 4.1 A) a lot of the samples are accumulated in the medium region (between 6-8.5), which is where maximum of the background noise lies in an Affymetrix signal. However, exactly the opposite of that is happening in the case of UPC scores (Fig 4.1 B), where there are very few samples in the middle region.

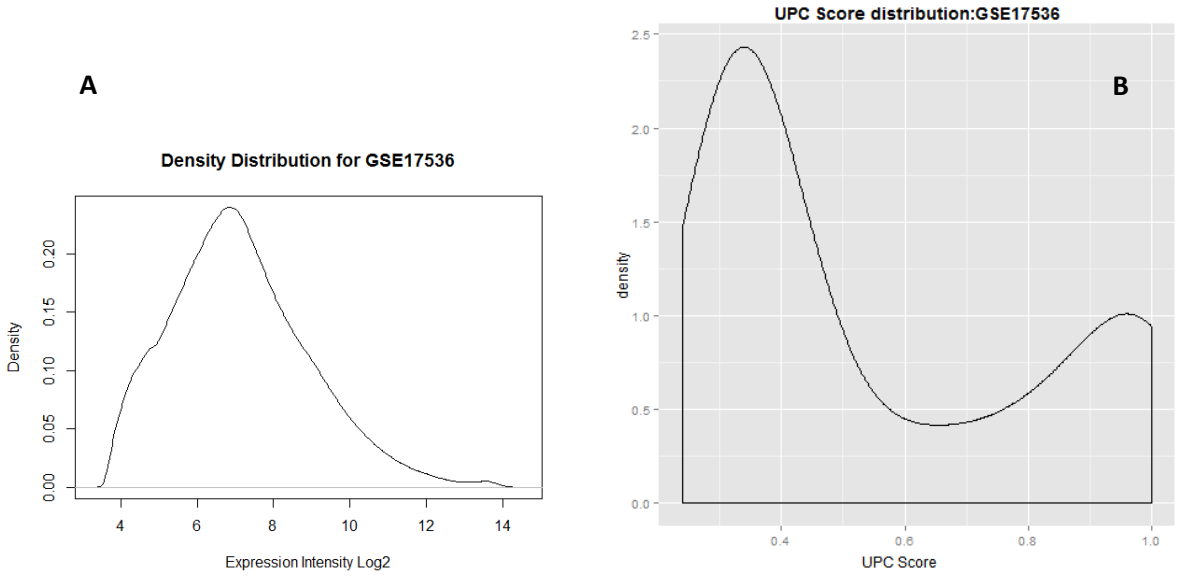


Fig. 4.1 – The density distribution plots for a gene in samples of dataset GSE17536 **A)** RMA normalized intensity for ESC-1. **B)** UPC scores for the same gene.

```
# Computing UPC scores
library(SCAN.UPC)
UPCresult = UPCfast(celFilePattern="*.CEL")

# UPCresult is the object with UPC scores
# Retreiving UPC scores as expression set

mat=exprs(UPCresult) mat=mat[,sort(colnames(mat))]
```

UPC however, is a very processor and primary memory (RAM) intensive normalization method, much more than RMA. Therefore, I divided the datasets into batches of 30-50 samples and then ran UPC and later on assembled the results (UPC scores for all the probes/genes in each sample) from those individual runs into a matrix for one dataset.

4.3 *Statistical Terms*

Some of the statistical terms are explained here with small examples, to facilitate quick understanding or recap.

4.3.1 *Mean, Median & Quartiles*

Let us consider a set V = {$0 \leq x \leq 100$, $x \epsilon N$} so that "V" is a set of Nautral numbers from 0 to 100. Now **mean** (μ) is the average of observations i.e. *Sum of observations divided by total number of observations*. In this case the mean is 50. The **median** of a set of observations is *the observation that divides the data into two equal sized halves*, i.e. lower half which comprises of set of observations less than it and upper half which comprises of observations higher than it [67]. If the set of observations is finite and can be written in an ordered fashion then the median would be the middle one. For set V, the median is 50. This can only be done when there are odd number of observations, but in case of even number of observations there is never going to be a middle element. In that case the median is the mean of the two centre-most elements.

Quartiles are the statistical quantities that divide a *dataset into quarters* ($1/4^{th}$) i.e. into *four groups of equal sizes* just like the median divides into two groups of equal sizes [67]. First Quartile (Q_1) is the middle observation between the smallest observation of the set and the median, Median is the second Quartile (Q_2), and the Third Quartile (Q_3) is the middle observation between the median and the highest observation. For our set V the Q_1 is 25 , Q_2 also known as median is 50 and Q_3 is 75. **Inter Quartile Range** (IQR) is the difference of Q_1 and Q_3, IQR =Q_3 - Q_1. In our case IQR is 50.

Let us redefine our variable V = {$1 \leq x \leq 100$, $x \epsilon N$} so that "V" is a set of Natural numbers from 1 to 100. Now this has 100 observations as opposed to 101, as in the previous definition of set V. It can be seen how this affects the values of our statistics. μ is $\frac{50+51}{2} = 50.5$. The Quartiles also change in the same manner, Q_1 is 25.75 and Q_3 is 75.25 and IQR is 49.5.

Median is more resistant as compared to mean to the presence of outliers in the data. The reason is that mean computes the sum of the observations (i.e. includes those extreme observations). However, median orders the observations and picks the middle one, and as the order is unaffected by the extreme observations, median manages to stay robust.

4.3.2 *Standard Deviation and Variance*

Standard Deviation (σ) is a statistical quantity representing the amount of dispersion of the data around mean [67]. Low σ means that observations tend to be scattered very close to the mean (μ), and a high σ means that the observations are scattered far away from the mean. The Standard Deviation is defined as: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$, where n is the number of observations in sample and μ is the true mean of the population. If population mean is used then we get Population Standard Deviation. However, in most cases we do not know the true mean μ and we are bound to use the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{i=n} x_i$ (where n is the number of observations in sample). The Sample Standard Deviation can then be calculated as: $\sigma_s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\hat{\mu})^2}{n-1}}$, where n is number of observations in sample and $\hat{\mu}$ is sample mean. Variance is the Square of Standard deviation i.e. $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\hat{\mu})^2}{n-1}$.

4.4 *Plots and Diagrams*

Some of the plots and diagrams used in this thesis are discussed here along with their features to facilitate better comprehension of those images and to instigate critical thinking.

4.4.1 *Boxplot*

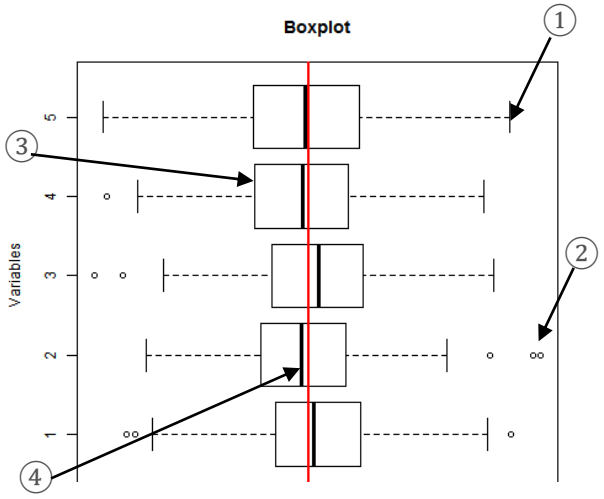


Fig. 4.3 – Boxplot of pseudo-random data. 1-Whisker, 2-Outlier, 3-Box , 4-Median.

Boxplot is a graphical way to represent a dataset through the use of quartiles [67]. In the example boxplot (Fig. 4.3) of random data the variables are on the Y axis and the value of the observations are on the X axis, (this orientation does not matter it can be the other way around as well). Here we have a pseudo-randomly generated matrix with dimensions 200 x 5. The values are drawn from the standard normal distribution [67] with mean μ = 0 and standard deviation σ =1. Now in the individual box plot for each variable we have a box marked with *First Quartile* (Q_1) forming the lower bound and the *Third Quartile* (Q_3) forming the upper bound. The thick line within the box is the Median (Q_2). Dotted lines radiating from the box are called "*whiskers*". The

whiskers generally extend from the nearest quartile to 1.5 times the IQR (or till the maximum value, if it lies within Quartile + 1.5 x IQR). However, if there are observations outside this mark, then they are denoted by circles and those observations are called "*outliers*". The red line here is drawn to show the mean of the whole data, unlike the median which is calculated for each variable. It is an additional statistical feature in this plot which is conventionally not drawn in a boxplot.

4.4.2 *Heatmap*

Heatmap is a 2D representation of a matrix, with cells organized in rows and columns. The heatmap allows us to see in a single plot how each observation in a particular sample fares with respect to others of the same variable as well as other variables. Fig 4.4 shows an example of a heatmap for a data matrix (10 x 10) which was generated pseudo randomly from standard normal distribution with the $\mu = 0$ and $\sigma = 1$. The colour key shown in the top left corner can be seen and each cell representing an observation in the matrix. Rows were named to be samples, so that it is analogous to a gene expression matrix, where there are samples on one margin and genes (represented by probes) on the other margin. We can also see the dendrogram which is a representation how the rows cluster if we use hierarchical clustering. Hierarchical clustering makes a cluster by two approaches 1) Agglomerative, i.e. keep adding samples one by one in a cluster till the algorithm has incorporated all and made a cluster containing all samples, or 2) Divisive, i.e. the algorithm keeps removing samples from one large cluster, until all the samples are removed. The order of incorporating or removing the samples is represented using a dendrogram (tree like structure with binary splits).

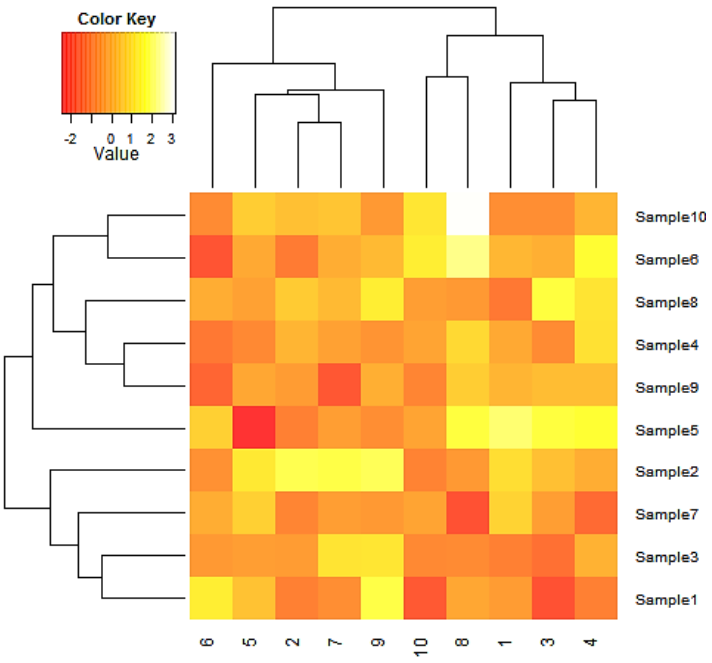


Fig. 4.4 – Heatmap of a pseudo-random data matrix with colour key and dendrograms (showing hierarchical clustering of rows, as well as of columns).

4.4.3 *Scatterplot*

Scatterplot is another graphical tool used to visualize each observation in a dataset on two orthogonal dimensions. It plots points, representing each observation, while representing its value in the variables under consideration on the axes. For an example of a scatterplot see Fig. 4.6 (under section 4.5.1 Principal Component Analysis).

4.4.4 *Venn diagram*

Venn diagram is from Set Theory [68], and shows all possible logical relations between sets under consideration (generally speaking, they are finite sets). The conception of Venn Diagrams and their use in mathematics is attributed to John Venn, a British Philosopher and logician.

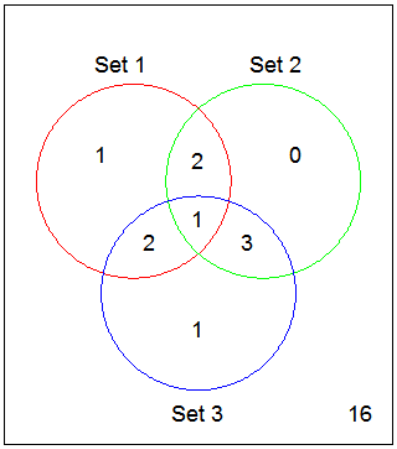


Fig. 4.5 – Venn diagram for three sets.

Let us define:
Set 1 = {"a", "b", "c", "d", "e", "f"};
Set 2 = {"d", "e", "f", "g", "h", "i"};
Set 3 = {"a", "b", "f", "g", "h", "i", "j"};

And let us consider the universal set (ξ) made of all lower case English alphabet, which are 26 in number. Now the Venn diagram (Fig. 4.5) illustrates that there is one letter that is present in all the 3 sets, written as $Set\ 1 \cap Set\ 2 \cap Set3$, and if one examines the three sets they can see that it indeed is the letter "f". There are two letters common between Set 1 and Set 2, "d" and "e", that is not common with Set 3. It can be written as $(Set\ 1\ \cap Set\ 2) \setminus Set3$ [68].

A table with common symbols used in set theory is presented here as Table IV.III.

Symbol	Operation Name	Description of operation
∑	Universal Set	Contains everything under consideration.
ϵ	Element of	1 ϵ Set A, implies 1 is a member of Set A.
∩	Intersection	A ∩ B, implies, common elements of A and B.
∪	Union	A ∪ B, implies all unique elements of A and B.
-	Difference	A - B, means unique items of A that are not in B.
‘	Complement	A' means, Everything but not A.
\	Relative complement	A \ B means elements in A that are not in B. It is the same as the Difference operator "-".

Table IV.III – List of commonly used symbols in Set theory and their mathematical interpretation.

4.5 *Dimensionality Reduction*

Gene expression experiments deal with tens of thousands of probes and even hundreds of samples. In this case visualization of the data in such high dimensions is not possible. This is because we are limited by the 3 dimensional space we live in. In such a scenario we require methods which scale down the data while preserving as much variance as possible. Dimensionality reduction [69] is the mathematical technique which is used to do so. There are several methods for this, but I used Principal Component Analysis and implemented it in R.

4.5.1 *Principal Component Analysis*

Principal Component Analysis (PCA) reduces the dimensionality of the data by applying orthogonal linear transformation [69]. The objective of this method is to find Principal components (or dimensions) which represent the maximum amount of variance of the data, in simpler words, it aims at finding the attributes that can account for most variation or differences in the dataset. The unique property of these Principal components (PCs) is that they are orthogonal to each other, i.e. perpendicular. Each PC represents a specific proportion of the Total Variance in the data [69] and it is orthogonal to its preceding components (to make sure that there is no correlation). The number of PCs that we get is less than or equal to the number of variables. In large datasets with thousands of variables the first two components are chosen to represent the scaled down version of the data because they are the two components explaining the largest proportion of the total variance of the data. When the number of PCs approaches the number of variables all the variance in the data is accounted for by all the preceding components.

If we want to get a bird's eye view of the data and find out if there is any organization in the data, we can use PCA as it is an unsupervised method. If there is a correlation/clustering, among the samples then it will be visible. However, if it is none, then there will be no organization among the samples in the plot.

We can understand this by an example. There is a dataset called 'iris' in R (it can be loaded using the command `data(iris)` which contains 150 observations of Sepal length and Width and Petal Length and Width. It has 50 flowers each from the three species of *Iris* genus viz. *I. setosa*, *I. versicolor* and *I. virginica*. This dataset was collected by Sir Ronald Fisher in the year 1936 [70].

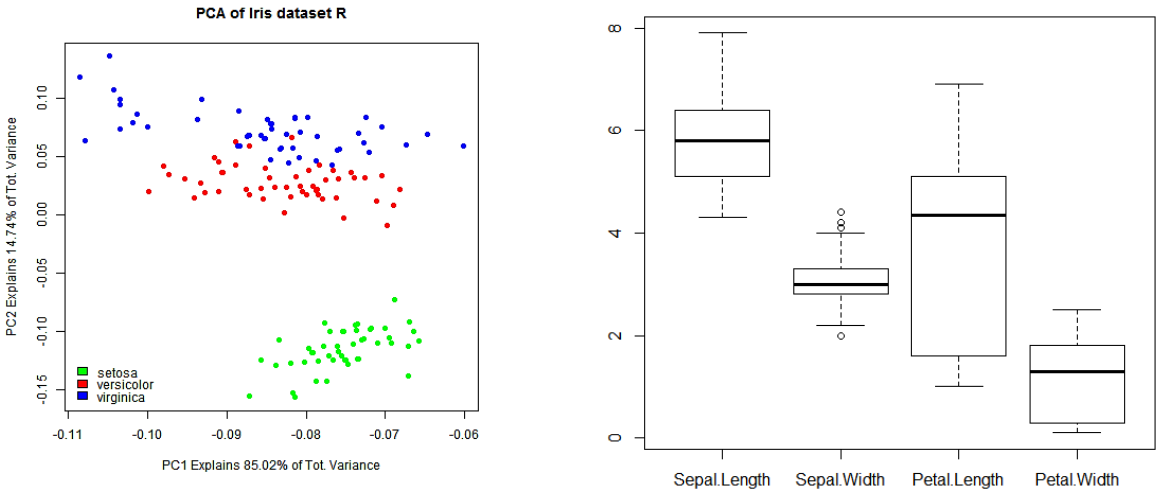


Fig. 4.6 – A) Scatter plot of PCA Eigen values on the First two Principal Components of the iris dataset. The colours represent the species of the flower, legend on the bottom left of the plot. **B)** Boxplot of the 4 parameters in the iris dataset (for all 150 samples). Y-axis values are in centimetres.

In the scatter plot (Fig. 4.6 A) we can see that the first two components are explaining the differences well. The *I. setosa* species are separated very well from the other two. The plot also suggests that there is not such a clear difference between the other two species. Although most of them are well separated with just a few samples in the twilight zone. But we must remember that this is a separation just based on four parameters. We can also have a look at the box plot (Fig. 4.6 B) of these four parameters to visualize the data space and values of parameters.

PCA does not tell us what the different components mean or represent, but we can attempt to find it out. PCA on the iris dataset (four numeric parameters) returns us four components, the standard deviations of which are:

```
> pca$sdev #prcomp returns standard deviations of PCs as sdev.
PC1             PC2             PC3             PC4
2.365402e+01    9.850791e+00    1.224728e+00    3.266697e-15
```

Now we can compute the variance by squaring these standard deviations, the variances are:

```
> (v=pca$sdev^2) #Std. Dev is square root if Variance, therefore squaring Std.Dev.
PC1             PC2             PC3             PC4
5.595128e+02    9.703808e+01    1.499959e+00    1.067131e-29
```

The proportion of the total variance that each component explains is:

```
> v*100/sum(v)
PC1             PC2             PC3             PC4
85.02577        14.74629        0.2279398       1.621654e-30
```

Now we should look at the variance of the four parameters of the iris dataset and the percentage each makes up for the total variance.

```
> (var2=apply(iris[,1:4],2,var))
Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
   0.6856935    0.1899794       3.1162779       0.5810063

> apply(iris[,1:4],2,var)*100/sum(var2)
Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
   14.994532    4.154411        68.145793    12.705264
```

We can see that Petal Length represents the most variance in the dataset followed by Sepal Length which in turn is closely followed by Petal Width. This is also evident from the Boxplot (Fig. 4.6 B). But a key thing to note here is that amount of variance represented by the Principal Components does not necessarily comply with the proportion of variance the individual variables represent in the dataset e.g. here the PC1 represents 85% of the total variance of the dataset but the maximum variance is represented by any of the actual parameters is 68%. And most often we encounter this kind of a situation with gene expression datasets, where a particular PC comprises of different proportions of several variables and not just one. However, we can conclude that the Petal Length forms a large portion of the PC1 here.

4.6 *Differential Expression Analysis*

Genes in living systems have a lot of heterogeneity when it comes to their activity and expression level. The differences in these levels give rise and sustain the complex system Life is. All cells in our body have the complete DNA that is unique to us, but each tissue functions differently, and therefore has different gene expression. It is an aggregated effect of these differences that, for instance makes cells and tissue in our intestines work in a way different than our heart which in turn works in a completely different way as compared to our Brain. Apart from these tissue specific differences, the genes are not always expressed exactly at the same level in our tissues, instead the expression changes depending on the task the tissue is

doing or if there is any disorder then there is a difference due to that. This gives us two scenarios when the gene expression in a tissue undergoes variation.

1) Under normal circumstances the gene expression varies, but stays within particular bounds, and there are no drastic changes in the expression under healthy conditions.
2) In the case of a disease or a condition like cancer there are changes of different degrees in the gene expression as a lot of mechanisms that are involved in controlling and regulating it are disrupted.

A common method to detect if there is anything different going on in a tumour tissue as compared to a healthy tissue is to do a Differential Expression (DE) analysis [71] using microarray of RNA sequencing data. Generally, the observations are compared taking into account the range of the expression of a particular gene in different individuals under the same condition (i.e. either healthy, or affected by the same disorder). The most popular method is using a *T*-test to compare the means of groups under study [67].

4.6.1 *T-test*

The *T*-test is a parametric statistical test which assumes the data has a Student's *T* distribution which is very close to Normal distribution (also called Gaussian distribution) in properties [67]. The Null hypothesis (H_0) here when doing DE analysis is that the means of the two groups under consideration are equal (i.e. $\mu_1 = \mu_2$), the alternate hypothesis (H_1) is that the means are not equal (i.e. $\mu_1 \neq \mu_2$). This is done by computing the test statistic called t-statistic in this case as: $T = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$, where the μ_1 and μ_2 are means of two groups under study and σ_1^2 and σ_2^2 are the variances of the two groups under study. The number of samples in two groups is denoted by n_1 and n_2. Then, we compare the values of "*T*" with a *T*-distribution with $n_1 + n_2 - 2$ degrees of freedom to find out the *P*-value. More about this can be found in a statistics course book[67].

After performing the *T*-test we get the *P*-value, which tells us how significant the result is. The p-value is on a scale from 0 to 1, the closer the p-values is to 0, the more significant the differences between the two groups are. *The smaller the P-value, the more significant is the result*. Generally speaking we consider a result significant if the *P*-value < 0.05.

When doing DE analysis, conventionally there are tens of thousands of probes in a

dataset and we encounter a problem which is called "multiple-testing" problem in statistics because we are making so many comparisons. One needs to control for this problem when doing analysis. In this study the FDR (False Discovery Rate) method of *P*-value correction was used.

4.6.2 *ROTS (Reproducibility Optimized Test Statistic)*

ROTS is a test statistic developed to retain the good features of the *T*-test while neutralizing the fact that it assumes the data to have a certain distribution (i.e. parametric) [71]. ROTS facilitates the optimization of a test statistic of *T*-type based on the data itself. Let's consider a gene having its expression levels in two groups as $x_g = \{x_{g1}, x_{g2}, \ldots, x_{gn_x}\}$ and $y_g = \{y_{g1}, y_{g2}, \ldots, y_{gn_y}\}$. The sample mean is defined as $\bar{x}_g = \frac{1}{n_x}\sum_{i=1}^{n_x} x_{gi}$ and the sample variance as $\sigma^2 = \frac{1}{n_x - 1}\sum_{i=1}^{n_x}(x_{gi} - \bar{x}_g)^2$.

ROTS computes the test statistic, for a gene *g* as $d_\alpha(g) = \frac{|\bar{x}_g - \bar{y}_g|}{\alpha_1 + \alpha_2 . \sigma_g}$ where the estimated parameters are $\alpha_1 \in \{0, \infty)$ and $\alpha_2 \in \{0,1\}$. $\bar{x}_1$and $\bar{x}_2$ are means of the two groups for the gene g, and σ_g= standard deviation. In case $\alpha_2 = 0$, the statistic $d_\alpha(g)$ is essentially the same as the signal log-ratio if logarithmic expression values are used. Setting α_2 = 1 gives the SAM-statistic with a regularization constant α_1. The standard T-statistic is a special case where $\alpha_1 = 0, \alpha_2 = 1$.

ROTS returns FDR corrected *P*-value, which are calculated by randomly permuting the sample labels. Also ROTS computes the Reproducibility denoted as $R_k(d_\alpha)$, which is the average overlap of *k* top-ranked genes over several bootstrapped datasets. 500 bootstrapped datasets were considered in this work. The reproducibility statistic indicates how robust the findings are. ROTS is available for R [72] on MacOS and Windows from the website of University of Turku at:
http://www.utu.fi/en/units/sci/units/math/Research/biomathematics/projects/Pages/rots.aspx.

4.7 *Co-Expression analysis*

A method to test the co-expression for a set of genes in-silico is to test the correlation between the expression levels of genes across samples. If they have a high positive correlation then it is likely that they are being co-expressed. But it should be kept in mind that it is an

indirect quantification of co-expression, and actual co-expression can be determined by co-staining samples for the two proteins or mRNA transcripts.

There can be several types of correlation measures [67] between observations. In this study, *Spearman Rank Correlation was chosen because it is a non-parametric method* and it is unaffected by the magnitude of the observation. *It Ranks the observations and computes correlation* based on the difference in ranks of the observations. It signifies the statistical dependence of one variable over the other. Spearman Rank correlation coefficient is denoted by Greek letter rho (ρ), as $\rho = 1 - \frac{6(\sum_{i=1}^{n} d_i^2)}{(n^3 - n)}$, where n is number of observations and d_i is the difference in ranks of the observation.

	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student 7	Student 8	Student 9	Student 10
X	52	67	75	64	89	76	72	82	53	58
Y	86	79	76	70	52	82	87	60	58	90
Rank x (Rx)	1	5	7	4	10	8	6	9	2	3
Rank y (Ry)	8	6	5	4	1	7	9	3	2	10
d=|Rx-Ry|	7	1	2	0	9	1	3	6	0	7
d^2	49	1	4	0	81	1	9	36	0	49

Table IV.IV – Table for example of calculating Spearman Rank correlation populated with fictitious data. X and Y are random integer observations, *Rx* and *Ry* are ranks of the observation in that variable, with rank 1 for the lowest and rank 10 for highest valued observation.

The calculation of rho is demonstrated with an example shown as Table IV.IV. Here we have computed the d_i^2 for each student (see last row in Table IV.IV). Now we can calculate $\sum_{i=1}^{n} d_i^2 = 230$, now filling this in the formula for rho ρ we get:

$$\rho = 1 - \frac{(6 * 230)}{(10^3) - 10} = 1 - \frac{1380}{990} = -0.39394$$

By this we learn that there is a *negative correlation* between X and Y in our example set, which means that as value of X increases the value of Y tends to decrease in this example set (Table IV.IV). *Correlation can be between -1 to +1 with 0 implying there is no linear correlation between the observations* [67].

4.8 *Annotating the data matrix after normalization*

The data matrix for a typical microarray dataset is a two-dimensional matrix with rows representing genes and columns representing samples. Conventionally the column names are the sample names and the row names are the Probe IDs. As mentioned earlier in section 2.5.1, probes correspond to genes which are the focus of studies in almost all cases. But checking the gene which a particular probe detects can get cumbersome therefore two columns were added to the data matrices, one for the official gene symbol and one for the gene name. Both of these information are provided by the manufacturer of the chip and are freely available from platform descriptions on GEO as well as in R through Bioconductor. R annotation databases were used to annotate as follows.

```
# Computing UPC scores
library(SCAN.UPC)
UPCresult = UPCfast(celFilePattern="*.CEL")

# Retreiving UPC scores as expression set

mat=exprs(UPCresult) # UPCresult is the object with UPC scores
mat=mat[,sort(colnames(mat))]

# Installing the annotation library corresponding to the platform, in the case of
the Affymetrix datasets, all except few were from Affymetrix Human GeneChip U133
plus 2.0 array therefore the database “hgu133plus2.db”

source("http://bioconductor.org/biocLite.R")
biocLite("hgu133plus2.db")

library(hgu133plus2.db);
symbol<-gsub("\'", "", data.frame(unlist(as.list(get(paste("hgu133plus2",
"SYMBOL", sep="")))))[rownames(mat),])
genename<-gsub("\'", "", data.frame(unlist(as.list(get(paste("hgu133plus2",
"GENENAME", sep="")))))[rownames(mat),])
genename <- gsub("#", "", genename);symbol <- gsub("'", "", symbol);genename <-
gsub("'", "", genename)

# The annotated matrix is assembled as a data frame because in R matrix can only
have one type of elements while data frame can be composed of several types of
objects.

annotated_mat=data.frame(symbol, description=genename, mat)
```

4.9 *Filtering out Genes from a Heatmap*

Occasionally, the heatmaps are composed of genes whose expression levels are homogenous across the dataset, which although give some information but occasionally cause problems with clustering and visualizing the results as they make it more crowded. In this study when such situation (refer Fig 5.13) was encountered, the genes/probes in the heatmap were

filtered based on the standard deviation, σ, so that we eliminate the ones that are uniform (uniformly high, low or medium, i.e. any gene that was homogenous). Filtering was performed using two different cut-offs. First, the σ_g, the standard deviation for each gene g was calculated based on the UPC scores (using the same matrix which was used to create the heatmap). Then the genes were then filtered keeping the genes (rows in the matrix) that were having $\sigma >$ median (σ_g) or then those that were having $\sigma > Q_3$ (σ_g). Below there is an example to explain this process. In the Table IV.V we have UPC scores for genes for samples and standard deviations for each row computed.

	Sample 1	**Sample 2**	**Sample 3**	**Sample 4**	**Standard deviation σ_g**
Gene 1	0.66	0.58	0.68	0.62	0.044347116
Gene 2	0.15	0.14	0.86	0.75	0.38370996
Gene 3	0.5	0.3	0.75	0.92	0.273053719
….	…	…	…	…	…
Gene Nth	0.79	0.88	0.76	0.92	0.075

Table IV.V – Example Table the UPC scores for genes (as rows) for samples (as columns) and the standard deviation, σ, computed for each gene (i.e. each row), as a measure of heterogeneity among the samples for that gene. The higher the standard deviation, the more heterogeneous the gene's expression is.

Now if we compute the median (σ_g) we get a value, and if the σ for a gene (represented by row) is higher than the median then we keep that row (gene) as it implies that there is heterogeneity in that gene's expression among the samples.

4.10 *Software used*

Majority of the computation in this project was done using **R and Bioconductor**. Text processing e.g. editing sample names or gene names for convenience was done using **Python**. The input for python scripts (generally they are probe ids or sample names that need to be fixed like removing a repeating part to make them shorter) is written on disk to a file and then python is called from that directory executed with a structure like `system("python script.py")` and then python code is written to process and store its output to a file which is then read in R and then the variable that was to be edited is updated with the new values.

N.B. As the work in this thesis is yet to be published, the source code is not being released.

Section 5) Results & Discussion

5.1 *Oncomine Database and initial leads for ESC-1 differential expression*

The Oncomine database showed at a glance (Fig. 5.1) the expression of ESC-1 in different datasets and it can be seen in which of those datasets it is differentially expressed between groups. The Oncomine search for ESC-1 showed that colorectal cancer and seminoma would be interesting starting point for the investigation.

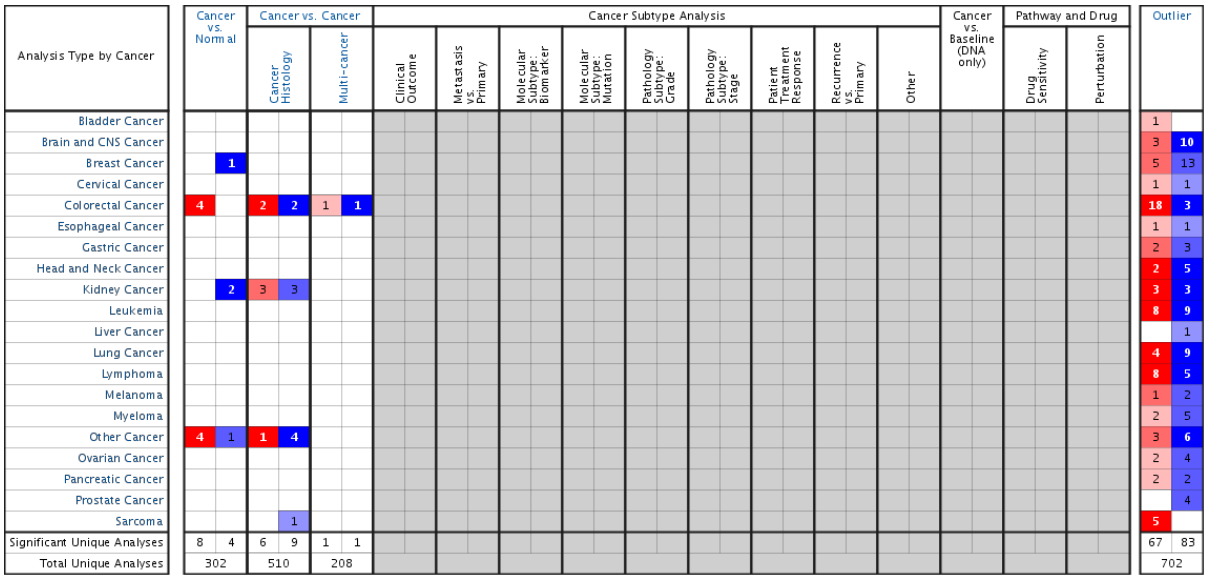


Fig. 5.1 – Summary View of Oncomine for ESC-1. Red represents studies where the gene is highly expressed and blue where gene is lowly expressed. The numbers are representatives of the number of different studies/analyses conducted with the dataset and not number of datasets, i.e. there are studies based on subtype and survival status on the same dataset.

5.2 *Overall expression of ESC-1 and its interactome from Oncomine data*

Before starting investigating individual datasets and analysing the expression of ESC-1 it was informative to see how its interaction partners behave in different cancers and for that the data in Oncomine [73] was retrieved for ESC-1 and 18 of its interaction partners. These interaction partners were chosen from a list of 306 validated interaction partners, by our collaborators. The complete 306 partners were used for investigation later on with the GEO datasets as in Oncomine the data had to be retrieved manually and it was not feasible on a large scale.

In the heatmap (Fig. 5.2) the expression pattern of ESC-1 and its partners across several cancer types can be seen. The colour red corresponds to high expression and blue corresponds to low and wheat colour corresponds to medium. White regions indicate missing data as some of the datasets for a particular cancer organ/tissue type did not contain probes or measurement for

that particular gene. In the heatmap were cluster of the genes based on similarity of expression profile. The samples were kept in original order so that the samples from same cancer type were together. It can be seen that there is a lot of heterogeneity within Colon Cancer group and which was the motivation to look into this cancer type. This heatmap (Fig. 5.2) represents a summary from 4614 samples across multiple datasets and cancer types.

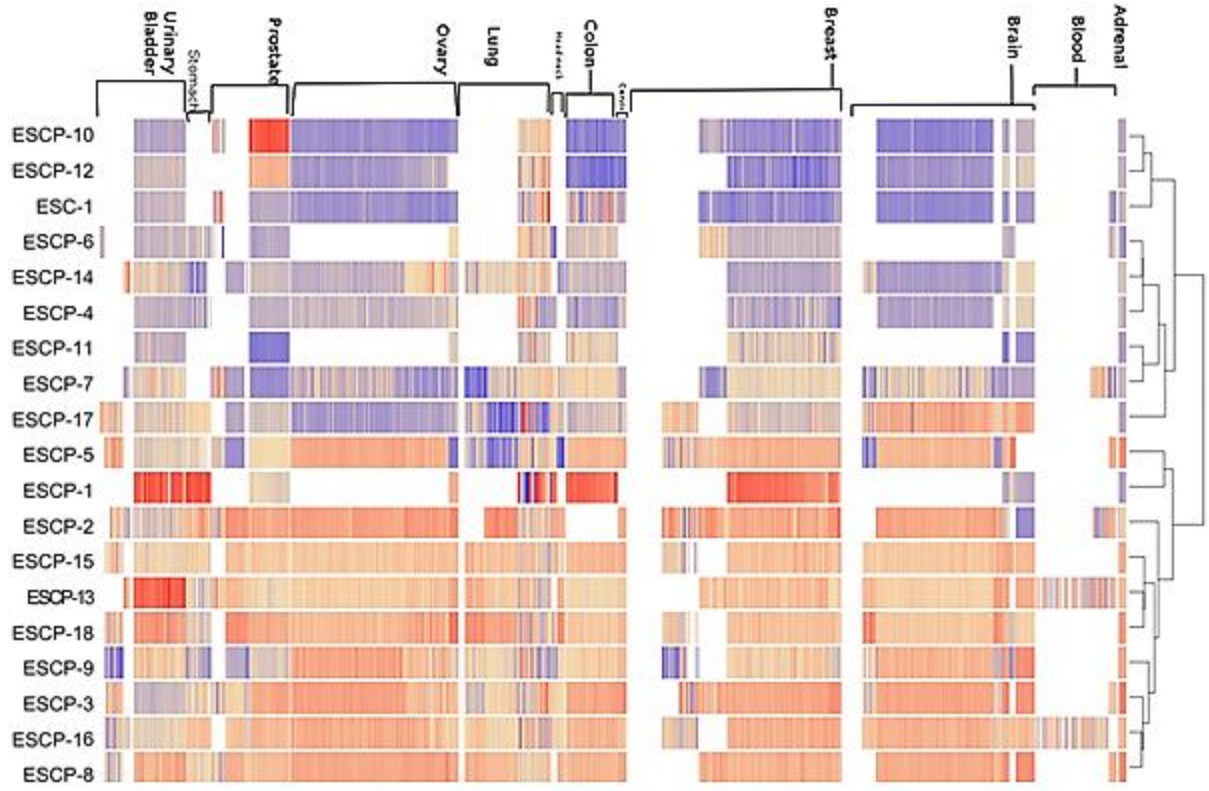


Fig. 5.2 Heatmap for data from 4614 samples across various cancer types (samples are placed alphabetically based on the original tissue/system). Red represents high expression and blue represents low, wheat colour represents average expression or medium expression. Heterogeneity e.g. within colon cancer and lung cancer for ESC-1 can be seen here.

I would also like to mention that seminoma datasets cannot be seen here in the heatmap (Fig. 5.2) above because initially there were three conditions on which the datasets in Oncomine were filtered:

a) Dataset should be an mRNA expression study (there were other study types too e.g. copy number variation study etc.).
b) Dataset should have survival information (to see whether ESC-1 has any effect on prognosis)
c) Dataset should have normal samples in the study (to use as a reference for base line expression in healthy tissue).

5.3 *High positive correlation between ESC-1 and its partners*

Our collaborators had evaluated the effect of suppressing ESC-1 in cell lines and had seen that some stem cell pluripotency factors are downregulated and I wanted to test this hypothesis in actual cancer samples too. Also I wanted to know that are some of its partners co-expressed with ESC-1 or not. To check co-expression I measured the statistical correlation between the expression levels of two genes on matched samples, i.e. expression value for Gene 1 on a set of samples, then expression level of Gene 2 is taken on the same set of samples. (Note: All the genes were not measured in all the datasets and samples.)

ESC-1 x ESCP-10	0.786296987
ESCP-6 x ESC-1	0.718597423
ESC-1 x ESCP-12	0.665228209
ESC-1 x ESCP-14	0.605962362
ESC-1 x ESCP-11	0.587469136
ESCP-7 x ESC-1	0.50288815
ESCP-4 x ESC-1	0.502513074
ESC-1 x ESCP-13	0.429524675
ESC-1 x ESCP-17	0.362293953
ESC-1 x ESCP-18	0.329618695
ESCP-9 x ESC-1	0.285209016
ESCP-5 x ESC-1	0.27456626
ESCP-8 x ESC-1	0.272217547
ESC-1 x ESCP-15	0.235972907
ESC-1 x ESCP-16	0.190357291
ESCP-2 x ESC-1	0.11459401
ESCP-3 x ESC-1	0.074644636
ESCP-1 x ESC-1	-0.036683977

Table V.I – The Spearman Rank Correlation scores of ESC-1 with its 18 interaction partners across thousands of matched samples (sample count ranging from 1270 being lowest common count, up to 3688 being highest common number of samples).

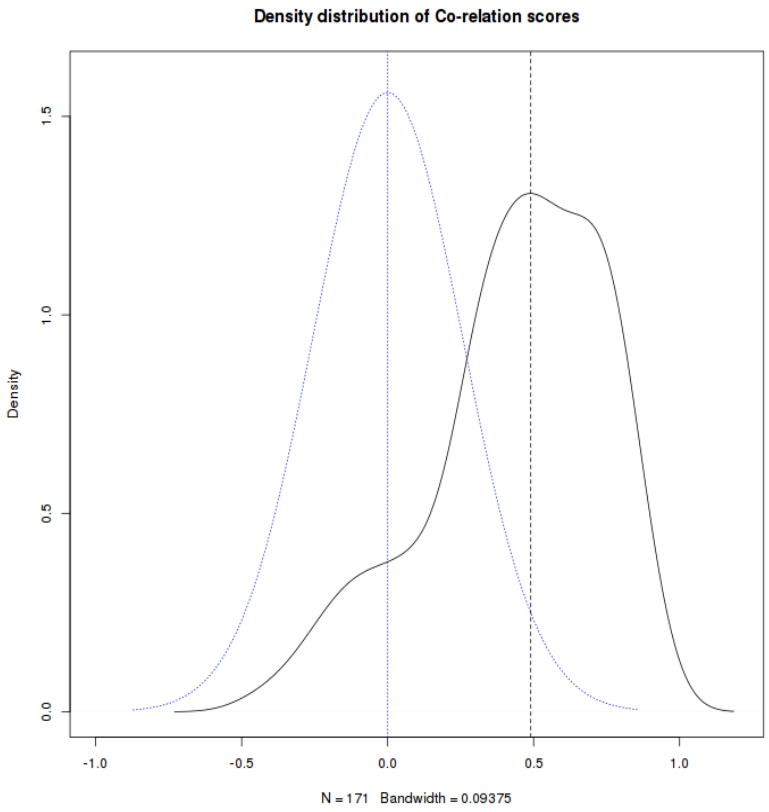


Fig. 5.3 Density distribution of spearman rank correlation scores for the Blue dotted line is for the correlation scores between a random data. The Black solid curve is for the correlation scores of for the 171 possible combination of the 19 genes of interest, [$^{19}C_2$ = 171 unique pairs].

Table V.I shows the Spearman rank correlation scores for the partners of ESC-1 are among the highest scores. These are the partners which seem to be key in the findings of our collaborators. It is not known yet exactly how do they work, but it was nice to see a high correlation in actual cancer samples. The density distribution in the

Fig. 5.3 shows that the mean of the correlation scores is far away from random (0) and is instead positive, being +0.5. Also it can be seen that the bulk of the population lies in the positive region (i.e. greater than zero). Both of these indicate that the set of genes being investigated have high coherence in the samples that were collected in our study.

5.4 *Interactome of ESC-1 in cancer*

I further investigated the coherence in results by making networks (e.g. Fig. 5.4) in Cytoscape using the data from Oncomine and computing the correlation between two genes for matched samples (e.g. Expression level of Gene X and Gene Y in the same sample). The network showed that ESC-1 had statistically high correlation scores with its most important partners ESCP-14, and ESCP-12, two proteins which are down-regulated when ESC-1 is depleted.

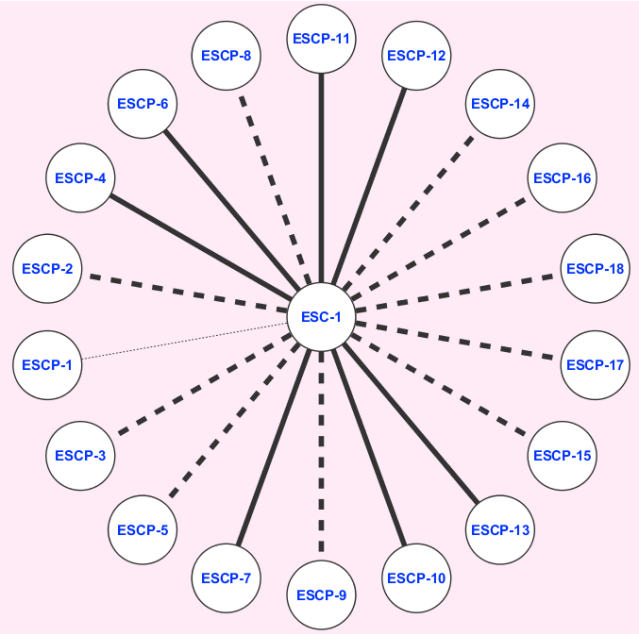


Fig. 5.4 – Network made from spearman rank correlation scores. Thick lines are statistically significant (FDR corrected p-value<0.05) correlation, thin lines statistically insignificant scores. Straight lines represent that correlation was >=+0.5, dotted lines are for pairs with the correlation <+0.5 for matched samples.

5.5 *Problem with Oncomine data and our Analysis model*

I was sceptical about the approach of analysing all the gathered data together which are from *different tissue and cancer type* but also are from *different research laboratories* and *different platforms*. There were a couple of more hurdles in our way as well. Oncomine is a great resource to get the overall report of how a gene behaves in different cancers but it had two major drawbacks for our study:

a) In Oncomine the *data is pre-analysed* and they have performed median-centering to bring data from different datasets to a comparable level. But median-centering is performed on the expression level of all the genes in the dataset and the basic approach behind it is to alter the levels of the expression such that the median of the whole dataset is 0 so that the data is centred around the median. *This method introduces ambiguity* e.g. if a gene is shown as slightly negative it cannot be concluded with certainty that whether

the gene is being expressed or not, of course if the scores are towards the extremes i.e. high positive or negative score one can surely make conclusions about the gene's expression. This was of importance to this study as the gene of interest is an embryonic stem cell pluripotency factor and the samples which were expressing it slightly were equally interesting as much as the samples expressing ESC-1 highly.

b) In Oncomine if one wants to *look at the gene expression of all the genes in the study, there is no convenient way to do so*. The results are shown as plots (i.e. images) and the values are stored in the code of the webpage. If I wanted to get the expression values for all the genes in a dataset I would have had to search each gene and then manually look for and copy code showing the values and then filter them. I used this approach for our 19 genes (ESC-1 + 18 partners), which took a couple of weeks to completely get the data as it was a manual process, but it surely is not a feasible approach if one wants to analyse thousands of genes.

Regarding the analysis model with Oncomine data, i.e. *analysing all the datasets spanning different cancer types together*, I investigated its robustness. The scatter plot (Fig5.5) shows the expression score distribution between ESC-1 (on x-axis) and ESCP-14 (one of ESC-1's important interaction partners on y-axis). Here it can be seen that the expression pattern of

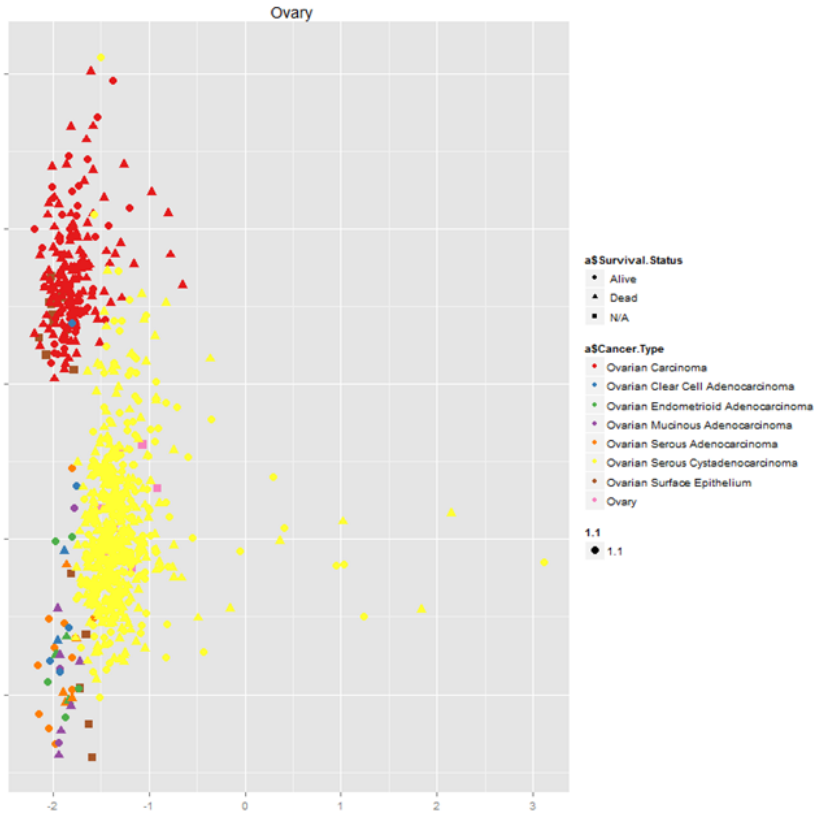


Fig 5.5 – Scatter plot showing the expression score of the samples (each point) in two genes (on the two axes). The Colours are for different cancer subtypes, e.g. in the figure the two biggest groups are ovarian carcinoma (red) and ovarian serous adenocarcinoma (yellow). The shapes represent the survival status of the patient: circle for alive, triangle for dead and square for normal samples.

ovarian carcinoma (shown in red) is different from ovarian serous adenocarcinoma (in yellow). Similar results were found when the expression of genes was compared in almost all the cancers where there were more than one subtype (with substantial number of samples, e.g. more than 20). I was also looking at the survival status of the samples when comparing the expression in the datasets and therefore there are different point shapes i.e. square, circle and triangles.

In the plot (Fig. 5.6 Part A) I was very surprised to see that some of the points are concentrated in a specific region and some are scattered. The points with blue are for the normal prostate tissue sample and the red is for the tumour. I was surprised to see that even the normal tissue samples are having such large differences. Then I investigated whether these differences are due to the samples being from two different datasets, and as soon as I plotted the image (Fig. 5.6 Part B) it was revealed that *the two data clusters were from two different studies (and of course 2 datasets)*. Similar results from other cancer types suggested that it is not practical to analyse different datasets together, instead a better approach would be to analyse them independently and then compare their results.

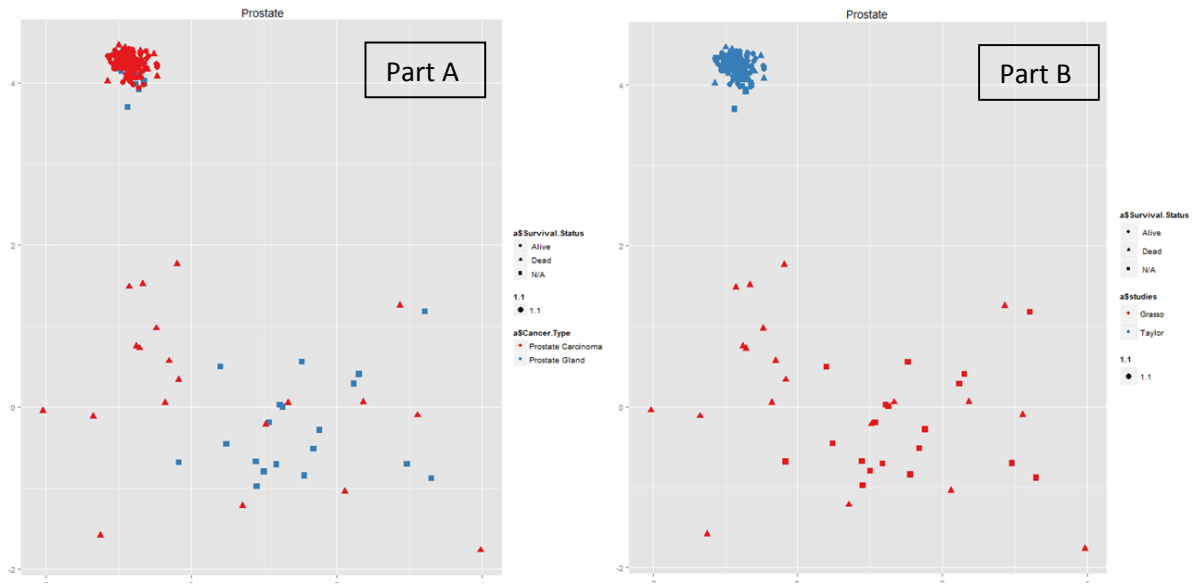


Fig. 5.6 – Scatter plots showing the expression score of the samples (each point) in two genes (on the two axes). **Part A**- Colour represents tissue Blue for Prostate gland, Red for the Prostate Carcinoma. In **Part B**- the colours represent the study/dataset. The shapes of points are representing the survival status as in Fig 5.4.

5.6 *Analysis of TCGA datasets*

TCGA or The Cancer Genome Atlas [48] was the first destination of choice for datasets as it is a repository of data from various experiment types e.g. Sequencing, Microarray experiments, Methylation study, etc. on cancer samples from various cancer types. They have data for more than 30 types of cancer types (as on 02-07-2014) and have a large collection of samples too e.g. for Colon adenocarcinoma they have 462 samples, for Breast invasive carcinoma they have 1101 samples. In addition to their large sample collection, they contain clinical information too e.g. tumour stage, tumour size (in some cases), and most importantly the survival information.

I analysed the datasets mentioned in section 4.1.1, for Brain Cancer, Colon Cancer, Lung Cancer and Ovarian Cancer. I chose to investigate these datasets based on the findings from analysing the Oncomine data. *In the TCGA datasets I was not able to find any cancer type to be having enough samples with high expression or even medium expression to test the hypothesis*. Table V.II summarizes the results with few statistics about the expression level of ESC-1 in the TCGA datasets. The table also shows the number of samples in the datasets and the Range of whole dataset (i.e. min to max). For all of them it can be seen that the *median and mean are low as well as the third quartile is also low in all the studies*.

TCGA Colon Adenocarcinoma, Agilent G4502A-07-3							
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	**Samples**	**Range**
-7.494	-4.4	-2.439	-2.341	-0.6239	3.996	174	-13.29 : +17.35
TCGA GlioBlastoma Multiforme (Brain Cancer), Affymerix HG-U133							
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	**Samples**	**Range**
3.285	3.604	3.728	3.769	3.892	8.919	519	+2.80 : +14.47
TCGA GlioBlastoma Multiforme (Brain Cancer), Agilent (G4502A-07-1)							
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	**Samples**	**Range**
-7.088	-5.704	-4.917	-4.826	-4.143	-0.071	89	-9.83 : +12.09
TCGA GlioBlastoma Multiforme (Brain Cancer) Agilent (G4502A-07-2)							
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	**Samples**	**Range**
-8.695	-5.915	-5.167	-5.125	-4.543	-0.877	473	-13.73 : +16.04
TCGA Lung Adenocarcinoma, Agilent G4502A-07-3							
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	**Samples**	**Range**
-7.134	-5.413	-5.004	-4.934	-4.464	-2.159	32	-13.78 : +12.00
TCGA Lung SC Carcinoma, Agilent G4502A-07-3							
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	**Samples**	**Range**
-7.095	-5.705	-5.202	-5.152	-4.656	-1.138	153	-12.61 : +13.08

TCGA Ovarian serous cystadenocarcinoma, Affymetrix HG-U133							
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	**Samples**	**Range**
2.475	2.753	2.877	2.957	3.002	7.413	586	+2.07 : +13.75
TCGA Ovarian serous cystadenocarcinoma, Agilent G4502A-07-3							
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	**Samples**	**Range**
-8.416	-6.074	-5.461	-5.275	-4.693	0.53	588	-14.66 : +12.96

Table V.II – Table showing the 5 number summary for ESC-1 (Minimum, 1st Quartile, Median, 3rd Quartile, Maximum) and mean too. The table also shows the number of samples in each dataset and the Range of the expression scores for all the probes (for different genes) in the dataset in the format Minimum observation : Maximum observation.

I found some samples that might be feebly expressing ESC-1 (Fig. 5.7) but there is no certainty in those cases and it is hard to make inferences. Thus, I decided to investigate more datasets to get a clearer understanding.

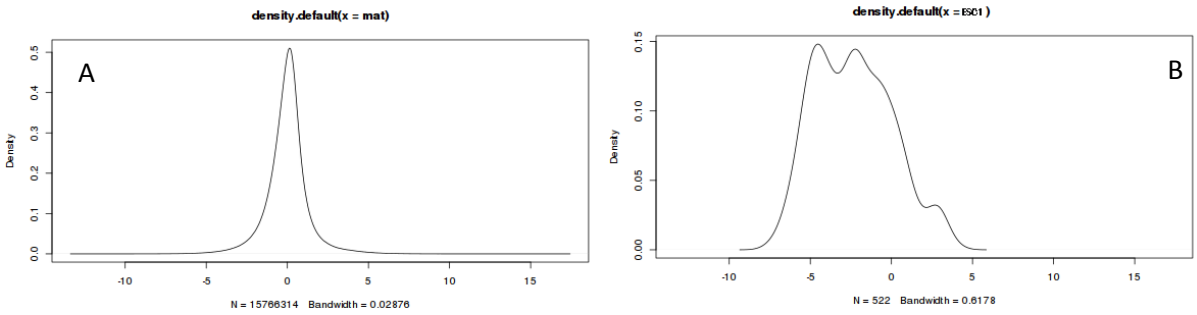


Fig. 5.7 – Density distribution plots for all the samples from TCGA Colon Adenocarcinoma **(A)** For expression values of all the genes across all the samples. **(B)** Expression values for gene ESC-1 (comprising of three probes, as the Agilent platform has three probes for measuring ESC-1) across all the samples. It can be seen that the majority of the samples lie in the negative region, with only a small fraction in the positive region.

The heatmap (Fig. 5.8) for ESC-1 and its partners gave me motivation to look more in Colon Adenocarcinoma as among the samples that were expressing ESC-1 (slightly, but in the set that was in the TCGA dataset, they are the highest ESC-1 expressing samples, marked by a black box in Fig. 5.8), it can be seen that all of them are from cancer samples. But it is also baffling that a lot of normal samples are present in the region between 0 to -2.5 (see colour key in Fig. 5.8). There is no clustering of samples in Fig. 5.8 when the samples are ordered on the ESC-1 expression levels, which makes it clear that ESC-1 is not playing a role in defining the cancer stage.

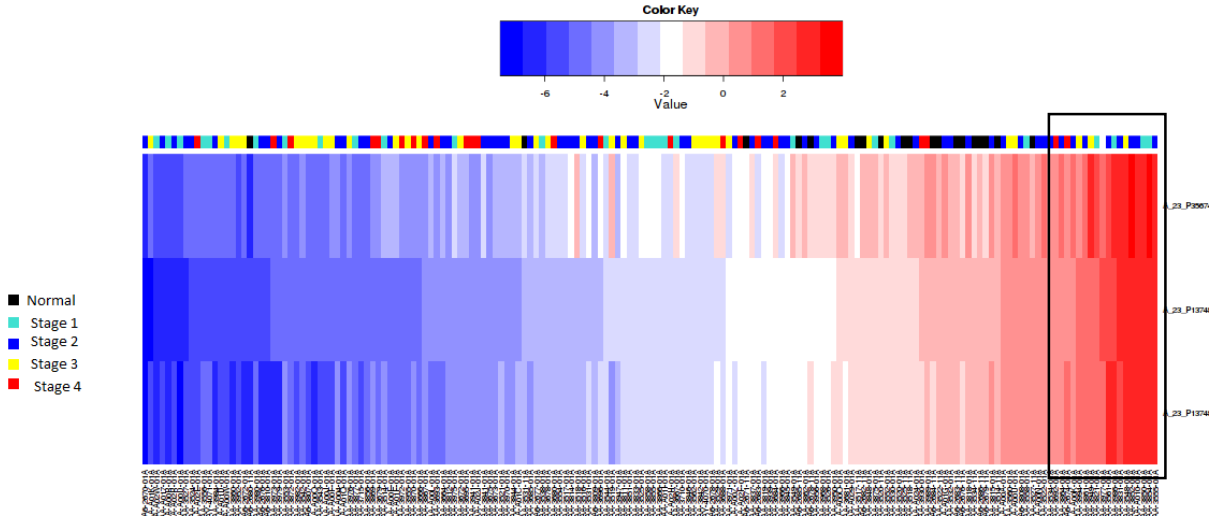


Fig. 5.8 – Heatmap for three probes measuring ESC1 in the TCGA Colon Adenocarcinoma where the expression is ordered in ascending order and a colour code on top represents the stage of the tumour (Black, Cyan, Blue, Yellow and Red). It shows that samples that are expressing ESC-1 , though at low levels (marked with black box) are mostly cancers of higher stage (blue, yellow and red) and none of them are normal (black).

5.7 *Results from GSE8671 (Colon Cancer) and GSE3218 (Seminoma)*

I went back to my Oncomine [73] results (Fig. 5.2) to look for datasets where some of the samples are expressing the gene of interest, ESC-1. To test the hypotheses datasets that have samples expressing the gene of interest are needed. I found some datasets (description in section 4.1.2) which had samples expressing ESC-1 but none of them had survival information or any clinical information for that matter. However, I still wanted to proceed with them as they gave me the opportunity to test whether any of the partners of ESC-1 are differentially expressed between the sample groups or not. These datasets were GSE8671 for Colon cancer [57] (colorectal adenomas) and GSE3218 Seminoma (Male Germ Cell tumour) [56]. In both of these datasets (as see in Fig. 5.9 and Fig. 5.10) there were subsets of samples expressing ESC-1 highly.

The heatmaps (Fig. 5.9 and Fig. 5.10) also show a *colour code for each sample based on the expression level of ESC-1 in that sample*, Red represents high expression i.e. RMA normalized [63] intensity (RNI) >10, Cyan represents medium expression where $7 \leq RNI \leq 10$, Blue represents under expressed $RNI < 7$, and Green is for representing normal tissue samples regardless of their expression level. It can be seen that in Seminoma as well as Colon Cancer

dataset *the ESC-1 Expressing samples have a different expression profile than the normal samples* based on this gene set. One can also see this by the way the samples are clustered (through the dendrogram at the top of the heatmaps, they show the hierarchical clustering of samples (columns) and genes (rows)). The genes in the heatmap apart from the evident ESC-1 are the partners of ESC-1 (only those that were measured in that particular dataset) which do not show much heterogeneity in expression. Also, I included some housekeeping genes as reference e.g. NDUFA1 (NADH Dehydrogenase), POL2RA (RNA polymerase II) and PSAT-1 (Phosphoserine aminotransferase 1).

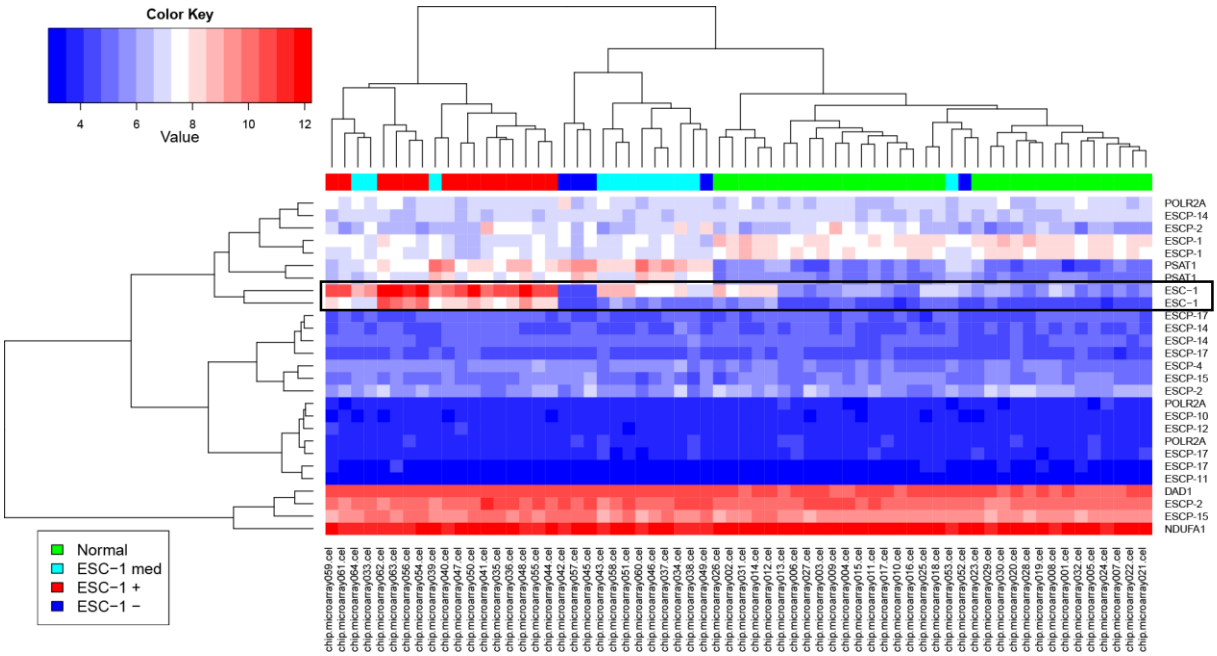


Fig. 5.9 – Heatmap showing expression levels of ESC-1 in colorectal adenoma tumour samples and matched normal colon tissue controls from dataset GSE8671, showing a subset of samples having high ESC-1 expression. There is a colour code on top marking samples based on their expression status for ESC-1.

Through the Fig. 5.10 one can see that *ESC-1, ESCP-12 and ESCP-14 cluster together which signifies that their expression levels are highly correlated*, a fact that is for the first time shown in cancer patient samples. I included KRAS, the classical oncogene, to see if there is any correlation between its expression levels and that of ESC-1, but from Fig. 5.10 it seems clear.

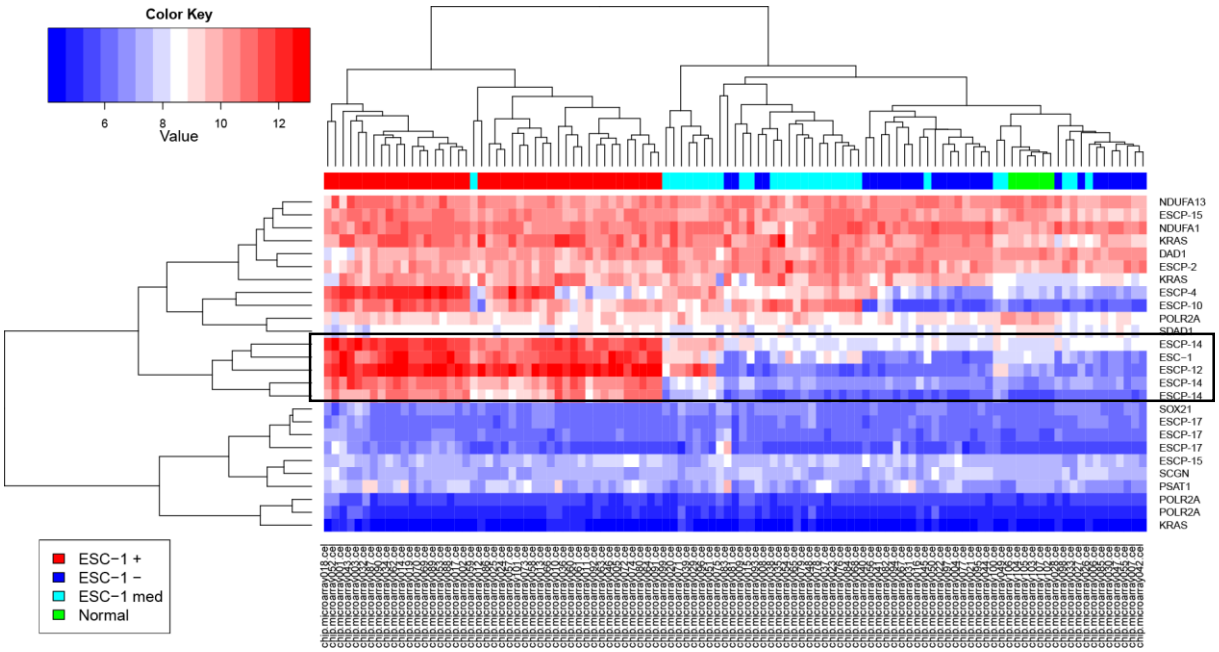


Fig. 5.10 – Heatmap showing expression levels of ESC-1 in Seminoma samples and 6 normal tissue controls from dataset GSE3218, showing a subset of samples having high ESC-1 expression. The colour code on top categorizes samples based on their expression intensity for ESC-1 (legend is on bottom left side).

In the scatterplots (Fig. 5.11) from the Principal Component Analysis (PCA), where the points represent samples in the two datasets and the X axis represents the First Principal Component and the Y axis represents the second principal component, one can see that the normal samples are different than the tumour samples. The difference is very clear in the case of Colon Cancer (Fig. 5.11 A*), as it can be seen that the normal samples (shown by green points) are present far away from the tumour samples*. Similarly in the case of Seminoma dataset (Fig. 5.11 B) it can be seen that the normal samples are clustered distinctly, although there is not such a big degree of separation. This fact does not imply any important and noteworthy deductions but it is just a visual interpretation of the plots. They should also not be compared directly with each other as the number of normal samples and the type of normal samples in both the datasets is different. For the Colon cancer there were matched control (i.e. normal) samples which means that the normal samples (32 in number) were taken from the same patients from whom the tumour samples (32 in number) were derived from, while in the case of Seminoma dataset there were six normal organ samples as control tissue. But still it indicates that there are definite differences in the gene expression profile of the normal tissue and the tumour tissue.

One more point to note was that in the Seminoma dataset (Fig. 5.11 B) *the samples not expressing ESC-1 (show in blue colour) cluster together*, and the samples expressing ESC-1 (shown in red) form a group although it still has several samples expressing ESC-1 at medium levels (shown in cyan colour). This cannot however be seen clearly in the colon cancer dataset (Fig. 5.11 A), partly because there are few samples tagged as ESC-1 negative as *most of the normal samples are also not expressing ESC-1* which can be seen clearly in the heatmap (Fig. 5.9), this was a good sign and kept me motivated to look further into Colon Cancer datasets.

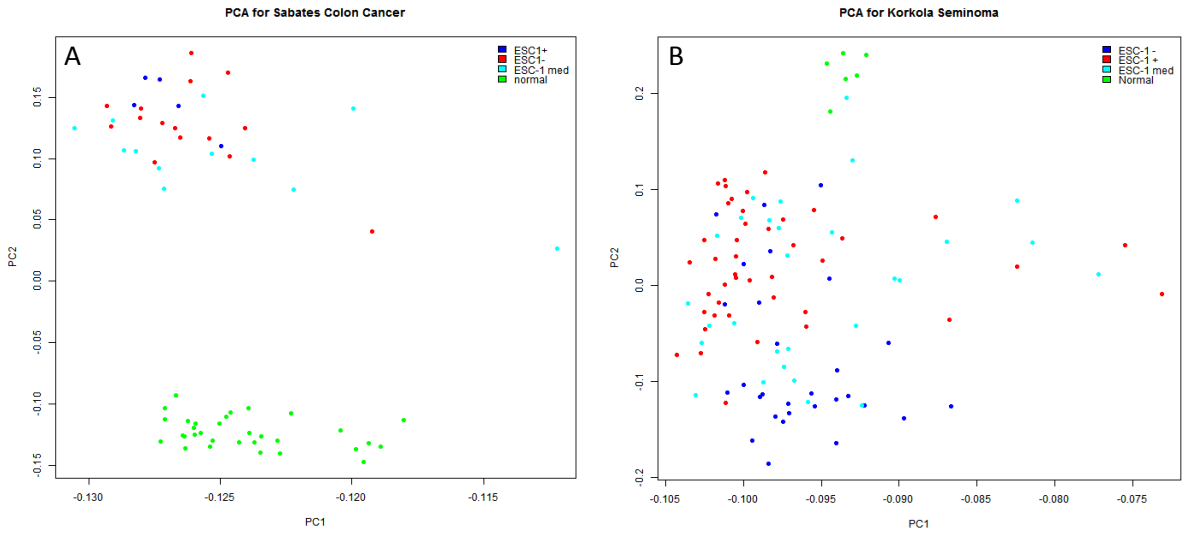


Fig. 5.11 – Scatter plot showing the first 2 Principle components for the complete expression matrix (i.e. all the probes) with the points representing the samples and the colours representing the expression level of ESC-1 (legend is on top right side) and green for normal samples. **A)** Samples from colon cancer dataset GSE8671, **B)** Samples from Seminoma (Male Germ Cell Tumour) dataset GSE3218.

On the other hand for the Seminoma dataset the normal samples lied in the twilight zone or almost at the lower edge of the medium expressing zone. This can be inferred from the heatmap for Seminoma dataset (Fig. 5.10), therefore I paused with the Seminoma dataset.

5.8 *Results from Colon Cancer Datasets & Prognosis*

I decided to investigate more into Colon Cancer dataset after I came across the dataset from International Genomics Consortium's Expression Project for Oncology on GEO with accession number GSE2109 [58]. The dataset had 2158 samples from various cancer types on the platform GPL570 Affymetrix Human Genome U133 Plus 2.0 Array. I was interested to see if there are cancer types other than Seminoma or Colon Cancer where ESC-1 is highly expressed.

Therefore I analysed this dataset after normalizing it through RMA and $\log_2$ transforming the intensities. The boxplot (Fig. 5.12) shows the expression intensities for ESC-1 in different cancer types, it can be seen that *colon cancer* (3rd from left, marked with a black dotted box) *has a large number of samples expressing ESC-1*. Also the *median expression level for samples of Colon Cancer is highest among all cancer types*. The boxplot shows 15 cancer types that had largest number of samples. A wide range was covered by this method as the highest is 381 samples for Breast cancer and the lowest in this plot is 17 samples for both Pancreatic and Cervical cancer.

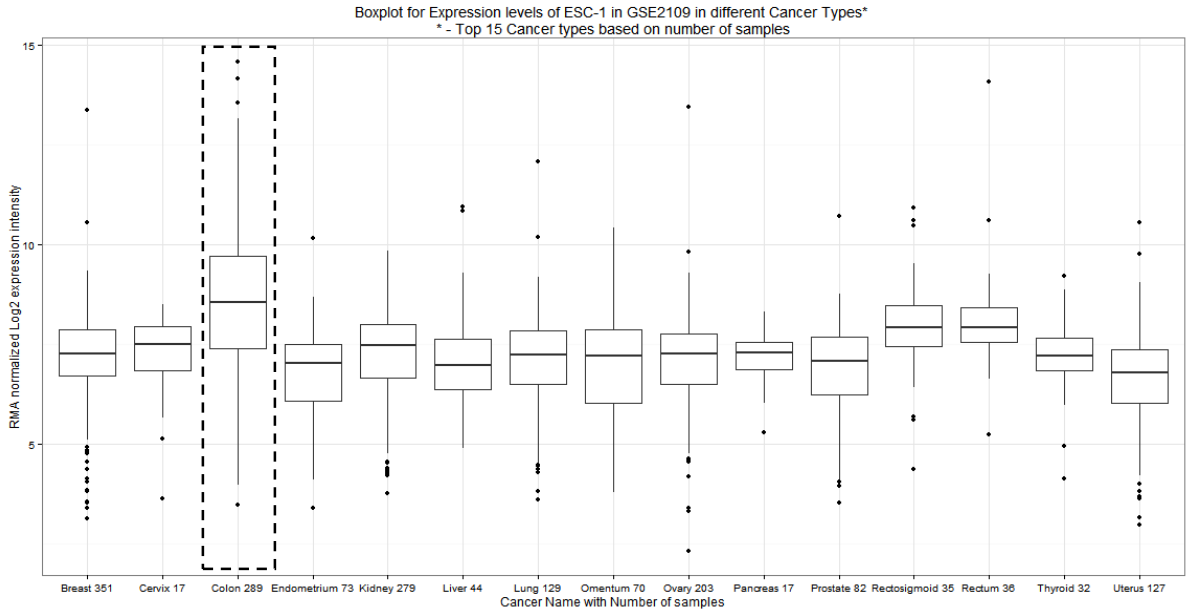


Fig. 5.12 – Boxplot showing RMA normalized intensity for ESC-1 from the samples in dataset GSE2109 from the 15 cancer types that had most number of samples. Colon Cancer (black dotted box) has a large number of samples in the range of medium to high expression, which is also reflected in the median of the expression intensities. The labels on X axis also include the number of samples for that particular cancer type.

The results from datasets GSE8671 (Colon1) (Fig.5.9 and Fig. 5.11A) and GSE2109 (ExPO) (Fig. 5.12) I found and worked on colon cancer datasets that had survival information available.

I found four more datasets with GEO accession numbers GSE14333(Colon2) [59], GSE17536 (Colon3) [60], GSE17537 (Colon4) [60] and GSE33113 (Colon5) [61] (description in section 4.1.2) that had the survival information available for the samples. This was of immense importance as MIAME (Minimum Information About a Microarray Experiment) standards do not require this information to be provided and there is a severe lack of datasets with such accessory information.

Another concern while working with Microarray datasets is that for the very high and very low signal intensities one can conclude the expression state with a great deal of certainty, but not enough can be said about samples whose intensities lie in the middle of the scale. Therefore, I chose to use UPC (see section 4.2.2) normalization. To be sure to a higher level, I chose UPC ≥ 0.75 as expressing the gene of interest ESC-1 and samples with UPC < 0.75 were treated as not expressing ESC-1 (shown as blue dots in Fig. 5.13).

Also it is worth mentioning here that in case there were two probes corresponding to a gene in Affymetrix HG U133 Plus 2.0 and only one in Affymetrix HG U133 microarray platform, the probe (Probe 1) that was *common between the two platforms* was chosen for deciding if the sample is expressing a gene or not (i.e. Probe 1 is common and Probe 2 is additional probe in the Plus 2.0 array for the gene). Fig 5.13 shows the correalation of UPC scores between the probes.

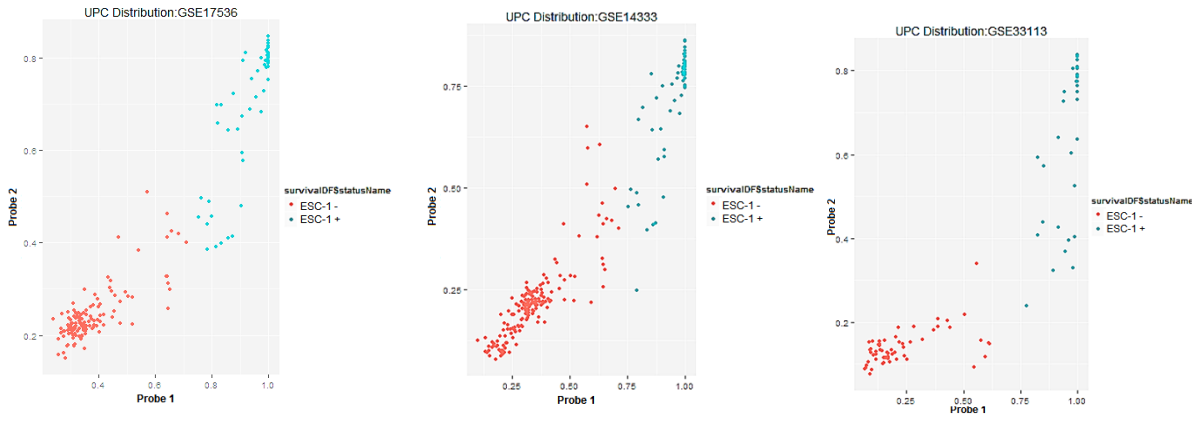


Fig. 5.13 – Scatter plot in three Colon Cancer datasets for UPC scores of a gene for 2 probes. The dots represent samples coloured blue for 'Active' if UPC score for Probe 1 ≥ 0.75.

The heatmap in Fig. 5.14 shows the UPC scores for ESC-1 and its partners along with some housekeeping genes like NDUFA1 (NADH Dehydrogenase), POL2RA (RNA polymerase II) and PSAT-1 (Phosphoserine aminotransferase 1) in the dataset GSE17537 (colon5) for expression levels based on. I also included a well-known oncogene KRAS to see if it is perturbed in the samples, but here *KRAS had a low signal in all the samples* in GSE17537. The heatmap (Fig. 5.14) is also having a colour code for Alive (pink) and Dead (Black) for the samples. From the heatmap it can be seen that there are several genes/probes that are homogenous throughout the samples.

I performed filtering (section 4.9) of the genes based on the Standard deviation (σ), so

that these rows that are uniformly high or low can be get rid of, because I was interested in heterogeneity in gene expression.

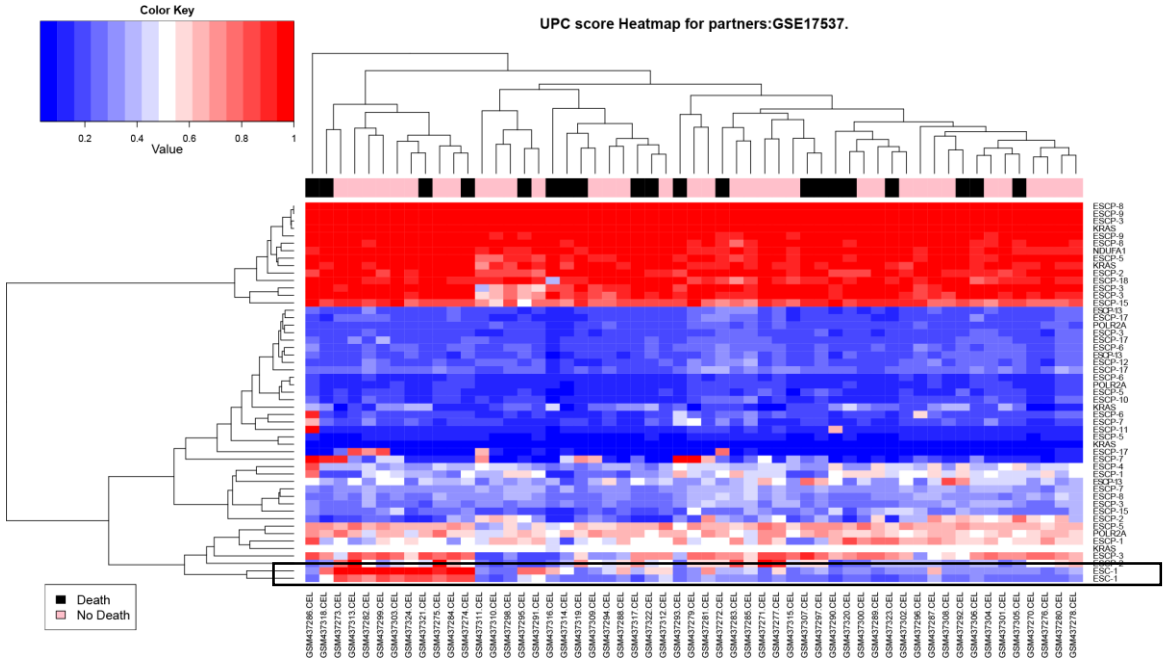


Fig. 5.14 – Heatmap of GSE17537 (Colon5) for ESC-1 and its partners along with some housekeeping genes like POL2RA, NDUFA1, etc. and KRAS (a well-known oncogene). The two probes for ESC-1 are present as the last 2 rows in the heatmap above, with only a subset of samples expressing it (as seen earlier in Fig 5.9 and 5.10).

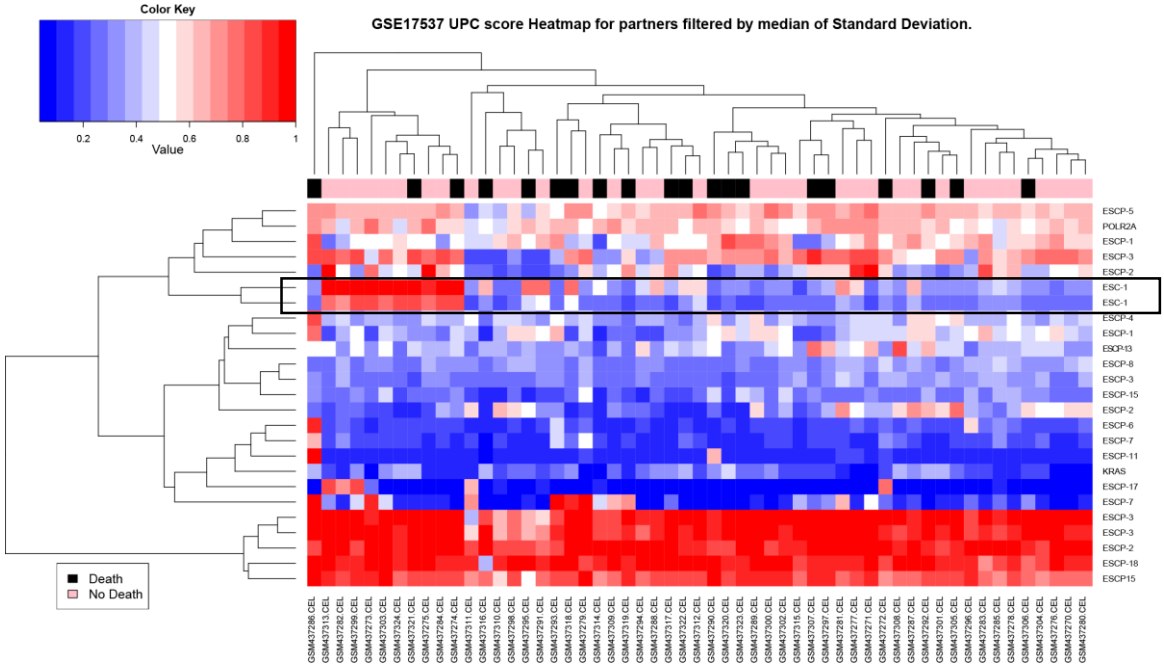


Fig. 5.15 – Heatmap of GSE17537 (Colon5) for ESC-1 and its partners filtered based on standard deviation, by keeping the genes (represented by rows) that have $\sigma_g > \text{median}(\sigma_g)$. The colour coding on top, pink represents "no death" and black represents "death". ESC-1 rows highlighted with a black rectangle.

As evident in the heatmap (Fig.5.15) even after filtering by taking the median as

threshold I still had rows with homogeneity, so I *performed a more strict filtering* (see section 4.9) by keeping rows (genes) that had standard deviation σ_g higher than the third Quartile (Q_3) of the standard deviations for all genes, in the matrix used to create the original unfiltered heatmap (Fig. 5.14). After this I got a heatmap (Fig. 5.16) that is *comprised of most heterogeneous genes*. I applied the same procedure to other datasets as well and here I present only those heatmaps (Fig. 5.17, 5.18 and 5.19) here.

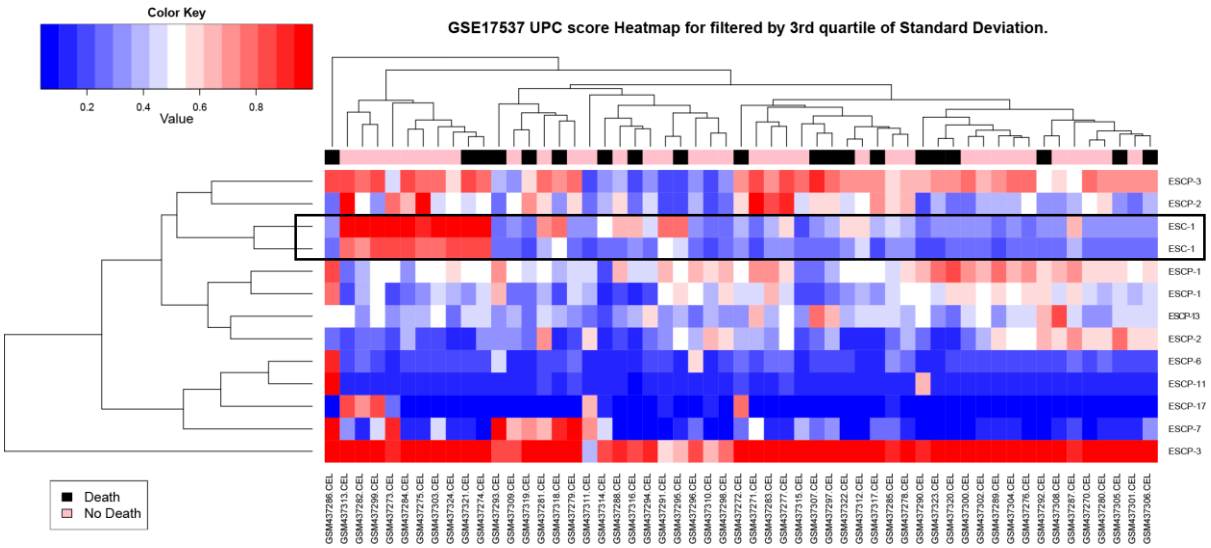


Fig. 5.16 – Heatmap of GSE17537 (Colon5) for ESC-1 and its partners filtered based on standard deviation, by keeping the genes (represented by rows) that have $\sigma > Q_3(\sigma_g)$. The colour coding on top, pink represents "no death" and black represents "death". ESC-1 rows highlighted with a black rectangle.

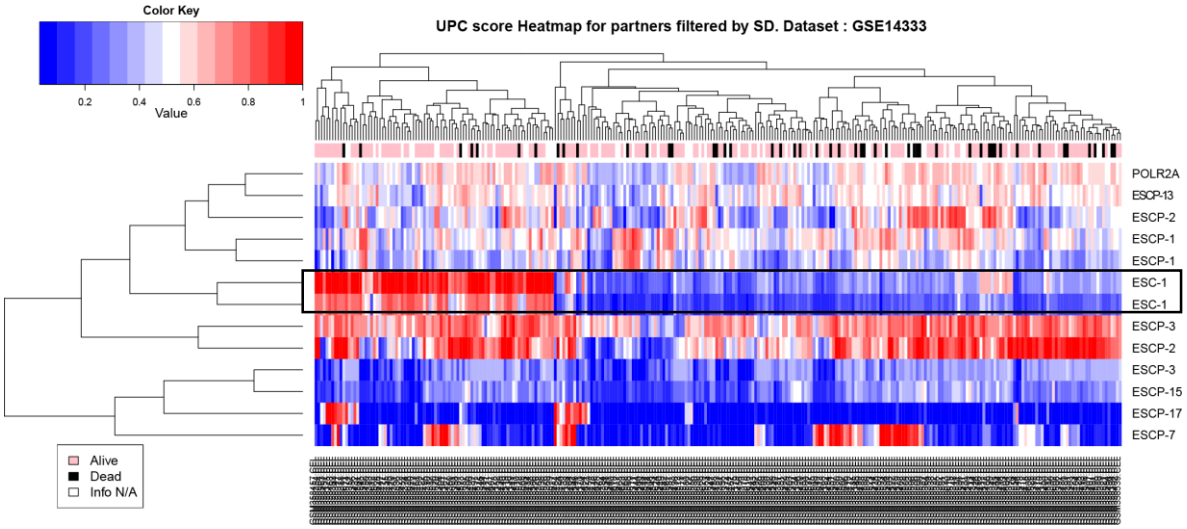


Fig. 5.17 – Heatmap of GSE14333 (Colon2) for ESC-1 and its partners filtered based on standard deviation, by keeping the genes (represented by rows) that have $\sigma > Q_3(\sigma_g)$. The colour coding on top, pink represents "Alive" and black represents "death", white is for samples with missing status. ESC-1 rows are highlighted in black.

From these four heatmaps (Fig. 5.16 - 5.19) *it seems that there is no proper segregation of the samples based on this gene list of partners of ESC-1*. Also it is interesting to point out that in line with the finding from GSE8671 (Colon1) for colon cancer (heatmap in Fig. 5.9) *ESCP-12 or ESCP-14 are not there in the heatmaps with the set of genes* that are having a heterogeneous expression. Instead they are having at low signal for intensity across all samples. This came as a surprise, it was expected based on the previous findings, that they would be co-expressing with ESC-1 (section 5.3) i.e. having a high correlation. However, those results were from a study including a variety of cancers, and as we know cancer is a disorder which is full of unique properties of its own and one should always keep a sense of caution in mind when extrapolating findings. But to my surprise I found *ESCP-17 among the heterogeneous genes, in all the 4 colon cancer datasets*. Although, here it did not have a high correlation with ESC-1, which was expected based on my previous findings (section 5.3, also see Table V.I). It is also interesting to mention that in the dataset GSE8671 (Colon1) (Fig. 5.9) ESCP-17 seems to be homogenously giving a signal in medium range.

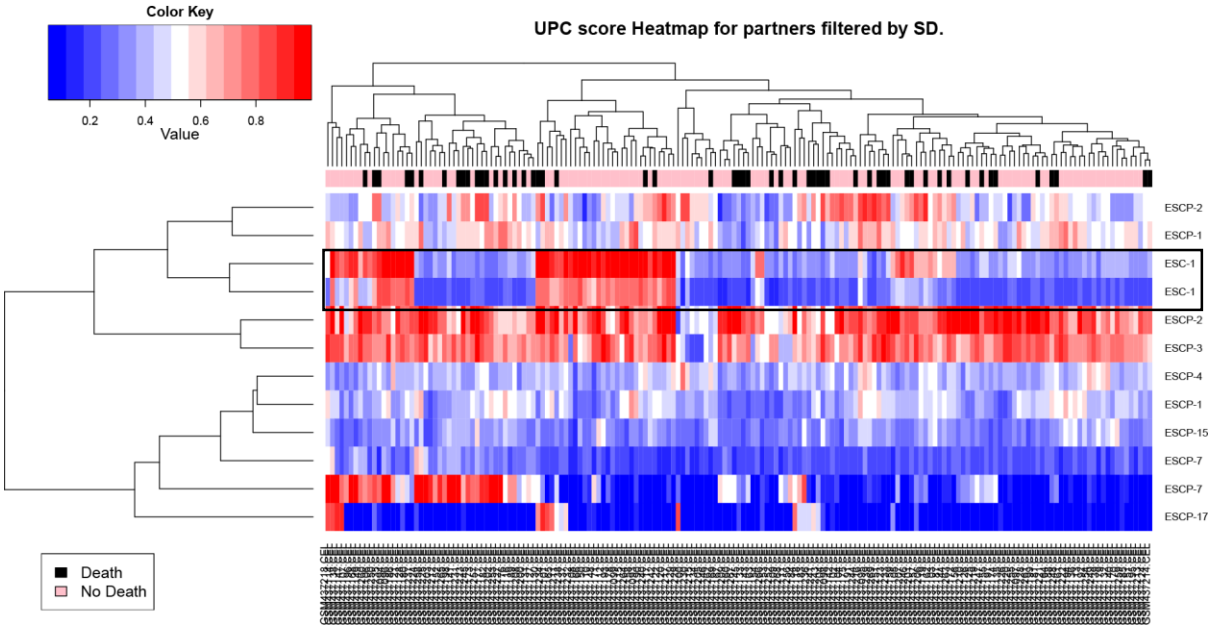


Fig. 5.18 – Heatmap of GSE17536 (Colon3) for ESC-1 and its partners filtered based on σ (std. dev.), by keeping the genes (represented by rows) that have $\sigma_g > Q_3(\sigma_{\text{all probes}})$. The colour coding on top, pink represents "no death" and black represents "death". Rows corresponding to ESC-1 are highlighted in black.

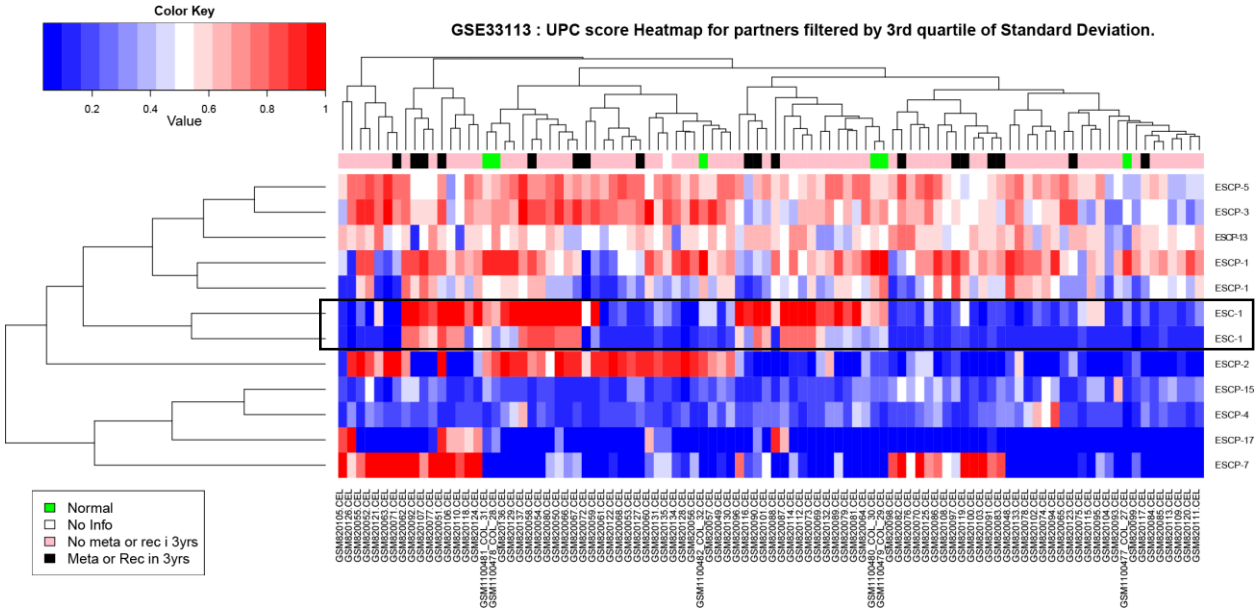


Fig. 5.19 – Heatmap of GSE33113 (Colon4) for ESC-1 and its partners filtered based on standard deviation, by keeping the genes (represented by rows) that have $\sigma_g > Q_3(\sigma_g)$. The colour coding on top, pink represents "no metastasis or reoccurrence in 3 years" and black represents "metastasis or reoccurrence in 3 years". Normal samples are shown in green. Rows corresponding to ESC-1 are highlighted in black.

From these heatmaps (Fig. 5.16 through to 5.19) one can see that among the samples expressing ESC-1 there are larger proportion of Alive samples or non-metastatic samples while very few have their status as Dead or metastatic. This is a theme that reoccurs during my investigation in the colon cancer datasets, and it came as a surprise that *it might be that ESC-1 is a positive prognostic factor after all*, instead of being a negative one. But I have done further analyses like Kaplan-Meier survival analysis with these datasets and am trying to find out reasons or factors that could explain this phenomenon. Also I am looking if it is actually true and significant or is it just a mere coincidence. But I am not at the liberty to disclose those results yet.

5.9 *Differential expression analysis and overlap among results*

I also performed analysis to find out differentially expressed genes (only significant results reported i.e. FDR p-value < 0.05) between the samples expressing ESC-1 and the samples not expressing ESC-1 (based on the UPC score, as mentioned earlier I decided that UPC ≥ 0.75 as active or expressing and UPC < 0.75 as inactive or not expressing). In the Venn diagram one can see the overlap between differentially expressed genes between the four Colon Cancer

datasets GSE14333 (Colon2), GSE17536 (Colon3), GSE17537 (Colon4) and GSE33113 (Colon5) [59]–[61]. The analysis was done using the UPC scores for the datasets by the method ROTS (Reproducibility Optimized Test Statistic) [71], where the statistic is optimized, based on the data, among a family of T-type statistics. I chose this method because it does not assume the sample (i.e. the data provided) to have a particular type of distribution e.g. the T-test assumes the data to follow a normal distribution, which the data, or in general any microarray data does not, see Fig. 4.1 A and 4.1 B.

It can be seen in Fig. 5.20 that there were a lot more detections in the dataset GSE14333 (Colon2) than in the other datasets, which might be attributed to the fact that it had most samples, 290 tumour samples, among the four datasets under consideration. In fact this trend can be observed in other datasets too, after GSE17536 (Colon3) which had 177 tumour samples, then there is GSE33113 (Colon4) with 90 tumour samples and the least number of detections in GSE17537 (Colon5) with 57 tumour samples. Similar pattern was observed with the T-test results too.

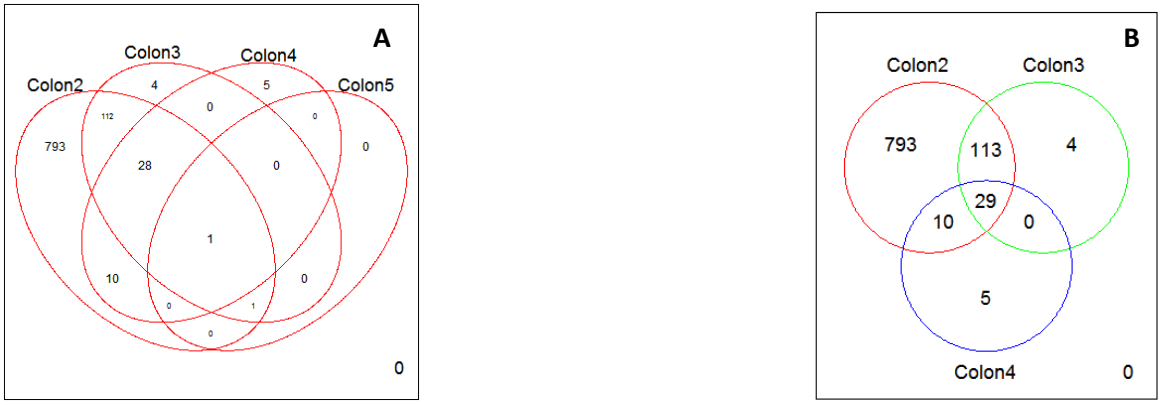
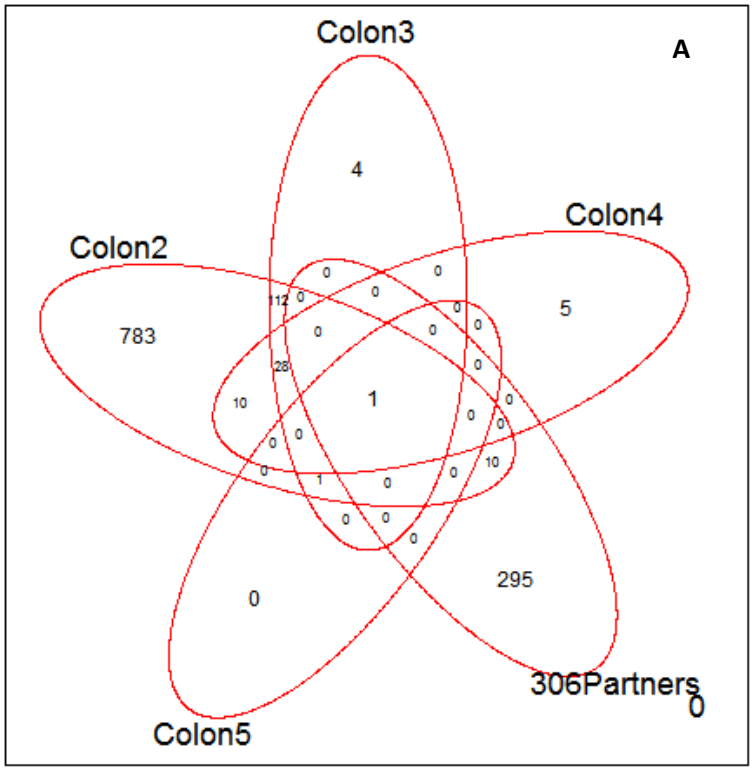


Fig. 5.20 – Venn diagram for showing overlap among the four colon cancer datasets. Differentially expressed genes detected using ROTS for the four datasets, **(B)** ROTS if the dataset GSE17537 (Colon4) is excluded as it had only two detections. For names refer to Table IV.II.

I also wanted to check how many of the interaction partners of ESC-1 were differentially expressed between the ESC-1 expressing and ESC-1 not expressing samples. It can be seen in the Fig. 5.21 that the result looks very disappointing as there is just *1 among the 306 partners that is common to all four Datasets, and even that is ESC-1 itself*. But when I compared the findings to Fig. 5.21A I saw that the dataset size may be a factor behind this. *The dataset GSE17537 had just two Differentially Expressed genes being detected*. I was motivated to do this kind of comparison because of the previous result (Fig. 5.22) with GSE8671 (Colon Cancer) and GSE3218 (Seminoma).



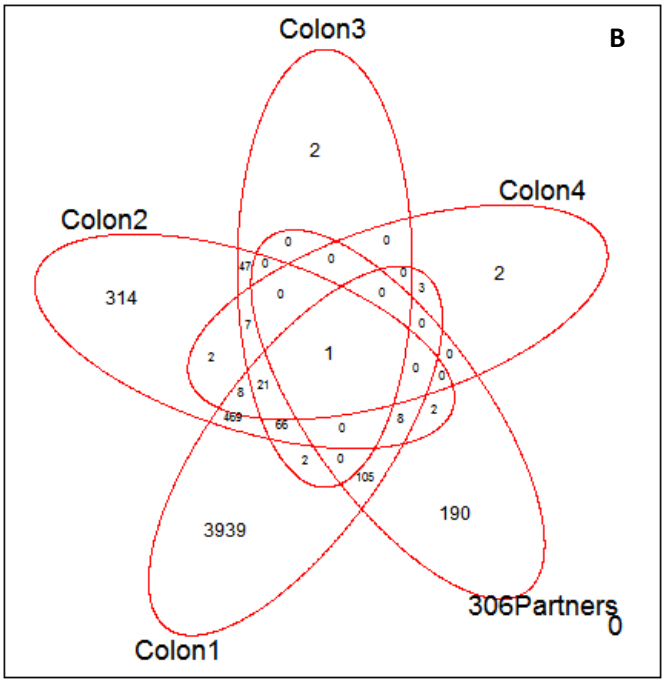


Fig. 5.21 – Venn diagram for showing overlap among the 306 partners of ESC1 and the colon cancer datasets differentially expressed genes detected using ROTS. Datasets labelled with their nicknames similar to Fig5.19, for more information refer Table IV.II, and 306Partner is the list of interaction partners I received from our collaborators. **A)** The overlap among the 4 new datasets. **B)** As Colon5 dataset had very few differentially expressed genes, the dataset colon1 from section 5.7 was used to see the overlap. But as there are few detections in colon3 and colon4 datasets the intersection is still 1, but there are more common interaction partners differentially expressed in colon1.

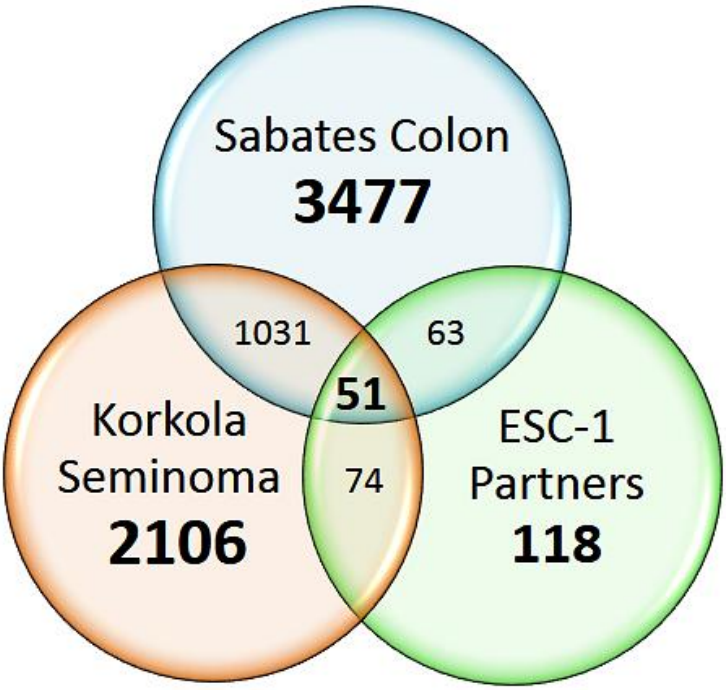


Fig. 5.22 – Venn diagram for showing overlap among the 306 partners of ESC1 and the differentially expressed genes in Seminoma (GSE3218) and Colon Cancer (GSE8671 or Colon1) detected using ROTS. It can be seen all the 306 partners were differentially expressed in one or more of the datasets. Colour of the circles is just for aesthetics. ESC-1 partners are interaction partners of ESC-1 validated by our collaborators (they were 306 in total).

As evident from the Venn diagram (Fig. 5.22) for the analyses on Colon Cancer (Colon1, GSE8671 [57]) and Seminoma dataset (GSE3218 [56]) (Section 5.7), I saw that quite a lot of partners of ESC-1 were differentially expressed in these two datasets. Therefore I decided to investigate more. *There were 51 partners of ESC-1 being differentially expressed in the two datasets* (Fig. 5.22). Also in Fig. 5.21 A *it can be seen that there are 10 of the partners of ESC-1 that are differentially expressed in the dataset with largest number of samples (290) GSE14333* (labelled as Colon2) and in Fig. 5.21 B it can be that there are several partners (113 to be precise) being differentially expressed in GSE8671 (labelled as Colon1). This again is a testimony to the heterogeneity in cancer and demands more research.

5.10 *Impact on Survival ?*

The prime focus of this study has been to investigate if there exists any relationship between ESC-1 and the survival of the patient. If there isn't any direct impact then finding out if there is any impact of some of its partners together with ESC-1 on the survival of a patient.

I have been working with the colon cancer datasets (GSE17537, GSE17536, GSE 14333 and GSE33113 [59]–[61]; results from them shown in section 5.8) that had survival information. Not being at the liberty to disclose much details, I would present a simple percentage table (Table V.IV) for these four colon cancer datasets.

Dataset	Property	ESC-1 positive		ESC-1 negative		All Samples
GSE14333	Number	64	28.3%	162	71.7%	226
	Mean DFS Time ^	52.92	-	39.8	-	43.52
	Dead	7	14.0%	43	86.0%	50
	Alive	57	32.4%	119	67.6%	176
GSE17536	Number	53	29.9%	124	70.1%	177
	Mean DFS Time *	48.65	-	32.79	-	37.54
	Dead	12	21.8%	43	78.2%	55
	Alive	41	33.6%	81	66.4%	122
GSE17537	Number	29	52.7%	26	47.3%	55
	Mean DFS Time	44.35	-	29.75	-	32.94
	Dead	3	15.0%	17	85.0%	20
	Alive	26	74.3%	9	25.7%	35
GSE33113	Number	32	36.0%	57	64.0%	89
	Median Time to Meta	1222	-	1175	-	1184
	Metastasis Yes	7	38.9%	11	61.1%	18
	Metastasis No	25	35.2%	46	64.8%	71

Note: For ESC-1+ve and -ve groups T-Test P value ^ = 0.00442, *= p value : 0.05405.

Table V.III – Table summarizing the survival information in the respective datasets to give a bird's eye view. It is worth keeping in mind that in the datasets the ESC-1 +ve samples are less in number (~ 30% of total).

It seems that ESC-1 negative patients of colon cancer have a poorer survival related statistics i.e. higher percentage of dead, low Disease Free Survival, higher occurrence of metastasis, but this requires more investigation. Since cancer is so heterogeneous[74] it is rational to test a hypothesis on different cancers instead of attempting to extrapolate findings from one cancer type to others. This is what is going to be done in the future course of time (Future plans mentioned in section 6).

Section 6) Conclusions & Future

6. *Conclusions & Future*

Throughout the project work different challenges were encountered which required changes to be made to the research strategy to overcome them. This project gave insight into how the gene ESC-1 and its partners are expressed in various cancer types especially colon cancer, in which we found a subset of samples expressing ESC-1 at high levels which was previously unknown. We also found that it is not necessary that ESC-1 expressing samples will also be expressing ESCP-12 or ESCP-14, the two important embryonic pluripotency factors.

There were several cases where there was no presence of ESCP-12 and ESCP-14 in colon cancer (Fig 5.9 and 5.11), but there were certain cases in seminoma where ESCP-14 and ESCP-12 both were being expressed in the same set of samples where ESC-1 was being highly expressed (Fig. 5.10), which is a new finding as well. We found the differentially expressed genes between the samples expressing ESC-1 and the samples not expressing ESC-1, where we found that there are several interaction partners of ESC-1 that are differentially expressed (Fig. 5.19). We also established that many interaction partners are co-expressed in a wide range of samples, by demonstrating a positive statistical co-relation (Fig. 5.3 and Table V.I). The studies with the survival information show that ESC-1 alone is not able to explain the poor survival of patients, instead it might be a positive prognostic factor after all.

But as *Statistical significance does not imply causation*, more research is being done with the survival information so that it is possible to narrow down the list of genes that explain the survival of the patients along with ESC-1 (because ESC-1 alone is not able to explain the survival of the patients). Our collaborators will then co-stain tumour samples with similar properties (e.g. Type of cancer, Stage at diagnosis, Age, etc.) to verify if those genes are actually expressed in the same samples.

If they do succeed in verifying this *in-vitro*, then we plan to build a model that would predict the probability of survival of the patients based on the gene expression profile. Initially it would be cancer specific because we do not want to generalize findings too soon. However, on further investigation we hope to gain more knowledge and build a unified model that would predict the survival based on gene expression.

The amount of sequencing data publicly available for analysis is increasing with time. Also, there is a trend where the world is inching away from microarray based studies towards

next generation sequencing (NGS)[75]. This is because NGS it is a one stop solution for sequence information, gene expression quantification with high sensitivity[75], and it also assists in analysing mutations, so in the future the findings of this thesis and the methodology too, might be adapted to incorporate Next Generation Sequencing information. But, these are rather long term goals and applications of the findings of this thesis.

References

[1] "**Cancer Research UK, Cancer Statistics.**" [Online]. Available: http://www.wcrf.org/cancer_statistics/world_cancer_statistics.php.

[2] "**Cancer Society of Finland.**" [Online]. Available: http://www.cancer.fi/syoparekisteri/en/statistics/cancer-statistics/koko-maa/.

[3] T. Lapidot, C. Sirard, J. Vormoor, B. Murdoch, T. Hoang, J. Caceres-Cortes, M. Minden, B. Paterson, M. A. Caligiuri, and J. E. Dick, "**A cell initiating human acute myeloid leukaemia after transplantation into SCID mice.,**" *Nature*, vol. 367, no. 6464, pp. 645–8, Feb. 1994.

[4] C. A. O'Brien, A. Pollett, S. Gallinger, and J. E. J. Dick, "**A human colon cancer cell capable of initiating tumour growth in immunodeficient mice.,**" *Nature*, vol. 445, no. 7123, pp. 106–10, Jan. 2007.

[5] B. M. Boman and M. S. Wicha, "**Cancer stem cells: a step toward the cure.,**" *J. Clin. Oncol.*, vol. 26, no. 17, pp. 2795–9, Jun. 2008.

[6] S. Zhang, C. Balch, M. Chan, and H. Lai, "**Identification and characterization of ovarian cancer-initiating cells from primary human tumors,**" *Cancer Res.*, vol. 68, no. 11, pp. 4311–20, Jun. 2008.

[7] I. Ben-Porath, M. Thomson, V. Carey, and R. Ge, "**An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors,**" *Nat. …*, vol. 40, no. 5, pp. 499–507, May 2008.

[8] S. M. Kumar, S. Liu, H. Lu, H. Zhang, P. J. Zhang, P. A. Gimotty, M. Guerra, W. Guo, and X. Xu, "**Acquired cancer stem cell phenotypes through Oct4-mediated dedifferentiation.,**" *Oncogene*, vol. 31, no. 47, pp. 4898–911, Nov. 2012.

[9] International Agency for Research on Cancer, "**Latest world cancer statistics Global cancer burden rises to 14 . 1 million new cases in 2012 : Marked increase in breast cancers must be addressed.,**" *Press Release, World Heal. Organ.*, no. December, 2013.

[10] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, "**Molecular Cell Biology.**" W. H. Freeman, p. 973, 2000.

[11] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of The Cell.*, 5th. ed. Garland Science, 2008, p. 1601.

[12] T. Strachan and A. P. Read, *Human Molecular Genetics*, 2nd Editio. New York: Wiley-Liss, 1999.

[13] N. Petrucelli, M. B. Daly, and G. L. Feldman, "**BRCA1 and BRCA2 Hereditary Breast and Ovarian Cancer.**" University of Washington, Seattle, 26-Sep-2013.

[14] D. Hanahan and R. A. Weinberg, "**The Hallmarks of Cancer Review,**" *Cell*, vol. 100, pp. 57–70, 2000.

[15] D. J. Slamon, G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, and W. L. McGuire, "**Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene.,**" *Science*, vol. 235, no. 4785, pp. 177–82, Jan. 1987.

[16] A. J. Butt, S. M. Firth, and R. C. Baxter, "**The IGF axis and programmed cell death,**" in *Immunology and Cell Biology*, 1999, vol. 77, pp. 256–262.

[17] G. Evan and T. Littlewood, "**A matter of life and cell death,**" *Science (80-.).*, vol. 281, pp. 1317–1322, 1998.

[18] L. Hayflick, "**Mortality and immortality at the cellular level. A review.,**" *Biochemistry. (Mosc).*, vol. 62, pp. 1180–1190, 1997.

[19] T. M. Bryan and T. R. Cech, "**Telomerase and the maintenance of chromosome ends,**" *Current Opinion in Cell Biology*, vol. 11. pp. 318–324, 1999.

[20] H. Vaziri and S. Benchimol, "**Reconstitution of telomerase activity in normal human cells leads to elongation of telomeres and extended replicative life span.,**" *Curr. Biol.*, vol. 8, pp. 279–282, 1998.

[21] A. G. Bodnar, M. Ouellette, M. Frolkis, S. E. Holt, C. P. Chiu, G. B. Morin, C. B. Harley, J. W. Shay, S. Lichtsteiner, and W. E. Wright, "**Extension of life-span by introduction of telomerase into normal human cells.,**" *Science*, vol. 279, pp. 349–352, 1998.

[22] M. B. Sporn, "**The war on cancer: A review,**" in *Annals of the New York Academy of Sciences*, 1997, vol. 833, pp. 137–146.

[23] D. Hanahan and J. Folkman, "**Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis,**" *Cell*, vol. 86. pp. 353–364, 1996.

[24] M. Veta, J. P. W. Pluim, P. J. Van Diest, and M. A. Viergever, "**Breast cancer histopathology image analysis: A review,**" *IEEE Transactions on Biomedical Engineering*, vol. 61. pp. 1400–1411, 2014.

[25] A. Buchali, D. Geismar, M. Hinkelbein, L. Schlenger, K. Zinner, and V. Budach, "**Virtual simulation in patients with breast cancer,**" *Radiother. Oncol.*, vol. 59, pp. 267–272, 2001.

[26] D. G. Fryback, N. K. Stout, M. A. Rosenberg, A. Trentham-Dietz, V. Kuruchittham, and P. L. Remington, "**The Wisconsin Breast Cancer Epidemiology Simulation Model.,**" *J. Natl. Cancer Inst. Monogr.*, pp. 37–47, 2006.

[27] S. Nobili, D. Lippi, E. Witort, M. Donnini, L. Bausi, E. Mini, and S. Capaccioli, "**Natural compounds for cancer treatment and prevention.,**" *Pharmacol. Res.*, vol. 59, pp. 365–378, 2009.

[28] M. J. V Vähä-Koskela, J. E. Heikkilä, and A. E. Hinkkanen, "**Oncolytic viruses in cancer therapy.,**" *Cancer Lett.*, vol. 254, no. 2, pp. 178–216, Sep. 2007.

[29] D. Melton, *StemBook*. Harvard Stem Cell Institute, 2008.

[30] S. F. Gilbert, *Developmental Biology*, 6th Editio. Sunderland (MA): Sinauer Associates, 2000.

[31] J. A. Thomson, "**Embryonic Stem Cell Lines Derived from Human Blastocysts,**" *Science (80-.).*, vol. 282, no. 5391, pp. 1145–1147, Nov. 1998.

[32] K. Takahashi and S. Yamanaka, "**Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.,**" *Cell*, vol. 126, no. 4, pp. 663–76, Aug. 2006.

[33] K. Takahashi, K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka, K. Tomoda, and S. Yamanaka, "**Induction of pluripotent stem cells from adult human fibroblasts by defined factors.,**" *Cell*, vol. 131, no. 5, pp. 861–72, Nov. 2007.

[34] J. Yu, M. A. Vodyanik, K. Smuga-Otto, J. Antosiewicz-Bourget, J. L. Frane, S. Tian, J. Nie, G. A. Jonsdottir, V. Ruotti, R. Stewart, I. I. Slukvin, and J. A. Thomson, "**Induced pluripotent stem cell lines derived from human somatic cells.,**" *Science*, vol. 318, no. 5858, pp. 1917–20, Dec. 2007.

[35] D. L. Clarke, C. B. Johansson, J. Wilbertz, B. Veress, E. Nilsson, H. Karlström, U. Lendahl, and J. Frisén, "**Generalized potential of adult neural stem cells.,**" *Science*, vol. 288, no. 5471, pp. 1660–3, Jun. 2000.

[36] B. Suárez-Álvarez, A. López-Vázquez, and C. López-Larrea, "**Mobilization and homing of hematopoietic stem cells.,**" *Adv. Exp. Med. Biol.*, vol. 741, pp. 152–70, Jan. 2012.

[37] R. Z. Yusuf and D. T. Scadden, "**Homing of hematopoietic cells to the bone marrow.,**" *Journal of visualized experiments : JoVE*. 2009.

[38] U. Lakshmipathy and C. Verfaillie, "**Stem cell plasticity.,**" *Blood Rev.*, vol. 19, no. 1, pp. 29–38, Jan. 2005.

[39] T. Reya, S. J. Morrison, M. F. Clarke, and I. L. Weissman, "**Stem cells, cancer, and cancer stem cells.,**" *Nature*, vol. 414, no. 6859, pp. 105–11, Nov. 2001.

[40] S. Heneidi, A. a Simerman, E. Keller, P. Singh, X. Li, D. a Dumesic, and G. Chazenbalk, "**Awakened by cellular stress: isolation and characterization of a novel population of pluripotent stem cells derived from human adipose tissue.,**" *PLoS One*, vol. 8, no. 6, p. e64752, Jan. 2013.

[41] S. a Mousa, T. Sudha, E. Dyskin, U. Dier, C. Gallati, C. Hanko, S. V Chittur, and A. Rebbaa, "**Stress resistant human embryonic stem cells as a potential source for the identification of novel cancer stem cell markers.,**" *Cancer Lett.*, vol. 289, no. 2, pp. 208–16, Mar. 2010.

[42] M. Ito, H. Hiramatsu, K. Kobayashi, K. Suzue, M. Kawahata, K. Hioki, Y. Ueyama, Y. Koyanagi, K. Sugamura, K. Tsuji, T. Heike, and T. Nakahata, "**NOD/SCID/gamma(c)(null) mouse: an excellent recipient mouse model for engraftment of human cells.,**" *Blood*, vol. 100, no. 9, pp. 3175–82, Nov. 2002.

[43] M. Dean, T. Fojo, and S. Bates, "**Tumour stem cells and drug resistance.,**" *Nat. Rev. Cancer*, vol. 5, no. 4, pp. 275–84, Apr. 2005.

[44] L. Seguin, S. Kato, A. Franovic, M. F. Camargo, J. Lesperance, K. C. Elliott, M. Yebra, A. Mielgo, A. M. Lowy, H. Husain, T. Cascone, L. Diao, J. Wang, I. I. Wistuba, J. V Heymach, S. M.

Lippman, J. S. Desgrosellier, S. Anand, S. M. Weis, and D. A. Cheresh, "**An integrin β_3-KRAS-RalB complex drives tumour stemness and resistance to EGFR inhibition.,**" *Nat. Cell Biol.*, vol. 16, no. 5, pp. 457–68, May 2014.

[45] L. Ricci-Vitiani, E. Fabrizi, E. Palio, and R. De Maria, "**Colon cancer stem cells,**" *J. Mol. …*, pp. 16–23, 2009.

[46] M. Schwede, D. Spentzos, and S. Bentink, "**Stem cell-like gene expression in ovarian cancer predicts type ii subtype and prognosis,**" *PLoS One*, vol. 8, no. 3, p. e57799, Jan. 2013.

[47] I. H. Janna Saarela, "**Affymetrix GeneChip System,**" in *DNA Microarray Data Analysis*, 2nd. ed., 2005, pp. 25–31.

[48] "**TCGA Data Portal.**" [Online]. Available: https://tcga-data.nci.nih.gov.

[49] "**TCGA Colon Adenocarcinoma.**" [Online]. Available: https://tcga-data.nci.nih.gov/tcga/tcgaCancerDetai.

[50] "**TCGA Lung Adenocarcinoma.**" [Online]. Available: https://tcga-data.nci.nih.gov/tcga/tcgaCancerDetails.jsp?diseaseType=LUAD&diseaseName=Lung adenocarcinoma.

[51] "**TCGA Lung Small Cell Carcinoma.**" [Online]. Available: https://tcga-data.nci.nih.gov/tcga/tcgaCancerDetails.jsp?diseaseType=LUSC&diseaseName=Lung squamous cell carcinoma.

[52] "**TCGA Ovarian Serous Cystadenocarcinoma.**" [Online]. Available: https://tcga-data.nci.nih.gov/tcga/tcgaCancerDetails.jsp?diseaseType=OV&diseaseName=Ovarian serous cystadenocarcinoma.

[53] "**TCGA Glioblastoma Multiforme.**" [Online]. Available: https://tcga-data.nci.nih.gov/tcga/tcgaCancerDetails.jsp?diseaseType=GBM&diseaseName=Glioblastoma multiforme.

[54] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "**NCBI GEO: archive for functional genomics data sets--update.,**" *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D991–5, Jan. 2013.

[55] R. Edgar, "**Gene Expression Omnibus: NCBI gene expression and hybridization array data repository,**" *Nucleic Acids Res.*, vol. 30, no. 1, pp. 207–210, Jan. 2002.

[56] J. E. J. Korkola, J. Houldsworth, R. S. V Chadalavada, A. B. Olshen, D. Dobrzynski, V. E. Reuter, G. J. Bosl, and R. S. K. Chaganti, "**Down-regulation of stem cell genes, including those in a 200-kb gene cluster at 12p13. 31, is associated with in vivo differentiation of human male germ cell tumors,**" *Cancer Res.*, vol. 66, no. 2, pp. 820–827, Jan. 2006.

[57] J. Sabates-Bellver and L. Van der Flier, "**Transcriptome profile of human colorectal adenomas,**" *Mol. Cancer …*, vol. 5, no. 12, pp. 1263–1275, Dec. 2007.

[58] "**Expression Project for Oncology.**" [Online]. Available: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2109.

[59] R. N. Jorissen, P. Gibbs, M. Christie, S. Prakash, L. Lipton, J. Desai, D. Kerr, L. A. Aaltonen, D. Arango, M. Kruhøffer, T. F. Orntoft, C. L. Andersen, M. Gruidl, V. P. Kamath, S. Eschrich, T. J. Yeatman, and O. M. Sieber, "**Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer.,**" *Clin. Cancer Res.*, vol. 15, no. 24, pp. 7642–7651, Dec. 2009.

[60] J. J. Smith, N. G. Deane, F. Wu, N. B. Merchant, B. Zhang, A. Jiang, P. Lu, J. C. Johnson, C. Schmidt, C. E. Bailey, S. Eschrich, C. Kis, S. Levy, M. K. Washington, M. J. Heslin, R. J. Coffey, T. J. Yeatman, Y. Shyr, and R. D. Beauchamp, "**Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer.,**" *Gastroenterology*, vol. 138, no. 3, pp. 958–68, Mar. 2010.

[61] K. Kemper, M. Versloot, K. Cameron, S. Colak, F. de Sousa e Melo, J. H. de Jong, J. Bleackley, L. Vermeulen, R. Versteeg, J. Koster, and J. P. Medema, "**Mutations in the Ras-Raf Axis underlie the prognostic value of CD133 in colorectal cancer.,**" *Clin. Cancer Res.*, vol. 18, no. 11, pp. 3132–41, Jun. 2012.

[62] and M. M. L. Jarno Tuimala, Ilana Saarikko, "**Normalization,**" in *DNA Microarray Data Analysis*, 2nd. ed., 2005, pp. 99–114.

[63] R. Irizarry, B. Hobbs, and F. Collin, "**Exploration, normalization, and summaries of high density oligonucleotide array probe level data,**" *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[64] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, "**affy--analysis of Affymetrix GeneChip data at the probe level.,**" *Bioinformatics*, vol. 20, no. 3, pp. 307–15, Feb. 2004.

[65] L. Gautier, R. Irizarry, L. Cope, and B. Bolstad, "**Description of affy,**" pp. 1–29, 2013.

[66] S. Piccolo and M. Withers, "**Multiplatform single-sample estimates of transcriptional activation,**" *Proc. …*, vol. 110, no. 44, pp. 17778–83, Oct. 2013.

[67] D. C. LeBlanc, *Statistics Concepts and Applications for Science*. Jones & Bartlett Learning, 2004, p. 382.

[68] R. R. Stoll, *Set Theory and Logic*. Courier Dover Publications, 2012, p. 496.

[69] O. Kramer, "**Principal Component analysis,**" in *Dimensionality Reduction with Unsupervised Nearest Neighbors*, Springer Science & Business Media, 2013, pp. 39–40.

[70] R. A. Fisher, "**THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS,**" *Ann. Eugen.*, vol. 7, no. 2, pp. 179–188, 1936.

[71] L. Elo, J. Hiissa, J. Tuimala, A. Kallio, E. Korpelainen, and T. Aittokallio, "**Optimized detection of differential expression in global profiling experiments: case studies in clinical transcriptomic and quantitative proteomic datasets.,**" *Brief. Bioinform.*, vol. 10, no. 5, pp. 547–55, Sep. 2009.

[72] L. Elo, S. Filén, R. Lahesmaa, and T. Aittokallio, "**ROTS Documentation,**" 2008. [Online]. Available: http://www.btk.fi/fileadmin/Page_files/Research/compbiomed/ROTS_1.1.1.pdf.

[73] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. M. Chinnaiyan, "**ONCOMINE: a cancer microarray database and integrated data-mining platform.,**" *Neoplasia*, vol. 6, no. 1, pp. 1–6.

[74] R. Fisher, L. Pusztai, and C. Swanton, "**Cancer heterogeneity: implications for targeted therapeutics.,**" *Br. J. Cancer*, vol. 108, no. 3, pp. 479–85, Feb. 2013.

[75] A. Sîrbu, G. Kerr, M. Crane, and H. J. Ruskin, "**RNA-Seq vs dual- and single-channel microarray data: sensitivity analysis for differential expression and clustering.,**" *PLoS One*, vol. 7, no. 12, p. e50986, Jan. 2012.

[76] "**Affymetrix GeneChip Output, Sample Image.**" [Online]. Available: http://media.affymetrix.com/_media/corporate/media/image_library/low_res/genechiparrayoutput.jpg.

[77] "**'Affymetrix-microarray'. Licensed under Creative Commons Attribution 2.5 via Wikimedia Commons.**"[Online]. Available: http://commons.wikimedia.org/wiki/File:Affymetrix-microarray.jpg#mediaviewer/File:Affymetrix-microarray.jpg.