

https://dspace.org/

# In silico discovery of transcription factors as breast cancer biomarkers

**Perumal, Shanen**

**2019**
https://hdl.handle.net/10539/27059

Downloaded from a DSpace repository. https://dspace.org/

School of Molecular and Cell Biology

University of the Witwatersrand,

Johannesburg, Private Bag 3, WITS

2050, South Africa

# *In Silico* Discovery of Transcription Factors as Breast Cancer Biomarkers

MSc Research Dissertation

February 2019

Shanen Perumal

705672

Supervisor: Professor Mandeep Kaur

Advisor: Dr Vanessa Meyer

Postgraduate co-ordinator: Dr Vanessa Meyer

Student Signature: 

A dissertation submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science.

# Table of Contents

DECLARATION ..... i
ACKNOWLEDGEMENTS ..... ii
DISCLAIMER ..... ii
LIST OF SYMBOLS ..... iii
LIST OF ABBREVIATIONS ..... iii
LIST OF FIGURES ..... v
LIST OF TABLES ..... v
ABSTRACT ..... 1
CHAPTER ONE – INTRODUCTION ..... 2
1.1 Background and Literature Review ..... 2
1.1.1 The Hallmarks of Cancer ..... 2
1.1.2 Prevalence of Breast Cancer ..... 2
1.1.3 Breast Cancer Subtypes ..... 3
1.1.4 Stages of Breast Cancer ..... 4
1.1.5 Mechanisms of Gene Expression ..... 5
1.1.6 Regulation of Gene Expression by TFs ..... 6
1.1.7 TFs as Biomarkers ..... 8
1.1.8 The Tumour Suppressor p53 ..... 9
1.1.9 The Estrogen Receptor ..... 9
1.1.10 Other TFs as Biomarkers ..... 10
1.1.11 Detection of TFs for use in Cancer Diagnostics ..... 11
1.1.12 Experimental Methods to Confirm TF binding ..... 11
1.1.13 The Transcriptome: Quantification and Analysis ..... 12
1.1.14 Bioinformatics for Transcriptome Analysis ..... 14
1.2 Introduction to the Present Study ..... 16
1.3 Aims and Objectives ..... 18
1.3.1 Aim ..... 18
1.3.2 Objectives ..... 18
CHAPTER TWO - MATERIALS AND METHODS ..... 19
2.1 Methodology Workflow ..... 19
2.2 Retrieval of Gene Expression Data ..... 20
2.3 Differential Expression Analysis ..... 22
2.4 Promoter Sequence Acquisition ..... 24
2.5 ERE Prediction ..... 24
2.6 Functional Enrichment ..... 25
2.7 TFBS enrichment ..... 26

2.8 Oncomine Validation ..... 29
2.9 Survival Analysis ..... 29
2.10 Network Construction ..... 30
2.11 Predictive Value of Prospective Biomarkers ..... 31
CHAPTER THREE – RESULTS ..... 32
3.1 Differential Expression ..... 32
3.2 Functional Enrichment ..... 39
3.3 TFBS Enrichment ..... 40
3.4 Oncomine Validation ..... 44
3.5 Survival Analysis ..... 46
3.6 TF-Gene Networks ..... 47
3.7 Predictive Value of Prospective Biomarkers ..... 53
CHAPTER FOUR – DISCUSSION AND CONCLUSIONS ..... 54
4.1 Altered Gene Expression in Breast Cancer ..... 54
4.2 TFs as Controllers of Cancer Gene Expression ..... 57
4.3 Prediction of Known Breast Cancer Biomarkers ..... 57
4.3.1 E2F1 as a Breast Cancer Biomarker ..... 57
4.3.2 MYC and MAX as Breast Cancer Biomarkers ..... 59
4.4 Identification of Unknown Breast Cancer Biomarkers ..... 60
4.4.1 INSM1 as a Breast Cancer Biomarker ..... 60
4.4.2 FOXD1 as a Breast Cancer Biomarker ..... 61
4.4.3 TAL1 as a Breast Cancer Biomarker ..... 62
4.4.4 RUNX1 as a Breast Cancer Biomarker ..... 64
4.5 Conclusions ..... 65
4.6 Troubleshooting ..... 66
4.6.1 Differential Expression ..... 66
4.6.2 ERE Prediction ..... 66
4.6.3 Scripting ..... 67
4.7 Future Studies ..... 67
4.8 Limitations ..... 68
Appendix ..... 69
References ..... 70

# DECLARATION

I, Shanen Perumal (705672), am a student registered for the degree of Master of Science (MSc) by research in the academic year 2017-2018.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where explicitly indicated otherwise and acknowledged.
- I have not submitted this work before for any other degree or examination at this or any other University.
- The information used in the Dissertation HAS NOT been obtained by me while employed by, or working under the aegis of, any person or organisation other than the University.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

NRF Declaration

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged.

Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be contributed to the NRF.

Signature:  5th day of February 2019

## ACKNOWLEDGEMENTS

I would like to acknowledge the following people and institutions for their support and assistance without which this research project would not be possible:

I am grateful to my supervisor, Professor Mandeep Kaur, for her guidance, mentorship, and persistent efforts which have made this research possible. Thank you for your support and motivation.

I thank my close family and friends for supporting me emotionally and standing by me throughout this journey. A special thanks goes to my mother, Sharlin Perumal, for her constant support throughout my life.

My gratitude goes out to my colleagues in GH527 for their advice and support. I am grateful to have had the opportunity to work with all of you.

I acknowledge the University of Witwatersrand for providing the facilities in which I worked and for financial assistance through the Postgraduate Merit Award. I also acknowledge the National Research Foundation for financial support (NRF Freestanding, Innovation and Scarce Skills Development Masters Scholarship).

## DISCLAIMER

Please treat the contents of this dissertation as confidential. The information contained in this document may not be used, published or redistributed without the prior written consent of the author.

# LIST OF SYMBOLS

| | |
|---|---|
| α | Alpha |
| β | Beta |
| κ | Kappa |
| % | Percent |

# LIST OF ABBREVIATIONS

| | |
|---|---|
| A | Adenine |
| BCV | Biological Coefficient of Variation |
| BLAST | Basic Local Alignment Search Tool |
| bp | Base Pair |
| C | Cytosine |
| CAGE | Cap Analysis Gene Expression |
| cDNA | Complimentary DNA |
| ChIP | Chromatin Immunoprecipitation |
| CPM | Counts Per Million |
| DAVID | Database for Annotation, Visualization and Integrated Discovery |
| DBD | DNA-Binding Domain |
| DE | Differential Expression |
| DNA | Deoxyribonucleic Acid |
| e.g. | For Example |
| EGFR | Epidermal Growth Factor Receptor |
| EMSA | Electrophoretic Mobility Shift Assay |
| EMT | Epithelial to Mesenchymal Transition |
| EPD | Eukaryotic Promoter Database |
| ERα | Estrogen Receptor Alpha |
| ER- | Estrogen Receptor Negative |
| ER+ | Estrogen Receptor Positive |
| ERE | Estrogen Response Element |
| FC | Fold Change |
| FDA | Food and Drug Administration |
| FDR | False Discovery Rate |
| FTP | File Transfer Protocol |
| G | Guanine |
| GEO | Gene Expression Omnibus |
| GO | Gene Ontology |
| GPL | GEO Platform |

| | |
|---|---|
| GSE | GEO Series |
| GSM | GEO Sample |
| GTF | Gene Transfer Format |
| HER2 | Human Epidermal Growth Factor Receptor 2 |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| KM | Kaplan-Meier |
| logFC | log2 Fold Change |
| MD | Mean Difference |
| MDS | Multi-Dimensional Scaling |
| miRNA | Micro RNA |
| mRNA | Messenger RNA |
| NB | Negative Binomial |
| NCBI | National Centre for Biotechnology Information |
| NGS | Next Generation Sequencing |
| OC | Ovarian Cancer |
| PFM | Position Frequency Matrix |
| polyA | Poly Adenylation |
| PR | Progesterone Receptor |
| PSFM | Position Specific Frequency Matrix |
| PWM | Position Weight Matrix |
| QL | Quasi Likelihood |
| RNA | Ribonucleic Acid |
| RNA pol II | RNA Polymerase II |
| RNA-seq | RNA Sequencing |
| rRNA | Ribosomal RNA |
| SAM | Sequence Alignment Map |
| SERM | Selective Estrogen Receptor Modulator |
| SRA | Sequence Read Archive |
| SSA | Single Site Analysis |
| SSD | Signal Sensing Domain |
| T | Thymine |
| TAD | Trans-Activating Domain |
| TAF | TATA-Binding Protein Associated Factor |
| TBP | TATA-Binding Protein |
| TCGA | The Cancer Genome Atlas |
| TF | Transcription Factor |
| TFBS | Transcription Factor Binding Site |
| TIC | Transcription Initiation Complex |
| TMM | Trimmed Mean of M-Values |
| TSS | Transcription Start Site |

## LIST OF FIGURES

Figure 1: Workflow of Methodology.....19
Figure 2: Representation of a PSFM.....27
Figure 3: MDS plot representing the differences in mRNA levels between RNA-seq samples .....33
Figure 4: MD plot showing logFC compared to average abundance of each gene. .....34
Figure 5: Heat maps of the expression levels of the most DE genes across all samples .....36
Figure 6: Bar graphs representing GO enrichment in upregulated gene sets .....37
Figure 7: Bar graphs representing KEGG pathway in upregulated gene sets.....38
Figure 8: Survival analysis of predicted biomarkers in MCF7 cells .....48
Figure 9: Survival analysis of predicted biomarkers in MDA-MB-231 cells.....49
Figure 10: Analysis of prognostic value of predicted biomarkers for breast cancer metastasis .....50
Figure 11: Network of TF-gene interactions .....51
Figure 12: Network of TF-gene interactions .....52
Figure S1: Scatterplot of BCV against transcript abundance across samples. .....69

## LIST OF TABLES

Table 1: RNA-seq samples retrieved from GEO .....21
Table 2: Number of up- and downregulated DE genes.....34
Table 3: TFBSs Enriched in DE genes upregulated in MCF7 vs MCF10A.....41
Table 4: TFBSs Enriched in DE genes upregulated in MDA-MB-231 vs MCF10A .....42
Table 5: TFBSs Enriched in Predicted ER-regulated genes upregulated in MCF7 vs. MCF10A .....43
Table 6: TFBSs Enriched in Predicted ER-regulated genes upregulated in MDA-MB-231 vs. MCF10A .....44
Table 7: Predicted Biomarkers for MCF7 Showing DE in Oncomine Breast Cancer vs. Normal Datasets.....45
Table 8: Predicted Biomarkers for MDA-MB-231 Showing DE in Oncomine Breast Cancer vs. Normal Datasets .....45
Table 9: Predictive value of prospective biomarkers.....53

# ABSTRACT

Breast cancer is the most prevalent type of cancer affecting women. This disease is grouped into subtypes with different gene expression profiles, which affect the response to treatment and the prognosis of breast cancer patients. Estrogen receptor negative (ER-) breast cancer subtypes generally have a poor patient prognosis due to the lack of targeted treatment options and the high relapse rate after chemotherapy. The present study is aimed at computationally evaluating the differences in gene transcription between ER positive (ER+) and ER- breast cancer subtypes and identifying transcription factors (TFs) controlling these differences. RNA-sequencing data was obtained from publically available databases for MCF7, MDA-MB-231 and MCF10A cell lines, representing ER+ breast cancer, ER- breast cancer and non-tumorigenic breast cells respectively. Differentially expressing genes were selected by comparing the gene expression profiles of cancer cell lines to non-tumorigenic cells. Functional enrichment was performed using gene ontology and KEGG pathways to identify the biological roles the differentially expressing genes play in breast cancer. The promoters of differentially expressing genes were assessed for TF binding site (TFBS) enrichment to identify the transcriptional controllers of breast cancer-related gene expression. The expression of the TFs selected as key regulators was validated in breast cancer patient datasets. The prognostic value of the TFs upregulated in breast cancer patient data was evaluated using patient survival data. Potential biomarkers were selected based on prognostic value. E2F1, INSM1 and MYC were predicted as potential biomarkers from MCF7 expression data and FOXD1, TAL1, RUNX1 and MAX were predicted using MDA-MB-231 data. Finally, networks were constructed to visualise the interactions between potential TF biomarkers and the genes that they regulate. This preliminary prediction of TF biomarkers could provide a better understanding of the molecular mechanisms governing the characteristics of different breast cancer subtypes, and could be used as novel biomarkers for breast cancer with diagnostic and therapeutic potential after further validation using patient tumour samples.

# CHAPTER ONE – INTRODUCTION

## 1.1 Background and Literature Review

### 1.1.1 The Hallmarks of Cancer

Cancer is a group of diseases attributed to the loss of control mechanisms governing the normal growth and proliferation of cells in the body. Hundreds of varieties of cancer exist, affecting different tissues, each having vastly different characteristics, symptoms and prognoses. Cancer has affected humans for thousands of years and yet still remains a leading cause of death in both developing and developed countries today (Faltas, 2011). Development of cancer is a multi-step process, requiring multiple exposures to carcinogenic agents and influence from several genetic predispositions, ultimately leading to a loss of the homeostatic properties of cells and tissues. The complex characteristics of neoplastic disease can be summarised into a group of hallmarks that define the behaviour of cancer cells in contrast to normal cells. These include chronic proliferation, failure to respond to growth suppressors, and avoiding cell death and senescence, each enabling the accumulation of cells required for tumour formation (Hanahan and Weinberg, 2011). Invasion and metastasis, as well as the induction of angiogenesis, further allow cancer to grow and spread throughout the body. Although these hallmarks are common to most cancers, the molecular mechanisms controlling each type of cancer differ, leading to varying levels of aggression.

### 1.1.2 Prevalence of Breast Cancer

Breast cancer is the most prevalent cancer among women globally, with an estimated 2.088 million cases diagnosed in 2018 (Bray et al., 2018). It has the highest mortality rate among females globally, with an estimated 626 679 deaths attributed to breast cancer in 2018 alone (Bray et al., 2018). In South Africa, breast cancer is the leading cancer in females, making up 21.78% of total cancer cases diagnosed in the year 2014 (National Cancer Registry, 2014). In the year 2018, 14 097 cases of breast cancer were diagnosed and it is possible that many cases remain undiagnosed due to limitations in our developing healthcare system (Bray et al., 2018). According to the National Cancer Registry, 1 in 27 South African females is at risk of developing breast cancer, illustrating an urgent need for improved clinical solutions (National Cancer Registry, 2014).

### 1.1.3 Breast Cancer Subtypes

Breast cancer is a heterogeneous group of diseases in which cells in the breast exhibit abnormal growth and division. The human breast is composed primarily of adipose and glandular tissue. Within the breast, milk-producing cells called lactocytes are contained within alveoli, which cluster to form lobules. These lobules form the lobes, which connect to lactiferous ducts to drain the milk from the lobules toward the nipple (Ramsay et al., 2005). Breast cancer predominantly develops from the lobules and ducts, known as lobular carcinoma and ductal carcinoma respectively. Breast cancer cases each differ vastly in their prognosis and response to therapeutic agents. Similar cases of breast cancer are grouped into subtypes, which helps physicians decide on therapeutic strategies and predict patient outcome for each different case (Cancer Genome Atlas Network, 2012).

Breast cancer subtypes are commonly classified based on the gene expression profiles of the tumours – most notably the expression of the estrogen receptor alpha (ERα), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) proteins (Perou et al., 2000). Using these markers, breast cancer is subdivided into the intrinsic subtypes: luminal A, luminal B, basal-like, and HER2-expressing breast cancers (Perou et al., 2000). The proteins used to classify these subtypes offer a useful diagnostic tool for breast cancer classification, but have limited utility as therapeutic targets in many cases.

Luminal subtypes are the most common breast cancers (O'Brien et al., 2010). These tumours overexpress ERα and PR and are therefore usually responsive to hormone therapy, but respond poorly to most chemotherapy (Brenton et al., 2005). HER2 over-expressing tumours often do not express ERα and PR, but over-express HER2 – a cell-surface receptor which promotes cell proliferation and growth (O'Brien et al., 2010). They are not responsive to hormone therapy and generally have a poor prognosis. They are, however, more responsive to taxane and anthracycline-based chemotherapeutic agents than luminal subtypes (Brenton et al., 2005). HER2 over-expressing cancers can also be treated with HER2-tageting drugs, such as trastuzumab, but this is not effective in all cases (Nagata et al., 2004). The basal subtype includes triple negative tumours, which do not express ERα, PR or HER2. These tumours have a high expression of basal markers, such as EGFR (epidermal growth factor receptor), keratins and genes related to cell proliferation (Sotiriou et al., 2003). Triple negative tumours are usually aggressive and lack targeted therapeutic options. Patients with this type of cancer have a high relapse rate and a poor prognosis (Rakha et al., 2006). The only standard treatment for basal

subtypes is chemotherapy, such as anthracyclines (e.g. doxorubicin and epirubicin) and taxanes (e.g. paclitaxel and docetaxel) (Brenton et al., 2005). The lack of targeted treatment options for triple negative breast cancer contributes to its poor prognosis, highlighting a need for better understanding of the regulatory network that governs breast cancer aetiology. Numerous studies have been conducted to discover novel therapeutic markers for breast cancer (Morettin et al., 2015, Rizzo et al., 2008). These targets are usually regulators of gene expression, such as transcription factors (TFs), which can play distinct roles in different subtypes of breast cancer.

### 1.1.4 Stages of Breast Cancer

Breast cancer is also classified into stages, which describe the size, type and progression of the tumour (Heim et al., 1997). Stage 0 of breast cancer describes both cancerous and non-cancerous tumours which are non-invasive and small in size. This is the earliest stage of breast cancer and is usually easily treatable. Stage 0 is often observed in ductal carcinoma in situ, which is an early form of breast cancer with low risk of becoming invasive (Bednarek et al., 1997). At stage 1 of breast cancer, invasion is possible and the size of the tumour increases up to 2 cm (Segal et al., 2001). This stage is divided into stage 1A and 1B, with no lymph node involvement in stage 1A and groups of tumorigenic cells larger than 0.2 mm in the lymph nodes in stage 1B (Segal et al., 2001). In stage 2A of breast cancer, tumours are found in the axillary or sentinel lymph nodes, between 2 and 5 cm in size. Stage 2B tumours can be larger than 5 cm, but are not able to reach the axillary lymph nodes (Moran et al., 2014). There are three types of stage 3 breast cancer. Stage 3A indicates that no tumour is detected in the breast, but tumorigenic cells are found in between four and nine axillary or sentinel lymph nodes (Jacquillat et al., 1990). Stage 3B tumours have spread to up to nine lymph nodes and swelling of the skin of the breast is observed, indicating inflammation of the breast. Stage 3C tumours have spread to 10 or more lymph nodes and to tissues surrounding the clavicle (Jacquillat et al., 1990).

Stage 4 breast cancer is an advanced form of breast cancer with metastatic characteristics. During this stage of breast cancer, the tumour has spread to distant organs such as the brain, liver, bones or lungs. This advanced stage of cancer is difficult to treat and patients have a very poor prognosis (Neuman et al., 2010). Breast cancer at an advanced stage is often fatal, as the spread of tumorigenic cells from the primary tumour makes it difficult to treat. For this reason, early detection of breast cancer results in a favourable patient outcome. The use of biomarkers

is essential for the early diagnosis of breast cancer, as the small size and lack of symptoms at an early stage makes it difficult to detect the tumour directly.

### 1.1.5 Mechanisms of Gene Expression

The hereditary information required for the development and functioning of an organism is contained in sequences of DNA called genes. The genetic material each cell contains is identical; however, only a specific subset of genes is expressed in different cell types or under different conditions, resulting in a diverse range of phenotypes. The subset of genes expressed largely defines the state of a biological system. Gene expression begins with transcription, in which DNA is used as a template for the synthesis of an RNA product. Of the many types of RNA products synthesised during transcription, only a small proportion code for proteins (Palazzo and Lee, 2015). These coding RNAs are called messenger RNAs (mRNAs), and are used as an intermediate for the process of translation. During translation, mRNA is used as a template for the synthesis of proteins, which carry out structural and functional roles in the organism (Latchman, 1997).

Gene expression is commonly measured at the level of transcription, by quantifying the mRNA present in a cell at a given time. Transcription of mRNA is carried out by a twelve subunit enzyme, RNA polymerase II (RNA pol II), in three steps: initiation, elongation and termination. For transcription to occur, RNA pol II needs to bind to a specific region of the gene called the promoter. General transcription factors assemble at the promoter to form a transcription initiation complex (TIC) (Kornberg, 2007). The TIC recruits RNA pol II to the promoter and facilitates the unwinding of DNA, initiating transcription. During the elongation step of transcription, RNA pol II attaches nucleotides complementary to the DNA template to the mRNA chain by phosphodiester bonds (Brueckner et al., 2009). mRNA is also processed during the elongation step, with the addition of a 7-methylguanosine cap on the 5’ end and the removal of non-coding intron sequences through splicing. RNA pol II enzyme moves along the DNA template, continuously adding nucleotides to the growing mRNA chain. This continues until a specific DNA sequence is reached, which causes the activation of termination factors, which cause RNA pol II to cleave the 3’ end of the mRNA and attach a polyadenylated (polyA) tail, consisting of a stretch of adenine bases (Richard and Manley, 2009). This event marks the termination of transcription. The mature mRNA is exported from the nucleus into the cytoplasm, where it is used as a template for the synthesis of a protein product by ribosomes in a process known as translation.

### 1.1.6 Regulation of Gene Expression by TFs

The expression of genes is regulated by a number of molecules such as microRNAs (miRNAs) and TFs. TFs are proteins that regulate the rate of gene transcription by altering the stability of the TIC (Semenza, 1994). TFs regulate gene expression by binding specific sequences on *cis*-regulatory elements or motifs on the DNA of target genes and either promoting or repressing gene expression (Lelli et al., 2012). These *cis*-regulatory elements include promoters, enhancers, silencers and insulators, which each play a role in gene control. TFs bind to these motifs, form complexes by interacting with other TFs and recruit RNA polymerase II to initiate gene transcription. In eukaryotes, the promoter is a region of DNA that can contain multiple TF binding sites (TFBSs). These TFBSs are regions of DNA containing a highly conserved nucleotide sequence that is recognised by specific TFs. The presence of multiple TFBSs allows for gene regulation by many TFs, leading to a variety of expression patterns through combinatorial TF control (Wray et al., 2003). The action of TFs is governed by the DNA-binding domain (DBD), trans-activating domain (TAD) and the signal-sensing domain (SSD). These affect the DNA sequence recognised by the TF, interaction with co-regulators, and response to external signals respectively (Latchman, 1997). The different domains can either be on the same TF protein, or on separate subunits which form a heterodimeric structure when active. The domains present in a TF determine which genes it regulates and the conditions in which it is active.

The DBD binds short sequences of DNA – usually 6-12 base pairs (bp) long (Spitz and Furlong, 2012). This provides the TF with specificity in the genes that it targets. The consensus sequences recognised by the DBD are present in multiple gene regulatory elements, meaning that a single TF can control the expression of multiple different genes involved in various cellular pathways. In general, α-helices or β-strands of the DBD insert into the major groove of the DNA helix with high affinity to the specific target sequences (Wray et al., 2003). The protein commonly interacts with DNA through Van der Waals forces, and through contact with the sugar-phosphate backbone of DNA. Through this binding, TFs are able to recognise their target genes and regulate their expression.

Because TFs in the same family have similar DBDs, they often have binding affinity for similar sequences of DNA. One mechanism of increasing the binding affinity of a TF for a specific TFBS is though cooperative DNA binding (Slattery et al., 2011). Transcription co-regulators bind TADs on the TF, which can increase DNA-binding affinity and change the specificity of the TF to a specific TFBS. This is evident in the Hox TFs, which recognise similar sequences

*in vitro*, but have distinct functions *in vivo* due to cooperative DNA binding with different coregulators (Noyes et al., 2008). The TAD is also involved in recruiting components of the TIC to the promoter via protein-protein interactions (Stringer et al., 1990). The interaction of TFs with different coregulators allows for the specific control of gene expression programs in different developmental phases, tissues, and under different conditions where the availability of cofactors differs.

An additional, optional mechanism controlling the activation of TFs under different conditions is through the activity of the SSD. The SSD is a protein domain that recognises external chemical signals or ligands which, upon binding, alter the ability of the TF to control gene expression. The SSD and DBD can either be on the same subunit or on separate subunits. ERα is an example of a TF activated by a ligand. The SSD binds the hormone estrogen with high affinity, activating the transcriptional activity of the TF (Li et al., 2004). The DBD of the TF then recognises and binds estrogen response elements (EREs) on target genes, regulating their expression. Another TF that responds to external signals, albeit through a different mechanism, is HIF-1α. HIF-1α is rapidly degraded in the presence of oxygen, due to an oxygen-dependent degradation domain (Hu et al., 2013). It is constitutively expressed in the cells, but degraded under normal conditions. In hypoxic conditions, when oxygen levels drop, it is no longer degraded and can rapidly activate genes in response to a potentially harmful situation. The SSD facilitates the activation or deactivation of TFs in response to hormones and other chemical signals. This allows for a change in gene expression patterns in response to changing environmental conditions or in different tissues.

TFs activate transcription through two fundamental mechanisms. They recruit the components of the TIC to the promoter of the target gene and they alter the structure of chromatin to facilitate the process of transcription (Barberis et al., 1995, Stringer et al., 1990). The TAD of the TF interacts with various components of the TIC, including TATA-binding protein (TBP), TATA-binding protein associated factor (TAF) and other transcription initiation factors (Goodrich et al., 1993, Stringer et al., 1990, Takahashi et al., 1995). Formation of the TIC allows for the initiation of transcription. TFs also play a role in chromatin remodelling by promoting an open or closed chromatin structure for gene activation or repression respectively. TFs recruit components of chromatin remodelling complexes to their target regions through transient interactions with the TAD (McManus et al., 2011). Additionally, they can integrate into chromatin remodelling complexes as components, directly contributing to the remodelling

of chromatin (Hogan et al., 2010). With the ability to activate or repress transcription, TFs play a major role in controlling gene expression.

The TFs expressed in a particular cell control the gene expression patterns in that cell, thereby determining the phenotype of the cell under certain conditions. By modulating gene expression, TFs and other regulatory molecules, such as miRNAs, play a central role in the diversity observed in different tissues, cellular states and disease. Aberrant expression or activity of these regulators can lead to dysregulated expression of genes, which is the cause of many chronic diseases, such as cancer, neurological disorders and obesity (Lee and Young, 2013). Small sets of key TFs control the gene expression profiles of large systems of cells and tissues. The mutation or misregulation of these TFs has a deleterious effect on the regulatory networks governing homeostasis and normal cellular function. This effect is observed in the liver and the pancreas, where a set of master TFs largely controls the gene expression of these organs and when mutated, result in various forms of diabetes (Odom et al., 2004). Additionally, genome instability is an enabling characteristic for many of the hallmarks of cancer (Hanahan and Weinberg, 2011). Altered gene profiles give tumours a survival advantage by enabling rapid proliferation and the downregulation of tumour suppressing pathways. Aberrant expression of TFs is known to play a major role in tumorigenesis. The substantial effect TFs have on the regulation of the human body warrants further investigation into their molecular mechanisms and the role they play in disease.

### 1.1.7 TFs as Biomarkers

The regulation of biological processes involved in the maintenance of an organism relies heavily on the combined gene expression patterns at a cellular level. Gene expression patterns are commonly used as a marker of disease type and stage. TFs play an integral role in controlling gene expression and their deregulation is associated with many diseases in humans. As controllers of gene expression, TFs can provide insight into the controlling mechanisms of disease. Additionally, the aberrant expression of a small number of specific TFs has a large effect on gene expression patterns of a cell. This makes TFs a logical target for prognostic prediction and therapeutic interference as biomarkers of disease.

Biomarkers are biological characteristics that can be measured to diagnose disease and predict the patient prognosis and response to treatment (Mayeux, 2004). TFs are a useful type of biomarker due to their ability to alter gene expression and their involvement in human disease. As biomarkers, TFs can be used to as diagnostic and prognostic markers, and as targets for

therapeutic intervention (Ordonez, 2012, Yeh et al., 2013). Cancer in particular has been linked to the aberrant expression of numerous TFs, resulting in cellular transformation to a malignant state. A few TFs have already been identified as reliable biomarkers for cancer (Duffy et al., 2017, Littlepage et al., 2012), but with an increasing number of cancer cases lacking targeted therapy and non-invasive diagnostic methods, there is a need to identify additional biomarkers of cancer to fully elucidate the molecular mechanisms of the disease and contribute to available therapeutic and diagnostic strategies.

### 1.1.8 The Tumour Suppressor p53

The most well-known TF aberrantly expressed in cancer is the tumour suppressor p53. The protein p53 is coded by the *TP53* gene, which is directly mutated in over 50% of human cancer cases (Robles and Harris, 2010). In most other cancer cases, pathways involving p53 are altered indirectly. The primary function of p53 as a TF is activating numerous pathways involved in apoptosis, cell cycle arrest and senescence (Vogelstein et al., 2000). Activation of p53 is generally in response to cellular stress signals, such as ionising radiation, hypoxia, DNA mutations or chromosomal breakages (Bieging et al., 2014). p53 prevents replication until the damage is repaired or, if necessary, induces apoptosis to prevent the replication of mutated cells. Loss of p53 function leads to the accumulation of damaged cells, which may have increased proliferative capabilities, leading to the formation of tumours. *TP53* is frequently mutated in breast cancer, and the type of p53 deactivation (e.g. point mutation, insertion/deletion mutation or *MDM2* amplification) correlates with different molecular subtypes of breast cancer (Silwal-Pandit et al., 2014). p53 shows prognostic value in a variety of cancer, especially when combined with other biomarkers, but since its function is perturbed in such a wide variety of cancers, it lacks the specificity needed to differentiate between specific subtypes of breast cancer on its own.

### 1.1.9 The Estrogen Receptor

A biomarker commonly used to differentiate between breast cancer subtypes is the nuclear receptor ERα. As described above, this TF is regulated by the steroid hormone estrogen. It has a variety of functions in neuroendocrine, skeletal and cardiovascular processes, but is primarily a regulator of the female reproductive system (Swedenborg et al., 2009). ERα mediates proliferative pathways, linking it to estrogen-dependent tumorigenesis primarily in the breast and ovaries, and occasionally in the prostate and colon (Shang, 2007). ERα forms a homodimer when activated though estrogen binding, allowing it to bind ERE regions on DNA and activate

its target genes (Jerry et al., 2010). ERα can also be activated in the absence of estrogen via interactions with other TFs such as p53, RUNX1 and NF-κB (Jerry et al., 2010, Stender et al., 2010). ERα positive (ER+) breast cancer subtypes overexpress ERα, increasing sensitivity to the proliferation-stimulating effects of estrogen (Weigel and Dowsett, 2010). This leads to an increase in proliferation and an accumulation of cells required for tumorigenesis. In contrast, ERα-negative (ER-) breast cancer cells do not overexpress ERα and do not require estrogen for proliferation. They are therefore not responsive to endocrine therapy, such as selective estrogen receptor modulators (SERMs) (e.g. Tamoxifen), or aromatase inhibitors (e.g. Letrozole), which are usually used to treat ER+ breast cancer types. This makes ERα an important biomarker in breast cancer, as it indicates the sensitivity of the tumour to endocrine therapy. ER status also predicts the response to chemotherapy, with ER- tumours responding better to neoadjuvant chemotherapeutic treatment (Weigel and Dowsett, 2010). As mentioned previously, ERα is also a valuable prognostic marker with ER- cancer generally having higher relapse rates and poor prognoses (Rakha et al., 2006). The limitation of ERα as a therapeutic target is that it is not present in ER- cancer and therefore cannot be targeted. In addition, even resistance to endocrine treatment can arise in ER+ breast cancer patients, rendering endocrine treatments ineffective (Fan et al., 2015). Nevertheless, ERα is an important biomarker in breast cancer and combined with other biomarkers, provides insight into the subtype of breast cancer and therapeutic strategies to treat it.

### 1.1.10 Other TFs as Biomarkers

Of the hundreds of biomarkers discovered, relatively few have been approved for use by the Food and Drug Administration (FDA). This may be because of the low specificity and reliability of many potential biomarkers due to the non-standardised methods of biomarker discovery. As mentioned above, ERα, PR and HER2 are the most commonly used predictive markers for breast cancer. Recently, ERβ has also demonstrated potential as a marker of positive prognosis in breast cancer, especially in ER- patients, where ERα is absent (Tan et al., 2016). Other well-established biomarkers are the BRCA1 and BRCA2 tumour suppressors. These proteins are involved in DNA repair and are mutated in a large proportion of breast and ovarian cancers (Ferla et al., 2007). Mutation of the *BRCA* genes are present in over 20% of hereditary breast cancer and approximately 15% of hereditary ovarian cancer, making them a useful biomarker for screening, diagnosis and prognosis of breast and ovarian cancers (Szabo and King, 1995). The number of FDA approved biomarkers for specific cancers is limited, highlighting a need for the development and validation of more reliable biomarkers.

Many novel biomarkers have been proposed which may be useful in clinically assessing breast cancer cases. One potential biomarker is Ki67, a nuclear protein expressed among proliferating cells (Lopez et al., 1991). This protein could be used in combination with ER, PR and HER2 to differentiate between luminal A and B tumours (Cheang et al., 2009). It is also being investigated for use as a predictive marker for neoadjuvant chemotherapy outcome both pre- and post-treatment of breast cancer (Jones et al., 2009). Other potential biomarkers include cyclins D1 and E, which are under investigation as predictive markers for the prognosis and response to chemotherapy of breast cancer patients (Bilalović et al., 2005, Smith and Seo, 2000). The discovery and investigation of breast cancer biomarkers is of great clinical importance. Most cancers reach an advanced stage before they are detected due to the limited number of reliable biomarkers for early diagnosis. Early detection of cancers is essential, as they are easily treatable in most cases, as opposed to late stage cancers which are usually incurable. The diversity of cancer subtypes calls for the discovery of novel biomarkers for screening and early diagnosis of cancer. Specific and reliable biomarkers are required to accurately classify cancer cases and prescribe the appropriate treatment.

### 1.1.11 Detection of TFs for use in Cancer Diagnostics

An important property of a biomarker is the ability to detect and measure it in patients. The importance of TFs in cancer is evident, but without the ability to quantify their expression in clinical samples, they have little use as diagnostic or prognostic markers. The expression levels of certain TFs have been shown to be detectable by blood-based diagnostic assays. These include VDR, E2F3 and CREB1, which have been measured in blood-based assays as markers of various diseases (Belzeaux et al., 2010, Pipinikas et al., 2007, Reimer et al., 2006). This is promising, as it suggests a non-invasive method of assessing TF biomarkers in a clinical setting. A limitation of such methods is the inability to assess interactions of TFs with their target DNA sequences, which is important to determine the specific activities of the TF in the patient.

### 1.1.12 Experimental Methods to Confirm TF binding

Conventional methods of measuring protein-DNA interactions are often expensive and time-consuming, and have little value in a clinical setting. The two common methods of assessing protein-DNA interactions are the electrophoretic mobility shift assay (EMSA) and chromatin immunoprecipitation (ChIP). EMSA involves the separation of protein-DNA fragment mixtures based on size and charge via native polyacrylamide or agarose gel electrophoresis

(Fried, 1989). The TF of interest can be labelled with an antibody to alter the weight of the target complex and radioactive labelling can be used for better sensitivity. ChIP involves the crosslinking of TF-DNA interacting complexes, followed by the precipitation of the target TF-DNA complex using a TF-specific antibody conjugated to an agarose or magnetic bead (Solomon et al., 1988). Following reversal of the TF-DNA crosslinks, the DNA can be analysed using PCR, microarray or sequencing methods. Both of these methods are time consuming, expensive and require large cell or tissue samples, making them clinically inviable.

A more efficient method of identifying TF-DNA interactions is through the use of electrochemical biosensors. This was demonstrated by Gorodetsky *et al.* for the detection of TBP, a TF that binds to the TATA box in the promoter region (Gorodetsky et al., 2008). This assay uses DNA-modified microelectrodes to detect the binding of TBP by measuring interruptions in DNA-mediated charge transport to a Nile Blue probe attached to the DNA. Nanomolar concentrations of TBP were detected in the presence of micromolar concentrations of protein contaminants, demonstrating the sensitivity of this assay. The assay can be adapted to detect any DNA-binding protein and can be multiplexed on a single DNA chip for high throughput analysis of multiple biomarkers from one sample. The speed, accuracy and compact size of this technology makes it an ideal candidate for clinically testing TFs as biomarkers.

### 1.1.13 The Transcriptome: Quantification and Analysis

The rapid development of high-throughput gene technologies has facilitated the genomic, proteomic and transcriptomic profiling of tumour subtypes, enabling the elucidation of the molecular mechanisms controlling different diseases. The transcriptome is the collection of RNA transcripts in the cell at a given time. It represents actively expressed genes in the cell under specific conditions (Wang et al., 2009). Analysis of the transcriptome provides insight into gene expression patterns under different conditions and in different disease states. The transcriptome is, however, extremely complex due to the variation in transcript splicing, polymorphisms, and the biological variance in transcript abundance (Pan et al., 2008). This means that huge quantities of information are generated when a transcriptome is quantified, requiring powerful computational techniques to decipher.

Several techniques are available for analysis of the transcriptome. The most widely used of these techniques is the microarray (Casneuf et al., 2007). This method uses a chip with an array of specific DNA sequence probes attached. The sequences of these probes correspond to genes or DNA elements of interest. The RNA transcripts in a cell population are converted to

complimentary DNA (cDNA), fluorescently labelled, and passed over the probes, allowing for hybridisation based on sequence complementarity (Adomas et al., 2008). The fluorescence intensity – corresponding to number of transcripts bound to a particular probe – can be quantified as a measure of gene expression. This allows for the profiling of thousands of transcripts simultaneously in order to analyse the effects of various conditions, drug treatments or disease states on gene expression. This technology is inexpensive and allows for high-throughput analysis of gene expression in cells under different conditions with multiple replicates. There are, however, numerous limitations to microarray technology. The microarray chip uses predetermined DNA probes to measure gene expression, which makes this assay susceptible to bias based on the probes selected. Additionally, it is unable to detect novel transcript variants or polymorphisms (Wang et al., 2009). Due to the fluorescent method of quantification, the analysis of genes with low and high expression is unreliable. Careful design is also required to avoid cross-hybridisation between probes (Casneuf et al., 2007). In addition, due to specific chip design, it is difficult to compare microarray data generated from different experiments. Due to these limitations, alternative methods of transcript quantification are preferred.

Recent advances in next generation sequencing (NGS) have enabled the analysis of the transcriptome through RNA sequencing (RNA-seq). A general strategy for RNA-seq is as follows. The first step of RNA-seq is the isolation and purification of RNA from cell or tissue samples. This RNA can be separated into components if the study of one type of RNA is desired. The study of mRNA is common due to its direct link to protein synthesis. mRNA transcripts can be isolated using their 3' polyA tails as a unique feature of identification, although this ignores non-coding RNA and may introduce a 3' bias in the sequenced library (Chen et al., 2014). Additionally, ribosomal RNA (rRNA) is usually depleted from the sample, as it comprises over 90% of cellular RNA and is not useful for gene expression quantification. cDNA is synthesised through reverse transcription and random priming from isolated RNA which has been fragmented to reduce 5' bias in the cDNA library. cDNA fragments of uniform size are sequenced using NGS technology, generating millions of short sequences called reads (Wang et al., 2009). RNA reads are mapped to a reference genome or assembled *de novo* and annotated with gene feature names using bioinformatics tools such as HTSeq (Anders et al., 2015), with the number of reads mapped to a gene corresponding to calculate gene-specific transcript abundance.

RNA-seq introduces numerous advantages over microarray technology. By directly sequencing RNA transcripts instead of relying on fluorescent measurements, changes in gene expression can be measured in much greater depth and with higher sensitivity. In addition, other characteristics of the RNA molecule can be assessed, such as alternative splicing, sequence variation, transcription start site (TSS) mapping and expression of different alleles (Heap et al., 2009, Pan et al., 2008, Sultan et al., 2008, Trapnell et al., 2010). Furthermore, RNA-seq allows for the discovery of novel transcripts and genes, as it does not rely on predetermined probes for transcript detection (Pepke et al., 2009). These advantages make RNA-seq an ideal tool for the study of altered gene expression patterns in disease and the discovery of biomarkers based on this altered expression. The reason microarray technology is still widely used is because of the significantly higher costs of RNA-seq analysis. For this reason, the results of RNA-seq analysis are usually deposited in public repositories and are freely available for public use. This has led to the availability of vast quantities of expression data.

Cap analysis gene expression (CAGE) is a technique that allows for the use of RNA-seq technology at a lower cost. This technique allows for high throughput analysis of gene transcription without having to sequence the entire RNA transcript. RNA is extracted from the cells or tissue and cDNA is synthesised. The 5' end of the cDNA is then captured, using the cap-trapper method (Carninci and Hayashizaki, 1999), and cleaved by class II restriction endonucleases to produce 20 – 27 nucleotide CAGE tag, which correspond to the TSS of the mRNA. The CAGE tags are amplified, sequenced, and aligned to a reference genome (Lizio et al., 2015). The 5' nucleotide is considered the TSS and clusters of TSSs, known as CAGE peaks, are identified using decomposition peak identification (Lizio et al., 2015). This method allows for the quantification of transcription in a similar way to RNA-seq, and also allows for the identification of different TSSs used under varying conditions. The magnitude of size and complexity of gene expression data calls for the use of statistical and computational tools in order to derive biological meaning. This has brought about the development of numerous bioinformatics tools to assist in interpreting biological datasets too complex to be manually analysed by humans.

### 1.1.14 Bioinformatics for Transcriptome Analysis

The vast quantity of complex data generated through microarray and sequencing technologies challenges researchers to create methods of interpretation using computational tools. The field of bioinformatics has been established to address this challenge and derive biological meaning

from otherwise uninterpretable datasets. Bioinformatics combines computational and statistical methods to provide an understanding of complex biological information. Bioinformatics is a multidisciplinary field used to interpret biological sequences, gene and protein expression, biological and chemical structures, genotype distributions among populations, and many other complex biological systems. For the purposes of this study, it is used to interpret large gene expression datasets to identify candidate regulators in an efficient and cost-effective way before committing resources on laboratory validation.

Most bioinformatics projects rely on a combination of pre-written software tools. One of the largest repositories of bioinformatics tools is Bioconductor (Huber et al., 2015). Bioconductor is an open-source software project that provides tools for the analysis and understanding of complex biological data. It contains a wide variety of tools developed by bioinformaticists primarily in the R programming language. edgeR, a popular Bioconductor tool, is used to calculate differential gene expression using RNA-seq profiles of different cell populations (Robinson et al., 2010). The edgeR software package provides a range of statistical methodology pre-written in R to facilitate the comparison of gene expression data under differing biological conditions.

In addition to software packages for use in programming environments, there are many web-based applications used for analysis of biological data. The Database for Annotation, Visualization and Integrated Discovery (DAVID) is an online tool which provides tools for functionally annotating lists of genes (Huang et al., 2008). DAVID classifies genes into functionally related groups which are easier to interpret than flat lists of genes. It also provides various enrichment tools to identify over-represented functional classifications in a user-provided gene list. These functional classifications are based on gene ontology (GO), biological pathways, and disease associations, and a variety of other classifications. The lists of genes provided are generally associated with the cell state in study. In this study, differentially expressing (DE) genes are analysed in DAVID to identify over-represented functional annotations to elucidate cellular mechanisms altered in different breast cancer cell lines.

Functional annotations are useful for classifying the biological role of groups of genes in order to interpret large datasets. The GO project aims to represent the growing knowledge of genes in a controlled, machine-readable set of vocabularies (Ashburner et al., 2000). This project has created ontologies to represent the role of genes in thousands of biological concepts based on experimental evidence from published research articles. These ontologies fall into three

categories: biological process, molecular function and cellular component. Each GO term has a unique identifier and a definition with sources of evidence listed. Another database used for functional annotation is the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). KEGG is a database aimed at assigning functional meaning to genes. It contains networks of interacting molecules and pathway maps which are useful for assigning genes of interest to molecular pathways active in the cell. Both GO and KEGG allow for the analysis of genes not only at the gene level, but also at the level of molecular interactions and biological interactions, which is more useful for determining functions of genes derived from complex experiments. Using these databases, it is possible to computationally assign biologically relevant annotations to genes of interest, representing their role in biological systems. This is useful in the study of disease, where they facilitate the identification of perturbed biological molecular mechanisms underlying disease phenotypes.

In previous studies, computational methods have successfully aided in the discovery of cancer biomarkers in various cancers. In a study by Kaur, *et al.* in 2011, a computational method was developed to discover potential biomarkers for ovarian cancer (OC) (Kaur et al., 2011). This study found 17 TFs as potential biomarkers for OC and 3 unique TFs that regulate estrogen-controlled genes associated with OC. Another study by Kaczkowski, *et al.* used a computational method to analyse FANTOM5 CAGE and RNAseq data for genes and DNA-regulatory elements in cancer (Kaczkowski et al., 2016). This study revealed multiple potential pan-cancer biomarkers, including protein-coding genes, non-coding transcripts, which could be useful therapeutic targets. The growing number of bioinformatics publications highlights the importance of computational tools in analysing genomic and transcriptomic data.

### 1.2 Introduction to the Present Study

This study was carried out with the aim of identifying sets of genes deregulated in ER+ and ER- breast cancer and the TFs that control their expression. In this study, a computational approach was used to observe the changes in gene expression that occur in breast cancer cells compared to normal cells, and to predict TFs that control these changes. These TFs were analysed for their prognostic value in breast cancer to determine how well they can predict patient outcome. This allows for the discovery of biomarkers for breast cancer that could be used for diagnosis and screening of breast cancer, predicting the patient's response to drugs, predicting patient outcome, and as therapeutic targets to replace or enhance available breast cancer treatment options.

Bioinformatics tools were used to analyse RNA-seq data to elucidate the changes in gene transcription between breast cancer cell lines and a non-tumorigenic breast control cell line. Publically available RNA-seq datasets for the cell lines MCF7, MDA-MB-231 and MCF10, corresponding to ER+ breast cancer, ER- breast cancer and non-tumorigenic control breast cells respectively, were analysed using various computational tools. Differential expression (DE) analysis was performed to determine the gene expression profiles specific to ER+ and ER- breast cancer cell lines relative to non-tumorigenic cells. The DE genes were analysed for functional enrichment of GO and pathways to determine the biological roles of these genes. The promoters of the genes were assessed for the enrichment of EREs and TFBSs to elucidate the transcriptional mechanisms regulating the gene expression profiles of different breast cancer subtypes. The TFs that bind to enriched TFBSs were analysed for their expression levels in breast cancer patients and their prognostic value in predicting patient outcome. Finally, TF-gene networks were constructed to visualise interactions between genes and the key regulators of gene expression in ER+ and ER- breast cancer.

The TFs predicted in this study could act as potential biomarker with prognostic, diagnostic and therapeutic potential in breast cancer. These potential biomarkers could be used in addition to available clinical tools to better diagnose breast cancer at an early stage, thus increasing the probability of treatment success. They can also assist in determining the correct treatment for breast cancer patients, avoiding cases of ineffective treatment or drug resistance. Furthermore, biomarkers can be used as targets for novel treatments strategies to replace non-specific cytotoxic drugs, which often have numerous side effects. Furthermore, the methodology used in this study can be applied to data from any disease caused by altered gene expression to elucidate the mechanisms of the disease and develop better diagnostic and therapeutic strategies.

## 1.3 Aims and Objectives

### 1.3.1 Aim

To develop a computational pipeline for identifying a set of genes or TFs that can serve as biomarkers for ER+ and ER- breast cancer. These will serve as candidates for diagnostic and therapeutic markers and will contribute to the understanding of the mechanisms of breast cancer.

### 1.3.2 Objectives

1. To identify ER+ and ER- breast cancer-associated genes by conducting differential expression analysis on RNA-seq data.

2. To determine the function of DE genes by performing GO and pathway enrichment analysis.

3. To identify genes in both ER+ and ER- breast cancer potentially controlled by estrogen by predicting ERα sites in the DE gene promoters.

4. To determine which TFBSs are over-represented in the promoters of breast cancer-associated genes by performing TFBS enrichment analysis.

5. To construct a network of genes regulated by individual TFs and identify TFs that could be potential breast cancer biomarkers.

6. To validate the expression and prognostic value of identified biomarkers *in silico* using patient datasets.

# CHAPTER TWO - MATERIALS AND METHODS

## 2.1 Methodology Workflow

The following workflow diagram describes the methodology used in this study:

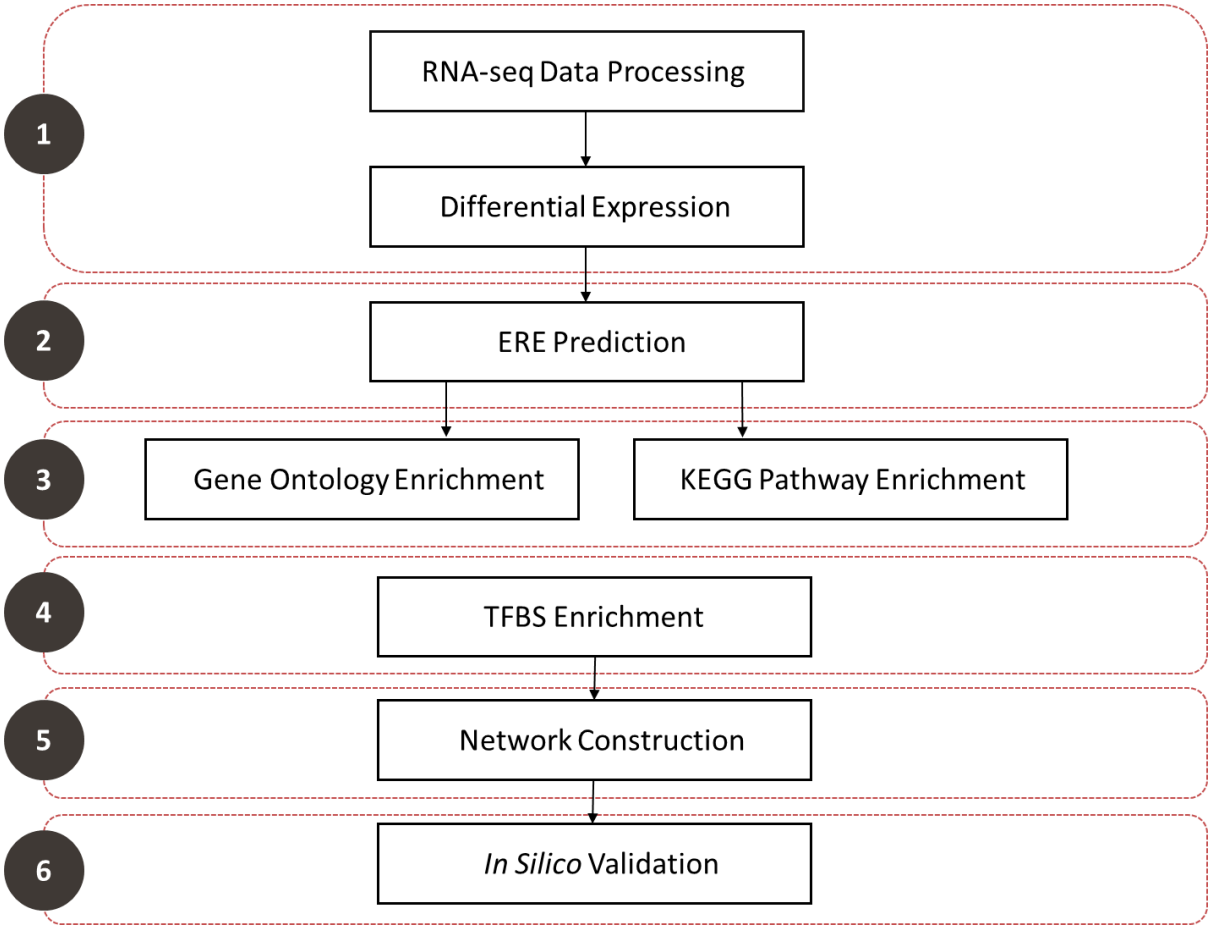



**Figure 1: Workflow of Methodology** showing the methods used to determine potential breast cancer biomarkers. The numbers correspond to the objectives in section 1.3.2 above.

## 2.2 Retrieval of Gene Expression Data

### Gene Expression Omnibus (GEO)

The GEO project was initiated by the National Centre for Biotechnology Information (NCBI) to address the growing need for a central public repository of gene expression data generated through a variety of high-throughput technologies (Edgar et al., 2002). The GEO web interface facilitates submission and retrieval of both raw and processed data, which is made available to the public for download and analysis. The data is stored in a relational database with each entry having a unique and constant accession number with different prefixes specifying attributes of the entry, facilitating extensive search and retrieval based on user-defined criteria:

- Platform (GPL) describes the experimental technology and array or sequencing platforms.
- Sample (GSM) describes the individual sample being assayed in terms of sample type and experimental conditions.
- Series (GSE) comprises all the samples belonging to a particular experiment organized as a data set.

For this study, RNA-seq data was retrieved from GEO (https://www.ncbi.nlm.nih.gov/geo/) from the series GSE75168. The data was generated using the Illumina HiSeq 1500 platform (GPL18460) with single-end 100 base sequencing for a study of breast cancer subtypes for a study on histone modifications in breast cancer (Messier et al., 2016). The raw sequencing data was downloaded in FASTQ format from the Sequence Read Archive (SRA). Table 1 shows the samples downloaded and their corresponding accession numbers.

**Table 1: RNA-seq samples retrieved from GEO**

| Sample | Accession Number | SRA Accession |
|---|---|---|
| MCF10A (replicate 1) | GSM1944515 | SRX1438068 |
| MCF10A (replicate 2) | GSM1944516 | SRX1438069 |
| MCF10A (replicate 3) | GSM1944517 | SRX1438070 |
| MCF7 (replicate 1) | GSM1944518 | SRX1438071 |
| MCF7 (replicate 2) | GSM1944519 | SRX1438072 |
| MCF7 (replicate 3) | GSM1944520 | SRX1438073 |
| MDA-MB-231 (replicate 1) | GSM1944521 | SRX1438074 |
| MDA-MB-231 (replicate 2) | GSM1944522 | SRX1438075 |
| MDA-MB-231 (replicate 3) | GSM1944523 | SRX1438076 |

## Data Pre-processing

Raw data generated by RNA sequencing requires processing before meaningful analysis can be performed. The reads from the FASTQ file need to be aligned to a reference genome and assigned to specific genes using an annotation file. This was performed using available tools implemented in the Linux Ubuntu 18.04 Bash shell environment:

- **Indexing the human genome:** The complete human genome sequence (GRCh38) was downloaded in FASTA format from the Ensembl file transfer protocol (FTP) site (https://www.ensembl.org/info/data/ftp/index.html). An index of this genome was built for read mapping using the hisat2-build program (Kim et al., 2015).
- **Read Alignment:** RNA-seq reads were aligned to the indexed human genome using the HISAT2 alignment program (Kim et al., 2015). The FASTQ files for each sample listed in Table 1 were used as an input for alignment and a sequence alignment map (SAM) file was generated.
- **Read annotation:** Aligned reads were assigned to genes and the number of reads mapped to each gene was counted. This was performed using HTSeq version 0.10.0 (Anders et al., 2015). The annotation file containing gene models corresponding to the GRCh38 genome was downloaded from the Ensembl FTP site in gene transfer format (GTF). The GTF file and SAM files were used as input for HTSeq and a matrix of genes and their corresponding read counts was generated in plaintext format for each RNA-

seq sample, representing gene expression for each sample. Separate count matrix files were combined into one file for further analysis.

### 2.3 Differential Expression Analysis

DE analysis is a statistical method used to quantify the differences in gene expression between different experimental groups by comparing the normalized gene expression levels generated through RNA quantification experiments. For analysis of high-throughput RNA-seq data, the normalized read counts are compared between groups using a statistical model. To identify the gene expression signatures of ER+ and ER- breast cancers, the pairwise differences in gene expression were compared between the MCF7, MDA-MB-231, and MCF10A cell lines. The Bioconductor package edgeR was used in the R programming language to perform DE analysis between the samples above (Robinson et al., 2010).

The edgeR program is used to measure DE between genomic features such as genes, transcripts, tags, or exons. It models read counts for each sample using a negative binomial (NB) distribution, a discrete probability distribution of counts that is accurate at low expression levels (Lun et al., 2016). This allows for the calculation of a NB dispersion parameter to estimate and account for variability between biological replicates. Empirical Bayes methods are applied to estimate gene-specific biological variation, even when few biological replicates are available. Pairwise DE between MCF7, MDA-MB-231, and MCF10A samples was calculated using edgeR in the R programming environment.

The annotated gene counts were loaded into the R programming environment as an array with genewise counts as rows and samples as columns. Three replicates each of MCF7, MDA-MB-231, and MCF10A were used as samples. To assess the consistency between replicates, a multi-dimensional scaling (MDS) plot was generated. This unsupervised clustering method calculates the log2 fold changes (logFC) between genes in each sample and computes the root-mean-square average of the largest 500 logFC values across all sample (Cox and Cox, 2000). This value is used to calculate distances between samples, which are converted to coordinates and plotted on the MDS plot. Genes with low expression across all replicates were filtered out, as DE cannot accurately be calculated at low count numbers. For this, genes with less than 0.25 counts-per-million (CPM) across at least three samples were excluded, which corresponds to approximately 10 counts for the average library size of the samples used. The data were then normalized for RNA composition. This is important because RNA-seq only measures the relative abundance of each gene in each sample. This means that when a gene is highly

expressed in one sample, that gene will make up a large proportion of the sequenced library, causing remaining genes in that sample to appear to be downregulated (Robinson and Oshlack, 2010). The libraries for each sample therefore need to be scaled to account for RNA composition bias caused by highly expressed genes in certain samples. Assuming that most genes in a sample will not be differentially expressed, scaling factors are calculated using a trimmed mean of M-values (TMM) between each pair of samples, which minimizes fold changes for the majority of genes between samples (Robinson and Oshlack, 2010). The effective library sizes are calculated by multiplying the scale factors corresponding to each sample to the original library sizes. The common, trended, and tagwise NB dispersion estimates were then calculated to account for biological and technical variation between samples. This is visualized in a scatterplot of the biological coefficient of variation (BCV) compared to average gene abundance.

DE was tested using a quasi-likelihood (QL) F-test. The QL F-test (Wedderburn, 1974) was chosen instead of the likelihood ratio test as it provides a more robust error rate control when replicate number is low compared to the likelihood ratio test. To test for DE genes, pairwise contrasts were made between MCF7, MDA-MB-231 and MCF10A cell line expression data. The null hypothesis is defined as a difference of zero in gene expression value between cell lines being compared. DE genes significantly differ from the null hypothesis according to the QL F-test with a p-value of less than 0.05. To account for multiple comparisons and control the rate of false positives detected, the Benjamini-Hochberg method is applied to the p-values to calculate the false discovery rate (FDR) (Benjamini and Hochberg, 1995). Genes with an FDR greater than 5% were not considered significant. DE genes were reported with a logFC value, which represents how different the expression of the gene is between the cell lines compared. Positive logFC values indicate that a gene is overexpressed in one cell line compared to the other, and negative values indicate that the gene is underexpressed. The glmTreat statistical test in edgeR was used to set a logFC threshold of 1, which discards DE genes with logFC that is below 1 and above -1. This method was used instead of manually selecting genes with logFC above 1 and below -1 because it makes use of an exact test to test the logFC threshold, which does not favour highly variable genes with low expression and preserves the FDR. Significantly DE genes for each comparison which conformed to the imposed criteria were written to files and used for further analysis.

## 2.4 Promoter Sequence Acquisition

The gene promoter sequences for the DE genes were obtained from the Eukaryotic Promoter Database (EPD) (https://epd.vital-it.ch/index.php). EPD is a database of experimentally validated promoter sequences for a variety of eukaryotic species. The EPDnew database was selected, as it contains combines promoters from published articles with those validated through high-throughput experiments, such as CAGE and Oligocapping (Dreos et al., 2016). The data is passed through an EPD analysis pipeline, which assesses the quality of the data and assembles it into a preliminary TSS collection. Quality evaluation of the TSS collections are performed automatically using motif enrichment tests. Randomly selected promoters are then manually evaluated before the final TSS collection is released to the public.

The significantly DE genes were separated into lists of upregulated and downregulated genes for each pairwise comparison. These lists were used to query the EPDnew database using the EPD selection tool. Only the most representative TSS was chosen if multiple were available for the query genes. The sequences starting 1000 bp upstream and ending 200 bp downstream of the TSS for each gene, using the GRCh38 genome assembly, were extracted in FASTA format for each set of DE genes. These sequences were used to represent the promoter region of each gene for further analysis. The region 1000 bp upstream and 200 bp downstream of the TSS was used, as this is generally accepted as the promoter region in which most TFs bind. One study shows that 86% of TFBSs occur within this region of the promoter (Yu et al., 2016).

## 2.5 ERE Prediction

To predict the presence of EREs in the promoter regions of DE genes, the Dragon ERE Finder version 2 program was used (Bajic et al., 2003). This is used to predict genes involved in the estrogen signalling pathway where ER proteins are activated either estrogen or other growth factors and bind directly to the DNA of target genes. This program implements the basic local alignment search tool (BLAST) algorithm to compare the DNA sequences of promoter regions to a database of EREs. The program detects ERE patterns using probabilistic models, which are applied to input sequences to predict the presence of known novel EREs. The program was run on Linux Ubuntu 18.04 64-bit, using the promoter regions downloaded from the EPD for each set of upregulated and downregulated genes in MCF7 and MDA-MB-231 cells compared to the non-tumorigenic MCF10A cell line. Even though MDA-MB-231 does not express ERα, it is still useful to analyse genes containing EREs, which would normally be regulated by ERα.

This gives insight into the alternate TFs that activate genes involved in ERα-mediated proliferation in the absence of ERα. The program was run with the default sensitivity option of 0.85 and the reverse strand option enabled so that EREs were detected on both the forward and reverse strands of the promoters. The program generated a text file containing information on each gene queried in the input file. This included the number of EREs predicted, as well as the sequence that was detected as the ERE. This file was then processed using a script in the python programming language, resulting in lists of genes containing an ERE for each DE gene list. These genes in these lists were assumed to be regulated by ER, due to the presence of EREs in their promoter regions.

### 2.6 Functional Enrichment

Functional enrichment was performed using DAVID (https://david.ncifcrf.gov/) to determine over-representation of functional characteristics in sets of genes (Huang et al., 2008). This is used to determine the shared function of lists of co-expressed genes instead of identifying the function of each gene product individually. Enrichment, also called over-representation, is used to compare functional categories that are present at a higher proportion in a list of genes than in a background set of genes. To determine the altered biological processes regulated by genes in the gene lists generated above, GO enrichment was performed using DAVID. GO is a consistent and controlled vocabulary system used to describe the biological role of genes and proteins (Ashburner et al., 2000). It is based on accumulating knowledge about biological function generated through published studies. GO is separated into three sets of ontologies: biological process, molecular function and cellular component. Biological process describes the function of the gene product in a biological system, e.g. “signal transduction”. Molecular function describes what the gene product does on a biochemical level does on a biochemical level, e.g. “receptor activity”. Cellular component describes the location in the cell in which the gene product is active, e.g. “nuclear membrane”. Each set of ontologies has a hierarchy of differing levels of specificity with specific terms branching off of broader terms, e.g. “endocytosis” is a part of “vesicle-mediated transport”, which is a part of the broad “transport” term. In this study, only the biological process ontology set was used, as the biological functions of cancer-associated genes is of interest.

Similarly, KEGG pathway enrichment was performed using DAVID to determine the biological pathways altered in lists of genes generated above. KEGG is a database of gene functions, which includes pathways representing various cellular processes (Kanehisa and

Goto, 2000). KEGG pathways are also manually drawn, allowing genes of interest to be mapped onto a pathway to assess its role in a specific biological pathway and its interaction with other gene products.

Lists of genes generated through DE analysis (genes upregulated in MCF7 and MDA-MB-231 compared to MCF10A) and ERE prediction (ERE-containing gene subset of upregulated gene lists) were uploaded to the DAVID web-interface. The whole Homo sapiens genome was used as a background for comparison. GO enrichment was performed in the biological processes category at level 4. GO levels indicate the specificity of annotations used with level one being the most general and level 5 being most specific. Level 4 is an ideal compromise between specificity and coverage, as the stringency of level 5 has low term coverage. KEGG pathway enrichment was also performed on the same gene lists. The top 10 terms were selected based on statistical significance. DAVID calculates significance using an EASE score calculation, which is a modified version of the Fisher Exact test. This is represented as a P-value which lies between 0 and 1 with statistically significant enrichment having a P-value less than 0.05. The 10 most enriched GO biological processes and KEGG pathways in the gene lists were plotted on graphs using $Log_{10}(1/P\text{-value})$. This was done to invert the P-value so that higher significance is represented by a larger value.

### 2.7 TFBS enrichment

A TFBS is a conserved DNA sequence recognised and bound by specific TFs. TFBSs can be represented as a consensus sequence – a specific sequence of DNA (Schneider, 2002). This method of representation is easy to read and interpret, but due to the variable nature of DNA, there are deviations from the consensus sequence. For this reason, a TFBS is more accurately represented as a position specific frequency matrix (PSFM), which is a 2-dimensional matrix which gives information about the frequency that each DNA base appears at a position of the DNA-binding motif, shown in Figure 2A (Stormo, 2000). This is computer-readable and can be used to predict TFBSs in DNA sequences. A PSFM can be represented as a sequence logo for human readability, with letters representing the DNA bases – adenine (A), thymine (T), guanine (G) and cytosine (C) – found at each location of the motif (Figure 2B). The size of each letter is a representation of the frequency at which the DNA base appears in the motif, with bigger letters representing higher frequency. PSFMs are used to predict the presence of TFBSs in sequences of DNA to determine TF binding and regulation of specific genes.

A

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| A | 1 | 4 | 1 | 2 | 0 | 17 | 13 |
| C | 28 | 5 | 5 | 0 | 3 | 3 | 2 |
| G | 0 | 0 | 4 | 0 | 25 | 1 | 7 |
| T | 2 | 22 | 21 | 29 | 4 | 10 | 9 |

B

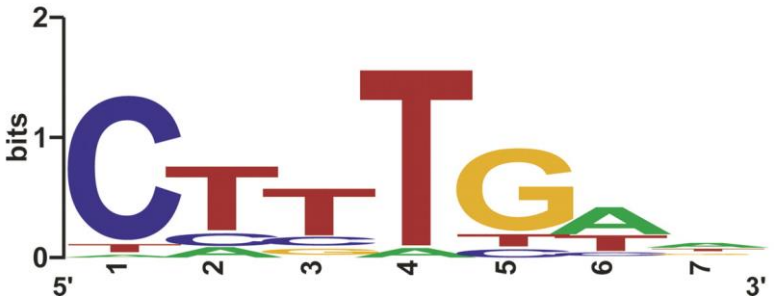



**Figure 2: Representation of a PSFM. A.** A 2-dimensional matrix of DNA base frequencies at each position of the motif. **B.** A visual representation of the PSFM as a sequence logo with the size of the DNA base letters corresponding to the frequency of their appearance at a specific position of the motif. Figure adopted from https://sites.google.com/site/iiserbioinformatics/tutorials (accessed 2018/12/06).

TFBS enrichment was performed using the oPOSSUM-3 web-based software system available at http://opossum.cisreg.ca (Kwon et al., 2012). oPOSSUM-3 is a tool used for the identification of enrichment or over-representation of TFBSs in nucleotide sequence-based data. In this study, the Single Site Analysis (SSA) procedure was used. This procedure identifies and counts TFBSs in both user-defined sequences (referred to as the foreground set) and a set of background sequences (all the genes in the oPOSSUM database). Based on the number of counts, Z-scores and Fisher scores are calculated for each TFBS hit, comparing the foreground counts to the background counts. Evolutionarily conserved genomic regions are defined using the PhastCons program, which calculates the probability that each nucleotide is part of a conserved region (Siepel et al., 2005). Only motifs overlapping with conserved regions determined by PhastCons are searched for TFBSs.

To calculate motif over-representation in foreground sequences vs background sequences, Z-score and Fisher score are used. The Z-score employs a binomial distribution model to calculate whether occurrence rate of each individual TFBS in the foreground set significantly differs from that of the background set. The Z-score, therefore, tests for the significance at the level of individual TFBS hits and is expressed as a magnitude of standard deviation. In contrast, the

Fisher score tests for significance at the gene or sequence level, only taking into account the presence or absence of a TFBS and not the number of individual TFBSs present in the sequence. It compares the proportion of genes or sequences in the foreground set containing a specific TFBS to the proportion of background genes or sequences with the same TFBS. In this way, the probability of non-random association between a specific TFBS and the user-defined gene set is calculated. The Fisher score is based on a hypergeometric distribution and is expressed as the negative natural logarithm of the calculated probability of significance (Kwon *et al.,* 2012). Both Z-score and Fisher score are subjected to the Benjamini-Hochberg method to correct for multiple testing and an FDR value is reported (Benjamini and Hochberg, 1995).

In this study, TFBS profiles from the JASPAR database (http://jaspar.genereg.net/) were used with oPOSSUM-3. JASPAR is a database of curated TFBS profiles for TFs in multiple species (Khan et al., 2018). JASPAR stores TFBS profiles as positional frequency matrices (PFMs), which can be converted to position weight matrices (PWMs). PWMs are probabilistic models commonly used to predict TFBSs in sequences of DNA. The JASPAR CORE collection of TFBS profiles was used in this study. This is a manually curated collection of DNA binding models for each TF. These models are used with oPOSSUM-3 to predict the over-representation of TFBSs in DE gene profiles determined for MCF7 and MDA-MB-231 cell lines.

Sets of genes significantly upregulated in the comparisons of RNA-seq expression of MCF7 and MDA-MB-231 with the normal control cell line, MCF10A, were uploaded to the oPOSSUM-3 web interface. All 24 752 genes in the oPOSSUM database were used as a background set of genes. The JASPAR CORE vertebrate TFBS profiles were used as models for TFBS enrichment. A conservative cut-off score of 0.4 was selected, corresponding to the PhastCons probability score described above. The matrix score threshold was set to 85%, corresponding to the minimum PWM score considered to be significant when scanning each position of the query sequence. The region 1000 bp upstream and 200 bp downstream of the TSS of the target gene was used. TFBSs significantly over-represented in the query gene sets were reported with corresponding Z-scores, Fisher scores and the query genes containing each enriched TFBS. Predicted TFBS motifs were sorted based on Fisher score, as this analysis focussed more on individual genes than individual TFBS hits. The recommended Fisher score cut-off of 7 was used, which is based on empirical studies (Ho Sui et al., 2005). A further cut-off of 7 was used for Z-score, which is less stringent than the proposed cut-off of 10, but adds enough stringency to exclude insignificant results (Ho Sui et al., 2005). The TFs corresponding

to the selected enriched TFBSs were used as prospective biomarkers and subjected to further validation steps.

### 2.8 Oncomine Validation

Oncomine was used to validate the expression of the TFs selected above in breast cancer patients. Oncomine (https://www.oncomine.org/) is an online platform that unifies gene expression datasets from different cancer publications (Rhodes et al., 2004). Beginning as a microarray database, Oncomine now also includes sequencing data of both DNA and RNA. Oncomine provides gene expression data of many different tissue types, including both normal and tumours, and many different types of cancers. DE analysis is performed comparing cancer vs normal, cancer vs cancer and normal vs normal tissues. The data can be explored using filters for specific genes, drug interactions and clinical annotations. Data can also be visualised in various different ways.

Oncomine was used to compare the expression of selected TFs in breast cancer vs normal breast tissue. Filters for "Breast Cancer" and "Cancer vs. Normal Analysis" were selected to limit the data to the relevant criteria. The gene names of the selected TFs were individually entered as a dataset filter. Datasets containing breast cancer vs normal DE analysis including the specified gene were displayed. A dataset that included DE information of the gene of interest was selected and if the P-value indicated significance ($p < 0.05$), the fold change (FC) value comparing gene expression in the breast cancer samples vs normal samples was used. If the FC value was greater than 1, indicating over-expression of the gene in cancer cells, the TF encoded by the gene was used for further analysis for prognostic properties.

### 2.9 Survival Analysis

To assess the prognostic value of selected TFs the online tool PROGgeneV2 was used (http://watson.compbio.iupui.edu/chirayu/proggene/database/index.php) (Goswami and Nakshatri, 2014). This tool compiles datasets from public repositories such as GEO, The Cancer Genome Atlas (TCGA) and EBI Array Express. Patient gene expression data assessing survival and metastasis for various cancer types is available to assess the prognostic value of potential biomarkers. The tool generates Kaplan Meier (KM) prognostic plots comparing the effect of high and low expression of the selected gene on patient survival. The KM estimator is a non-parametric test used to assess patient survival over time (Kaplan and Meier, 1958). The KM plot is a line of probability that an event of interest, such as patient relapse, metastasis

or death, takes place. There are generally multiple lines representing different populations of patients separated into categories, such as cancer subtypes or expression of a gene or gene signature.

PROGgeneV2 was used to assess the prognostic value of Oncomine-validated TFs in patient survival and metastasis. The 9 prospective biomarkers validated for MCF7 cells and the 13 prospective biomarkers validated for MDA-MB-231 cells were entered as input genes in the PROGgeneV2 online interface. Breast cancer was chosen from a list of cancer types. The genes were measured for prognostic value in both "death" (overall survival) and "metastasis" (metastasis-free survival). The data was bifurcated by median expression value, meaning the samples were divided into groups of high and low expression by the median of expression of the target gene. The KM plots were corrected for the presence of ER to ensure that only the prognostic value of the TF of interest was being evaluated. A representative dataset containing the gene of interest was chosen based on the significance of the KM plot. The NKI and TCGA datasets were used, as together they contained survival information on all of the genes of interest. The KM plot was generated and analysed for prognostic value.

### 2.10 Network Construction

To visualise the predicted interactions between TFs selected as potential biomarkers and the genes they are predicted to regulate, Cytoscape (https://cytoscape.org/) was used. Cytoscape is an application used to visualise and analyse interaction networks generated using high-throughput expression data (Shannon et al., 2003). It has the ability to generate networks of a variety of interactions, such as protein-protein, protein-DNA and genetic interactions. Using different plug-ins, it can also be linked to databases of molecular interactions to integrate information from different sources.

Cytoscape was used to construct networks using TFs selected as potential biomarkers and the genes they regulate according to TFBS enrichment analysis. A network file was created in plain-text form, assigning the TFs as source nodes and the genes they regulate as target nodes. For each cell line, the TFs and the genes they regulate were combined to visualise overlapping regulatory function of the TFs. A circular layout was applied to the networks to clearly view the TF-gene interactions. Networks were exported as images.

### 2.11 Predictive Value of Prospective Biomarkers

After assessing the expression patterns and prognostic value of the potential biomarkers and reviewing the literature surrounding these TFs, TAL1, FOXD1 and INSM1 were selected as prospective novel biomarkers with potential to predict a negative outcome in breast cancer patients. Although other prospective biomarkers are discussed further, they were not included in this analysis because they either lack novelty or lack the ability to predict negative outcome. The predictive values of the three selected TFs in breast cancer were measured using statistical measurements based on their expression levels in breast cancer patients on the Oncomine platform using the datasets described in tables 7 and 8.

The sensitivity, specificity and precision of the three selected TFs in predicting breast cancer were calculated (Parikh et al., 2008). These calculations use the number of true positives (TP), which is the number of breast cancer patients expressing the TF, false positives (FP), which is the number of patients without breast cancer that express the TF, true negatives (TN), which is the number of patients without breast cancer that do not express the TF, and false negatives (FN), which is the number of breast cancer patients that do not express the TF. Sensitivity calculates the likelihood that the TF will be expressed if the disease is present. It is calculated as follows:

$$\frac{TP}{TP + FN} \times 100$$

Specificity calculates the likelihood that the TF will not be expressed if the disease is absent and is calculated as follows:

$$\frac{TN}{TN + FP} \times 100$$

Precision is the positive predictive value of the TF. It determines whether a patient testing positive for TF expression actually has breast cancer and is not a false positive. Precision is calculated as follows:

$$\frac{TP}{TP + FP} \times 100$$

These measurements were calculated using expression data from the Oncomine platform for each selected TF to determine the predictive potential for breast cancer.

# CHAPTER THREE – RESULTS

## 3.1 Differential Expression

DE analysis was performed to compare gene expression at the transcriptional level between different cell lines. The mRNA expression levels of MCF7 and MDA-MB-231 cell lines – representing ER+ and ER- breast cancer respectively – were both compared to the non-tumorigenic control breast cell line MCF10A. RNA-seq data obtained from the GEO series GSE75168 was analysed for DE using the edgeR Bioconductor package implemented in the R programming language. A total of 20 572 genes were annotated for each cell line dataset and used for DE analysis. A MDS plot was generated to assess the consistency of gene expression between replicates and to cluster samples with similar gene expression levels. In Figure 3, the MDS plot shows the clustering of RNA-seq samples used for the DE analysis. The three replicates of MCF7, MDA-MB-231 and MCF10A, represented in red, blue and green respectively, cluster very closely, confirming that the expression levels are similar in each sequencing replicate. Additionally, cell line clusters were separate from each other, showing that the mRNA levels differ between the cell lines compared.

Dispersion parameters were estimated for the RNA-seq samples to account for variability between replicates. The average common dispersion calculated was 0.087, which is low. This is expected when using genetically identical organisms or cell lines (Yoon and Nam, 2017). The distribution of dispersion is represented in Figure S1 in the appendix, showing BCV (biological coefficient of variation) against average transcript abundance. The blue trend line shows that dispersion is higher for genes with low expression and decreases to a constant value as the abundance reaches a higher value of approximately 5 CPM. This is also expected, as transcripts with low counts are highly variable and therefore cannot be used to calculate DE.

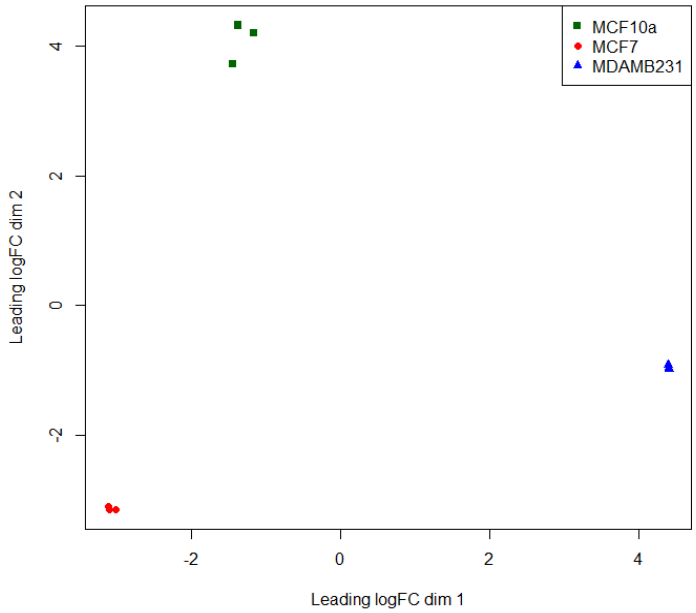



**Figure 3: MDS plot representing the differences in mRNA levels between RNA-seq samples.** The axes represent the average difference in leading logFC values between samples. Replicates of MCF7, MDA-MB-231 and MCF10A are shown in red, blue and green respectively.

Pairwise comparisons were performed, comparing the gene expression between MCF7 and MCF10A, MDA-MB-231 and MCF10A, and MCF7 and MDA-MB-231. Significantly DE genes for each comparison were calculated with a p-value below 0.05 and an FDR below 5%. An absolute logFC threshold of 1 was applied using the glmTreat algorithm. Mean-difference (MD) plots (figure 4) show significantly DE genes for each comparison with upregulated genes shown in red, downregulated genes shown in blue and non-DE genes shown in black relative to MCF7 (A), MDA-MB-231 (B) and MCF7 (C). Non-DE genes did not meet the criteria for p-value, FDR or logFC threshold. The results of the DE analysis showed 1894 genes upregulated and 2502 genes downregulated in MDA-MB-231 compared to MCF10A, 2484 genes upregulated and 2484 genes downregulated in MCF7 compared to MDA-MB-231, and 1924 genes upregulated and 2543 genes downregulated in MCF7 compared to MCF10A. This is shown in table 2. The expression of these genes is altered in breast cancer cell lines and is associated with changes in phenotype observed in tumours compared to normal tissue.

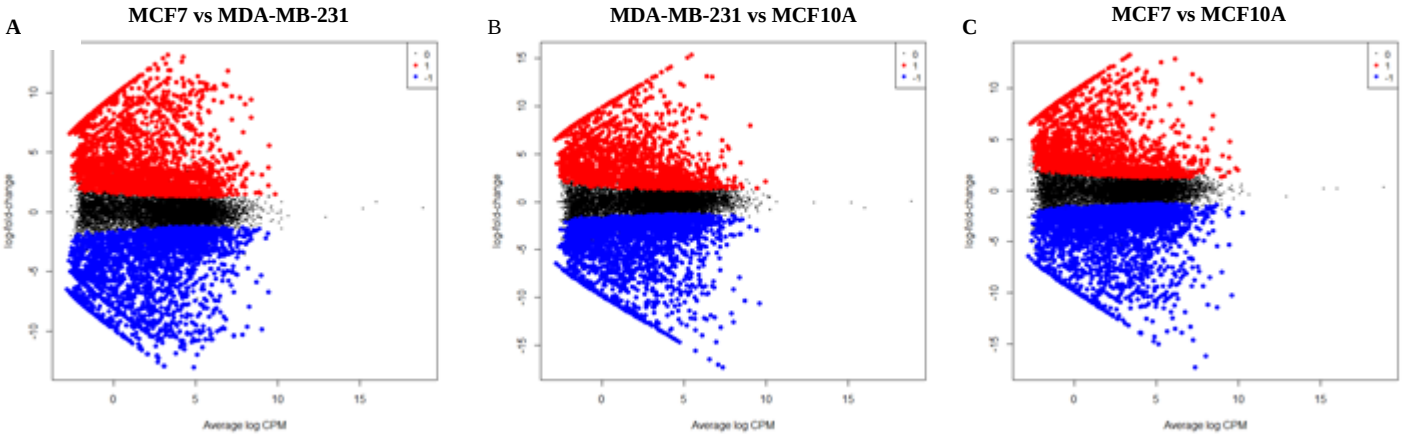



**Figure 4: MD plot showing logFC compared to average abundance of each gene.** Significantly up- and downregulated genes are represented in red and blue respectively. Comparisons of gene expression between MCF7 and MDA-MB-231 (A), MDA-MB-231 and MCF10A (B), and MCF7 and MCF10A(C).

**Table 2: Number of up- and downregulated DE genes.** Significantly up- and downregulated genes with a logFC greater than (upregulated) or less than (downregulated) 1, relative to the first cell line in each DE comparison. The number of input genes was 20 572 and three replicates of each cell line were used.

| Cell Lines Compared | Number of Upregulated Genes | Number of Downregulated Genes |
|---|---|---|
| MCF7 vs MCF10A | 1924 | 2543 |
| MDA-MB-231 vs MCF10A | 1894 | 2502 |
| MCF7 vs MDA-MB-231 | 2484 | 2484 |

The relative expression levels of the 50 most significantly DE genes for each comparison are displayed in heat maps in figure 5. The dendrogram on the x-axis shows hierarchical clustering between samples. As expected, the three replicates of each cell line clustered together very closely, showing consistency in their gene expression. Hierarchical clustering of individual genes is represented by the dendrogram on the y-axis, which groups genes with correlated expression patterns. This shows genes with specific expression patterns, either upregulated or downregulated in one cell line and the opposite expression pattern in other cell lines. Clusters of gene expression seen in figure 5 could represent genes that are expressed under similar conditions, suggesting control by the same set of regulatory molecules. Lists of DE genes were

separated into upregulated and downregulated lists for each pairwise comparison and written to separate files for further analysis.

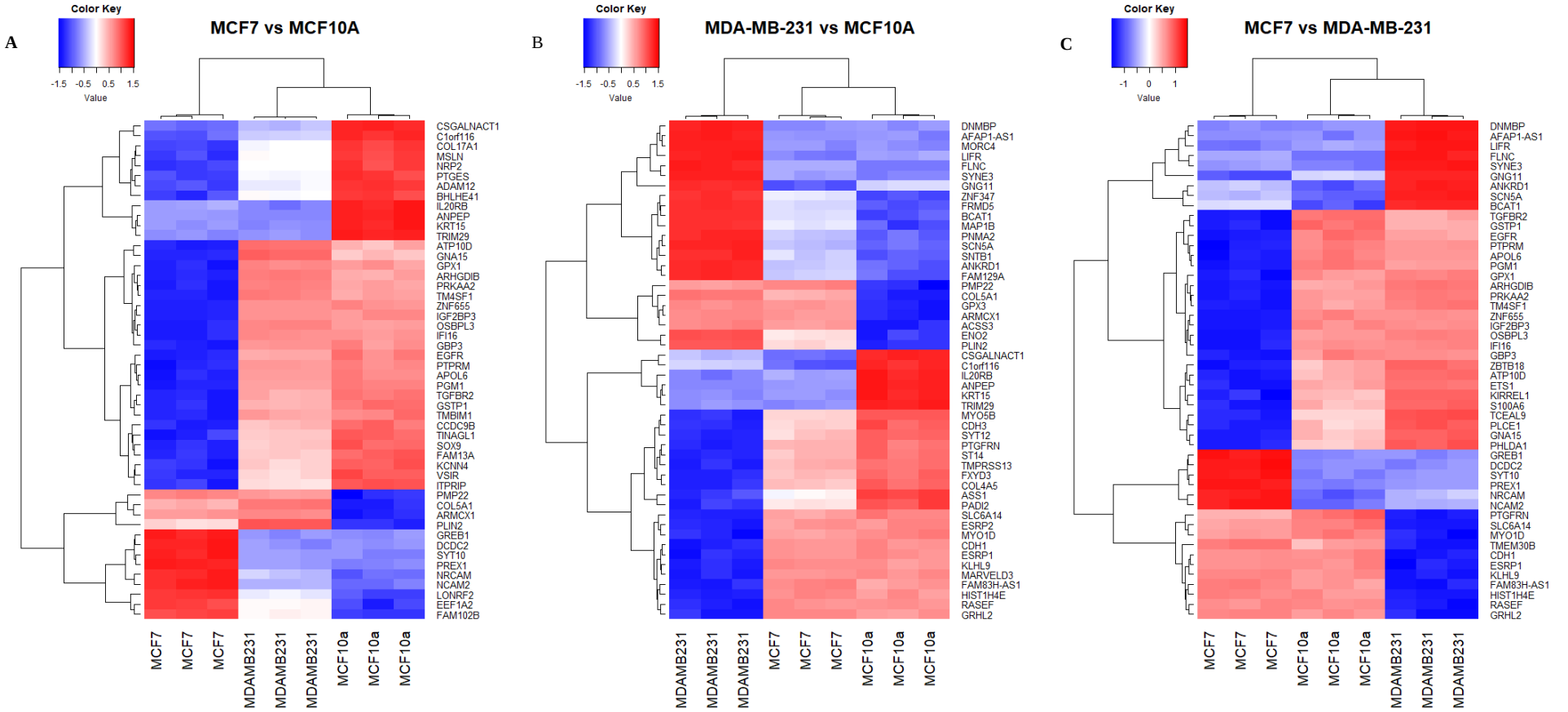



**Figure 5: Heat maps of the expression levels of the most DE genes across all samples.** Expression levels of the top 50 most DE genes between MCF7 and MCF10A (A), MDA-MB-231 and MCF10A (B), and MCF7 and MDA-MB-231(C). Expression levels across all cell lines are shown for the DE genes selected. High expression levels are depicted in red and a low expression levels are depicted in blue. Sample replicates are displayed on the x-axis and gene symbols are displayed on the y-axis. Dendrograms show hierarchical clustering between samples on the x-axis and genes on the y-axis.

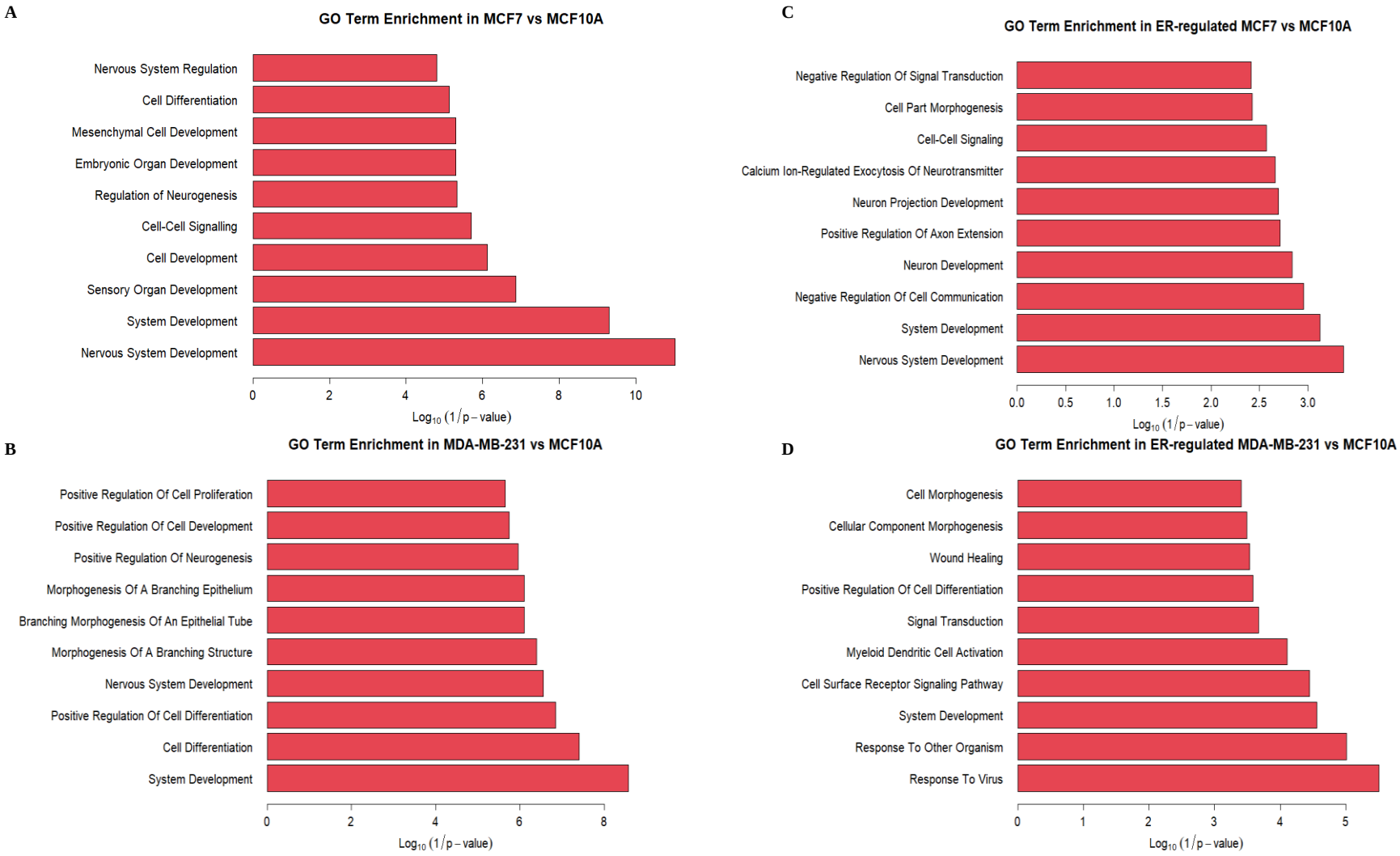



**Figure 6: Bar graphs representing GO enrichment in upregulated gene sets.** Top 10 most significantly enriched GO terms in DE genes upregulated in MCF7 vs MCF10A (A), MDA-MB-231 vs MCF10A (B) and predicted ER-regulated genes upregulated in MCF7 vs MCF10A (C), MDA-MB-231 vs MCF10A (D). The size of the bars indicates $\log_{10}(1/\text{p-value})$, representing higher significance with increasing size.

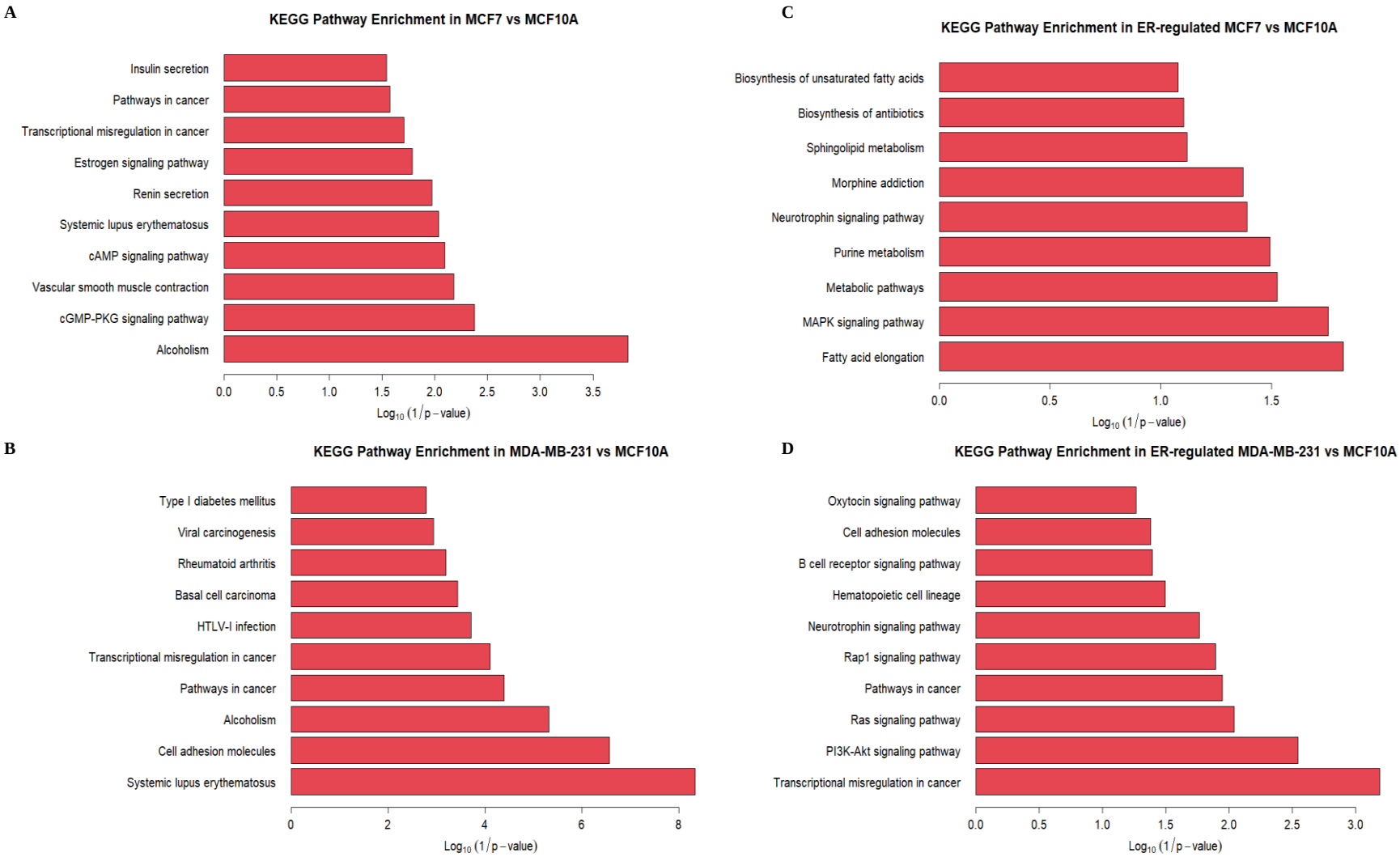



**Figure 7: Bar graphs representing KEGG pathway in upregulated gene sets.** Top 10 most significantly enriched KEGG pathways in DE genes upregulated in MCF7 vs MCF10A (A), MDA-MB-231 vs MCF10A (B) and predicted ER-regulated genes upregulated in MCF7 vs MCF10A (C), MDA-MB-231 vs MCF10A (D). The size of the bars indicates $\log_{10}(1/\text{p-value})$, representing higher significance with increasing size.

### 3.2 Functional Enrichment

Functional enrichment was performed on gene lists generated through DE analysis and ERE prediction to determine GO biological processes and KEGG pathways that are statistically over-represented compared to a background set of genes (all the genes in the human genome). The top 10 most enriched GO terms for each gene list are represented as graphs in figure 6 and the top 10 enriched KEGG terms are represented in figure 7, based on statistical significance (lower P-value). In both figures, the name of the enriched term is shown on the y-axis and the $Log_{10}$(1/p-value) is shown on the x-axis, representing the statistical significance of enrichment. A significant P-value of 0.05 corresponds to a $Log_{10}$(1/p-value) of 1.301 and the higher the $Log_{10}$(1/p-value) is, the more significant the enrichment is. All enriched terms shown in figures 6 and 7 are statistically significant.

Figure 6A represents GO biological process enrichment in genes upregulated in MCF7 compared to MCF10A. The terms "mesenchymal cell development" and "embryonic organ development" are over-represented within upregulated MCF7-associated genes, both of which are associated with a stem-like state. Notably, many biological processes involved in with the nervous system are also over-represented, such as "nervous system development", "nervous system regulation" and "regulation of neurogenesis", suggesting a link between the nervous system and breast cancer. Figure 6B represents biological processes enriched in genes upregulated in MDA-MB-231 compared to MCF10A. Terms related to cancer development, such as "positive regulation of cell proliferation" and "positive regulation of cell development" are enriched. Interestingly, there are also enriched biological processes involved in growth of ductal cells of the mammary gland, namely "morphogenesis of a branching epithelium", "branching morphogenesis of an epithelial tube" and "morphogenesis of a branching structure", which could be involved in increased breast tissue growth and proliferation. Similar to MCF7-associated genes, nervous system development is also an observed function of genes associated with MDA-MB-231. Figures 6C and 6D represent the GO biological processes in ERE-containing genes upregulated in MDA-MB-231 compared to MCF10A and MCF7 compared to MCF10A respectively. These show similar enrichment as the DE gene lists with an emphasis on nervous system development in ERE-containing MCF7-associated genes. Notably, ERE-containing MDA-MB-231-associated genes are involved in the "wound healing" process, which is linked to growth, inflammation and a stem-like state observed in aggressive cancers.

Figures 7A, 7B, 7C and 7D represent the same gene sets as Figure 6A, 6B, 6C and 6D respectively, but Figure 7 represents KEGG pathway enrichment. In Figure 7A, representing upregulated MCF7-associated genes, “pathways in cancer” and “transcriptional misregulation in cancer” KEGG pathways are enriched, which is expected of a cancer cell line. The “Estrogen signalling pathway” is also enriched, as this represents ER+ breast cancer. There is also an enrichment of genes involved in the “cAMP signalling pathway” and the “cGMP-PKG signalling pathway”, which are both closely related to cancer. Increased “insulin secretion” and “renin secretion” is also observed, both of which have an oncogenic effect on cancer cells. In upregulated MDA-MB-231-associated genes (Figure 7B), “pathways in cancer” and “transcriptional misregulation in cancer” pathways are enriched, as well as the “PI3K-Akt signalling pathway” and the “Ras signalling pathway”, which are commonly associated with tumorigenesis. The enriched “Rap1 signalling pathway” is linked to MAPK and PI3K-Akt signalling, linking it to cancer-associated processes. Interestingly, pathways associated with leukaemia, such as “hematopoietic cell lineage” and “B cell receptor signalling pathway” are also enriched in MDA-MB-231-associated genes. Figures 7C and 7D, representing ERE-containing genes upregulated in MCF7 and MDA-MB-231 respectively, show similar cancer-related pathway enrichment. In ER-regulated genes upregulated in MCF7 (Figure 7C) exhibit increased “biosynthesis of unsaturated fatty acids” and “sphingolipid metabolism”. The GO terms and KEGG pathways enriched in each gene set represent biological mechanisms that may be controlled by the altered gene expression patterns observed in different cancer gene sets compared to non-tumorigenic genes. The relevance of key processes and pathways in breast cancer will be elaborated further in the discussion section.

### 3.3 TFBS Enrichment

Using the oPOSSUM tool, JASPAR TFBS motif enrichment was identified in the promoter regions of upregulated DE genes. For this project, only upregulated genes were considered, as we were interested in the regulation of oncogenes. Downregulated genes could be considered at a later stage to study the regulation of tumour suppressing genes. TFs corresponding to these TFBSs are displayed in tables 3-6. Within the 1924 genes upregulated in MCF7 vs MCF10A (table 3), 13 enriched TFs met the Fisher score and Z-score criteria and were selected as potential biomarkers of ER+ BC. In the 1894 genes upregulated in MDA-MB-231 vs MCF10A (table 4), 28 TFs were selected as potential biomarkers. These TFs are potential controllers of the expression of genes upregulated in each comparison. Among the upregulated genes, subsets

of genes with predicted EREs were analysed for TFBS enrichment. In the 440 upregulated MCF7 genes containing predicted ERE motifs (table 5), 8 TFBSs were enriched and 4 TFBSs were identified in the 497 ERE-containing upregulated genes in MDA-MB-231 (table 6). It should be noted that Myf was identified as an enriched motif, which represents members of the MyoD and Myog TF families, therefore these TFs were used for further analysis in place of Myf. In Tables 3-6, the official gene symbols of the TFs corresponding to enriched TFBSs were listed, along with target gene hits, which is the number of genes in the gene set containing TFBSs for the specified TF, Z-score, Fisher score, and the proportion of gene hits, which is the percentage of total genes in the gene set used that contain TFBS hits for the specified TF. The TFs listed in the below tables were used for further analysis of their use as prognostic markers in BC.

**Table 3: TFBSs Enriched in DE genes upregulated in MCF7 vs MCF10A.** Significantly over-represented TFBSs in the promoter regions of the set of genes upregulated in MCF7 cells compared to MCF10A cells. TF symbol represents the gene symbol of the TF corresponding to the TFBS. JASPAR ID shows the ID of the TF in the JASPAR database. Target gene hits represents the number of genes in the gene set predicted to have the TFBS in the promoter region. Z-score and Fisher score show statistical significance of the enrichment of the TFBS. Proportion of gene hits is the percentage of the total genes in the DE gene set that contain the TFBS.

| TF Symbol | JASPAR ID | Target Gene Hits | Z-Score | Fisher Score | Proportion of Gene Hits (%) |
|---|---|---|---|---|---|
| SP1 | MA0079.2 | 700 | 13.015 | 78.047 | 36.38 |
| MZF1 | MA0056.1 | 789 | 11.41 | 55.167 | 41.01 |
| HIF1A | MA0259.1 | 525 | 8.452 | 52.089 | 27.29 |
| EBF1 | MA0154.1 | 442 | 11.163 | 40.818 | 22.97 |
| E2F1 | MA0024.1 | 338 | 8.874 | 37.601 | 17.57 |
| Mycn | MA0104.2 | 357 | 11.488 | 34.728 | 18.56 |
| Myf | MA0055.1 | 352 | 7.39 | 31.932 | 18.3 |
| Myc | MA0147.1 | 329 | 10.07 | 30.218 | 17.1 |
| NHLH1 | MA0048.1 | 203 | 10.465 | 24.778 | 10.55 |
| INSM1 | MA0155.1 | 286 | 7.273 | 24.358 | 14.86 |
| Stat3 | MA0144.1 | 303 | 7.07 | 23.205 | 15.75 |
| NFYA | MA0060.1 | 249 | 7.149 | 20.207 | 12.94 |
| NFKB1 | MA0105.1 | 139 | 8.654 | 16.473 | 7.22 |

**Table 4: TFBSs Enriched in DE genes upregulated in MDA-MB-231 vs MCF10A.** Significantly over-represented TFBSs in the promoter regions of the set of genes upregulated in MDA-MB-231 cells compared to MCF10A cells. TF symbol represents the gene symbol of the TF corresponding to the TFBS. JASPAR ID shows the ID of the TF in the JASPAR database. Target gene hits represents the number of genes in the gene set predicted to have the TFBS in the promoter region. Z-score and Fisher score show statistical significance of the enrichment of the TFBS. Proportion of gene hits is the percentage of the total genes in the DE gene set that contain the TFBS.

| TF Symbol | JASPAR ID | Target Gene Hits | Z-Score | Fisher Score | Proportion of Gene Hits (%) |
|---|---|---|---|---|---|
| Myf | MA0055.1 | 394 | 9.623 | 30.922 | 20.8 |
| RUNX1 | MA0002.2 | 515 | 9.435 | 26.764 | 27.19 |
| FOXA1 | MA0148.1 | 451 | 9.288 | 26.34 | 23.81 |
| Sox17 | MA0078.1 | 510 | 9.05 | 24.809 | 26.93 |
| Foxd3 | MA0041.1 | 355 | 10.745 | 24.658 | 18.74 |
| MEF2A | MA0052.1 | 202 | 13.131 | 24.439 | 10.67 |
| TBP | MA0108.2 | 353 | 8.468 | 23.929 | 18.64 |
| CEBPA | MA0102.2 | 433 | 10.532 | 23.182 | 22.86 |
| Gfi | MA0038.1 | 504 | 11.234 | 22.258 | 26.61 |
| ARID3A | MA0151.1 | 576 | 14.967 | 22 | 30.41 |
| HOXA5 | MA0158.1 | 708 | 9.996 | 21.504 | 37.38 |
| SRY | MA0084.1 | 509 | 11.539 | 21.254 | 26.87 |
| Nkx2-5 | MA0063.1 | 635 | 13.739 | 19.673 | 33.53 |
| Foxa2 | MA0047.2 | 353 | 8.581 | 19.136 | 18.64 |
| NKX3-1 | MA0124.1 | 394 | 7.142 | 19.133 | 20.8 |
| Prrx2 | MA0075.1 | 520 | 7.838 | 19.007 | 27.46 |
| Pdx1 | MA0132.1 | 545 | 9.204 | 18.723 | 28.78 |
| FOXO3 | MA0157.1 | 444 | 10.393 | 18.231 | 23.44 |
| Gata1 | MA0035.2 | 492 | 10.412 | 18.157 | 25.98 |
| FOXI1 | MA0042.1 | 353 | 11.14 | 17.868 | 18.64 |
| Foxq1 | MA0040.1 | 211 | 13.097 | 17.525 | 11.14 |
| TAL1::TCF3 | MA0091.1 | 192 | 7.393 | 16.438 | 10.14 |
| FOXD1 | MA0031.1 | 418 | 8.466 | 15.933 | 22.07 |
| Nobox | MA0125.1 | 431 | 9.311 | 13.777 | 22.76 |

| | | | | | |
|---|---|---|---|---|---|
| NFIL3 | MA0025.1 | 173 | 9.746 | 13.707 | 9.13 |
| Lhx3 | MA0135.1 | 137 | 7.46 | 12.286 | 7.23 |
| IRF1 | MA0050.1 | 159 | 7.493 | 11.356 | 8.39 |
| SRF | MA0083.1 | 31 | 9.525 | 9.404 | 1.64 |

**Table 5: TFBSs Enriched in Predicted ER-regulated genes upregulated in MCF7 vs. MCF10A.** Significantly over-represented TFBSs in the promoter regions of the set of genes upregulated in MCF7 cells compared to MCF10A cells that are predicted to contain EREs. TF symbol represents the gene symbol of the TF corresponding to the TFBS. JASPAR ID shows the ID of the TF in the JASPAR database. Target gene hits represents the number of genes in the gene set predicted to have the TFBS in the promoter region. Z-score and Fisher score show statistical significance of the enrichment of the TFBS. Proportion of gene hits is the percentage of the total genes in the DE gene set that contain the TFBS.

| TF Symbol | JASPAR ID | Target Gene Hits | Z-Score | Fisher Score | Proportion of Gene Hits (%) |
|---|---|---|---|---|---|
| MZF1 | MA0057.1 | 187 | 10.06 | 17.048 | 42.5 |
| Myf | MA0055.1 | 117 | 7.198 | 15.577 | 26.59 |
| MZF1_1-4 | MA0056.1 | 237 | 8.281 | 14.111 | 53.86 |
| EBF1 | MA0154.1 | 137 | 10.03 | 13.996 | 31.14 |
| NF-kappaB | MA0061.1 | 96 | 8.898 | 12.327 | 21.82 |
| INSM1 | MA0155.1 | 93 | 10.247 | 10.998 | 21.14 |
| RELA | MA0107.1 | 71 | 7.704 | 9.667 | 16.14 |
| PLAG1 | MA0163.1 | 29 | 10.789 | 9.55 | 6.59 |

**Table 6: TFBSs Enriched in Predicted ER-regulated genes upregulated in MDA-MB-231 vs. MCF10A** Significantly over-represented TFBSs in the promoter regions of the set of genes upregulated in MDA-MB-231 cells compared to MCF10A cells that are predicted to contain EREs. TF symbol represents the gene symbol of the TF corresponding to the TFBS. JASPAR ID shows the ID of the TF in the JASPAR database. Target gene hits represents the number of genes in the gene set predicted to have the TFBS in the promoter region. Z-score and Fisher score show statistical significance of the enrichment of the TFBS. Proportion of gene hits is the percentage of the total genes in the DE gene set that contain the TFBS.

| TF Symbol | JASPAR ID | Target Gene Hits | Z-Score | Fisher Score | Proportion of Gene Hits (%) |
|---|---|---|---|---|---|
| RUNX1 | MA0002.2 | 169 | 10.864 | 15.631 | 34 |
| MAX | MA0058.1 | 95 | 9.497 | 11.529 | 19.11 |
| Gata1 | MA0035.2 | 153 | 8.003 | 8.227 | 30.78 |
| IRF1 | MA0050.1 | 53 | 7.462 | 6.88 | 10.66 |

### 3.4 Oncomine Validation

Oncomine was used to validate the expression of selected TFs in BC patient datasets. TFs with significantly altered expression levels in BC samples compared to normal samples are shown in tables 7 and 8 for biomarker TFs predicted in MCF7 and MDA-MB-231 cell lines respectively. Out of 13 TFs controlling upregulated genes in MCF7 and 8 TFs controlling the ER-regulated subset of these genes, 9 TFs met the criteria of having an FC > 1 in BC samples vs normal samples with a P-value < 0.05 (table 7). Out of 28 TFs controlling upregulated genes in MDA-MB-231 and 8 TFs in the ER-regulated subset, 13 TFs met the aforementioned criteria (table 8). Tables 7 and 8 list the official gene symbols of the qualifying TFs, the fold change of gene expression in BC vs normal DE analysis, the P-value specifying statistical significance of the DE analysis, and the dataset from which samples were compared. These TFs are prospective biomarkers for BC and are further tested for prognostic value.

**Table 7: Predicted Biomarkers for MCF7 Showing DE in Oncomine Breast Cancer vs. Normal Datasets**

| Biomarker | Fold Change | P-value | Dataset |
|---|---|---|---|
| SP1 | 1.845 | 1.29E-24 | TCGA-BRCA |
| E2F1 | 2.734 | 1.68E-22 | TCGA-BRCA |
| MYC | 2.222 | 1.54E-19 | Finak-Breast |
| NHLH1 | 1.278 | 4.92E-134 | TCGA-BRCA 2 |
| INSM1 | 1.981 | 5.04E-20 | Finak-Breast |
| STAT3 | 1.566 | 5.50E-8 | TCGA-BRCA |
| NFYA | 1.952 | 3.47E-21 | TCGA-BRCA |
| NFKB1 | 2.236 | 3.66E-14 | Finak-Breast |
| MYOD1 | 5.974 | 2.16E-6 | Zhao Breast |

**Table 8: Predicted Biomarkers for MDA-MB-231 Showing DE in Oncomine Breast Cancer vs. Normal Datasets**

| Biomarker | Fold Change | P-value | Dataset |
|---|---|---|---|
| RUNX1 | 2.102 | 2.03E-17 | TCGA-BRCA |
| FOXA1 | 5.769 | 8.84E-9 | TCGA-BRCA |
| SOX17 | 1.203 | 1.78E-59 | TCGA-BRCA 2 |
| MEF2A | 2.271 | 7.92E-7 | Richardson Breast 2 |
| GFI1 | 1.596 | 7.88E-5 | Zhao Breast |
| ARID3A | 1.845 | 8.48E-6 | TCGA-BRCA |
| FOXO3 | 1.154 | 1.93E-6 | Curtis Breast |
| TAL1 | 3.445 | 3.38E-17 | Finak-Breast |
| FOXD1 | 4.665 | 2.43E-16 | Finak-Breast |
| LHX3 | 3.786 | 6.06E-19 | Finak-Breast |
| IRF1 | 2.051 | 1.39E-12 | TCGA-BRCA |
| MAX | 1.445 | 7.66E-8 | TCGA-BRCA |
| MYOD1 | 5.974 | 2.16E-6 | Zhao Breast |

### 3.5 Survival Analysis

To assess the prognostic value of TFs that were identified as prospective biomarkers, survival analysis was performed using patient datasets on the PROGgeneV2 tool. Genes validated using breast patient datasets on Oncomine were used as input genes in PROGgeneV2 and KM plots were generated for each, showing the effect of target gene expression on overall and metastasis-free survival in breast cancer patients over time. Plots of overall survival for prospective biomarkers identified in MCF7 and MDA-MB-231 are shown in figures 8 and 9 respectively. Figures 8D-I show the effect of high and low gene expression on overall patient survival of *MYOD1, NFKB1, NFYA, NHLH1, SP1* and *STAT3*. Expression of these genes did not show significant prognostic value for overall survival in breast cancer patients. The expression of the *MYC* gene showed a decrease in overall survival with high expression, but with a P-value of 0.2602, this was not statistically significant. This could be due to a low sample number; therefore, this gene will be further analysed for prognostic value. The *E2F1* and *INSM1* genes (figures 8A and 8B) showed significant prognostic value with high expression leading to a decrease in overall survival with a P-value of 9.31e-07 and 0.024 respectively.

Among prospective biomarkers identified using MDA-MB-231 expression data, the *ARID3A, FOXA1, FOXO3, GFI1, IRF1, LHX3, MEF2A* and *SOX17* genes (figures 9A, B, D, E, F, G, I and K) did not show significant prognostic value for overall survival in breast cancer patients. The *FOXD1* and *TAL1* genes (figure 9C and L) showed statistically significant prognostic value, with a high gene expression predicting a low overall survival with P-values of 0.016 and 0.0024 respectively. The *MAX* and *RUNX1* (figure 9H and J) genes are also predictive of patient survival, with low expression correlating with a decrease in overall patient survival with a P-value of 0.025 and 0.25 respectively. The prognostic value of *RUNX1* is not statistically significant, but again this could be due to a low sample number and this gene will be considered as a marker of positive prognosis based on effect size.

Based on their prognostic value, *MYC, E2F1, INSM1, FOXD1, TAL1, MAX* and *RUNX1* were selected as prospective biomarkers for further investigation. These genes were used to predict metastasis-free survival in breast cancer using PROGgeneV2 to assess their ability to predict breast cancer metastasis. KM-plots of metastasis-free survival are shown in figure 10. High expression of the *E2F1, FOXD1* and *INSM1* genes correlates with an increased probability of metastasis in breast cancer patients with P-values of 3.99e-07, 0.028 and 0.033 respectively (figures 10A-C). High expression of the *MYC* gene correlates with an increased probability of

metastasis, but with a P-value of 0.47, this is not statistically significant (figure 10E). *RUNX1* and *TAL1* were not found to be prognostic of metastasis in breast cancer (figures 10F and G). Higher expression of *MAX* correlated with a decreased probability of metastasis, but with a P-value of 0.12, this was not statistically significant (figure 10D).

### 3.6 TF-Gene Networks

Cytoscape was used to construct networks of showing the interaction between TFs selected as potential biomarkers and the genes they control. Figure 11 shows the interactions of biomarkers proposed using the set of genes upregulated in MCF7 cells, MYC, E2F1 and INSM1, with their predicted target genes. It can be seen that each of these TFs are predicted to regulate the expression of small groups of unique genes and the majority of genes are regulated by all three TFs. Similarly, in figure 12 showing the interactions of biomarkers proposed using the set of genes upregulated in MDA-MB-231 cells, FOXD1, MAX, TAL1 and RUNX1, small groups of unique genes are controlled by individual TFs, but the majority of genes are regulated in combination.

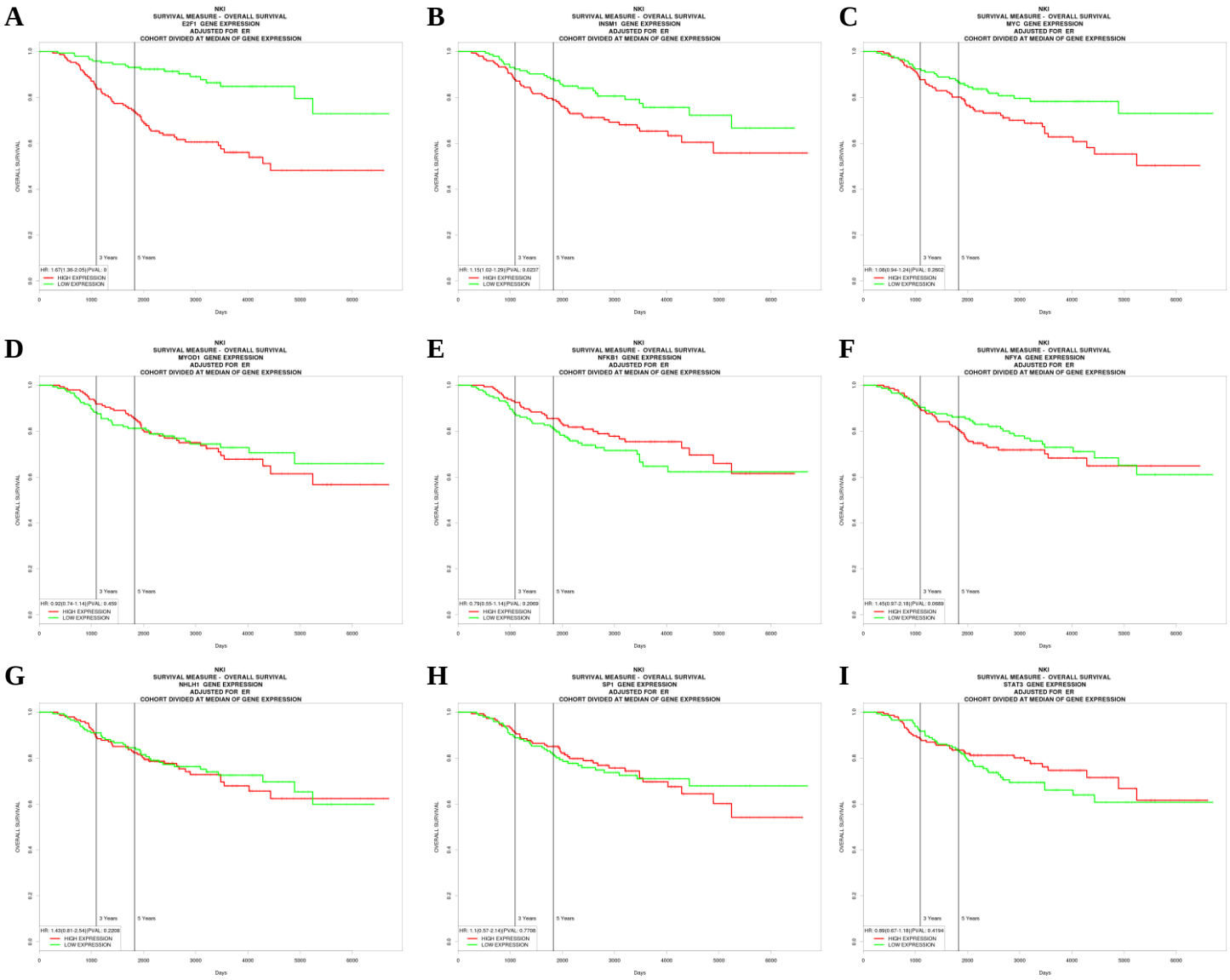



**Figure 8: Survival analysis of predicted biomarkers in MCF7 cells.** KM plots of overall survival of breast cancer patients. Probability of overall survival is represented on the y-axis and time in days is represented on the x-axis. The green line indicates low expression of the predicted biomarker and the red line indicates high expression.

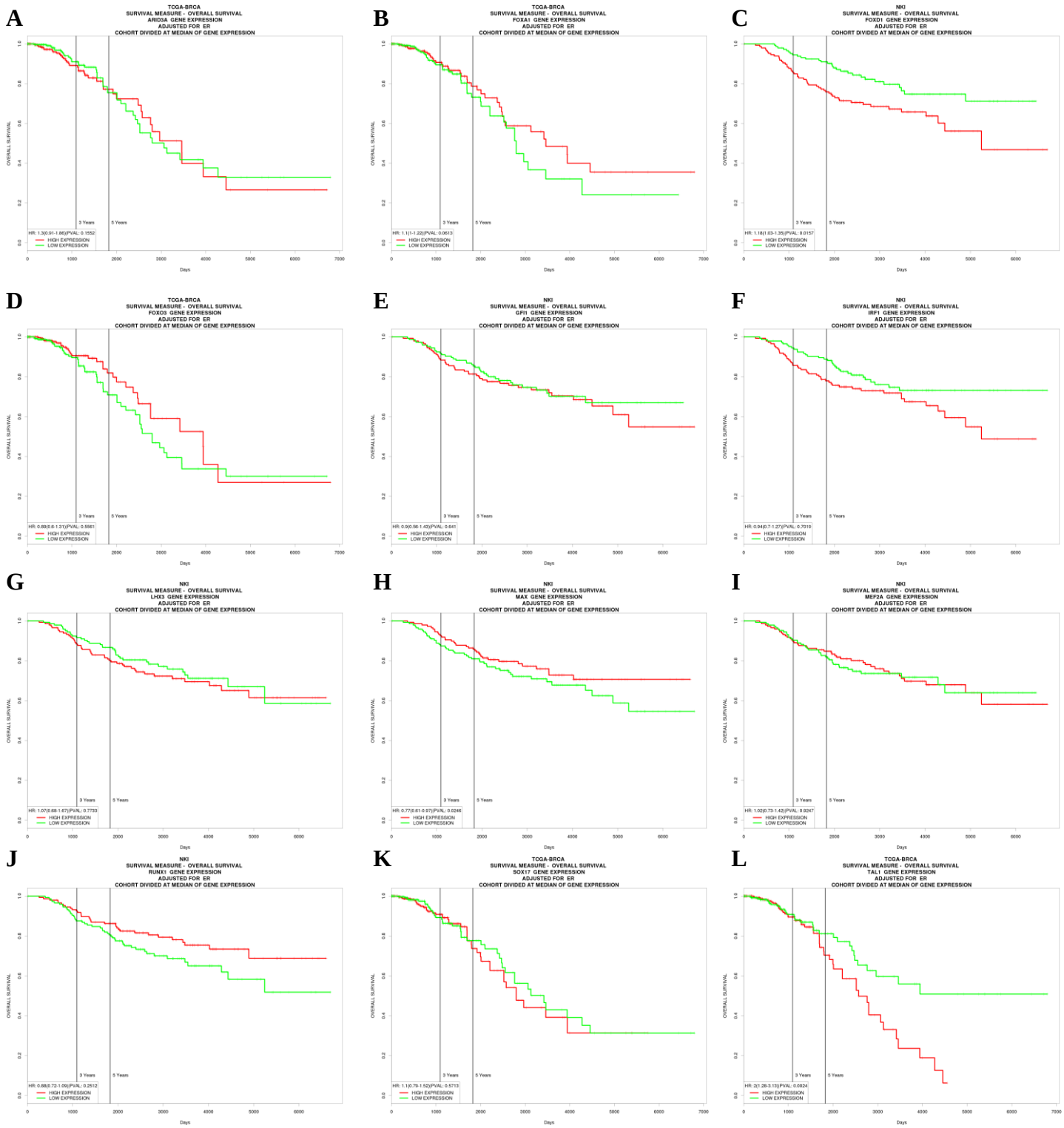



**Figure 9: Survival analysis of predicted biomarkers in MDA-MB-231 cells.** KM plots of overall survival of breast cancer patients. Probability of overall survival is represented on the y-axis and time in days is represented on the x-axis. The green line indicates low expression of the predicted biomarker and the red line indicates high expression.

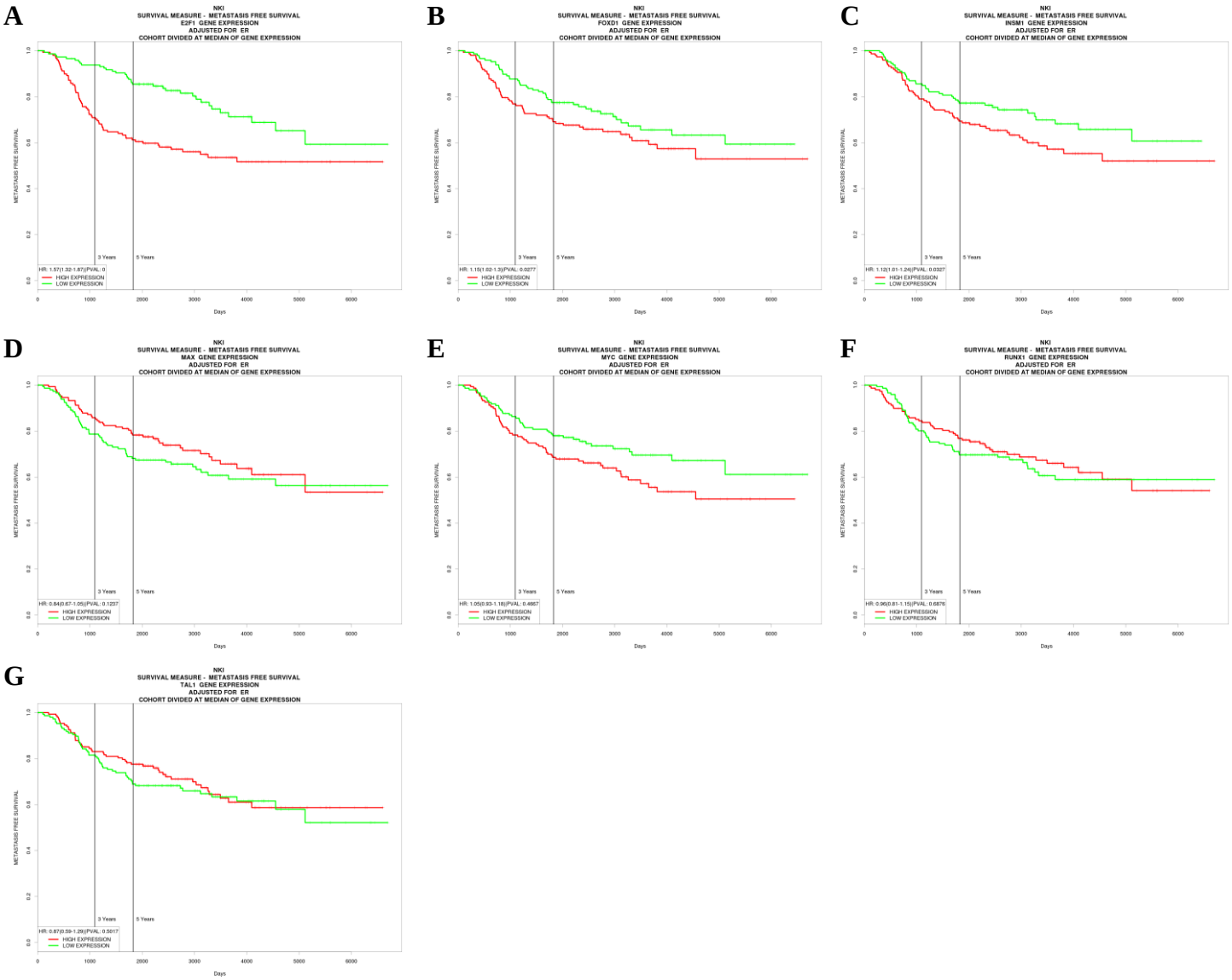



**Figure 10: Analysis of prognostic value of predicted biomarkers for breast cancer metastasis.** KM plots of metastasis-free survival of breast cancer patients. Probability of metastasis-free survival is represented on the y-axis and time in days is represented on the x-axis. The green line indicates low expression of the predicted biomarker and the red line indicates high expression.

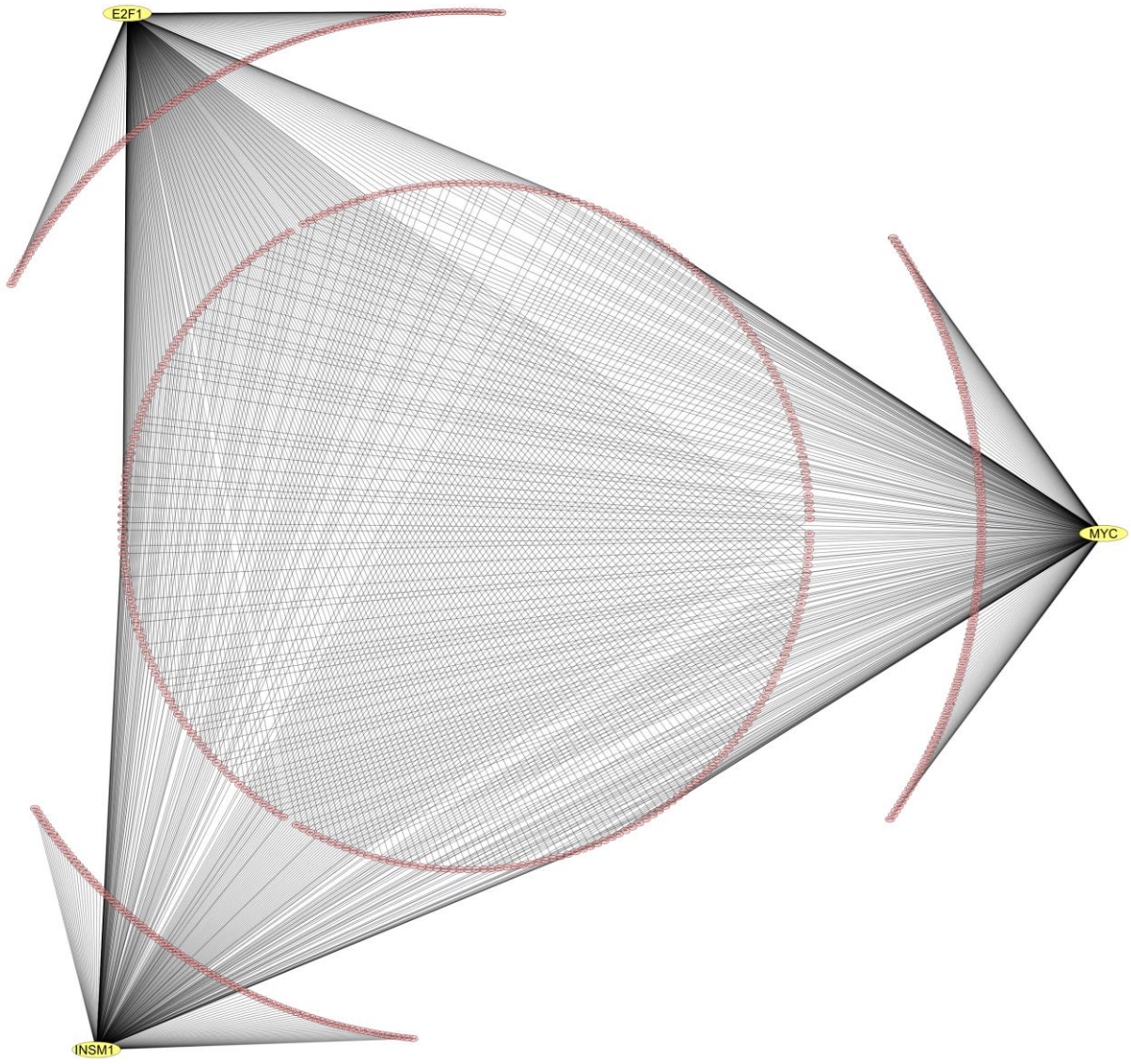



**Figure 11: Network of TF-gene interactions.** TFs predicted as biomarkers from MCF7 expression data are coloured yellow. The genes that they control from the set of genes upregulated in MCF7 cells compared with MCF10A cells are coloured in pink. The black lines represent a predicted regulatory interaction.

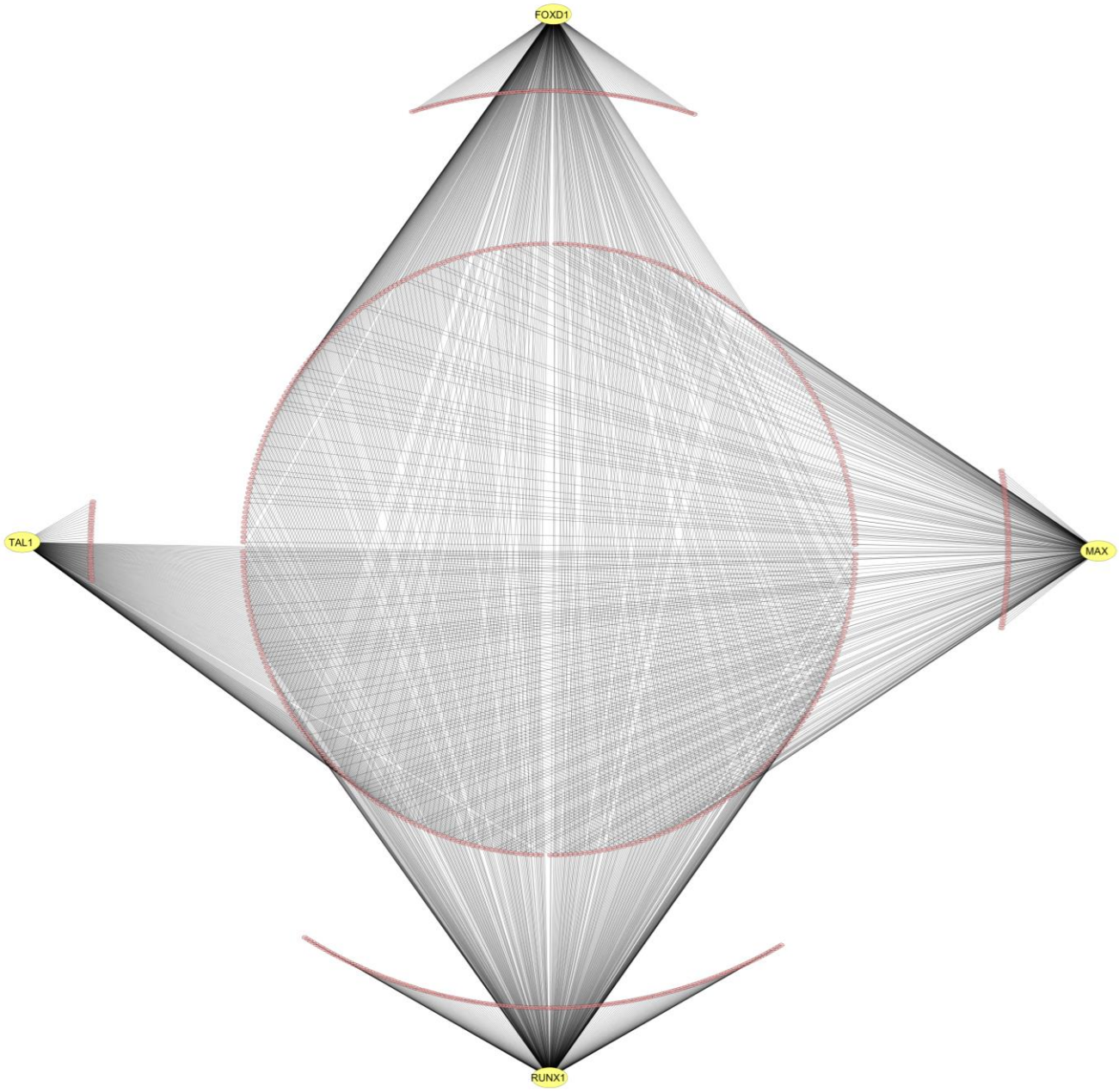



**Figure 12: Network of TF-gene interactions.** TFs predicted as biomarkers from MDA-MB-231 expression data are coloured yellow. The genes that they control from the set of genes upregulated in MDA-MB-231 cells compared with MCF10A cells are coloured in pink. The black lines represent a predicted regulatory interaction.

### 3.7 Predictive Value of Prospective Biomarkers

The sensitivity, specificity and precision of three prospective biomarkers, TAL1, FOXD1 and INSM1 were assessed using breast cancer patient datasets on the Oncomine platform. Table 9 shows these predictive measurements. TAL1 and FOXD1 both have a sensitivity of 98.11%, showing the high likelihood that these TFs will be expressed in breast cancer tissue. INSM1 has a sensitivity of 25%, which means that it may not be present in all breast cancer tissue. FOXD1 and INSM1 have specificity values of 100%, suggesting that these TFs are unlikely to be expressed in non-cancerous breast tissue. TAL1 has a specificity value of 66.67%, which suggests that it is sometimes expressed in non-cancerous breast tissue, but not in most cases. TAL1, FOXD1 and INSM1 have precision values of 96.3%, 100% and 100% respectively, suggesting that when these TFs are present, it is highly likely that the patient has breast cancer. These calculations are based on datasets with a limited sample number and need to be performed using larger datasets for improved accuracy.

**Table 9: Predictive value of prospective biomarkers.** The sensitivity, specificity and precision of selected TFs in breast cancer

| TF | Sensitivity (%) | Specificity (%) | Precision (%) |
|---|---|---|---|
| TAL1 | 98.11 | 66.67 | 96.3 |
| FOXD1 | 98.11 | 100 | 100 |
| INSM1 | 25 | 100 | 100 |

# CHAPTER FOUR – DISCUSSION AND CONCLUSIONS

Altered gene expression is known to play an important role in controlling the characteristics of cellular diseases such as cancer. The activation of genes involved in proliferation, angiogenesis and migration, and the deactivation of genes involved in growth suppression and cell death governs the aggressive growth and invasive characteristics associated with cancerous cells (Hanahan and Weinberg, 2011). The role of TFs as key controllers of gene regulatory networks suggests that they play a major part in the deviation from normal gene expression seen in tumours. Understanding the changes in gene expression patterns observed in cancer and the underlying regulatory mechanisms governing these changes is important in understanding the cause of this complex disease. Because a small subset of TFs regulate the expression of numerous genes, TFs are an ideal candidate for the study of cancer gene regulation. Identifying TFs controlling the characteristic hallmarks of cancer aids in predicting diagnostic markers for the early detection of cancer, prognostic markers for predicting patient outcome and response to therapeutic agents, and therapeutic targets for the development of targeted therapeutics as alternatives to the non-specific cytotoxic treatments used today. It was therefore the aim of this research project to identify gene expression changes in different breast cancer cell line models compared to a non-tumorigenic model and predict the TFs that control this expression. The biological roles of these TFs were studied and their prognostic value was predicted using patient sample data. The identified TFs could serve as biomarkers of breast cancer with diagnostic, prognostic and therapeutic potential.

## 4.1 Altered Gene Expression in Breast Cancer

Aberrant gene expression enables the acquisition of the abnormal characteristics observed in breast cancer. When comparing the gene expression patterns of breast cancer cell lines, MCF7 and MDA-MB-231, with the non-tumorigenic cell line, MCF10A, thousands of genes were found to be expressed at different levels (table 2). To visualise the expression differences between cell lines, heatmaps of the top 50 most DE genes were plotted for each DE comparison with genes clustered based on expression patterns (figure 5). Figure 5 shows clusters of gene expression, grouped using dendrograms on the y-axis, representing genes that are co-expressed and their expression patterns are characteristic to the cell lines in which they appear.

In MCF7 cells, the gene most significantly upregulated was the *PREX1* gene with a P-value of 4.071e-12 and a logFC of 10.03 compared to the non-tumorigenic MCF10A cell line. This gene

is implicated in the increased growth and metastasis of breast cancer cells and is shown to have higher expression levels in ER+ breast cancer than ER- breast cancer (Marotti et al., 2017). In MDA-MB-231 cells, the *BCAT1* gene is upregulated with a logFC of 13.09 compared to MCF10A and a P-value of 6.31e-11. This gene promotes the growth of breast cancer cells through promoting mitochondrial biogenesis and decreasing mitochondrial reactive oxygen species in breast cancer cells (Zhang and Han, 2017). These genes are among thousands with altered expression in breast cancer cells and contribute to the altered phenotype observed in breast tumours. The known roles these genes play in breast cancer progression validates the reliability of results obtained by DE analysis, but due to the high number of DE genes found, this project aims to find TF regulators of altered gene expression in order to find the smaller subset of TF regulating large changes in gene expression.

Due to the high number of DE genes, it is not viable to analyse the function of these genes individually. For this reason, GO and KEGG enrichment was performed to assess the collective functions of co-expressed DE genes (Figures 6 and 7). Figure 6A shows the GO biological processes over-represented in genes upregulated in MCF7 cells vs MCF10A cells, compared to a background set of genes. Of note, the GO terms, “Mesenchymal Cell Development” and “Embryonic Organ Development” were enriched in this gene set. Cancer cells are phenotypically similar to human embryonic cells in terms of their stem-like and highly proliferative nature. It is observed that embryonic genes are re-expressed in cancer cells, promoting the tumour phenotype (Monk and Holding, 2001). Mesenchymal cells are multipotent and an epithelial-to-mesenchymal transition (EMT) is observed in tumours, in which tumour cells lose epithelial markers and regress into a less differentiated, stem-like state, promoting invasion and migration and metastasis (Micalizzi and Ford, 2009). Interestingly, there is an enrichment of terms involved in nervous system development, such as “Nervous System Regulation” and “Regulation of Neurogenesis”. Studies have shown that the nervous system facilitates tumour metastasis and an inhibition of cell death through neural factors (Douma et al., 2004). Furthermore, there is an inverse correlation between cancer and neurodegenerative disorders, with tumours expressing genes related to neurogenesis, which are lost in neurodegenerative disorders (Plun-Favreau et al., 2010). This is demonstrated by the over-representation of nervous system related genes overexpressed in MCF7 cells.

Figure 6B shows the GO biological processes over-represented in genes upregulated in MDA-MB-231 cells vs MCF10A cells. The GO terms, “positive regulation of cell proliferation” and “positive regulation of cell development” are enriched in this gene set and are characteristic of

the highly proliferative nature of cancer cells. There is also an enrichment of nervous system related genes, which is explained above. The enriched terms, “branching morphogenesis of an epithelial tube”, “morphogenesis of a branching structure” and “morphogenesis of a branching epithelium” refer to the formation of branched ducts from the epithelial bud, which occurs during breast development in the embryo and during puberty (Sternlicht, 2006). This is a key stage in breast development, but aberrant expression of these genes at inappropriate times causes abnormal development of breast tissue which is characteristic of breast tumours.

Figure 7A shows the KEGG pathways over-represented in genes upregulated in MCF7 cells vs MCF10A cells. Enrichment of “transcriptional misregulation in cancer” and “pathways in cancer” is expected in genes upregulated in cancer. “Estrogen signalling pathway” genes were enriched, which is characteristic of an ER+ breast cancer. Enrichment of the “insulin secretion” pathway corroborates studies showing increased levels of the insulin receptor in breast cancer tissue compared to normal tissue (Papa et al., 1990). Enrichment of the “cAMP signalling pathway” could be linked to the increased glucose uptake observed in tumours. cAMP regulates glucose transporter GLUT3, which is shown to increases glucose uptake in breast cancer cells (Meneses et al., 2008).

Figure 7B shows the KEGG pathways over-represented in genes upregulated in MDA-MB-231 cells vs MCF10A cells. Again, there is enrichment of “transcriptional misregulation in cancer” and “pathways in cancer”, which is characteristic of any cancer cell line. The “cell adhesion molecules” pathway is enriched in this gene set. The process of cell adhesion is facilitated by interactions between tumour cells and cells of the endothelium. Cell adhesion molecules facilitate these interactions, allowing for invasion and metastatic spread of cancer cells (Bendas and Borsig, 2012). Since MDA-MB-231 is an aggressive cell line, it is expected that genes involved in metastasis are upregulated. The enriched GO biological processes and KEGG pathways in sets of upregulated genes give insight into the collective functions of these genes and the biological processes that are deregulated in each cancer cell type. Furthermore, by comparing the enriched functions with studies on cancer cell lines and tumours, we are able to validate the DE analysis performed, as the functions represented by the DE gene sets correlate with characteristics seen in breast cancer tumours.

### 4.2 TFs as Controllers of Cancer Gene Expression

TFs are proteins that control the expression of many genes. oPOSSUM was used to predict the TFs controlling genes upregulated in MCF7 (table 3) and MDA-MB-231 (table 4) compared to MCF10A cells. The TFs were predicted by analysing JASPAR TFBS motifs in the promoter regions of upregulated genes. TFs predicted were validated in patient datasets using the Oncomine platform and further validated for prognostic value by performing survival analysis. Networks of TF-gene interactions constructed in Cytoscape are shown in figures 11 and 12 using TFs predicted to regulate genes upregulated in MCF7 and MDA-MB-231 respectively. Each TF is predicted to regulate the expression of a unique set of genes, but the majority of genes are regulated by all the TFs in combination. This shows that although each proposed biomarker may have a unique effect on the characteristics of breast cancer, they often act in combination to affect the overall phenotype of the disease. This suggests that the main action of each proposed biomarker is involved in controlling similar biological processes, which may contribute to breast cancer. The small sets of unique genes that each proposed biomarker regulates may be responsible for the unique effect each TF has on the patient prognosis and behaviour of the tumour in which it is expressed. The functions of the TFs selected as prospective biomarkers will be discussed in detail in the following pages.

### 4.3 Prediction of Known Breast Cancer Biomarkers

This study was aimed at identifying TFs that control the aberrant gene expression observed in breast cancer cells. It was expected that the methodology would identify known and unknown TFs that are involved in breast cancer. The prediction of E2F1, MYC and MAX as potential biomarkers confirmed this expectation, as these TFs are known prognostic markers of breast cancer. The identification of known biomarkers supports the validity of the methodology used in the present study, providing confidence that the unknown biomarkers predicted have similar functions and are likely to be involved in breast cancer.

#### 4.3.1 E2F1 as a Breast Cancer Biomarker

The E2F1 protein is a member of the E2F TF family, a family involved in the control of the cell cycle, DNA damage repair and the induction of apoptosis. Another member of this family (E2F5) has been shown to act as a biomarker for ovarian cancer (Kaur et al., 2011). E2F1 is shown to cooperate with p53 to induce apoptosis in the absence of growth factors (Wu and Levine, 1994). This is carried out through direct interaction with the tumour suppressor RB1.

It is, therefore, surprising that E2F1 expression is associated with tumour progression, invasion and migration in various cancers (Liang et al., 2016, Vuaroqueaux et al., 2007). The normal pro-apoptotic function of E2F1 seems to be lost in cancer, where instead, it promotes aggressive tumorigenesis. A possible mechanism of this has been proposed by Wang *et al.* in a 2015 study showing that through direct binding to the EPC1 (enhancer of polycomb homolog 1) protein, E2F1 upregulates the expression of genes such as *BCL-2* and *BIRC5,* which are involved in anti-apoptotic cell survival (Wang et al., 2016). Co-operation between EPC1 and E2F1 is shown to induce metastasis-related genes and predict poor patient outcome, which is consistent with the results observed in this study (Figures 8A and 10A).

In breast cancer, low levels of transcription of the *E2F1* gene correlates with a favourable patient outcome (Vuaroqueaux et al., 2007). Furthermore, a high expression of the AAA nuclear coregulator cancer-associated protein (ANCCA), a co-activator of E2F1 that recruits E2F1 to specific chromatin locations, is highly expressed in cases of tumour metastasis, recurrence, poor survival and triple-negative breast cancer (Kalashnikova et al., 2010). NCOA3 is another co-activator of E2F1 implicated in cancer progression. The co-operation of these proteins is shown to promote estrogen-independent proliferation in breast cancer cells and additionally, high expression of NCOA3 both enhances cell sensitivity to the proliferative effects of estrogen and negates the effects of Tamoxifen, a commonly used breast cancer drug, which may be through interaction with E2F1 (Louie et al., 2004). In addition to anti-estrogen resistance, the E2F1 signalling network has been shown to be upregulated in chemotherapy-resistant cell lines (Andersen et al., 2010). These studies show the prognostic value of E2F1 in predicting patient outcome, metastasis and drug resistance in breast cancer patients.

The expression of E2F1 in the breast seems to be cancer specific, which is important for diagnostic testing. In 2000, Zhang *et al.* showed that in patient samples, only 1.9% of cells were positive for E2F1 expression in normal breast tissue, but this percentage increased to 6.3% in ductal carcinoma tissue and 15.3% in invasive ductal carcinomas (Zhang et al., 2000). Furthermore, the expression of E2F1 increased with more advanced stages of breast cancer. This study shows the value of E2F1 as a diagnostic marker for breast cancer. Additionally, expression of this gene correlates with the expression of genes in the MammaPrint 70-gene signature panel and as a single gene, shows similar value in predicting patient outcome (Vuaroqueaux et al., 2007). This demonstrates the importance of E2F1 as a biomarker for breast cancer. In addition, the role of E2F1 in apoptotic pathways suggests that when it is expressed, anti-apoptotic drugs may not be suitable for cancer treatment and through activating pathways

which may include E2F1, may actually have the undesired effect of supporting cancer cell proliferation. It is therefore important to study the mechanisms of E2F1 signalling and measure its expression prior to prescribing treatment to breast cancer patients.

Identification of E2F1 also highlights that the methodology used to predict biomarkers in the current study is able to identify known proposed biomarkers. This gene is shown to control the genes upregulated in the ER+ breast cancer cell line, MCF7 (table 3). It is also shown to be upregulated in breast cancer patient datasets (table 7). High expression of E2F1 correlates with a decrease in overall patient survival and a decrease in metastasis-free survival (figures 8A and 10A). The results of the present study combined with previous research demonstrate that the utility of E2F1 as a potential breast cancer biomarker, which could be used to predict patient outcome and response to certain anti-proliferative drugs, and could be used as a therapeutic target to improve the efficacy of available breast cancer treatments.

### 4.3.2 MYC and MAX as Breast Cancer Biomarkers

*MYC* and *MAX* (MYC Associated Factor X), are genes encoding basic helix-loop-helix leucine zipper family TFs. MAX forms homodimers or heterodimers with members of the same family. These include Mad, Mxi1 and MYC. The formation of different dimers adds complexity to the transcriptional regulation carried out by MYC and MAX. Together, these TFs regulate many genes involved in proliferation and their role in cancer is well-established (Xu et al., 2010).

The role of MAX on its own in cancer is not well-described in literature, as it primarily acts to enhance the function of MYC through heterodimerization. The MYC-MAX heterodimeric complex binds to the consensus sequence CACGTG within the E-box (enhancer box) of target genes, either activating or inhibiting transcription (Blackwood and Eisenman, 1991). MYC has many roles in breast cancer progression, including increased proliferation and cell cycle activation (Dang et al., 2006). MYC is also involved in promoting EMT, angiogenesis, endocrine resistance, and blocking differentiation of cancer cells (Xu et al., 2010). This explains the decrease in overall survival seen in figure 8C and the decrease in metastasis-free survival seen in figure 10E with increased expression of MYC. The expression of MYC generally correlates with breast cancer subtype, with high expression observed in ER- basal-like tumours and lower expression observed in ER+ luminal tumours (Sorlie et al., 2001). The present study found that the target genes of MYC were highly expressed in the ER+ cell line (table 3), but MAX was found to regulate ERE-containing genes in the ER- cell line (table 6). This suggests that the MYC-MAX complex may activate genes normally controlled by ERα in

ER- breast cancer where ERα is not expressed. The seemingly contradicting result that shows MYC as a regulator of ER+ breast cancer can be explained by the role of MYC as a downstream effector in the ERα pathway (Dubik and Shiu, 1992). A study by Dubik *et al.* shows that MYC is upregulated in ER+ breast cancer cells in an estrogen-dependant manner (Dubik, 1987). This confirms the role of MYC as a controller of ER+ breast cancer characteristics. The role of MYC and its heterodimerization with MAX is well-studied, further confirming the validity of methods used in the present study. The disruption of the interaction between these two TFs has been proposed as a potential therapeutic strategy (Vita and Henriksson, 2006). The results of the present study further support the established research showing the potential of MYC and MAX as diagnostic markers and targets for therapeutic intervention.

### 4.4 Identification of Unknown Breast Cancer Biomarkers

This study has resulted in the prediction of potential breast cancer biomarkers that are not well-studied. These biomarkers include INSM1, FOXD1, TAL1 and RUNX1. Many of these TFs are known markers of different cancer types, such as leukaemia, but have not been studied as breast cancer biomarkers. These potential biomarkers are not completely unknown, but few studies link them to breast cancer and their role as a breast cancer biomarker has not yet been established. The prediction of these TF as potential biomarkers for breast cancer may lead to better diagnostic or therapeutic strategies after further experimental investigation.

#### 4.4.1 INSM1 as a Breast Cancer Biomarker

INSM1 (Insulinoma-associated protein 1) is a zinc-finger TF that plays a role in neuroendocrine differentiation and neurogenesis during embryonic development (Xie et al., 2002). It acts as a transcription repressor by recruiting histone deacetylases, such as HDAC1-3, KDM1A and RCOR1, to its target genes. INSM1 usually promotes cell cycle arrest and inhibition of proliferation. Its expression is rarely detected outside of the embryonic stages under normal circumstances. This makes it highly specific for the detection of ectopic expression. INSM1 is found to be highly expressed in tumours of neuroendocrine origin, such as small-cell lung cancer and its expression was absent in tumours of non-neuroendocrine origin, making it a highly specific marker of neuroendocrine tumours (Pedersen et al., 2006).

The role of INSM1 has not been studied in breast cancer. A study by Rosenbaum *et al.* showed the expression patterns of INSM1 in various cancer tissues (Rosenbaum et al., 2015). This study showed that expression of ISNM1 was almost exclusive to neuroendocrine neoplasms,

including neuroendocrine carcinoma of the breast. The only non-neuroendocrine patient sample that expressed INSM1 was a breast adenocarcinoma, suggesting that this protein could have a role in a small proportion of breast cancers. The only proposed link between INSM1 and breast cancer is through the breast cancer and salivary gland expression (*BASE*) gene, a relatively unstudied gene with that is almost exclusively expressed in breast tumours (Bretschneider et al., 2008). It is found to be expressed in many breast cancer cell lines and in 5 in 8 breast tumour samples. Binding sites for both ERα and INSM1 are found in the promoter of the *BASE* gene and expression of *INSM1* correlates with *BASE* when the expression of *BASE* is not regulated by estrogen (Bretschneider et al., 2008). This suggests that INSM1 plays a role in regulating the expression of breast-cancer associated genes by either interacting with ERα or activating its target genes when ERα is absent. In the present study, INSM1 was predicted to control genes involved in ER+ breast cancer (table 3), which suggests a co-operative action with ERα. Based on the available research, no conclusions can be made on the role of INSM1 in breast cancer or as a biomarker for breast cancer cases without neuroendocrine origin. Regardless of that, INSM1 shows potential as a prognostic marker of survival and metastasis in breast cancer patient data (Figure 8B and 10C) and should be studied further to elucidate its role in breast cancer.

INSM1 shows high specificity and precision values for predicting breast cancer in patients (table 9). The high specificity shows that INSM1 is unlikely to be expressed in patients without breast cancer and the high precision value shows that if INSM1 is present, the patient is likely to have breast cancer. INSM1 has a low sensitivity value, meaning that it is not present in all breast cancer patients, and is therefore not useful in detecting breast cancer in general. This is because the expression of INSM1 is specific to certain subtypes of breast cancer. This TF may be useful as a biomarker for certain types of breast cancer, such as neuroendocrine-derived breast cancer, and can therefore help in deciding the treatment strategy and predicting the prognosis of breast cancer patients.

### 4.4.2 FOXD1 as a Breast Cancer Biomarker

FOXD1 is a member of the forkhead family of TFs, which share a 100 amino acid winged-helix DNA-binding domain. FOXD1 plays an important role in kidney development and its expression is usually restricted to cortical interstitial cells, which give rise to various components of the kidney during development (Levinson et al., 2005). The role of FOXD1 in cancer is not well-established. FOXD1 expression is found to be upregulated in breast cancer

cells (Zhao et al., 2015). FOXD1 overexpression in the MCF7 cell line caused an increase in cell proliferation and chemoresistance, and a decrease in expression in MDA-MB-231 cells showed decreased proliferation and chemoresistance (Zhao et al., 2015). FOXD1 also induces cell cycle progression from G1 to S phase by directly regulating the expression of p27 (Zhao et al., 2015). This suggests that FOXD1 increases proliferation and treatment resistance in breast cancer, making it a valuable marker of prognosis and a potential therapeutic target. The expression of this gene is also highly restricted to certain tissues of the kidney, making detection of ectopic expression easier. This adds to the specificity of the biomarker, as the protein is not usually expressed in the breast. The present study shows that FOXD1 controls the expression of genes upregulated in the ER- cell line MDA-MB-231 (table 4), which could be involved in increased proliferation and chemoresistance, as observed in the study discussed above (Zhao et al., 2015). The increased expression of FOXD1 observed in breast cancer patient data (table 8) and the decrease in overall patient survival (figure 9C) and metastasis-free survival (figure 10B) observed with increased expression of this TF implicate it as a possible controller of the breast cancer phenotype. It is therefore worth further investigating the role of this TF as a potential biomarker for ER- breast cancer.

FOXD1 has a high sensitivity, specificity, and precision in predicting breast cancer (table 9). The sensitivity shows that it is likely to be present in breast cancer patients, making it valuable for diagnosis and screening. The high specificity shows that it is not likely to be expressed in non-cancerous breast tissue. The high precision value shows that FOXD1 is has positive predictive value in breast cancer and could be used for diagnosis and screening of the disease.

### 4.4.3 TAL1 as a Breast Cancer Biomarker

TAL1 is a basic helix-loop-helix TF predominantly expressed in erythroid cells, which are precursors to erythrocytes and megakaryocytes. It is an important regulator of haematopoiesis and is required for the development of endothelial tissue (Sanda and Leong, 2017). TAL1 forms a heterodimer with class 1 basic helix-loop-helix proteins which regulates gene expression in erythroid cells through binding of GATA1, LDB1 and LMO2 proteins (Kassouf et al., 2010). The TAL1 complex activates several genes including the MYB and TRIB2 oncogenes (Mansour et al., 2013). TAL1 also inhibits the expression of certain genes, such as the tumour suppressor FBXW7, through interaction with various miRNAs (Mansour et al., 2013). Deregulation of TAL1 promotes tumorigenesis through both the activation of oncogene expression and the repression of tumour suppressor expression. The interaction of TAL1 with

these genes may lead to increased tumour aggression and a worse patient outcome, which would explain the decrease in overall patient survival observed in figure 9L.

High expression or gain-of-function mutations of TAL1 are associated with diseases such as leukaemia (O'Neil et al., 2004). This is related to the role of TAL1 as a master regulator of haematopoiesis. The role of TAL1 in breast cancer is not well-studied, although some targets of TAL1 are involved in breast cancer progression. The *MYB* and *TRIB2* oncogenes and the *FBXW7* tumour suppressor mentioned above are important targets of TAL1. The expression of the *MYB* gene activates the Wnt/β-catenin signalling pathway, which increases cell proliferation in breast cancer (Li et al., 2016). It also enhances metastasis in breast cancer through interaction with Axin2 (Li et al., 2016). TRIB2 activates the PI3K/Akt pathway in cancer cells, which is also associated with increased proliferation in cancer (Hill et al., 2017). TRIB2 is implicated in resistance to chemotherapy and PI3K inhibitors in various cancers through the activation of Akt. This affects the efficacy of drugs that target upstream components of the PI3K/Akt pathway, such as trastuzumab, and are used for treatment of breast cancer. *MYB* and *TRIB2* expression are both activated by TAL1. TAL1 also inhibits the expression of *FBXW7,* which is a tumour suppressor that inhibits proliferation and promotes apoptosis in breast cancer and is prognostic of a better outcome in breast cancer patients (Chen et al., 2018). A further link to breast cancer is through the miRNA, miR-140-5p. This miRNA has tumour suppressive characteristics in breast cancer and is downregulated by ERα signalling (Zhang et al., 2012). Expression of miR-140-5p downregulates the expression of TAL1 (Correia et al., 2016). The downregulation of this miRNA in breast cancer could therefore increase the expression of TAL1, leading to increased activation of oncogenes, inhibition of tumour suppressors, and possibly increased angiogenesis through the activation of endothelial-related genes. Furthermore, breast cancer patients are at high risk for developing acute myeloid leukaemia as a secondary malignancy (Martin et al., 2009). This is often attributed to an interaction between certain breast cancer therapies and a genetic predisposition. TAL1 could therefore be useful in predicting the development of leukaemia as a secondary malignancy and also help physicians avoid prescribing therapies that are related to leukaemia development, such as alkylating agents, anthracyclines and radiotherapy, when TAL1 expression is high. Additionally, Figure 9L shows that a high expression of TAL1 leads to a drastic decrease in overall survival of breast cancer patients. It is therefore important to further elucidate the role of TAL1 in breast cancer and to investigate its potential as a possible target for therapeutic intervention and as a diagnostic marker.

TAL1 has high sensitivity and precision values for predicting breast cancer (table 9). This demonstrates that it is likely to be expressed in breast cancer tissue and could be a useful marker for the diagnosis and screening of breast cancer. It also had a moderately high specificity value of 66.67% which suggests that it may be present in some breast tissue that is not cancerous. Combined with other biomarkers, TAL1 is still useful in the diagnosis of breast cancer. The value of this TF as a biomarker is in its ability to predict negative outcome in the breast cancer patients in which it is expressed (figure 9L). After further elucidation of its role in breast cancer, TAL1 could be a valuable prognostic marker in patients and could also be a useful therapeutic target.

### 4.4.4 RUNX1 as a Breast Cancer Biomarker

RUNX1 is a TF involved in development of haematopoietic stem cells and their differentiation into mature myeloid and lymphoid cells (Okuda et al., 2001). The tumour suppressive role of RUNX1 in leukaemia is well-established (Ito et al., 2015). Recently, RUNX1 expression has been detected in non-tumorigenic breast epithelial tissue, suggesting a role in regulating normal breast cell behaviour (Ito et al., 2015). In ER+ breast cancer, loss-of-function mutations are observed, usually due to a mutation in the DBD (van Bragt et al., 2014). In the breast, RUNX1 represses the expression of *ELF5*, a regulator of alveolar cells. ELF5 has been shown to suppress ER-driven gene expression patterns and decreases estrogen sensitivity in breast cancer (Kalyuga et al., 2012). Knockdown of *ELF5* in ER- breast cancer cell lines suppressed basal gene expression patterns and shifted the molecular subtype towards a normal-like state (Kalyuga et al., 2012). In ER+ breast cancer cells, high expression of *ELF5* caused an increase in acquired Tamoxifen resistance, as the repression of ER-related gene expression caused cells to become dependent on ELF5 instead of estrogen and ERα (Kalyuga et al., 2012). Therefore, the repression of *ELF5* expression by RUNX1 could decrease Tamoxifen resistance and reduce the proliferative capacity of breast cancer cells.

Although the role of RUNX1 in breast cancer has not been fully elucidated, studies suggest that in ER+ breast cancer, the loss of RUNX1 function combined with either the loss of RB1 or p53 leads to an increase in ER-mediated proliferation (van Bragt et al., 2014). The loss of RUNX1 is also associated with increased metastasis and poor clinical outcome (Hong et al., 2017). This is consistent with the results shown in Figure 9J, which shows a decrease in overall patient survival with a decreased expression of RUNX1. The loss of RUNX1 function allows TGF-β to induce EMT, suggesting a role for RUNX1 in inhibiting EMT by increasing E-

Cadherin expression (Hong et al., 2017). It is worth noting that a very high expression also correlates with poor patient outcome and increased metastasis, especially in triple negative breast cancer (Browne et al., 2015). This could be due to an increased expression of E-cadherin, which induces metastasis through interaction with Twist1 (Shamir et al., 2014). Therefore, abnormally high or low levels of RUNX1 promote tumour progression and invasion through its regulation of E-Cadherin. This makes RUNX1 an important marker of tumour progression, as any deviation from normal expression causes an increase in tumour aggression and may lead to poor patient outcome.

### 4.5 Conclusions

This study aimed to determine the changes in gene expression in ER+ and ER- breast cancer compared to non-tumorigenic cells and the TFs that control these changes in gene expression. DE analysis showed that 1894 genes were upregulated and 2502 genes were downregulated in MDA-MB-231 compared to MCF10A, 2484 genes were upregulated and 2484 genes were downregulated in MCF7 compared to MDA-MB-231, and 1924 genes were upregulated and 2543 genes were downregulated in MCF7 compared to MCF10A. Many of these genes are involved in biological processes that contribute to the hallmarks of cancer. This confirms that gene expression patterns are significantly altered in cancer cells, which may be the primary controller of the cancer phenotype. TFBS enrichment analysis was used to predict the TFs responsible for this change in gene expression. The expression of these TFs and their prognostic value were validated using publicly available breast cancer patient datasets. This resulted in the selection of the TFs E2F1, INSM1 and MYC as potential biomarkers specific to MCF7 cells, and FOXD1, TAL1, MAX and RUNX1 as potential biomarkers specific to MDA-MB-231 cells. These biomarkers could have utility in breast cancer diagnosis, prognosis and therapy. Of these selected TFs, E2F1, MYC and MAX are well-studied in breast cancer and are shown to have prognostic value. This validates the methodology used, as it was able to predict known breast cancer-associated TFs. The role of RUNX1 in breast cancer is less clear, but at least one study has linked it to breast cancer and further investigation will confirm its role in this disease. Predicted biomarkers TAL1, FOXD1 and INSM1 have not yet been directly linked to breast cancer, but play a role in different cancer types. Proposed links to breast cancer for these potential biomarkers have been discussed above and they show potential as prognostic markers based on their effect on patient survival. Further investigation is required to elucidate the roles these proposed biomarkers have in breast cancer, but due to the ability of the

methodology used to predict known and novel biomarkers, and the proposed mechanisms in breast cancer, they show promise as novel biomarkers which may have clinical use. In conclusion, the biomarkers proposed in this study could be promising in the development of novel diagnostic and therapeutic strategies or the improvement of existing strategies for better breast cancer diagnosis and treatment.

## 4.6 Troubleshooting

### 4.6.1 Differential Expression

Initially, the aim of the present study was to evaluate the differential expression and TSS usage of ER+, ER- and non-tumorigenic cells using FANTOM5 CAGE data. The FANTOM5 project aimed to investigate transcription activity in different types cells. They accomplished this by first performing CAGE analysis using hundreds of human and mouse primary cells, cell lines and tissues (Forrest et al., 2014). Data from this project is publically available for use at http://fantom.gsc.riken.jp/5/. Initially, CAGE data for MCF7, MDA-MB-453 and HEK293 cell lines were used to represent ER+, ER- and non-tumorigenic cells respectively. The first problem encountered was that HEK293 was the only non-tumorigenic cell line available in this study. This is a human embryonic kidney cell line and while it can be used as a non-tumorigenic cell line to compare the effect of drugs and other compounds with cancer cells, non-tumorigenic breast cells are required to perform a DE comparison with breast cancer cells. The reason for this is that we want to observe the changes in gene expression between cancer and normal cells and the changes in expression between breast and kidney cells would interfere with this. The second problem was that there was only one replicate for each cell line, meaning the DE analysis would not be statistically significant. For these reasons, it was decided to use RNA-seq data published on GEO, which contained 3 replicates each of MCF7, MDA-MB-231 and MCF10A, representing ER+ breast cancer, ER- breast cancer and non-tumorigenic breast cells respectively (Messier et al., 2016).

### 4.6.2 ERE Prediction

The tool used for detection of EREs in the promoters of target genes, Dragon ERE Finder version 2, was published in the year 2003 (Bajic et al., 2003). For this reason, it was designed for use on older computer platforms and did not work on Ubuntu 18.04 64-bit edition. In order to run this program, support for 32-bit architecture had to be enabled through installation of multiple software packages including: libc6:i386, libncurses5:i386, libstdc++6:i386 and

lib32z1, and the i386 architecture had to be enabled. After this was done, the Dragon ERE Finder Version 2 software was able to run on Ubuntu 18.04 64-bit edition.

### 4.6.3 Scripting

The data generated by certain methods used in this study was not compatible with other methods. An example of this is Dragon ERE Finder 2, which generated an unstructured text output, which could not be read by any other programs. To solve this issue, programming scripts written in the Python programming language were used to modify and process the data into a usable format.

## 4.7 Future Studies

The biomarkers proposed in this study are based on computational predictions and need to be validated experimentally before conclusions about their clinical utility can be drawn. For this, future work will be focused on elucidating the molecular mechanisms of the proposed biomarkers in breast cancer cells. Firstly, the expression levels of the proposed biomarkers will be measured in breast cancer tissue samples obtained from patients using real-time PCR. These samples will be of different breast cancer subtypes, allowing the expression of the TFs to be compared across subtype. Biomarkers showing promise will need to be studied further to elucidate their molecular mechanisms in breast cancer. To do this, the cell lines MCF7, MDA-MB-231 and a non-tumorigenic cell line will be used. Protein-level expression will be measured using Western blotting. Knockdown and overexpression of the proposed biomarkers will be performed to further elucidate their role in breast cancer. The effects of knockdown and overexpression will be analysed using Western blotting and a variety of cell viability assays, such as the MTT assay, APOPercentage assay, mitochondrial outer membrane potential assay and the caspase 3/7 assay, in combination with drug treatments. These *in vitro* experiments will help to elucidate the role the proposed biomarkers have in breast cancer and their effects on cell viability and drug efficacy. The final testing of these biomarkers needs to be done in breast cancer tumour samples obtained from patients at different stages of the disease. This will provide a complete validation of the outcome of the present study. Nonetheless, a computational pipeline was established, which can be used to predict biomarkers for various types of cancers in future, once the results of the present study are experimentally validated.

### 4.8 Limitations

One of the major limitations of this study is the use of gene expression data originating from immortalised cell lines. Normal human tissue does not proliferate indefinitely and is therefore difficult to maintain in culture. Immortalised cell lines have accumulated mutations that allow for the evasion of cellular senescence and growth restrictions present in normal cells. For this reason, immortalised cell lines are commonly used in cancer research, as they can be grown *in vitro* for prolonged periods of time. The disadvantage of using immortalised cell lines is that they are not a perfect representation of human tumour tissue (Gillet et al., 2013). The frequency of DNA mutations observed in immortalised cell lines is higher than that observed in tumours (Dai et al., 2017). This leads to differences in gene expression patterns between immortalised cell lines and tumour tissue. The increase in DNA mutations lead to phenotypical changes in the cells, causing behavioural changes between different immortalised cell lines and tumour tissue of the same type. For this reason, the results of this study, and of any study relying primarily on the use of immortalised cell lines, should be considered preliminary. The predicted breast cancer biomarkers discovered in this study are at an early stage of development and should not be considered reliable prognostic markers until they are validated using patient tumour samples.

An additional limitation of *in silico* research such as the present study is the inconsistency of results obtained when using the computational tools with different parameters or using different computational tools. Where possible, different parameters and tools were tested to obtain reliable results, but this was not presented in the methodology or results of this dissertation. The tools used in the present study were assumed to be reliable based on numerous publications using them. Additionally, multiple gene expression datasets should have been tested with the tools to confirm that results were consistent. Due to time constraints and lack of availability of suitable data, this was not done. As mentioned above, the biomarkers identified in this study are preliminary predictions and require validation before they can be considered reliable biomarkers. In this case, the expression of the biomarkers in patient tumour samples of different subtypes will be assessed before publication of results.

# Appendix

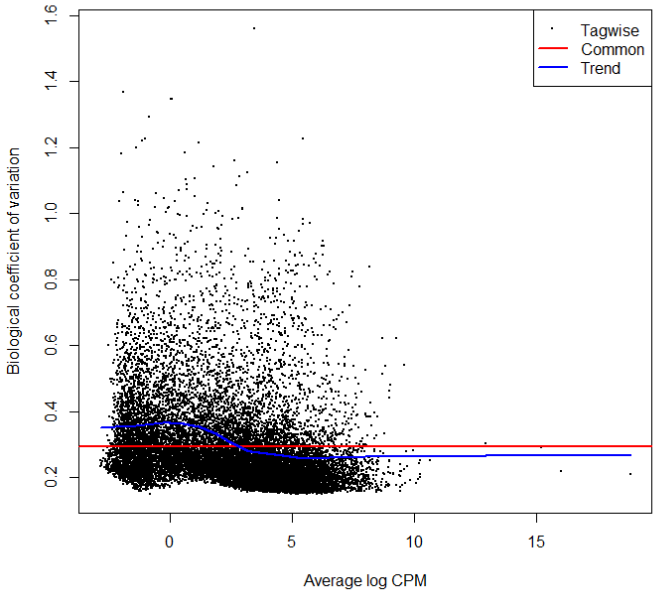



**Figure 13: Scatterplot of BCV against transcript abundance across samples.** Estimates of common, trended and tagwise NB dispersions are represented in red, blue and black respectively. The y-axis represents BCV and the x-axis represents average gene abundance as average logCPM.

## References


ADOMAS, A., HELLER, G., OLSON, Å., OSBORNE, J., KARLSSON, M., NAHALKOVA, J., VAN ZYL, L., SEDEROFF, R., STENLID, J. & FINLAY, R. 2008. Comparative analysis of transcript abundance in Pinus sylvestris after challenge with a saprotrophic, pathogenic or mutualistic fungus. *Tree physiology,* 28**,** 885-897.

ANDERS, S., PYL, P. T. & HUBER, W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics,* 31**,** 166-169.

ANDERSEN, J. B., FACTOR, V. M., MARQUARDT, J. U., RAGGI, C., LEE, Y. H., SEO, D., CONNER, E. A. A. T. & S.S 2010. An integrated genomic and epigenomic approach predicts therapeutic response to zebularine in human liver cancer. *Science translational medicine,* 2**,** 54-77.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S. & EPPIG, J. T. 2000. Gene Ontology: tool for the unification of biology. *Nature genetics,* 25**,** 25.

BAJIC, V. B., TAN, S. L., CHONG, A., TANG, S., STRÖM, A., GUSTAFSSON, J.-A., LIN, C.-Y. & LIU, E. T. 2003. Dragon ERE Finder version 2: A tool for accurate detection and analysis of estrogen response elements in vertebrate genomes. *Nucleic acids research,* 31**,** 3605-3607.

BARBERIS, A., PEARLBERG, J., SIMKOVICH, N., FARRELL, S., REINAGEL, P., BAMDAD, C., SIGAL, G. & PTASHNE, M. 1995. Contact with a component of the polymerase II holoenzyme suffices for gene activation. *Cell,* 81**,** 359-368.

BEDNAREK, A. K., SAHIN, A., BRENNER, A. J., JOHNSTON, D. A. & ALDAZ, C. M. 1997. Analysis of telomerase activity levels in breast cancer: positive detection at the in situ breast carcinoma stage. *Clinical cancer research,* 3**,** 11-16.

BELZEAUX, R., FORMISANO-TRÉZINY, C., LOUNDOU, A., BOYER, L., GABERT, J., SAMUELIAN, J.-C., FÉRON, F. & NAUDIN, J. 2010. Clinical variations modulate patterns of gene expression and define blood biomarkers in major depression. *Journal of psychiatric research,* 44**,** 1205-1213.

BENDAS, G. & BORSIG, L. 2012. Cancer cell adhesion and metastasis: selectins, integrins, and the inhibitory potential of heparins. *Int J Cell Biol,* 2012**,** 676731.

BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B,* 289-300.

BIEGING, K. T., MELLO, S. S. & ATTARDI, L. D. 2014. Unravelling mechanisms of p53-mediated tumour suppression. *Nature Reviews Cancer,* 14**,** 359.

BILALOVIĆ, N., VRANIĆ, S., BAŠIĆ, H., TATAREVIĆ, A. & SELAK, I. 2005. Immunohistochemical evaluation of cyclin D1 in breast cancer. *Croatian medical journal,* 46.

BLACKWOOD, E. M. & EISENMAN, R. N. 1991. Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. *Science,* 251**,** 1211-7.

BRAY, F., FERLAY, J., SOERJOMATARAM, I., SIEGEL, R. L., TORRE, L. A. & JEMAL, A. 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*.

BRENTON, J. D., CAREY, L. A., AHMED, A. A. & CALDAS, C. 2005. Molecular classification and molecular forecasting of breast cancer: ready for clinical application? *Journal of clinical oncology,* 23**,** 7350-7360.

BRETSCHNEIDER, N., BRAND, H., MILLER, N., LOWERY, A. J., KERIN, M. J., GANNON, F. & DENGER, S. 2008. Estrogen induces repression of the breast cancer and salivary gland expression gene in an estrogen receptor alpha-dependent manner. *Cancer Res,* 68**,** 106-14.

BROWNE, G., TAIPALEENMAKI, H., BISHOP, N. M., MADASU, S. C., SHAW, L. M., VAN WIJNEN, A. J., STEIN, J. L., STEIN, G. S. & LIAN, J. B. 2015. Runx1 is associated with breast cancer progression in MMTV-PyMT transgenic mice and its depletion in vitro inhibits migration and invasion. *J Cell Physiol,* 230**,** 2522-32.

BRUECKNER, F., ORTIZ, J. & CRAMER, P. 2009. A movie of the RNA polymerase nucleotide addition cycle. *Current opinion in structural biology,* 19**,** 294-299.

CANCER GENOME ATLAS NETWORK 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature,* 487**,** 330.

CARNINCI, P. & HAYASHIZAKI, Y. 1999. [2] High-efficiency full-length cDNA cloning. *Methods in enzymology*. Elsevier.

CASNEUF, T., VAN DE PEER, Y. & HUBER, W. 2007. In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC bioinformatics,* 8**,** 461.

CHEANG, M. C., CHIA, S. K., VODUC, D., GAO, D., LEUNG, S., SNIDER, J., WATSON, M., DAVIES, S., BERNARD, P. S. & PARKER, J. S. 2009. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *JNCI: Journal of the National Cancer Institute,* 101**,** 736-750.

CHEN, E. A., SOUAIAIA, T., HERSTEIN, J. S., EVGRAFOV, O. V., SPITSYNA, V. N., REBOLINI, D. F. & KNOWLES, J. A. 2014. Effect of RNA integrity on uniquely mapped reads in RNA-Seq. *BMC research notes,* 7**,** 753.

CHEN, X., LI, X. Y., LONG, M., WANG, X., GAO, Z. W., CUI, Y., REN, J., ZHANG, Z., LIU, C., DONG, K. & ZHANG, H. 2018. The FBXW7 tumor suppressor inhibits breast cancer proliferation and promotes apoptosis by targeting MTDH for degradation. *Neoplasma,* 65**,** 201-209.

CORREIA, N. C., MELAO, A., POVOA, V., SARMENTO, L., GOMEZ DE CEDRON, M., MALUMBRES, M., ENGUITA, F. J. & BARATA, J. T. 2016. microRNAs regulate TAL1 expression in T-cell acute lymphoblastic leukemia. *Oncotarget,* 7**,** 8268-81.

COX, T. F. & COX, M. A. 2000. *Multidimensional scaling*, Chapman and hall/CRC.

DAI, X., CHENG, H., BAI, Z. & LI, J. 2017. Breast cancer cell line classification and its relevance with breast tumor subtyping. *J Journal of Cancer,* 8**,** 3131.

DANG, C. V., O'DONNELL, K. A., ZELLER, K. I., NGUYEN, T., OSTHUS, R. C. A. L. & F 2006. The c-Myc target gene network. *Seminars in cancer biology.* Academic Press.

DOUMA, S., VAN LAAR, T., ZEVENHOVEN, J., MEUWISSEN, R., VAN GARDEREN, E. & PEEPER, D. S. 2004. Suppression of anoikis and induction of metastasis by the neurotrophic receptor TrkB. *Nature,* 430**,** 1034-9.

DREOS, R., AMBROSINI, G., GROUX, R., CAVIN PÉRIER, R. & BUCHER, P. 2016. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic acids research,* 45**,** D51-D55.

DUBIK, D. & SHIU, R. P. 1992. Mechanism of estrogen activation of c-myc oncogene expression. *Oncogene,* 7**,** 1587-94.

DUBIK, D. D., T. C.; AND SHIU, R.P 1987. Stimulation of c-myc oncogene expression associated with estrogen-induced proliferation of human breast cancer cells. *Cancer research,* 47**,** 6517-6521.

DUFFY, M., HARBECK, N., NAP, M., MOLINA, R., NICOLINI, A., SENKUS, E. & CARDOSO, F. 2017. Clinical use of biomarkers in breast cancer: Updated guidelines from the European Group on Tumor Markers (EGTM). *European journal of cancer,* 75**,** 284-298.

EDGAR, R., DOMRACHEV, M. & LASH, A. E. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research,* 30**,** 207-210.

FALTAS, B. 2011. Cancer is an ancient disease: the case for better palaeoepidemiological and molecular studies. *Nature Reviews Cancer,* 11**,** 76.

FAN, W., CHANG, J. & FU, P. 2015. Endocrine therapy resistance in breast cancer: current status, possible mechanisms and overcoming strategies. *Future medicinal chemistry* 7**,** 1511-1519.

FERLA, R., CALO, V., CASCIO, S., RINALDI, G., BADALAMENTI, G., CARRECA, I., SURMACZ, E., COLUCCI, G., BAZAN, V. & RUSSO, A. 2007. Founder mutations in BRCA1 and BRCA2 genes. *Annals of Oncology,* 18**,** vi93-vi98.

FORREST, A. R., KAWAJI, H., REHLI, M., BAILLIE, J. K., DE HOON, M. J., HABERLE, V., LASSMANN, T., KULAKOVSKIY, I. V., LIZIO, M., ITOH, M., ANDERSSON, R., MUNGALL, C. J., MEEHAN, T. F., SCHMEIER, S., BERTIN, N., JORGENSEN, M., DIMONT, E., ARNER, E., SCHMIDL, C., SCHAEFER, U., MEDVEDEVA, Y. A., PLESSY, C., VITEZIC, M., SEVERIN, J., SEMPLE, C., ISHIZU, Y., YOUNG, R. S., FRANCESCATTO, M., ALAM, I., ALBANESE, D., ALTSCHULER, G. M., ARAKAWA,

T., ARCHER, J. A., ARNER, P., BABINA, M., RENNIE, S., BALWIERZ, P. J., BECKHOUSE, A. G., PRADHAN-BHATT, S., BLAKE, J. A., BLUMENTHAL, A., BODEGA, B., BONETTI, A., BRIGGS, J., BROMBACHER, F., BURROUGHS, A. M., CALIFANO, A., CANNISTRACI, C. V., CARBAJO, D., CHEN, Y., CHIERICI, M., CIANI, Y., CLEVERS, H. C., DALLA, E., DAVIS, C. A., DETMAR, M., DIEHL, A. D., DOHI, T., DRABLOS, F., EDGE, A. S., EDINGER, M., EKWALL, K., ENDOH, M., ENOMOTO, H., FAGIOLINI, M., FAIRBAIRN, L., FANG, H., FARACH-CARSON, M. C., FAULKNER, G. J., FAVOROV, A. V., FISHER, M. E., FRITH, M. C., FUJITA, R., FUKUDA, S., FURLANELLO, C., FURINO, M., FURUSAWA, J., GEIJTENBEEK, T. B., GIBSON, A. P., GINGERAS, T., GOLDOWITZ, D., GOUGH, J., GUHL, S., GULER, R., GUSTINCICH, S., HA, T. J., HAMAGUCHI, M., HARA, M., HARBERS, M., HARSHBARGER, J., HASEGAWA, A., HASEGAWA, Y., HASHIMOTO, T., HERLYN, M., HITCHENS, K. J., HO SUI, S. J., HOFMANN, O. M., HOOF, I., HORI, F., HUMINIECKI, L., et al. 2014. A promoter-level mammalian expression atlas. *Nature,* 507**,** 462-70.

FRIED, M. G. 1989. Measurement of protein-DNA interaction parameters by electrophoresis mobility shift assay. *Electrophoresis,* 10**,** 366-376.

GILLET, J.-P., VARMA, S. & GOTTESMAN, M. M. 2013. The clinical relevance of cancer cell lines. *J Journal of the National Cancer Institute,* 105**,** 452-458.

GOODRICH, J. A., HOEY, T., THUT, C. J., ADMON, A. & TJIAN, R. 1993. Drosophila TAFII40 interacts with both a VP16 activation domain and the basal transcription factor TFIIB. *Cell,* 75**,** 519-530.

GORODETSKY, A. A., EBRAHIM, A. & BARTON, J. K. 2008. Electrical detection of TATA binding protein at DNA-modified microelectrodes. *Journal of the American Chemical Society,* 130**,** 2924-2925.

GOSWAMI, C. P. & NAKSHATRI, H. 2014. PROGgeneV2: enhancements on the existing database. *BMC cancer,* 14**,** 970-970.

HANAHAN, D. & WEINBERG, R. A. 2011. Hallmarks of cancer: the next generation. *Cell,* 144**,** 646-674.

HEAP, G. A., YANG, J. H., DOWNES, K., HEALY, B. C., HUNT, K. A., BOCKETT, N., FRANKE, L., DUBOIS, P. C., MEIN, C. A. & DOBSON, R. J. 2009. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Human molecular genetics,* 19**,** 122-134.

HEIM, E., VALACH, L. & SCHAFFNER, L. 1997. Coping and psychosocial adaptation: longitudinal effects over time and stages in breast cancer. *Psychosomatic Medicine,* 59**,** 408-418.

HILL, R., MADUREIRA, P. A., FERREIRA, B., BAPTISTA, I., MACHADO, S., COLACO, L., DOS SANTOS, M., LIU, N., DOPAZO, A., UGUREL, S., ADRIENN, A., KISS-TOTH, E., ISBILEN, M., GURE, A. O. & LINK, W. 2017. TRIB2 confers resistance to anti-cancer therapy by activating the serine/threonine protein kinase AKT. *Nat Commun,* 8**,** 14687.

HO SUI, S. J., MORTIMER, J. R., ARENILLAS, D. J., BRUMM, J., WALSH, C. J., KENNEDY, B. P. & WASSERMAN, W. W. 2005. oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic acids research,* 33**,** 3154-3164.

HOGAN, C. J., ALIGIANNI, S., DURAND-DUBIEF, M., PERSSON, J., WILL, W. R., WEBSTER, J., WHEELER, L., MATHEWS, C. K., ELDERKIN, S. & OXLEY, D. 2010. Fission yeast Iec1-ino80-mediated nucleosome eviction regulates nucleotide and phosphate metabolism. *Molecular cellular biology,* 30**,** 657-674.

HONG, D., MESSIER, T. L., TYE, C. E., DOBSON, J. R., FRITZ, A. J., SIKORA, K. R., BROWNE, G., STEIN, J. L., LIAN, J. B. & STEIN, G. S. 2017. Runx1 stabilizes the mammary epithelial cell phenotype and prevents epithelial to mesenchymal transition. *Oncotarget,* 8**,** 17610-17627.

HU, Y., LIU, J. & HUANG, H. 2013. Recent agents targeting HIF-1α for cancer therapy. *Journal of cellular biochemistry,* 114**,** 498-509.

HUANG, D. W., SHERMAN, B. T. & LEMPICKI, R. A. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols,* 4**,** 44.

HUBER, W., CAREY, V. J., GENTLEMAN, R., ANDERS, S., CARLSON, M., CARVALHO, B. S., BRAVO, H. C., DAVIS, S., GATTO, L. & GIRKE, T. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods,* 12**,** 115.

ITO, Y., BAE, S. C. & CHUANG, L. S. 2015. The RUNX family: developmental regulators in cancer. *Nat Rev Cancer,* 15**,** 81-95.

JACQUILLAT, C., WEIL, M., BAILLET, F., BOREL, C., AUCLERC, G., DE MAUBLANC, M., HOUSSET, M., FORGET, G., THILL, L. & SOUBRANE, C. 1990. Results of neoadjuvant chemotherapy and radiation therapy in the breast-conserving treatment of 250 patients with all stages of infiltrative breast cancer. *Cancer,* 66**,** 119-129.

JERRY, D. J., DUNPHY, K. A. & HAGEN, M. J. 2010. Estrogens, regulation of p53 and breast cancer risk: a balancing act. *Cellular molecular life sciences,* 67**,** 1017-1023.

JONES, R. L., SALTER, J., A'HERN, R., NERURKAR, A., PARTON, M., REIS-FILHO, J. S., SMITH, I. E. & DOWSETT, M. 2009. The prognostic significance of Ki67 before and after neoadjuvant chemotherapy in breast cancer. *Breast cancer research treatment,* 116**,** 53-68.

KACZKOWSKI, B., TANAKA, Y., KAWAJI, H., SANDELIN, A., ANDERSSON, R., ITOH, M., LASSMANN, T., HAYASHIZAKI, Y., CARNINCI, P. & FORREST, A. R. 2016. Transcriptome analysis of recurrently deregulated genes across multiple cancers identifies new pan-cancer biomarkers. *Cancer research,* 76**,** 216-226.

KALASHNIKOVA, E. V., REVENKO, A. S., GEMO, A. T., ANDREWS, N. P., TEPPER, C. G., ZOU, J. X., CARDIFF, R. D., BOROWSKY, A. D. & CHEN, H. W. 2010. ANCCA/ATAD2 overexpression identifies breast cancer patients with poor prognosis, acting to drive proliferation and survival of triple-negative cells through control of B-Myb and EZH2. *Cancer Res,* 70**,** 9402-12.

KALYUGA, M., GALLEGO-ORTEGA, D., LEE, H. J., RODEN, D. L., COWLEY, M. J., CALDON, C. E., STONE, A., ALLERDICE, S. L., VALDES-MORA, F., LAUNCHBURY, R., STATHAM, A. L., ARMSTRONG, N., ALLES, M. C., YOUNG, A., EGGER, A., AU, W., PIGGIN, C. L., EVANS, C. J., LEDGER, A., BRUMMER, T., OAKES, S. R., KAPLAN, W., GEE, J. M., NICHOLSON, R. I., SUTHERLAND, R. L., SWARBRICK, A., NAYLOR, M. J., CLARK, S. J., CARROLL, J. S. & ORMANDY, C. J. 2012. ELF5 suppresses estrogen sensitivity and underpins the acquisition of antiestrogen resistance in luminal breast cancer. *PLoS Biol,* 10**,** e1001461.

KANEHISA, M. & GOTO, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research,* 28**,** 27-30.

KAPLAN, E. L. & MEIER, P. 1958. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association,* 53**,** 457-481.

KASSOUF, M. T., HUGHES, J. R., TAYLOR, S., MCGOWAN, S. J., SONEJI, S., GREEN, A. L., VYAS, P. & PORCHER, C. 2010. Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res,* 20**,** 1064-83.

KAUR, M., MACPHERSON, C. R., SCHMEIER, S., NARASIMHAN, K., CHOOLANI, M. & BAJIC, V. B. 2011. In Silico discovery of transcription factors as potential diagnostic biomarkers of ovarian cancer. *BMC systems biology,* 5**,** 144.

KHAN, A., FORNES, O., STIGLIANI, A., GHEORGHE, M., CASTRO-MONDRAGON, J. A., VAN DER LEE, R., BESSY, A., CHÈNEBY, J., KULKARNI, S. R., TAN, G., BARANASIC, D., ARENILLAS, D. J., SANDELIN, A., VANDEPOELE, K., LENHARD, B., BALLESTER, B., WASSERMAN, W. W., PARCY, F. & MATHELIER, A. 2018. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic acids research,* 46**,** D260-D266.

KIM, D., LANGMEAD, B. & SALZBERG, S. L. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature methods,* 12**,** 357.

KORNBERG, R. D. 2007. The molecular basis of eukaryotic transcription. *Proceedings of the National Academy of Sciences,* 104**,** 12955-12961.

KWON, A. T., ARENILLAS, D. J., WORSLEY HUNT, R. & WASSERMAN, W. W. 2012. oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3 (Bethesda, Md.),* 2**,** 987-1002.

LATCHMAN, D. S. 1997. Transcription factors: an overview. *The international journal of biochemistry cell biology,* 29**,** 1305-1312.

LEE, T. I. & YOUNG, R. A. 2013. Transcriptional regulation and its misregulation in disease. *Cell,* 152**,** 1237-1251.

LELLI, K. M., SLATTERY, M. & MANN, R. S. 2012. Disentangling the many layers of eukaryotic transcriptional regulation. *Annual review of genetics,* 46**,** 43-68.

LEVINSON, R. S., BATOURINA, E., CHOI, C., VORONTCHIKHINA, M., KITAJEWSKI, J. & MENDELSOHN, C. L. 2005. Foxd1-dependent signals control cellularity in the renal capsule, a structure required for normal renal development. *Development,* 132**,** 529-39.

LI, X., HUANG, J., YI, P., BAMBARA, R. A., HILF, R. & MUYAN, M. 2004. Single-chain estrogen receptors (ERs) reveal that the ERα/β heterodimer emulates functions of the ERα dimer in genomic estrogen signaling pathways. *J Molecular cellular biology* 24**,** 7681-7694.

LI, Y., JIN, K., VAN PELT, G. W., VAN DAM, H., YU, X., MESKER, W. E., TEN DIJKE, P., ZHOU, F. & ZHANG, L. 2016. c-Myb Enhances Breast Cancer Invasion and Metastasis through the Wnt/beta-Catenin/Axin2 Pathway. *Cancer Res,* 76**,** 3364-75.

LIANG, Y. X., LU, J. M., MO, R. J., HE, H. C., XIE, J., JIANG, F. N., LIN, Z. Y., CHEN, Y. R., WU, Y. D., LUO, H. W., LUO, Z. & ZHONG, W. D. 2016. E2F1 promotes tumor cell invasion and migration through regulating CD147 in prostate cancer. *Int J Oncol,* 48**,** 1650-8.

LITTLEPAGE, L. E., ADLER, A. S., KOUROS-MEHR, H., HUANG, G., CHOU, J., KRIG, S. R., GRIFFITH, O. L., KORKOLA, J. E., QU, K. & LAWSON, D. A. 2012. The transcription factor ZNF217 is a prognostic biomarker and therapeutic target during breast cancer progression. *Cancer discovery*.

LIZIO, M., HARSHBARGER, J., SHIMOJI, H., SEVERIN, J., KASUKAWA, T., SAHIN, S., ABUGESSAISA, I., FUKUDA, S., HORI, F. & ISHIKAWA-KATO, S. 2015. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome biology,* 16**,** 22.

LOPEZ, F., BELLOC, F., LACOMBE, F., DUMAIN, P., REIFFERS, J., BERNARD, P. & BOISSEAU, M. 1991. Modalities of synthesis of Ki67 antigen during the stimulation of lymphocytes. *Cytometry: The Journal of the International Society for Analytical Cytology,* 12**,** 42-49.

LOUIE, M. C., ZOU, J. X., RABINOVICH, A. A. C. & H.W 2004. ACTR/AIB1 functions as an E2F1 coactivator to promote breast cancer cell proliferation and antiestrogen resistance. *Molecular and cellular biology,* 24**,** 5157-5171.

LUN, A. T., CHEN, Y. & SMYTH, G. K. 2016. It's DE-licious: a recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. *Statistical Genomics.* Springer.

MANSOUR, M. R., SANDA, T., LAWTON, L. N., LI, X., KRESLAVSKY, T., NOVINA, C. D., BRAND, M., GUTIERREZ, A., KELLIHER, M. A., JAMIESON, C. H., VON BOEHMER, H., YOUNG, R. A. & LOOK, A. T. 2013. The TAL1 complex targets the FBXW7 tumor suppressor by activating miR-223 in human T cell acute lymphoblastic leukemia. *J Exp Med,* 210**,** 1545-57.

MAROTTI, J. D., MULLER, K. E., TAFE, L. J., DEMIDENKO, E. & MILLER, T. W. 2017. P-Rex1 Expression in Invasive Breast Cancer in relation to Receptor Status and Distant Metastatic Site. *International journal of breast cancer,* 2017**,** 4537532-4537532.

MARTIN, M. G., WELCH, J. S., LUO, J., ELLIS, M. J., GRAUBERT, T. A. & WALTER, M. J. 2009. Therapy related acute myeloid leukemia in breast cancer survivors, a population-based study. *Breast Cancer Res Treat,* 118**,** 593-8.

MAYEUX, R. 2004. Biomarkers: potential uses and limitations. *NeuroRx,* 1**,** 182-188.

MCMANUS, S., EBERT, A., SALVAGIOTTO, G., MEDVEDOVIC, J., SUN, Q., TAMIR, I., JARITZ, M., TAGOH, H. & BUSSLINGER, M. 2011. The transcription factor Pax5 regulates its target genes by recruiting chromatin-modifying proteins in committed B cells. *The EMBO journal,* 30**,** 2388-2404.

MENESES, A. M., MEDINA, R. A., KATO, S., PINTO, M., JAQUE, M. P., LIZAMA, I., GARCIA MDE, L., NUALART, F. & OWEN, G. I. 2008. Regulation of GLUT3 and glucose uptake by the cAMP signalling pathway in the breast cancer cell line ZR-75. *J Cell Physiol,* 214**,** 110-6.

MESSIER, T. L., GORDON, J. A., BOYD, J. R., TYE, C. E., BROWNE, G., STEIN, J. L., LIAN, J. B. & STEIN, G. S. 2016. Histone H3 lysine 4 acetylation and methylation dynamics define breast cancer subtypes. *Oncotarget,* 7**,** 5094.

MICALIZZI, D. S. & FORD, H. L. 2009. Epithelial-mesenchymal transition in development and cancer. *Future Oncol,* 5**,** 1129-43.

MONK, M. & HOLDING, C. 2001. Human embryonic genes re-expressed in cancer cells. *Oncogene,* 20**,** 8085-91.

MORAN, M. S., SCHNITT, S. J., GIULIANO, A. E., HARRIS, J. R., KHAN, S. A., HORTON, J., KLIMBERG, S., CHAVEZ-MACGREGOR, M., FREEDMAN, G. & HOUSSAMI, N. 2014. Society of Surgical Oncology–American Society for Radiation Oncology consensus guideline on margins for breast-conserving surgery with whole-breast irradiation in stages I and II invasive breast cancer. *Annals of surgical oncology,* 21**,** 704-716.

MORETTIN, A., BALDWIN, R. M. & CÔTÉ, J. 2015. Arginine methyltransferases as novel therapeutic targets for breast cancer. *Mutagenesis,* 30**,** 177-189.

NAGATA, Y., LAN, K.-H., ZHOU, X., TAN, M., ESTEVA, F. J., SAHIN, A. A., KLOS, K. S., LI, P., MONIA, B. P. & NGUYEN, N. T. 2004. PTEN activation contributes to tumor inhibition by trastuzumab, and loss of PTEN predicts trastuzumab resistance in patients. *Cancer cell,* 6**,** 117-127.

NEUMAN, H. B., MORROGH, M., GONEN, M., VAN ZEE, K. J., MORROW, M. & KING, T. A. 2010. Stage IV breast cancer in the era of targeted therapy: does surgery of the primary tumor matter? *Cancer: Interdisciplinary International Journal of the American Cancer Society* 116**,** 1226-1233.

NOYES, M. B., CHRISTENSEN, R. G., WAKABAYASHI, A., STORMO, G. D., BRODSKY, M. H. & WOLFE, S. A. 2008. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell,* 133**,** 1277-1289.

O'BRIEN, K. M., COLE, S. R., TSE, C.-K., PEROU, C. M., CAREY, L. A., FOULKES, W. D., DRESSLER, L. G., GERADTS, J. & MILLIKAN, R. C. 2010. Intrinsic breast tumor subtypes, race, and long-term survival in the Carolina Breast Cancer Study. *Clinical Cancer Research,* 16**,** 6100-6110.

O'NEIL, J., SHANK, J., CUSSON, N., MURRE, C. & KELLIHER, M. 2004. TAL1/SCL induces leukemia by inhibiting the transcriptional activity of E47/HEB. *Cancer Cell,* 5**,** 587-96.

ODOM, D. T., ZIZLSPERGER, N., GORDON, D. B., BELL, G. W., RINALDI, N. J., MURRAY, H. L., VOLKERT, T. L., SCHREIBER, J., ROLFE, P. A. & GIFFORD, D. K. 2004. Control of pancreas and liver gene expression by HNF transcription factors. *Science,* 303**,** 1378-1381.

OKUDA, T., NISHIMURA, M., NAKAO, M. & FUJITA, Y. 2001. RUNX1/AML1: a central player in hematopoiesis. *Int J Hematol,* 74**,** 252-7.

ORDONEZ, N. G. 2012. Value of thyroid transcription factor-1 immunostaining in tumor diagnosis: a review and update. *Applied Immunohistochemistry Molecular Morphologyn*20**,** 429-444.

PALAZZO, A. F. & LEE, E. S. 2015. Non-coding RNA: what is functional and what is junk? *Frontiers in genetics,* 6**,** 2.

PAN, Q., SHAI, O., LEE, L. J., FREY, B. J. & BLENCOWE, B. J. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics,* 40**,** 1413.

PAPA, V., PEZZINO, V., COSTANTINO, A., BELFIORE, A., GIUFFRIDA, D., FRITTITTA, L., VANNELLI, G. B., BRAND, R., GOLDFINE, I. D. & VIGNERI, R. 1990. Elevated insulin receptor content in human breast cancer. *The Journal of clinical investigation,* 86**,** 1503-1510.

PARIKH, R., MATHAI, A., PARIKH, S., SEKHAR, G. C. & THOMAS, R. 2008. Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology,* 56**,** 45.

PEDERSEN, N., PEDERSEN, M. W., LAN, M. S., BRESLIN, M. B. & POULSEN, H. S. 2006. The insulinoma-associated 1: a novel promoter for targeted cancer gene therapy for small-cell lung cancer. *Cancer Gene Ther,* 13**,** 375-84.

PEPKE, S., WOLD, B. & MORTAZAVI, A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nature methods,* 6**,** S22.

PEROU, C. M., SØRLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., REES, C. A., POLLACK, J. R., ROSS, D. T., JOHNSEN, H. & AKSLEN, L. A. 2000. Molecular portraits of human breast tumours. *Nature,* 406**,** 747.

PIPINIKAS, C. P., NAIR, S. B., KIRBY, R. S., CARTER, N. D. & FENSKE, C. D. 2007. Measurement of blood E2F3 mRNA in prostate cancer by quantitative RT-PCR: a preliminary study. *Biomarkers,* 12**,** 541-557.

PLUN-FAVREAU, H., LEWIS, P. A., HARDY, J., MARTINS, L. M. & WOOD, N. W. 2010. Cancer and neurodegeneration: between the devil and the deep blue sea. *PLoS genetics,* 6**,** e1001257-e1001257.

RAKHA, E., PUTTI, T., EL-REHIM, D. A., PAISH, C., GREEN, A., POWE, D., LEE, A., ROBERTSON, J. & ELLIS, I. 2006. Morphological and immunophenotypic analysis of breast carcinomas with basal and myoepithelial differentiation. *The Journal of pathology,* 208**,** 495-506.

RAMSAY, D., KENT, J., HARTMANN, R. & HARTMANN, P. 2005. Anatomy of the lactating human breast redefined with ultrasound imaging. *Journal of anatomy,* 206**,** 525-534.

REIMER, D., SADR, S., WIEDEMAIR, A., GOEBEL, G., CONCIN, N., HOFSTETTER, G., MARTH, C. & ZEIMET, A. G. 2006. Expression of the E2F family of transcription factors and its clinical relevance in ovarian cancer. *Annals of the New York Academy of Sciences,* 1091**,** 270-281.

RHODES, D. R., YU, J., SHANKER, K., DESHPANDE, N., VARAMBALLY, R., GHOSH, D., BARRETTE, T., PANDEY, A. & CHINNAIYAN, A. M. 2004. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia (New York, N.Y.),* 6**,** 1-6.

RICHARD, P. & MANLEY, J. L. 2009. Transcription termination by nuclear RNA polymerases. *Genes development,* 23**,** 1247-1269.

RIZZO, P., MIAO, H., D'SOUZA, G., OSIPO, C., YUN, J., ZHAO, H., MASCARENHAS, J., WYATT, D., ANTICO, G. & HAO, L. 2008. Cross-talk between notch and the estrogen receptor in breast cancer suggests novel therapeutic approaches. *Cancer research,* 68**,** 5226-5235.

ROBINSON, M. D., MCCARTHY, D. J. & SMYTH, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26**,** 139-140.

ROBINSON, M. D. & OSHLACK, A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology,* 11**,** R25.

ROBLES, A. I. & HARRIS, C. C. 2010. Clinical outcomes and correlates of TP53 mutations and cancer. *Cold Spring Harbor perspectives in biology,* 2**,** a001016.

ROSENBAUM, J. N., GUO, Z., BAUS, R. M., WERNER, H., REHRAUER, W. M. & LLOYD, R. V. 2015. INSM1: A Novel Immunohistochemical and Molecular Marker for Neuroendocrine and Neuroepithelial Neoplasms. *Am J Clin Pathol,* 144**,** 579-91.

SANDA, T. & LEONG, W. Z. 2017. TAL1 as a master oncogenic transcription factor in T-cell acute lymphoblastic leukemia. *Exp Hematol,* 53**,** 7-15.

SCHNEIDER, T. D. 2002. Consensus sequence Zen. *Applied bioinformatics,* 1**,** 111-119.

SEGAL, R., EVANS, W., JOHNSON, D., SMITH, J., COLLETTA, S., GAYTON, J., WOODARD, S., WELLS, G. & REID, R. 2001. Structured exercise improves physical functioning in women with stages I and II breast cancer: results of a randomized controlled trial. *Journal of clinical oncology,* 19**,** 657-665.

SEMENZA, G. L. 1994. Transcriptional regulation of gene expression: mechanisms and pathophysiology. *Human mutation,* 3**,** 180-199.

SHAMIR, E. R., PAPPALARDO, E., JORGENS, D. M., COUTINHO, K., TSAI, W. T., AZIZ, K., AUER, M., TRAN, P. T., BADER, J. S. & EWALD, A. J. 2014. Twist1-induced dissemination preserves epithelial identity and requires E-cadherin. *J Cell Biol,* 204**,** 839-56.

SHANG, Y. 2007. Hormones and cancer. *Cell research,* 17**,** 277.

SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N. S., WANG, J. T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B. & IDEKER, T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research,* 13**,** 2498-2504.

SIEPEL, A., BEJERANO, G., PEDERSEN, J. S., HINRICHS, A. S., HOU, M., ROSENBLOOM, K., CLAWSON, H., SPIETH, J., HILLIER, L. W., RICHARDS, S., WEINSTOCK, G. M., WILSON, R. K., GIBBS, R. A., KENT, W. J., MILLER, W. & HAUSSLER, D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research,* 15**,** 1034-1050.

SILWAL-PANDIT, L., VOLLAN, H. K. M., CHIN, S.-F., RUEDA, O. M., MCKINNEY, S., OSAKO, T., QUIGLEY, D. A., KRISTENSEN, V. N., APARICIO, S. & BØRRESEN-DALE, A.-L. 2014. TP53 mutation spectrum in breast cancer is subtype specific and has distinct prognostic relevance. *Clinical Cancer Research*.

SLATTERY, M., RILEY, T., LIU, P., ABE, N., GOMEZ-ALCALA, P., DROR, I., ZHOU, T., ROHS, R., HONIG, B. & BUSSEMAKER, H. J. 2011. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell,* 147**,** 1270-1282.

SMITH, M. L. & SEO, Y. R. 2000. Sensitivity of cyclin E-overexpressing cells to cisplatin/taxol combinations. *Anticancer research,* 20**,** 2537-2539.

SOLOMON, M. J., LARSEN, P. L. & VARSHAVSKY, A. 1988. Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. *Cell,* 53**,** 937-947.

SORLIE, T., PEROU, C. M., TIBSHIRANI, R., AAS, T., GEISLER, S., JOHNSEN, H., HASTIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., THORSEN, T., QUIST, H., MATESE, J. C., BROWN, P. O., BOTSTEIN, D., LONNING, P. E. & BORRESEN-DALE, A. L. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A,* 98**,** 10869-74.

SOTIRIOU, C., NEO, S.-Y., MCSHANE, L. M., KORN, E. L., LONG, P. M., JAZAERI, A., MARTIAT, P., FOX, S. B., HARRIS, A. L. & LIU, E. T. 2003. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences,* 100**,** 10393-10398.

SPITZ, F. & FURLONG, E. E. 2012. Transcription factors: from enhancer binding to developmental control. *Nature reviews genetics,* 13**,** 613.

STENDER, J. D., KIM, K., CHARN, T. H., KOMM, B., CHANG, K. C., KRAUS, W. L., BENNER, C., GLASS, C. K. & KATZENELLENBOGEN, B. S. 2010. Genome-wide analysis of estrogen receptor α DNA binding and tethering mechanisms identifies Runx1 as a novel tethering factor in receptor-mediated transcriptional activation. *Molecular cellular biology,* 30**,** 3943-3955.

STERNLICHT, M. D. 2006. Key stages in mammary gland development: the cues that regulate ductal branching morphogenesis. *Breast Cancer Res,* 8**,** 201.

STORMO, G. D. 2000. DNA binding sites: representation and discovery. *Bioinformatics,* 16**,** 16-23.

STRINGER, K. F., INGLES, C. J. & GREENBLATT, J. 1990. Direct and selective binding of an acidic transcriptional activation domain to the TATA-box factor TFIID. *Nature,* 345**,** 783.

SULTAN, M., SCHULZ, M. H., RICHARD, H., MAGEN, A., KLINGENHOFF, A., SCHERF, M., SEIFERT, M., BORODINA, T., SOLDATOV, A. & PARKHOMCHUK, D. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science,* 321**,** 956-960.

SWEDENBORG, E., POWER, K. A., CAI, W., PONGRATZ, I. & RÜEGG, J. 2009. Regulation of estrogen receptor beta activity and implications in health and disease. *Cellular Molecular Life Sciences,* 66**,** 3873-3894.

SZABO, C. I. & KING, M. C. 1995. Inherited breast and ovarian cancer. *Human molecular genetics,* 4**,** 1811-1817.

TAKAHASHI, H., KOBAYASHI, H., HASHIMOTO, Y., MATSUO, S. & IIZUKA, H. 1995. Interferon-γ-dependent stimulation of Fas antigen in SV40-transformed human keratinocytes: modulation of the apoptotic process by protein kinase C. *Journal of investigative dermatology,* 105.

TAN, W., LI, Q., CHEN, K., SU, F., SONG, E. & GONG, C. 2016. Estrogen receptor beta as a prognostic factor in breast cancer patients: a systematic review and meta-analysis. *Oncotarget,* 7**,** 10373.

TRAPNELL, C., WILLIAMS, B. A., PERTEA, G., MORTAZAVI, A., KWAN, G., VAN BAREN, M. J., SALZBERG, S. L., WOLD, B. J. & PACHTER, L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology,* 28**,** 511.

VAN BRAGT, M. P., HU, X., XIE, Y. & LI, Z. 2014. RUNX1, a transcription factor mutated in breast cancer, controls the fate of ER-positive mammary luminal cells. *Elife,* 3**,** e03881.

VITA, M. & HENRIKSSON, M. 2006. The Myc oncoprotein as a therapeutic target for human cancer. *Semin Cancer Biol,* 16**,** 318-30.

VOGELSTEIN, B., LANE, D. & LEVINE, A. J. 2000. Surfing the p53 network. *Nature,* 408**,** 307.

VUAROQUEAUX, V., URBAN, P., LABUHN, M., DELORENZI, M., WIRAPATI, P., BENZ, C. C., FLURY, R., DIETERICH, H., SPYRATOS, F., EPPENBERGER, U. & EPPENBERGER-CASTORI, S. 2007. Low E2F1 transcript levels are a strong determinant of favorable breast cancer outcome. *Breast Cancer Res,* 9**,** R33.

WANG, Y., ALLA, V., GOODY, D., GUPTA, S. K., SPITSCHAK, A., WOLKENHAUER, O., PUTZER, B. M. & ENGELMANN, D. 2016. Epigenetic factor EPC1 is a master regulator of DNA damage response by interacting with E2F1 to silence death and activate metastasis-related gene signatures. *Nucleic Acids Res,* 44**,** 117-33.

WANG, Z., GERSTEIN, M. & SNYDER, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics,* 10**,** 57.

WEDDERBURN, R. W. 1974. Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika,* 61**,** 439-447.

WEIGEL, M. T. & DOWSETT, M. 2010. Current and emerging biomarkers in breast cancer: prognosis and prediction. *Endocrine-related cancer,* 17**,** R245-R262.

WRAY, G. A., HAHN, M. W., ABOUHEIF, E., BALHOFF, J. P., PIZER, M., ROCKMAN, M. V. & ROMANO, L. A. 2003. The evolution of transcriptional regulation in eukaryotes. *Molecular biology evolution,* 20**,** 1377-1419.

WU, X. & LEVINE, A. J. 1994. p53 and E2F-1 cooperate to mediate apoptosis. *Proc Natl Acad Sci U S A,* 91**,** 3602-6.

XIE, J., CAI, T., ZHANG, H., LAN, M. S. & NOTKINS, A. L. 2002. The zinc-finger transcription factor INSM1 is expressed during embryo development and interacts with the Cbl-associated protein. *Genomics,* 80**,** 54-61.

XU, J., CHEN, Y. A. O. & O.I 2010. MYC and breast cancer. *Genes &,* 1**,** 629-640.

YEH, J. E., TONIOLO, P. A. & FRANK, D. A. 2013. Targeting transcription factors: promising new strategies for cancer therapy. *Current opinion in oncology,* 25**,** 652-658.

YOON, S. & NAM, D. 2017. Gene dispersion is the key determinant of the read count bias in differential expression analysis of RNA-seq data. *BMC genomics,* 18**,** 408-408.

YU, C.-P., LIN, J.-J. & LI, W.-H. 2016. Positional distribution of transcription factor binding sites in Arabidopsis thaliana. *Scientific reports,* 6**,** 25164.

ZHANG, L. & HAN, J. 2017. Branched-chain amino acid transaminase 1 (BCAT1) promotes the growth of breast cancer cells through improving mTOR-mediated mitochondrial biogenesis and function. *Biochem Biophys Res Commun,* 486**,** 224-231.

ZHANG, S. Y., LIU, S. C., AL-SALEEM, L. F., HOLLORAN, D., BABB, J., GUO, X. A. K.-S. & A.J 2000. E2F-1: a proliferative marker of breast neoplasia. *Cancer Epidemiology and Prevention Biomarkers,* 9**,** 395-401.

ZHANG, Y., EADES, G., YAO, Y., LI, Q. & ZHOU, Q. 2012. Estrogen receptor alpha signaling regulates breast tumor-initiating cells by down-regulating miR-140 which targets the transcription factor SOX2. *J Biol Chem,* 287**,** 41514-22.

ZHAO, Y. F., ZHAO, J. Y., YUE, H., HU, K. S., SHEN, H., GUO, Z. G. & SU, X. J. 2015. FOXD1 promotes breast cancer proliferation and chemotherapeutic drug resistance by targeting p27. *Biochem Biophys Res Commun,* 456**,** 232-7.