

Gene expression analysis in breast cancer

Jai Prakash Mehta

Ph.D. Thesis 2009

A thesis submitted for the degree of Ph.D.

Dublin City University

By

Jai Prakash Mehta, M.Sc.

The research work described in thesis was performed

under the supervision of

Dr. Padraig Doolan

and

Prof. Martin Clynes

National Institute for Cellular Biotechnology

Dublin City University

2009

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work. Signed: _____

(Candidate) ID No.: _____ Date: _____

Acknowledgement

I bow my head with great reverence to Him who is omnipresent, omnipotent and omniscient and the cause behind every effort.

I would like to express my indebtedness and veneration to Prof. Martin Clynes for giving me an opportunity to work in the exciting area of cancer genomics and providing his meticulous guidance throughout my research. I express my gratitude to Dr. Padraig Doolan for his supervision, advice and guidance for this research. I also wish to record my gratitude to Dr. Niall Barron for perspicacity and inculcating the scientific temper. I would also like to thank Dr. Lorraine O'Driscoll for her advice and indelible inspiration.

My special thanks to Dr Sweta Rani, Dr Mohan Muniyappa, Mr Kishore Katikireddy, Ms Irene Oglesby and Dr. Laura Breen for teaching me and helping me through various lab techniques.

I am grateful to Mr. Joseph Carey for preparing media and autoclaving all lab materials. Thanks to Dr. Verena Murphy, Alex Eustace and Brigid Browne for providing me with cell lines for my study. I take this opportunity to extend my thanks to Ms. Carol McNamara, Ms. Yvonne Reilly and Ms. Mairead Callan for helping me in day to day work.

Words would never be able to fathom the depth of feelings for my reverend parents, Dr. R.C.Mehta and Mrs Pratibha Mehta; my wife Dr Sweta Rani and my daughter Nehal, an indelible inspiration, and affection during my research.

I am thankful to all my friends for making my day to day life enjoyable during the course of my research.

Finally, I would also like to thank everyone in NICB.

Abbreviations

BSA	Bovine Serum Albumin
cDNA	Complementary DNA
DE	Differentially expressed
DMEM	Dulbecco's Minimal Essential Medium
DMSO	Dimethyl Sulphoxide
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleotide Triphosphate (N = A, C, T, G)
EDTA	Ethylene Diamine Tetraacetic Acid
FC	Fold change
FCS	Fetal Calf Serum
GAPDH	Glyceraldehyde-6-Phosphate Dehydrogenase
IMS	Industrial Methylated Spirits
kDA	Kilo Daltons
mRNA	messenger RNA
NS	Non significant
OD	Optical Density
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PLIER	Probe Logarithmic Intensity ERror Estimation
qRT-PCR	Quantitative real-time PCR
Rcf	Rotational Speed at the Relative Centrifugal Force
RNA	Ribonucleic Acid
RNase	Ribonuclease
RNasin	Ribonuclease Inhibitor
RPM	Revolutions Per Minute
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
TBE	Tris-Boric Acid EDTA Buffer
TBS	Tris Buffered Saline
TV	Trypsin Versene
UHP	Ultra High Pure Water

ABSTRACT.....	6
1.0 INTRODUCTION.....	8
1.1 Breast cancer.....	9
1.2 Sub-types in Breast cancer.....	9
1.2.1 Estrogen receptor (ER)	9
1.2.2 Progesterone Receptor (PR).....	11
1.2.3 HER2/neu.....	12
1.2.4 Triple Negative (TN) and basal type of breast cancer	13
1.3 Prognostic markers in breast cancer	13
1.3.1 Grade.....	13
1.3.2 Lymph node metastasis.....	14
1.4 Invasion and Metastasis	15
1.4.1 Microenvironment of Breast Cancer.....	16
1.4.2 Angiogenesis.....	18
1.4.3 Invasion.....	18
1.4.4 Transport	19
1.4.5 Arrest.....	19
1.4.6 Extravasation.....	19
1.4.7 Breast cancer metastasis-associated genes.....	20
1.5 Gene expression profiling.....	21
1.5.1 Affymetrix microarrays	22
1.5.2 Microarrays and Breast cancer.....	23
1.5.3 Gene expression in diagnostics.....	24
1.6 Microarray Data Analysis	27
1.6.1 Normalization	27
1.6.2 Clustering.....	27
1.6.3 Principal component analysis	32
1.7 Classification.....	33
1.8 Representative nature of cell line models to clinical conditions.....	35
1.9 Small interfering RNA (siRNA).....	37
1.9.1 Mechanism of action of siRNA	37
1.10 Basal cell carcinoma	40
1.11 Aims.....	42
2.0 MATERIALS AND METHODS	43
2.1 Microarray data used in this study.....	44
2.1.1 Breast cancer clinical microarray dataset generated at NICB	44
2.1.2 Data obtained from public repositories.....	45
2.1.3 Basal cell carcinoma cancer clinical microarray dataset generated at NICB ..	47
2.2.1 Normalization and Quantification.....	47
2.2.2 Quality inspection	47
2.2.3 Standard deviation filtration	50
2.2.4 Hierarchical Clustering.....	50
2.2.5 Finding significant genes	51
2.2.6 Identifier conversion	52

2.2.7 Gene list comparison.....	53
2.2.8 GenMAPP	53
2.2.9 Genesis	55
2.2.10 Dev C++.....	56
2.2.11 C# (C-sharp).....	56
2.2.12 Kaplan-Meier survival function.....	56
2.2.13 CLUSTALW	57
2.2.14 BLAST.....	57
2.3 Cell Culture Methods.....	57
2.3.1 Water.....	57
2.3.2 Treatment of Glassware	58
2.3.3 Sterilisation	58
2.3.4 Media Preparation.....	58
2.4 Maintenance of cell lines	59
2.4.1 Safety Precautions.....	59
2.4.2 Culture of Adherent Cell Lines.....	59
2.4.3 Cell Counting.....	60
2.4.4 Cell freezing.....	60
2.4.5 Cell Thawing.....	61
2.4.6 Sterility Checks.....	61
2.4.7 Mycoplasma Analysis.....	62
2.5 Analytical Techniques	63
2.5.1 Preparation of total RNA from cells using RNeasy Mini Prep Kit.....	63
2.5.2 RNA Quantification using NanoDrop	63
2.5.3 RNA amplification, labelling and fragmentation of cRNA in preparation for hybridisation to Affymetrix array chips.....	65
2.5.4 Reverse Transcription of RNA from cells (cDNA Synthesis).....	71
2.5.5 Quantitative real time RT-PCR (qRT-PCR).....	72
2.5.6 Large scale plasmid preparation	74
2.5.7 Plasmid transfection protocol	76
2.5.8 RNA interference (RNAi).....	76
2.5.9 Western Blot analysis	77
2.5.10 Invasion assay	82
2.5.11 Motility assay.....	83
2.5.12 Microsoft PowerPoint	83
3.0 RESULTS	84
3.1 Breast Cancer clinical specimens.....	85
3.1.1 Data Normalization and Quantification	85
3.1.2 Data Filtration	86
3.1.3 Hierarchical Clustering	86
3.1.3.1 Two-way Hierarchical clustering.....	98
3.1.3.1.1 Up-regulated Genes	98
3.1.3.1.2 Down-regulated Genes.....	100
3.1.4 Comparison criteria: Normal vs. cancer specimens.....	104
3.1.5 Comparison criteria: Estrogen receptor-negative vs. Estrogen receptor-positive	112

3.1.6 Comparison criteria: Lymph node-negative vs. Lymph node-positive	118
3.1.7 Comparison criteria: Grade 1 vs. Grade 2	122
3.1.8 Comparison criteria: Grade 2 vs. Grade 3	127
3.1.9 Comparison criteria: Tumour Size < 2.8cm vs. > 2.8 cm	135
3.1.10 Comparison criteria: Patients who did not relapse vs. patients who did relapse (Overall relapse).....	140
3.1.11 Comparison criteria: Patients who survived vs. patients who did not survive	147
3.1.12 Comparison criteria: Patients who did not relapse within 5 years vs. patients who did relapse within 5 years.....	152
3.1.13 Comparison criteria: Patients who survived for 5 years vs. patients who did not survive for 5 years.....	158
3.1.14 Comparing gene lists to identify bad prognosis genes.....	164
3.1.15 Non-parametric analysis	168
3.1.16 Summary	171
3.2 Comparing in-house gene lists with publicly available datasets	172
3.2.1 Comparison with public datasets on the Affymetrix GeneChip platform	172
3.2.2 Comparison in-house relapse and lymph node gene list with OncotypeDx genes	175
3.2.3 Comparison in-house relapse and lymph node gene list with MammaPrint genes	178
3.2.4 Summary	182
3.3 Meta analysis of Estrogen receptor pathway genes using gene expression data. .	184
3.3.1 Up-regulated gene transcripts	185
3.3.2 Down-regulated transcripts.....	190
3.3.3 Genes correlated with ESR1	195
3.3.4 Correlation patterns among genes.....	196
3.4.5 Hierarchical clustering, Principal component analysis and k-means analysis on ESR1 correlated genes.	202
3.4.6 Summary	205
3.4 Gene expression signature for HER2.....	206
3.4.1 HER2-positive vs. HER2-negative	206
3.4.2 Summary	208
3.5 Development of MLPERCEP, a software tool for predicting relapse in breast cancer	209
3.5.1 Design	209
3.5.2 Architecture.....	209
3.5.2.1 Algorithm.....	210
3.5.3 Software Modules	212
3.5.4 Results.....	224
3.5.5 Summary	228
3.6 Functional analysis on ROPN1B	229
3.6.1 Similarity and difference among ROPN1B and ROPN1	229
3.6.2 Expression of ROPN1 and ROPN1B in normal and cancerous breast tissue	240
3.6.3 siRNA knockdown of ROPN1 and ROPN1B in melanoma cell lines	249
3.6.4 ROPN1 and ROPN1B cDNA over-expression studies.....	262

3.6.5 Summary	272
3.7 How Representative are Cell line models of clinical conditions?	273
3.7.1 Data filtration	273
3.7.2 Clustering.....	273
3.7.3 Significant genes.....	275
3.7.4 Gene ontology and pathway analysis.....	275
3.7.5 Estrogen receptor analysis	277
3.7.6 Summary	278
3.8 Molecular profile of basal cell carcinoma	280
3.8.1 Data Normalization and Quantification	280
3.8.2 Data Filtration	280
3.8.3 Hierarchical Clustering.....	280
3.8.4 Normal specimens vs. Basal cell carcinoma.....	281
3.8.5 Summary	287
4.0 DISCUSSION.....	288
4.1 Clinical heterogeneity in breast cancer	289
4.1.1 High level of correlation between Normal samples (Cluster A).....	290
4.1.2 Samples closest in character to Normal samples enriched for ER- & Grade I	291
291	
4.1.3 ER-negative samples (Cluster C) display three distinct sub-clusters	292
4.1.4 ER-positive tumors sub-divide as two groups.	293
4.2 Gene expression differences between Normal and Cancer tissue	293
4.2.1 Cell cycle pathway up-regulated in tumors	294
4.2.2 Embryonic stem cell pathway up-regulated in cancer.....	294
4.2.3 Fatty acid biosynthesis pathway down-regulated in cancer.....	296
4.3 Genes up-regulated in Estrogen Receptor-positive breast patients.....	297
4.3.1 In-house study.....	297
4.3.2 Meta analysis	298
4.4 Gene interaction network for ESR1 gene	304
4.5 Genes up-regulated in ER-negative breast patients	306
4.6 Genes up-regulated in HER2-positive breast cancers.....	307
4.7 Lymph node-negative vs. Lymph node-positive	308
4.8 Tumour Grade.....	309
4.9 Tumour size	311
4.10 Genes associated with relapse and survival.....	312
4.10.1 In-house study.....	312
4.10.2 Meta-analysis	313
4.10.3 Comparison our in-house result with OncotypeDx	315
4.10.4 Comparison our in-house result with MammaPrint.....	316
4.11 Relapse prediction.....	316
4.12 Identification and functional validation of Ropporin.....	319
4.12.1 Affymetrix probe annotation for Ropporin.....	320
4.12.2 Ropporin expression in our in-house breast dataset.	322
4.12.3 Confirmation of Ropporin expression by qPCR.....	322
4.12.4 Functional validation using <i>in-vitro</i> cell line models	323
4.12.5 Ropporin expression in cancers and normal tissues	324

4.12.6 Previous studies identifying Ropporin expression in breast cancer.....	325
4.13 Conclusion	326
4.13 Discussion of some peripheral research projects	327
4.13.1 How Representative are cell line models of clinical conditions?	327
4.13.2 Basal cell carcinoma	329
5.0 SUMMARY AND CONCLUSIONS	333
5.1 Hierarchical clustering analysis identified clinical heterogeneity in breast cancer	334
5.2 Association of clinical parameters with genes, functions and pathways	334
5.3 Comparing our in-house genelists with publicly available datasets	335
5.4 Meta analysis for estrogen receptor pathway genes using gene expression data .	336
5.5 Development of MLPERCEP, a software tool for predicting relapse in breast cancer	337
5.6 Functional analysis on Ropporin	339
6.0 FUTURE WORK	341
6.1 Validation of novel groups of specimens in independent studies.....	342
6.2 Diagnostic models.....	342
6.3 Validation of gene interaction network.....	342
6.4 Ropporin as biomarker and targeted therapy	343
REFERENCES.....	344

Abstract

Gene expression analysis in breast cancer

Jai Prakash Mehta

Breast cancer is the most common type of cancer among females, both in incidence and death. As meaningful biological understanding of the disease is confounded by the existence of various molecular groups and sub-groups, the challenge for targeted drug development may lie in understanding the molecular mechanisms of various sub-groups in breast cancer.

An in-house breast cancer gene expression dataset comprising 17 normal and 104 tumour samples was analysed to identify important genes and pathways relevant to various clinical parameters. Our results identified groups of patients with similar expression profiles, the possible biology driving them and the clinical implications. Comparing Normal and Cancer specimens' gene expression profiles, TP53, along with cell cycle genes, were up-regulated in cancer samples. Embryonic stem cell pathway genes were up-regulated, while fatty acid biosynthesis pathways were down-regulated in tumors vs normal.

The cancer specimens largely clustered with respect to ER status. Meta-analysis was performed on in-house datasets along with five public datasets to identify ER pathway genes. The analysis identified novel genes which had not been previously associated with ER-related pathways in cancer. Nuclear receptor pathways were up-regulated in ER-positive tumors/cell lines. Mining for ESR1-correlated genes across 5897 specimens identified FOXA1, SPDEF, C1ORF34 and GATA3 expression to be highly correlated.

Three sub-clusters were identified among the ER-negative cluster. One represented ERBB2 over-expressing cluster. Additionally two unique groups of patients, with significant differences in survival, previously un-identified by other studies, were identified among the ER-negative cluster; a good prognosis cluster with high expression of Immune response genes; and a bad prognosis cluster with high expression of Ropporin, over-expression of which was also linked to high incidence of relapse in our study. siRNA knockdown of Ropporin (ROPN1 and ROPN1B) in the M14 melanoma cell line impaired cancer cell motility and invasion. Knockdown of ROPN1B in MDA-MB-435s reduced motility. In the first study of its kind our results validated the role of Ropporin in cancer cell motility and invasion.

A list of 162 relapse-associated prognostically-important genes was used to develop a Neural Network back propagation model to predict the clinical outcomes. The model was successful in predicting relapse with 97.8% accuracy and outperformed existing models, indicating a strong possibility of its use as diagnostic model.

1.0 Introduction

1.1 Breast cancer

In Ireland an average of 2368 new cases of malignant breast cancer are diagnosed in females and 21 in males each year. In females, breast cancer is the most frequent cancer after skin cancer. Females are estimated to have a 1-in-13 chance of developing cancer of the breast by the age of 74 and 969 deaths among females and 6 deaths among males are attributed to breast cancer each year, on average. For every five incidences of the disease, two deaths occur, and it is the most frequent cause of cancer-related deaths among females (The National Cancer Registry Ireland; <http://www.ncri.ie>).

1.2 Sub-types in Breast cancer

Breast cancer is considered a highly heterogeneous group of cancers arising from different cell types and each having its own clinical implications. Currently, all breast cancers are tested for expression of Estrogen Receptor (ER), Progesterone Receptor (PR) and HER2/neu proteins. ER and PR tests are usually done by immunohistochemistry whereas HER2/neu is accessed by FISH. This protein profiling of tumors helps to predict the eventual prognosis and can assist in the determination of the most appropriate treatment for the individual.

1.2.1 Estrogen receptor (ER)

The ER is a member of the nuclear hormone family of intracellular receptors which is activated by the hormone 17 β -estradiol (Dahlman-Wright *et al.*, 2006). The main function of ER is as a DNA-binding transcription factor which regulates gene expression (Levin 2005).

There are two different forms of ER, referred as α and β , each encoded by a separate gene. The α isoform is encoded by the ESR1 and the β isoform is encoded by the ESR2 gene (Cowley *et al.*, 1997). Hormone-activated ERs form dimers (Pace *et al.*, 1997). These two forms of ERs are co-expressed in various cell types including thyroid, bone, adrenals and female rat brain (Greco *et al.*, 2003; Arts *et al.*, 1997; Couse *et al.*, 1997; Kuiper *et al.*, 1997). This may lead to the formation of homodimer ER α ($\alpha\alpha$) or ER β ($\beta\beta$)

or heterodimer ER $\alpha\beta$ ($\alpha\beta$) (Li *et al.*,2004; Cowley *et al.*,1997). There is significant overall sequence homology among the two isoforms (Hall, Couse and Korach 2001).

ESR1 is encoded on chromosome 6 (6q25.1) and ESR2 is encoded on chromosome 14 (14q) (Menasce *et al.*, 1993; Sluysers *et al.*,1988). Both ERs are widely expressed in different tissue types, however, there are some differences in their expression patterns (Couse *et al.*,1997). ER α is expressed in endometrial, breast cancer cells, ovarian stroma cells and in the hypothalamus. ER β is expressed in kidney, brain, bone, heart, lungs, intestinal mucosa, prostate, and endothelial cells. The ER α proteins are regarded as being cytoplasmic receptors in their unliganded state, but visualization research has shown that a fraction of the ER α resides in the nucleus of ER-negative breast cancer epithelial cells (Htun *et al.*, 1999). The ER's helix 12 domain plays an important role in determining interactions with co-activators and co-repressors and thereby affecting the respective agonist or antagonist effect of the ligand (Ascenzi, Bocedi and Marino 2006, Bourguet, Germain and Gronemeyer 2000).

ERs are over-expressed in around 70% of breast cancer cases, and are referred to as "ER-positive" tumors. Binding of estrogen to ER stimulates proliferation of mammary cells, with the resulting increase in cell division and DNA replication and increases mutation rate. This causes disruption of the cell cycle, apoptosis and DNA repair processes eventually leading to tumour formation. Additionally, estrogen metabolism leads to the production of genotoxic by-products that could directly damage DNA, resulting in point mutations (Deroo and Korach 2006). ER α expression is associated with more differentiated tumors, while evidence that ER β is involved is controversial (Herynk and Fuqua 2004). However, recent research suggests that ER β is associated with proliferation and a poor prognosis (Rosa *et al.*, 2008). Different versions of the ESR1 gene have been identified (with single-nucleotide polymorphisms) and are associated with different risks of developing breast cancer (Deroo and Korach 2006).

Patients with high levels of ER are treated with endocrine therapy (Normanno *et al.*, 2005). Endocrine therapy for breast cancer involves Selective ER Modulators (SERMS) which act as ER antagonists in breast tissue or aromatase inhibitors which work by

inhibiting the action of the enzyme aromatase which converts androgens into estrogens (Osborne 1999, Tamoxifen for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group 1998). ER status is used to determine sensitivity of breast cancer lesions to tamoxifen and aromatase inhibitors (Fabian and Kimler 2005). Raloxifene, which has anti-estrogenic behaviour has been used as a preventative chemotherapy for women judged to have a high risk of developing breast cancer (Oseni *et al.*, 2008).

1.2.2 Progesterone Receptor (PR)

The progesterone receptor (PR) also known as NR3C3 (nuclear receptor subfamily 3, group C, member 3), is an intracellular steroid receptor that binds progesterone. PR is encoded by the PGR gene which lies on chromosome 11 (11q22) (Law *et al.*, 1987). This gene has two main forms, A and B that differ in their molecular weight (A: 94kDa and B: 114kDa) (Horwitz and Alexander 1983). These two isoforms are transcribed from distinct, estrogen-inducible promoters within a single-copy PR gene; the only difference between them is that the first 164 amino acids of B are absent in A (Giangrande and McDonnell 1999).

PR is expressed in reproductive tissue and has important roles in folliculogenesis, ovulation, implantation and pregnancy (Gadkar-Sable *et al.*, 2005). Estrogen is necessary to induce the progesterone receptors (PRs) activity (Horwitz, Koseki and McGuire 1978). PRs become hyperphosphorylated upon binding of the steroid ligand. PR phosphorylation is complex, occurring in different cellular compartments and perhaps requiring multiple serine kinases (Takimoto and Horwitz, 1993). After progesterone binds to the receptor, restructuring with dimerization follows and the complex enters the nucleus and binds to DNA. There, transcription takes place, resulting in formation of messenger RNA that is translated by ribosomes to produce specific proteins (Edwards *et al.*, 1995, Li and O'Malley 2003).

About 65% of ER-positive breast cancers are also PR-positive and about 5% of breast cancers are ER-negative and PR-positive. If cells have receptors for both hormones or receptors for one of the two hormones, the cancer is considered hormone-receptor-

positive. Co-regulators of PR either enhance or suppress transcription activity and thereby modulate the function of the PR. Chromatin high-mobility group protein 1, chromatin high-mobility group protein 2, TIP60 (Tat-interacting protein), proline-rich nuclear receptor coregulatory protein 1, proline-rich nuclear receptor coregulatory protein 2, Cdc25B, and GT198 enhance PR transcription activity as demonstrated by transient transfection assays (Ko *et al.*, 2002; Ma *et al.*, 2001; Zhou *et al.*, 2000; Brady *et al.*, 1999; Verrijdt *et al.*, 2002). Nuclear receptor corepressor, BRCA1 and Ubiquitin-activating enzyme 3 suppress PR transcription activity (Fan *et al.*, 2002; Gao and Nawaz 2002). A mutation or change in expression of the co-regulators affects the normal function of the PR and may disrupt the normal development of the mammary gland, thereby leading to breast cancer (Gao and Nawaz 2002).

1.2.3 HER2/neu

HER2/neu (also known as ErbB-2, ERBB2) stands for "Human Epidermal growth factor Receptor 2" and is a protein giving higher aggressiveness in breast cancers (Quenel *et al.*, 1995). It is a member of the ErbB protein family, more commonly known as the epidermal growth factor receptor family. HER2/neu belongs to a family of four transmembrane receptor tyrosine kinases involved in signal transduction pathways that regulate cell growth and proliferation (Zhou and Hung 2003).

HER2/neu is notable for its role in the pathogenesis of breast cancer and as a target of treatment. It is a cell membrane surface-bound receptor tyrosine kinase and is normally involved in the signal transduction pathways leading to cell growth and differentiation. HER2 is thought to be an orphan receptor, with none of the EGF family of ligands able to activate it. However, ErbB receptors dimerise on ligand binding, and HER2 is the preferential dimerisation partner of other members of the ErbB family (Olayioye 2001). The HER2 gene is a proto-oncogene located at the long arm of chromosome 17 (17q11.2-q12).

Approximately 30% of breast and ovarian cancers have an amplification of the HER2/neu gene or over-expression of its protein product (Zhou and Hung 2003). Over-expression of this receptor in breast cancer is associated with increased disease recurrence and worse

prognosis. The poor prognosis may be due to global genomic instability as cells with high frequencies of chromosomal alterations have been associated with increased cellular proliferation and aggressive behaviour (Ellsworth *et al.*, 2008).

HER2 is co-localized, and thus most of the time co-amplified, with another proto-oncogene GRB7 (Vinatzer *et al.*, 2005). Clinically, HER2/neu is important as the target of the monoclonal antibody trastuzumab (marketed as Herceptin). Trastuzumab is only effective in breast cancer where the HER2/neu receptor is over-expressed. One of the mechanisms of how trastuzumab works after it binds to HER2 is by increasing p27, a protein that halts cell proliferation (Le, Pruefer and Bast 2005).

1.2.4 Triple Negative (TN) and basal type of breast cancer

Breast cancer is termed triple negative (TN) when there is absence of ER, PR and HER2 receptor proteins. This type of cancer accounts for nearly 20% of all breast cancers (Rhee *et al.*, 2008). TN is a heterogeneous group of breast cancer and is commonly associated with the worst prognosis (Stockmans *et al.*, 2008). TN breast cancer is associated with younger age and more aggressive tumour type. TN breast cancers are generally negative for bcl-2 expression but positive for the epidermal growth factor receptor and have a high level of p53 and Ki67 expression (Rhee *et al.*, 2008).

The basal subtype of breast cancer is accompanied by the expression of cytokeratin and P-cadherin markers (Paredes *et al.*, 2007). Basal-like carcinomas typically express one or more of the basal cytokeratins such as CK5 and CK5/6. CK5 is more sensitive in identifying basal-like tumors than CK5/6 (Bhargava *et al.*, 2008, Bryan, Schnitt and Collins 2006). The majority of TN breast cancers display a "basal-like" molecular profile on gene expression arrays (Anders and Carey 2008). The majority of BRCA1-associated breast cancers are TN and basal-like (Anders and Carey 2008).

1.3 Prognostic markers in breast cancer

1.3.1 Grade

The histological grade of a tumour is determined by a pathologist under a microscope. A well-differentiated (low grade) tumour resembles normal tissue. A poorly differentiated

(high grade) tumour is composed of disorganized cells and, therefore, does not look like normal tissue. Moderately differentiated (intermediate grade) tumors are somewhere in between.

The Bloom-Richardson grade (BR grade) (BLOOM and RICHARDSON 1957) is a histological grade assigned by pathologists to breast cancers. It is the most common type of cancer grade system currently used. It is a semi-quantitative grading method based on three morphologic features of invasive breast cancers. The morphologic features that are used are:

- 1) Degree of tumour tubule formation *i.e.* percentage of cancer composed of tubular structures
- 2) Tumour mitotic activity or rate of cell division.
- 3) Nuclear polymorphism of tumour cells, nuclear grade, change in cell size and uniformity.

Each of these features is assigned a score ranging from 1 to 3. The scores are then added together for a final sum that will be in the range of 3 to 9. This value is then used to grade the tumour as follows:

Value: 3-5 Grade 1 tumors (well-differentiated): Tumors with Grade 1 are associated with a good prognosis.

Value: 6-7 Grade 2 tumors (moderately-differentiated): Tumors with Grade 2 are associated with an intermediate prognosis.

Value: 8-9 Grade 3 tumors (poorly-differentiated): Tumors with Grade 3 are associated with a bad prognosis.

1.3.2 Lymph node metastasis

Lymph node metastasis is considered an important prognostic parameter in treating breast cancer patients. The sentinel node is the first lymph node reached by metastasising cells

from a primary tumour. A sentinel node biopsy is a minimally invasive technique to identify lymph node metastases (Tanis *et al.*, 2001). Involvement of a lymph node in breast cancer significantly correlates with worse prognosis compared with no lymph node involvement (Colleoni *et al.*, 2005). Such patients have a higher incidence of death due to disease (Jatoi *et al.*, 1999) and should therefore be treated more aggressively.

1.4 Invasion and Metastasis

One of the most lethal aspects of breast tumors is their ability to **invade** the surrounding normal mammary tissue and re-locate to other sites in the body distal to the primary tumour, whereby tumour growth begins anew (**metastasis**). While the process by which cancer cells lose adherence to the primary tumour and develop migratory and invasive capacities has been well-described at the cellular level, the progression of invasion is still poorly understood at the molecular level. Cancer cells from a primary tumour enter lymphatic and blood vessels, circulate through the bloodstream, and settle down to grow within normal tissues elsewhere in the body. Most tumors, if left un-treated can metastasize to other parts of body.

However, there are cancers with very low metastatic potential such as glioma and basal cell carcinoma. When tumour cells metastasize, the new tumour is known as a secondary or metastatic tumour, and often displays properties of the original (primary) tumour. Metastasis can occur long after the apparent elimination of the primary tumour. In breast cancer, metastases have been known to occur decades after the primary treatment (Karrison, Ferguson and Meier 1999). Cancer cells can exist in three separate states in a secondary site, solitary cells in quiescence, active pre-angiogenic micrometastases, in which proliferation is balanced with apoptosis and no net increase in tumour size occurs, and vascularised metastases, either small and clinically undetectable, or large and detectable by current technology (Demicheli 2001).

Metastatic tumors are very common in the late stages of cancer. The spread of cancer cells may occur via the blood or the lymphatic system or through both routes. There is also a propensity for certain tumors to metastasize to particular organs (Chambers, Groom and MacDonald 2002). Successful formation of metastases requires angiogenesis

at the primary tumour, down regulation of cohesive molecules, increased motility of tumour cells, invasion into neo-vessels, tumour cell embolism, arrest and attachment in capillary beds of distant organs, extravasations and proliferation in the organ parenchyma and re-establishment of angiogenesis when the tumour reaches > 1-2 mm in size (Li *et al.*, 2000).

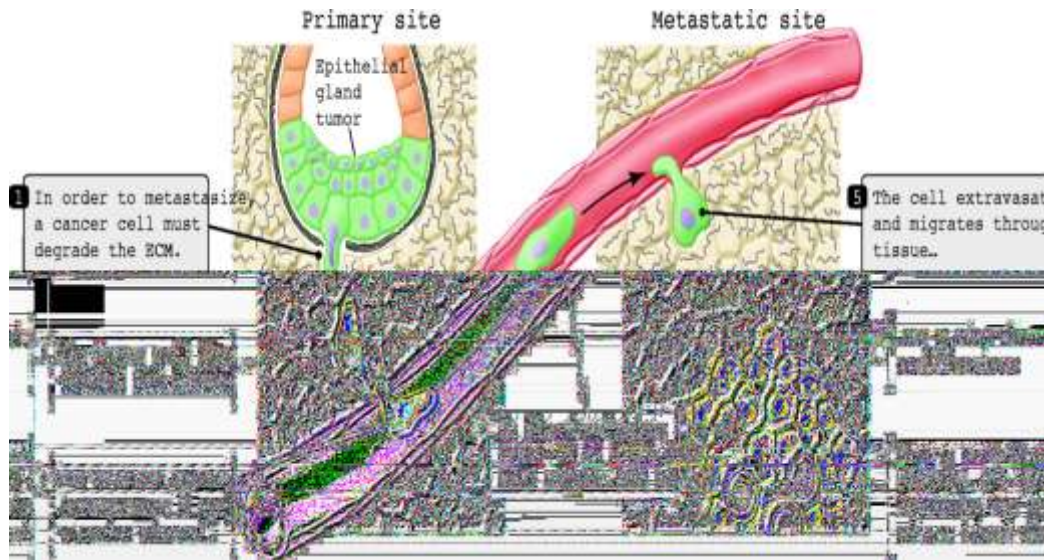


Fig 1.4: The metastatic sequence (Geho *et al.*, 2005)

1.4.1 Microenvironment of Breast Cancer

In vivo, every cell functions within its microenvironment. The mammary duct consists of epithelial cells surrounded by stroma, including fibroblast cells and other support components. A thin layer of extracellular matrix (ECM) lies between epithelial cells and stroma (Woodward, Xie and Haslam 1998). The proliferation and phenotype of breast epithelial cells are the results of the epithelial-epithelial cell, epithelial-stromal cell and epithelial cell-ECM interactions (Haslam and Woodward 2003). Carcinogenesis of the breast cells causes both transformation of cells and changes to their microenvironment. Four kinds of cell connections are known to be important in maintaining the epithelial layer: tight junctions, adherens junctions, desmosomes and gap junctions (Ehmann *et al.*, 1998).

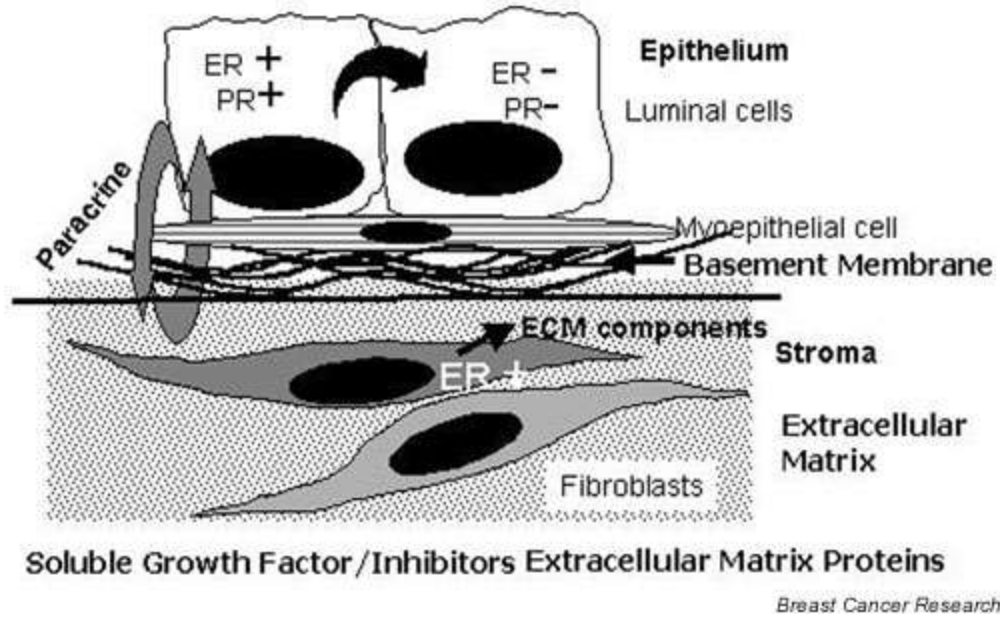


Fig 1.4.1: Model of epithelial-cell–stromal-cell interactions. ECM, extracellular matrix; ER, estrogen receptor; PR, progesterone receptor (Haslam and Woodward 2003).

The function of these structures is to restrict both free cell movement and infinite proliferation. Gap junctions also act as the pathways for cell-cell communication, helping the biological signals of a cell, such as cytokines or the products of tumour suppressor genes, to pass into the neighbouring cells. Thus, one prerequisite of cell transformation is the breakdown of physical connections between cells, which means the loss of cell proliferation restriction. Epithelial cells are also found to form physical junctions with the ECM and the stroma. In fact, during metastasis, migrating cells alternatively attach to and detach from ECM and stroma to move forward (Price, Bonovich and Kohn 1997). However, ECM has been shown to serve as a natural physical obstacle of metastasis. Some studies have demonstrated that Matrigel, on ECM extract from tumors, promotes the formation of tumors and blood vessels in mouse models (Noel and Foidart 1998).

Fibroblasts are the main cell type in stroma and have a similar influence on epithelial cells. When co-injected with Matrigel and mammary tumour cells, fibroblasts accelerated tumour formation in a mouse model (Noel and Foidart 1998). Studies have also shown that fibroblasts stimulate the movement and proliferation of cancerous epithelial cells

when they have direct cell contact in co-culture (Korohoda and Madeja 1997, Olumi, Dazin and Tlsty 1998).

The last phase of primary tumour development is progression, which usually starts with growth of the dormant tumour following the promotion phase. The rapid progression phase is triggered once the new blood vessels are formed in the primary tumour. As a result, the cancer cells acquire the ability to metastasize.

1.4.2 Angiogenesis

With the proliferation of the primary tumour, angiogenesis, or the generation of new blood vessels, becomes necessary. Tumors induce blood vessel growth (angiogenesis) by secreting various growth factors (e.g. vascular endothelial growth factor (VEGF) and basic fibroblast growth factor (bFGF)) (Barinaga 1997). Growth factors such as bFGF and VEGF can induce capillary growth into the tumour, which some researchers suspect supply required nutrients, allowing for tumour expansion (Hanahan and Folkman 1996; Sato *et al.*, 2000; Ferrara 2001). Angiogenesis also increases the possibility of the tumour cells to enter into the circulation. Thus, angiogenesis promotes tumour cell invasion.

1.4.3 Invasion

To access the circulation, tumour cells must cross the ECM. This active process is called invasion. Invasion is a process that includes proteolysis of the ECM, pseudopodial extension and cell migration (Palecek *et al.*, 1997; Wolf *et al.*, 2007b). It usually happens when the tumour size is relatively large. On the other hand, ECM and interstitial stroma act integrally as the barriers that must be overcome for invasion (Price, Bonovich and Kohn 1997, Nicolson 1988; Woodhouse, Chuaqui and Liotta 1997). For example, matrix metalloproteinases (MMPs) are over-expressed by most kinds of metastatic cells and are essential for degradation of the ECM (Price, Bonovich and Kohn 1997; Sengupta and MacDonald 2007).

1.4.4 Transport

Cancer cells can enter into the circulatory system indirectly via the lymphatic system and are thus transported to distant tissues (Chambers, Groom and MacDonald 2002). Once tumour cells enter the blood circulation, they are exposed to shear stress and interactions with leukocytes which could lead to their destruction. Cancer cells are capable of resisting leucocyte mediated destruction by forming a thrombus, adhering to the endothelia of ductal structures and thereby protecting themselves from the immune system (Bacac and Stamenkovic 2008). Considering the ability of cancer cells to proliferate infinitely, formation of secondary tumour growth is not a rare event once tumour cells have entered the circulation (Price, Bonovich and Kohn 1997).

1.4.5 Arrest

During this step, circulating cancer cells embed into the vascular endothelia forming a secondary site for tumour growth. Several factors contribute to this stage; mechanical trapping of tumour cells at a secondary site by small capillary beds; clusters of cancer cells are blocked at very narrow blood vessels; tumour cell adhesion at a secondary site by the expression of appropriate cell surface proteins; cancer cells are recognized and bound by receptors on the endothelial duct (Price, Bonovich and Kohn 1997; Nicolson 1988; Horak and Steeg 2005).

1.4.6 Extravasation

Extravasation can be taken as the reverse process of invasion, during which the arrested cells enter into the secondary sites and are followed by formation of a new tumour (Price, Bonovich and Kohn 1997; Nicolson 1988). Taken together, metastasis is a multi-variable process and demonstrates diverse behaviour in different kinds of cancer (Price, Bonovich and Kohn 1997; Nicolson 1988). The malignant cell's metastatic properties are influenced by expression of many genes related to degradative enzymes or their inhibitors, cell adhesion components, growth factor receptors, programmed cell death or apoptosis, cell-cell communication components, cell motility components and host surveillance mechanisms (Price, Bonovich and Kohn 1997; Demicheli *et al.*, 1997; Ben-Baruch 2008).

1.4.7 Breast cancer metastasis-associated genes

A number of genes have been investigated for changes in expression level during progression of breast cancer. An ability to circumvent the defences of the microenvironment is critical for progression of mammary tumors to malignancy. This process was first investigated by transferring a dominant oncogene into susceptible cells and then following progression of malignancy in animal models, such as mouse. Later, it was confirmed that the slow, stepwise changes in mammary cancer progression are not only qualitative, but can be quantified. Some of these changes can be reversible and do not involve dominantly acting oncogenes and tumour suppressor genes (Nicolson 1998). Thus, there is still no common cascade of changes to gene expression levels found in breast cancer, like that of colon cancer. In addition, oncogenes and tumour suppressor genes, genes regulating cell cycle, growth factors and their receptors and intercellular communication have become targets of research to better understand the progression of breast cancer. A brief description of some of the better known target genes will be outlined here.

ER and PR and their ligands play important roles in the development and function of the mammary gland. Normal human mammary epithelial cells express very low or no levels of ER and PR. But, in breast cancer patients, about two-thirds of tumour tissues are ER-positive by immunohistochemical analysis (Allred *et al.*, 1998; Lapidus, Nass and Davidson 1998).

p53 usually functions as a tumour suppressor by regulating transcription, cell cycle, and apoptosis. Mutations of p53 detected in breast cancers are primarily point mutations that often lead to loss of function of wild type p53 and over-expression of mutant p53 in malignant cells (Lacroix, Toillon and Leclercq 2006; Ravaioli *et al.*, 1998).

The c-erbB-2 or HER2/neu, gene codes for a transmembrane tyrosine kinase and acts as a receptor of a group of peptide ligands that can stimulate cell growth, cellular differentiation, adhesion and motility. Over-expression of HER2/neu is detected in 20-30% metastatic breast cancer (Ravaioli *et al.*, 1998; Hyun *et al.*, 2008).

The *bcl-2* gene is involved in the regulation of cell death, inhibiting apoptosis, and is found over expressed in breast cancer. Bcl-2 has been retrospectively considered as a potential prognostic factor of breast cancer (Callagy *et al.*, 2006).

Cyclins are a group of proteins that regulate cell cycle and deregulation of cell cycle control is one of the most evident alterations in cancer cell growth. Aberrations in sequence and expression of cyclins B1, D1, E, etc., are often detected in breast cancer (Ravaioli *et al.*, 1998).

Kai-1 is one of the few metastasis-suppressor genes discovered and was first mapped out in prostate cancer (Dong *et al.*, 1995). Later, it was shown that transfection of the Kai-1 gene into breast cancer cells suppresses their metastatic ability and may be a useful marker for staging human breast diseases (Phillips *et al.*, 1998).

MTAL, a novel gene identified in 1998, is associated with mammary tumour metastasis and may also be involved in human breast cell motility and growth regulation. Antisense blocking experiments showed that MTAL may stimulate the highly malignant breast cancer cells to move into and grow in distant sites such as bone and brain, which are common sites for breast cancer metastasis (Nicolson 1998).

1.5 Gene expression profiling

Expression microarray profiling is a high throughput technology used in molecular biology and biotechnology to simultaneously access the gene expression profile of thousands of genes. A typical microarray chip consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, each containing a small amount of a specific DNA sequence. This can be a short section of a gene or other DNA element that are used as probes to hybridize a cDNA or cRNA sample under appropriate conditions. The hybridization is detected and quantified by fluorescence-based detection of fluorophore-labeled targets to determine relative abundance of nucleic acid sequences in the sample.

In standard microarrays, the probes are attached to a solid surface made of glass or silicon by a covalent bond to a chemical matrix via epoxy-silane, amino-silane, lysine and polyacrylamide (Derisi 2001). Affymetrix technology uses a photolithographic technology to synthesize 25-mer oligonucleotides on a silica wafer (<http://www.affymetrix.com>). Other microarray platforms, such as Illumina, use microscopic beads, instead of the large solid support (<http://www.illumina.com/>).

1.5.1 Affymetrix microarrays

The microarray experiments carried out in our study employed the Affymetrix GeneChip system. Affymetrix probes are designed using publicly available information. The sequences, from which the probe sets were derived, were selected from GenBank, dbEST, and RefSeq. The sequence clusters were created from the UniGene database (Build 133, April 20, 2001) and then refined by analysis and comparison with a number of other publicly available databases, including the Washington University EST trace repository and the University of California, Santa Cruz Golden-Path human genome database (April 2001 release). Sequences from these databases were collected and clustered into groups of similar sequences.

The probes are manufactured on the chip using photolithography (a process of using light to control the manufacture of multiple layers of material), which is adapted from the computer chip industry. Each GeneChip contains approximately 1,000,000 features. Each probe is spotted as a pair, one being a perfect match (PM), and the other with a mismatch (MM) at the centre. These probe pairs allow the quantitation and subtraction of signals caused by non-specific cross-hybridisation. The differences in hybridisation signals between the partners, as well as their intensity ratios, serve as indicators of specific target abundance. Each gene or transcript is represented on the GeneChip by 11 probe pairs. The probe sets are given different suffixes to describe their uniqueness and/ or their ability to bind different genes or splice variants.

- “_at” describes probes set that are unique to one gene
- “_a_at” describes probe sets that recognise multiple transcripts from the same gene

- “_s_at” describes probe sets with common probes among multiple transcripts from separate genes. The _s_at probe sets can represent shorter forms of alternatively polyadenylated transcripts, common regions in the 3’ ends of multiple alternative splice forms, or highly similar transcripts. Approximately 90% of the _s_at probe sets represent splice variants. Some transcripts will also be represented by unique _at probe sets.
- “_x_at” designates probe sets where it was not possible to select either a unique probe set or a probe set with identical probes among multiple transcripts. Rules for cross-hybridisation are dropped in order to design the _x_at probe sets. These probe sets share some probes identically with two or more sequences and therefore, these probe sets may cross-hybridise in an unpredictable manner.

A sample must be registered and an experiment defined in GCOS (GeneChip Operating Software) before processing a probe array in the fluidics station or scanning. Once the array is scanned, an image file is created called a “.dat” file. The software then computes cell intensity data (“.cel” file) from the image file. It contains a single intensity value for each probe cell delineated by the grid (calculated by the Cell Analysis algorithm). The amount of light emitted at 570nm from stained chip is proportional to the amount of labelled RNA bound to each probe. Each spot correspond to individual probe (either perfect match or mismatch). The probes for each gene are distributed randomly across the chip to nullify any region specific bias. Following this, data analysis algorithms combine the probes to the respective intensity of individual transcripts (see section 1.6).

1.5.2 Microarrays and Breast cancer

Microarray analyses of clinical breast cancer specimens and cell lines have identified gene expression profiles which separated the tumors into various groups and sub-groups. These sub-groups have been associated with different clinical outcomes. The various sub-groups that have been defined using a microarray approach are Luminal A, Luminal B, ERBB2 over-expressing, Basal sub-type and Normal-like (Sorlie *et al.*, 2001). The Luminal sub-type A identified has a higher ESR1 and ER partner gene over-expression than Luminal sub-type B. Luminal sub-type A is considered to have a better prognosis

and response to Endocrine therapy than Luminal sub-type B. The ERBB2 and Basal sub-groups of patients display a more aggressive form of cancer. The ERBB2 group of patients respond well to Herceptin. Additionally, an Apocrine (Androgen receptor positive) group has been defined based on microarray studies (Farmer *et al.*, 2005). Apocrine sub-type is defined as Androgen receptor-positive and negative for Estrogen, Progesterone and HER2 protein. Apocrine sub-type of breast cancer is regulated by androgen.

It was previously thought that a few cells from a tumour attain metastatic potential and move to different parts of body, where they develop as secondary tumors (Fidler and Kripke 1977; Poste and Fidler 1980). With the advent of microarray-based studies, this hypothesis has changed and it is now believed that metastasis potential is the property of the whole tumour rather than a sub-set of cells as previously thought (Ma *et al.*, 2003; Weigelt *et al.*, 2003). These findings resulted in studies aimed at identifying genes which may be involved in metastasis, relapse and shorter survival. These genes have been used to develop prognostic models to predict long term relapse (van 't Veer *et al.*, 2002; Huang *et al.*, 2003; Karlsson *et al.*, 2008). Following successful attempts to identify the prognostic important genes and develop prediction models, microarrays have evolved into diagnostic assays. Studies based on gene expression were later translated to diagnostic assays has been approved by the US Food and Drug Administration (FDA; <http://www.fda.gov/>) for routine use on breast cancer patients (see section 1.5.3). Apart from predicting clinical outcomes, these kits can also indicate therapeutic options for patients.

1.5.3 Gene expression in diagnostics

1.5.3.1 OncotypeDx

OncotypeDx, developed by Genomic Health (<http://www.genomichealth.com>), is a diagnostic kit that aims to quantify the likelihood of disease recurrence in women with early-stage breast cancer (Paik *et al.*, 2004) and also assesses the likely benefit from certain types of chemotherapy (Paik *et al.*, 2006). The OncotypeDx diagnostic assay is suitable for women with early-stage invasive breast cancer, who are ER-positive and

lymph node-negative. Typically in these cases, treatment with hormonal therapy, such as tamoxifen, is indicated. OncotypeDx is not suitable for patients with carcinoma *in-situ* or metastatic breast cancer.

250 candidate genes possibly associated with breast cancer tumour behaviour were identified from published literature, genomic databases and microarray experiments. These genes were analyzed in 447 patients from three independent clinical studies in order to identify a panel of 21 genes strongly correlated with distant recurrence-free survival (Paik *et al.*, 2004).

Group	Genes
Proliferation	Ki67, STK15, Survivin, CCNB1, MYBL2
Invasion	MMP11, CTSL2
HER2	GRB7, HER2
Estrogen	ER, PGR, BCL2, SCUBE2
Others	GSTM1, CD68, BAG1
Controls	ACTB, GAPDH, RPLPO, GUS, TFRC

Table 1.5.3.1: List of genes on the OncotypeDx assay

OncotypeDx analyzes expression of these 21 genes from tumour mRNA to determine a prognostic recurrence score. The recurrence score is a number between 0 and 100 and corresponds to a likelihood of breast cancer recurrence within 10 years of the initial diagnosis (Paik *et al.*, 2004). The result was later validated on a very large study of 4,964 node-negative breast cancer patients (Habel *et al.*, 2006). If successful in onward trials, this information would help doctors choose the right combination and medicinal dose for individual patients. Despite being an expensive test, it could result in considerable cost saving considering the fact that chemotherapy can cost thousands of euros per year, per patient (Hornberger, Cosler and Lyman 2005).

OncotypeDx is a non-invasive test that is performed on a small amount of the tissue removed during the original lumpectomy, mastectomy, or core biopsy. The tissue sample is fixed in formalin and embedded in paraffin so it can be preserved and send to Genomic

Health for further diagnostic testing where the RNA is isolated from sectioned tissue blocks using the MasterPure Purification kit (Epicenter, Madison, WI) and subsequently the assay is performed.

1.5.3.2 MammaPrint

MammaPrint is a microarray-based molecular diagnostic test that is used to assess the risk that a breast tumour will metastasize to other parts of the body. MammaPrint is marketed by Agendia (<http://agendia.com>) and assesses the risk factor for distant metastasis within five to 10 years (Glas *et al.*, 2006). The test is suitable for lymph node-negative breast cancer patients under 61 years of age with tumors of less than 5cm in diameter.

MammaPrint uses a 70 gene signature, obtained by microarray studies to classify patients as low or high risk for recurrence of the disease. The 70 gene signature was previously identified by analysing microarray data from 34 patients who developed distant metastasis within five years and 44 patients who remained disease free for at least five years (van 't Veer *et al.*, 2002). The results were later validated in independent studies (van de Vijver *et al.*, 2002; Buyse *et al.*, 2006).

MammaPrint estimates the expression of 70 identified genes in the tumour sample and compares the gene expression profile to reference expression profiles of 'Low Risk' or 'High Risk' profiles. The risk of tumour recurrence is then determined according to the degree of similarity between the tumour gene expression profile and reference profiles (van de Vijver *et al.*, 2002).

A low risk patient has a 95% chance of being metastasis-free within the following five years and 90% chance of being metastasis-free within the following 10 years, whereas a high risk patient has a 78% chance of being metastasis-free within the following five years and 71% chance of being metastasis-free within the following 10 years (van de Vijver *et al.*, 2002).

1.6 Microarray Data Analysis

1.6.1 Normalization

Normalization is the first step in the data analysis process. Normalization is the adjusting of microarray data to remove variations that arise from the technology rather than from biological differences between the RNA samples. Some of the common normalization algorithms for Affymetrix arrays are MAS5 (Pepper *et al.*, 2007), RMA (Irizarry *et al.*, 2003) and dChip (Li and Hung Wong 2001).

MAS5 developed by Affymetrix uses a reference (baseline) chip which is used to normalise all the experimental chips. The procedure is to adjust the intensity of each probes against the corresponding probes on the baseline chip; eliminate the highest 1% of probes (and for symmetry the lowest 1%), and fit a regression line to the middle 98% of probes.

dChip uses an array with median overall intensity as the baseline array against which other arrays are normalised at probe level intensity. Subsequently a subset of PM (“perfect match”) probes, with small within-subset rank difference in the two arrays (also known as invariant set), serves as the basis for fitting a normalisation curve.

RMA employs normalization at probe level using the quantile method. This normalization method makes the chips have identical intensity distribution

1.6.2 Clustering

Clustering is the grouping of objects based on similarity. In other words it is the partitioning of a data set into subsets, so that the data in each subset share some common trait. The measure for a common trait is defined before the clustering is performed and is often a distance metric defining the relative similarity among the two objects. Data clustering is a common technique for statistical data analysis, and has applications to many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics.

Clustering gene expression data helps in identifying genes of similar function. These co-expressed genes with poorly characterized or novel genes may provide a simple means of gaining insight to the functions of many genes for which information is not available currently (Eisen *et al.*, 1998). Co-regulated families of genes cluster together, as was demonstrated by the clustering of ribosomal genes as a group (Alon *et al.*, 1999). Clustering is also used to identify the grouping patterns of specimens and has been widely used in studying the heterogeneity of cancer. Clinical breast cancers cluster as distinct groups based on their gene expression profiles and can be correlated with clinical outcomes (Sorlie *et al.*, 2001).

Primarily, most clustering techniques use a distance metric to define the similarity or difference among the two objects. Some of the most common distance metrics used are Euclidean distance, Manhattan distance and Correlation distance. Euclidean distance is the distance between two points that would be measured with a simple ruler, and can be also calculated by repeated application of the Pythagorean Theorem. Thus the distance measure would be:

$$\text{Distance} = \sqrt{(\sum (X_i - Y_i)^2)}$$

X and Y are expression vectors of genes or samples.

Manhattan distance is the distance between two points expressed as the sum of the absolute differences of their coordinates. Therefore the distance between point P₁ with coordinates (x₁, y₁) and the point P₂ at (x₂, y₂) would be |x₁ - x₂| + |y₁ - y₂|.

Correlation distance measures the similarity between two points expressed as the correlation between the two objects. Often the Pearson correlation measure is taken as distance measure for most of the microarray data clustering. Correlation measure value range from -1 to +1. Positive values indicate a positive correlation (*i.e.* increase in value of one corresponds to increase in the value of the other). Negative values indicate a negative correlation (*i.e.* increase in value of one corresponds to decrease in value of the other and *vice versa*). A correlation value of 0 indicates no relation between the two values.

1.6.2.1 Hierarchical clustering

Hierarchical clustering is a technique to generate a hierarchy among objects based on their similarity or differences. The similarity or difference is measured based on the distance criteria explained above. Hierarchical clustering may be constructed using an agglomerative or divisive approach. The representation of this hierarchy is a tree also known as dendrogram, with individual elements at one end and a single cluster containing every element at the other (Fig 1.6.2.1.1). Agglomerative algorithms begin at the leaves of the tree, whereas divisive algorithms begin at the root. Agglomerative clustering can be single linkage clustering, complete linkage clustering or average linkage clustering.

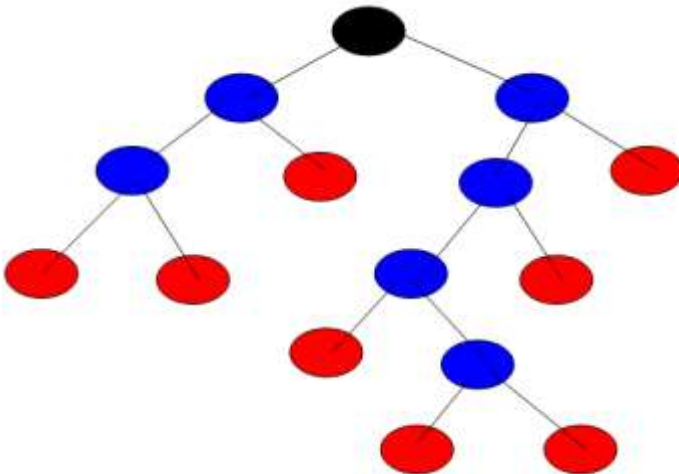


Fig 1.6.2.1.1: An example of a tree or dendrogram. The leaves are shown in red and the nodes are shown in blue. A leaf reflects the entity and a node reflects the relationship between two entities, one entity and one node or among two nodes.

Single linkage clustering: The distance between groups is defined as the distance between the closest pair of objects, and only pairs consisting of one object from each group are considered (Fig 1.6.2.1.2).

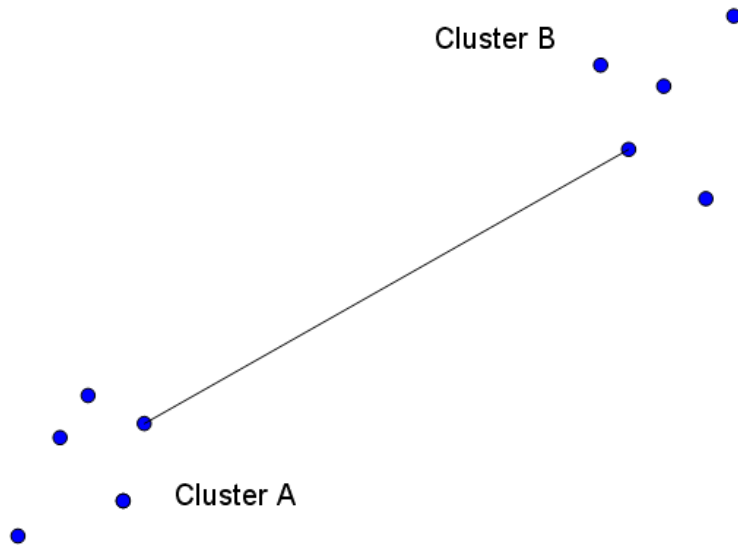


Fig 1.6.2.1.2: Single linkage clustering. The closest element in the cluster is used to calculate the reference distance among the two clusters.

Complete linkage clustering: The complete linkage, also called farthest neighbour clustering method is the opposite of single linkage. The distance between groups is defined as the distance between the most distant pair of objects, one from each group (Fig 1.6.2.1.3).

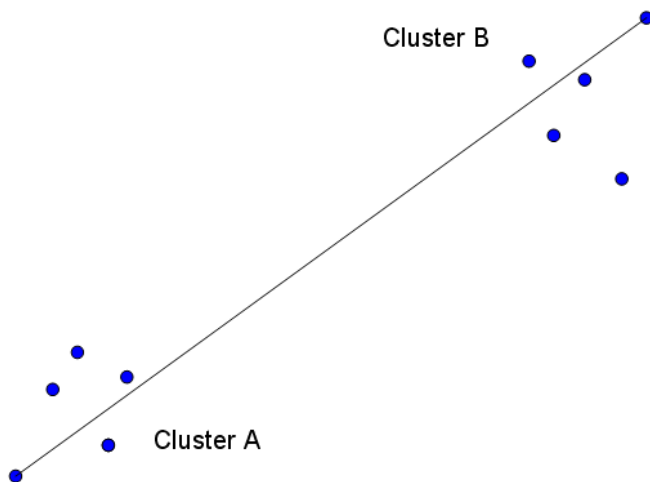


Fig 1.6.2.1.3: Complete linkage clustering. The most distant element in the cluster is used to calculate the reference distance among the two clusters.

Average linkage clustering: Distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group (Fig 1.6.2.1.4).

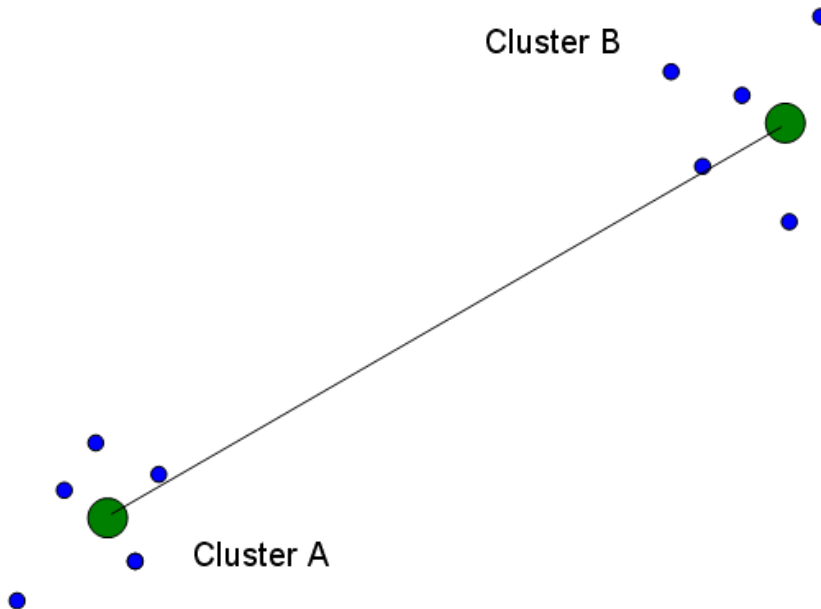


Fig 1.6.2.1.4: Average linkage clustering. The average of the element in the cluster is used to calculate the reference distance among the two clusters. The green is the average or centroid of the cluster.

Hierarchical clustering has been extensively used in cancer research to identify relationship among genes and samples. Hierarchical clustering using multiple markers can group breast cancers into various classes with clinical relevance and is superior to individual prognostic markers (Makretsov *et al.*, 2004). Hierarchical clustering has been widely used in studying the sub-groups in breast cancer (Sorlie *et al.*, 2001; Charafe-Jauffret *et al.*, 2006; Weigelt *et al.*, 2005; Hu *et al.*, 2006).

1.6.2.2 K-Means clustering

The k-means algorithm is an algorithm to cluster objects into k partitions using the similarity between the objects. k is the number of partitions/clusters and is provided by the user. The algorithm starts by partitioning the input points into k initial sets randomly

or by using some heuristic approaches. It then calculates the centroid (mean point), of each set. Thereafter, it constructs a new partition by associating each object with the closest centroid. The centroids are then recalculated for the new clusters, and the process repeated by alternate application of these two steps until convergence, which is obtained when the objects no longer switch clusters or the centroids no longer change. K-means is one of the most commonly used clustering methods and has a wide application in microarray studies (Do and Choi 2008).

Limitations of k-means clustering (MacKay 2003)

- 1) Since k-means clustering starts with random seed points, the end result will not be the same and will depend on the initial random vector.
- 2) K-means clustering needs the number of clusters from the user and forces all the genes/samples to fit on those defined number of clusters.
- 3) Does not work well with non-globular clusters. Non-globular clusters are those whose boundaries are not well defined.

1.6.3 Principal component analysis

Principal component analysis (PCA) is a method to reduce multidimensional data sets to lower dimensions for easier analysis and visualization. PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

PCA can be used for dimensionality reduction in a data set by retaining those features of the data set that contribute most to its variance, by keeping lower-order principal components and ignoring higher-order ones. Lower-order components contain the most important essence of the data and higher-order components contain the least important

essence of the data. However, this may not be the case with all types of datasets. PCA is used in microarray experiments to identify the most significant patterns in the data.

Raychaudhuri, Stuart and Altman (2000), working with yeast sporulation data, concluded that much of the observed variability in the experiment was captured by the first two components corresponding to overall induction level and change in induction level over time.

1.7 Classification

Class predictions are supervised learning algorithms which learn the outcomes from the known (“Training”) dataset, in order to accurately predict the outcomes on the new (“Test”) datasets. Class prediction has a wide applicability both in research and in diagnosis. Some of the most popular algorithms include k-nearest neighbor, Support Vector Machines (SVM), Linear Discriminant Analysis (LDA) classification and Neural Networks.

Back-propagation is a class of neural network algorithm which can be used to accurately predict the outcomes based on its learning on a known dataset. Neural networks have been used extensively in gene-finding (Sherriff and Ott, 2001), protein structure prediction (Cai, Liu and Chou 2003), drug screening (Jaiswal and Naik, 2008), cancer class prediction and clinical outcome prediction in cancer (De Laurentiis *et al.*, 1999) and other diseases.

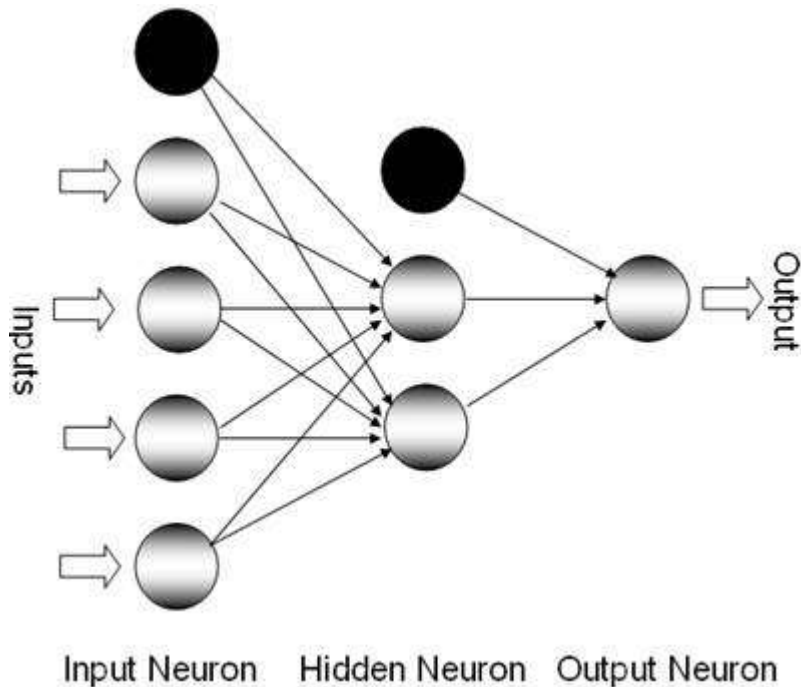


Fig 1.7.1: Generalized representation of the Multiple Layer Perceptron Architecture. Input layer gets the input and the information is processed in the network and the output is obtained on the output layer.

Back-propagation uses the Multiple Layer Perceptron architecture to learn complex patterns. The Multiple Layer Perceptron is an architecture whereby the neurons are in layers; an input neuron layer where the network gets the input and an output neuron layer which gets the output. In between can be n layers of hidden neurons. The neurons in each layer are interconnected with all the neurons in the previous and next layer of neurons. These interconnections are associated with weights which helps in learning complex problems (Haykin 1998). The weights are numbers and contain information on the positive or negative regulation of any particular neuron on the closest neuron. The weights are adjusted in a way that the more important interconnections attain a higher value than the less important interconnections.

Back propagation, which as the name suggests, is the propagation of error to the previous layer of network, is a very efficient method of training the artificial neural networks to

perform a particular task. It was first described by Paul Werbos in 1974, but it wasn't until 1986, through the work of David E. Rumelhart, Geoffrey E. Hinton and Ronald J. Williams, that it gained recognition leading to a renaissance in the field of artificial neural network (Rumelhart *et al.*, 1986).

Back propagation is a supervised learning method and implements the “delta rule”. The delta rule is a gradient descent learning rule for updating the weights of the artificial neurons in a single-layer perceptron. The algorithm requires a guided training that knows, or can calculate, the desired output for any given input. The difference in the expected and actual result is termed as error and this error is back-propagated to the network and weight correction measures are done in a way to minimize the error.

1.8 Representative nature of cell line models to clinical conditions

Cell line models are routinely studied to understand particular biological phenomena, with the expectation that discoveries made in these models will provide insight into human biology. These models are widely used to explore potential causes and treatments for human disease, where experimentation on humans would be unfeasible or unethical. Breast cancer cell lines are generated from cells isolated from breast tumour specimens and have the capability to divide indefinitely when grown *in-vitro* under stringent growth conditions. This potential makes these cell lines an excellent model of study for understanding the basic biology of breast cancer. Many studies which are not possible on animal models can be carried out relatively easily on these cell lines.

There is, however, a great difference in the growth environment of the cancer cells *in-vivo* to that of *in-vitro*. Despite the relatively large number of cancer cell lines currently under study in a variety of clinical settings worldwide, so far studies aiming at investigating the similarity of cell line models to their respective clinical conditions have been very limited. A previous study (Gazdar *et al.*, 1998) found that only a small subset of primary breast cancers that display certain features of advanced tumour and poor prognosis can be cultured for a lengthy time. This group (Wistuba *et al.*, 1998) also reported that there was an excellent correlation among the cell lines to their clinical specimens, in terms of morphological features, presence of aneuploidy,

immunohistochemical expression of ER, HER2/neu, p53 proteins, allelic loss at all of the chromosomal regions analysed and TP53 gene mutations. A more recent study (Burdall *et al.*, 2003), concluded that most of the currently used cell lines are derived from metastatic sites rather than primary tumour and therefore may not be representative of the diverse nature of breast cancer.

The advent of large-scale expression profiling experiments heralded by developments in microarray technology has facilitated a whole-genome analysis approach to this question. Large-scale expression profiling has made it possible to quantify the gene expression profiles of thousands of genes in a single experiment, thereby allowing the comparison of different samples on the basis of their full genomic expression profile, rather than on a selected number of genes. A previous study (Chang, Hilsenbeck and Fuqua 2005) reviewed the role of microarrays in management and treatment of breast cancer, and observed that a combined genomic approach should be taken to understand the heterogeneity of breast cancer.

Given the novelty of microarrays, the number of studies utilizing this technology to investigate the similarity between the gene expression profiles of cell lines and clinical specimens is limited. Previously (Ross and Perou 2001), it has been found that cell lines and tumour specimens have distinct gene expression patterns which need to be considered for their appropriateness for each subtype of clinical conditions. Another study (Dairkee *et al.*, 2004) compared gene expression profiles of early passage tumour cultures and immortal cell lines and observed that epithelial cultures isolated from primary breast tumors retain the characteristics of the tumour, but these characteristics are eliminated following *in vitro* selection of the rapidly proliferating cell population. In a similar comparative study of gene expression profiles of lung cancer cell lines and their respective clinical specimens (Wang *et al.*, 2006), it was observed that 51 of 59 cell lines represented their presumed tumors of origin.

1.9 Small interfering RNA (siRNA)

Small interfering RNA (siRNA) is also known as short interfering RNA or silencing RNA and covers a class of 20-25 nucleotide-long double-stranded RNA molecules. In the late nineties, RNA silencing was discovered in plants during the course of transgenic experiments that eventually led to the silencing of the introduced transgene and, in some cases, of homologous endogenous genes or resident transgenes (Matzke *et al.*, 1989; Linn *et al.*, 1990; Napoli, Lemieux and Jorgensen 1990; Smith *et al.*, 1990; van der Krol *et al.*, 1990). However, this approach could not be used in mammalian cells as the long double-stranded RNAs (dsRNAs) triggered a cytotoxic reaction leading to cell death (Hunter *et al.*, 1975). This cytotoxic reaction, mediated by the interferon system, protected the organism from RNA viruses by sacrificing the infected cell and thus preventing the spread of the virus (Stark *et al.*, 1998). It was later reported that the dsRNAs shorter than 30 nucleotides do not trigger the interferon response; therefore artificially produced siRNAs and their delivery into mammalian cells were able to efficiently induce RNA silencing (Elbashir *et al.*, 2001).

1.9.1 Mechanism of action of siRNA

Long double-stranded RNA (dsRNA) (typically >200 nt) (upon introduction), enters a cellular pathway that is commonly referred to as the RNA interference (RNAi) pathway. During the initiation stage, long dsRNA is cleaved into siRNA (Hamilton *et al.*, 2002), mediated by type III RNase Dicer enzyme. RNase III family members are among the few nucleases that show specificity for dsRNAs (Hamilton *et al.*, 2002) and are evolutionarily conserved in worms, flies, fungi, plants, and mammals (Aggarwal *et al.*, 2006). Complete digestion, by RNase III enzyme results in dsRNA fragments of 23- to 28-mer diced siRNA products (Blaszczyk *et al.*, 2001).

During the effector stage, the siRNAs assemble into endoribonuclease-containing complexes known as RNA-induced silencing complexes (RISCs). siRNAs undergo unwinding before being incorporated into a high-molecular-weight protein complex called RISC (Hammond *et al.*, 2000). Dicers are part of the RISC complex, which includes several different proteins such as the Argonaute gene family members and an

ATP-dependant RNA helicase activity that unwinds the two strands of RNA. Functional RISCs contain only single stranded siRNA (Martinez *et al.*, 2002). The siRNA strands subsequently guide the RISC to complementary RNA molecules, where base pairing takes place between the antisense strand of the siRNA and the sense strand of the target mRNA. This leads to endonuclease cleavage of the target RNA (Novina and Sharp 2004). Gene silencing by RISC is accomplished via homology-dependent mRNA degradation (Tuschl *et al.*, 1999; Hamilton *et al.*, 2002), translational repression (Grishok *et al.*, 2001) or transcriptional gene silencing (Pal-Bhadra, Bhadra and Birchler 2002) (Fig 1.9.1.1).

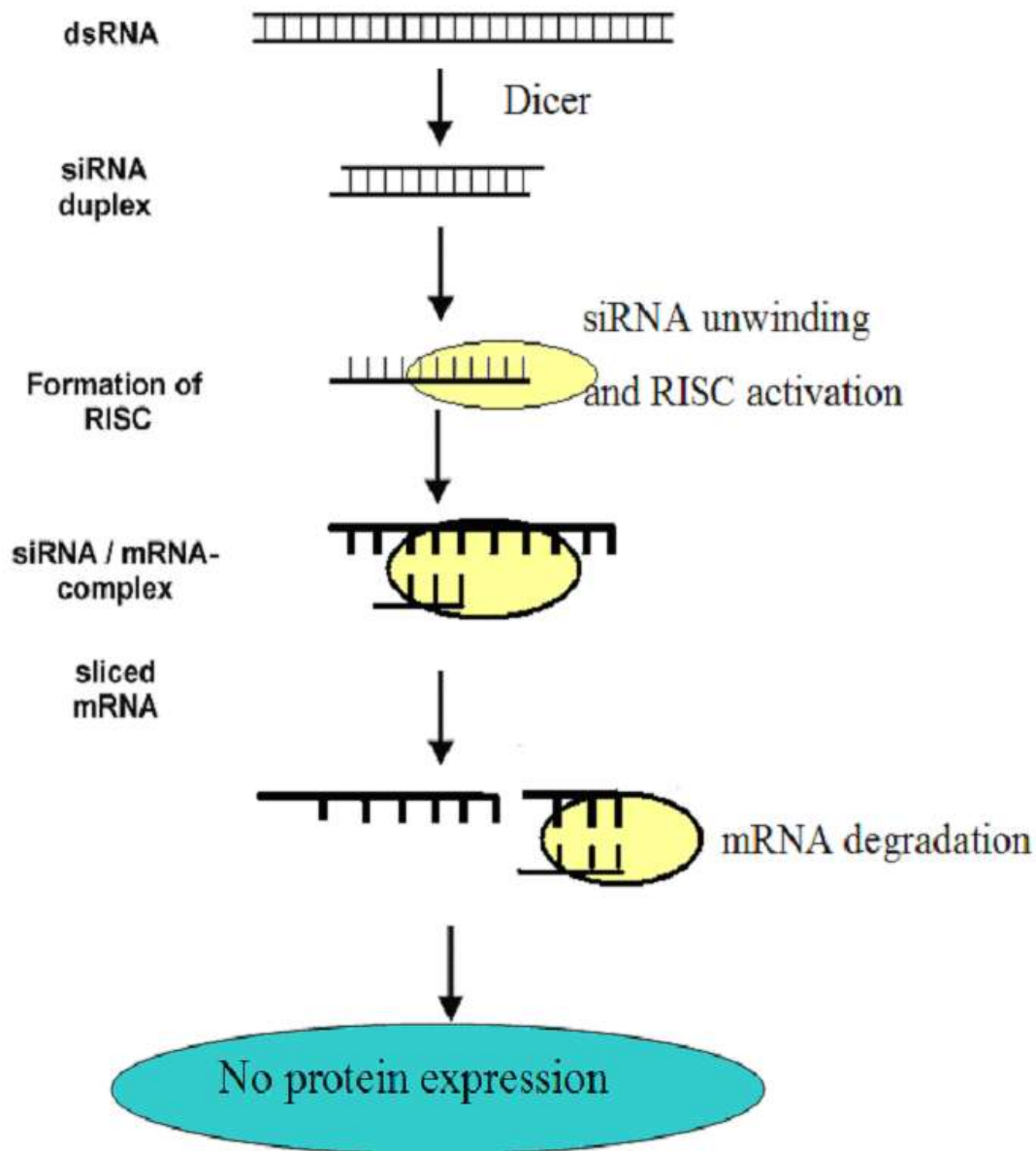


Fig 1.9.1.1: siRNA mechanism of action. dsRNAs are processed by a host Dicer enzyme to form siRNAs. Dicer-processed siRNAs and synthetic siRNAs undergo ATP dependent unwinding before being incorporated into a protein complex called RISC (RNA-induced silencing complex) that contains single stranded siRNAs. The RISC is reconfigured to active RISC which contains the proteins required for cleaving the target mRNA at the point where the antisense siRNA binds. After the cleavage the active RISC is released to cleave additional mRNA molecules whereas the cleaved mRNA is degraded by cellular ribonucleases.

1.10 Basal cell carcinoma

Basal cell carcinoma (BCC) is the most common malignancy in the white-skinned population with an estimated 750,000 cases per year in the US, with 175 cases per 100,000 reported in American men (Chuang *et al.*, 1990). Although not lethal, tumors are locally invasive with disfiguring growth in surrounding tissues causing morbidity due to prevalent localization of tumors in facial skin. Accordingly, this disease typically has a favorable prognosis, as complete surgical excision is almost always curative (Walling *et al.*, 2004).

BCC is a slowly growing tumor occurring in hair-growing squamous epithelium. The transformation of basal stem cells located in the hair follicles or basal epidermis gives rise to BCC (Backvall *et al.*, 2005). Ultraviolet (UV) radiation is considered the main carcinogen (Corona *et al.*, 2001; Krickler *et al.*, 1995) and approximately 80% of the tumors occur on the head and neck. If detected early, it can be treated and cured without serious side effects (Dua *et al.*, 2004).

Two main approaches to classify BCC have been suggested based on histopathological growth pattern and histological differentiation. To date, no universally agreed classification exists and it is regarded that classification based on growth pattern has the greatest biological significance (Saldanha, Fletcher and Slater 2003). Several sub-types of BCC have been identified- nodular-ulcerated BCC, superficial BCC, sclerosing BCC, cystic BCC, linear BCC and micronodular BCC. BCC rarely metastasize with rates ranging from 0.003 to 0.55% (Kapucuoglu *et al.*, 2009; Walling *et al.*, 2004). Up to 85% of metastasis has the neck or head as the site of the primary tumor (von Domarus and Stevens 1984), with at least two-thirds of cases originating from the face (Snow *et al.*, 1994). The most frequent site of BCC metastasis is regional lymph nodes, followed by bone, lung, and liver (Snow *et al.*, 1994; Lo *et al.*, 1991; Martin *et al.*, 2000). Furthermore, people with BCC are at higher risks of developing further BCCs and other malignancies, including squamous cell carcinomas, malignant melanomas and possibly also non-cutaneous malignancies (Wong, Strange and Lear 2003).

Previous studies have indicated the role of Sonic Hedgehog (Shh) pathway, via Patched (PTCH) gene mutations (Johnson *et al.*, 1996, Gailani *et al.*, 1996), as a key cellular signaling event in BCC tumorigenesis. However, relatively little is known about the molecular events involved in this disease. A single study of BCC has been reported using a cDNA microarray representing 1,718 genes (Howell *et al.*, 2005). Immunohistochemical techniques were used to study the expression of several proteins including CD10 (Pham *et al.*, 2006; Yada *et al.*, 2004), p63 (Park *et al.*, 2004), low expression levels of CD44 (Baum *et al.*, 1996) to associate with the presence of BCC. Expression level of the Ki67 antigen differs in BCCs that recur and BCC that do not recur (Healy *et al.*, 1995).

1.11 Aims

The overall goal of this project was to study breast cancer with reference to gene expression analysis using clinical specimens and *in-vitro* models. The specific aims of this project were:

- To advance the understanding of the heterogeneity of breast cancer using gene expression analysis, and to identify gene expression differences among the Normal and Cancer breast tissue and comparing various clinical parameters such as ER status, Grade, LN status and Tumour size.
- To identify gene expression changes that may be linked to clinical outcomes such as relapse-free survival and overall survival.
- To compare our results with various other similar or related studies and identify precise genelists which may be linked to disease progression, relapse-free survival and overall survival.
- To compare our results with two of the FDA approved prognostic assays MammaPrint and OncotypeDX to identify common genes in both studies which may be of common diagnostic importance.
- To develop a model based on gene expression signatures to predict clinical outcomes for breast cancer patients using Back Propagation Neural Network algorithm
- To identify sets of genes whose expression correlate with ER status using gene expression data generated in-house & publicly-available clinical and cell line datasets
- To identify prognostically important genes from our microarray study and validate their functions in the laboratory using molecular biology techniques such as siRNA and cDNA transfection in cancer cell lines, in particular focussing on invasion and motility.
- To identify the representative nature of cell lines to clinical conditions using gene expression data.
- Basal Cell Carcinoma data analysis

2.0 Materials and Methods

2.1 Microarray data used in this study

The gene expression profiles of 104 tumors and 17 normal specimens were generated in-house as a starting point for this study. To complement this analysis, public datasets were also downloaded from GEO (Gene Expression Omnibus) for comparison; as outlined.

2.1.1 Breast cancer clinical microarray dataset generated at NICB

A total of 104 tumour specimens and 17 normal specimens were obtained from Dr. Susan Kennedy (Consultant Histopathologist, St. Vincent's University Hospital (SVUH), Dublin. The patients underwent potentially curative resection at the hospital and after pathological examination; the tumors were snap frozen in liquid nitrogen. The tumors were subsequently stored at $-70/-80^{\circ}\text{C}$ and were later processed onto microarray chips (see section 2.5.11) by Dr. Lorraine O'Driscoll and Dr. Padraig Doolan.

Clinical information was obtained from the hospital for all patients. The dataset contains information on the following clinical parameters for individual patients:

- Estrogen Receptor status
- Censored relapse free survival for 7 years.
- Type of cancer e.g. lobular, ductal
- Overall Relapsed status
- RIP (Event of death due to disease)
- Relapse within 5 years
- Survival for 5 years
- Age at diagnosis
- Tumour type
- Tamoxifen treatment status
- Chemotherapy status
- Tumour size
- Tumour grade
- Lymph Node Status

2.1.2 Data obtained from public repositories

Several published datasets relating to breast cancer were downloaded from the GEO (Gene Expression Omnibus) (<http://www.ncbi.nlm.nih.gov/geo/>), Array Express (<http://www.ebi.ac.uk/microarray-as/ae/>) as well as from independent sources (Table 2.1.2.1). Datasets from GEO carry a unique GEO ID and more information can be obtained by searching for the specified GEO number. For some experiments, gene expression values were available as raw data files, while for others they were available as processed data. For all the experiments for which raw data is available the data was processed using dChip algorithm. The summary of the experiments taken for analysis is depicted in Table 2.1.2.1.

Experiment	Chip type	No. of samples	Comments
GEO: GSE3156 (Bild <i>et al.</i> , 2006)	U133 Plus2.0	19	Breast Cell line
GEO: GSE3744 (Richardson <i>et al.</i> , 2006)	U133 Plus2.0	47	Breast clinical specimens
GEO: GSE2034 (Wang <i>et al.</i> , 2005; Carroll <i>et al.</i> , 2006)	U133A	286	Breast clinical specimens
GEO: GSE2990 (Sotiriou <i>et al.</i> , 2006)	U133A	193	Breast clinical specimens
GEO: GSE4922 (Ivshina <i>et al.</i> , 2006)	U133A+B	347	Breast clinical specimens
GEO: GSE1456 (Pawitan <i>et al.</i> , 2005)	U133A+B	159	Breast clinical specimens
GEO: GSE4570 (Hoek <i>et al.</i> , 2004)	U133A	8	Melanoma cell lines
GEO: GSE4587 (Smith, Hoek and Becker 2005)	U133 Plus2.0	19	Melanoma clinical specimens
GEO: GSE5720 (Shankavaram <i>et al.</i> , 2007)	U133A+B	60	Cell lines of different origin
GEO: GSE1133 (Su <i>et al.</i> , 2004)	U133A	79	Various tissue
(van 't Veer <i>et al.</i> , 2002)	Hu25K	117	Breast clinical specimens
(Paik <i>et al.</i> , 2004) (genes taken from paper)	PCR-based assay	668	Breast clinical specimens
Array Express E-TABM-185	U133A	5897	Various tissue and cell lines

Table 2.1.2.1: Summary of individual experiments included in study

2.1.3 Basal cell carcinoma cancer clinical microarray dataset generated at NICB

Tissue specimens from twenty cases of BCC were procured at the Blackrock Clinic and the Bons Secours Hospitals, Dublin, were examined macroscopically, immediately snap-frozen in liquid nitrogen, and were subsequently stored at -80°C until required for analysis. Five normal skin specimens (from consenting male and female volunteers of a similar age range who never had skin cancer) were also included in these studies. Following this microarray analysis was performed (see section 2.5.11) by Dr. Lorraine O'Driscoll and Dr. Pdraig Doolan using Affymetrix U133plus chips

2.2.1 Normalization and Quantification

For experiments where raw data files were available, normalisation using the dChip algorithm (www.dchip.org) was carried out. DNA-Chip Analyzer (dChip) is a software package implementing model-based expression analysis of oligonucleotide arrays and several high-level analysis procedures. This model-based approach allowed probe-level analysis on multiple arrays. In this normalisation procedure, an array with median overall intensity is chosen as the baseline array against which other arrays are normalised at probe level intensity. Subsequently a subset of PM (“perfect match”) probes, with small within-subset rank difference in the two arrays (also known as invariant set), serves as the basis for fitting a normalisation curve. PM (“perfect match”) is the exact match is a section of the mRNA sequence whereas MM (“mismatch”) is identical except for one base difference from its exact match counterpart.

2.2.2 Quality inspection

The following Quality Control measures are reported by dChip

- Median Intensity: This is the middle intensity (when all chip probe intensities are ordered from low to high intensity) of the un-normalized probe values. Normalization process brings the median intensity to a comparable level.
- P (Present) call %: Calls indicate if the transcript is expressed or not. It can be ‘P’ for “Present”, ‘M’ for “Marginal” and ‘A’ for “Absent”. Total P calls in an array

can vary widely based on the different nature of samples, but usually range from 40-60% for a cell line experiment and 25-35% for clinical tissue.

- % Signal outlier: The signal value greater than 80th percentile multiplied by 3 is taken as signal outlier and is represented as % signal outlier.
- % Array outlier: The array-outliers are the arrays whose probe pattern for selected probe sets are different from the consensus probe patterns seen in most arrays. If the array outlier value increases above 5%, the chip is marked as an outlier chip and is marked by ‘*’; indicating potential image contamination or sample hybridization problem of that array.

Manual inspection: Apart from the above parameters manual inspection was also performed to estimate the quality of individual chips. An example of a good and bad quality chip is shown overleaf (Fig 2.2.2.1 and Fig 2.2.2.2).

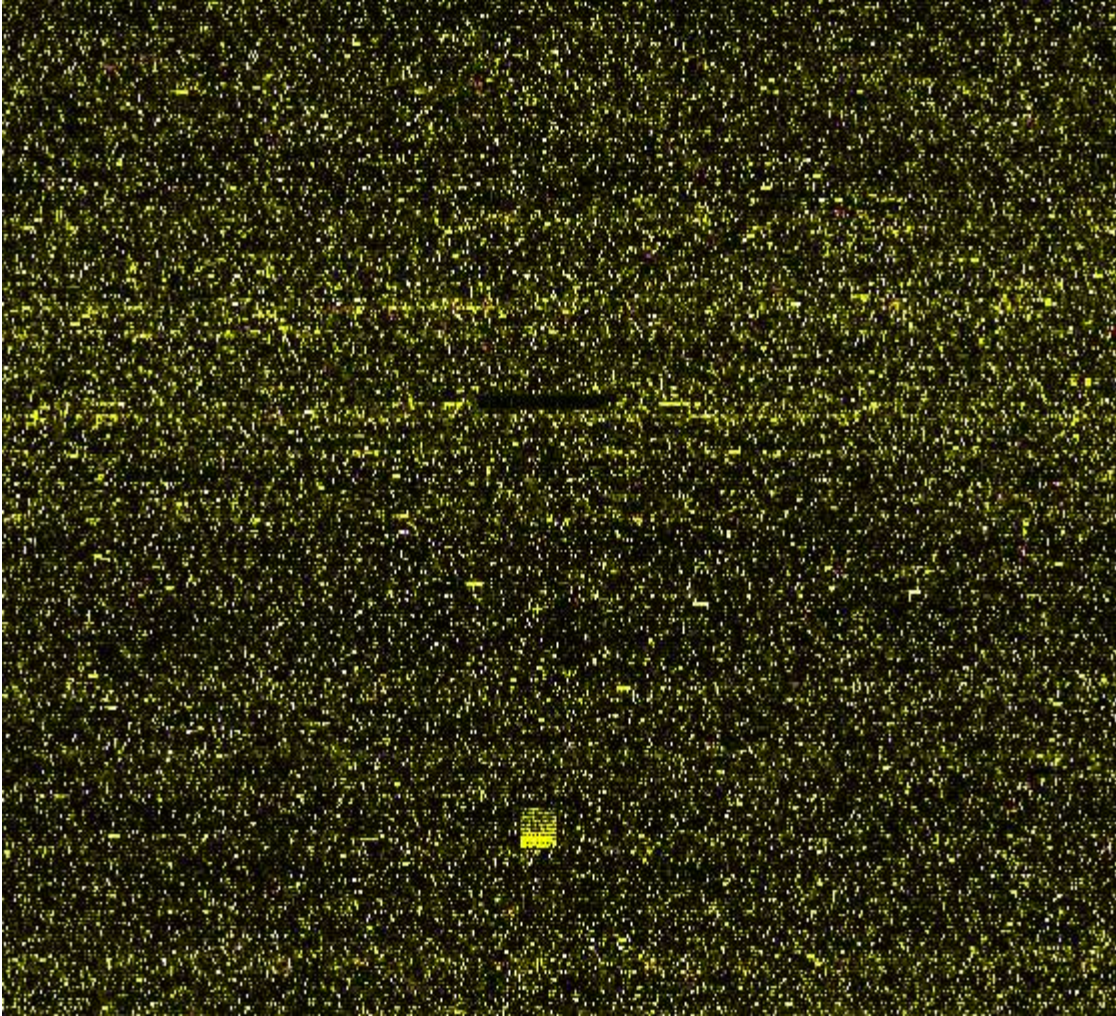


Fig 2.2.2.1: An example of a good quality scan image (image simulated by dChip)



Fig 2.2.2.2: A chip with very high background (image simulated by dChip)

2.2.3 Standard deviation filtration

This filtration was applied as a prerequisite to Hierarchical clustering. The aim of this filtration was to remove genes which had a standard deviation divided by mean i) less than 1 across samples or ii) more than 1000 across samples. This process removed genes which i) did not fluctuate significantly across samples ii) fluctuated too highly across samples to be prognostically valuable

2.2.4 Hierarchical Clustering

Hierarchical clustering is a mathematical technique whereby the analysed samples/genes are connected iteratively based on their similarity. Samples/genes with similar expression patterns are grouped together and are connected by a series of branches, which is called a

dendrogram (or clustering tree). Hierarchical clustering was used to see how well samples clustered together, identify any sub-groups in the samples and identify correlated genes. Cluster analysis was also used to identify significant clinical parameter (enrichment analysis) associated with the cluster. The observed and expected clinical parameter was calculated for each cluster and the whole data and a hyper geometric distribution was used to calculate the p-value for individual clinical parameter in each cluster. Similar analysis was performed for clusters of genes enriched for particular Gene Ontology and Pathways. Genes in the cluster were compared to genes not in cluster to find clusters enriched with gene belonging to a particular gene ontology or pathway.

2.2.5 Finding significant genes

The following criterion was used to generate genelists.

2.2.5.1 Fold change

Fold change is the ratio of the mean of the experimental group to that of the baseline. It's a metric to define the gene's mRNA-expression level between two distinct experimental conditions.

2.2.5.2 Difference

The difference of Affymetrix expression units (gene expression values obtained after dChip processing) was also incorporated for finding differentially regulated genes.

2.2.5.3 T-test

The t-test assesses whether the means of two groups are statistically different from each other. The t-statistic is calculated as follows:

$$t = (X_T - X_C) / \sqrt{(\text{var}_T / n_T + \text{var}_C / n_C)}$$

X_T \rightarrow mean of Treatment samples

X_C \rightarrow mean of Control samples

$\text{var}_T \rightarrow$ variance of Treatment samples

$\text{var}_C \rightarrow$ variance of Control samples

$n_T \rightarrow$ number of Treatment samples

$n_C \rightarrow$ number of Control samples

Subsequently, the p-value is calculated from the t-test.

The purpose of the t-test is to evaluate the null hypothesis that there is no difference between the means of two samples. The t-test is a parametric test which is used to analyse the mean and standard deviation of two or more groups of samples based on a number of underlying assumptions, including a normal distribution of the data within the test.

Therefore, hypothesis testing facilitates the calculation of the probability of the observed value of the t-statistic occurring based on the assumption that the null hypothesis is true.

For calculation of the probability, the data is assumed to be normally distributed. By convention, a p-value of ≤ 0.05 is usually considered sufficient to reject the null hypothesis, i.e. that there is a real difference between the means (≤ 0.01 would be considered strong evidence) (Stekel, 2003).

For gene ontology and pathway analysis, the filtration criteria used was as follows: $FC > 2$, Difference of means > 100 and p-value ≤ 0.05 . For developing the MLPERCEP classifier (see section 3.5), a p-value ≤ 0.001 was used. For gene list generation purposes, $FC > 1.2$, Difference of means > 100 and p-value ≤ 0.05 was utilised.

2.2.6 Identifier conversion

NetAffx from Affymetrix and David and Ease was used for gene identifier conversion. This was essential wherever microarray genelists from different platforms were compared.

2.2.6.1 NetAffx

NetAffx is provided by Affymetrix (www.affymetrix.com) and provides detailed annotation of its probe sets on various chips. Individual and batch query was used at various places to convert the Affymetrix identifier to a different identifier or *vice versa*.

2.2.6.2 David and Ease

David and Ease (<http://david.abcc.ncifcrf.gov/>) is an online tool for gene identifier conversion and was used at many places for converting the gene identifier.

2.2.7 Gene list comparison

Microsoft Access and Venny were used to compare various genelists.

2.2.7.1 Microsoft access

MSAccess is a database-building package that was used to compare different gene lists. MSAccess allowed comparison of like genes across multiple lists. It allowed comparison of genes and also relevant information such as probe sets, difference of means and p-values. It was also used as a repository and used for database queries. It was also widely used for merging tables and adding annotations to genelists.

2.2.7.2 Venny

Venny (<http://bioinfogp.cnb.csic.es/tools/venny/index.html>) is a Venn diagram-drawing software tool that was used to overlap and compare genelists using Venn diagrams.

2.2.8 GenMAPP

GenMAPP (<http://www.genmapp.org>) is a computer application designed to visualize gene expression and other genomic data on maps representing biological pathways and groupings of genes. It overlays gene-expression data on the pathways incorporating colour-coding according to user-defined parameters. Additionally, the MappFinder module identifies significant Gene ontologies and pathways affected by the submitted genelists.

MappFinder

MAPPFinder is an accessory program that works with GenMAPP and the annotations from the Gene Ontology (GO) consortium to identify significant GO and MAPPs. The calculations made by MAPPFinder (Fig 2.2.8.1) are intended to give an idea of the relative amount of genes meeting the criterion that are present in each GO term or Local MAPP.

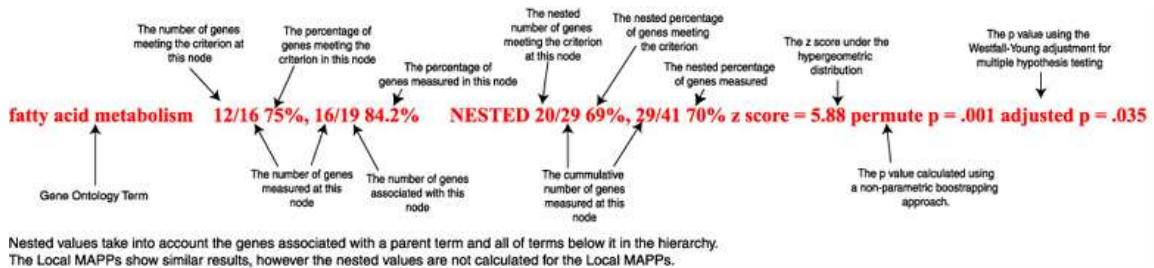


Fig 2.2.8.1: Calculations made by GenMAPP to calculate the significant of individual GO and MAPPs. (Obtained from GenMAPP website)

Genes meeting the criterion: The number of distinct genes that met the user-defined criterion in the Expression Dataset. This may also be referred to as "genes changed."

Genes measured: The number of distinct genes in the submitted expression dataset that were found to link to this GO term or MAPP.

Genes associated with this GO term or MAPP: The number of genes assigned to this GO term or on this MAPP. Also referred to as the number of "Genes in GO" for a specific term.

% genes meeting the criterion: $\text{Genes meeting the criterion} / \text{genes measured} * 100$

% genes measured: $\text{Genes measured} / \text{genes associated} * 100$

Nested numbers: The same 5 calculations are repeated, but as nested numbers.

Z score: The standard statistical test under the hypergeometric distribution (Fig 2.2.8.2).

$$zscore = \frac{(r - n \frac{R}{N})}{\sqrt{n(\frac{R}{N})(1 - \frac{R}{N})(1 - \frac{n-1}{N-1})}}$$

Fig 2.2.8.2: Z-score calculation. (Obtained from GenMAPP website)

Where N is the total number of genes measured, R is the total number of genes meeting the criterion, n is the total number of genes in this specific MAPP, and r is the number of genes meeting the criterion in this specific MAPP. A positive Z score indicates that there are more genes meeting the criterion in a GO term/MAPP than would be expected by random chance. A negative Z score indicates that there are fewer genes meeting the criterion than would be expected by random chance.

Hypergeometric distribution: The hypergeometric distribution is a discrete probability distribution that quantifies the number of successes in a sequence of n draws from a finite population without replacement. An example of the hypergeometric distribution is an urn with some red marbles and some black marbles and we have knowledge of the ratio of them. A handful of marbles is taken and analysed for significant difference between the ratio of red to black in the sample and the total population in the urn.

Permute P and Adjusted P: p-value is calculated based on the Z score and the hypergeometric distribution. A p-value of 0 indicates a value < 0.001.

2.2.9 Genesis

Genesis is comprehensive software for microarray data analysis. It is available at <http://genome.tugraz.at/>. This software was used to perform k-means clustering (see section 1.6.2.2) and principal component analysis (see section 1.6.3). K-means clustering is a mathematical technique where the similar experiments/genes are grouped together. Principal component analysis is a mathematical technique to reduce the dimensionality of the data, thus giving a deeper insight into hidden patterns that influence the level of variation within the dataset.

2.2.10 Dev C++

Dev C++ is an environment and compiler code used to write and execute C and C++ programs. This software is available at <http://www.bloodshed.net/devcpp.html> MLPERCEP was developed using Dev C++.

2.2.11 C# (C-sharp)

Borland C# was used to build user interface for the MLPERCEP programs. The software is available at <http://www.codegear.com/products/bds2006>.

2.2.12 Kaplan-Meier survival function

The Kaplan-Meier estimate (KM) (Kaplan and Meier 1958) estimates the survival function from life-time data. In life science research, it is used to compare two groups of patients or treatments for differences in survival. Kaplan Meier curves represent the proportion of the study population surviving at successive times. Kaplan-Meier curves for the parameters of interest and the outcomes are represented on the graph and the p-value is used to determine the likelihood that there is no difference between the two survival curves.

Kaplan-Meier plots of the estimate of the survival function as a series of steps of reducing magnitude. The X-axis normally depicts the time of survival and the y-axis represents the percent of patients surviving. The algorithm takes account of censored data; loss of part of the sample before the final outcome is observed, e.g. patients leaving the study or patients dying due to different causes before the study is completed. The survival functions are compared for significant differences using Chi-squared statistics.

SPSS (<http://www.spss.com/>) software was used for performing Kaplan-Meier analysis which was performed by Dr. Lorraine O'Driscoll, NICB.

2.2.13 CLUSTALW

CLUSTALW (<http://www.ebi.ac.uk/Tools/clustalw2/index.html>) was used to performing multiple sequence alignment. CLUSTALW is an online tool to perform sequence alignment.

2.2.14 BLAST

BLAST was used to search for homologues sequences.

(<http://blast.ncbi.nlm.nih.gov/Blast.cgi>)

2.2.15 Non-parametric analysis

Non-parametric analysis was performed using MeV (<http://www.tm4.org/>). The p-value was calculated using Wilcoxon test. Wilcoxon rank-sum test is a nonparametric test similar to the two sample t-test and is based on the rank order in which the observations from the two samples fall. The test is based upon ranking the two sets of sample observations as a combined ranking. Each observation therefore has a rank; the smallest has rank 1 and so on. The Wilcoxon rank-sum test statistic is the sum of the ranks for observations from one of the samples. The genes were termed as significant if $FC > 1.2$, Difference of means > 100 and p-value ≤ 0.05 .

2.3 Cell Culture Methods

2.3.1 Water

Ultra high pure water (UHP) was used in the preparation of all media and solutions. Pre-treatment of water, involving activated carbon, pre-filtration and anti-scaling was first carried out. This water was then purified by a reverse osmosis system (Millipore Milli-RO 10 Plus, Elgastat UHP), which is low in organic salts, organic matter, colloids and bacteria with a standard of 12 - 18 M Ω /cm resistance.

2.3.2 Treatment of Glassware

All solutions for use in cell culture and maintenance were prepared and stored in sterile glass bottles. Bottles, lids and all other glassware used for any cell-related work were prepared as follows: all glassware and lids were soaked in a 2% (v/v) solution of RBS-25 (AGB Scientific, 83460) for at least 1hrs. This is a deproteinising agent, which removes proteineous material from the bottles. Glassware was scrubbed and rinsed several times in tap water; the bottles were then washed by machine using Neodisher detergent, an organic, phosphate-based acid detergent. The bottles were then rinsed twice with distilled water, once with UHP water and sterilised by autoclaving.

2.3.3 Sterilisation

Water, glassware and all thermostable solutions were sterilised by autoclaving at 121°C for 20 min under 15 p.s.i. pressures. Thermolabile solutions were filtered through a 0.22 µm sterile filter (Millipore, millex-gv, SLGV-025BS). Low protein-binding filters were used for all protein-containing solutions. Acrodisc (Pall Gelman Laboratory, C4187) 0.8/0.2 µm filters were used for non-serum/protein solutions.

2.3.4 Media Preparation

Medium was routinely prepared and sterility checked by Mr. Joe Carey (technician) as in SOP NCTCC 003-02. 10X media were added to sterile UHP water buffered with HEPES (N- [2-Hydroxyethyl]-N'- [2-ethanesulphonic acid]) (Sigma, H-9136) and NaHCO₃ (BDH, 30151) and adjusted to a pH of 7.45 - 7.55 using sterile 1.5M NaOH and 1.5M HCl. The media were then filtered through sterile 0.22 µm bell filters (Gelman, 121-58) and stored in 500 ml sterile bottles at 4°C.

The basal media were stored at 4°C up to their expiry dates as specified on each individual 10X medium container. Working stocks of culture media were prepared as 100 ml aliquots and supplemented as required. These were stored for up to 3 weeks at 4°C; after this time, fresh culture medium was prepared.

2.4 Maintenance of cell lines

2.4.1 Safety Precautions

All cell culture work was carried out in a class II down-flow re-circulating laminar flow cabinet (Nuair Biological Cabinet) and any work which involved toxic compounds was carried out in a cytoguard (Gelman). Strict aseptic techniques were adhered to at all times. The laminar flow cabinet was swabbed with 70% industrial methylated spirits (IMS) before and after use, as were all items used in the cabinet. Each cell line (including low and high passage cells) was assigned specific media and waste bottles and only one cell line was used at a time in the cabinet, which, was allowed to clear for 15 min between different cell lines. The cabinet and incubators were cleaned each week with industrial detergents (Virkon, Antec. International; TEGO, TH. Goldschmidt Ltd.). A separate Laboratory coat was kept for aseptic work and gloves were worn at all times during cell work.

2.4.2 Culture of Adherent Cell Lines

The cell lines used during the course of this study, their sources and their basal media requirements are listed in Table 2.4.2.1. Cell lines were generally maintained in 25 cm² (Costar, 3056) and 75 cm² flasks (Costar, 3376) and fed every two to three days.

Cell Line	Media	Cell Type
MDA-MB-435S	RPMI with 10% FCS	Breast/Melanoma
M14	RPMI with 10% FCS	Melanoma
MDA-MB-231	RPMI with 10% FCS	Breast

Table 2.4.2.1: Cell Lines used in this study

MDA-MB-435s was earlier thought to be a breast cell line, but recent analysis using gene expression and clustering of this cell line with melanoma cell lines indicates the origin of this cell line to be melanoma (Rae *et al.*, 2007; Ellison *et al.*, 2002).

2.4.2.1 Subculture of Adherent Cell Lines

Prior to subculture cells were always monitored for any contamination and were only sub-cultured when the cells were 70-80% confluent. During routine sub-culturing or harvesting of adherent cell lines, cells were removed from their flasks by enzymatic detachment.

Medium were emptied from cell culture flasks and rinsed with a pre-warmed (37°C) trypsin/EDTA (Trypsin Versene - TV) solution (0.25% trypsin (Gibco, 25090-028), 0.01% EDTA (Sigma, E-5134) solution in PBS (Oxoid, BR14a)). The purpose of this was to inhibit any naturally occurring trypsin inhibitor which would be present in residual serum. Fresh TV was then placed on the cells (2ml/25cm² flask or 3ml/75cm² flask) and the flasks incubated at 37°C until the cells were detached (5-10 min). The flasks were struck once, roughly, to ensure total cell detachment. The trypsin was deactivated by addition of an equal volume of growth medium (*i.e.* containing 10% serum). The entire solution was transferred to a 20ml sterile universal tube (Greiner, 201151) and centrifuged at 1,000 rpm for 5 min. The resulting cell pellet was resuspended in pre-warmed (37°C) fresh growth medium, counted (see section 2.4.3) and used to re-seed a flask at the required cell density or to set up an assay.

2.4.3 Cell Counting

Sample of this mixture was applied to the chamber of a haemocytometer over which a glass cover slip had been placed. Cells in the 16 squares of the four outer corner grids of the chamber were counted microscopically. An average per corner grid was calculated with the dilution factor being taken into account. Final cell numbers were multiplied by 10⁴ to determine the number of cells per ml (volume occupied by sample in chamber is 0.1cm x 0.1cm x 0.01cm *i.e.* 0.0001cm³; therefore cell number x 10⁴ is equivalent to cells per ml).

2.4.4 Cell freezing

Cryoprotective medium or freezing medium was prepared in complete culture medium containing 10% dimethylsulfoxide (DMSO) (Sigma, D-5879) and filter sterilised using

0.22 µm filter and syringe. The appropriate number of cryogenic vials (Greiner, 122 278) were labelled with the cell line, passage no and date. Cells were trypsinised as outlined previously (see section 2.4.2.1). The supernatant from the centrifuged cells were removed and resuspend the cell pellet in 1 ml of fresh media. Freezing medium (10% DMSO) was slowly added drop wise to the cell suspension to give a final concentration to 5% of DMSO, and a final cell concentration of $5 \times 10^6 - 1 \times 10^7$ cells/ml. This step was very important, as DMSO is toxic to cells. When added slowly, the cells had a period of time to adapt to the presence of the DMSO, otherwise cells may have lysed. 1.5-1.8 mls of the DMSO-containing cell suspension was then added to each of the vials. The cryovials were which were quickly placed at -80°C . To allow long term storage of cell stocks, cells were frozen and cryo-preserved in liquid nitrogen at temperatures below -180°C .

2.4.5 Cell Thawing

Prior to the removal of a cryovial from the liquid nitrogen stores for thawing, a sterile universal tube containing growth medium was prepared for the rapid transfer and dilution of thawed cells to reduce their exposure time to the DMSO freezing solution which is toxic at room temperature. The cryovial was removed and thawed quickly by rubbing by hand. When almost fully thawed, the DMSO-cell suspension was quickly transferred to the media-containing universal and centrifuged at 1,000 rpm for 5 min. the DMSO-containing supernatant removed and the pellet re-suspended in fresh growth medium. Thawed cells were then placed into 25cm^2 tissue culture flasks with 5mls of the appropriate type of medium and allowed to attach overnight. After 24hrs, the cells were re-fed with fresh medium to remove any residual traces of DMSO.

2.4.6 Sterility Checks

Sterility checks were routinely carried out on all media, supplements and TV used for cell culture. Samples of basal media were inoculated into Columbia blood agar plates (Oxoid, CM331), Sabouraud dextrose (Oxoid, CM217) and Thioglycollate broth (Oxoid, CM173) which when combined detect most contaminants including bacteria, fungus and yeast. Growth media (*i.e.* supplemented with serum) were sterility checked at least 3 days prior to use by incubating samples at 37°C . These were subsequently examined for turbidity

and other indications of contamination. Freshly thawed cells were also subjected to sterility checks.

2.4.7 Mycoplasma Analysis

Mycoplasma examinations were carried out routinely (at least every 3 months) on all cell lines used in this study. This analysis was performed by Michael Henry and Shane Kelly at the National Institute for Cellular Biotechnology (NICB).

2.4.7.1 Indirect Staining Procedure

In this procedure, Mycoplasma-negative NRK cells (a normal rat kidney fibroblast line) were used as indicator cells and incubated with supernatant from test cell lines to test for Mycoplasma contamination. NRK cells were used for this procedure because cell integrity is well maintained during fixation. A fluorescent Hoechst stain was utilised which binds specifically to DNA and so will stain the nucleus of the cells in addition to any Mycoplasma DNA present. A Mycoplasma infection would thus be seen as small fluorescent bodies in the cytoplasm of the NRK cells and occasionally outside the cells.

NRK cells were seeded onto sterile cover slips in sterile Petri dishes (Greiner, 633 185) at a cell density of 2×10^3 cells per ml and were allowed to attach overnight at 37°C in a 5% CO₂ humidified incubator. 1 ml of cell-free supernatant (cleared by centrifugation at 1,000 rpm for 5 min) from each test cell line was then inoculated onto a NRK cover slip and incubated as before until the cells reached 20-50% confluency (4-5 days). After this time, the waste medium was removed from the Petri dish; the cover slips (Chance Propper, 22 x 22 mm) were washed twice with sterile PBS, once with a cold PBS/Carnoy's (50/50) solution and fixed with 2 ml of Carnoy's solution (acetic acid: methanol - 1:3) for 10 min. The fixative was then removed and after air-drying, the cover slips were washed twice in deionised water and stained with 2 ml of Hoechst 33258 dye (BDH) (50 ng/ml) for 10 min.

From this point on, work proceeded in the dark to limit quenching of the fluorescent stain. The cover slips were rinsed three times in PBS. They were then mounted in 50%

(v/v) glycerol in 0.05 M citric acid and 0.1 M disodium phosphate and examined using a fluorescence microscope with a UV (ultraviolet) filter.

Prior to removing a sample for Mycoplasma analysis, cells were be passaged a minimum of 3 times after thawing to facilitate the detection of low-level infection. Optimum conditions for harvesting supernatant for analysis occur when the culture is in log-phase near confluency and the medium has not been renewed in 2-3 days.

2.5 Analytical Techniques

2.5.1 Preparation of total RNA from cells using RNeasy Mini Prep Kit

High quality RNA was isolated from cells using the RNeasy mini-kit (Qiagen, 74104). Cell pellets for RNA extraction (stored at -80°C) were re-suspended in 1.2 ml of buffer RLT (supplemented with $10\mu\text{l/ml}$ of β -mercaptoethanol) and vortexed to loosen the pellets. The samples were completely homogenised by passing the lysate at least 5 times through a blunt 20-gauge needle (0.9 mm diameter) fitted to an RNase-free syringe. One volume (1.2 ml) of 70% ethanol was added to the homogenised samples and mixed well by pipetting. This mixture was then loaded in $700\mu\text{l}$ aliquots on to an RNeasy mini column, which was placed in a collection tube and centrifuged at $8,000 \times g$ for 15 sec (this was continued until the entire mixture had been passed through the column). Once all the homogenised cells had been passed through the column, the washes were carried out. Initially $700\mu\text{l}$ RW1 was loaded on to the column and centrifuged at $8,000 \times g$ for 15 sec. This was closely followed by 2 washes in buffer RPE (also followed by centrifuging at $8,000\text{ rpm}$ for 15 sec). To completely dry the spin column, it was placed in a fresh collection tube and centrifuged at full speed for 1 min. The RNA was eluted by passing two lots of $25\mu\text{l}$ RNase free water (supplied) through the column by centrifuging it at $8,000\text{ rpm}$ for 1 min. The eluted RNA was then quantified (see section 2.5.2).

2.5.2 RNA Quantification using NanoDrop

The NanoDrop ND-1000 is a full-spectrum (220-750nm) spectrophotometer that measures $1\mu\text{l}$ samples with high accuracy and reproducibility. It uses a sample retention technology that relies on surface tension alone to hold the sample in place eliminating the

need for cuvettes and other sample containment devices. In addition, the NanoDrop has the capability to measure highly concentrated samples without dilution (50X higher concentration than the samples measured by a standard cuvette spectrophotometer).

To quantify an RNA sample, 1 μ l of the sample is pipetted onto the end of a fibre optic cable (the receiving fibre, Fig 2.5.2.1 (A)). A second fibre optic cable (the source fibre, Fig 2.5.2.1 (B)) is then brought into contact with the liquid sample causing the liquid to bridge the gap between the fibre optic ends. The gap is controlled to a 1mm path (Fig 2.5.2.1 (C)). A pulsed xenon flash lamp provides the light source and a spectrometer utilising a linear CCD array is used to analyse the light after passing through the sample. The instrument is controlled by special software run from a computer, and the data is logged in an archive file on the computer.

When measurement of the sample is complete, the sample can be simply wiped away using a soft laboratory wipe. This is sufficient to prevent sample carryover because each measurement pedestal is a highly polished end of a fibre optic cable, with no cracks or crevices for leftover sample to reside.



Fig 2.5.2.1: Samples are quantified by loading 1 μ l onto the receiving fibre (A), the source fibre, connected to the sampling arm (B) is brought down into contact with the sample allowing a 1mm gap between the upper and lower pedestal (C), through which the light is passed. (Pictures adapted from ND-1000 Spectrophotometer users manual V 3.1.0).

RNA (like DNA) was quantified using an ND-1000 spectrophotometer. The ND-1000 software automatically calculated the quantity of RNA in the samples using the OD₂₆₀.

i.e. $OD_{260} \times 40 \times \text{Dilution Factor}/1000 = \text{RNA content } (\mu\text{g}/\mu\text{l})$

The software simultaneously measured the OD_{280} of the samples allowing the purity of the sample to be estimated.

$$\text{Purity} = OD_{260}/OD_{280}$$

This was typically in the range of 1.8-2.0. A ratio of <1.6 indicated that the RNA may not be fully in solution. The RNA was diluted to $1\mu\text{g}/\mu\text{l}$ stocks for the subsequent reverse transcription (RT) protocol (see section 2.5.4).

2.5.3 RNA amplification, labelling and fragmentation of cRNA in preparation for hybridisation to Affymetrix array chips

Components required for this protocol were included in the Two-Cycle Target Labelling and Control Reagents (Affymetrix, P/N 900494) and MEGAscript High Yield Transcription Kit, Ambion Inc, P/N 1334 with the exception of Ethanol (Sigma, E7023).

The positive control Poly-A RNA is firstly diluted before spiking in with the sample RNA. Affymetrix supply a Eukaryotic Poly-A RNA Control Kit along with the Two-Cycle Target Labelling and Control Reagents. The kit is designed specifically to provide exogenous positive controls to monitor the entire GeneChip eukaryotic target labelling process. It is important to note that the Poly-A spikes were made up in non-stick RNase/DNase free tubes (Ambion cat no. 12450(1.5ml)/ 12350(0.5ml)), which prevents the Poly-A spikes from sticking to the sides of the tubes and interfering with the final concentration of the positive controls.

The tube with first strand cDNA synthesis master mix (Table 2.5.3.1) was then flicked and centrifuged briefly. $2\mu\text{l}$ of this mix was added to $2\mu\text{l}$ of the $50\text{ng}/\mu\text{l}$ RNA sample. The tubes were flicked and centrifuged briefly before being incubated for 6 min at 70°C . They were then incubated for 2 min on ice and centrifuged briefly.

Reagents	Amount
Poly-A RNA Control	2 μ l
T7-Oligo (dT) Primer (50 μ M)	2 μ l
RNA + RNase-free water	16 μ l

Table 2.5.3.1: First-Strand cDNA Synthesis

The first cycle, first strand master mix was prepared as in Table 2.5.3.2. It is worth noting that Affymetrix suggest that if there are more than 2 samples that it is prudent to include extra to compensate for potential pipetting inaccuracy or solution lost during the process. The 5 μ l of first strand master mix was added to each sample, the tube gently flicked, briefly centrifuged and placed immediately at 42°C for 1hr and 72°C for 10 min before being placed on ice for 2 min.

Reagents	Amount
5X 1st strand buffer	2 μ l
DTT (0.1M)	1 μ l
dNTP (10mM)	0.5 μ l
RNase Inhibitor	0.5 μ l
Superscript II	1 μ l
Total	7 μ l

Table 2.5.3.2: 1st strand master mix

The first cycle second strand master mix was prepared by adding following reagents (Table 2.5.3.3). This 10 μ l mix was then added to each tube which were then flicked and centrifuged briefly before being placed at 16°C for 2hrs and 75°C for 10 min and then ice for 2 min.

Reagents	Amount
RNase-free Water	4.8µl
dNTP (10mM)	0.4µl
MgCl ₂ (17.5mM)	4µl
E. coli DNA Polymerase	0.6µl
RNase H	0.2µl
Total Volume	10µl

Table 2.5.3.3: Second-strand master mix

The components for the first cycle IVT amplification (Ambion Megascript T7 kit) were assembled at room temperature. 5µl of each of the components ATP, CTP, UTP, GTP, enzyme mix and 10 x reaction buffers were added together for each sample included before being added to each sample. The tubes were gently flicked, centrifuged and placed at 37°C for 16hrs.

Reagents	Amount
10X Reaction Buffer	5µl
CTP Solution	5µl
GTP Solution	5µl
10X Reaction Buffer	5µl
CTP Solution	5µl
GTP Solution	5µl
Total Volume	30µl

Table 2.5.3.4: First-Cycle, IVT Master Mix

The cRNA was then purified using the GeneChip Sample Cleanup module (Affymetrix, 900371) as recommended by the manufacturers instructions.

The quantity of the cRNA was subsequently checked by diluting 2µl of cRNA in 18µl of

H₂O and reading the quantity by using the NanoDrop (see section 2.5.2). 600ng is required for the second cycle of the protocol. 2µl of freshly diluted random primers (3µg/µl) was added to each sample and the tubes were flicked, centrifuged briefly and placed at 70°C for 10 min before being placed on ice for 2 min.

The second cycle first strand mix was prepared by adding each reagent together as in Table 2.5.3.5 for each sample included. This 9µl was added to each sample before they were flicked, centrifuged and placed at 42°C for 1hr and ice for 2 min. After this, 1µl of RNase H was added to each sample before they were flicked, spun, and placed at 37°C for 20 min, 95°C for 5 min and ice for 2 min.

Reagents	Amount
5X 1st Strand Reaction Mix	4µl
DTT, 0.1M	2µl
RNase Inhibitor	1µl
dNTP (10mM)	1µl
SuperScript II	1µl
Total Volume	9µl

Table 2.5.3.5: Second-Cycle, First-Strand Master Mix

4µl of a freshly prepared aliquot of T7 Oligo dT primer was added to each sample before they were flicked, centrifuged briefly and incubated at 70°C for 6 min and ice for 2 min. The second cycle second strand master mix was prepared by adding the following in a tube for each sample required (Table 2.5.3.6). This 125µl master mix was added to each sample before being flicked, centrifuged briefly and incubated for 2 hrs at 16°C. T4 DNA polymerase (2µl) was then added to each sample before incubating at 16°C for a further 10 min. After incubating the samples at 4°C for 2 min, they were immediately purified using the GeneChip Sample Cleanup module (Affymetrix, 900371) following the manufacturers instructions.

Reagents	Amount
RNase-free Water	88µl
5X 2nd Strand Reaction Mix	30µl
dNTP, 10mM	3µl
E. coli DNA Polymerase I	4µl
Total Volume	125µl

Table 2.5.3.6: Second-Cycle, Second-Strand Master Mix

All 12µl of cDNA were used for the second IVT step. The reagents required for this step were assembled at room temperature. The master-mix included 8µl RNase- free water, 4µl of IVT labelling buffer, 12µl IVT labelling NTP mix and 4µl labelling enzyme mix for each sample included. The 28µl volume was added to each sample before flicking, centrifuging briefly and incubating at 37°C for 16 hrs.

The biotin-labelled cRNA was purified using the GeneChip Sample Cleanup module (Affymetrix, 900371), as recommended by the manufacturers, and quantified using a NanoDrop. For quantification of cRNA when using total RNA as starting material, an adjusted cRNA yield needed to be calculated to reflect carryover of unlabeled total RNA. Using an estimate of 100% carryover, the formula below was used to determine adjusted cRNA yield:

$$\text{Adjusted cRNA yield} = \text{RNAm} - (\text{total RNAi}) (y)$$

Where, RNAm = amount of cRNA measured after IVT (µg), total RNAi = starting amount of total RNA (µg), y = fraction of cDNA reaction used in IVT

The final step of the entire process was to fragment 20µg of the biotin-labelled cRNA by adding 8µl of fragmentation buffer to 20µg of cRNA and bringing the total volume of the reaction to 40µl, so that the concentration of the cRNA is 0.5µg/µl. This mix was incubated for 35 min at 94°C. From the fragmented cRNA 30µl (=15µg) was hybridised to the Affymetrix U133-plus-2 chip.

2.5.3.1 Probe Array Scan

After staining and washing, the chips were scanned using an Affymetrix GeneChip Scanner 3000 (Affymetrix, 00-0186). The sample door on the scanner was opened and the probe array was inserted into the holder. Affymetrix GeneChip Operating Software (GCOS) runs all aspects of the array process, saving images of the scanned probe array in a data file (*.dat). GCOS automatically calculates the *.cel (Cell Intensity File) file from each *.dat file, which contains a single intensity value for each probe cell delineated by the grid (calculated by the Microarray Suite 5.0 (MAS5.0) algorithm) (ref- Affymetrix, I. Statistical Algorithms Description Document. 2002).

<http://www.affymetrix.com/support/technical/whitepapers.affx>). The chip File (*.chp) generated from the analysis of a probe array contains qualitative and quantitative analysis for every probe set. The report file (*.rpt) generated by GCOS summarizes the data quality information for a single experiment. The report is generated from the analysis output file (*.chp).

2.5.3.2 Quality assessment of Affymetrix microarray chips

The quality of the data generated with Affymetrix microarray chips was assessed based on different criteria including the scaling factor, background and noise levels, GAPDH 3'/5' ratios and the % Present call.

Scaling factor: The scaling factor was the multiplication factor applied to each signal value on an array. A scaling factor of 1.0 indicates that the average array intensity was equal to the target intensity. Scaling factors vary across different samples and so there were no set guidelines for any particular sample type. However, Affymetrix advise that for replicates and comparisons involving a relatively small number of changes, the scaling/normalization factors (calculated by the global method) should be comparable among arrays. Larger discrepancies among scaling/normalization factors (e.g., three-fold or greater) may indicate significant assay variability or sample degradation leading to noisier data.

Background and noise levels: Although there are no official guidelines regarding background, Affymetrix has found that typical Average Background values range from 20 to 100 for arrays scanned with the GeneChip® Scanner 3000. Arrays being compared should ideally have comparable background values. A high background (over 60%) implies that impurities, such as cell debris and salts, are binding to the probe array in a non-specific manner, and that these substances are fluorescing at 570nm (the detection wavelength). This non-specific binding (noise) causes a low signal-to-noise ratio (SNR), meaning that transcripts present at very low levels in the sample may be incorrectly called as “Absent”. High background creates an overall loss of sensitivity in the experiment

GAPDH 3’/ 5’ ratios: β -actin and GAPDH are used to assess RNA sample and assay quality for the majority of GeneChip expression arrays. Specifically, the Signal values of the 3’ probe sets for β -actin and GAPDH are compared to the Signal values of the corresponding 5’ probe sets. The ratio of the 3’ probe set to the 5’ probe set is generally no more than 3 for the 1-cycle assay. A high 3’ to 5’ ratio may indicate degraded RNA or inefficient transcription of ds cDNA or biotinylated cRNA. 3’ to 5’ ratios for internal controls are displayed in the Expression Report (.rpt) file.

%Present call: The number of probe sets called “Present” relative to the total number of probe sets on the array is displayed as a percentage in the Expression Report (.rpt) file. Percent present (%P) values depend on multiple factors including cell/tissue type, biological or environmental stimuli, probe array type, and overall quality of RNA. Replicate samples should have similar %P values. Extremely low %P values are a possible indication of poor sample quality. In practice, % present calls averaged between 40-60% for cell line RNA and 15-25% for clinical specimens.

2.5.4 Reverse Transcription of RNA from cells (cDNA Synthesis)

For cDNA synthesis High-Capacity cDNA Reverse Transcription Kit was used (Applied BioSystems, 43755750). The components of the kit were allowed to thaw on ice. RT master mix was also prepared on ice as described in Table 2.5.4.1.

Component	Volume (μ l)
10X RT Buffer	2
25X dNTP Mix (100 mM)	0.8
10X Random Primers	2
MultiScribe Reverse Transcriptase	1
Nuclease-free H ₂ O	4.2
Total	10

Table 2.5.4.1: 2X RT Master Mix

10 μ l of 2X RT master mix were pipetted into each 0.5 ml eppendorf tube (Eppendorf, 0030 121 023). To this eppendorf tube 10 μ l of RNA sample (400ng) was mixed by pipetting up and down few times. The tubes were then briefly centrifuged to spin down the contents and to eliminate any air bubbles and placed on ice until ready to be loaded in the thermal cycler. The thermal cycler was programmed as per Table 2.5.4.2.

	Step 1	Step 2	Step 3	Step 4
Temperature ($^{\circ}$C)	25	37	85	4
Time	10 min	120 min	5 sec	∞

Table 2.5.4.2: Programme for thermal cycler

2.5.5 Quantitative real time RT-PCR (qRT-PCR)

TaqMan probes are oligonucleotides that have fluorescent reporter dyes attached to the 5' end and a quencher moiety coupled to the 3' end. These probes are designed to hybridize to an internal region of a PCR product. In the unhybridized state, the proximity of the fluor and the quench molecules prevents the detection of fluorescent signal from the probe. During PCR, when the polymerase replicates a template on which a TaqMan probe is bound, the 5'- nuclease activity of the polymerase cleaves the probe. This decouples the fluorescent dye thus, increasing the fluorescence in each cycle, proportional to the amount of probe cleavage.

2.5.5.1 Primer Design

Primer design was done using Primer Express from Applied BioSystems (<http://www.appliedbiosystems.com>) and the primers were ordered from MWG (<http://www.mwg-biotech.com/>). The primers were designed across the introns, making them specific for detection of RNA. Since the two genes under study (ROPN1 and ROPN1B) was 95% homologous, the region with maximum variability among the two genes was taken for the design of forward primers, reverse primers and probe.

2.5.5.2 qRT-PCR

The TaqMan quantitative Real time PCR (qRT-PCR) analysis was performed using the Applied BioSystems. In order to exclude any amplification product derived from genomic DNA or any other contaminant that could contaminate the RNA preparation, total RNA without reverse transcription was used as a negative control. Water on its own was used as a negative control to detect the presence of any contaminating RNA or DNA.

Reagents	Volume
Nuclease-Free water (Ambion, 9930)	5 μ l
TaqMan® Fast Universal PCR master mix (2 X) (Applied BioSystems, 4352042)	10 μ l
Forward primer	1 μ l
Reverse primer	1 μ l
Probe	1 μ l
Total	18 μ l

Table 2.5.5.2.1: qRT-PCR Reaction Mixture

18 μ l of reaction master mix (Table: 2.5.5.2.1) was added to the MicoAmp fast optical 96-well reaction plate (Applied BioSystems, 4346906) followed by 2 μ l of the cDNA.

Step	Denature	PCR	
	HOLD	CYCLE (40 cycles)	
		Denature	Anneal/Ext
Time	20 sec	3 sec	30 sec
Temp	95°C	95°C	60°C

Table 2.5.5.2.2: Thermal cycling conditions used in this study

Cycle threshold: The Threshold is the level of detection or the point at which a reaction reaches a fluorescent intensity above background. The threshold line is set in the exponential phase of the amplification for the most accurate reading. The cycle at which the sample reaches this level is called the Cycle Threshold, Ct.

Relative Quantification: Relative quantification determines the change in expression of a nucleic acid sequence (target) in a test sample relative to the same sequence in a calibrator sample (control).

2.5.6 Large scale plasmid preparation

Luria-Bertani (LB Broth) was prepared as per Table 2.5.6 and was autoclaved. An aliquot of 10 mls LB Broth was taken in a 20 ml universal. Ampicillin was added to this broth at a concentration of 100µg/ml and inoculated with 10-20µl of glycerol stock for one unique clone. This was grown for 6-7hrs in an upright shaker at 37°C and ~300rpm. This was further inoculated into a 1000 ml flask with 400 mls of LB Broth containing ampicillin antibiotic (100µg/ml). The culture was incubated at 37°C with vigorous shaking (~300rpm) for ~8hrs. The bacterial cells were then harvested by centrifugation at 6000xg for 15 min at 4°C. Plasmid DNA was then extracted using the Qiagen Endofree Plasmid Purification Kit (Qiagen, 12362) (see section 2.5.6.1). DNA concentration was determined by measuring using NanoDrop at OD260nm (see section 2.5.2).

Reagents	Volume
Peptone	20g/L
Yeast Extract	10g/L
NaCl	5g/L

Table 2.5.6: 2X-LB broth (low-salt) media preparation

2.5.6.1 Isolation of plasmid DNA using Qiagen Endofree Plasmid Purification Kit

Pelleted bacterial cells were resuspended in 250µl Buffer P1 with RNase A until no clumps were visible. This was then transferred to a microcentrifuge tube. 250µl of Buffer P2 was added to this mixture and mixed thoroughly by inverting the tube gently 4–6 times until the solution became viscous and slightly clear. To this mixture 350µl of Buffer N3 was added and mixed immediately and thoroughly by inverting the tube 4–6 times till the solution became cloudy. This solution was then centrifuged for 10 min at 13,000 rpm (~17,900 x g). During centrifugation the vacuum manifold and QIAprep spin columns (Qiagen, 27104) were prepared. The supernatant from this was pipetted to the QIAprep spin column. The vacuum source was then switched on to draw the solution through the QIAprep spin columns. The QIAprep spin column was washed by adding 0.5 ml Buffer PB and the vacuum source was switched on. After the solution had moved through the column, the vacuum source was switched off. This wash step removed trace nuclease activity. The QIAprep spin column was again washed by adding 0.75 ml Buffer PE. The vacuum source was again switched on to draw the wash solution through the column and then switched off. The QIAprep spin columns were then transferred to a microcentrifuge tube and centrifuged for 1 min. This step removed residual Buffer PE and ethanol from Buffer PE that may inhibit subsequent enzymatic reactions. The QIAprep column was then placed in a clean 1.5 ml microcentrifuge tube and DNA was eluted by adding 50µl Buffer EB (10mM Tris·Cl, pH 8.5) the center of the QIAprep spin column. The QIAprep spin column was allowed to stand for 1 min and then centrifuged for 1 min.

2.5.7 Plasmid transfection protocol

To determine the optimal conditions for plasmid transfection in 6-well plates, an optimisation with GFP plasmid was carried. Cell suspensions were prepared at 2×10^5 cells per ml of plating media and plated one day in advance in 6-well plate (2 ml per well). Solutions of GFP plasmid at a concentration of $2 \mu\text{g}/\mu\text{l}$ were prepared in optiMEM (Gibco, 31985). Lipofectamine 2000 (2, 4, $6 \mu\text{l}$) solutions were prepared in $500 \mu\text{l}$ optiMEM and incubated at room temperature for 5 min. After incubation, the lipofectamine- optiMEM solution was added to each GFP plasmid. These solutions were mixed well and incubated for a further 20 min at room temperature. $500 \mu\text{l}$ of the plasmid/lipofectamine solutions were added to a 6-well plate. The plates were mixed gently and incubated at 37°C for 5hrs. After 5hrs, the transfection mixture was removed from the cells and the plates were fed with fresh medium. After 72hrs, cells were observed under the fluorescent microscope. The conditions chosen for large-scale use were those that showed the most fluorescent cells and fewest dead floating cells.

The plasmid in bacteria was obtained from Open Biosystems (<http://www.openbiosystems.com>). Details are in Table 2.5.7.

Gene	Accession	Vector	Catalogue
ROPN1	BC132744.1, BC132744	PCR4-TOPO	MHS4426- 99240150
ROPN1B	BC015413, BC015413.1, BG034740.1, BG034740	PCMV-SPORT6	MHS1010-58339

Table 2.5.7: Details of the cDNA used in the study

2.5.8 RNA interference (RNAi)

RNAi was carried out using small interfering RNAs (siRNAs) to silence specific genes. The siRNAs used were purchased from Ambion Inc (USA). Details of the individual siRNA are listed in Table 2.5.7.1. These siRNAs were 21-23 bp in length and were

introduced to the cells via reverse transfection with the transfection agent NeoFx (Ambion Inc., 4511). The details of siRNA used are listed in Table 2.5.8. The siRNA obtained was 5nmol and was diluted with 100µl nuclease free water to obtain a stock of 50µM. The transfection solution was prepared using 30nm siRNA and 4µl of NeoFx.

The amount of siRNA used was calculated as follows:

Amount required = (required concentration / existing concentration) x total transfection volume

In this case; Amount required = $(30 \times 10^{-9} \text{ g} / 50 \times 10^{-6} \text{ g}) \times 1100\mu\text{l} = 0.66\mu\text{l/well}$

This amount was diluted with 50µl of optiMEM. Similarly NeoFx was diluted with 50µl optiMEM. Both were incubated for 10 min, mixed, again incubated for 10 min and then placed in the wells of 6 well plates. Cell suspension @ 2×10^5 per well was placed on top of the siRNA solution and mixed. This was placed in incubator at 37°C. The media was changed after 24hrs and the assay was performed after 72hrs following transfection.

Gene	Ambion Catalogue	Ambion siRNA Id
ROPN1B	16708A	279600
ROPN1B	16708A	279601
ROPN1B	16708A	279602
ROPN1	4392420	S29402
ROPN1	4392420	S29404
Negative Control	4611	Negative control 1

Table 2.5.8: Ambion siRNA details used in the study

2.5.9 Western Blot analysis

2.5.9.1 Lysis of cell pellet

A stock of lysis buffer was prepared using the reagents as shown in table 2.5.9.1.1. Working lysis buffer was prepared fresh every time before use as shown in the table 2.5.9.1.2.

Reagents	Volume
1M Tris pH 7.4	2 ml
3M NaCl	1.66 ml
1M NaF	5 ml
10% NP40	1 ml
H2O	90.34 ml
Total	100 ml

Table 2.5.9.1.1: NP40 Lysis Buffer Stock

Reagents	Volume
100mM Sodium Orthovanidate	10 μ l
100mM PMSF	10 μ l
Protease Inhib Cocktail	40 μ l
NP40 Lysis Buffer	940 μ l
Total	1000 μ l

Table 2.5.9.1.2: Working Lysis Buffer

After 72hrs of transfection, the cells were lysed in the 6-well plate. Prior to lysis, the plate was washed twice with cold PBS and drained of all supernatant. Cold lysis buffer (40 μ l) was added to the cells, dropping it evenly over the whole well/plate. The cells were scraped from the well and all the lysed/scraped cells were gathered in one corner of the well. The solution was pipetted up and down without frothing and placed in a pre-chilled eppendorf tube (Eppendorf, 0030 121 023). The tube was vortexed for 30-60 sec until the solution was homogenised. The tube was placed on ice for 20 min and centrifuged at maximum speed in a microfuge for 15 min. The supernatant was transferred to a fresh chilled eppendorf tube and immediately quantified for protein (see section 2.5.9.2). The samples were prepared with 2-5X loading buffer (Sigma, S-3401) and water was added to make the all the samples at the same concentration. Parafilm was wrapped around the lids of the eppendorfs (to avoid evaporation) and the samples were boiled for 3-5 min. If not used immediately, the samples were stored at -20°C until needed.

2.5.9.2 Protein Quantification

Lysed samples were removed from the freezer and placed on ice. A BSA standard of 1 mg/ml was prepared in UHP. Diluted BSA Standards was prepared and 5µl of standard and 5µl of sample were placed in triplicate wells on a 96-well plate (Costar, 3596). The Biorad D_c Protein Assay (Biorad, 500-0116) was used for protein quantification. 25µl of Reagent A (containing 20µl Reagent S (Biorad, 500-0115) per ml of Reagent A (500-0113)) followed by 200µl of Reagent B (Biorad, 500-0114) were added to each test well. The plate was kept at room temperature for 15 min prior to reading on the Spectra Max Plus using a softmax Lowry protein assay (750nm) program.

2.5.9.3 Gel electrophoresis

Proteins for Western blot analysis were separated by SDS-polyacrylamide gel electrophoresis (SDS-PAGE). Resolving and stacking gels were prepared as outlined in table 2.5.8.4.1 and poured into clean 10 cm x 8 cm gel cassettes, separated by 0.75 cm plastic spacers. The plates were cleaned by tap water, followed by UHP. After drying, the plates were wiped down in one direction using tissue paper soaked in 70% Industrial Methylated Spirits (IMS). The spacers and comb used were also cleaned in this way. After these had dried, the resolving gel was poured first and allowed to set for 20 min at room temperature. The stacking gel was then poured and a comb was placed into the stacking gel in order to create wells for sample loading. Once set, the gels could be used immediately or wrapped in wet tissue paper and stored at 4°C for 24hrs.

1X running buffer (Table 2.5.9.3) was added to the running apparatus before samples were loaded. The samples were loaded onto the stacking gels, in equal amounts relative to the protein concentration of the sample. The empty wells were loaded with loading buffer.

Reagents	Volume
Glycine	14.4 g
Tris	3.03 g
SDS	1g
H ₂ O	1L

Table 2.5.9.3: Running Buffer Components

The samples were loaded including 7µl of molecular weight colour protein markers (New England Biolabs, P7708S). The gels were run at 200 V, 45mA for approximately 1.5hrs. When the bromophenol blue dye front was seen to have reached the end of the gels, electrophoresis was stopped.

2.5.9.4 Western blotting

Following electrophoresis, the acrylamide gels were equilibrated in transfer buffer (2mM Tris, 192mM glycine (Sigma, G-7126) pH 8.3-8.5 without adjusting) for 10 min. The protein in the gels was transferred onto nitrocellulose membranes (Boehringer Mannheim, 1722026) by semi-dry electroblotting. Eight sheets of Whatman 3 mm filter paper (Whatman, 1001824) were soaked in transfer buffer and placed on the cathode plate of a semi-dry blotting apparatus (Biorad, 170-3940). Excess air was removed from between the filters by rolling a universal tube (Sterilin, 128a) over the filter paper. A piece of nitrocellulose membrane, cut to the same size of the gel, was prepared for transfer (soaked in transfer buffer) and placed over the filter paper, making sure there were no air bubbles.

The acrylamide gel was placed over the nitrocellulose membrane and eight more sheets of pre-soaked filter paper were placed on top of the gel. Excess air was again removed by rolling the universal over the filter paper. The proteins were transferred from the gel to the nitrocellulose at a current of 34mA at 15V for 24-25 min.

All incubation steps from then on, including the blocking step, were carried out on a revolving apparatus (Stovall, Bellydancer) to ensure even exposure of the blot to all

reagents. The nitrocellulose membranes were blocked for 2hrs at room temperature with fresh 5% non-fat dried milk (Cadburys, Marvel skimmed milk) in Tris-buffered saline (TBS) with 0.5% Tween (Sigma, P-1379). After blocking, the membranes was washed 3 x 5 min using 1X TBS/PBS. The membrane was then incubated with 5 to 10 mls primary antibody (concentration of primary antibody was used as in Table 2.5.9.4.2) for 1hr. Details of the antibody used is listed in Table 2.5.9.4.2. The membrane was again was washed 3 x 5 min using 1X TBS/PBS and then incubated in secondary antibody (Mouse antibody diluted at 1/1000) (DakoCytomation, P 0260). Finally the membrane was washed 3 x 5 min using 1X TBS/PBS. Bound antibody was detected using enhanced chemiluminescence (ECL) (Amersham, RPN2109) (see section 2.5.9.5).

Components	Resolving gel (7.5%)	Resolving gel (12%)	Stacking
Acrylamide stock(Sigma, A3574)	3.8 mls	5.25 mls	0.8 mls
Ultra pure water	8.0 mls	6.45 mls	3.6 mls
1.5M-Tris/HCl, pH 8.8(BioRad, 161-0798)	3.0 mls	3.0 mls	-
1.25M-Tris/HCl, pH 6.8(BioRad, 161-0799)	-	-	0.5 mls
10% SDS(Sigma, L-4509)	150 mls	150 mls	50 mls
10% Ammonium persulphate(Sigma, A-1433)	60 mls	60 mls	17 mls
TEMED(Sigma, T-8133)	10 mls	10 mls	6 mls

Table 2.5.9.4.1: Preparation of electrophoresis gels

Note: *Acrylamide stock solution consists of 29.1g acrylamide (Sigma, A8887) and 0.9g NN'-methylene bis-acrylamide (Sigma, 7256) dissolved in 60ml UHP water and made up to 100ml final volume. The solution was stored in the dark at 4°C for up to 1 month. All components were purchased from Sigma, SDS (L-4509), NH₄-persulphate (A-1433) and TEMED, N,N,N,N'-tetramethylethylenediamine (T-8133).

Antibody	Dilution/Concentration	Supplier	Cat. No.
Ropporin	1/1000	Abnova	H00054763-M03
α tubulin	1/5000	Abcam	ab7291

Table 2.5.9.4.2: List of primary antibodies used for western blot analysis

2.5.9.5 Enhanced chemiluminescence detection

Protein bands were developed using the Enhanced Chemiluminescence Kit (ECL) (Amersham, RPN2109) according to the manufacturer's instructions. The blot was removed to a darkroom for all subsequent manipulations. A sheet of parafilm was flattened over a smooth surface, e.g. a glass plate, making sure all air bubbles were removed. The membrane was placed on the parafilm, and excess fluid removed. 1.5 mls of ECL detection reagent 1 and 1.5 mls of reagent 2 were mixed and covered over the membrane. Charges on the parafilm ensured the fluid stayed on the membrane. The reagent was removed after one minute and the membrane covered in cling film. The membrane was exposed to autoradiographic film (Boehringer Mannheim, 1666916) in an autoradiographic cassette for various times, depending on the signal (30s – 15 min). The autoradiographic film was then developed.

The exposed film was developed for 5 min in developer (Kodak, LX24, and diluted 1:6.5 in water). The film was briefly immersed in water and fixed (Kodak, FX-40, diluted 1:5 in water), for 5 min. The film was transferred to water for 5 min and then air-dried.

2.5.10 Invasion assay

2.5.10.1 Reconstitution of ECM proteins.

Matrigel (Sigma, E-1270) was diluted to a working stock of 1 mg/ml in serum free DMEM. Aliquoted stocks were stored at -20°C .

2.5.10.2 *In vitro* invasion assays

100 μl of matrigel were placed into each insert (Falcon, 3097) (8.0 μm pore size, 24-well format) and kept at 4°C for 24hrs. The insert and the plate were then incubated for one

hour at 37°C to allow the proteins to polymerize. Cells were harvested and resuspended in culture media at 1×10^6 cells/ml. Excess media/PBS was removed from the inserts, and they were rinsed with culture media. 100µl of the cell suspension was added to each insert and 500µl of culture medium was added to the well underneath the insert. Cells were incubated for 24hrs. After this time period, the inside of the insert was wiped with a cotton swab dampened with PBS, while the outer side of the insert was stained with 0.25% crystal violet for 10 min and then rinsed in distilled water and allowed to dry. The inserts were then viewed and photographed under the microscope. The invasion assays were quantified by counting cells in 10 random fields within a grid at 20x objectives and graphed as the total number of cells invading at 200 x magnifications.

2.5.11 Motility assay

Motility assays were carried out as described in section 2.5.10.2, without the addition of ECM.

2.5.12 Microsoft PowerPoint

Microsoft PowerPoint is a presentation program developed by Microsoft. All the non-referenced diagrams (Fig 1.6.2.1.1, Fig 1.6.2.1.2, Fig 1.6.2.1.3, Fig 1.6.2.1.4, Fig 1.7.1, Fig 1.10.1.1 and Fig 3.5.2.1) were drawn using Microsoft PowerPoint.

3.0 Results

3.1 Breast Cancer clinical specimens

Microarray gene expression profiling of 17 normal breast tissue specimens and 104 breast cancer specimens was performed using Affymetrix U133 Plus2.0 arrays as outlined in section 2.5.3. The aim was to study the clinical heterogeneity among breast tumors using gene expression data. Additionally the following comparisons based on clinical data were performed to study important genes, ontologies and pathways affected by the changing gene lists:

- 1) Normal vs. tumour
- 2) Estrogen receptor-negative vs. Estrogen receptor-positive
- 3) Lymph node-negative vs. Lymph node-positive
- 4) Grade 1 vs. Grade 2
- 5) Grade 2 vs. Grade 3
- 6) Tumour Size < 2.8cm vs. > 2.8 cm
- 7) Patients who did not relapse vs. patients who did relapse (Overall relapse)
- 8) Patients who survived vs. patients who did not survive (Overall survival)
- 9) Patients who did not relapse within 5 years vs. patients who did relapse within 5 years (Relapse 5yrs)
- 10) Patients who survived for 5 years vs. patients who did not survive beyond 5 years (Survival 5 yrs)

3.1.1 Data Normalization and Quantification

The microarray raw data files were normalized and quantified using the dChip algorithm as outlined in section 2.2.1).

3.1.2 Data Filtration

Data filtration was applied on 54,675 genes present on U133 Plus2.0 chip (see section 2.2.3), to remove genes which i) did not fluctuate very highly across samples and ii) fluctuated too highly across samples to be trustworthy. Genes with a Standard deviation / Mean i) below 1 or ii) above 1000 were removed from further analysis. This set of genes was used for Hierarchical clustering. 10,243 genes passed this criterion and were used to carry out clustering analysis of clinical specimens.

3.1.3 Hierarchical Clustering

Hierarchical clustering (see section 2.2.4) was performed on the 10,243-member filtered gene list. The correlation values between the samples were calculated and two-way clustering was performed using the correlation values. The distance metric used was 1-correlation and the clustering algorithm used was Average linkage clustering. Prior to clustering, the individual samples were standardised as follows: the expression of the individual genes was subtracted from their means for that sample and divided by their respective standard deviation. The results are shown in Figs 3.1.3.1-3.1.3.4. Fig 3.1.3.1 shows all the specimens and Fig 3.1.3.2 – Fig 3.1.3.4 displays the individual sub-clusters. The dendrogram has specimens on both axes and the intersection point between any two samples represents the correlation value among the two specimens. The colours on the heat map represent the correlation values among the specimens. Red colours indicate positive correlation, blue colours indicate negative correlation and white colours indicate zero correlation. Different shades of red and blue reflect the relative correlation values. The diagonal red line is because of the correlation values of 1 among the identical samples. Blocks of red therefore represents similar specimens and also indicate how homogenous or heterogeneous the various groups of specimens are.

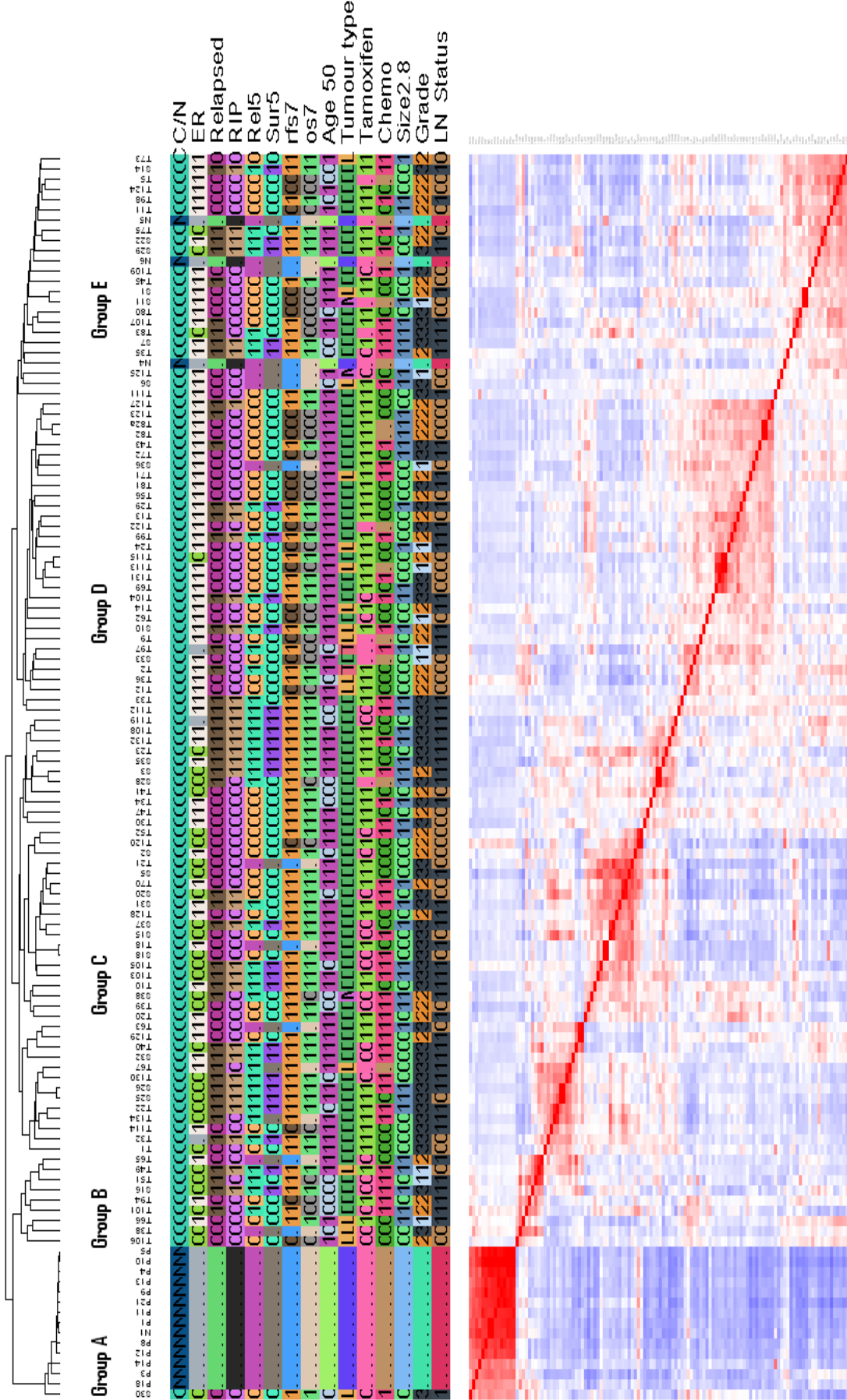


Fig 3.1.3.1: The above figure represents the two-way clustering of the specimens. This clustering was generated by calculating the correlation value among the samples based on the 10243 filtered genes and performing two-way clustering on those correlation values. The clustering method was average linkage clustering and the distance metric used was 1-correlation. This analysis identified five distinct clusters and is labelled as A, B, C, D, and E as shown below the tree.

The hierarchical clustering results identify several groups of samples which correlate with different clinical parameters. The Normal samples largely clustered together. Among the tumors, the samples largely clustered based on the ER status. The clustering results have been divided into groups represented as A, B, C, D and E and are denoted just below the tree in Fig 3.1.3.1 and are all shown separately in greater detail in Figs 3.1.3.2 - Fig 3.1.3.4.

The first observation to be made is that the tumour specimens largely clustered as a separate group. However, there was one cancer specimen (S30) which clustered with the normal specimens. Additionally, there were 3 normal specimens (N4, N5 and N6) which clustered within group E of the cancer specimens. A very high correlation was observed between the normal specimens in Cluster A compared to the various cancer groups and sub-groups. The other important criterion on which the specimens clustered was Estrogen receptor status. Cluster C is enriched with ER-negative specimens whereas clusters D and E are highly enriched with ER-positive specimens. The details of the individual clusters and their correlation with clinical parameters are detailed below.

Individual clusters were compared to identify genes important to that cluster. A nearest cluster comparison approach was used so as to mask the confounding factors. Additionally this approach helped in keeping the sample size of each group to a comparable level. Bigger groups were also compared to each other with an aim to obtain the hierarchy of the heterogeneity of breast cancer.

Cluster A: This group comprised 14 normal specimens and one cancer specimen (S30). This group represents a very tight cluster as the level of correlation among samples as indicated by the strength of red colour is very high (Fig 3.1.3.1, Fig 3.1.3.2). However, three of the normal specimens (N4, N5 and N6) did not cluster with this group.

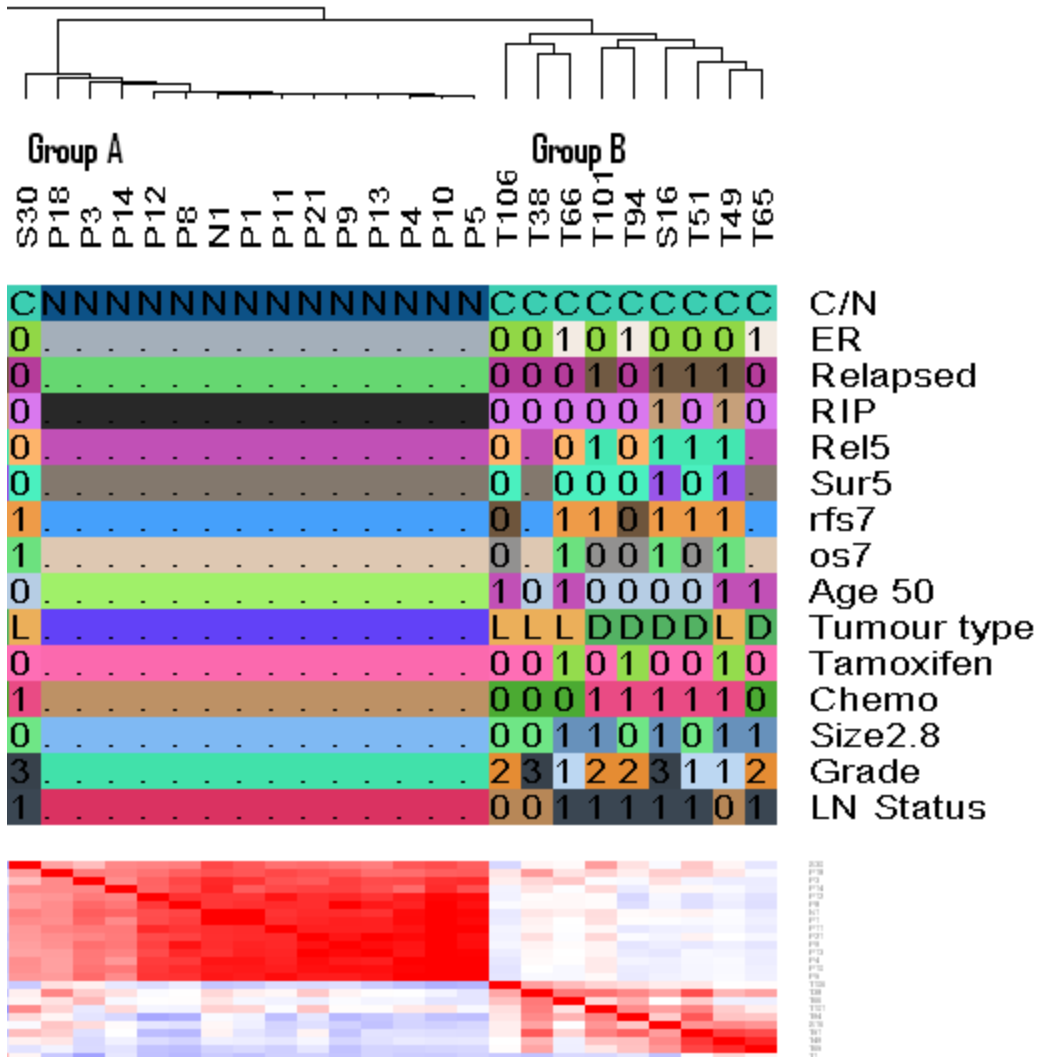


Fig 3.1.3.2: Enlarged view of the cluster A and B from the Fig 3.1.3.1. The colours on the heat map represent the correlation values among the specimens. Red colours indicate positive correlation, blue colours indicate negative correlation and white colours indicate zero correlation. Different shades of red and blue reflects the relative correlation values. Group A is composed of 14 normal samples and 1 tumour sample (S30). Group B is the cluster closest to the normal sample group.

Cluster B: This group was enriched for patients who are ER-negative, relative to the other categories. The other linked clinical parameters are depicted in Table 3.1.3.1. The group that clusters closest to cluster B is the Normal specimen cluster (A)

Significance analysis (as described in section 2.2.4) was carried out on this cluster and the significant clinical parameters specific to this cluster was identified (Table 3.1.3.1).

Clinical	Status	Represented	Total	p-value
ER	Negative	6/9	34/122	0.0139
Age50	<50	5/9	27/122	0.0248
Type	Lobular	4/9	17/122	0.0215
Tamoxifen	Not taken	6/9	26/122	0.0030
Grade	1	3/9	11/122	0.0343

Table 3.1.3.1: Clinical parameters over-represented in Cluster B. ‘Clinical’ refers to the parameter of interest. ‘Status’ indicates the particular frame of reference for that clinical parameter. The ‘Represented’ column indicates the number of specimens of the total number of specimens within Cluster B that display those values for that parameter. ‘Total’ indicates the total number of specimens for that clinical status over the whole sample dataset. Based on the ratio of ‘Represented’ and ‘Total’, the p-value is calculated.

While comparing cluster A and B, cluster B did not display a higher expression of ESR1 as this cluster is enriched for ER-negative specimens; however ERBB2 expression was found to be high (FC: 2.46) in Cluster B specimens. ESR1 gene and ERBB2 gene are important in breast cancer classification.

Cluster C: This group was enriched for patients who are ER-negative and have undergone chemotherapy (see Table 3.1.3.2). By looking at the hierarchical clustering (Fig 3.1.3.3), it seems that many of the tumors isolated from patients in this group are of histologic Grade 3 (31 specimens are of Grade 3 out of total of 43 specimens in cluster C), despite the fact that grade was not identified by the significance analysis. None of the samples in this group was of Grade 1. This group was also enriched by patients who relapsed within

7 years (41/43) and who did not survive beyond 7 years (39/43). Close examination of this group (based on the dendrogram and heat map) reveals that there are 3 sub-groups in this main group. However, no clinical parameters associated with any of the sub-groups were identified as statistically-relevant by significance analysis.

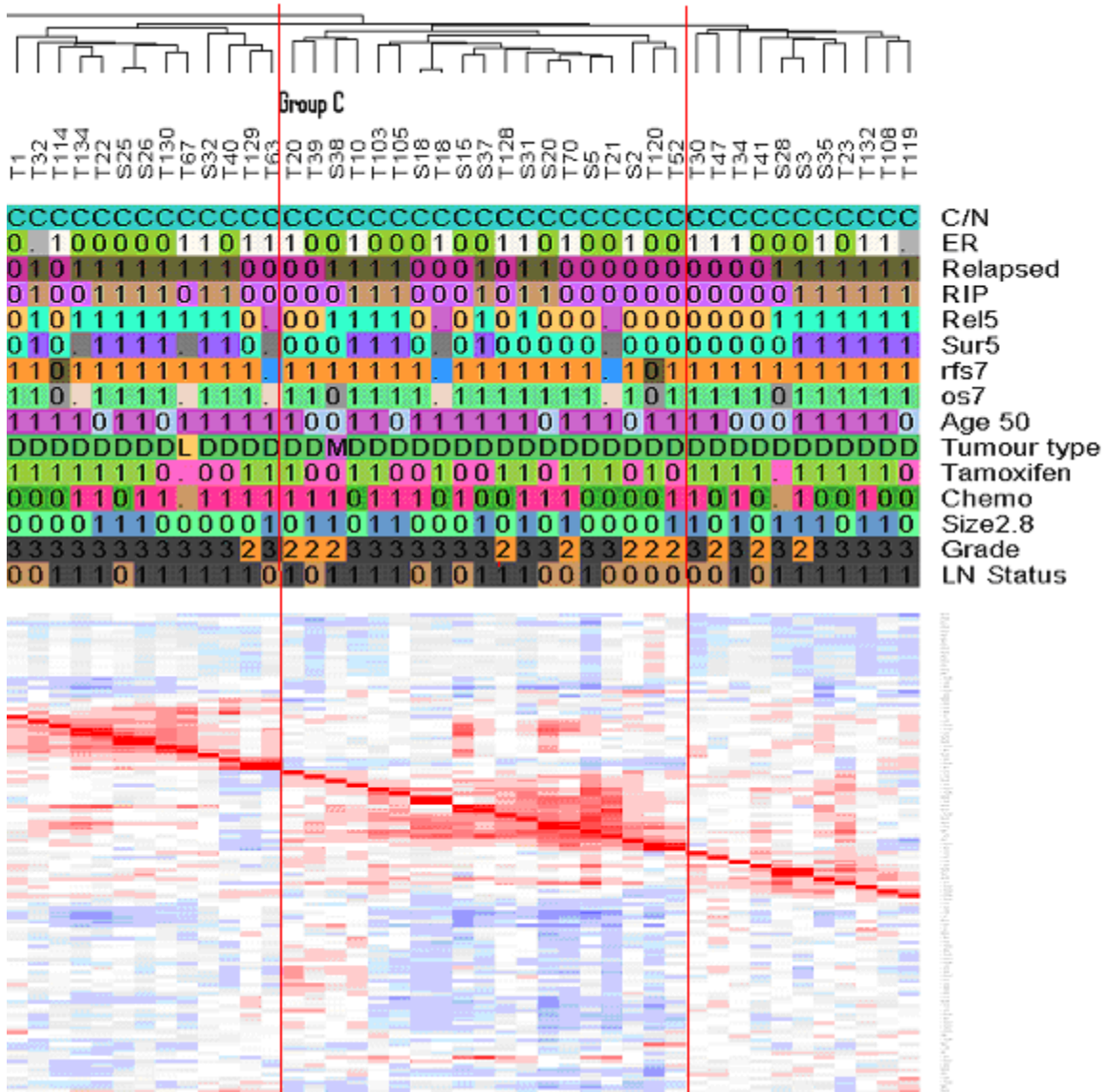


Fig 3.1.3.3: Enlarged view of the cluster C from Fig 3.1.3.1. This group of specimens is enriched for ER-negative patients (52.4%). However there are many ER-positive specimens in this group (40.5%). This group signifies a very high level of diversity as can be seen from the clustering patterns. There are three sub-clusters in this cluster (marked by red lines).

Clinical	Status	Represented	Total	p-value
ER	negative	23/43	34/122	0
Chemo	yes	23/43	50/122	0.0304

Table 3.1.3.2: Clinical parameters over-represented in cluster C

Cluster C is highly heterogeneous groups with three distinct sub-clusters. The left cluster was compared to (middle + right) cluster. Similarly the middle cluster was compared to (left + right) cluster and the right cluster was compared to (left + middle) cluster. The nearest cluster comparison approach was used so as to mask out the confounding factors. Table 3.1.3.3, 3.1.3.4 and 3.1.3.5 lists the top 10 genes (based on Fold change) among each comparison.

Probe set	Gene	Baseline	Experiment	FC	Difference	p-value
205475_at	SCRG1	11.91	297.67	24.98	285.76	0.02769
229341_at	TFCP2L1	18.83	382.88	20.33	364.04	0.028253
242488_at	---	13.63	244.75	17.96	231.13	0.00699
213456_at	SOSTDC1	13	224.4	17.27	211.4	0.021923
220425_x_at	ROPN1B	18.82	316.25	16.81	297.43	0.028281
224191_x_at	ROPN1	24.19	401.57	16.6	377.38	0.03536
231535_x_at	ROPN1	30.22	441.08	14.6	410.86	0.032765
220559_at	EN1	24.27	311.89	12.85	287.62	0.023241
204733_at	KLK6	21.32	222.4	10.43	201.08	0.036002
204855_at	SERPINB5	89.09	786.11	8.82	697.02	0.001439

Table 3.1.3.3: Genes up-regulated in left cluster in comparison to (middle + right) cluster

The left sub-cluster is enriched with patients who relapsed (9/13). One of the important and novel genes identified in the table above is Ropporin (ROPN1, ROPN1B). Further work on this gene is presented in section 3.6

Probe set	gene	Baseline	Experiment	FC	Difference	p-value
205213_at	CENTB1	5.22	147.31	28.21	142.08	0.002377
208450_at	LGALS2	10.24	193.39	18.89	183.15	0.00811
216510_x_at	IFI6	18.12	309.82	17.1	291.7	0.005858
211650_x_at	IL8	24.77	373.07	15.06	348.3	0.003931
211637_x_at	LOC652128	17.04	249.1	14.62	232.06	0.006901
214777_at	---	31.89	464.36	14.56	432.47	0.004748
216365_x_at	IGL@	27.27	368.26	13.5	340.99	0.02424
211908_x_at	IL8	24.93	334.96	13.44	310.03	0.005076
216430_x_at	SCGB2A2	10.06	124.5	12.37	114.44	0.013686
211634_x_at	IGHM	11.83	145.43	12.3	133.6	0.04216

Table 3.1.3.4: Genes up-regulated in middle cluster in comparison to (left + right) cluster

The middle cluster is enriched with patients who did not relapse (12/19). The genes in the middle cluster are enriched for immune response function (IFI6, IF8, LOC652128, IGL, and IGHM).

Probe set	gene	Baseline	Experiment	FC	Difference	p-value
207802_at	CRISP3	31.13	1080.8	34.72	1049.67	0.037581
232547_at	SNIP	33.68	645.21	19.16	611.53	0.002929
213557_at	CRKRS	84.5	807.21	9.55	722.71	0.002197
234354_x_at	ERBB2	27.95	221.59	7.93	193.64	0.003465
238360_s_at	---	16.19	127.13	7.85	110.94	0.00764
213551_x_at	PCGF2	145.63	1078.29	7.4	932.66	0.004555
226727_at	LOC284106	183.4	1352.26	7.37	1168.85	0.010993
214239_x_at	PCGF2	273.49	1956.67	7.15	1683.19	0.003642
239224_at	FBXL20	23.02	148.81	6.47	125.79	0.003134
203496_s_at	PPARBP	248.16	1585.54	6.39	1337.37	0.005646

Table 3.1.3.4: Genes up-regulated in right cluster in comparison to (left + middle) cluster

The right cluster is enriched with patients who relapsed (7/11). Among other important genes, this cluster over-expresses ERBB2 which is known to confer bad prognosis on breast cancer patients.

Cluster D: This group was enriched for tumors which are mainly ER-positive (all samples apart from one), Grade 1 and lobular.

Clinical	Status	Represented	Total	p-value
Type	Lobular	8/31	17/122	0.0324
Grade	1	6/31	11/122	0.0302

Table 3.1.3.5: Clinical parameters over-represented in cluster D

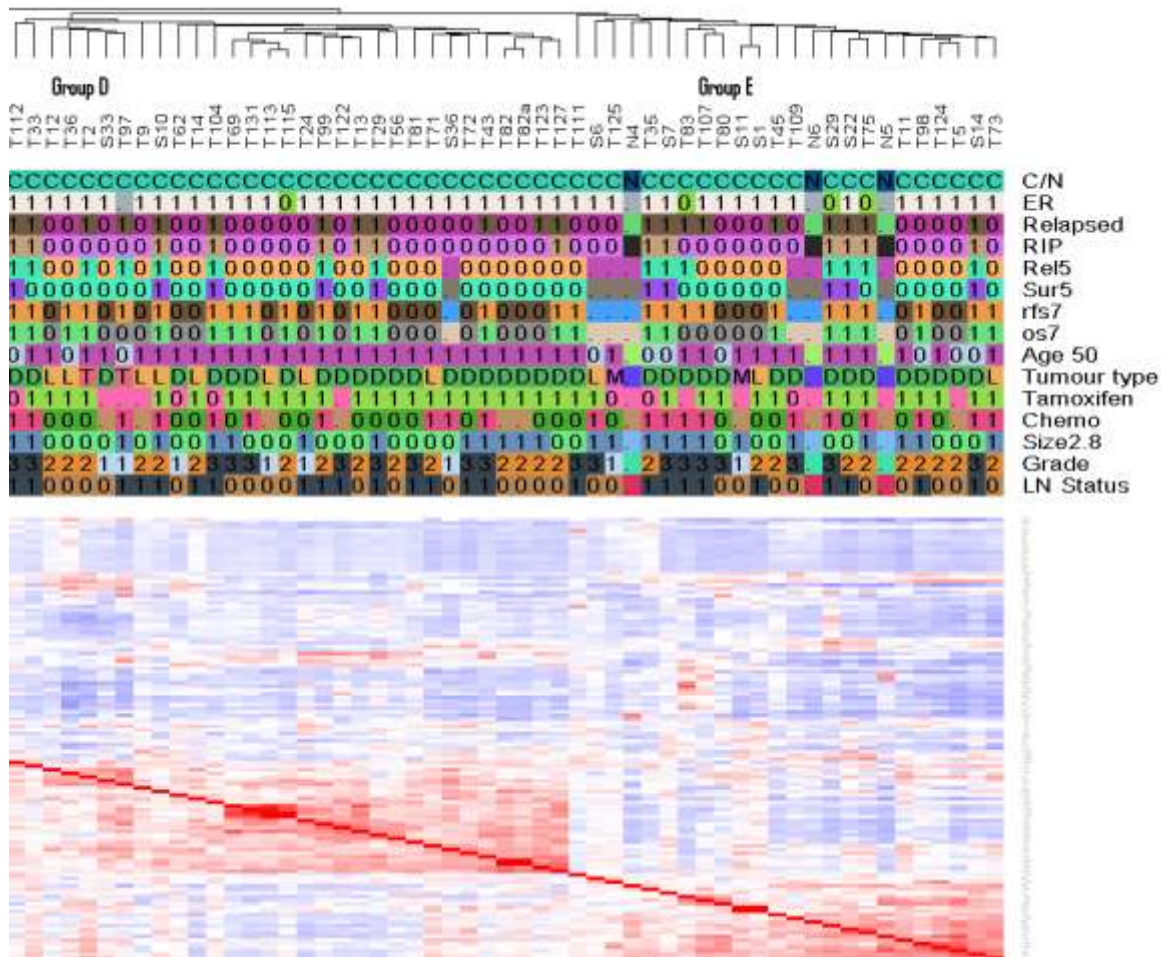


Fig 3.1.3.4 Enlarged view of cluster D and E from the Fig 3.1.3.1. These clusters are highly enriched for ER-positive specimens. Cluster D has only 1 ER-negative specimen and the Cluster E has 3 ER-negative specimens.

Cluster E: This group was also enriched for patients with ER-positive tumors. However there were 3 ER-negative and 3 Normal specimens in this group. No particular clinical parameter was identified by significance analysis for this group, however when combined with Cluster D, the significance analysis identified the combination group to be associated with overall survival and censored relapse free survival (Table 3.1.3.6). Despite both clusters D and E being enriched for ER-positive specimens, ER was not identified by the significance analysis, possibly due to its prevalence in Cluster C and B.

Clinical	Status	Represented	Total	p-value
CenRFC7	0	21/55	25/122	0
OSSur7	0	24/55	32/122	0.001

Table 3.1.3.6: Clinical parameters over-represented in clusters D and E. Both the clusters were combined to perform this analysis.

Cluster D and Cluster E both are enriched with ER-positive patients, however cluster D have an overall higher expression of ESR1 (FC: 2.65) and other ER related genes, including GATA3 (FC: 1.64), FOXA1 (FC: 1.65), SPDEF (FC: 1.67) and TFF3 (FC: 1.59).

Cluster C was compared to Cluster D+E. As expected, ESR1 (FC: 3.04) and other ER related genes e.g. GATA3 (FC: 2.1), FOXA1 (FC: 1.73), SPDEF (FC: 1.6) and TFF3 (FC: 1.9) were up-regulated in Cluster D+E.

The above results correlating different clusters with clinical parameters was also corroborated by the results of Kaplan Myer analysis performed by Dr Lorraine O'Driscoll (Fig 3.1.3.5) and described in section 2.2.12.

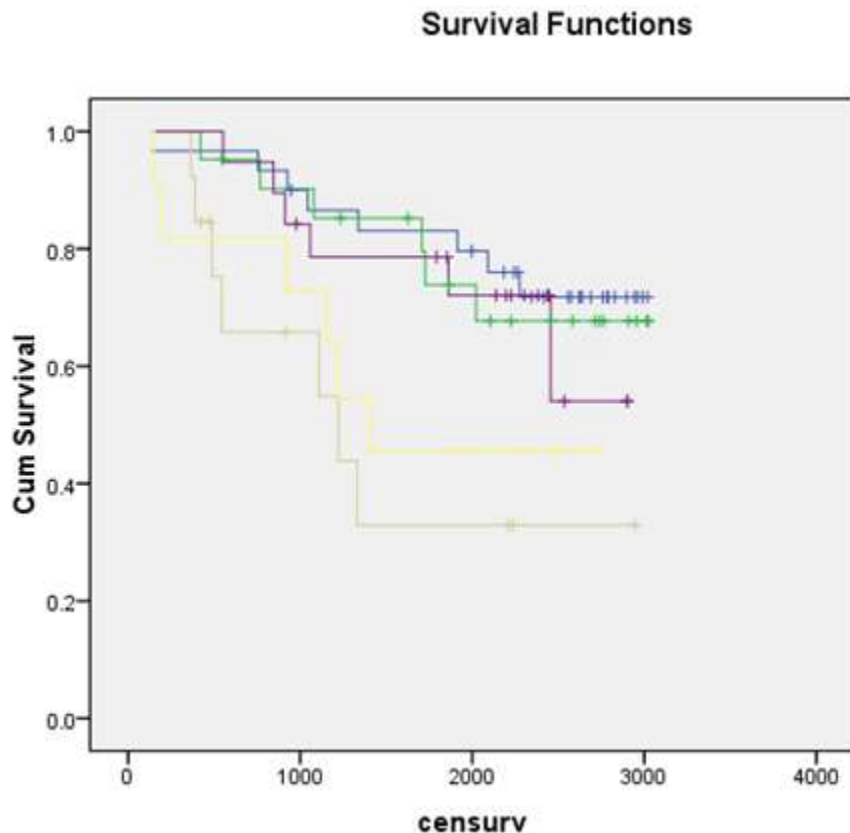


Fig 3.1.3.5: KM analysis on various groups identified by Hierarchical clustering. The different colour line represents the survival of patients of each group. BLUE: Cluster E; GREEN: Cluster D; VIOLET: Cluster C (Middle sub-cluster); YELLOW: Cluster C (Right sub-cluster); GREY: Cluster C (Left sub-cluster)

The results indicate that the Ropporin-enriched cluster (Grey) has the worst survival compared to patients belonging to other groups. This was followed by ERBB2 over-expressing cluster (Yellow). These two groups of patients were enriched for ER-negative specimens. Another ER-negative enriched group (Violet) with a high expression of immune response genes displayed a much better survival. Among the groups of ER-positive patients, one with high expression of ER partner genes (Blue) had a marginally better prognosis than the other ER-positive patients with relatively low level of ER genes (Green).

3.1.3.1 Two-way Hierarchical clustering

Two-way clustering is a highly computationally-intensive process if the number of genes and/or number of samples included is very large. For this reason, a more stringent differentially expressed genelist ($p \leq 0.05$, Fold Change (FC) ≥ 2 and Difference > 100) was generated comparing the normal vs. tumour gene expression profiles, yielding a total of 3924 differentially expressed (2166 up-regulated and 1758 down-regulated) genes and separate 2-way clustering analyses were performed. The gene clusters with enriched genes for similar functions and pathways was identified (see section 2.2.4).

3.1.3.1.1 Up-regulated Genes

Two-way hierarchical clustering was performed on all specimens using the differentially expressed genelist ($p \leq 0.05$, Fold Change (FC) > 2 and Difference > 100) comparing cancer specimens and normal specimens. The distance metric used was 1-correlation and the clustering algorithm used was Average linkage clustering. Prior to clustering, the individual samples were standardised as follows: the expression of the individual genes was subtracted from their respective means and divided by their respective standard deviation. 2166 genes passed the filtration criteria and the two-way hierarchical clustering was performed using this list (Fig 3.1.3.1.1)

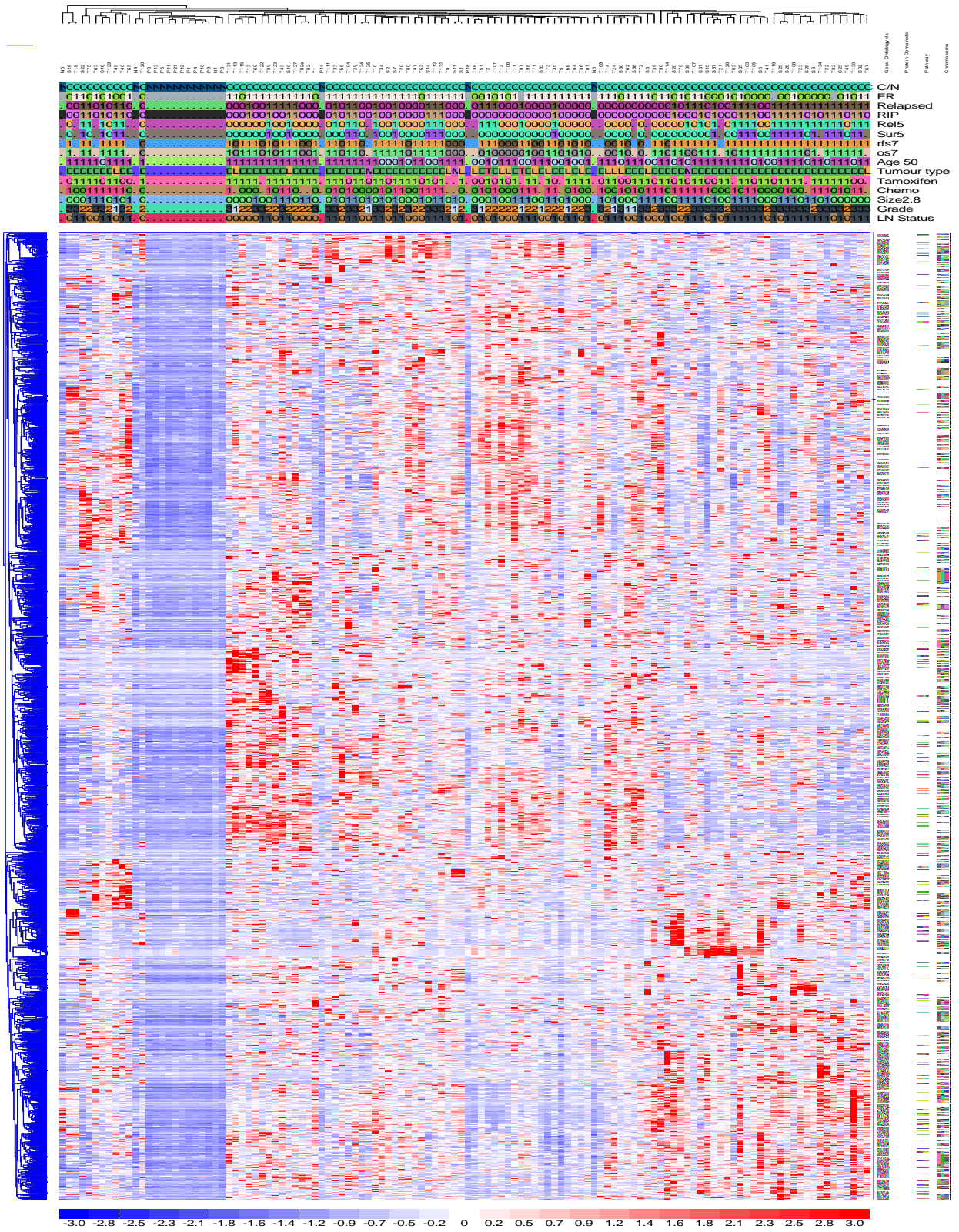


Fig 3.1.3.1.1: Two-way clustering of samples and genes using up-regulated genes only.

As can be seen from Fig. 3.1.3.1.1, normal samples and tumour samples largely clustered separately. Four (n5, 14, p18, n6) of the normal specimens did not cluster with the rest of the normals. The second most important clinical parameter on which the clustering was visually observed was Estrogen Receptor. Clusters of co-expressed genes with similar functions/pathways were identified in the dendrogram. There were a total of 487 clusters of genes with significant Gene ontology terms ($p < 0.001$). The functions along with the number of genes are listed in Appendix 1. 10 pathways were also identified which were significant ($p < 0.001$) in various clusters Table 3.1.3.1.1. The individual clusters are shown in Appendix 2 (Fig A1 - Fig A9). "*" indicates this cluster includes all genes. Therefore they have not been shown in Appendix 2

Pathway	Changed	Measured	p-value
Cell cycle *	22	220	0.000014
DNA replication	7	103	0.000862
Glycosphingolipid metabolism	6	102	0.000905
Inflammatory Response Pathway	7	15	0
Circadian Exercise	4	15	0.000289
Androgen and estrogen metabolism	4	30	0.000271
Proteasome Degradation	5	17	0.000045
Nitrogen metabolism	5	69	0.000892
O-Glycan biosynthesis	4	58	0.000695
Nuclear Receptors	4	14	0.000335

Table 3.1.3.1.1: Significant pathways in clusters of genes among the up-regulated genes.

3.1.3.1.2 Down-regulated Genes

Similarly, two-way clustering was performed separately on down regulated genes ($p \leq 0.05$, Fold Change (FC) > -2 and Difference > -100) comparing cancer specimens and normal specimens. The distance metric used was 1-correlation and the clustering algorithm used was Average linkage clustering. Prior to clustering, the individual

samples were standardised as follows: the expression of the individual genes was subtracted from their means for that sample and divided by their respective standard deviation. 1758 genes passed the filtration criteria and the two-way hierarchical clustering was performed on them (Fig 3.1.3.1.2)

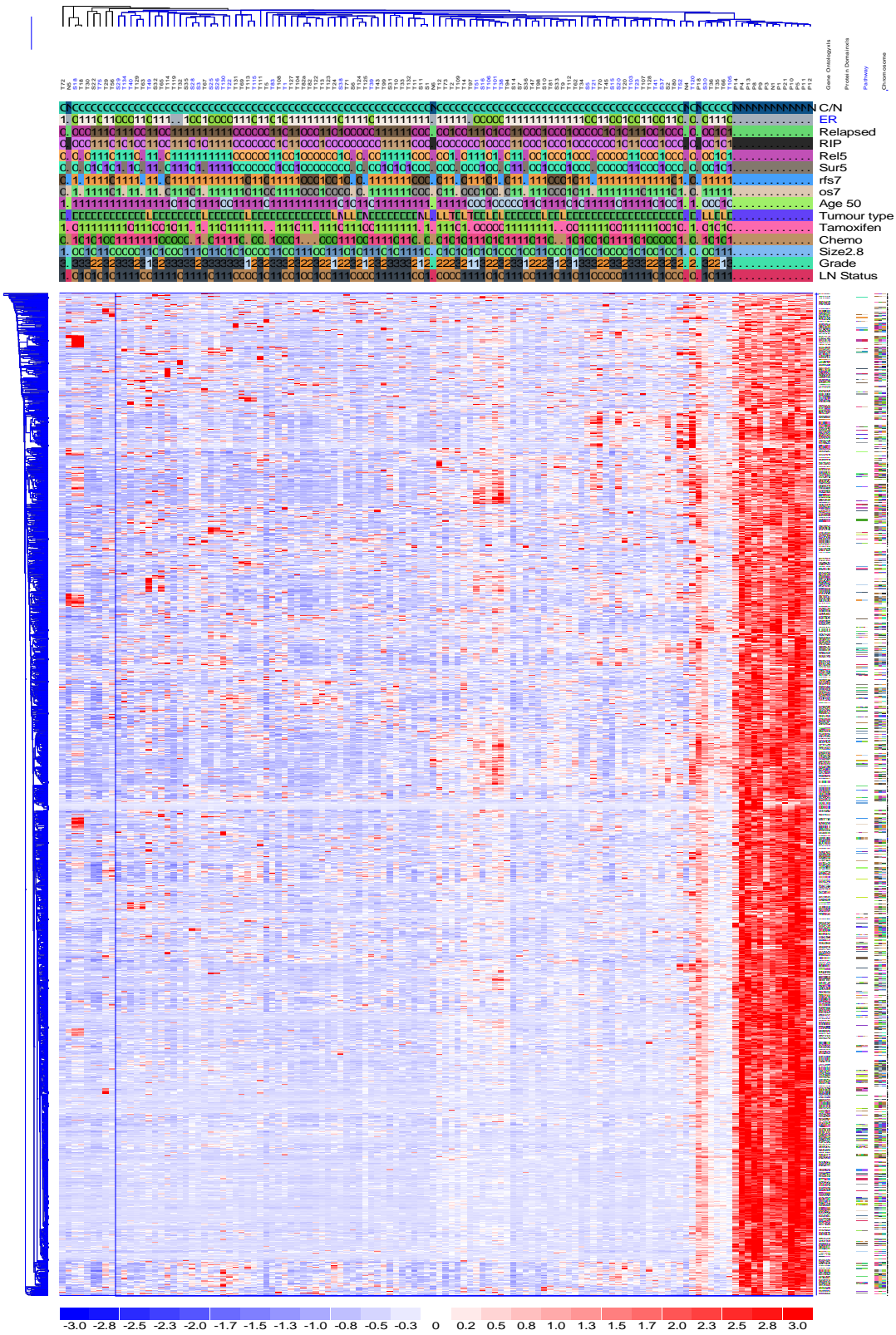


Fig 3.1.3.1.2: Two-way clustering of sample and genes on down-regulated genes.

As before, the normals and tumors largely clustered as distinct groups, with 4 of the normals (n5, n4, p18, n6) not clustering with the rest of the normals. No clinical parameter was found to be significant among the clusters (see section 2.2.4). Clusters of co-expressed genes with similar functions/pathways were identified in the dendrogram. There were a total of 382 clusters of genes with significant Gene ontology terms ($p < 0.001$). The functions along with the number of genes are listed in Appendix 1. 25 pathways were also identified which were significant ($p < 0.001$) in various clusters Table 3.1.3.1.2. The individual clusters are shown in Appendix 3 (Fig A1 - Fig A6). "*" indicates this cluster includes all genes. Therefore they have not been shown in Appendix 3. These seem to be normal metabolic processes which are down-regulated in tumour samples.

Pathway	Changed	Measured	p-value
Glycolysis / Gluconeogenesis *	18	272	0.000064
Fatty acid metabolism *	24	272	0
Glycogen Metabolism *	15	272	0.000511
Pyruvate metabolism *	13	272	0.000004
Propanoate metabolism *	10	272	0.000001
Tyrosine metabolism *	14	272	0.000035
Fatty Acid Degradation *	11	272	0.000058
Glutathione metabolism *	11	272	0.000348
Krebs-TCA Cycle *	12	272	0.000163
Citrate cycle / TCA cycle *	11	272	0.000003
Glycerolipid metabolism *	22	272	0.000041
Lysine degradation *	10	272	0.000626
Fatty Acid Synthesis *	12	272	0.000001
Bile acid biosynthesis *	15	272	0
Valine, leucine and isoleucine degradation *	17	272	0
Ascorbate and aldarate metabolism *	5	272	0.000183
Small ligand GPCRs *	10	272	0.000101
Methane metabolism	6	210	0.000985

Histidine metabolism	5	104	0.00087
Tryptophan metabolism	7	71	0.000827
Glycine, serine and threonine metabolism	5	60	0.000881
Phenylalanine metabolism	4	49	0.000889
Arginine and proline metabolism	4	29	0.00065
Integrin-mediated cell adhesion	5	18	0.000975
GPCRs Class A Rhodopsin-like	4	4	0.000019

Table 3.1.3.1.2: Significant pathways in clusters of genes among the up-regulated genes.

3.1.4 Comparison criteria: Normal vs. cancer specimens

Identifying genes up-regulated or down-regulated in cancer vs. normal helps us in better understanding the cancer dynamics and help identify markers and treatment targets for breast cancer.

A total of 17 normal breast specimens and 104 breast cancer specimens were compared for gene expression changes using data generated from an in-house microarray experiment as previously detailed.

Up-regulated gene transcripts:

4,213 genes were identified as significantly up-regulated ($p \leq 0.05$, Fold Change (FC) > 1.2 and Difference > 100) in cancer specimens compared to normal specimens. Genes were ranked by fold change and, based on this criterion, the top 20 genes are listed in Table 3.1.4.1.

Gene ontology analysis was performed on the up-regulated genes ($p \leq 0.05$, FC >2 , and Difference > 100). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions represented are listed in Table 3.1.4.2.

Pathway analysis was performed using GenMAPP on the up-regulated genes ($p \leq 0.05$, FC >2 , and Difference > 100). Significant pathways were identified based on p-value ($p \leq$

0.05) and the 10 most significant pathways are listed in Table 3.1.4.3. Embryonic stem cells (Fig 3.1.4.1) and Cell cycle (Fig 3.1.4.2) pathways were observed to be enriched by the up-regulated genes.

Down-regulated gene transcripts:

3235 genes were identified as significantly down-regulated ($p \leq 0.05$, $FC < -1.2$ and $Difference < -100$) in cancer specimens compared to normal specimens. Genes were ranked by fold change and, based on this criterion, the top 20 genes are listed in Table 3.1.4.4.

Gene ontology analysis was performed on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and $Difference < -100$). Significant functions were identified based on p-value ($p \leq 0.05$) and 10 most significant functions represented are listed in Table 3.1.4.5.

Pathway analysis was performed using GenMAPP on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and $Difference < -100$). Significant pathways were identified based on p-value ($p \leq 0.05$) and the 10 most significant pathways are listed in Table 3.1.4.6. Fatty acid Biosynthesis (Fig 3.1.4.3) pathways were observed to be enriched by the down-regulated genes.

Probe set	gene	Baseline	Experiment	FC	p-value
205916_at	S100A7	1	406.87	406.87	0.000514
205941_s_at	COL10A1	2.65	478.64	180.62	0
226548_at	SBK1	1.17	160.03	136.8	0
213201_s_at	TNNT1	2.31	314.01	135.96	0
208502_s_at	PITX1	1	106.63	106.21	0
207802_at	CRISP3	2.01	182.44	90.77	0.004065
239983_at	SLC30A8	5.27	350.19	66.44	0.000373
204915_s_at	SOX11	3.4	186.95	54.92	0
236885_at	LOC92312	3.16	160.97	50.96	0
220318_at	EPN3	2.9	138.45	47.74	0
204351_at	S100P	16.14	757.64	46.94	0
229341_at	TFCP2L1	2.31	103.09	44.61	0.000727
231352_at	SLC22A8	3.62	141.75	39.12	0
217428_s_at	COL10A1	12.95	498.17	38.47	0
206502_s_at	INSM1	4.43	151.87	34.32	0.026509
228969_at	AGR2	20.64	705.66	34.19	0
220414_at	CALML5	8.25	279.3	33.86	0.000001
1558281_a_					
at	MGC9712	3.94	132.13	33.58	0
229158_at	WNK4	4.39	143.38	32.65	0
204913_s_at	SOX11	4.77	153.22	32.14	0.00001

Table 3.1.4.1: Genes up-regulated in cancer specimens in comparison to normal specimens

GOID	GO Name	Changed	Measured	p-value
5581	Collagen	12	30	0
785	Chromatin	30	124	0
278	Mitotic cell cycle	38	177	0
7067	Mitosis	31	135	0
87	M phase of mitotic cell cycle	31	137	0
5584	Collagen type I	3	3	0
5694	Chromosome	43	230	0
279	M phase	35	176	0
6333	Chromatin assembly or disassembly	25	114	0
8094	DNA-dependent ATPase activity	11	35	0

Table 3.1.4.2 Functions enriched among genes up-regulated in cancer in comparison to normal specimens

MAPP Name	Changed	Measured	p-value
1-Tissue-Embryonic Stem Cell	15	47	0
Cell cycle KEGG	22	89	0
M phase of mitotic cell cycle	25	124	0
Chromatin	29	152	0
mRNA processing Reactome	25	125	0
2-Tissues-Blood and Lymph	16	78	0
Rhodopsin-like receptor activity	5	240	0
GPCRDB Class A Rhodopsin-like	3	253	0
Chromatin assembly or disassembly	24	139	0.001
Establishment and or maintenance of chromatin architecture	34	217	0.001

Table 3.1.4.3: Pathways enriched among genes up-regulated in cancer in comparison to normal specimens

Embryonic Stem Cell

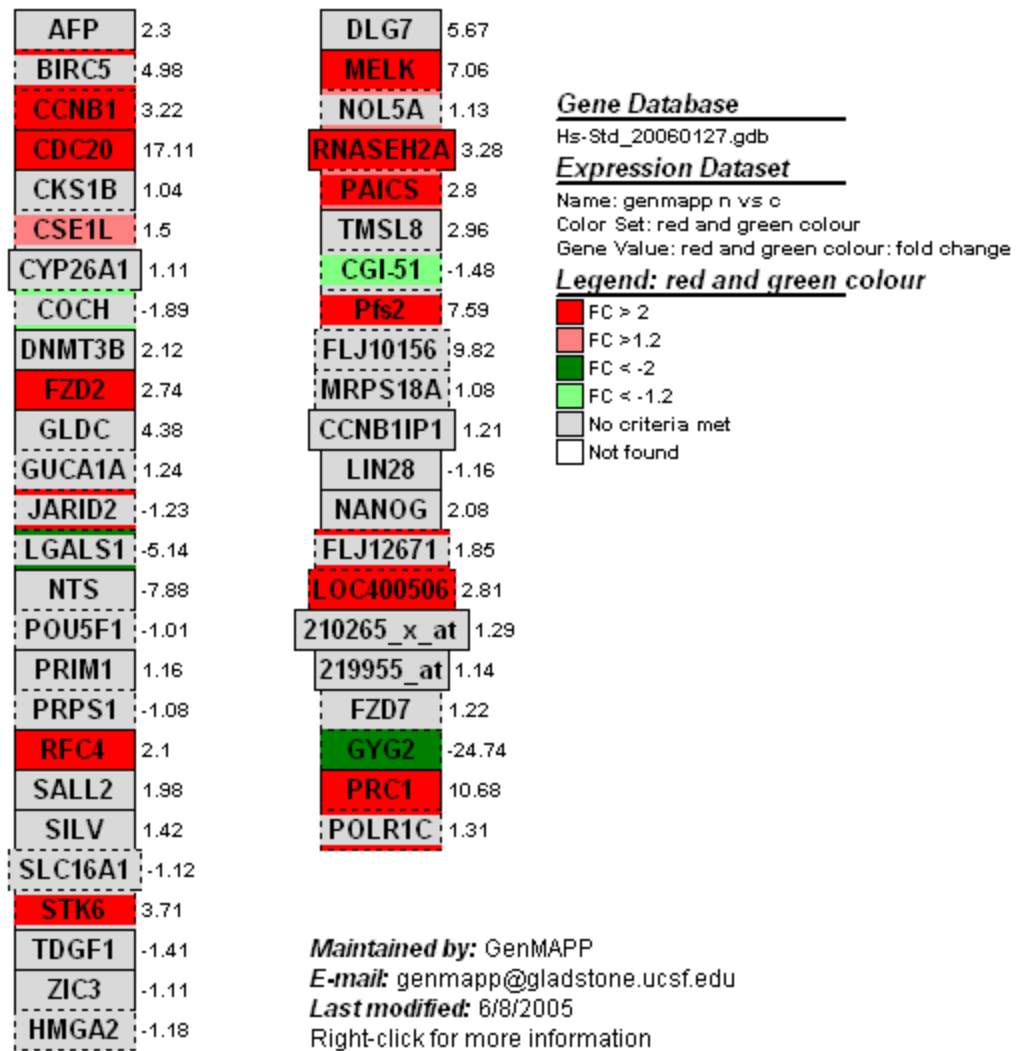


Fig 3.1.4.1: Embryonic stem cells pathway. Red indicates up-regulated genes in cancer specimens in comparison to normal specimens. Green indicates down-regulated genes in cancer specimens in comparison to normal specimens.

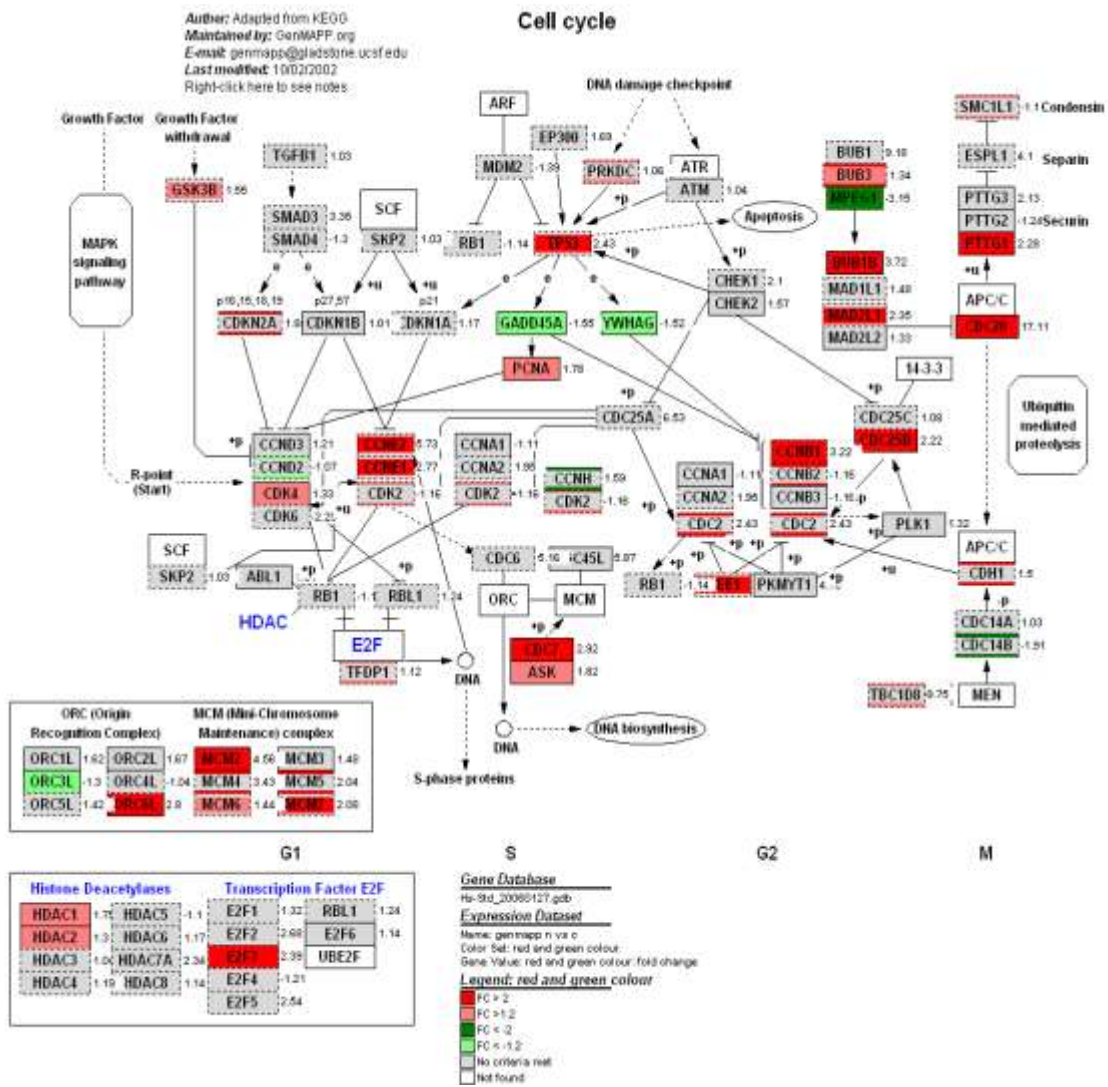


Fig 3.1.4.2: Cell cycle pathways. Red indicates up-regulated genes in cancer specimens in comparison to normal specimens. Green indicates down-regulated genes in cancer specimens in comparison to normal specimens.

Probe set	Gene	Baseline	Experiment	FC	p-value
228880_at	LOC339984	248.01	1.42	-174.51	0.000056
222083_at	GLYAT	445.31	2.58	-172.89	0.000075
204997_at	GPD1	721.75	5.14	-140.45	0.000013
1558421_a_at	LOC400258	1347.45	13.29	-101.39	0.000089
220736_at	SLC19A3	366.24	3.63	-100.91	0.000028
213515_x_at	HBG1, HBG2	1036.62	11.04	-93.92	0.022274
201539_s_at	FHL1	1048.5	12.75	-82.21	0.000024
243311_at	RP5- 1103G7.6	1258.94	18.22	-69.09	0.000183
222226_at	SAA3P	145.13	2.13	-68.07	0.00477
243813_at	---	129.99	1.95	-66.7	0.004182
210298_x_at	FHL1	1198.5	18.47	-64.88	0.000029
210106_at	RDH5	831.47	12.95	-64.21	0.000011
235708_at	KLB	736.68	12.68	-58.09	0.00003
234943_at	---	108.54	1.87	-57.96	0.005823
237154_at	HSD11B1	1378.6	24.14	-57.11	0.000022
219140_s_at	RBP4	3687.79	65	-56.74	0.000016
228168_at	ATP5G3	102.45	1.83	-55.88	0.002188
208383_s_at	PCK1	1328	25.05	-53.01	0.000014
207092_at	LEP	4466.38	88.5	-50.47	0.000006
209980_s_at	SHMT1	111.96	2.42	-46.31	0.000878

Table 3.1.4.4: Genes down-regulated in cancer specimens in comparison to normal specimens

GOID	GO Name	Changed	Measured	p-value
9109	Coenzyme catabolism	12	22	0
46356	Acetyl-CoA catabolism	11	21	0
6099	Tricarboxylic acid cycle	11	21	0
16628	Oxidoreductase activity, acting on the CH-CH Group of donors, NAD or NADP as acceptor	9	15	0
51187	Cofactor catabolism	12	26	0
6629	Lipid metabolism	82	535	0
9060	Aerobic respiration	12	28	0
16491	Oxidoreductase activity	85	579	0
4300	Enoyl-CoA hydratase activity	5	6	0

Table 3.1.4.5: Functions enriched among genes down-regulated in cancer in comparison to normal specimens

MAPP Name	Changed	Measured	p-value
Fatty Acid Synthesis BiGCaT	15	22	0
Propanoate metabolism	14	27	0
Citrate cycle TCA cycle	13	24	0
Mitochondrial fatty acid betaoxidation	10	16	0
Valine leucine and isoleucine degradation	17	39	0
Fatty Acid Beta Oxidation Meta BiGCaT	14	32	0
Adipogenesis	34	130	0
Pyruvate metabolism	14	34	0
1-Tissue-Muscle fat and connective	20	65	0
Fatty acid metabolism	20	66	0

Table 3.1.4.6: Pathways enriched among genes down-regulated in cancer in comparison to normal specimens

Author: Kim Dahlquist and Chris Ewels et al
 Maintained by: Chris Ewels for NCI
 E-mail: mapps@bioprot.uminn.edu
 Last modified: 11/22/05
 http://www.bioprot.uminn.edu/mapps/
 Copyright © Gladstone Institute and Biocat Bioinformatics

Fatty Acid Biosynthesis

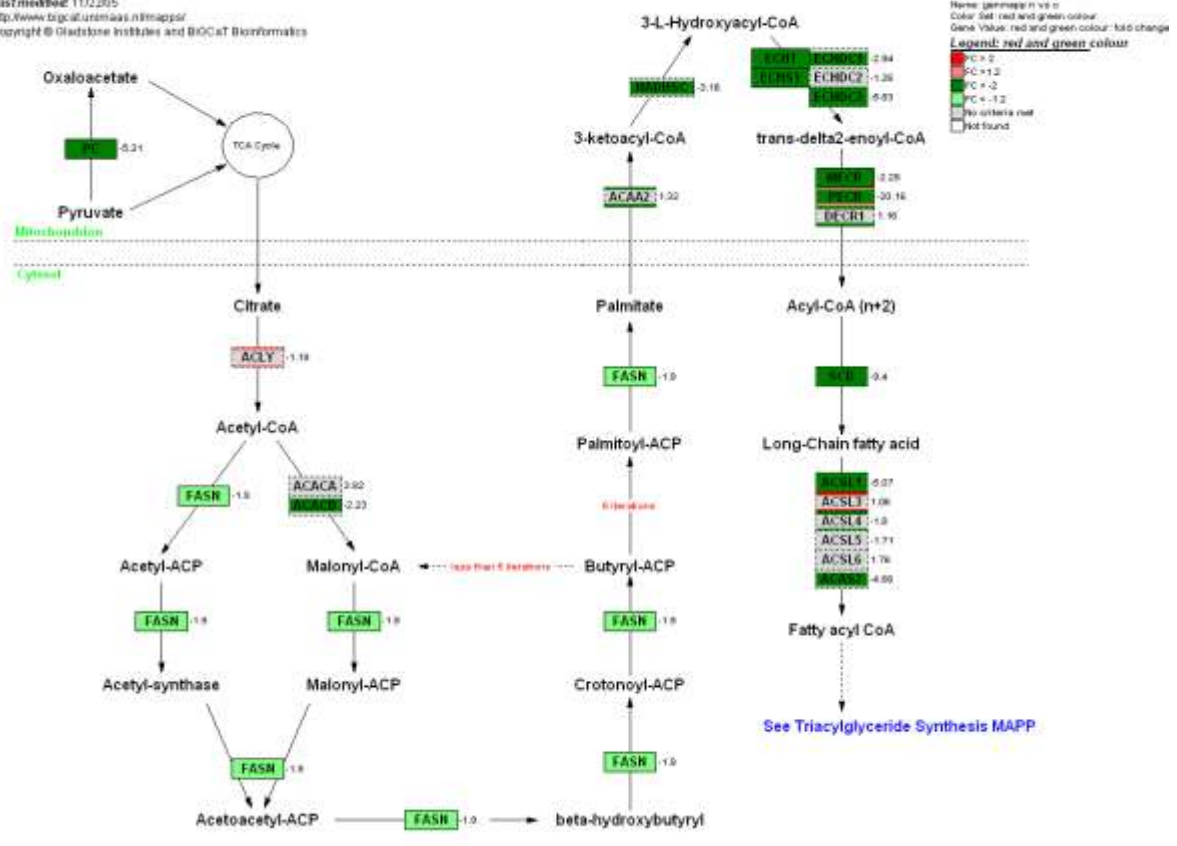


Fig 3.1.4.3: Fatty acid biosynthesis pathways. Red indicates up-regulated genes in cancer specimens in comparison to normal specimens. Green indicates down-regulated genes in cancer specimens in comparison to normal specimens.

3.1.5 Comparison criteria: Estrogen receptor-negative vs. Estrogen receptor-positive

Estrogen receptor (ER) status is important in identifying patients who are likely to respond from endocrine therapy. Identifying genes up and down regulated in ER-positive vs. ER-negative patients is important to get a better understanding of the significance of ER pathway.

A total of 34 ER-negative breast specimens and 67 ER-positive breast cancer specimens were compared for gene expression changes.

Up-Regulated gene transcripts:

855 genes were identified as significantly up-regulated ($p \leq 0.05$, Fold Change (FC) > 1.2 and Difference > 100) in ER-positive breast cancer specimens compared to ER-negative specimens. Genes were ranked by fold change and, based on these criteria, the top 20 genes are listed in Table 3.1.5.1.

Gene ontology analysis was performed on the up-regulated genes ($p \leq 0.05$, FC >2 , and Difference > 100). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions are listed in Table 3.1.5.2.

Pathway analysis was performed using GenMAPP on the up-regulated genes ($p \leq 0.05$, FC >2 , and Difference > 100). Significant pathways were identified based on p-value ($p \leq 0.05$) and significant pathways are listed in Table 3.1.5.3. The Nuclear receptors pathway was observed to be enriched by the up-regulated genes.

Down-regulated gene transcripts:

1145 genes were identified as significantly down-regulated ($p \leq 0.05$, Fold Change (FC) < -1.2 and Difference < -100) in ER-positive breast cancer specimens compared to ER-negative specimens. Genes were ranked by fold change and, based on these criteria, the top 20 genes are listed in Table 3.1.5.4.

Gene ontology analysis was performed on the down-regulated genes ($p \leq 0.05$, FC <-2 , and Difference < -100). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions are listed in Table 3.1.5.5.

Pathway analysis was performed using GenMAPP database on the down-regulated genes ($p \leq 0.05$, FC <-2 , and Difference < -100). Significant pathways were identified based on p-value ($p \leq 0.05$) and the significant pathways are listed in Table 3.1.5.6. Genes specific to blood and lymph tissue pathway were observed to be enriched by the down-regulated genes.

Probe set	Gene	Baseline	Experimental	FC	p-value
236445_at	LOC731986	6.71	264.52	39.43	0.000901
1494_f_at	CYP2A6	49.93	606.61	12.15	0.024179
242301_at	CBLN2	10.41	116.75	11.21	0.012938
210272_at	CYP2B7P1	39.93	438.38	10.98	0.000165
1552507_at	KCNE4	65.09	511.13	7.85	0.000406
1562821_a_at	DSCAM	69.35	536.03	7.73	0.017704
206754_s_at	CYP2B7P1	189.29	1445.09	7.63	0.000069
205696_s_at	GFRA1	42.88	291.22	6.79	0
230163_at	LOC143381	82.42	558.14	6.77	0
239983_at	SLC30A8	76.2	505.49	6.63	0.004899
218332_at	BEX1	79.39	510.9	6.44	0.00405
240192_at	FLJ45983	24.51	135.67	5.54	0.000012
222379_at	---	60.8	331.69	5.46	0.000691
220540_at	KCNK15	42.8	215.18	5.03	0.000035
211712_s_at	ANXA9	52.94	251.19	4.74	0.000007
1555997_s_at	IGFBP5	51.64	230.38	4.46	0.001196
227550_at	LOC143381	211.8	933.96	4.41	0.000002
203999_at	---	35.44	147.18	4.15	0.012644
241368_at	LSDP5	55.83	222.53	3.99	0.000033
226271_at	GDAP1	79.83	307.42	3.85	0.017169

Table 3.1.5.1: Genes up-regulated in ER-positive specimens in comparison to ER-negative specimens

GOID	GO Name	Changed	Measured	p-value
30027	Lamellipodium	2	15	0
902	Cellular morphogenesis	7	236	0
5006	Epidermal growth factor receptor activity	2	7	0.002
31252	Leading edge	2	18	0.002
9653	Morphogenesis	9	522	0.002
7584	Response to nutrient	2	15	0.003
17162	Aryl hydrocarbon receptor binding	1	1	0.004
42995	Cell projection	3	65	0.004
1786	Phosphatidylserine binding	1	1	0.005
31667	Response to nutrient levels	2	19	0.005

Table 3.1.5.2: Functions enriched among genes up-regulated in ER-positive specimens in comparison to ER-negative specimens.

MAPP Name	Changed	Measured	p-value
Nuclear Receptors	2	38	0.039
IL-3 NetPath 15	3	101	0.039
Synthesis and Degradation of Ketone Bodies KEGG	1	5	0.04
Butanoate metabolism	2	38	0.042
2-Tissues-Endocrine and CNS	3	103	0.042

Table 3.1.5.3 Pathways enriched among genes up-regulated in ER-positive specimens in comparison to ER-negative specimens.

Probe set	Gene	Baseline	Experimental	FC	p-value
216365_x_at	IGL@, CPVL	228.98	31.35	-7.3	0.042385
220425_x_at	ROPN1B	164.84	24.84	-6.64	0.012712
205363_at	BBOX1	205.39	33.76	-6.08	0.004225
224191_x_at	ROPN1	187.87	31.7	-5.93	0.027217
213711_at	KRT81	587.61	103.04	-5.7	0.022424
231535_x_at	ROPN1	210.24	37.59	-5.59	0.023043
219225_at	PGBD5	169.43	31.97	-5.3	0.033009
206165_s_at	CLCA2	175.85	34.81	-5.05	0.044217
212531_at	LCN2	219.97	43.92	-5.01	0.030348
210147_at	ART3	126.27	25.23	-5	0.040628
237625_s_at	---	427.38	86.05	-4.97	0.038949
220559_at	EN1	141	29.54	-4.77	0.03141
214777_at	---	273.81	59.04	-4.64	0.033383
217528_at	CLCA2	232.94	50.85	-4.58	0.039492
211881_x_at	IGLJ3	532.05	118.48	-4.49	0.047763
223468_s_at	RGMA	482.4	108.43	-4.45	0.00109
202037_s_at	SFRP1	1122.95	256.41	-4.38	0.005489
209396_s_at	CHI3L1	802.87	188.9	-4.25	0.022891
235209_at	RPESP	316.05	76.29	-4.14	0.029573
211798_x_at	IGLJ3	611.25	150.68	-4.06	0.044385

Table 3.1.5.4: Genes down-regulated in ER-positive specimens in comparison to ER-negative specimens

GOID	GO Name	Changed	Measured	p-value
45012	MHC class II receptor activity	3	13	0
19884	Antigen presentation, exogenous antigen	3	13	0
19886	Antigen processing, exogenous antigen via MHC class II	3	14	0
19221	Cytokine and chemokine mediated signaling pathway	3	15	0
9607	Response to biotic stimulus	24	845	0
6952	Defense response	22	806	0
6955	Immune response	19	716	0
50896	Response to stimulus	34	1754	0
6032	Chitin catabolism	2	7	0.001
4568	Chitinase activity	2	7	0.001

Table 3.1.5.5: Functions enriched among genes down-regulated in ER-positive specimens in comparison to ER-negative specimens.

MAPP Name	Changed	Measured	p-value
2-Tissues-Blood and Lymph	14	78	0
1-Tissue-Blood and Lymph	13	168	0
Phosphatidylinositol signaling system	7	122	0.004
Kit-Receptor NetPath 6	4	67	0.017
B Cell Receptor NetPath 12	6	158	0.039

Table 3.1.5.6: Pathways enriched among genes down-regulated in ER-positive specimens in comparison to ER-negative specimens.

3.1.6 Comparison criteria: Lymph node-negative vs. Lymph node-positive

Positive lymph node status indicates the spread of disease and is an indicator of aggressive disease. Identifying genes up and down regulated in Lymph node-positive vs. lymph node-negative patients may help identify biomarkers and targets for aggressive disease.

A total of 45 lymph node-negative specimens and 59 lymph node-positive specimens were compared for gene expression changes.

Up-regulated gene transcripts:

102 genes were identified as significantly up-regulated ($p \leq 0.05$, Fold Change (FC) > 1.2 and Difference > 100) in lymph node-positive specimens compared to lymph node-negative specimens. Genes were ranked by fold change and, based on this criterion, the top 20 genes are listed in Table 3.1.6.1.

Gene ontology analysis was performed on the up-regulated genes ($p \leq 0.05$, FC >2 , and Difference > 100). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions are listed in Table 3.1.6.2.

Pathway analysis was performed using GenMAPP on the up-regulated genes ($p \leq 0.05$, FC >2 , and Difference > 100). Significant pathways were identified based on p-value ($p \leq 0.05$) and listed in Table 3.1.6.3

Down-regulated gene transcripts:

126 genes were identified as significantly down-regulated ($p \leq 0.05$, Fold Change (FC) < -1.2 and Difference < -100) in lymph node-positive specimens compared to lymph node-negative specimens. Genes were ranked by fold change and, based on these criteria, the top 20 genes are listed in Table 3.1.6.4.

Gene ontology analysis was performed on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and Difference < -100). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions are listed in Table 3.1.6.5.

Pathway analysis was performed using GenMAPP on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and Difference < -100). No pathway was found to be significantly affected ($p \leq 0.05$).

Probe set	Gene	Baseline	Experiment	FC	P value
208161_s_at	ABCC3	114.65	288.11	2.51	0.030769
232547_at	SNIP	78.47	197.18	2.51	0.031976
201467_s_at	NQO1	104.07	252.98	2.43	0.014538
213557_at	CRKRS	105.28	252.03	2.39	0.013519
213551_x_at	PCGF2	172.75	355.64	2.06	0.009378
204942_s_at	ALDH3B2	367.35	737.26	2.01	0.012173
201525_at	APOD	607.61	1214.76	2	0.03805
236885_at	LOC92312	104.67	204.99	1.96	0.003213
201080_at	PIP5K2B	460.62	888.4	1.93	0.013447
202991_at	STARD3	126.81	241.75	1.91	0.005253
226727_at	LOC284106	230.36	437.84	1.9	0.03065
214239_x_at	PCGF2	354.75	670.7	1.89	0.012442
228854_at	---	129.08	243.04	1.88	0.037461
210519_s_at	NQO1	424.43	779.09	1.84	0.011794
224784_at	MLLT6	199.44	367.65	1.84	0.01422
204351_at	S100P	528.01	935.78	1.77	0.045632
227512_at	LOC92312	160.25	281.39	1.76	0.003296
226346_at	LOC92312	266.36	463.27	1.74	0.006202
222706_at	CCDC49	175.35	302.95	1.73	0.013138
201400_at	PSMB3	1252.09	2107.56	1.68	0.003089

Table 3.1.6.1: Genes up-regulated in lymph node-positive specimens in comparison to lymph node-negative specimens

GOID	GO Name	Changed	Measured	p-value
3955	NAD(P)H dehydrogenase (quinone) activity	1	2	0.001
4128	Cytochrome-b5 reductase activity	1	5	0.002
4030	Aldehyde dehydrogenase [NAD(P)+] activity	1	5	0.003
46209	Nitric oxide metabolism	1	13	0.003
6809	Nitric oxide biosynthesis	1	13	0.003
7271	Synaptic transmission, cholinergic	1	8	0.004
16652	Oxidoreductase activity, acting on NADH or NADPH, NAD or NADP as acceptor	1	8	0.004
7270	Nerve-nerve synaptic transmission	1	10	0.004
8514	Organic anion transporter activity	1	10	0.006
4028	Aldehyde dehydrogenase activity	1	13	0.006

Table: 3.1.6.2: Functions enriched among genes up-regulated in lymph node-positive specimens in comparison to lymph node-negative specimens

MAPP Name	Changed	Measured	p-value
Sterol biosynthesis	1	19	0.015
Phenylalanine metabolism	1	21	0.023
Nuclear receptors in lipid metabolism and toxicity	1	33	0.029
Oxidative Stress	1	28	0.031
Tyrosine metabolism	1	50	0.046
Histidine metabolism	1	43	0.048

Table: 3.1.6.3: Pathways enriched among genes up-regulated in lymph node-positive specimens in comparison to lymph node-negative specimens

Probe set	Gene	Baseline	Experiment	FC	P value
1562309_s_at	PHF21B	203.88	33.16	-6.15	0.029485
205710_at	LRP2	186.27	51.63	-3.61	0.027245
226269_at	GDAP1	439.22	124.8	-3.52	0.048222
229947_at	PI15	1154.42	334.98	-3.45	0.006649
221796_at	NTRK2	342.72	102.47	-3.34	0.047688
205794_s_at	NOVA1	496.28	183.17	-2.71	0.023049
230863_at	---	484.5	185.15	-2.62	0.024507
232687_at	---	241.09	91.97	-2.62	0.021802
205567_at	CHST1	253.67	112.66	-2.25	0.043266
211421_s_at	RET	321.52	144.02	-2.23	0.017844
205472_s_at	DACH1	518.9	258.04	-2.01	0.032869
213832_at	---	364.12	185	-1.97	0.01482
225123_at	---	432.05	227.79	-1.9	0.014714
1570344_at	---	314.92	168.16	-1.87	0.01455
205471_s_at	DACH1	381.97	205.01	-1.86	0.035402
228915_at	DACH1	474.53	254.77	-1.86	0.024403
225613_at	MAST4	691.74	373.56	-1.85	0.031486
227192_at	PRRT2	291.04	161.4	-1.8	0.04585
1554007_at	ZNF483	250.9	140.49	-1.79	0.009076
244696_at	AFF3	379.29	219.24	-1.73	0.016035

Table 3.1.6.4: Genes down-regulated in lymph node-positive specimens in comparison to lymph node-negative specimens

GOID	GO Name	Changed	Measured	p-value
45130	Keratan sulfotransferase activity	1	1	0
7456	Eye development (sensu Endopterygota)	1	1	0
43121	Neurotrophin binding	1	3	0
7497	Posterior midgut development	1	1	0.001
7494	Midgut development	1	1	0.001
42339	Keratan sulfate metabolism	1	2	0.001
48565	Gut development	1	2	0.002
30304	Trypsin inhibitor activity	1	3	0.002
6012	Galactose metabolism	1	7	0.003
1654	Eye development	1	9	0.004

Table: 3.1.6.5: Functions enriched among genes down-regulated in lymph node-positive specimens in comparison to lymph node-negative specimens

3.1.7 Comparison criteria: Grade 1 vs. Grade 2

Higher Grade cancers are more aggressive. Identifying genes up and down regulated in patients with high grade tumors vs. low grade tumors may help identify biomarkers and targets for aggressive disease.

A total of 11 specimens with Grade 1 cancer and 40 specimens with Grade 2 cancers were compared for gene expression changes.

Up-regulated gene transcripts:

275 genes were identified as significantly up-regulated ($p \leq 0.05$, Fold Change (FC) > 1.2 and Difference > 100) in Grade 2 specimens compared to Grade 1 specimens. Genes were ranked by fold change and, based on this criterion, the top 20 genes are listed in Table 3.1.7.1.

Gene ontology analysis was performed on the up-regulated genes ($p \leq 0.05$, $FC > 2$, and Difference > 100). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions are listed in Table 3.1.7.2.

Pathway analysis was performed using GenMAPP on the up-regulated genes ($p \leq 0.05$, $FC > 2$, and Difference > 100). Significant pathways were identified based on p-value ($p \leq 0.05$) listed in Table 3.1.7.3.

Down-regulated gene transcripts:

75 genes were identified as significantly down-regulated ($p \leq 0.05$, Fold Change (FC) < -1.2 and Difference < -100) in Grade 2 specimens compared to Grade 1 specimens. Genes were ranked by fold change and, based on these criteria, the top 20 genes are listed in Table 3.1.7.4.

Gene ontology analysis was performed on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and Difference < -100). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions are listed in Table 3.1.7.5.

Pathway analysis was performed using GenMAPP on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and Difference < -100). No pathway was found to be significantly affected ($p \leq 0.05$).

Probe set	Gene	Baseline	Experiment	FC	p-value
221107_at	CHRNA9	11.95	163.02	13.64	0.02427
210576_at	CYP4F8	15.69	162.6	10.36	0.038742
202917_s_at	S100A8	117.23	819.95	6.99	0.027742
203915_at	CXCL9	95.42	616.34	6.46	0.01134
230966_at	IL4I1	30.59	168.93	5.52	0.015138
1562821_a_at	DSCAM	82.14	447.18	5.44	0.017192
202672_s_at	ATF3	102.7	416.69	4.06	0.000924
204533_at	CXCL10	64.57	261.66	4.05	0.042573
221667_s_at	HSPB8	154.17	605.27	3.93	0.004196
203645_s_at	CD163	64.84	249.67	3.85	0.044332
202768_at	FOSB	75.36	286.7	3.8	0.001608
203936_s_at	MMP9	229.25	870.22	3.8	0.016992
210163_at	CXCL11	42.15	156.18	3.71	0.017085
217388_s_at	KYNU	119.89	416.78	3.48	0.007713
221491_x_at	HLA-DRB1	85.24	296.62	3.48	0.000384
215049_x_at	CD163	86.73	292.72	3.37	0.042521
203290_at	HLA-DQA1	58.37	187.41	3.21	0.021995
202988_s_at	RGS1	56.62	177.43	3.13	0.001173
211143_x_at	NR4A1	62.24	193.85	3.11	0.000432
229476_s_at	THRSP	89.79	274	3.05	0.012882

Table 3.1.7.1: Genes up-regulated in Grade 2 specimens in comparison to Grade 1 specimens

GOID	GO Name	Changed	Measured	p-value
42379	Chemokine receptor binding	5	44	0
8009	Chemokine activity	5	44	0
1664	G-protein-coupled receptor binding	5	52	0
19363	Pyridine nucleotide biosynthesis	2	10	0
6955	Immune response	17	716	0
9607	Response to biotic stimulus	18	845	0
42364	Water-soluble vitamin biosynthesis	2	13	0
6952	Defense response	17	806	0
42330	Taxis	6	116	0
6935	Chemotaxis	6	116	0

Table 3.1.7.2: Functions enriched among genes up-regulated in Grade 2 specimens in comparison to Grade 1 specimens

MAPP Name	Changed	Measured	p-value
1-Tissue-Blood and Lymph	5	168	0.001
BCell Receptor NetPath_12	3	158	0.034

Table 3.1.7.3: Pathways enriched among genes up-regulated in Grade 2 specimens in comparison to Grade 1 specimens

Probe set	Gene	Baseline	Experiment	FC	p-value
221207_s_at	NBEA	202.86	79.62	-2.55	0.038106
206509_at	PIP	4995.37	1998.45	-2.5	0.031477
204014_at	DUSP4	1377.66	585.14	-2.35	0.034649
229331_at	SPATA18	196.9	84.09	-2.34	0.015777
205009_at	TFF1	3627.58	1582.24	-2.29	0.024715
226034_at	---	1763.7	800.52	-2.2	0.023362
201445_at	CNN3	353.94	178.34	-1.98	0.039975
204633_s_at	RPS6KA5	289.74	147.1	-1.97	0.012653
204635_at	RPS6KA5	243.95	126.37	-1.93	0.006307
204686_at	IRS1	603.2	330.01	-1.83	0.023378
243495_s_at	---	274.03	154.92	-1.77	0.01535
208978_at	CRIP2	243.56	139.62	-1.74	0.031256
204623_at	TFF3	3889.12	2265.98	-1.72	0.046233
238044_at	---	260.61	154.2	-1.69	0.047571
227769_at	---	254.05	151.79	-1.67	0.024086
202936_s_at	SOX9	838.74	508.58	-1.65	0.012404
227856_at	C4orf32	882.22	533.75	-1.65	0.047376
226989_at	RGMB	408.82	249.54	-1.64	0.02876
228496_s_at	CRIM1	1155.66	722.16	-1.6	0.021534
229478_x_at	BIVM	296.93	190.88	-1.56	0.025046

Table 3.1.7.4: Genes down-regulated in Grade 2 specimens in comparison to Grade 1 specimens

GOID	GO Name	Changed	Measured	p-value
8330	Protein tyrosine/threonine phosphatase activity	1	3	0.001
51018	Protein kinase A binding	1	9	0.002
17017	MAP kinase phosphatase activity	1	10	0.002
5802	Golgi trans face	1	14	0.002
6892	Post-Golgi transport	1	15	0.006
8138	Protein tyrosine/serine/threonine phosphatase activity	1	39	0.012
19901	Protein kinase binding	1	43	0.014
7586	Digestion	1	52	0.014
19900	Kinase binding	1	49	0.017
4725	Protein tyrosine phosphatase activity	1	76	0.019

Table 3.1.7.2: Functions enriched among genes down-regulated in Grade 2 specimens in comparison to Grade 1 specimens

3.1.8 Comparison criteria: Grade 2 vs. Grade 3

Higher Grade cancers are more aggressive. Identifying genes up- and down-regulated in patients with high grade tumors vs. low grade tumors may help identify biomarkers and targets for aggressive disease.

A total of 40 specimens with Grade 2 cancer and 53 specimens with Grade 3 cancer were compared for gene expression changes.

Up-regulated gene transcripts:

930 genes were identified as significantly up-regulated ($p \leq 0.05$, Fold Change (FC) > 1.2 and Difference > 100) in Grade 3 specimens compared to Grade 2 specimens. Genes were ranked by fold change and, based on these criteria, the top 20 genes are listed in Table 3.1.8.1.

Gene ontology analysis was performed on the up-regulated genes ($p \leq 0.05$, $FC > 2$, and Difference > 100). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions are listed in Table 3.1.8.2.

Pathway analysis was performed using GenMAPP database on the up-regulated genes ($p \leq 0.05$, $FC > 2$, and Difference > 100). Significant pathways were identified based on p-value ($p \leq 0.05$) and are listed in Table 3.1.8.3.

Down-regulated gene transcripts:

596 genes were identified as significantly down-regulated ($p \leq 0.05$, Fold Change (FC) < -1.2 and Difference < -100) in Grade 3 specimens compared to Grade 2 specimens. Genes were ranked by fold change and, based on these criteria, the top 20 genes are listed in Table 3.1.8.4.

Gene ontology analysis was performed on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and Difference < -100). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions are listed in Table 3.1.8.5.

Pathway analysis was performed using GenMAPP database on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and Difference < -100). Significant pathways were identified based on p-value ($p \leq 0.05$) and are listed in Table 3.1.8.6.

Probe set	Gene	Baseline	Experiment	FC	p-value
215729_s_at	VGLL1	8.32	121.89	14.65	0.004585
204602_at	DKK1	28.8	212.85	7.39	0.002297
213711_at	KRT81	73.65	492.39	6.69	0.007681
234764_x_at	LOC96610	47.16	206.44	4.38	0.016944
231535_x_at	ROPN1	32.27	138.4	4.29	0.037289
212531_at	LCN2	40.45	171.13	4.23	0.01685
237625_s_at	---	77.24	319.08	4.13	0.027354
223062_s_at	PSAT1	67.35	257.38	3.82	0.000293
204751_x_at	DSC2	41.88	152.19	3.63	0.002025
215189_at	KRT86	45	159.32	3.54	0.009247
213060_s_at	CHI3L2	121.29	427.27	3.52	0.00917
220625_s_at	ELF5	149.21	512.18	3.43	0.005103
226960_at	UNQ473	47.89	150.55	3.14	0.027938
206714_at	ALOX15B	86.79	267.16	3.08	0.03856
216401_x_at	---	122.68	374.9	3.06	0.036479
201195_s_at	SLC7A5	157	461.06	2.94	0.006608
1560818_at	LOC387895	120.25	348.45	2.9	0.023133
213680_at	KRT6B	197.82	570.54	2.88	0.01229
204855_at	SERPINB5	89.41	254.12	2.84	0.015386
222549_at	CLDN1	154.73	438.34	2.83	0.016607

Table 3.1.8.1: Genes up-regulated in Grade 3 specimens in comparison to Grade 2 specimens

GOID	GO Name	Changed	Measured	p-value
5882	Intermediate filament	5	92	0
45111	Intermediate filament cytoskeleton	5	92	0
278	Mitotic cell cycle	4	177	0
5856	Cytoskeleton	9	710	0
15288	Porin activity	2	20	0.001
19212	Phosphatase inhibitor activity	2	25	0.002
51301	Cell division	4	153	0.002
5198	Structural molecule activity	8	673	0.002
16781	Phosphotransferase activity, paired acceptors	1	1	0.003
4756	Selenide, water dikinase activity	1	1	0.003

Table 3.1.8.2: Functions enriched among genes up-regulated in Grade 3 specimens in comparison to Grade 2 specimens

MAPP Name	Changed	Measured	p-value
1-Tissue-Embryonic Stem Cell	4	47	0
2-Tissues-Muscle Fat and Connective	4	82	0
2-Tissues-Blood and Lymph	4	78	0.001
Cell cycle KEGG	3	89	0.02
Streptomycin biosynthesis	1	4	0.022
Vitamin B6 metabolism	1	6	0.034
Cell Cycle-G1 to S control Reactome	2	67	0.05

Table 3.1.8.3: Pathways enriched among genes up-regulated in Grade 3 specimens in comparison to Grade 2 specimens

Probe set	Gene	Baseline	Experiment	FC	p-value
221107_at	CHRNA9	163.02	13.2	-12.35	0.025506
241811_x_at	---	122.75	10.29	-11.93	0.013787
210576_at	CYP4F8	162.6	20.88	-7.79	0.045351
1560850_at	---	193.12	47.72	-4.05	0.042593
206799_at	SCGB1D2	1482.2	461.3	-3.21	0.006931
219602_s_at	FAM38B	149.67	48.6	-3.08	0.004821
205440_s_at	NPY1R	1193.02	388.98	-3.07	0.01136
233059_at	---	472.38	157.13	-3.01	0.04971
241368_at	LSDP5	277.55	93.45	-2.97	0.002462
235976_at	SLITRK6	636.62	217.17	-2.93	0.039497
213651_at	PIB5PA	308.74	105.94	-2.91	0.000372
227550_at	LOC143381	928.69	327.47	-2.84	0.000501
229975_at	---	1105.68	390.97	-2.83	0.006145
203980_at	FABP4	1345.62	486.59	-2.77	0.009236
228766_at	---	671.85	245.4	-2.74	0.024059
204018_x_at	HBA1	193.91	72.2	-2.69	0.046376
209458_x_at	HBA1	187.28	69.6	-2.69	0.04667
243241_at	---	447.87	170.16	-2.63	0.024914
229580_at	---	418.9	160.07	-2.62	0.002053
205696_s_at	GFRA1	278.61	108.16	-2.58	0.003903

Table 3.1.8.4: Genes down-regulated in Grade 3 specimens in comparison to Grade 2 specimens

GOID	GO Name	Changed	Measured	p-value
3867	4-aminobutyrate transaminase activity	1	1	0.002
47298	(S)-3-amino-2-methylpropionate transaminase activity	1	1	0.002
16167	Glial cell line-derived neurotrophic factor receptor activity	1	2	0.002
15674	Di-, tri-valent inorganic cation transport	3	112	0.002
4060	Arylamine N-acetyltransferase activity	1	2	0.003
5010	Insulin-like growth factor receptor activity	1	3	0.003
30284	Estrogen receptor activity	1	2	0.004
9448	Gamma-aminobutyric acid metabolism	1	2	0.004
6631	Fatty acid metabolism	3	135	0.004
45839	Negative regulation of mitosis	1	2	0.005

Table 3.1.8.5: Functions enriched among genes down-regulated in Grade 3 specimens in comparison to Grade 2 specimens

MAPP Name	Changed	Measured	p-value
1-Tissue-Muscle_fat_and_connective	5	65	0
2-Tissues-Endocrine_and_CNS	3	103	0.007
Circadian_Exercise	2	48	0.017
Bile_acid_biosynthesis	2	37	0.02
Fatty_acid_metabolism	2	66	0.04

Table 3.1.8.6: Pathways enriched among genes down-regulated in Grade 3 specimens in comparison to Grade 2 specimens

The transcripts in the Grade 1 vs. 2 comparisons and Grade 2 vs. 3 comparisons were cross compared to identify transcripts which progressively increase with grade. Table 3.1.8.7 list the transcripts which progressively increased or decreased with grade. Transcripts with different trends of expression were removed from the list.

Probe	Name	Grade 1 vs. 2	Grade 2 vs. 3
230966_at	IL4I1	5.52	2.26
204655_at	CCL5	2.05	1.81
209644_x_at	CDKN2A	1.9	2.43
219806_s_at	C11orf75	1.89	1.51
204994_at	MX2	1.77	1.52
216397_s_at	BOP1	1.77	1.38
219402_s_at	DERL1	1.71	1.51
219202_at	RHBDF2	1.6	1.65
201201_at	CSTB	1.52	1.48
200632_s_at	NDRG1	1.4	1.72
201433_s_at	PTDSS1	1.4	1.36
218499_at	RP6-213H19.1	1.4	1.3
222977_at	SURF4	1.39	1.32
218151_x_at	GPR172A	1.36	1.45
201587_s_at	IRAK1	1.35	1.49
225751_at	RBM17	1.34	1.3
208691_at	TFRC	1.34	1.24
201772_at	AZIN1	1.32	1.26
208693_s_at	GARS	1.3	1.4
200844_s_at	PRDX6	1.3	1.24
201527_at	ATP6V1F	1.3	1.24
222992_s_at	NDUFB9	1.26	1.22
217835_x_at	C20orf24	1.26	1.21
225334_at	C10orf32	-1.41	-1.3
227856_at	C4orf32	-1.65	-1.37

Table 3.1.8.7: Transcripts progressively increasing or decreasing with increase in grade

3.1.9 Comparison criteria: Tumour Size < 2.8cm vs. > 2.8 cm

Large size tumors are more aggressive. Identifying genes up- and down-regulated in patients with large tumors vs. small tumors may help identify biomarkers and targets for aggressive disease.

A total of 56 specimens with tumour size less than 2.8cm and a total of 48 specimens with tumour size greater than 2.8cm were compared for gene expression changes.

Up-regulated gene transcripts:

36 genes were identified as significantly up-regulated ($p \leq 0.05$, Fold Change (FC) > 1.2 and Difference > 100) in tumour size greater than 2.8cm compared to tumour size less than 2.8cm. Genes were ranked by fold change and, based on these criteria, the top 20 genes are listed in Table 3.1.9.1.

Gene ontology analysis was performed on the up-regulated genes ($p \leq 0.05$, FC >2 , and Difference > 100). Significant functions were identified based on p-value ($p \leq 0.05$) and the significant functions are listed in Table 3.1.9.2.

Pathway analysis was performed using GenMAPP on the up-regulated genes ($p \leq 0.05$, FC >2 , and Difference > 100). No pathway was found to be significantly affected ($p \leq 0.05$).

Down-regulated gene transcripts:

139 genes were identified as significantly down-regulated ($p \leq 0.05$, Fold Change (FC) < -1.2 and Difference < -100) in tumour size greater than 2.8 compared to tumour size less than 2.8. Genes were ranked by fold change and, based on this criteria, the top 20 genes are listed in Table 3.1.9.3.

Gene ontology analysis was performed on the down-regulated genes ($p \leq 0.05$, FC <-2 , and Difference < -100). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions are listed in Table 3.1.9.4.

Pathway analysis was performed using GenMAPP database on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and $\text{Difference} < -100$). No pathway was found to be significantly affected ($p \leq 0.05$).

Probe set	Gene	Baseline	Experimental	FC	p-value
235210_s_at	RPESP	119.16	294.12	2.47	0.04609
223895_s_at	EPN3	269.69	631.28	2.34	0.003628
219300_s_at	CNTNAP2	87.71	195.71	2.23	0.037062
220318_at	EPN3	92.13	192.54	2.09	0.007949
235203_at	---	146.71	249.42	1.7	0.015045
243552_at	---	296.91	474.03	1.6	0.036136
1556316_s_at	LOC284889	263.36	410.07	1.56	0.000235
201562_s_at	SORD	353.7	540.08	1.53	0.04714
213971_s_at	SUZ12	196.02	299	1.53	0.041818
214295_at	KIAA0485	304.21	452.79	1.49	0.014113
210002_at	GATA6	210.51	311.08	1.48	0.031133
225203_at	PPP1R16A	255.24	367.22	1.44	0.009487
235079_at	---	266.76	384.43	1.44	0.023296
1553303_at	C16orf46	629.84	905.39	1.44	0.028459
213577_at	SQLE	588.55	841.36	1.43	0.037003
242824_at	NFIA	430.72	610.49	1.42	0.045806
200641_s_at	YWHAZ	289.37	393.34	1.36	0.043948
208972_s_at	ATP5G1	854.63	1162.89	1.36	0.010666
226616_s_at	NDUFV3	350.37	475.02	1.36	0.017362
208104_s_at	TSC22D4	339.09	459.32	1.35	0.02723

Table 3.1.9.1: Genes up-regulated in tumour size greater than 2.8 in comparison to tumour size less than 2.8

GOID	GO Name	Changed	Measured	p-value
8038	Neuron recognition	1	5	0.002
8037	Cell recognition	1	16	0.005
19226	Transmission of nerve impulse	1	226	0.022
8289	Lipid binding	1	195	0.027

Table 3.1.9.2: Functions enriched among genes up-regulated in tumour size greater than 2.8 in comparison to tumour size less than 2.8

Probe set	Gene	Baseline	Experimental	FC	p-value
205358_at	GRIA2	421.5	60.72	-6.94	0.048872
243722_at	PYDC1	164.35	36.57	-4.49	0.02198
203029_s_at	PTPRN2	220.94	88.15	-2.51	0.048149
220414_at	CALML5	383.44	161.31	-2.38	0.028932
202768_at	FOSB	244.75	113.18	-2.16	0.0116
202506_at	SSFA2	317.81	152.65	-2.08	0.020651
221667_s_at	HSPB8	531.42	255.95	-2.08	0.028217
204351_at	S100P	975.59	515.89	-1.89	0.021114
204363_at	F3	328.01	177.65	-1.85	0.019519
219440_at	RAI2	414.68	236.28	-1.76	0.017806
203423_at	RBP1	435.93	249.66	-1.75	0.019874
212771_at	C10orf38	251.59	147.36	-1.71	0.035128
218976_at	DNAJC12	943.16	562.33	-1.68	0.028794
208078_s_at	SNF1LK	361.28	220.95	-1.64	0.014417
211026_s_at	MGLL	282.48	175.3	-1.61	0.010454
207992_s_at	AMPD3	292.54	182.61	-1.6	0.009085
204489_s_at	CD44	439.87	278.4	-1.58	0.002035
219681_s_at	RAB11FIP1	390.62	252.8	-1.55	0.03929
204550_x_at	GSTM1	327.44	213.13	-1.54	0.030244
223251_s_at	ANKRD10	784.11	516.69	-1.52	0.005009

Table 3.1.9.3: Genes down-regulated in tumour size greater than 2.8 in comparison to tumour size less than 2.8

GOID	GO Name	Changed	Measured	p-value
4971	Alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionate selective glutamate receptor activity	1	3	0.003
15277	Kainate selective glutamate receptor activity	1	8	0.005
19198	Transmembrane receptor protein phosphatase activity	1	18	0.006
5001	Transmembrane receptor protein tyrosine phosphatase activity	1	18	0.006
4970	Ionotropic glutamate receptor activity	1	18	0.009
5234	Glutamate-gated ion channel activity	1	19	0.009
8066	Glutamate receptor activity	1	38	0.02
5231	Excitatory extracellular ligand-gated ion channel activity	1	45	0.02
122	Negative regulation of transcription from RNA polymerase II promoter	1	65	0.022
6986	Response to unfolded protein	1	41	0.023

Table 3.1.9.4: Functions enriched among genes down-regulated in tumour size greater than 2.8 in comparison to tumour size less than 2.8

3.1.10 Comparison criteria: Patients who did not relapse vs. patients who did relapse (Overall relapse)

Identifying genes up- and down-regulated in patients who relapse vs. those who did not may help identify biomarkers and targets for aggressive disease and can lead to development of diagnostic assays.

A total of 56 patients who did not relapse and 48 patients who relapsed were compared for gene expression changes

Up-regulated gene transcripts:

323 genes were identified as significantly up-regulated ($p \leq 0.05$, Fold Change (FC) > 1.2 and Difference > 100) in tumour specimens of the patients who relapsed compared to tumour specimens of the patients who did not relapse. Genes were ranked by fold change and, based on this criterion, the top 20 genes are listed in Table 3.1.10.1. As can be seen from this comparison, the Ropporin transcripts were significantly differentially-expressed in relapsed vs. non- relapsed tumour specimens.

Gene ontology analysis was performed on the up-regulated genes ($p \leq 0.05$, FC >2 , and Difference > 100). Significant functions were identified based on p-value ($p \leq 0.05$) and the 20 most significant functions are listed in Table 3.1.10.2.

Pathway analysis was performed using GenMAPP database on the up-regulated genes ($p \leq 0.05$, FC >2 , and Difference > 100). Significant pathways were identified based on p-value ($p \leq 0.05$) and are listed in Table 3.1.10.3.

Down-regulated gene transcripts:

476 genes were identified as significantly down-regulated ($p \leq 0.05$, Fold Change (FC) < -1.2 and Difference < -100) in tumour specimens of the patients who relapsed compared to tumour specimens of the patients who did not relapse. Genes were ranked by fold change and, based on this criterion, the top 20 genes are listed in Table 3.1.10.4.

Gene ontology analysis was performed on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and $\text{Difference} < -100$). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions are listed in Table 3.1.10.5.

Pathway analysis was performed using GenMAPP on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and $\text{Difference} < -100$). Significant pathways were identified based on p-value ($p \leq 0.05$) and listed in Table 3.1.10.6. Muscle, fat and connective tissue specific genes pathway (Fig 3.1.10.1) were observed to be enriched by the down-regulated genes.

Probe set	Gene	Baseline	Experimental	FC	p-value
224191_x_at	ROPN1	25.01	148.64	5.94	0.01968
220425_x_at	ROPN1B	24.85	125.74	5.06	0.015855
231535_x_at	ROPN1	32.75	162.84	4.97	0.021558
214595_at	KCNG1	58.43	228.4	3.91	0.004431
212531_at	LCN2	50.9	169.1	3.32	0.044731
206023_at	NMU	48.79	156.42	3.21	0.019771
220625_s_at	ELF5	176.3	536.45	3.04	0.010197
232547_at	SNIP	77.12	227.4	2.95	0.020363
235209_at	RPESP	86.61	254	2.93	0.040668
242350_s_at	ST8SIA6	81.48	227.73	2.79	0.017142
204855_at	SERPINB5	101.99	278.81	2.73	0.020136
235210_s_at	RPESP	115.69	303.22	2.62	0.039636
223062_s_at	PSAT1	95.93	243.48	2.54	0.00843
205044_at	GABRP	376.38	932.57	2.48	0.0153
208103_s_at	ANP32E	116.16	266.76	2.3	0.001988
213557_at	CRKRS	117.88	271.12	2.3	0.027907
204304_s_at	PROM1	215.84	492.38	2.28	0.022001
223748_at	SLC4A11	125.62	286.45	2.28	0.023205
202504_at	TRIM29	104.11	234	2.25	0.02805
213551_x_at	PCGF2	180.1	387.69	2.15	0.012842

Table 3.1.10.1: Genes up-regulated in tumour specimens of patients who relapsed in comparison to the tumour specimens of patients who did not relapse

GOID	GO Name	Changed	Measured	p-value
8603	cAMP-dependent protein kinase regulator activity	2	11	0
15288	Porin activity	2	20	0
15267	Channel or pore class transporter activity	4	376	0
4648	Posphoserine transaminase activity	1	1	0.002
42816	Vitamin B6 metabolism	1	2	0.002
8614	Pyridoxine metabolism	1	2	0.002
42819	Vitamin B6 biosynthesis	1	2	0.002
8615	Pyridoxine biosynthesis	1	2	0.002
19887	Potein kinase regulator activity	2	49	0.002
19867	Outer membrane	2	60	0.003
5215	Transporter activity	6	1412	0.004
19207	Kinase regulator activity	2	55	0.005
7340	Acrosome reaction	1	5	0.006
6564	L-serine biosynthesis	1	8	0.006
19861	Fagellum	1	9	0.008
15106	Bicarbonate transporter activity	1	10	0.008
15380	Anion exchanger activity	1	10	0.008
5452	Inorganic anion exchanger activity	1	10	0.008
6940	Regulation of smooth muscle contraction	1	9	0.009
15301	Anion:anion antiporter activity	1	11	0.009

Table 3.1.10.2: Functions enriched among genes up-regulated in tumour specimens of patients who relapsed in comparison to tumour specimens of patients who did not relapse

MAPP Name	Changed	Measured	p-value
1-Tissue-Endocrine and CNS	3	210	0.001
Vitamin B6 metabolism	1	6	0.008
Glycine serine and threonine metabolism	1	35	0.041

Table 3.1.10.3: Pathways enriched among genes up-regulated in tumour specimens of patients who relapsed in comparison to tumour specimens of patients who did not relapse

Probe set	Gene	Baseline	Experimental	FC	p-value
242301_at	CBLN2	134.72	15.37	-8.76	0.01888
210576_at	CYP4F8	124.3	16.15	-7.69	0.033726
229764_at	FAM79B	457.64	75.47	-6.06	0.001887
243929_at	ZNF533	200.31	52.69	-3.8	0.024215
235978_at	FABP4	142.81	40.78	-3.5	0.031338
205710_at	LRP2	160.38	50.4	-3.18	0.026702
205380_at	PDZK1	403.12	129.03	-3.12	0.008031
210222_s_at	RTN1	199.84	64.09	-3.12	0.017134
202833_s_at	SERPINA1	492.17	177.02	-2.78	0.012344
205794_s_at	NOVA1	455.57	165.34	-2.76	0.012041
203029_s_at	PTPRN2	222.84	85.37	-2.61	0.03321
203485_at	RTN1	688.43	264.94	-2.6	0.008909
214440_at	NAT1	1861.54	757.62	-2.46	0.000972
235976_at	SLITRK6	604.94	256.63	-2.36	0.034647
218398_at	MRPS30	621.53	265.65	-2.34	0.002144
231207_at	---	328.23	143.94	-2.28	0.004513
227600_at	---	305.13	134.82	-2.26	0.001608
205696_s_at	GFRA1	273.91	122.45	-2.24	0.00585
219197_s_at	SCUBE2	747.54	334.52	-2.23	0.016711
211429_s_at	SERPINA1	1076.28	485	-2.22	0.003786

Table 3.1.10.4: Genes down-regulated in tumour specimens of patients who relapsed in comparison to the tumour specimens of patients who did not relapse

GOID	GO Name	Changed	Measured	p-value
4421	Hdroxymethylglutaryl-CoA synthase activity	1	2	0.001
46912	Transferase activity, transferring acyl groups, acyl groups converted into alkyl on transfer	1	4	0.001
8393	Fatty acid (omega-1)-hydroxylase activity	1	2	0.002
6629	Lipid metabolism	5	535	0.002
4060	Arylamine N-acetyltransferase activity	1	2	0.003
16167	Glial cell line-derived neurotrophic factor receptor activity	1	2	0.006
51244	Regulation of cellular physiological process	0	2727	0.01
50791	Regulation of physiological process	0	2822	0.01
46847	Filopodium formation	1	6	0.011
43088	Regulation of Cdc42 GTPase activity	1	6	0.011

Table 3.1.10.5: Functions enriched among genes down-regulated in tumour specimens of patients who relapsed in comparison to tumour specimens of patients who did not relapse

MAPP Name	Changed	Measured	p-value
1-Tissue-Muscle fat and connective	3	65	0.001
Synthesis and Degradation of ketone Bodies KEGG	1	5	0.009
Synthesis and degradation of ketone bodies	1	6	0.013

Table 3.1.10.6: Pathways enriched among genes down-regulated in tumour specimens of patients who relapsed in comparison to tumour specimens of patients who did not relapse

3.1.11 Comparison criteria: Patients who survived vs. patients who did not survive

Identifying genes up- and down-regulated in patients who survived vs. those who did not survive may help identify biomarkers and targets for aggressive disease and can lead to development of diagnostic assays.

Tumour specimens of 69 patients who survived and a total of 35 patients who did not survive were compared for gene expression changes.

Up-regulated gene transcripts:

385 genes were identified as significantly up-regulated ($p \leq 0.05$, Fold Change (FC) > 1.2 and Difference > 100) in tumour specimens of the patients who did not survive compared to tumour specimens of patients who survived. Genes were ranked by fold change and, based on these criteria, the top 20 genes are listed in Table 3.1.11.1.

Gene ontology analysis was performed on the up-regulated genes ($p \leq 0.05$, FC >2 , and Difference > 100). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions are listed in Table 3.1.11.2.

Pathway analysis was performed using GenMAPP on the up-regulated genes ($p \leq 0.05$, FC >2 , and Difference > 100). Significant pathways were identified based on p-value ($p \leq 0.05$) and listed in Table 3.1.11.3.

Down-regulated gene transcripts:

993 genes were identified as significantly down-regulated ($p \leq 0.05$, Fold Change (FC) < -1.2 and Difference < -100) in tumour specimens of the patients who did not survive compared to tumour specimens of patients who survived. Genes were ranked by fold change and, based on these criteria, the top 20 genes are listed in Table 3.1.11.4

Gene ontology analysis was performed on the down-regulated genes ($p \leq 0.05$, FC <-2 , and Difference < -100). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions are listed in Table 3.1.11.5.

Pathway analysis was performed using GenMAPP database on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and $\text{Difference} < -100$). Significant pathways were identified based on p-value ($p \leq 0.05$) and listed in Table 3.1.11.6.

Probe set	Gene	Baseline	Experiment	FC	p-value
210147_at	ART3	23.95	129.59	5.41	0.030235
212531_at	LCN2	45.81	224.84	4.91	0.021047
232547_at	SNIP	79.47	282.82	3.56	0.021477
214595_at	KCNG1	79.42	249.59	3.14	0.020733
206023_at	NMU	57.62	178.51	3.1	0.047878
223748_at	SLC4A11	118.87	361.86	3.04	0.007922
242350_s_at	ST8SIA6	91.89	264.11	2.87	0.033159
204855_at	SERPINB5	117.94	302.9	2.57	0.038883
223062_s_at	PSAT1	108.16	272.54	2.52	0.014909
223075_s_at	C9orf58	362.21	878.02	2.42	0.013435
227512_at	LOC92312	156.91	375.07	2.39	0.000142
204914_s_at	SOX11	111.97	264.92	2.37	0.02351
226346_at	LOC92312	266.6	598.08	2.24	0.000878
210513_s_at	VEGFA	97.12	215.63	2.22	0.032837
236885_at	LOC92312	115.01	253.34	2.2	0.0041
204915_s_at	SOX11	135.5	290.79	2.15	0.042257
213523_at	CCNE1	141.89	303.3	2.14	0.041275
202991_at	STARD3	140.04	294.74	2.1	0.010777
203496_s_at	PPARBP	321.49	670.52	2.09	0.042681
213551_x_at	PCGF2	204.66	411.53	2.01	0.041232

Table 3.1.11.1: Genes up-regulated in tumour specimens of patients who did not survive in comparison to tumour specimens of patients who survived

GOID	GO Name	Changed	Measured	p-value
4648	Phosphoserine transaminase activity	1	1	0
30521	Androgen receptor signaling pathway	2	32	0
30518	Steroid hormone receptor signaling pathway	2	43	0
30522	Intracellular receptor-mediated signaling pathway	2	45	0
35257	Nuclear hormone receptor binding	2	43	0.001
51427	Hormone receptor binding	2	43	0.001
30947	Regulation of vascular endothelial growth factor receptor signaling pathway	1	1	0.002
48010	Vascular endothelial growth factor receptor signaling pathway	1	1	0.002
30949	Positive regulation of vascular endothelial growth Factor receptor signaling pathway	1	1	0.002
42816	Vitamin B6 metabolism	1	2	0.002

Table 3.1.11.2: Functions enriched among genes up-regulated in tumour specimens of patients who did not survive in comparison to tumour specimens of patients who survived.

MAPP Name	Changed	Measured	p-value
Id NetPath 5	2	51	0.001
Vitamin B6 metabolism	1	6	0.017
Hypertrophy model	1	20	0.025
Glycine serine and threonine metabolism	1	35	0.034

Table 3.1.11.3: Pathways enriched among genes up-regulated in tumour specimens of patients who did not survive in comparison to tumour specimens of patients who survived

Probe set	Gene	Baseline	Experiment	FC	p-value
206502_s_at	INSM1	230.98	7.19	-32.13	0.025417
206325_at	SERPINA6	120.98	18.06	-6.7	0.039756
229764_at	FAM79B	391.54	60.87	-6.43	0.001304
210222_s_at	RTN1	185.27	40.52	-4.57	0.002024
243929_at	ZNF533	178.14	41.83	-4.26	0.013087
205380_at	PDZK1	369.77	93.07	-3.97	0.002203
209706_at	NKX3-1	310.19	81.64	-3.8	0.001471
203485_at	RTN1	648.43	187.63	-3.46	0.00067
219197_s_at	SCUBE2	739.18	213.92	-3.46	0.000281
205794_s_at	NOVA1	420.34	125.93	-3.34	0.002229
229004_at	---	364.9	118.18	-3.09	0.000052
205913_at	PLIN	152.71	51.4	-2.97	0.02848
227742_at	CLIC6	461.51	156.68	-2.95	0.028478
227182_at	SUSD3	407.14	138.95	-2.93	0.000177
228554_at	---	519.6	177.93	-2.92	0.007057
227929_at	---	163.39	59.85	-2.73	0.03079
232176_at	SLITRK6	341.94	125.18	-2.73	0.02738
203413_at	NELL2	283.89	104.48	-2.72	0.003488
228390_at	---	515.22	189.51	-2.72	0.000006
210272_at	CYP2B7P1	375.11	141.21	-2.66	0.043449

Table 3.1.11.4: Genes down-regulated in tumour specimens of patients who did not survive in comparison to tumour specimens of patients who survived

GOID	GO Name	Changed	Measured	p-value
45010	Actin nucleation	1	1	0.001
5925	Focal adhesion	1	1	0.001
30027	Lamellipodium	2	15	0.001
31252	Leading edge	2	18	0.001
7494	Midgut development	1	1	0.002
7497	Posterior midgut development	1	1	0.002
9441	Glycolate metabolism	1	1	0.003
18445	Prothoracicotropic hormone activity	1	2	0.003
7388	Posterior compartment specification	1	2	0.004
7387	Anterior compartment specification	1	2	0.004

Table 3.1.11.5: Functions enriched among genes down-regulated in tumour specimens of patients who did not survive in comparison to tumour specimens of patients who survived

MAPP Name	Changed	Measured	p-value
1-Tissue-Muscle fat and connective	4	65	0.001
Adipogenesis	3	130	0.018
Circadian Exercise	2	48	0.019
Synthesis and Degradation of Ketone Bodies KEGG	1	5	0.025
Id NetPath 5	2	51	0.025
Synthesis and degradation of ketone bodies	1	6	0.029
2-Tissues-Internal Organs	3	137	0.034
Apoptosis	2	82	0.044

Table 3.1.11.6: Pathways enriched among genes down-regulated in tumour specimens of patients who did not survive in comparison to tumour specimens of patients who survived

3.1.12 Comparison criteria: Patients who did not relapse within 5 years vs. patients who did relapse within 5 years.

Identifying genes up- and down-regulated in patients who relapse vs. those who did not may help identify biomarkers and targets for aggressive disease and can lead to development of diagnostic assays.

Tumour specimens of 54 patients who did not relapse within 5 years and a total of 41 patients who relapsed within 5 years were compared for gene expression changes.

Up-regulated gene transcripts:

318 genes were identified as significantly up-regulated ($p \leq 0.05$, Fold Change (FC) > 1.2 and Difference > 100) in tumour specimens of the patients who relapsed compared to tumour specimens of patients who did not relapse. Genes were ranked by fold change and, based on these criteria, the top 20 genes are listed in Table 3.1.12.1. As can be seen from this comparison, the Ropporin transcripts were significantly differentially-expressed in the 5 year relapse vs. non- relapsed tumour specimens.

Gene ontology analysis was performed on the up-regulated genes ($p \leq 0.05$, $FC > 2$, and Difference > 100). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions are listed in Table 3.1.12.2.

Pathway analysis was performed using GenMAPP on the up-regulated genes ($p \leq 0.05$, $FC > 2$, and Difference > 100). Significant pathways were identified based on p-value ($p \leq 0.05$) and listed in Table 3.1.12.3.

Down-regulated gene transcripts:

680 genes were identified as significantly down-regulated ($p \leq 0.05$, Fold Change (FC) < -1.2 and Difference < -100) in tumour specimens of the patients who relapsed compared to tumour specimens of patients who did not relapse. Genes were ranked by fold change and, based on these criteria, the top 20 genes are listed in Table 3.1.12.4.

Gene ontology analysis was performed on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and $\text{Difference} < -100$). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions are listed in Table 3.1.12.5.

Pathway analysis was performed using GenMAPP on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and $\text{Difference} < -100$). Significant pathways were identified based on p-value ($p \leq 0.05$) and listed in Table 3.1.12.6.

Probe set	Gene	Baseline	Experiment	FC	p-value
224191_x_at	ROPN1	20.3	162.32	8	0.019245
220425_x_at	ROPN1B	17.82	139.52	7.83	0.010956
204437_s_at	FOLR1	21.98	151.16	6.88	0.014061
231535_x_at	ROPN1	26.52	180.61	6.81	0.017972
204855_at	SERPINB5	73.65	316.13	4.29	0.004791
229341_at	TFCP2L1	38.76	157.15	4.05	0.030912
209842_at	SOX10	56.05	226.31	4.04	0.015137
220625_s_at	ELF5	146.44	582.94	3.98	0.006595
202037_s_at	SFRP1	221.42	865.31	3.91	0.014187
206023_at	NMU	45.81	177.39	3.87	0.014196
212531_at	LCN2	50.64	195.86	3.87	0.031797
202036_s_at	SFRP1	169.73	623.66	3.67	0.00988
223748_at	SLC4A11	91.73	328.87	3.59	0.003251
219795_at	SLC6A14	44.12	155.01	3.51	0.005238
214595_at	KCNG1	68.69	232.46	3.38	0.012411
235209_at	RPESP	87.61	291.55	3.33	0.030122
232547_at	SNIP	79.78	259.39	3.25	0.022409
209466_x_at	PTN	269.74	870.2	3.23	0.040297
223468_s_at	RGMA	110.09	348.4	3.16	0.012154
235210_s_at	RPESP	115.26	349.68	3.03	0.02415

Table 3.1.12.1: Genes up-regulated in tumour specimens of patients who relapsed within 5 years in comparison to tumour specimens of patients who did not relapse within 5 years

GOID	GO Name	Changed	Measured	p-value
8603	cAMP-dependent protein kinase regulator activity	2	11	0
15288	Porin activity	2	20	0
30528	Transcription regulator activity	11	1171	0
19212	Phosphatase inhibitor activity	2	25	0.002
3712	Transcription cofactor activity	4	253	0.002
4648	Phosphoserine transaminase activity	1	1	0.003
3824	Catalytic activity	5	4622	0.003
19215	Intermediate filament binding	1	1	0.004
8614	Pyridoxine metabolism	1	2	0.004
8615	Pyridoxine biosynthesis	1	2	0.004

Table 3.1.12.2: Functions enriched among genes up-regulated in tumour specimens of patients who relapsed within 5 years in comparison to tumour specimens of patients who did not relapse within 5 years.

MAPP Name	Changed	Measured	p-value
2-Tissues-Muscle Fat and Connective	2	82	0.017
1-Tissue-Endocrine and CNS	3	210	0.018
Vitamin B6 metabolism	1	6	0.019
Wnt NetPath 8	2	109	0.033
2-Tissues-Internal Organs	2	137	0.04
Inositol phosphate metabolism	2	134	0.041

Table 3.1.12.3: Pathways enriched among genes up-regulated in tumour specimens of patients who relapsed within 5 years in comparison to tumour specimens of patients who did not relapse within 5 years.

Probe set	Gene	Baseline	Experiment	FC	p-value
206502_s_at	INSM1	287.37	9.91	-28.99	0.025471
214320_x_at	CYP2A6	816.58	36.86	-22.15	0.009574
1494_f_at	CYP2A6	695.91	33.63	-20.7	0.024309
221107_at	CHRNA9	127.61	9.11	-14	0.019025
1562309_s_at	PHF21B	190.53	15.67	-12.16	0.007264
206325_at	SERPINA6	136.76	13.02	-10.5	0.038544
242301_at	CBLN2	138.45	15.99	-8.66	0.021921
205357_s_at	AGTR1	590.42	69.39	-8.51	0.012981
210576_at	CYP4F8	126.86	17.06	-7.44	0.039621
210272_at	CYP2B7P1	487.73	67.47	-7.23	0.000966
229764_at	FAM79B	455.26	72.92	-6.24	0.002773
236445_at	LOC731986	266.11	44.94	-5.92	0.01803
226269_at	GDAP1	427.06	79.73	-5.36	0.010084
218332_at	BEX1	594.77	112.33	-5.29	0.008856
206754_s_at	CYP2B7P1	1573.54	302.78	-5.2	0.000651
240192_at	FLJ45983	154.22	30.82	-5	0.00002
226271_at	GDAP1	371.16	75.86	-4.89	0.012719
205509_at	CPB1	2047.11	450.59	-4.54	0.011659
205794_s_at	NOVA1	497.24	122.38	-4.06	0.001487
203029_s_at	PTPRN2	247.3	61.96	-3.99	0.004993

Table 3.1.12.4: Genes down-regulated in tumour specimens of patients who relapsed within 5 years in comparison to tumour specimens of patients who did not relapse within 5 years

GOID	GO Name	Changed	Measured	p-value
4867	Serine-type endopeptidase inhibitor activity	5	78	0
4866	Endopeptidase inhibitor activity	5	124	0
30414	Protease inhibitor activity	5	125	0
30027	Lamellipodium	2	15	0.001
31252	Leading edge	2	18	0.001
42995	Cell projection	3	65	0.001
4857	Enzyme inhibitor activity	5	220	0.001
45010	Actin nucleation	1	1	0.002
5925	Focal adhesion	1	1	0.002
7494	Midgut development	1	1	0.003

Table 3.1.12.5: Functions enriched among genes down-regulated in tumour specimens of patients who relapsed within 5 years in comparison to tumour specimens of patients who did not relapse within 5 years.

MAPP Name	Changed	Measured	p-value
Valine leucine and isoleucine degradation	2	39	0.021
1-Tissue-Internal Organs	5	244	0.021
Synthesis and Degradation of Ketone Bodies KEGG	1	5	0.024
Gamma Hexachlorocyclohexane degradation	2	45	0.025
1-Tissue-Endocrine and CNS	4	210	0.039
Synthesis and degradation of ketone bodies	1	6	0.041
Ethylbenzene degradation	1	8	0.042

Table 3.1.12.6: Pathways enriched among genes down-regulated in tumour specimens of patients who relapsed within 5 years in comparison to tumour specimens of patients who did not relapse within 5 years

3.1.13 Comparison criteria: Patients who survived for 5 years vs. patients who did not survive for 5 years.

Identifying genes up- and down-regulated in patients who survive vs. those who did not may help identify biomarkers and targets for aggressive disease and can lead to development of diagnostic assays

Tumour specimens of 64 patients who survived for 5 years and a total of 29 patients who did not survive for 5 years were compared for gene expression changes.

Up-regulated gene transcripts:

400 genes were identified as significantly up-regulated ($p \leq 0.05$, Fold Change (FC) > 1.2 and Difference > 100) in tumour specimens of the patients who did not survive for 5 years compared to tumour specimens of patients who did survive for 5 years. Genes were ranked by fold change and, based on these criteria, the top 20 genes are listed in Table 3.1.13.1.

Gene ontology analysis was performed on the up-regulated genes ($p \leq 0.05$, $FC > 2$, and Difference > 100). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions are listed in Table 3.1.13.2.

Pathway analysis was performed using GenMAPP on the up-regulated genes ($p \leq 0.05$, $FC > 2$, and Difference > 100). Significant pathways were identified based on p-value ($p \leq 0.05$) and listed in Table 3.1.13.3.

Down-regulated gene transcripts:

969 genes were identified as significantly down-regulated ($p \leq 0.05$, Fold Change (FC) < -1.2 and Difference < -100) in tumour specimens of the patients who did not survive for 5 years compared to tumour specimens of patients who did survive for 5 years. Genes were ranked by fold change and, based on these criteria, the top 20 genes are listed in Table 3.1.13.4.

Gene ontology analysis was performed on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and $\text{Difference} < -100$). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions are listed in Table 3.1.13.5.

Pathway analysis was performed using GenMAPP database on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and $\text{Difference} < -100$). Significant pathways were identified based on p-value ($p \leq 0.05$) listed in Table 3.1.13.6.

Probe set	Gene	Baseline	Experiment	FC	P value
213456_at	SOSTDC1	10.09	117.71	11.67	0.022369
210147_at	ART3	15.75	142.8	9.06	0.02646
204437_s_at	FOLR1	24.9	178.39	7.16	0.031434
220425_x_at	ROPN1B	22.83	153.17	6.71	0.043523
220559_at	EN1	19.75	129.15	6.54	0.043512
212531_at	LCN2	52.76	256.71	4.87	0.030136
220625_s_at	ELF5	147.31	652.97	4.43	0.011091
206373_at	ZIC1	31.32	137.98	4.41	0.005065
204855_at	SERPINB5	82.13	357.9	4.36	0.008081
204086_at	PRAME	39.82	172.45	4.33	0.048485
206023_at	NMU	49.78	208	4.18	0.030535
209842_at	SOX10	61.23	252.8	4.13	0.036135
232547_at	SNIP	87.64	323.76	3.69	0.026748
223748_at	SLC4A11	103.94	380.55	3.66	0.006712
242350_s_at	ST8SIA6	83.55	296.44	3.55	0.029189
1553613_s_at	FOXC1	188.81	651.73	3.45	0.008375
214595_at	KCNG1	78.07	262.27	3.36	0.03416
219795_at	SLC6A14	48.81	161.81	3.32	0.021036
202036_s_at	SFRP1	206.71	678.3	3.28	0.041266
226907_at	PPP1R14C	44.8	144.93	3.24	0.044567

Table 3.1.13.1: Genes up-regulated in tumour specimens of patients who did not survive for 5 years in comparison to tumour specimens of patients who did survive for 5 years

GOID	GO Name	Changed	Measured	p-value
19212	Phosphatase inhibitor activity	3	25	0
6563	L-serine metabolism	2	13	0
9070	Serine family amino acid biosynthesis	2	12	0.001
19888	Protein phosphatase regulator activity	3	45	0.001
19208	Phosphatase regulator activity	3	46	0.001
4864	Protein phosphatase inhibitor activity	2	24	0.002
8076	Voltage-gated potassium channel complex	3	80	0.002
9069	Serine family amino acid metabolism	2	27	0.003
5249	Voltage-gated potassium channel activity	3	96	0.003
4648	Posphoserine transaminase activity	1	1	0.004

Table 3.1.13.2: Functions enriched among genes up-regulated in tumour specimens of patients who did not survive for 5 years in comparison to tumour specimens of patients who did survive for 5 years

MAPP Name	Changed	Measured	p-value
Glycine serine and threonine metabolism	2	35	0.002
1-Tissue-Endocrine and CNS	3	210	0.008
Vitamin B6 metabolism	1	6	0.018
Methionine metabolism	1	14	0.026
Blood Clotting Cascade	1	20	0.049

Table 3.1.13.3: Pathways enriched among genes up-regulated in tumour specimens of patients who did not survive for 5 years in comparison to tumour specimens of patients who did survive for 5 years

Probe set	Gene	Baseline	Experiment	FC	P value
206502_s_at	INSM1	243.97	8.12	-30.05	0.028369
214320_x_at	CYP2A6	693.76	34.24	-20.26	0.009304
1494_f_at	CYP2A6	588.77	31.16	-18.89	0.024325
236538_at	GRIA2	178.03	10.14	-17.56	0.040355
205358_at	GRIA2	389.59	27.73	-14.05	0.020555
221107_at	CHRNA9	109.73	8.28	-13.25	0.017158
1562309_s_at	PHF21B	164.82	12.69	-12.99	0.006251
210576_at	CYP4F8	113.27	9.13	-12.4	0.022252
205357_s_at	AGTR1	516.17	50.59	-10.2	0.009019
236445_at	LOC731986	241.86	25	-9.67	0.006299
242301_at	CBLN2	119.14	16.62	-7.17	0.024079
240192_at	FLJ45983	140.56	21.15	-6.65	0.000002
219557_s_at	NRIP3	245.85	37.19	-6.61	0.006063
229764_at	FAM79B	399.62	65.3	-6.12	0.002428
210272_at	CYP2B7P1	419.42	78.88	-5.32	0.003092
233059_at	---	458.13	86.09	-5.32	0.002482
219197_s_at	SCUBE2	824.53	159.59	-5.17	0.00001
226269_at	GDAP1	374.8	73.26	-5.12	0.00918
210222_s_at	RTN1	186.12	41	-4.54	0.003962
203029_s_at	PTPRN2	224.07	51.34	-4.36	0.002727

Table 3.1.13.4: Genes down-regulated in tumour specimens of patients who did not survive for 5 years in comparison to tumour specimens of patients who did survive for 5 years

GOID	GO Name	Changed	Measured	p-value
50381	Unspecific monooxygenase activity	4	24	0
5006	Epidermal growth factor receptor activity	2	7	0
16712	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen	4	29	0
46906	Tetrapyrrole binding	4	77	0.001
20037	Heme binding	4	77	0.001
4497	Monooxygenase activity	4	86	0.001
50878	Regulation of body fluids	4	98	0.001
16705	Oxidoreductase activity, acting on paired donors, With incorporation or reduction of molecular oxygen	4	99	0.001
19752	Carboxylic acid metabolism	9	409	0.001
6082	Organic acid metabolism	9	411	0.001

Table 3.1.13.5: Functions enriched among genes down-regulated in tumour specimens of patients who did not survive for 5 years in comparison to tumour specimens of patients who did survive for 5 years.

MAPP Name	Changed	Measured	p-value
Synthesis and Degradation of Ketone Bodies KEGG	1	5	0.041
Valine leucine and isoleucine degradation	2	39	0.045
1-Tissue-Endocrine and CNS	5	210	0.045

Table 3.1.13.6: Pathways enriched among genes down-regulated in tumour specimens of patients who did not survive for 5 years in comparison to tumour specimens of patients who did survive for 5 years

3.1.14 Comparing gene lists to identify bad prognosis genes

The gene lists generated above (Relapse 5 years 0 vs. 1 (Rel5) (see section 3.1.12), Survival 5 years 0 vs. 1 (Sur5) (see section 3.1.13), Overall Relapsed 0 vs. 1 (Relapsed) (see section 3.1.10) and RIP 0 vs. 1 (see section 3.1.11)) were compared to identify genes common to all the gene lists. The DE genes in all these comparisons are linked to bad outcome and an overlap of these lists was carried out to identify any high-value common targets among them. The number of genes common to these gene lists is depicted in the Venn diagram in Fig 3.1.14.1. The total number of common genes in all the gene lists was 384. These gene lists were further compared with lymph node 0 vs. 1 (Fig: 3.1.14.2), as lymph node-positive is linked to bad prognosis, and a final list of 74 genes were identified common to all comparisons was identified. The genes common to all the five lists are shown in Table 3.1.14.1 (up-regulated genes) and Table 3.1.14.2 (down-regulated genes).

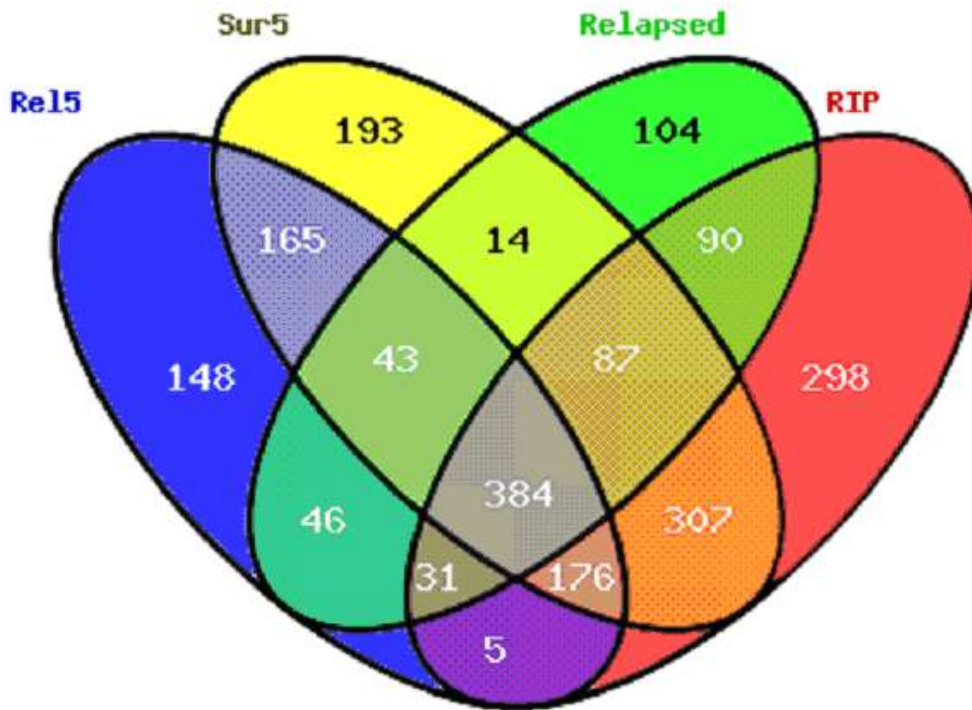


Fig 3.1.14.1: Venn diagram representing the number of genes common to two or more comparisons. There were a total of 384 genes common to all the 4 gene lists.

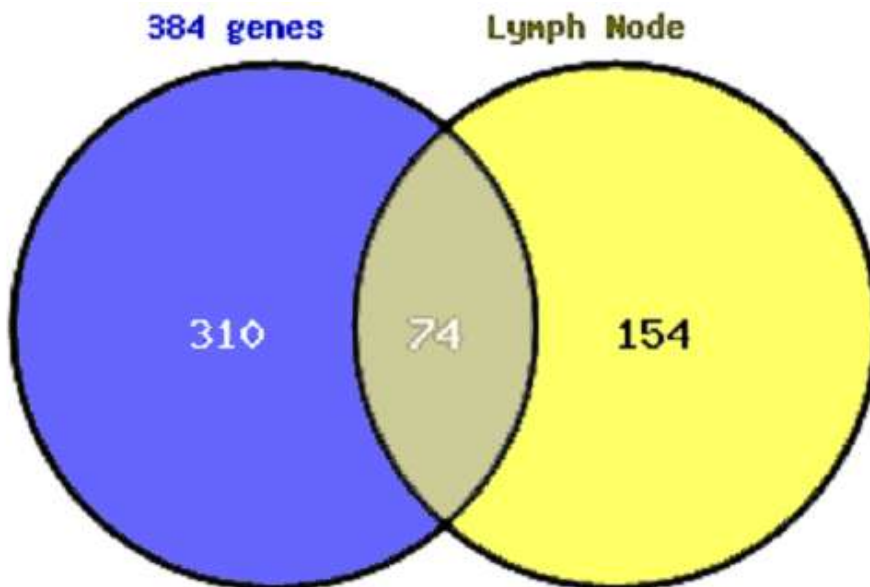


Fig 3.1.14.2: Venn diagram representing the number of genes common to the 384 genes identified earlier (Fig 3.1.14.1) with the Lymph node 0 vs. 1. There were a total of 74 genes common to all.

Probe Set	Gene Symbol	Rel5	Relapsed	RIP	Sur5	LN status
232547_at	SNIP	3.25	2.95	3.56	3.69	2.51
213551_x_at	PCGF2	2.36	2.15	2.01	2.25	2.06
236885_at	LOC92312	2.47	1.99	2.2	2.92	1.96
202991_at	STARD3	1.9	1.83	2.1	2.2	1.91
214239_x_at	PCGF2	2.13	1.93	1.94	2.14	1.89
227512_at	LOC92312	2.57	2.07	2.39	2.8	1.76
226346_at	LOC92312	2.42	2.06	2.24	2.79	1.74
222706_at	CCDC49	1.94	1.8	1.9	2.12	1.73
201400_at	PSMB3	1.71	1.73	1.78	1.84	1.68
216836_s_at	ERBB2	1.81	1.61	1.85	1.99	1.67
203287_at	LAD1	1.9	1.85	1.94	1.98	1.63
224447_s_at	C17orf37	1.63	1.59	1.78	1.9	1.58
213230_at	CDR2L	1.69	1.48	1.77	1.9	1.52
210827_s_at	ELF3	1.51	1.65	1.5	1.51	1.49
230660_at	SERTAD4	2.04	1.62	1.64	1.8	1.47
234464_s_at	EME1	1.68	1.54	1.76	1.84	1.44
212680_x_at	PPP1R14B	1.59	1.57	1.59	1.66	1.41
205107_s_at	EFNA4	1.63	1.41	1.51	1.64	1.41
203430_at	HEBP2	1.38	1.34	1.37	1.42	1.36
201584_s_at	DDX39	1.47	1.47	1.33	1.45	1.33
223993_s_at	CNIH4	1.31	1.28	1.32	1.34	1.32
200660_at	S100A11	1.3	1.36	1.42	1.38	1.31
209609_s_at	MRPL9	1.39	1.36	1.4	1.54	1.31
213668_s_at	SOX4	1.39	1.35	1.47	1.43	1.29
208540_x_at	LOC729659	1.3	1.33	1.44	1.38	1.28
222400_s_at	ADI1	1.35	1.24	1.28	1.27	1.28
203315_at	NCK2	1.34	1.32	1.33	1.43	1.27
200888_s_at	RPL23	1.37	1.33	1.3	1.42	1.26

222036_s_at	SLC12A4	1.29	1.39	1.41	1.37	1.26
222396_at	HN1	1.32	1.3	1.4	1.41	1.24
203956_at	MORC2	1.24	1.4	1.39	1.32	1.24
200882_s_at	PSMD4	1.23	1.23	1.21	1.21	1.21

Table 3.1.14.1: 32 gene transcripts, comprising 29 unique genes up-regulated and common to the 5 comparisons, making them indicative of a clinically poor prognosis

Probe Set	Gene Symbol	Rel5	Relapsed	RIP	Sur5	LN status
205794_s_at	NOVA1	-4.06	-2.76	-3.34	-3.77	-2.71
211421_s_at	RET	-3.27	-1.9	-2.31	-2.8	-2.23
205225_at	ESR1	-2.51	-1.64	-1.77	-2.51	-1.46
244696_at	AFF3	-2.45	-1.93	-2.24	-2.59	-1.73
227198_at	AFF3	-2.38	-1.71	-1.79	-2.34	-1.59
213832_at	---	-2.34	-2.2	-2.29	-2.96	-1.97
236194_at	---	-1.96	-1.98	-2.13	-2.24	-1.68
201311_s_at	SH3BGRL	-1.94	-1.48	-1.92	-1.91	-1.42
225123_at	---	-1.87	-1.8	-1.8	-1.81	-1.9
225613_at	MAST4	-1.84	-1.76	-1.81	-1.98	-1.85
204072_s_at	FRY	-1.79	-1.54	-1.74	-2.12	-1.51
227192_at	PRRT2	-1.79	-1.71	-1.83	-1.88	-1.8
222653_at	PNPO	-1.78	-1.54	-1.57	-1.71	-1.43
221874_at	KIAA1324	-1.77	-1.61	-2.04	-2.59	-1.47
226939_at	CPEB2	-1.73	-1.37	-1.45	-1.83	-1.48
238883_at	THRAP2	-1.67	-1.38	-1.46	-1.71	-1.53
223204_at	C4orf18	-1.65	-1.38	-1.49	-1.86	-1.41
201312_s_at	SH3BGRL	-1.58	-1.44	-1.75	-1.74	-1.29
227856_at	C4orf32	-1.58	-1.51	-1.68	-1.61	-1.36
225561_at	SELT	-1.54	-1.3	-1.51	-1.7	-1.25
212209_at	THRAP2	-1.54	-1.31	-1.33	-1.53	-1.42

201413_at	HSD17B4	-1.5	-1.35	-1.46	-1.5	-1.25
214924_s_at	TRAK1	-1.47	-1.41	-1.52	-1.59	-1.32
212208_at	THRAP2	-1.47	-1.36	-1.35	-1.47	-1.43
221918_at	PCTK2	-1.44	-1.34	-1.41	-1.54	-1.29
212207_at	THRAP2	-1.41	-1.35	-1.34	-1.49	-1.4
200940_s_at	RERE	-1.4	-1.32	-1.45	-1.51	-1.25
243993_at	PCTK2	-1.38	-1.41	-1.37	-1.36	-1.51
201334_s_at	ARHGEF12	-1.37	-1.36	-1.38	-1.36	-1.22
223342_at	RRM2B	-1.36	-1.25	-1.37	-1.35	-1.26
217122_s_at	SLC35E2	-1.35	-1.22	-1.38	-1.46	-1.23
225176_at	LNPEP	-1.34	-1.29	-1.34	-1.38	-1.22
225363_at	PTEN	-1.33	-1.38	-1.52	-1.41	-1.31
200850_s_at	AHCYL1	-1.32	-1.28	-1.38	-1.38	-1.21
218518_at	C5orf5	-1.3	-1.3	-1.43	-1.4	-1.26
227227_at	LOC728871	-1.29	-1.24	-1.41	-1.35	-1.34
224876_at	C5orf24	-1.28	-1.26	-1.39	-1.37	-1.22
218248_at	FAM111A	-1.27	-1.4	-1.44	-1.31	-1.23
200848_at	AHCYL1	-1.26	-1.29	-1.28	-1.3	-1.25
224928_at	SETD7	-1.25	-1.21	-1.42	-1.45	-1.2
200761_s_at	ARL6IP5	-1.24	-1.29	-1.4	-1.31	-1.22
243249_at	ACIN1	-1.23	-1.3	-1.26	-1.22	-1.29

Table 3.1.14.2: 42 gene transcripts, comprising 32 unique known genes and three unannotated transcripts, down-regulated and common to the 5 comparisons, making them indicative of a clinically good prognosis.

3.1.15 Non-parametric analysis

Non-parametric analysis was performed on the following clinical parameters as described in section 2.2.15. The null hypothesis tested was the mean of the two groups under

consideration are equal under an assumption that the data may not be normally distributed.

- Normals/Tumour specimens

Number of Up-Regulated genes: 4057 (4213 using parametric test)

Number of Down-Regulated genes: 3379 (3235 using parametric test)

- Estrogen Receptor status

Number of Up-Regulated genes: 869 (855 using parametric test)

Number of Down-Regulated genes: 1288 (1145 using parametric test)

- Overall Relapsed status

Number of Up-Regulated genes: 279 (323 using parametric test)

Number of Down-Regulated genes: 504 (476 using parametric test)

- RIP (Event of death due to disease)

Number of Up-Regulated genes: 429 (385 using parametric test)

Number of Down-Regulated genes: 1005 (993 using parametric test)

- Relapse within 5 years

Number of Up-Regulated genes: 288 (318 using parametric test)

Number of Down-Regulated genes: 724 (680 using parametric test)

- Survival for 5 years

Number of Up-Regulated genes: 471 (400 using parametric test)

Number of Down-Regulated genes: 946 (969 using parametric test)

- Tumour size

Number of Up-Regulated genes: 67 (36 using parametric test)

Number of Down-Regulated genes: 39 (139 using parametric test)

- Tumour grade 1 vs 2

Number of Up-Regulated genes: 108 (275 using parametric test)

Number of Down-Regulated genes: 197 (75 using parametric test)

- Tumour grade 2 vs 3

Number of Up-Regulated genes: 846 (930 using parametric test)

Number of Down-Regulated genes: 743 (596 using parametric test)

- Lymph Node Status

Number of Up-Regulated genes: 122 (102 using parametric test)

Number of Down-Regulated genes: 110 (126 using parametric test)

The genelists generated from this study are included on the CD with this thesis.

3.1.16 Summary

This section identified various sub-groups of breast cancer and its association with clinical and gene expression data. Our results confirmed many of the existing groups and also identified new groups with clinical relevance. The Ropporin-expressing clusters of patients had a bad prognosis, whereas immune response expressing cluster of patients had a good prognosis. Clinical parameters were compared to identify gene expression changes. Gene ontology and pathways analysis was performed on these gene lists. Ropporin gene was identified as expressed in patients who relapsed and was studied in detail.

3.2 Comparing in-house gene lists with publicly available datasets

This section deals with comparing our in-house results with publicly available datasets.

3.2.1 Comparison with public datasets on the Affymetrix GeneChip platform

The in-house data was compared to 4 publicly available datasets for genes commonly associated with bad prognosis. Overall Relapse was taken as a common measure for comparison across different experiments. A gene list was created for each individual experiment comparing the patients who relapsed vs. patients who did not relapse. The filtration criteria for identifying differentially-expressed genes as outlines in section 2.2.5 (FC>1.2, Difference>100 and p-value≤0.05) was taken for all the experiments. The number of patients who relapsed / did not relapse and the numbers of differentially-expressed (DE) genes in each experiment are listed in Table 3.2.1.1.

Experiment	No. of Non-relapsed patients	No. of Relapsed patients	No. of DE genes in each respective list
In-house dataset	57	48	799
GSE4922	160	89	67
GSE1456	119	40	500
GSE2990	139	40	153
GSE2034	107	179	377

Table 3.2.1.1: The above table represents the number of relapsed / non-relapsed patients in each experimental group and the number of differentially-expressed genes in each experimental group.

The gene lists generated were compared for genes common to all five experimental groups. However, no transcripts were found to be common across all the experiments. Therefore, a separate approach was taken to identify transcripts which changed in any 3 out of the 5 comparisons. To perform this task, a C program was written as this task was not possible to achieve using available software. Following this comparison, 22 transcripts were identified which were either up- or down-regulated in a minimum of 3

out of the 5 experimental groups (Table 3.2.1.2). There were 4 genes which were differentially-regulated in four out of five experiments; these are highlighted in bold.

Probe Set	Gene Symbol	In-house	GSE4922	GSE1456	GSE2990	GSE2034
201041_s_at	DUSP1	NS	-1.22	-1.48	-1.28	NS
201841_s_at	HSPB1	NS	1.26	1.56	1.29	1.21
202489_s_at	FXYD3	1.32	NS	1.42	1.23	NS
202503_s_at	KIAA0101	NS	1.27	1.49	1.29	1.23
202768_at	FOSB	-1.86	-1.46	-2.17	NS	NS
202954_at	PAK3	1.41	1.24	1.35	1.33	NS
204026_s_at	ZWINT	NS	1.26	NS	1.29	1.21
204607_at	HMGCS2	-2.02	-1.57	NS	-1.97	NS
208451_s_at	C4A	NS	-1.28	-1.47	-1.27	NS
209189_at	FOS	-1.42	-1.3	-1.67	-1.4	NS
209772_s_at	CD24	1.94	1.44	1.8	NS	NS
211429_s_at	SERPINA1	-2.22	NS	NS	-1.49	-1.82
212592_at	IGJ	NS	NS	-1.35	-1.52	-1.52
212593_s_at	PDCD4	-1.39	NS	-1.24	NS	-1.22
214428_x_at	C4A	-1.48	-1.25	-1.46	NS	NS
218039_at	NUSAP1	NS	NS	1.47	1.33	1.24
218336_at	PFDN2	1.26	NS	1.28	NS	1.21
218807_at	VAV3	-1.37	NS	-1.26	-1.24	NS
219956_at	GALNT6	-1.58	NS	1.55	1.36	NS
222077_s_at	RACGAP1	NS	NS	1.33	1.36	1.32
222453_at	CYBRD1	-1.61	-1.22	-1.55	NS	NS
227182_at	SUSD3	-1.9	-1.38	-1.51	NS	NS

Table 3.2.1.2: 22 transcripts DE in 3 out of the 5 experimental groups, incorporating the fold change in each experimental group when comparing the patients who relapsed vs. the patients who did not relapse. A positive fold change indicates that gene to be over-expressed in patients who relapsed in comparison to the patients who did not relapse. NS (non significant) = gene not differentially-expressed in that experimental group.

3.2.2 Comparison in-house relapse and lymph node gene list with OncotypeDx genes

OncotypeDx from Genomic Health (www.genomichealth.com) is a PCR based laboratory test that predicts the likelihood of breast cancer recurrence in women with newly diagnosed and early stage invasive breast cancer. It also estimates the benefits from hormone therapy and chemotherapy. The development of this test followed on from the research carried out by Paik *et al.*, (2004) where they identified 16 genes which have a prognostic importance in predicting relapse and survival. The aim of this study was to compare these 16 genes with our in-house study of microarray experiments on breast cancer.

The 16 genes of OncotypeDx were mapped to 41 Affymetrix transcripts (Table 3.2.2.1) using the NetAffx Analysis Center (<http://www.affymetrix.com/analysis/index.affx>) and these Affymetrix IDs were used to compare with our results of overall relapse (see section 3.1.10) and lymph node metastasis (see section 3.1.6).

Gene Symbol	Corresponding Affy Ids
MKI67	212020_s_at
	212021_s_at
	212022_s_at
	212023_s_at
AURKA	204092_s_at
	208079_s_at
	208080_at
BIRC5	202094_at
	202095_s_at
	210334_x_at
CCNB1	228729_at
	214710_s_at
MYB12	201710_at
GRB7	210761_s_at
HER2	210930_s_at

	216836_s_at
	234354_x_at
MMP11	235908_at
	203878_s_at
	203876_s_at
CTS12	210074_at
GSTM1	204550_x_at
	215333_x_at
CD68	203507_at
BAG1	202387_at
	229720_at
	211475_s_at
ESR1	205225_at
	217163_at
	211233_x_at
	211234_x_at
	211235_s_at
	211627_x_at
	215552_s_at
	217190_x_at
PGR	208305_at
BCL2	203685_at
	207004_at
	203684_s_at
	207005_s_at
SCUBE2	219197_s_at

Table 3.2.2.1: Affymetrix Probe set IDs for the 16 genes of OncotypeDx

The OncotypeDx genes were compared with in-house gene expression data of patients who relapsed vs. patients who did not relapse (Fold change > 1.2, Difference > 100 and p-value≤0.05).

There were a total of five genes common to the two lists, all with the same trend of expression. Over-expression of AURKA, BIRC5 and ERBB2 were associated with higher overall relapse in our study and were related to high risk of recurrence on OncotypeDx, whereas over-expression of ESR1 and SCUBE2 were associated with lower overall relapse in our study and were related to low risk of recurrence on OncotypeDx (Table 3.2.2.2).

Probe Set Name	Gene Symbol	In-House	OncotypeDX
208079_s_at	AURKA	1.32	+1
202095_s_at	BIRC5	1.55	+1
216836_s_at	ERBB2	1.61	+1
205225_at	ESR1	-1.64	-1
219197_s_at	SCUBE2	-2.23	-1

Table 3.2.2.2: Comparing gene expression values of in-house study on patients who relapsed vs. patients who did not relapse with the OncotypeDx gene lists. Positive value represents over-expression of gene to be association with higher recurrence of disease and negative value represents over-expression of gene to be association with lower recurrence of disease.

The OncotypeDx genes were also compared with gene expression data of patients with no lymph node metastasis and patients who had lymph node metastasis (Fold change > 1.2, Difference > 100 and p-value≤0.05) from our in-house dataset. There were only two genes common to these lists (Table 3.2.2.3). ERBB2 expression was up-regulated in lymph node-positive patients in our study and was related to high risk of recurrence on the OncotypeDx. ESR1 expression was down-regulated in lymph node-positive patients in our study and was related to low risk of recurrence on the OncotypeDx.

Probe Set Name	Gene Symbol	In-House	OncotypeDx
216836_s_at	ERBB2	1.67	+1
205225_at	ESR1	-1.46	-1

Table 3.2.2.3: Genes common to in-house study comparing gene expression values of lymph node-positive patients vs. lymph node-negative patients with the OncotypeDx gene lists. A positive value represent over-expression of that gene to be association with higher recurrence of disease and a negative value represents over-expression of that gene to be association with lower recurrence of disease.

3.2.3 Comparison in-house relapse and lymph node gene list with MammaPrint genes

MammaPrint (Agendia <http://www.agendia.com/>), like OncotypeDx, is a molecular diagnostic kit which is used to assess the risk of breast tumor spread to other parts of the body. The development of this test is a follow on from the research carried out by van't Veer *et al.*, (2002) where they identified 70 transcripts which had a prognostic importance in predicting metastasis. These genes were mapped to Affymetrix Ids using gene symbols in a batch query to the NetAffx Analysis Center (www.affymetric.com/netaffx). For the Contig Ids, their corresponding Accession nos. were obtained from the table provided by the van't Veer group. These Accession nos. were then searched in the unigene database using David and Ease (see section 2.2.6) to obtain the corresponding gene symbol which was then searched in the NetAffx Analysis center. In all, 45 unique genes from the van't Veer study was mapped to 81 Affymetrix transcripts (Table 3.2.3.1). The remaining 25 genes (AL080059, LOC51203, AA555029RC, DC13, AL137718, PK428, HEC, UCH37, KIAA1067, SERF1A, OXCT, L2DTL, AF052162, KIAA0175, SM20, DKFZP564D0462, MP1, FLJ11190, LOC57110, DHX58, AP2B1, CFFM4, HSA250839, CEGP1, ALDH4, and KIAA1442) were not identified by batch analysis in NetAffx as being present on the Affymetrix arrays. Following this, these 25 genes were searched manually and respective affymetrix identifier was allocated to them (Table 3.2.3.1). These were not identified in the batch analysis, because many were old gene names or less known alias of the genes. Still we

were not able to associate AL080059, AA555029RC, and AF052162 with any affymetrix identifier.

Gene Symbol	Probe Set ID
AKAP2	202759_s_at
	202760_s_at
	226694_at
AP2B1	200612_s_at
	200615_s_at
BBC3	211692_s_at
C9orf30	1552277_a_at
CCNE2	205034_at
	211814_s_at
CDCA7	224428_s_at
	230060_at
CENPA	204962_s_at
	210821_x_at
COL4A2	211964_at
	211966_at
DCK	203302_at
DIAPH3	220997_s_at
	229097_at
	232596_at
ECT2	219787_s_at
	234992_x_at
	237241_at
ESM1	208394_x_at
EXT1	201995_at
FBXO31	219784_at
	219785_s_at
	223745_at
	224162_s_at
FGF18	206987_x_at
	211029_x_at
	211485_s_at
	214284_s_at
	231382_at
FLT1	204406_at
	210287_s_at
	222033_s_at
	232809_s_at
GMPS	214431_at
GNAZ	204993_at
GPR180	231871_at

	232912_at
GSTM3	202554_s_at
	235867_at
IGFBP5	1555997_s_at
	203424_s_at
	203425_s_at
	203426_s_at
	211958_at
	211959_at
LOC643008	229740_at
LOC728492	219982_s_at
	223538_at
	223539_s_at
MCM6	201930_at
	238977_at
MMP9	203936_s_at
NMU	206023_at
ORC6L	219105_x_at
PECI	218025_s_at
	218009_s_at
QSOX2	227146_at
	235239_at
RAB6B	210127_at
	221792_at
	225259_at
RFC4	204023_at
RTN4RL1	230700_at
RUNDC1	226298_at
	235040_at
SLC2A3	202497_x_at
	202498_s_at
	202499_s_at
	216236_s_at
	222088_s_at
TGFB3	1555540_at
	209747_at
WISP1	206796_at
	211312_s_at
ZNF533	1555800_at
	1555801_s_at
	229019_at
LOC51203	218039_at
DC13	218447_at
AL137718	220997_s_at
PK428	214464_at

	240735_at
HEC	204162_at
UCH37	219960_s_at
	220083_x_at
	1570145_at
KIAA1067	215413_at
	212026_s_at
	212034_s_at
	212035_s_at
SERF1A	223539_s_at
	219982_s_at
	223538_at
OXCT	202780_at
	244134_at
L2DTL	222680_s_at
	218585_s_at
KIAA0175	204825_at
SM20	220956_s_at
	223083_s_at
	224314_s_at
	221497_x_at
	223045_at
	227147_s_at
DKFZP564D0462	213094_at
MP1	205273_s_at
	217971_at
FLJ11190	1552520_at
	1552521_a_at
LOC57110	219983_at
	219984_s_at
DHX58	219364_at
AP2B1	200615_s_at
	200612_s_at
CFFM4	223344_s_at
HSA250839	219686_at
CEGP1	219197_s_at
ALDH4	203722_at
KIAA1442	233850_s_at

Table 3.2.3.1: Affymetrix Id corresponding to genes on MammaPrint

The 123 mapped transcripts from the van't Veer study were compared with our in-house gene expression data of patients who relapsed (overall relapse) vs. patients who did not relapse (Fold change > 1.2, Difference > 100 and p-value ≤ 0.05) (see section 3.1.10).

There were a total of five genes common to the two lists. These 3 genes displayed the same trend of expression between the two studies. Over-expression of NMU, GMPS and MELK were associated with a poor prognosis (higher incidence of relapse in our study and greater chance of distant metastasis in the van't Veer study), whereas over-expression of PEGI and SCUBE2 was associated with good prognosis (lower incidence of relapse in our study and less chance of distant metastasis in the van't Veer study) (Table 3.2.3.2).

Probe Set ID	Gene Symbol	In-house	MammaPrint
206023_at	NMU	3.21	+1
214431_at	GMPS	1.28	+1
204825_at	MELK, KIAA0175	1.57	+1
218025_s_at	PEGI	-1.29	-1
219197_s_at	SCUBE2, CEGP1	-2.23	-1

Table 3.2.3.2: Genes common to in-house study comparing gene expression values of patients who relapsed vs. patients who did not relapse with the MammaPrint gene lists. A positive value represents over-expression of each gene and its association with poor prognosis while a negative value represents over-expression of gene and its association with good prognosis.

A similar overlap comparison was also performed using the MammaPrint 70 gene signature and our in-house generated gene list comparing lymph node-negative vs. lymph node-positive (Fold change > 1.2, Difference > 100 and p-value ≤ 0.05) (see section 3.1.6). However, no transcripts were identified common to both studies.

3.2.4 Summary

The analysis compared our findings to similar other studies. The analysis found HSPB1, KIAA0101, PAK3 genes to be up-regulated in patients who relapsed. AP1 transcriptional factor genes FOSA, FOSB were down-regulated in patients who relapsed. We also identified common genes to two of the breast cancer diagnostic assay (OncotypeDx and MammaPrint) and our in-house study. NMU, GMPS, MELK, PEGI and SCUBE2 were

common genes on MammaPrint and our in-house study. AURKA, BIRC5, ERBB2, ESR1 and SCUBE2 were common to OncotypeDx and our in-house study.

3.3 Meta analysis of Estrogen receptor pathway genes using gene expression data.

The in-house generated dataset, together with 5 published datasets from GEO (Table 3.3.0.1) were subjected to meta-analysis to identify genes significantly up or down-regulated in estrogen receptor-positive and negative specimens.

Five of these experiments were from clinical studies while one study was on breast cancer cell lines (GSE3156). All these experiments were carried out on Affymetrix GeneChip U133A or U133Plus2.0 arrays. Details of these experiments are listed in Materials and Methods (see section 2.1.2). The minimum number of common transcripts in all experiments was 22,283. That is the number of transcripts in U133A chip.

The normalization method used and number of ER-positive and ER-negative specimens in each experiment is listed in the Table 3.3.0.1.

GEO Id	No of ER(-) specimens	No of ER(+) specimens
In-house	34	68
GSE3156	11	8
GSE3744	18	15
GSE2034	77	209
GSE2990	34	149
GSE4922	34	211

Table 3.3.0.1: The above table lists the normalization type and the number of ER-positive (+) and ER-negative (-) samples in each experimental group. GSE3156 is the cell-line dataset while all others are clinical specimens.

Individual experiments were compared for differentially-expressed genes when comparing the ER-positive specimens to ER-negative specimens. Only those genes were taken for further analysis which had the p-value ≤ 0.05 , Fold change > 1.2 and Difference > 100 when comparing the ER-positive specimens to ER-negative specimens. The gene list generated by each experiment was compared to each other to find common transcripts that changed in two or more experiments. The numbers of common transcripts in any two comparisons are listed in Table 3.3.0.2.

GEO Id	In-house	GSE3156	GSE3744	GSE2034	GSE2990	GSE4922
In-house	2000	410	702	819	443	751
GSE3156		3715	481	576	315	414
GSE3744			2394	1031	639	968
GSE2034				3447	1027	1167
GSE2990					1607	780
GSE4922						2054

Table 3.3.0.2: The above table represents the number of transcripts which are common to any two comparisons.

Gene lists were also compared to identify common genes across all experiments. This analysis identified a set of 82 transcripts which were differentially up or down-regulated across all the six experiments. Out of these 82 transcripts, 62 were up-regulated and 20 transcripts were down-regulated in ER-positive specimens in comparison to ER-negative specimens.

3.3.1 Up-regulated gene transcripts

These transcripts, together with Affy IDs and associated fold changes are listed in Table 3.3.1.1.

Probe Set ID	Gene Symbol	In-House	GSE4922	GSE3744	GSE2990	GSE2034	GSE3156
211712_s_at	ANXA9	4.74	3.55	11.15	3.52	3.38	6.5
213234_at	KIAA1467	3.6	2.63	3.45	2.12	2.49	2.13
206401_s_at	MAPT	3.54	3.03	4.54	2.13	3.63	12.43
205225_at	ESR1	3.53	6.04	40.74	4.71	8.88	15.27
214440_at	NAT1	3.53	4.19	13.32	3.19	6.07	4.58
203928_x_at	MAPT	3.44	3.04	4.15	2.19	3.44	5.83
204540_at	EEF1A2	3.41	2.04	4.93	2.89	3.41	2.08
215304_at	THSD4	3.2	2.95	5.85	2.18	2.92	2.51
203929_s_at	MAPT	3.17	3.69	5.79	2.32	4.09	8.68

205009_at	TFF1	2.53	4.34	59.3	2.66	8.89	60.56
214053_at	ERBB4	2.49	2.61	6.06	2.37	3.01	5.28
203963_at	CA12	2.48	3.72	9.43	2.54	3.37	3.8
204508_s_at	CA12	2.48	5.03	11.83	3.16	4.03	5.1
209460_at	ABAT	2.43	3.54	5.9	3.41	4.07	10.54
209459_s_at	ABAT	2.23	3.31	6.29	3.28	4.33	11.58
209603_at	GATA3	2.14	3.39	5.78	2.25	4.75	5.08
219741_x_at	ZNF552	2.12	1.7	3.27	2.2	1.83	3.5
214164_x_at	CA12	2.06	3.68	11.48	2.58	3.57	3.23
215867_x_at	CA12	2.05	3.73	11.23	2.55	3.29	3.14
218195_at	C6orf211	2.03	2.65	3.47	1.73	3.27	5.13
201841_s_at	HSPB1	2.02	1.63	2.88	1.82	1.81	1.95
203571_s_at	C10orf116	2.01	2.06	7	1.77	3.2	13.28
202089_s_at	SLC39A6	2	3.51	6.11	2.22	3.85	4
205862_at	GREB1	1.99	5	16.28	1.88	4.07	60.55
209602_s_at	GATA3	1.99	3.21	7.02	2.13	4.44	4.4
218211_s_at	MLPH	1.98	2.01	6.23	1.78	2.69	1.95
209604_s_at	GATA3	1.97	2.72	6.62	1.99	3.61	3.9
41660_at	CELSR1	1.97	2.41	3.03	2	3.17	2.81
35666_at	SEMA3F	1.88	1.48	2.17	1.31	1.64	2.15
213441_x_at	SPDEF	1.87	1.27	2.69	1.48	1.64	3.2
204862_s_at	NME3	1.8	1.67	2.23	1.58	1.88	2.61
209681_at	SLC19A2	1.8	1.65	3.88	1.92	2.26	2.38
218931_at	RAB17	1.8	1.6	1.98	1.83	1.39	2.48
201349_at	SLC9A3R1	1.77	1.51	2.3	1.91	1.65	4.32
209623_at	MCCC2	1.74	1.8	2.12	1.67	1.95	1.62
216092_s_at	SLC7A8	1.74	1.8	2.88	1.46	1.9	4.35
205081_at	CRIP1	1.72	1.6	3.03	2	2.77	23.88
221139_s_at	CSAD	1.72	1.72	2.33	1.39	1.8	1.7
204798_at	MYB	1.68	2.51	2.57	1.65	2.9	5.18
212099_at	RHOB	1.66	2.03	3.96	1.53	1.95	2.43
202454_s_at	ERBB3	1.62	1.53	2.23	1.44	1.65	2.91

205074_at	SLC22A5	1.62	1.48	2.22	1.7	1.51	1.68
202088_at	SLC39A6	1.61	2.89	6.24	2.12	3.12	3.15
204667_at	FOXA1	1.61	1.91	4.97	1.67	2.58	3
201596_x_at	KRT18	1.6	1.6	2.71	1.73	2.02	1.67
201754_at	COX6C	1.6	1.53	1.97	1.56	1.75	1.4
204623_at	TFF3	1.6	2.91	20.95	2.05	4.55	59.8
205376_at	INPP4B	1.58	1.99	4	1.98	2.53	2.7
210652_s_at	C1ORF34	1.56	2.57	2.82	1.95	2.55	3.5
212446_s_at	LASS6	1.51	1.34	1.89	2.13	1.58	2.36
212208_at	THRAP2	1.49	1.43	1.5	1.51	1.48	2.2
212209_at	THRAP2	1.48	1.56	1.93	1.73	1.62	2.62
218259_at	MKL2	1.44	1.65	2.01	1.67	1.61	1.91
209008_x_at	KRT8	1.41	1.37	2.21	1.23	1.32	2.39
218807_at	VAV3	1.4	1.87	2.38	1.48	2.55	2.6
200670_at	XBP1	1.37	1.75	3.41	1.43	2.33	3.37
201650_at	KRT19	1.37	1.45	1.96	1.48	1.7	2.54
209110_s_at	RGL2	1.34	1.35	1.96	1.29	1.37	1.6
203476_at	TPBG	1.31	1.88	3.02	1.4	1.69	1.55
201236_s_at	BTG2	1.28	1.62	2.69	1.46	1.94	2.45
218966_at	MYO5C	1.28	1.31	1.76	1.46	1.62	1.86
217979_at	TSPAN13	1.21	1.37	3.23	1.59	1.82	3.57

Table 3.3.1.1: The above table lists the 62 transcripts which were up-regulated in ER-positive patients/cell-lines.

Gene ontology analysis was performed on genes up-regulated in a minimum of 3 out of 6 (50%) experimental groups. This approach was taken so as to increase the size of the DE genes for GO and Pathway analysis keeping the background gene list free from DE genes. The assumption taken was that if a gene is found DE in 50% of the experimental cohorts analysed, it is very likely to be involved in the ER metabolism. Similarly for the background gene list, only those genes were included which were not found to be DE in any of the experimental groups.

The background gene list included those genes which were not differentially-expressed in any of the experimental groups. There were a total of 935 genes which were up-regulated in a minimum of 3 out of 6 experimental groups. Similarly 15859 genes taken as background were found to be non-significant in all of the experimental groups.

GOID	GO Name	Changed	Measured	Z Score
6829	Zinc ion transport	3	6	4.272
15175	Neutral amino acid transporter activity	4	10	4.244
16461	Unconventional myosin	2	3	4.18
5010	Insulin-like growth factor receptor activity	2	3	4.18
51183	Vitamin transporter activity	2	3	4.18
4165	dDdecenoyl-CoA delta-isomerase activity	2	3	4.18
16675	Oxidoreductase activity, acting on heme group of donors	5	15	4.161
16676	Oxidoreductase activity, acting on heme group of donors, oxygen as acceptor	5	15	4.161
4129	Cytochrome-c oxidase activity	5	15	4.161
15002	Heme-copper terminal oxidase activity	5	15	4.161
15804	Neutral amino acid transport	3	7	3.854
5006	Epidermal growth factor receptor activity	3	7	3.854
4303	Estradiol 17-beta-dehydrogenase activity	3	7	3.854
4866	Endopeptidase inhibitor activity	14	82	3.819
30414	Protease inhibitor activity	14	83	3.767
41	Transition metal ion transport	6	23	3.756
50982	Detection of mechanical stimulus	1	1	3.753
50974	Detection of mechanical stimulus during sensory perception	1	1	3.753
9592	Detection of mechanical stimulus during sensory perception of sound	1	1	3.753
8389	Coumarin 7-hydroxylase activity	1	1	3.753

Fig 3.3.1.2: GO analysis on up-regulated genes

MAPP Name	Changed	Measured	Z Score
1-Tissue-Muscle fat and connective	9	44	3.55
Nuclear Receptors	7	34	3.148
Electron Transport Chain	10	58	3.113
Androgen-Receptor NetPath 2	13	88	2.927
Valine leucine and isoleucine biosynthesis	2	5	2.912
Pantothenate and CoA biosynthesis	3	11	2.659
Smooth muscle contraction	15	122	2.368
Ethylbenzene degradation	2	7	2.256
Glutamate metabolism	4	22	2.083
Wnt NetPath 8	11	89	2.038

Fig: 3.3.1.3: Pathways analysis on up-regulated genes

3.3.2 Down-regulated transcripts

Twenty transcripts were down-regulated in all experimental groups. They are listed in Table 3.3.2.1.

Probe Set ID	Gene Symbol	In-House	GSE4922	GSE3744	GSE2990	GSE2034	GSE3156
202037_s_at	SFRP1	-4.38	-2	-6.32	-2.04	-3.31	-6.13
202036_s_at	SFRP1	-4	-2.08	-8.44	-1.83	-4.05	-10.37
201012_at	ANXA1	-2.63	-1.21	-1.75	-1.3	-1.35	-3.74
212771_at	C10orf38	-2.17	-1.62	-4.14	-1.74	-2.14	-3.46
213113_s_at	SLC43A3	-2.08	-1.77	-2.03	-1.34	-1.69	-4.84
201300_s_at	PRNP	-1.9	-1.28	-1.84	-1.36	-1.55	-2.53
208627_s_at	YBX1	-1.83	-1.71	-2.81	-1.56	-1.61	-1.39
212276_at	LPIN1	-1.83	-1.44	-2.39	-1.62	-2.23	-2.13
202342_s_at	TRIM2	-1.79	-1.71	-3.48	-1.57	-2.22	-7.87
200600_at	MSN	-1.77	-1.39	-1.7	-1.26	-1.94	-9.79
221059_s_at	COTL1	-1.64	-1.68	-1.81	-1.47	-2.13	-4.08
200790_at	ODC1	-1.51	-1.68	-2.04	-1.44	-1.75	-1.98
218856_at	TNFRSF21	-1.5	-1.69	-1.82	-1.56	-2.13	-2.66
208628_s_at	YBX1	-1.47	-1.53	-2.06	-1.35	-1.54	-1.38
212274_at	LPIN1	-1.37	-1.77	-2.72	-1.64	-2.12	-2.32
212501_at	CEBPB	-1.37	-1.4	-1.58	-1.37	-1.58	-1.82
212263_at	QKI	-1.31	-1.22	-1.89	-1.36	-1.54	-2.27
201231_s_at	ENO1	-1.3	-1.33	-1.76	-1.24	-1.72	-1.62
218618_s_at	FNDC3B	-1.25	-1.23	-1.58	-1.29	-1.58	-2.98
212345_s_at	CREB3L2	-1.2	-1.27	-1.75	-1.27	-1.43	-1.86

Table 3.3.2.1: The above table lists the transcripts which were up-regulated in ER-negative patients/cell-lines

Gene ontology analysis was performed on genes down-regulated in a minimum of 3 out of 6 experimental groups. The background gene list included those genes which were not DE in any of the experimental groups. There were a total of 697 genes which were down-regulated in a minimum of 3 out of 6 experimental groups. Similarly 15859 genes taken as background were found non-significant in all of the experimental groups.

GOID	GO Name	Changed	Measured	Z Score
7051	Spindle organization and biogenesis	6	12	6.794
6270	DNA replication initiation	6	15	5.892
226	Microtubule cytoskeleton organization and biogenesis	9	33	5.524
16875	Ligase activity, forming carbon-oxygen bonds	8	28	5.39
16876	Ligase activity, forming aminoacyl-tRNA and related compounds	8	28	5.39
8452	RNA ligase activity	8	28	5.39
4812	tRNA ligase activity	8	28	5.39
6261	DNA-dependent DNA replication	12	55	5.355
9607	Response to biotic stimulus	64	633	5.337
278	Mitotic cell cycle	21	133	5.28
6418	tRNA aminoacylation for protein translation	8	29	5.252
43038	Amino acid activation	8	29	5.252
43039	tRNA aminoacylation	8	29	5.252
3690	Double-stranded DNA binding	6	18	5.209
49	tRNA binding	4	9	5.149
51301	Cell division	18	110	5.063
7049	Cell cycle	50	476	4.974
7017	Microtubule-based process	15	85	4.967
8283	Cell proliferation	43	389	4.962
6260	DNA replication	17	104	4.913

Table 3.3.2.2: Significant functions over-represented among ER-negative specimens.

Pathway analysis was performed using GenMAPP. Significantly affected pathways were identified using the enrichment analysis in a way similar to that for Gene ontology. The statistically-significant pathways (top 10 based on Z-score) are listed in Table 3.3.2.3. Ropporin gene, expression specific to testis was also found to be up-

regulated in ER-negative specimens in 3 out of 6 studies (Fig 3.3.2.1). However this pathway was not found to be significant.

MAPP Name	Changed	Measured	Z Score
Cell cycle KEGG	19	69	6.548
Aminoacyl tRNA biosynthesis	7	18	5.183
Streptomycin biosynthesis	2	2	5.055
Cell Cycle G1 to S control reactome	13	54	4.794
DNA replication reactome	9	32	4.567
Vitamin B6 metabolism	3	6	4.037
2-Tissues-Blood and lymph	12	59	3.9
One carbon pool by folate	4	13	3.272
Galactose metabolism	5	21	2.93
T-Cell-Receptor NetPath 11	15	107	2.731

Table 3.3.2.3: Significant pathways over-represented among ER-negative specimens.

Testis

CDKN3	ODF1
GAGE1	ODF2
GAGE2	PRM1
GAGE3	PRM2
GAGE4	SPINK2
GAGE5	TCP11
GAGE6	TNP1
GAGE7	TPTE
GAGE8	CRISP2
INSL3	TSPY1
LDHC	UTRN
CITED1	TKTL1
SLC6A16	PIAS2
DKFZP434P211	TBPL1
FSCN3	TCFL5
UBQLN3	ACTL7B
KCNK4	ACTL7A
YBX2	ZPBP
PHF7	ABHD2
OAZ3	LPIN1
CYB5R2	TSSK2
SPA17	PRND
SPATA6	CABYR
ROPN1	GAPDS
FAM46C	TBL2
WDR10	GAGE7B
ANKRD7	SKD3
SPINLW1	LOC81691
TSPY2	GSG1
FBX031	APH1B
215734_at	RPL39L
216323_x_at	ROPN1B

Fig 3.3.2.1: Testis specific genes up-regulated in ER-negative specimens. ROPN1B and ROPN1 has special significance and was studied in further detail (see section 3.6)

3.3.3 Genes correlated with ESR1

ESR1 is the key gene involved in estrogen receptor pathway. The estrogen receptor pathway is very complex involving interaction of large number of genes, including many transcriptional factors. Detailed study of these genes could help in better understanding of the disease and development of better treatment options. This analysis was aimed at identifying genes whose expression patterns are correlated with ESR1. Gene expression profiles of 5897 samples representing diseased, normal human specimens and cell lines were obtained from Array Express E-TABM-185 (see section 2.1.2). The challenge was to mine large datasets for gene interaction network analysis. Because of the extremely large dataset, C programs were written to process the data.

The first step in the data reduction effort was to filter out the genes which did not change much over the dataset, *i.e.* genes with low variation across samples. Only those genes which had a standard deviation greater than and equal to 1 across all samples were selected for further analysis. A total of 11,099 genes passed this criterion. The C program using Dev C++ (see section 2.2.10) was written for this analysis.

The ESR1 gene has been shown to play an essential role in estrogen receptor metabolism. Using the set of identified 11,099 genes, the dataset was mined to identify highly correlated genes (Correlation > 0.75). A C program using Dev C++ (see section 2.2.10) was written to generate the correlation values among genes. FOXA1, GATA3, SPDEF and C1ORF34 were the only genes which had a correlation greater than 0.75 with ESR1. To get a deeper insight into how other genes correlate with estrogen metabolism, genes correlated with FOXA1, GATA3, SPDEF and C1ORF34 were identified (Correlation > 0.75) Many genes were identified which correlated with FOXA1, SPDEF and C1ORF34 and this information was used to create the network shown in Fig 3.3.3.1. However, no genes other than ESR1 were found to be correlated with GATA3. All these genes were also found to be up-regulated in ER-positive tumour specimens in the previous analysis (see section 3.3.1).

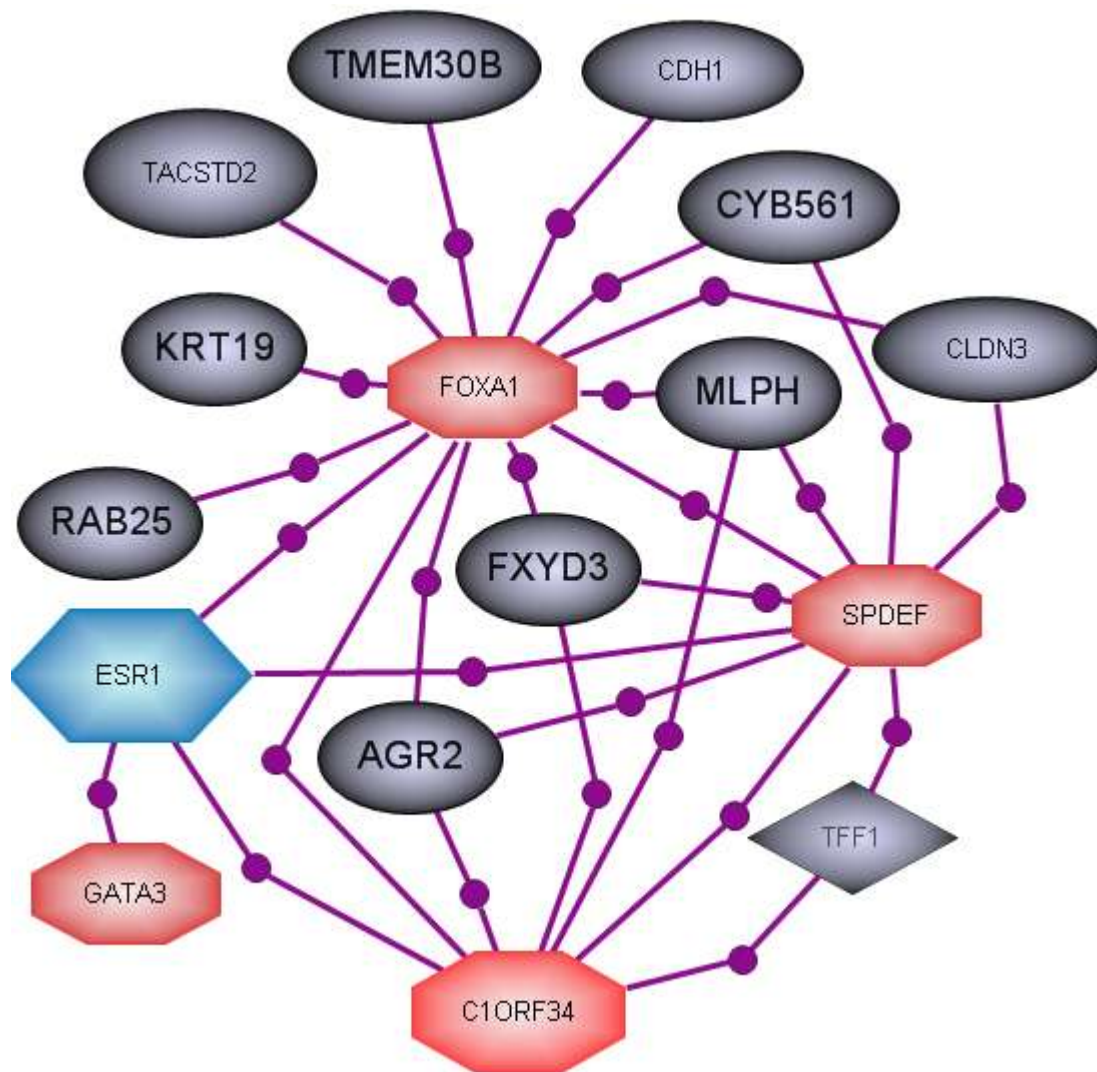


Fig 3.3.3.1: The above figure represents a correlation network among genes identified using 5897 samples. The correlation cut-off of 0.75 was used to filter genes. There were only 4 genes (FOXA1, GATA3, C1ORF34 and SPDEF) which correlated with ESR1 expression. However, there were many more genes which co-regulated with FOXA1, C1ORF34 and SPDEF.

3.3.4 Correlation patterns among genes

The correlation pattern among the ESR1-correlated genes (FOXA1, GATA3, C1ORF34, SPDEF) were plotted in excel to get a deeper understanding of the expression relationships between these genes (Fig 3.3.4.1 – Fig 3.3.4.9). The results indicate that, apart from one exception (that between genes SPDEF and FOXA1), there is independency of expression between each of the genes *i.e.* although the expression is correlated; the expression of one can be independent of the other.

However, there was an interesting relationship observed between the correlation patterns of SPDEF and FOXA1. The expression of FOXA1 was independent of the expression of SPDEF. However, the expression of SPDEF seems likely to be dependent on the expression of FOXA1 (Fig 3.3.4.1).

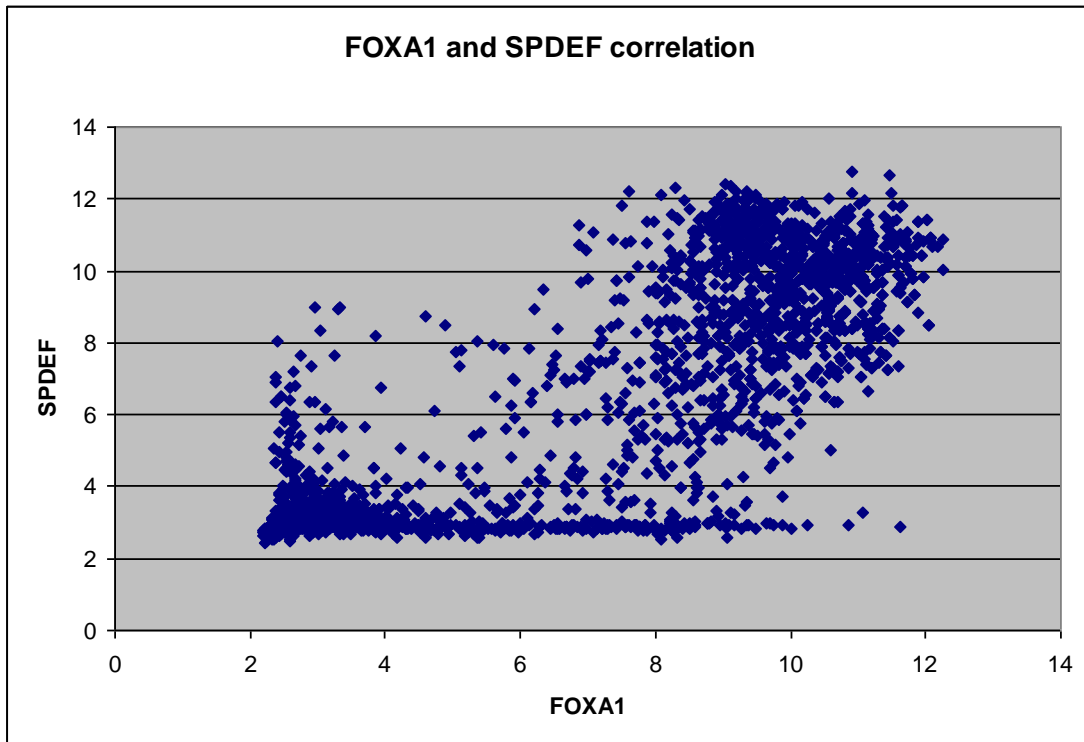


Fig 3.3.4.1: Correlation pattern between FOXA1 and SPDEF. Each axis represents the expression values of that individual gene.

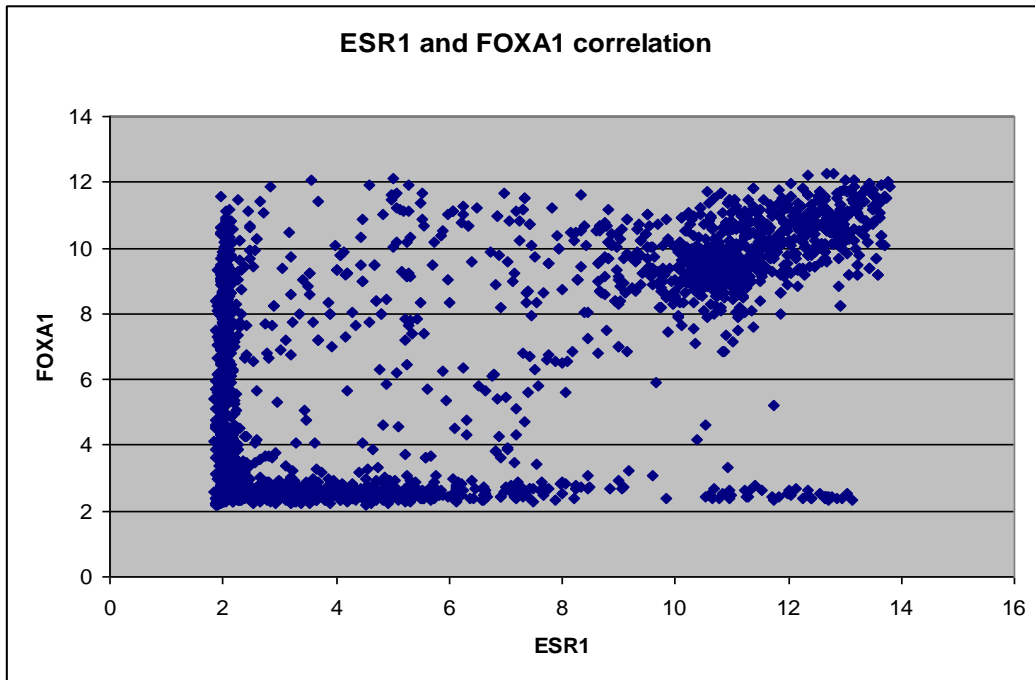


Fig 3.3.4.2: Correlation pattern of ESR1 and FOXA1. Each axis represents the expression values of that individual gene.

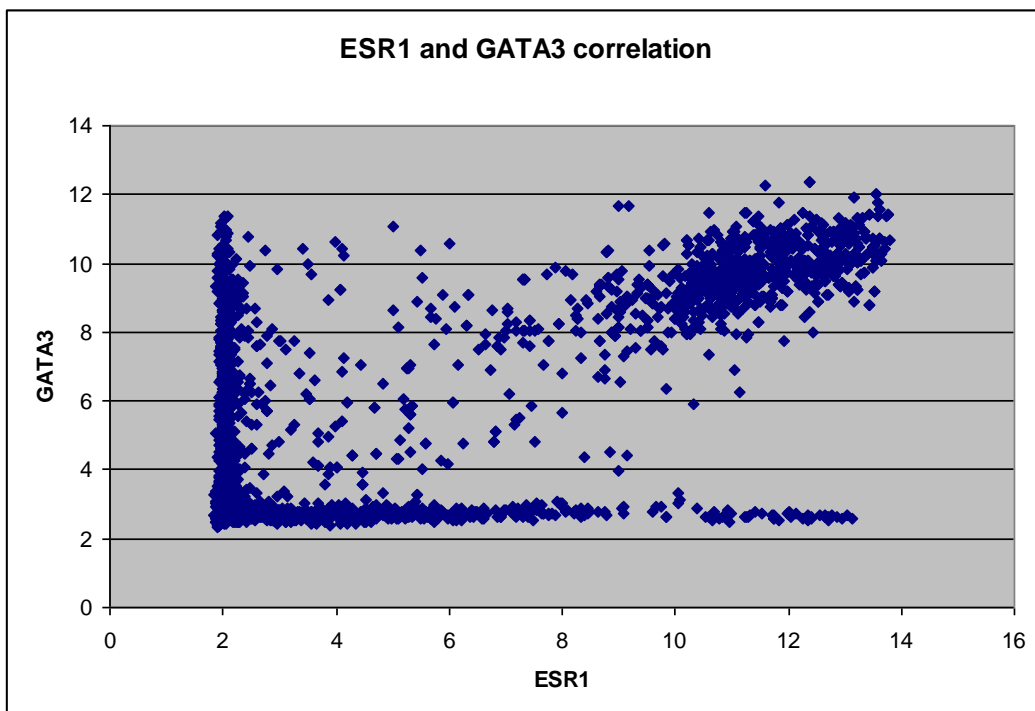


Fig 3.3.4.3: Correlation pattern between ESR1 and GATA3. Each axis represents the expression values of that individual gene.

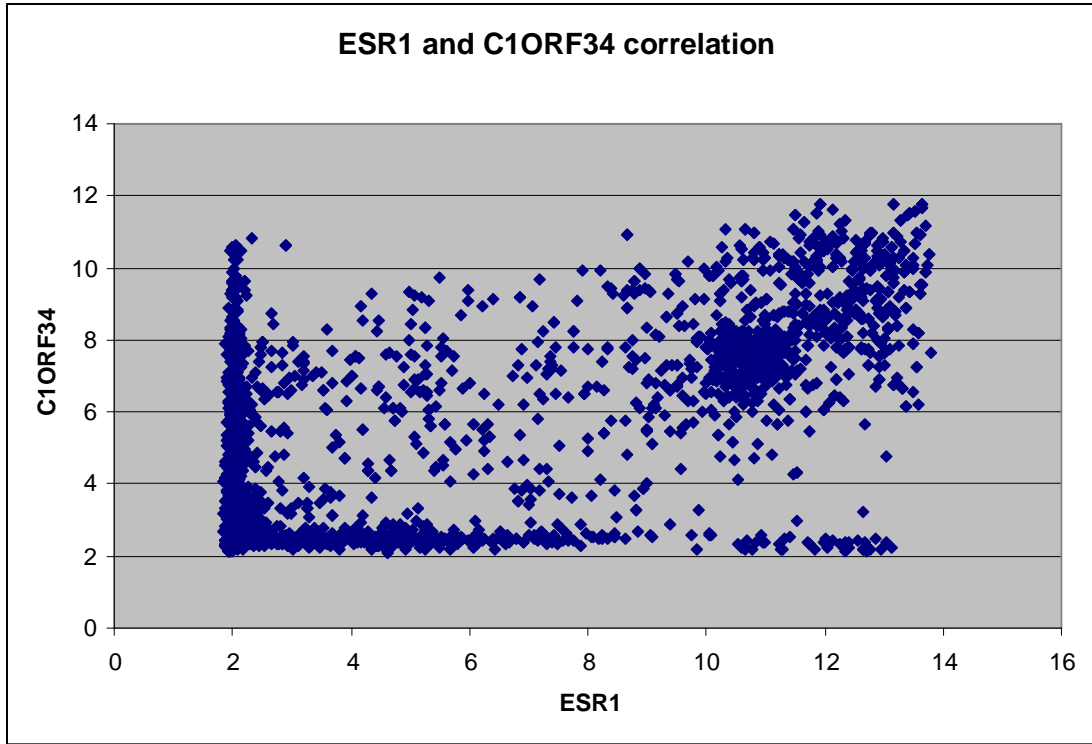


Fig 3.3.4.4: Correlation pattern between ESR1 and C1ORF34. Each axis represents the expression values of that individual gene.

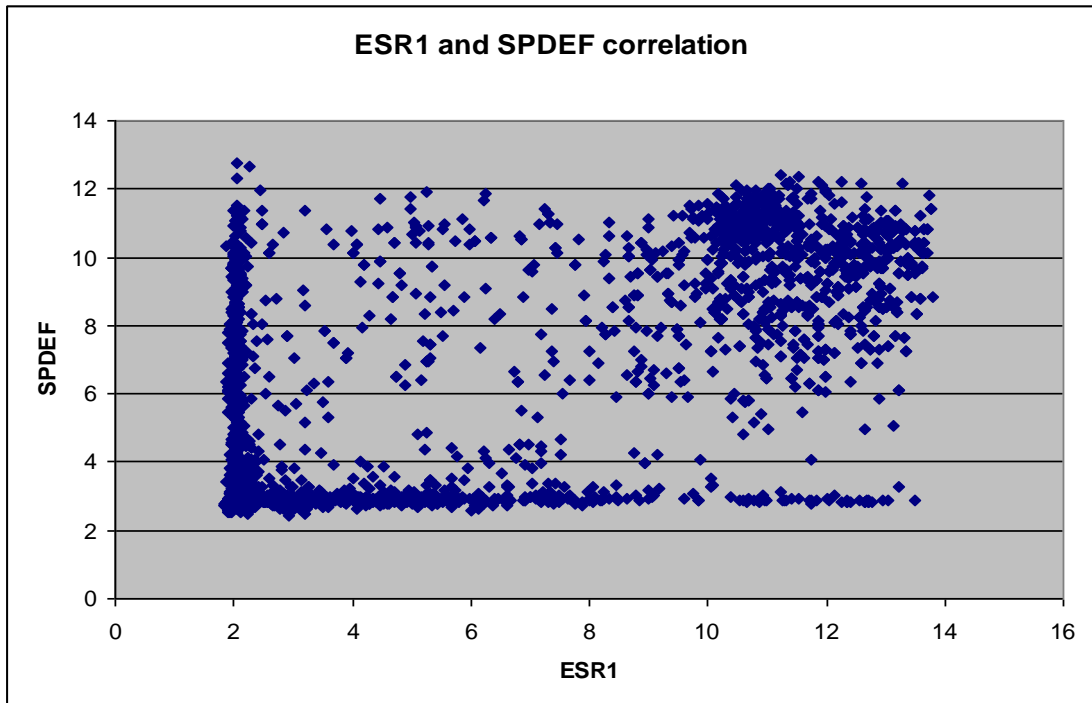


Fig 3.3.4.5: Correlation pattern between ESR1 and SPDEF. Each axis represents the expression values of that individual gene.

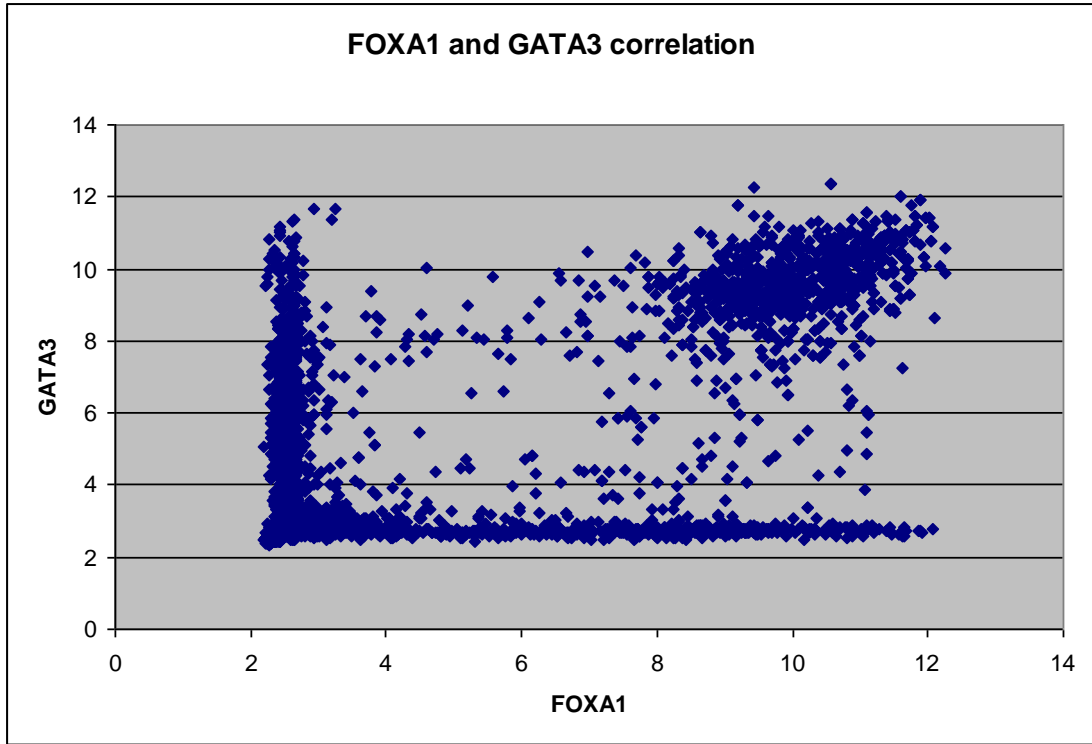


Fig 3.3.4.6: Correlation pattern in FOXA1 and GATA3. Each axis represents the expression values of that individual gene.

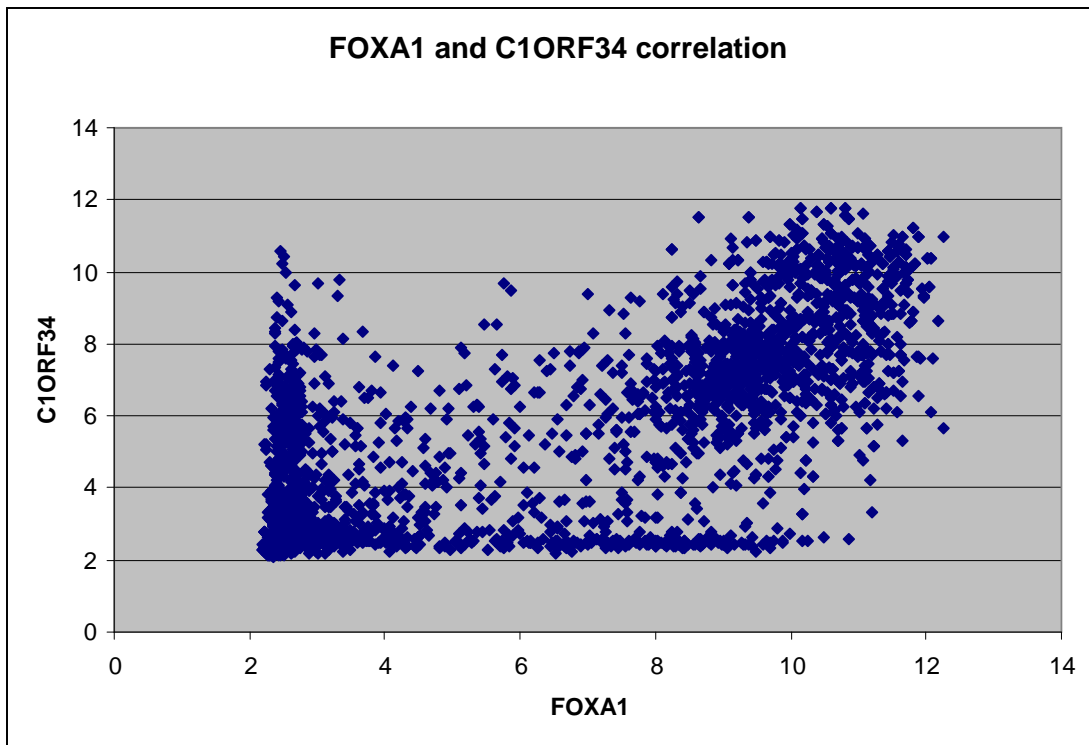


Fig 3.3.4.7: Correlation pattern between FOXA1 and C1ORF34. Each axis represents the expression values of that individual gene.

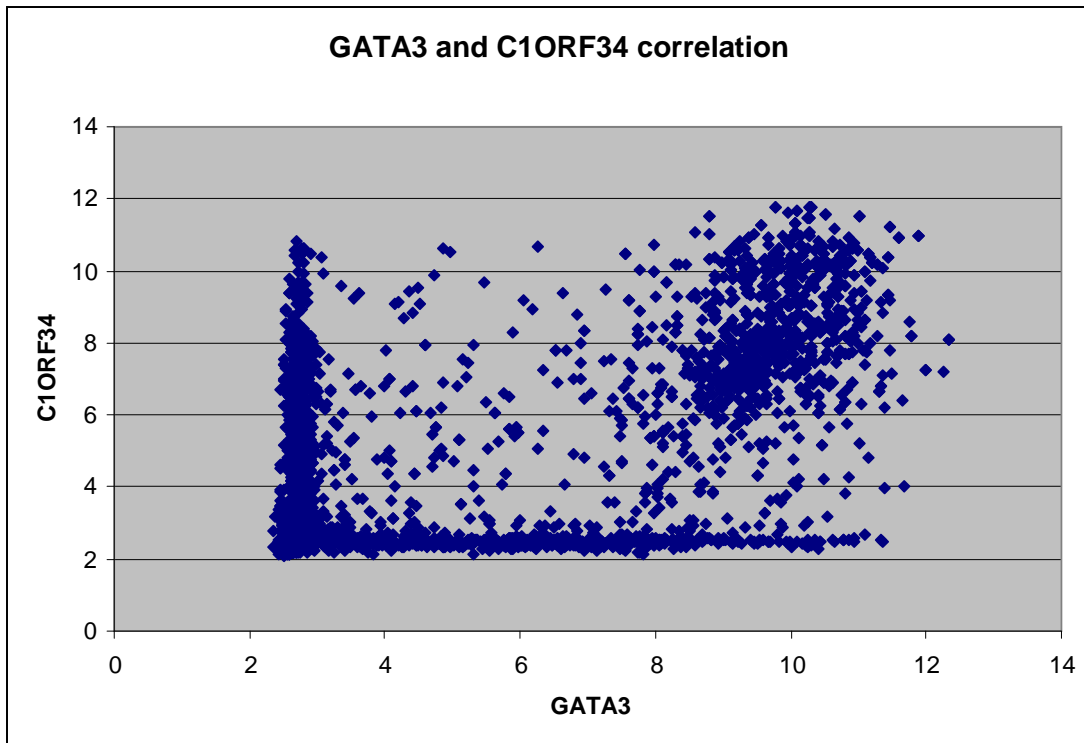


Fig 3.3.4.8: Correlation pattern between GATA3 and C1ORF34. Each axis represents the expression values of that individual gene.

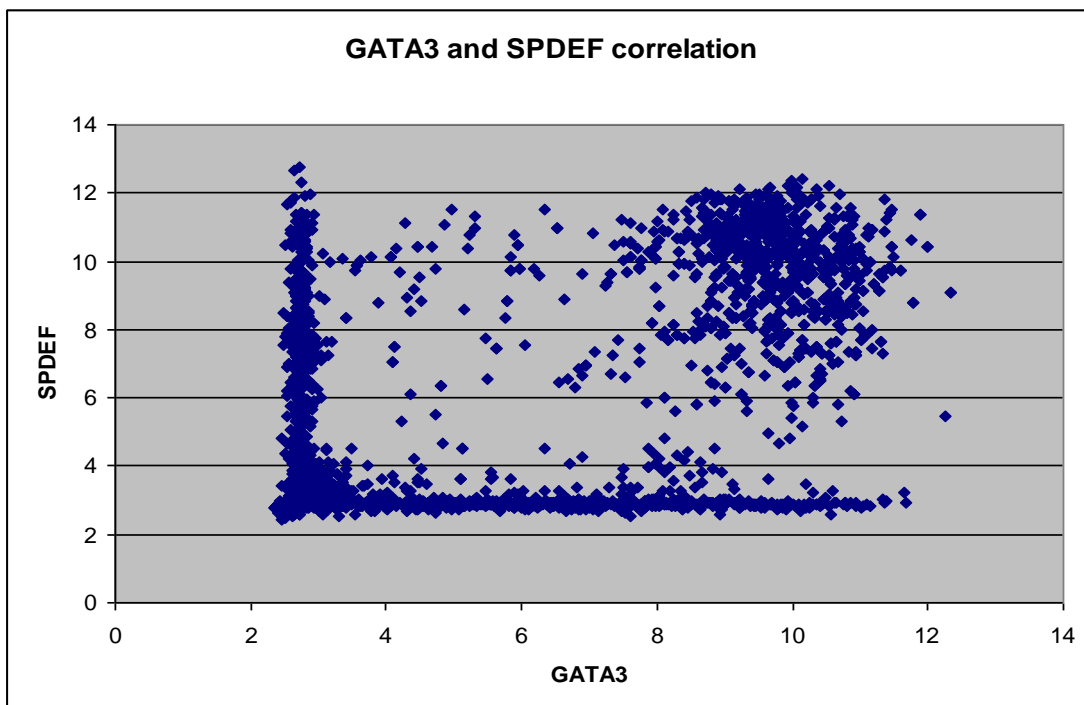


Fig 3.3.4.9: Correlation pattern between GATA3 and SPDEF. Each axis represents the expression values of that individual gene.

3.4.5 Hierarchical clustering, Principal component analysis and k-means analysis on ESR1 correlated genes.

Hierarchical clustering (Fig 3.4.5.1) and Principal component analysis (Fig 3.4.5.2) on the expression values of these 5 ESR1-correlated genes was utilized across all samples to identify similarities among the expression patterns of these genes. Both methods indicated that the FOXA1 and SPDEF expression were very similar to each other.

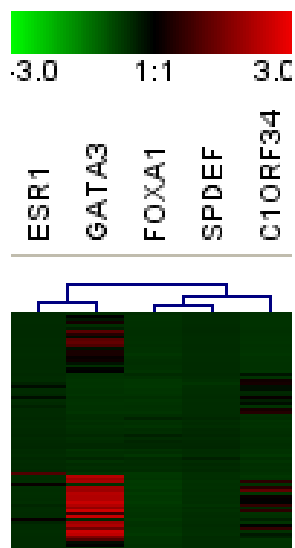


Fig 3.4.5.1: Hierarchical clustering showing relationship among genes.

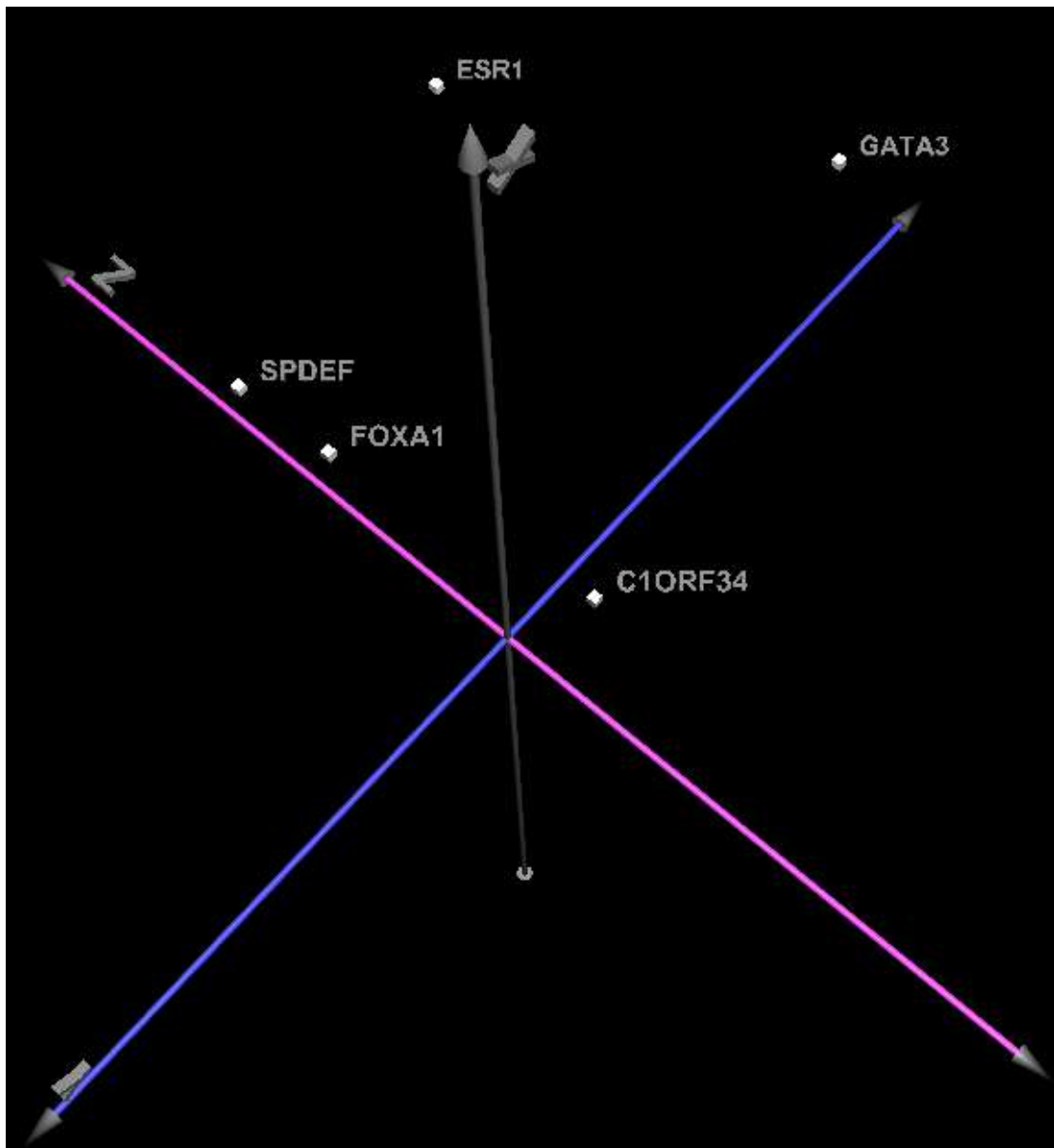


Fig 3.4.5.2: PCA showing relationship among genes. SPDEF and FOXA1 are close to each other on a 3-D representation of the data.

To get a deeper insight on how the expression patterns of these 5 genes are distributed across all samples, k-means clustering was performed on all samples (5897) using the expression values of these 5 genes only (Fig 3.4.5.3). The individual genes were mean centered and divided by standard deviation. The total number of clusters generated was 12, the number of iterations was 200 and the distance criteria were Euclidean distance.

The biggest cluster (3209 samples; Cluster 8) represented specimens which express low levels of these 5 genes. The next biggest cluster (899 samples; Cluster 9) represented specimens where all these 5 genes are highly expressed. This result also indicated that in some of the specimens, ESR1 can be highly expressed even when other four genes are not expressed very highly (Cluster 5). Another interesting result is that the GATA3 expression can be independent of that of the other genes (Clusters 2 and 6).

All the five genes expressing together is the most obvious result from this study (Cluster 9) indicating that mostly these genes are co-expressed. Individual genes can be expressed without depending on the expression of other genes. However, no cluster with high expression of SPDEF and low expression of FOXA1 was observed, indicating that the expression of SPDEF may be dependent on the expression of FOXA1.

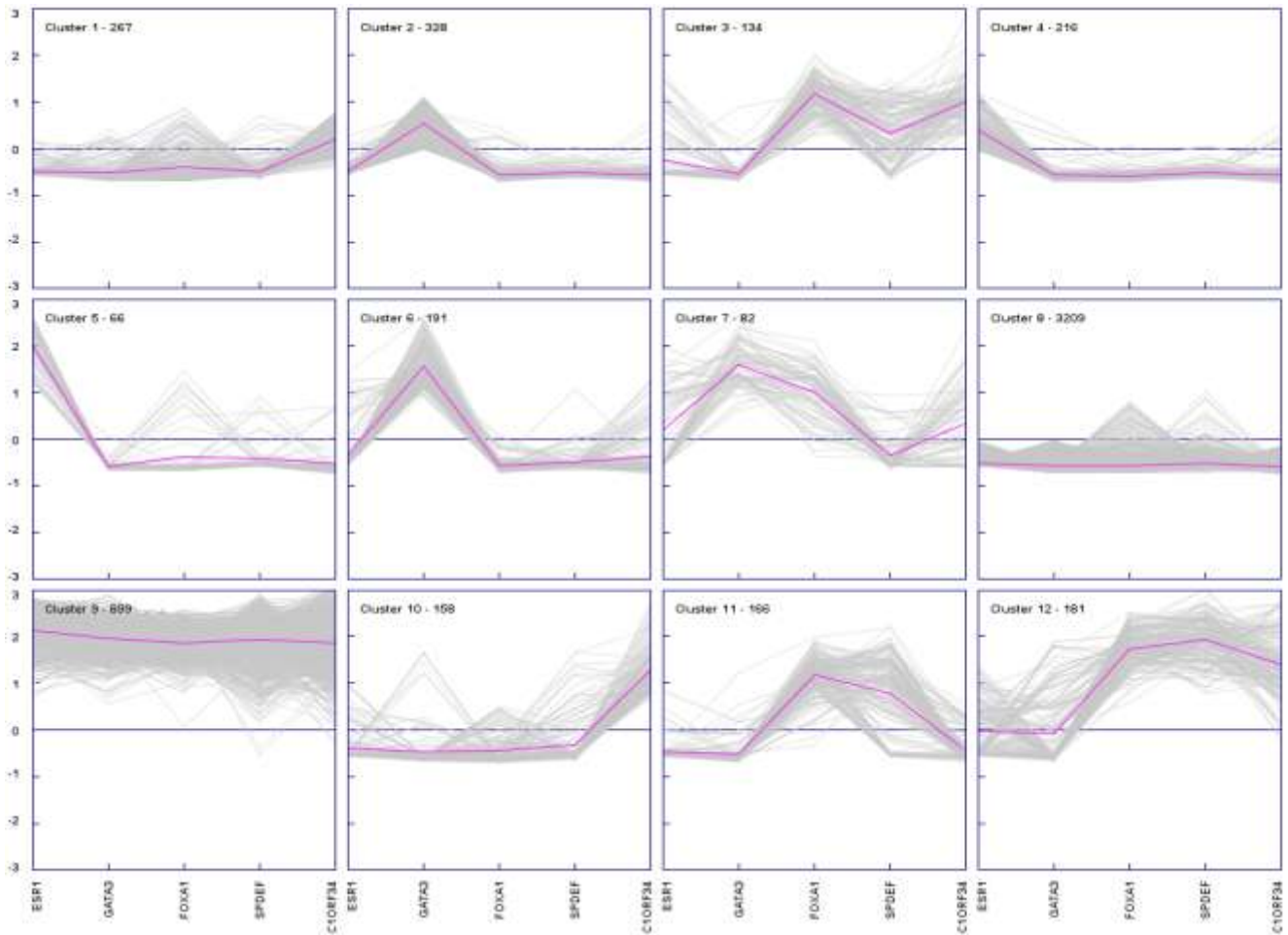


Fig 3.4.5.3: k-means clustering using all samples and 5 genes. The individual genes were mean centered and divided by standard deviation. The total number of clusters generated was 12. Number of iteration was 200 and the distance criteria were Euclidean distance. Total numbers of samples were 5897

3.4.6 Summary

The analysis identified important genes to ER pathway. ESR1, GATA3, FOXA1, SPDEF and C1ORF34 were found to be highly correlated and up-regulated in ER-positive specimens. Nuclear receptor pathway was found to be up-regulated in ER-positive tumors. Using gene expression data a gene interaction network was constructed around ESR1 gene, an important gene in the ER metabolism. The results also indicated that SPDEF expression may be dependent on the expression of FOXA1. SPDEF gene was identified to be up-regulated in ER-negative specimens.

3.4 Gene expression signature for HER2

This section analyses three publicly available datasets to identify diagnostic/prognostic markers associated with HER2-positive tumors.

3.4.1 HER2-positive vs. HER2-negative

Clinical (GSE-1456 and GSE-3744) and Cell Line (GSE-3156) datasets (see section 2.12) were cross compared to identify genes up- and down-regulated in HER2-negative vs. positive patients. The number of specimens in each group and the number of DE genes ($p \leq 0.05$, Fold Change (FC) > 1.2 and Difference > 100) are listed in Table 3.4.1.1. There were 13 transcripts common across all the experiments (Fig 3.4.1.1)

GEO Accession	HER2 -	HER2 +	Up-regulated	Down-regulated
GSE-1456	144	15	421	314
GSE-3744	24	8	101	290
GSE-3156	14	4	898	2616

Table 3.4.1.1: Number of HER2 samples in each experimental group and the number of DE genes in individual comparison

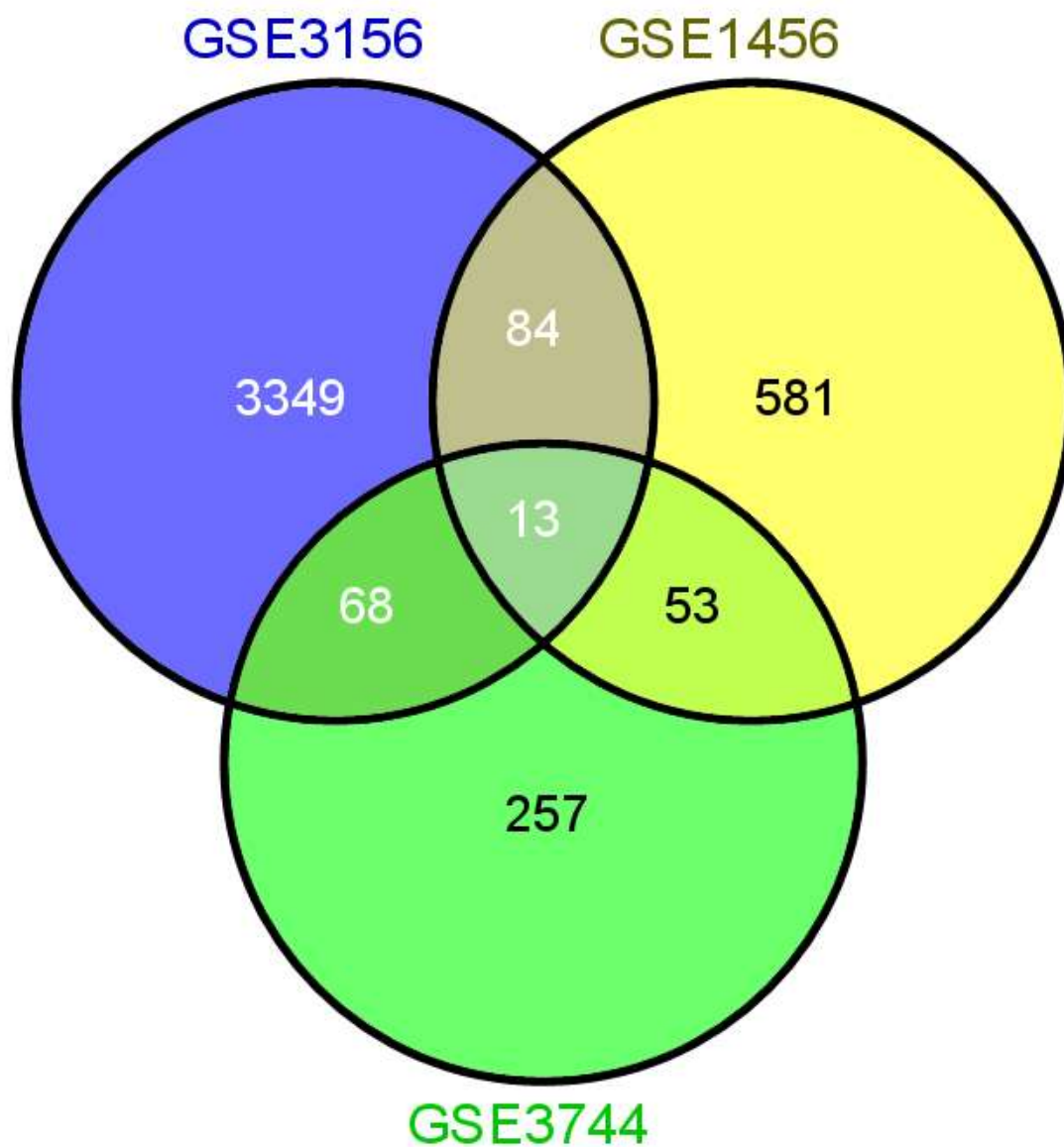


Fig 3.4.1.1: Common DE transcripts across experiments. Venny (see section 2.2.7.2) was used to draw this Venn diagram.

The 6 up-regulated transcripts in all experimental groups are listed in Table 3.4.1.2. The 7 down-regulated transcripts in all experimental groups are listed in Table 3.4.1.3.

probe set	gene	GSE-1456	GSE-3744	GSE-3156
216836_s_at	ERBB2	4.8	17.17	7.51
224447_s_at	C17orf37	4.42	9.54	6.24
202991_at	STARD3	3.36	6.3	5.49
224576_at	ERGIC1	1.58	1.98	2.23
215380_s_at	C7orf24	1.52	1.79	1.96
223847_s_at	ERGIC1	1.49	1.83	1.75

Table 3.4.1.2: Up-regulated transcripts in all experimental groups

probe set	gene	GSE-1456	GSE-3744	GSE-3156
223475_at	CRISPLD1	-2.69	-8.46	-7.53
202037_s_at	SFRP1	-2.5	-4.89	-7.37
202036_s_at	SFRP1	-2.37	-5.57	-22.38
218094_s_at	DBNDD2	-1.91	-2.88	-3.41
205383_s_at	ZBTB20	-1.53	-1.76	-2.92
235308_at	ZBTB20	-1.4	-2.01	-3.97
212190_at	SERPINE2	-1.36	-2.09	-9.01

Table 3.4.1.3: Down-regulated transcripts in all experimental groups

3.4.2 Summary

This section identified ERBB2, C17orf37, STARD3, ERGIC1 and C7orf24 as up-regulated among HER2-positive patients.

3.5 Development of MLPERCEP, a software tool for predicting relapse in breast cancer

Neural Network Multiple Layer Perception (MLPERCEP), using Back Propagation Algorithm, was used to accurately predict relapse in breast cancer patients. A stand-alone piece of software, MLPERCEP was developed to implement Back Propagation Algorithm. The software is available at <http://www.bioinformatics.org/mlpercep/>

3.5.1 Design

The MLPERCEP software is a collection of individual programs written in C language. Each of the C programs is complemented with the graphics user interface written in C# and the individual sets can be accessed from the main user interface also in C#. The C programs extensively use dynamic memory allocation and utilization of hard disk space to make the program practically handle extremely large networks and datasets. To run the software, the .NET runtime environment from Microsoft (<http://www.microsoft.com/downloads/details.aspx?FamilyID=0856EACB-4362-4B0D-8EDD-AAB15C5E04F5&displaylang=en>) must be installed.

3.5.2 Architecture

The architecture of the MLPERCEP back propagation algorithm is shown below. The software has an input neuron layer, a hidden neuron layer and an output neuron layer (Fig 3.5.2.1.)

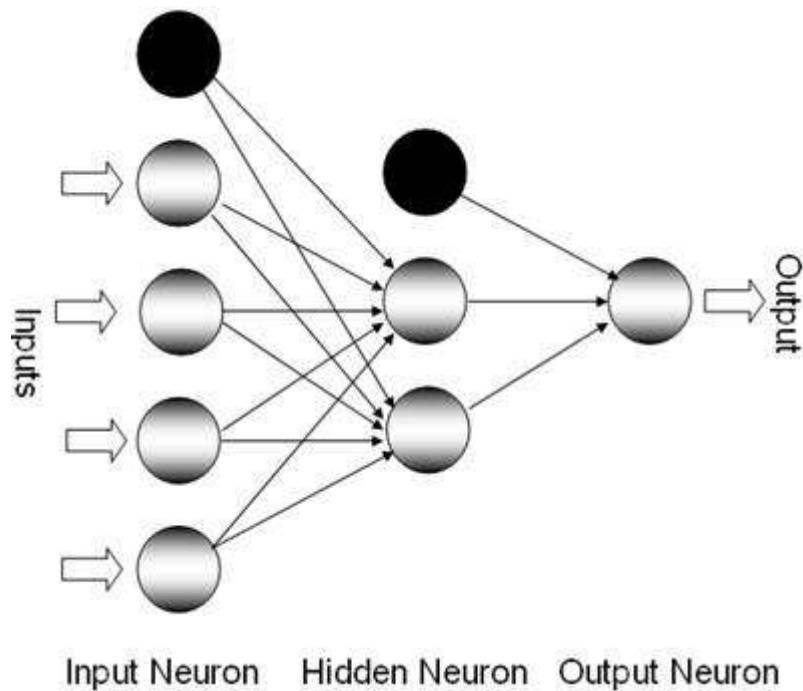


Fig 3.5.2.1: Architecture of back propagation algorithm.

The number of input neuron depends on the number of inputs. The inputs can be binary or floating numbers. However, the results are more accurate if values within -1 and 1 are used. Currently the software supports only binary outputs *i.e.* 0 or 1 or any value between them. The hidden layer is the generalization layer and can be varied by the user. The very high number of hidden neurons may lead to quicker learning, but may fail to generalize. However, a very small number of neurons may not allow the network to train at all. Some of the other fine tuning parameters are learning rate, momentum, error cut-off, maximum number of iterations and leave-one-out cross-validation.

3.5.2.1 Algorithm

The algorithm was adapted from the book by Simon Haykin “Neural Network. A comprehensive foundation” (Haykin 1998)

Forward propagation:

$$\text{Hidden layer output } H_i = F \left(\sum_{i=0}^{i < n} w_i \times I_i \right)$$

w is weight of input to hidden interconnection and I is the input signal, n is the number of input neuron.

$$\text{Output } O = F \left(\sum_{h=0}^{h<n} W_h \times H_h \right)$$

W is weight of hidden to output interconnection and H is the output of hidden neuron, n is the number of hidden neuron.

H is the output of hidden neuron and O is the output of output neuron

w and W are the weights of input-hidden and hidden-output connection.

F is sigmoid activation function $F(x) = 1/(1+e^{-x})$

Backward error propagation and weigh correction

$$\text{Output layer error vector } D = O(1-O)(T-O)$$

T is the desired output

Adjusting the hidden layer weights

$$\Delta W_i = \alpha HD + \theta \Delta W_{i-1}$$

α is the learning rate, θ is the momentum

$$\text{Hidden layer error vector } E_i = H_i(1-H_i)W_iD$$

$$\Delta w_i = \alpha I_i E + \theta \Delta w_{i-1}$$

The above process is iterated till the error sum of squares drops to user defined value or the maximum iteration is reached.

Added features include equal loading of positive and negative examples. This avoids the algorithm's propensity to predict more efficiently on class with more examples at the cost of the class with fewer examples. Another important feature is the run time randomization of input examples. This avoids the problem of the network falling in

local minima. Additional features include leave-one-out cross-validation. This has an advantage on judging the systems efficacy on small datasets where dividing the datasets as training and testing datasets becomes impractical.

3.5.3 Software Modules

Following installation, the application can be launched from Start → All programs → MLPERCEP or from a shortcut on the desktop if it has been made during the installation process. A window appears as shown in Fig 3.5.3.1.

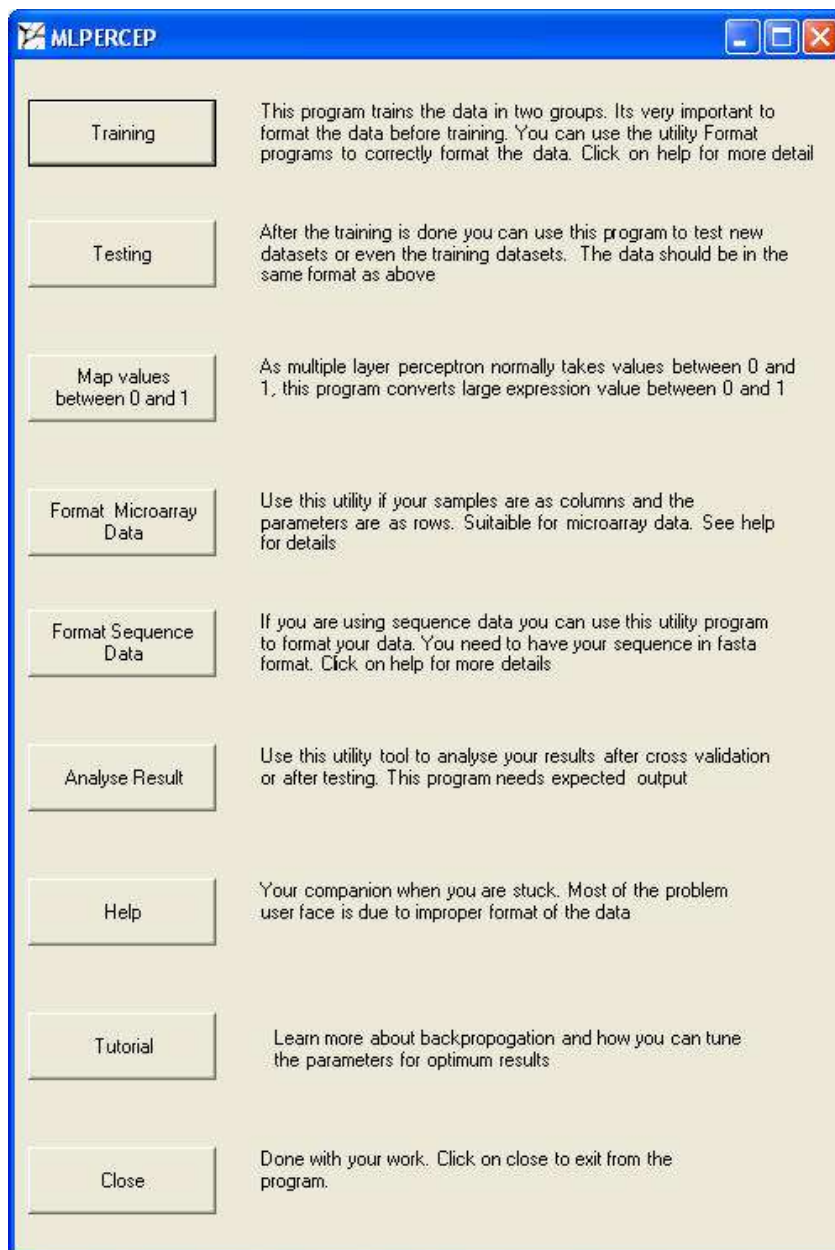


Fig 3.5.3.1: Start up screen of MLPERCEP.

Clicking on the individual button will activate the respective program.

3.5.3.1 Training program

This program trains the network. Clicking on the Training button will produce a window as shown in Fig 3.5.3.1.1.



Fig 3.5.3.1.1: Back propagation training program.

Training file: “Upload” button to select the file to be trained. A sample format is shown below. It should be a text file.

Samples	4		
Parameters	2		
Sample1	0	0	0
Sample2	1	0	1
Sample3	1	1	0
Sample4	0	1	1

The first line should contain the word “samples” or any other word followed by the total number of samples. In this example, it is 4.

The second line should contain the word “parameters” or any other word followed by the total number of input parameters (this will be also taken as the number of input neurons). In this example it is 2.

From the third line onwards, the first word should contain the sample name and the second column should be the expected output. The others in the row are the inputs to the system. It can be binary, e.g. 0 and 1, or it can be decimal (e.g. 0.02). The program will give optimum results only when the data range is between -1 and 1.

If the data has samples as columns and parameters as rows (e.g. microarray data), utility programs “Format Data” can be used. If the data is of DNA sequences, the utility tool “Format sequence data” can be used. This tool will easily convert data to the above-mentioned format.

Weight Output file: All the information of the training in the form of weights will be saved in this file.

No of hidden neurons: This denotes the total number of hidden neurons the network should have. Ideally it should be less than the number of input neurons. If the numbers of hidden neurons are very less, it is possible the system will never learn. However on the other hand if the number of hidden neurons is very high, it may lead to better learning but behave poorly in generalization.

Learning rate: Learning rate decides the speed by which the network will learn. A very small learning rate may take an infinite time for the network to learn. On the other hand a very high learning rate may lead to error bumping and the network not being trained at all. The optimum learning rate for most of the problems lies in the range of 0.4 to 0.6.

Momentum: During the training process, the error convergence reaches a local minima and no further convergence takes place. Placing a momentum (normally between 0.2 - 0.8) helps it come out of the local minima and proceed to the global minima.

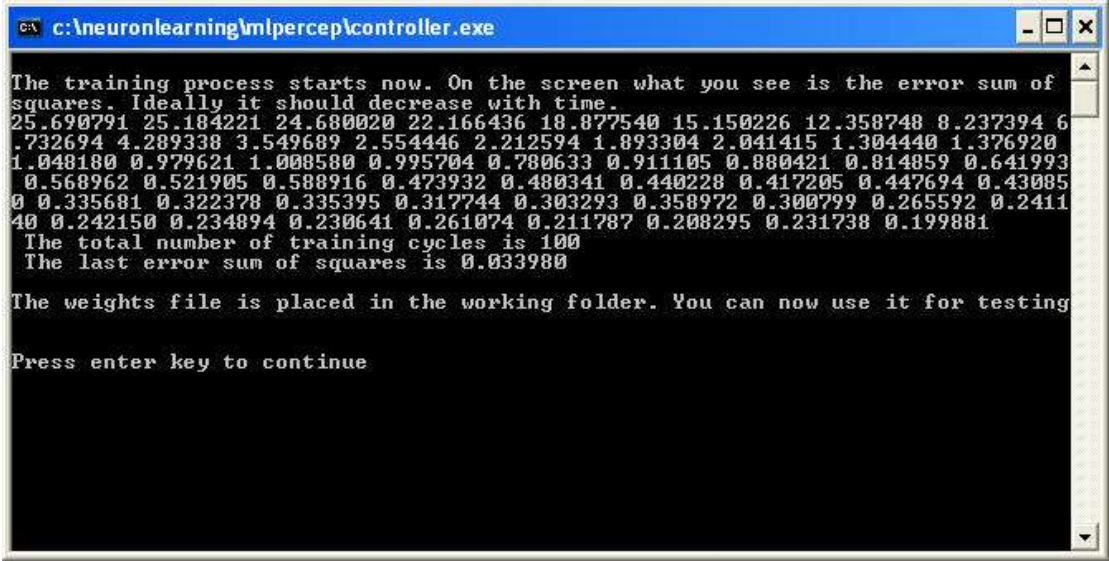
Error cut-off: Error reduces with iteration and the program will stop execution once it reaches the error cut-off or the maximum number of iterations.

Maximum iteration: The total number of training cycles performed by the program. However if the error reaches below error-cut-off, the program will stop. The lower the error cut-off, the better the results are.

Copy: The program makes a copy of the dynamic changing weights after “copy” number of iterations. So if the program is stopped between iterations, the user can obtain the previous copy of the weights.

Cross validation output file: By default, the system does not perform any cross-validation. If a file name is provided, the system will perform cross-validation. The system performs a leave-one-out cross-validation which is a very computationally-intensive process.

Train: Clicking on “Train” will open an MSDOS window as shown below (Fig 3.5.3.1.2).

A screenshot of a DOS command window titled 'c:\neuronlearning\mlpercep\controller.exe'. The window contains text describing the training process and a list of error sum of squares values. The text reads: 'The training process starts now. On the screen what you see is the error sum of squares. Ideally it should decrease with time.' followed by a list of 40 numerical values. Below the list, it says 'The total number of training cycles is 100' and 'The last error sum of squares is 0.033980'. At the bottom, it says 'The weights file is placed in the working folder. You can now use it for testing' and 'Press enter key to continue'.

```
c:\neuronlearning\mlpercep\controller.exe

The training process starts now. On the screen what you see is the error sum of
squares. Ideally it should decrease with time.
25.690791 25.184221 24.680020 22.166436 18.877540 15.150226 12.358748 8.237394 6
.732694 4.289338 3.549689 2.554446 2.212594 1.893304 2.041415 1.304440 1.376920
1.048180 0.979621 1.008580 0.995704 0.780633 0.911105 0.880421 0.814859 0.641993
0.568962 0.521905 0.588916 0.473932 0.480341 0.440228 0.417205 0.447694 0.43085
0 0.335681 0.322378 0.335395 0.317744 0.303293 0.358972 0.300799 0.265592 0.2411
40 0.242150 0.234894 0.230641 0.261074 0.211787 0.208295 0.231738 0.199881
The total number of training cycles is 100
The last error sum of squares is 0.033980

The weights file is placed in the working folder. You can now use it for testing

Press enter key to continue
```

Fig 3.5.3.1.2: MSDOS screenshot showing error sum of squares

These values are the error sum of squares. Ideally these values should decrease with time, indicating that the network is learning. The error sum of squares will reach zero when the software reaches the minimum error cut off or the maximum iteration is reached. If the error is not further reducing, the MSDOS window can be closed and the last set of weights will be taken for further calculations.

3.5.3.2 Testing

After training the data the network can be used for testing on the unknown samples. Clicking on “Testing” button will activate the window shown in Fig 3.5.3.2.1.



Fig 3.5.3.2.1: Back propagation testing program.

Weight File: The training program exports the knowledge in the form of a weight file. The weights are the optimised weights for correct prediction.

Testing File: The data format is the same as that for the training program. However, it does not need to have the target as the second row.

Output file: File where the results are to be saved.

Data contains expected output: If the test set data contains information on the output for each sample, check this option to test the algorithm efficiency

Click on “Test” and the testing program will start.

3.5.3.3 Map data between 0 and 1

Since the expression values may be quite high, this tool maps the data between 0 and 1. It uses the following calculation.

New value = (Old value - Minimum value for that gene)/ (Maximum value for that gene - Minimum value of that gene)

This tool should be used only after the feature selection; otherwise it will map the less changing data also between 0 and 1 thereby magnifying the error.

The data should be in the following format. It should be in a text file.

Identifier	s1	s2	s3	s4	s5	s6
Target	1	0	0	1	1	0
G1	3	8	2	9	1	2
G2	9	4	2	9	6	5
G3	3	2	7	9	3	2

First row and column should contain the word “Identifier”, or any other word, followed by the names of the identifiers.

Second row first column should contain the word “Target”, or any other word. Following this should contain the expected output of that sample, 0 for one class and 1 for the other class. If this row is present, the option “Second column contains target” should be checked.

The following rows should contain gene identifier followed by expression values.

Clicking on “Map values between 0 and 1” will activate the window shown in Fig 3.5.3.3.1.

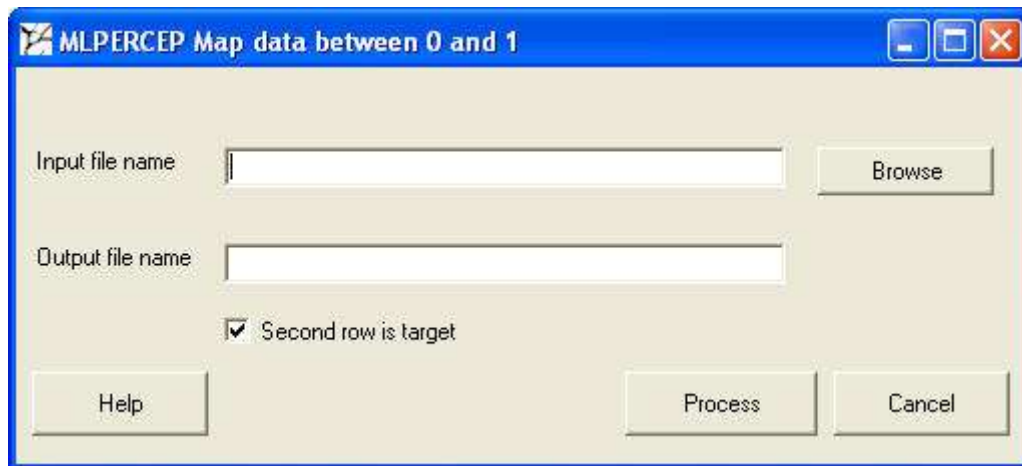


Fig 3.5.3.3.1: Map data between 0 and 1.

3.5.3.4 Formatting data

This program will convert gene expression data to the format in which the training program can accept. Clicking on “Format microarray data” will activate the window shown in Fig 3.5.3.4.1.

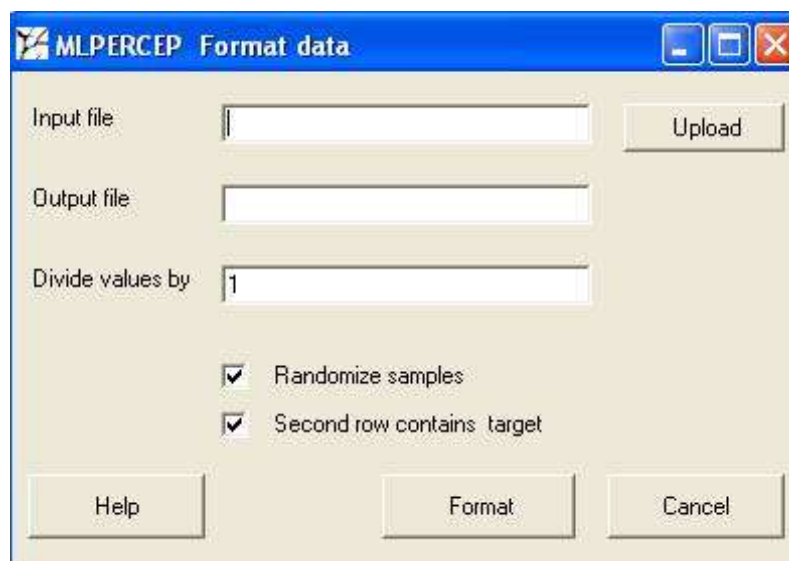


Fig 3.5.3.4.1: Format microarray data.

The data should be in the following format. It should be a text file.

Identifier	s1	s2	s3	s4	s5	s6
Target	1	0	0	1	1	0
G1	3	8	2	9	1	2
G2	9	4	2	9	6	5
G3	3	2	7	9	3	2

The first row and column should contain the word Identifier, or any other word followed by the names of the Identifier.

The second row, first column should contain the word “Target” or any other word. Following this should contain the expected output of that sample 0 for one class and 1 for the other class. This row is essential for the training data. However it is optional for the testing data. If this row is present the option “Second column contains target” should be checked or else unchecked. Target can be in the second row even in the testing file. In such case it will give a comparative result with the expected output. The following rows should contain gene identifier followed by expression values

Input file: Select the file to be formatted.

Output file: Specify the output file name.

Divide values by: The program trains best when the data range is in between 0 and 1. So if the data lies outside this range, divide the values by a particular number so as the values fall between 0 and 1.

Randomize samples: Randomizing samples is always helpful for the training. It randomizes the order of training examples fed to the program. If all positive examples are shown at one time and all the negative examples after that, then the system keeps on forgetting the previous class. Therefore it is essential to train the network with alternating positive and negative examples, so that the network remains balanced and does not deviate towards predicting one class at the cost of other.

Second row contains target: The second row should contain the known outcomes of individual samples as shown in the previous example and this is a must for the training file. However, it's optional for the testing set.

Click on "Format". The output file will be placed in the working folder.

3.5.3.5 Format Sequence data

If the data is in the form of a DNA sequence, this utility tool will format the data to numeric values. The submitted sequence should be in FASTA format.

Click on "Format Sequence Data" will activate the window shown in Fig 3.5.3.5.1.

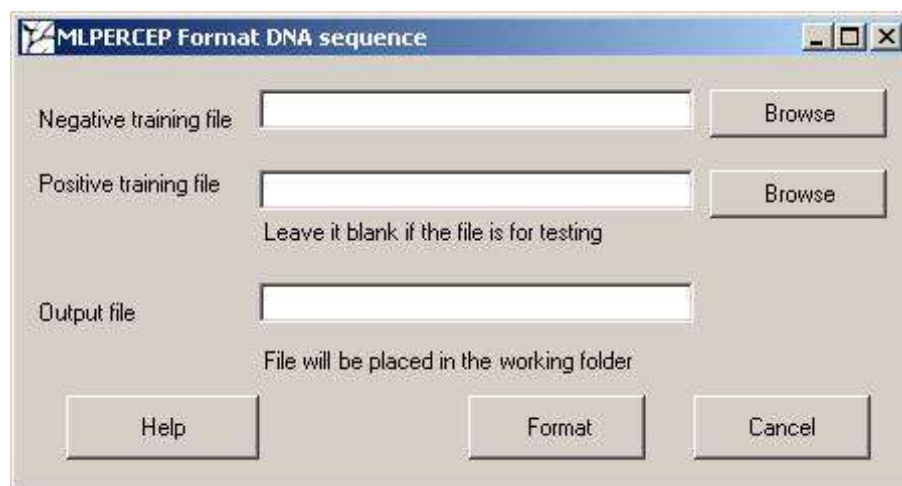


Fig 3.5.3.5.1: Format DNA sequence.

An example of input data is shown below:

>Seq1 Hypothetical gene

ATGCGTA

>Seq2 Kinase

TTCTAAC

This step will be repeated. The first word will be taken as the sequence name. For the above example it will be Seq1 and Seq2

Negative training file: Upload the file with sequences which belong to one group. Here it will be assigned 0.

Positive training file: Upload the file with sequences which belong to the other group. Here it will be assigned 1.

The sequence are numerically transferred as

A → 0 0

T → 0 1

G → 1 0

C → 1 1

Unidentified bases will be assigned 0 0. The length of sequences should be equal.

Output file: Specify the output file name.

Click on “format”. The output file will be placed in the working folder

For testing purposes, only one file is needed. Leave the field “Positive Training File” empty. The system will consider that the formatting is done for the testing purpose

3.5.3.5 Analyse Results

This tool can be used to analyse the results obtained from the cross validation study or after testing on a new result where the expected output is known. Clicking on “Analyse Results” will activate the window shown in Fig 3.5.3.5.1.

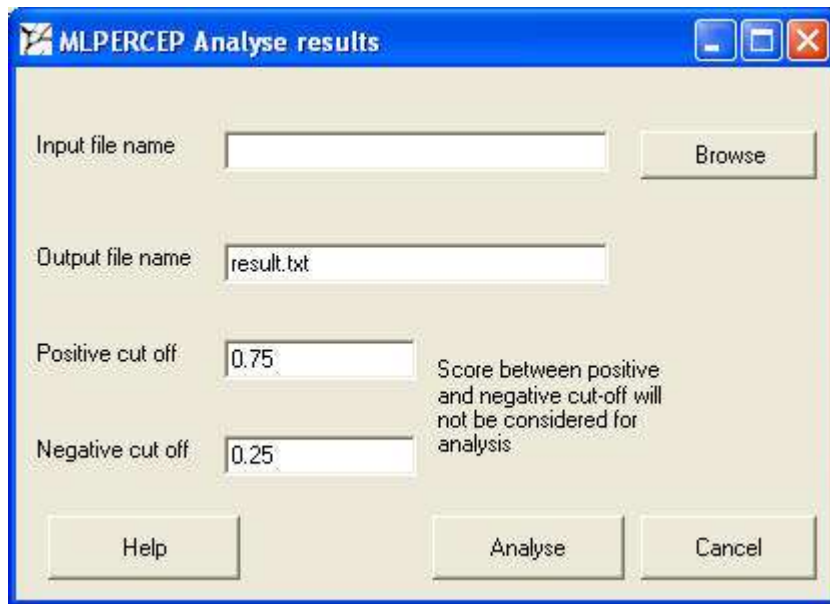


Fig 3.5.3.5.1: Analyze results.

Input file name: Input file name contains the actual output of each sample along with the expected value/class. It is the output of testing or cross-validation.

Positive and negative cut-off: Values between the positive and negative cut-off won't be considered for classification and their class would be considered as undermined. Having a stringent cut-off will increase the accuracy; however it will increase the number of samples as undetermined.

The output includes Overall accuracy, Positive and Negative accuracy, True and False positive rates. "Positive accuracy" (also known as true positive rate) is the accuracy of positive class. Similarly, negative accuracy (also known as true negative rate) is the accuracy of the other class. False positive rate and false negative rate are the error rates associated with the positive and negative class respectively.

An example of the result file is shown below

Positive cut-off:	0.750000
Negative cut-off:	0.250000
Total number of samples:	34
Total number of samples which could not be classified:	1
Total number of samples which were accounted:	33

a=19 b=0 c=1 d=13
 Positive accuracy: 0.928571
 Negative accuracy: 1.000000
 Accuracy: 0.969697
 True positive rate: 0.928571
 False positive rate: 0.000000
 True negative rate: 1.000000
 False negative rate: 0.071429

The calculations are shown below.

	Predicted	Actual
a	0	0
b	1	0
c	0	1
d	1	1

Positive accuracy = $d/(d+c)$
 Negative accuracy = $a/(a+b)$
 Overall Accuracy = $(a+d)/(a+b+c+d)$
 True positive rate = $d/(c+d)$
 False positive rate = $b/(a+b)$
 True negative rate = $a/(a+b)$
 False negative rate = $c/(c+d)$

3.5.4 Results

The aim was to develop a classifier to predict relapse in breast cancer patients. Differentially-expressed genes were generated by comparing the patients who relapsed (overall relapse) compared to patients who did not relapse. The total number

of samples was 105. The filtration criteria used was $p \leq 0.001$. There were a total of 162 genes meeting the criteria and a classifier was developed on those genes. The 162-member gene signature expression values was normalised to between 0 and 1, as discussed in section 3.5.3.3. The data was used to optimize the neural network back propagation program. The optimization was done by varying the number of hidden neurons, learning rate and the momentum. The results are indicated in Fig 3.5.4.1, Fig 3.5.4.2 and Fig 3.5.4.3

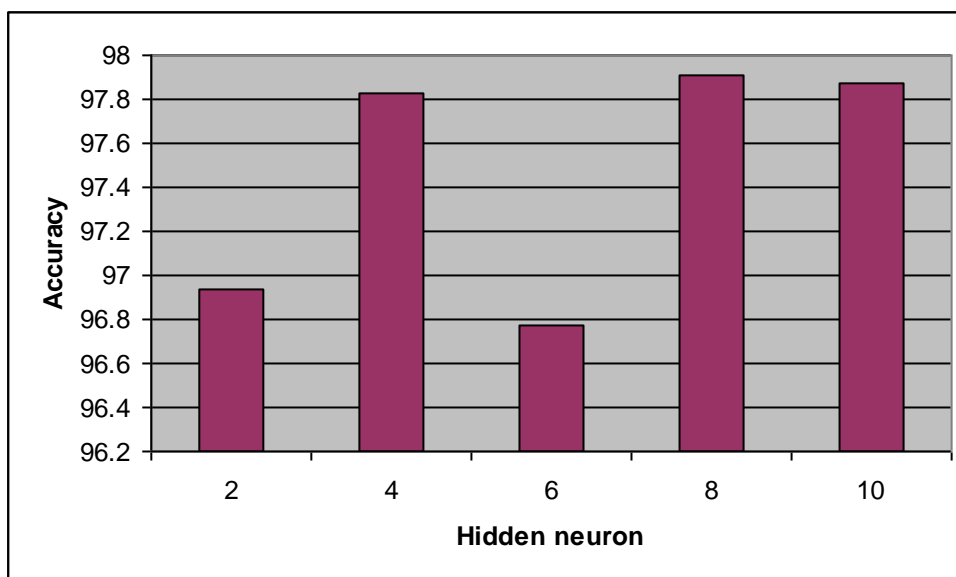


Fig 3.5.4.1: Leave-one-out cross validation accuracy on varying the hidden neurons

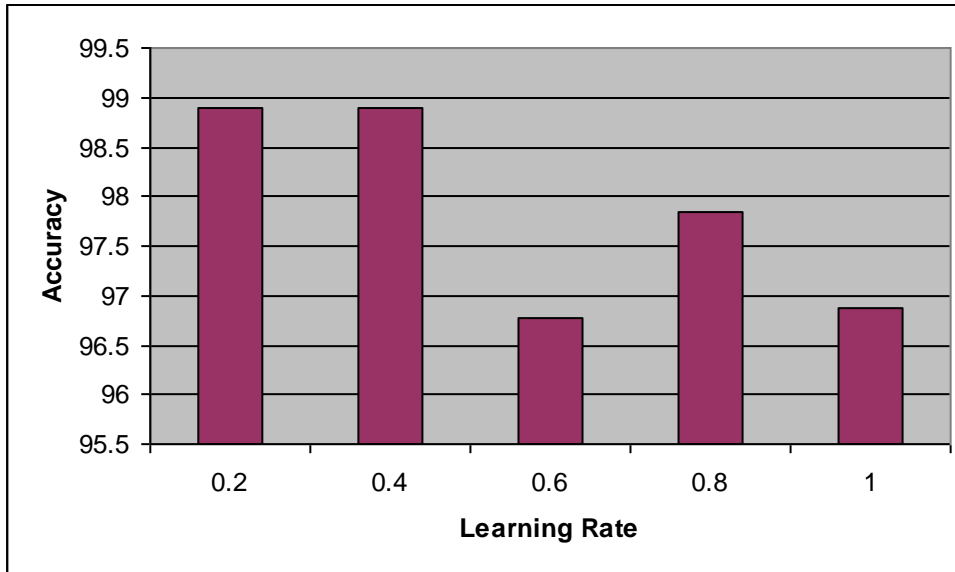


Fig 3.5.4.2: Leave-one-out cross validation accuracy on varying the learning rate

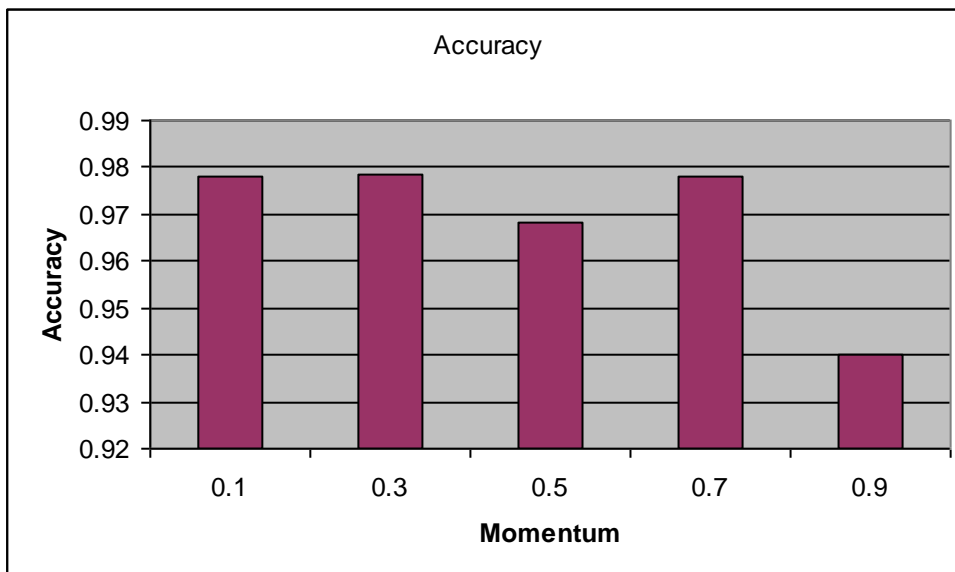


Fig 3.5.4.2: Leave-one-out cross validation accuracy on varying the momentum.

The results indicate that 4, 8 and 10 hidden neurons gave better accuracy. 4 hidden neurons were taken for further analysis as lower number of hidden neuron makes the training process fast and should perform better on independent validation. A learning rate of 0.2 and 0.4 was found to give a better accuracy. Therefore a learning rate of 0.4 was taken to obtain faster learning. Lower momentum performed better. Therefore a momentum of 0.1 was taken for subsequent analysis.

Optimised parameters were used: (Hidden neurons: 4; Learning rate: 0.4; Momentum: 0.1; Error cut-off: 0.04; Maximum number of iterations: 10000). The leave-one-out cross-validation method was used to estimate the accuracy of the model. Additionally, another classifier was developed using Support Vector Machines using GEMS software. Default parameters were used (SVM cost: 100; SVM kernel: Polynomial and SVM kernel parameter (degree): 1) using the same dataset.

When the cut-off 0.75 and 0.25 was used, the accuracy of the back propagation model was 97.8%. However there were a total of 11 specimens which could not be classified. When the cut-off was taken as 0.5, and all the specimens were grouped, the overall accuracy was 93.3%. The detail results are shown in Table 3.5.4.1. A support vector machine algorithm using the above data was able to classify with an accuracy of 93.3%.

Parameters	Positive cut-off: 0.75	Positive cut-off: 0.5
	Negative cut-off: 0.25	Negative cut-off: 0.5
A	51	53
B	2	4
C	0	3
D	41	45
Unclassified	11	0
Accuracy	0.978723	0.933333
True positive rate	1	0.937500
False positive rate	0.037736	0.070175
True negative rate	0.962264	0.929825
False negative rate	0	0.062500

Table 3.5.4.1: Results from Back propagation cross validation program.

A cDNA microarray dataset generated by van't Veer *et al.*, (2002), was used to develop a classifier to predict distant metastasis. This was done to judge how the algorithm performs on cDNA microarray data. The total number of specimens used was 78. DE genes ($p < 0.001$) were used to develop the classifier. The number of genes that passed this criterion was 117, and a classifier was developed on them. The

data was used without any transformation as the cDNA microarray values were close to 0 as the data was log ratio. Default parameters for both were used and leave-one-out cross-validation was performed. When the cut-off 0.75 and 0.25 was used, the accuracy of the back propagation model was 89.7%. However, there were a total of 10 specimens which could not be classified. When the cut-off was taken as 0.5, and all the specimens were grouped, the overall accuracy was 87.1%. The detail results are shown in Table 3.5.4.2. A support vector machine using the above data and default parameters was able to classify the data with an accuracy of 82.05%.

Parameters	Positive cut-off: 0.75 Negative cut-off: 0.25	Positive cut-off: 0.5 Negative cut-off: 0.5
A	37	40
B	2	4
C	5	6
D	24	28
Unclassified	10	0
Accuracy	0.897059	0.871795
True positive rate	0.827586	0.823529
False positive rate	0.051282	0.090909
True negative rate	0.948718	0.909091
False negative rate	0.172414	0.176471

Table 3.5.4.2: Results from Back propagation cross validation program.

3.5.5 Summary

A back propagation algorithm was successfully developed as a user-friendly software package which can be used to develop a prognostic model for breast cancer. The results generated were at par or better than Support Vector Machines in predicting relapse and distant metastasis in two of the datasets tested.

3.6 Functional analysis on ROPN1B

Ropporin is a sperm-specific protein and is associated with sperm motility (Fujita *et al.*, 2000). Its expression was also found in motile cilia helping them to move in one direction in a synchronised pattern (Newell *et al.*, 2008). Ropporin (ROPN1 and ROPN1B) was identified as differentially-expressed in several gene lists commonly associated with bad prognosis in our breast cancer investigation (see section 3.1) This gene was significantly up-regulated in patients who relapsed, patients who did not survive beyond 5 years, patients who relapsed within 5 years and patients with ER-negative tumors. Additionally this gene was up-regulated in one of the sub-groups of ER-negative breast cancer with a high incidence of relapse (see section 3.1.3). The gene was also up-regulated in the ER-negative subgroup in 3 out of 6 datasets comparing ER-negative and ER-positive specimens (see section 3.3.2). There was a total of 4 probe sets representing the Ropporin gene on the Affymetrix U133 Plus2.0 microarray chip, which when annotated in NetAffx, corresponded to two highly homologous gene targets, ROPN1 and ROPN1B, located at two different loci on the same chromosome.

3.6.1 Similarity and difference among ROPN1B and ROPN1

3.6.1.1 Sequence similarity

Using ClustalW, ROPN1B and ROPN1 were found to be 97% identical based on DNA (Coding sequence) sequence similarity (Fig 3.6.1.1.1) and 95.7% identical based on the predicted protein sequence similarity (Fig 3.6.1.1.2).

```

ROPN1      ATGGCTCAGACAGAT AAGCC AAC ATGC ATCCC GCCGGAGCTGCCG AAG ATGCTG AAGGAG 60
ROPN1B     ATGGCTCAGACAGAT AAGCC AAC ATGC ATCCC GCCGGAGCTGCCG AAAATGCTG AAGGAG 60
*****

ROPN1      TTTGCC AAAAGCCGCC ATT AAGGTGC AGCCGC AGGACCTCATCC AGTGGGCAGCCGATT AT 120
ROPN1B     TTTGCC AAAAGCCGCC ATT CGGGCGC AGCCGC AGGACCTCATCC AGTGGGGGCCGATT AT 120
*****

ROPN1      TTTGAGGCCCTGTCCCGTGGAGACAGCCCTCCGGTGAGAGAGCGGTCTGAGCCAGTCCGT 180
ROPN1B     TTTGAGGCCCTGTCCCGTGGAGAGACAGCCCTCCGGTGAGAGAGCGGTCTGAGCCAGTCCGT 180
*****

ROPN1      TTGTGT AACCGGCCAGAGCT AAC ACCTGAGCTGTT AAAAGATCCTGCATTCTCAGGTTGCT 240
ROPN1B     TTGTGT AACCGGCCAGAGCT AAC ACCTGAGCTGTT AAAAGATCCTGCATTCTCAGGTTGCT 240
*****

ROPN1      GGCAGACTGATCATCCGTGCAGAGGAGCTGGCCCA GATGTGGA AAGTGGTGAATCTCCCA 300
ROPN1B     GGCAGACTGATCATCCGTGCAGAGGAGCTGGCCCA GATGTGGA AAGTGGTGAATCTCCCA 300
*****

ROPN1      ACAGATCTGTTT AAT AGTGTGATGAATGTGGGTCGCTT CACGGAGGAGATCGAGTGGCTG 360
ROPN1B     ACAGATCTGTTT AAT AGTGTGATGAATGTGGGTCGCTT CACGGAGGAGATCGAGTGGCTG 360
*****

ROPN1      AAGTTTTT AGCCCTTGCTTG CAGCGCTCTGGGAGTT ACTATT ACCAAAACTCTCAAGATA 420
ROPN1B     AAGTTTTT AGCCCTTGCTTG CAGCGCTCTGGGAGTT ACTATT ACCAAAACTCTCAAGATA 420
*****

ROPN1      GTGTGTGAGGTC TTATCATGTGACCAT AATGGTGGGTCGCCCGGATCCCGTT CAGCACC 480
ROPN1B     GTGTGTGAGGTC TTATCATGTGACCAT AATGGTGGGTCGCCCGGATCCCGTT CAGCACC 480
*****

ROPN1      TTCCAGTTTCTCTACACGT ATATTGCC AAAAGTGGATGGGGAGATCTCTG CACATGTC 540
ROPN1B     TTCCAGTTTCTCTACACGT ATATTGCC AAAAGTGGATGGGGAGATCTCTG CACATGTC 540
*****

ROPN1      AGCAGGATGCT AAAC TACATGGAACAGGAAGT AATTGGCCCTGATGGT ATATCACAGTG 600
ROPN1B     AGCAGGATGCT AAAC TACATGGAACAGGAAGT AATTGGTCCTGATGGTTT AATCACGGTG 600
*****

ROPN1      AATGACTTTACCCAAAACCCAGGGTT CAGCTGGAGTAA 639
ROPN1B     AATGACTTTACCCAAAACCCAGGGTT CAGCTGGAGTAA 639
*****

```

Fig 3.6.1.1.1: DNA sequence alignment for ROPN1 and ROPN1B

```

ROPN1      MAQTDKPTCIPPELPKMLKEFAKAAIRVQPQDLIQWAADYFEALSRGETPPVRESEVA 60
ROPN1B    MAQTDKPTCIPPELPKMLKEFAKAAIRAQPQDLIQWGADYFEALSRGETPPVRESEVA 60
          *****_*****_*****

ROPN1      LCNRAELTPELLKILHSQVAGRLIIRAEELAQMVKVNLPTDLFNSVMNVGRFTEEIWL 120
ROPN1B    LCNWAELTPELLKILHSQVAGRLIIRAEELAQMVKVNLPTDLFNSVMNVGRFTEEIWL 120
          *** *****

ROPN1      KFLALACSA LGVTITKTLKIVCEVLS CDHNGGSPRIPFSTFQFLYTYIAKVDGEISASHV 180
ROPN1B    KFLALACSA LGVTITKTLKIVCEVLS CDHNGGLPRIPFSTFQFLYTYIAEVDGEICASHV 180
          *****:*****:*****_****

ROPN1      SRMLNYMEQEVI GPDGIITVNDFTQNPRVQLE 212
ROPN1B    SRMLNYIEQEVI GPDGLITVNDFTQNPRVWLE 212
          *****:*****:***** **

```

Fig 3.6.1.1.2: Protein sequence alignment for ROPN1 and ROPN1B

3.6.1.2 Design of gene specific primers for ROPN1 and ROPN1B

TaqMan primers were designed for ROPN1 and ROPN1B using primer express from ABI (see section 2.5.5.1). The primers were designed using the most variable region of the gene (close to 3' region). The forward and reverse primers were on different exons making it specific to detect mRNA only.

ROPN1 primers:

Sense: TGT CAG CAG GAT GCT AAA CTA CAT G Tm: 61.3, GC content 44%
Location: 878 – 902 on NM_017578.2

Antisense: ATT TTG GGT GGT ATA TGG GTT TCA Tm: 57.6, GC content: 37.5%
Location: 1062-1039

Probe: CAG CTG GAG TAA AAG CAC AAT TTT GGC AA Tm: 63.9, GC content 41.4%, 5' JOE 3' TAMARA Location: 969 – 997

ROPN1B primers:

Sense: TGT CAG CAG GAT GCT AAA CTA CAT T Tm: 59.7 GC content: 40%
Location: 765-789 in NM_001012337.1

Antisense: AGG TGG TAT ATG GGT TTA TCA TTC TGA Tm: 59, GC content 37% Location: 942-916

Probe: TGG CTG GAG TAA CAG CAC AAT TTT GGC Tm: 65. GC content: 48.1

5' FAM 3' TAMRA Location 856 - 882

Specificity of primers: ROPN1 and ROPN1B plasmids in bacteria were obtained from Open Biosystems (see section 2.5.7). The plasmids were isolated using mini-prep kit from Qiagen (see section 2.5.6.1). qRT-PCR was performed on these plasmids for both genes.

To determine the efficacy of the ROPN1 and ROPN1B primers in differentiating their respective cDNAs, qRT-PCR experiments were set up utilizing both sets of primers on both plasmid preps from both genes. As can be seen in Fig 3.6.1.2.1, the designed ROPN1 primers were specific to ROPN1 and did not amplify ROPN1B.

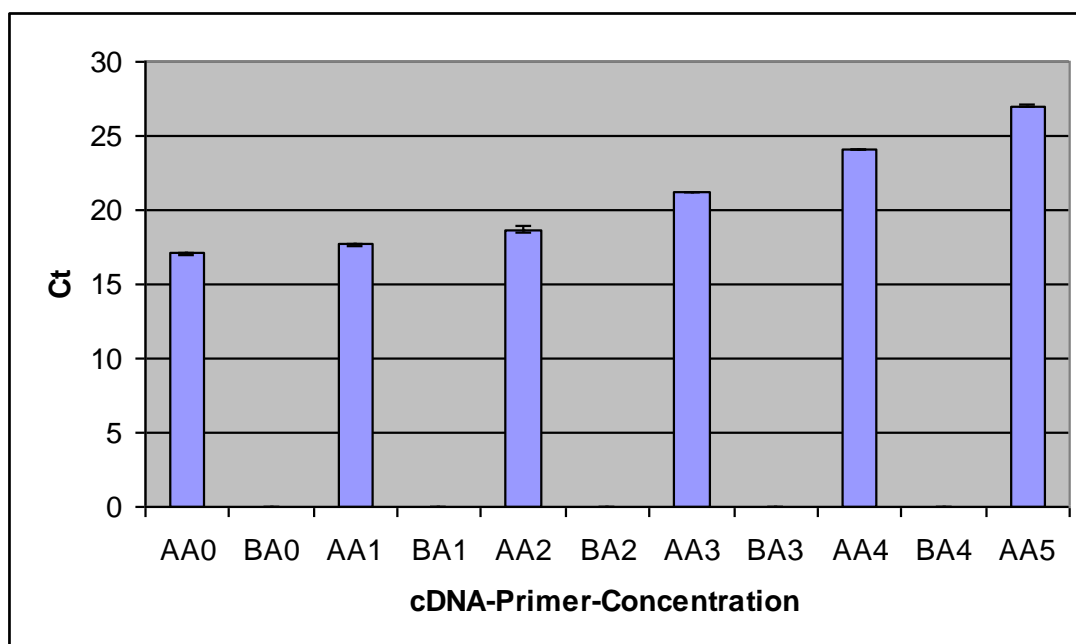


Fig 3.6.1.2.1: ROPN1 qRT-PCR on ROPN1 expression plasmid preps. ROPN1 primers amplified ROPN1 gene transcript from ROPN1 expression plasmid preps and did not amplify ROPN1B. On the X-axis, the first letter indicates the target (i.e. cDNA; A-ROPN1, B-ROPN1B) source, the second letter indicates the primer type and the number indicates the 1/x dilution. On the Y-axis, “Ct” refers to “Cycle Threshold” and is a measure of the cycle number at which the fluorescence generated within a reaction crosses the threshold. It is inversely correlated to the logarithm of the initial copy number.

The ROPN1B primers were efficient in detecting ROPN1B cDNA. However, ROPN1B primers did amplify ROPN1 but at a later cycle (~6 Ct), making the detection of ROPN1 by ROPN1B primer 100-fold less specific compared to its ability to detect ROPN1B (Fig 3.6.1.2.2). Every 3.2 Ct difference is equivalent to a 10 fold difference in gene expression. Due to the sequence conservation between ROPN1 and ROPN1B, it was impossible to design alternate primers. As a result, it was concluded that to accurately assess ROPN1B expression using these primers, all ROPN1B qRT-PCR experiments would be complemented with ROPN1 expression, to identify whether a given expression value was due to ROPN1B or ROPN1.

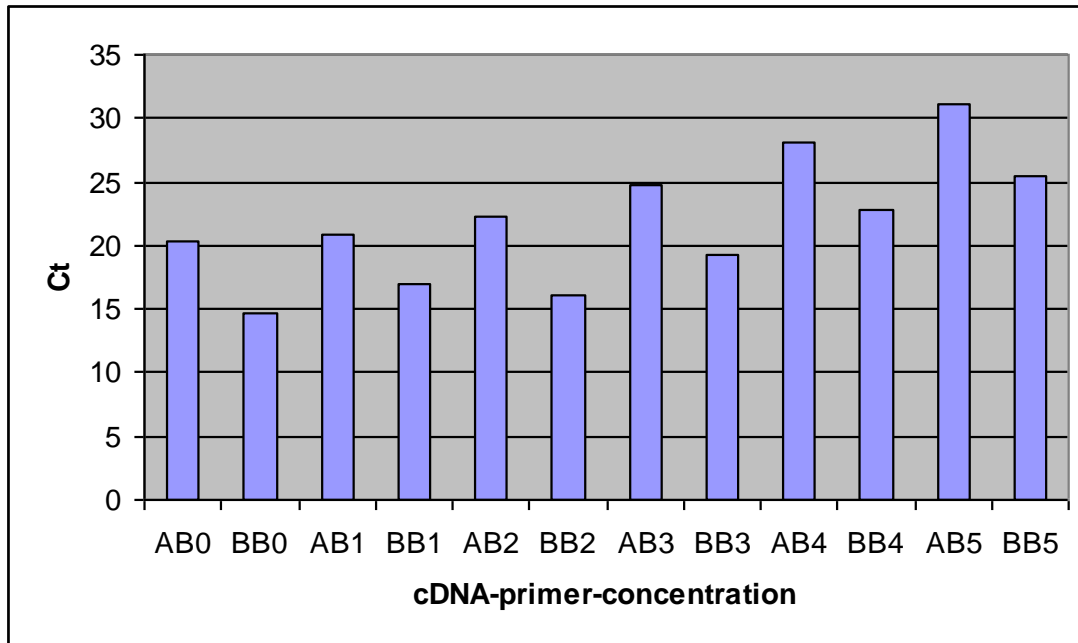


Fig 3.6.1.2.2: ROPN1B qRT-PCR on ROPN1/ROPN1B expression plasmid preps. The X/Y axes are as detailed for Fig 3.6.1.2.1. The detection of ROPN1 by ROPN1B primer was 100-fold less specific compared to its ability to detect ROPN1B.

3.6.1.3 Affymetrix probe presentation of Ropporin (ROPN1 and ROPN1B)

There are four probe sets representing Ropporin gene on Affymetrix U133 plus arrays, whereas there is only one on the U133A chip. To get a deeper understanding regarding which probes represent ROPN1 and which represented ROPN1B sequences for all these probes were BLASTED.

Probe: 233203_at: On Affymetrix arrays probe ID with “_at” are considered as unique to the specified gene and are not supposed to hybridize with other genes. BLASTING the sequence resulted in no match with any of the reference sequence of ROPN1 or ROPN1B. However, there was a match with ROPN1 gene (not the reference sequence) (Fig 3.6.1.3.1).

Two years previously when there was no reference sequence available for ROPN1, it was assumed that this probe represented ROPN1. And since this probe was not expressing in our study, it was assumed that the expression detected was due to ROPN1B and not ROPN1. With the updated annotation and qRT-PCR results, it was

concluded (at a much later date) that the unique probe for ROPN1 (233203_at) on Affymetrix chip is a faulty probe and does not represent ROPN1.

```
> gi|6599263|emb|AL133624.1|HSM801491 UEG Homo sapiens mRNA; cDNA DKFZp434B1222 (from clone DKFZp434B1222)
Length=2049

GENE ID: 54763 ROPN1 | ropporin, rhophilin associated protein 1 [Homo sapiens]
(10 or fewer PubMed links)

Score = 704 bits (780), Expect = 0.0
Identities = 404/404 (100%), Gaps = 0/404 (0%)
Strand=Plus/Plus

Query 1 ATGGGGGTGCTTTGCCTAGTAATCAGCCCTGTTAGTTCTCCATTTAAATTTTATATA 60
      |||
Sbjct 1524 ATGGGGGTGCTTTGCCTAGTAATCAGCCCTGTTAGTTCTCCATTTAAATTTTATATA 1583

Query 61 TTGCCTCTAAGATTGAACCTTAGCTTCAGAAGACTAATTTATTTGGCATA&GT&CAGATAG 120
      |||
Sbjct 1584 TTGCCTCTAAGATTGAACCTTAGCTTCAGAAGACTAATTTATTTGGCATA&GT&CAGATAG 1643

Query 121 ACACACTTTATCTTATACGTTAGTTTCAAACCCATTTGATTGTTCTTGTGCTGCCGTTTT 180
      |||
Sbjct 1644 ACACACTTTATCTTATACGTTAGTTTCAAACCCATTTGATTGTTCTTGTGCTGCCGTTTT 1703

Query 181 ATAGTAAGCAA&ACTGACATTGCTTAGTATTGTC&CTGACATTTATGGAAGAGGCATTCCA 240
      |||
Sbjct 1704 ATAGTAAGCAA&ACTGACATTGCTTAGTATTGTC&CTGACATTTATGGAAGAGGCATTCCA 1763

Query 241 CTATTGTAGACTGCTTTGCTCTCGTATTGTATCACGTGCTATAA&tttttttAGACATCCAT 300
      |||
Sbjct 1764 CTATTGTAGACTGCTTTGCTCTCGTATTGTATCACGTGCTATAA&TTTTTTAGACATCCAT 1823

Query 301 GAAA&ACTTAGTTTGTAGTTTAA&CC&CAATGAAA&AGG&AGGC&ATGACCTAA&GAG&CC&CCT 360
      |||
Sbjct 1824 GAAA&ACTTAGTTTGTAGTTTAA&CC&CAATGAAA&AGG&AGGC&ATGACCTAA&GAG&CC&CCT 1883

Query 361 CCATTCTATCCAGACATGAAA&AGC&ACTGCTCAGATGGTTGTATG 404
      |||
Sbjct 1884 CCATTCTATCCAGACATGAAA&AGC&ACTGCTCAGATGGTTGTATG 1927
```

Fig 3.6.1.3.1: BLAST result for 233203_at probe sequence. No BLAST hit was obtained for reference sequence for ROPN1 and ROPN1B

Probe 231535_x_at: BLAST results indicate this probe to represent ROPN1 gene (Fig 3.6.1.3.2). No hit was obtained for ROPN1B reference sequence.

```
> gi|21359919|ref|NM\_017578.2| UEG Homo sapiens ropporin, rhophilin associated protein 1 (ROPN1), mRNA
gi|15082272|gb|AF303889.2|AF303889 UEG Homo sapiens ropporin mRNA, complete cds
Length=1112

GENE ID: 54763 ROPN1 | ropporin, rhophilin associated protein 1 [Homo sapiens]
(10 or fewer PubMed links)

Score = 708 bits (784), Expect = 0.0
Identities = 392/392 (100%), Gaps = 0/392 (0%)
Strand=Plus/Plus

Query 1 ATGTGGGTCGCTTCACGGAGGAGATCGAGTGGCTGAAGTTTTAGCCCTTGCTTGCAGCG 60
      |||
Sbjct 667 ATGTGGGTCGCTTCACGGAGGAGATCGAGTGGCTGAAGTTTTAGCCCTTGCTTGCAGCG 726

Query 61 CTCTGGGAGTTACTATTACCAAACTCTCAAGATAGTGTGTGAGGTCTTATCATGTGACC 120
      |||
Sbjct 727 CTCTGGGAGTTACTATTACCAAACTCTCAAGATAGTGTGTGAGGTCTTATCATGTGACC 786

Query 121 ATAATGGTGGGTCGCCCCGGATCCCGTTCAGCACCTTCCAGTTTCTCTACACGTATATTG 180
      |||
Sbjct 787 ATAATGGTGGGTCGCCCCGGATCCCGTTCAGCACCTTCCAGTTTCTCTACACGTATATTG 846

Query 181 CCAAAGTGGATGGGAGATCTCTGCATCACATGTCAGCAGGATGCTAAACTACATGGAAC 240
      |||
Sbjct 847 CCAAAGTGGATGGGAGATCTCTGCATCACATGTCAGCAGGATGCTAAACTACATGGAAC 906

Query 241 AGGAAAGTAATTGGCCCTGATGGTATAATCACAGTGAATGACTTTACCCAAAACCCAGGG 300
      |||
Sbjct 907 AGGAAAGTAATTGGCCCTGATGGTATAATCACAGTGAATGACTTTACCCAAAACCCAGGG 966

Query 301 TTCAGCTGGAGTAAAAGCACAATTTTGGCAATTTTAAAGGAAGATACAGAGATGATTGTA 360
      |||
Sbjct 967 TTCAGCTGGAGTAAAAGCACAATTTTGGCAATTTTAAAGGAAGATACAGAGATGATTGTA 1026

Query 361 CTTCAGAATGACTGAAACCCATATACCACCCA 392
      |||
Sbjct 1027 CTTCAGAATGACTGAAACCCATATACCACCCA 1058
```

Fig 3.6.1.3.2: BLAST result for 231535_x_at probe sequence. BLAST hit was obtained for reference sequence for ROPN1 and no hit was obtained for reference sequence for ROPN1B.

Probe 224191_x_at: BLAST results indicate this probe to represent both ROPN1B and ROPN1 gene. A 100% match was obtained for ROPN1B reference sequence (Fig 3.6.1.3.3) and 99% match was obtained for ROPN1 reference sequence (Fig 3.6.1.3.4).

```
> gi|59891408|ref|NM\_001012337.1| UEG Homo sapiens ropporin, rophilin associated protein 1B (ROPN1B), mRNA
   gi|12539612|gb|AF231410.1|AF231410 UEG Homo sapiens AKAP-binding sperm protein ropporin mRNA, complete cds
   Length=1018

   GENE ID: 152015 ROPN1B | ropporin, rophilin associated protein 1B
   [Homo sapiens] (10 or fewer PubMed links)

   Score = 742 bits (822), Expect = 0.0
   Identities = 411/411 (100%), Gaps = 0/411 (0%)
   Strand=Plus/Plus

   Query  1  GGGCAGAGCTAACACCTGAGCTGTTAAAGATCCTGCATTCTCAGGTTGCTGGCAGACTGA  60
           |||
   Sbjct  419 GGGCAGAGCTAACACCTGAGCTGTTAAAGATCCTGCATTCTCAGGTTGCTGGCAGACTGA  478

   Query  61  TCATCCGTGCAGAGGAGCTGGCCAGATGTGGAAGTGGTGAATCTCCCAACAGATCTGT  120
           |||
   Sbjct  479  TCATCCGTGCAGAGGAGCTGGCCAGATGTGGAAGTGGTGAATCTCCCAACAGATCTGT  538

   Query  121  TTAATAGTGTGATGAATGTGGGTCGCTTCACGGAGGAGATCGAGTGGCTGAAGTTTTAG  180
           |||
   Sbjct  539  TTAATAGTGTGATGAATGTGGGTCGCTTCACGGAGGAGATCGAGTGGCTGAAGTTTTAG  598

   Query  181  CCCTTGCTTGCAGCGCTCTGGGAGTTACTATTACAAAACCTCAAGATAGTGTGTGAGG  240
           |||
   Sbjct  599  CCCTTGCTTGCAGCGCTCTGGGAGTTACTATTACAAAACCTCAAGATAGTGTGTGAGG  658

   Query  241  TCTTATCATGTGACCACAATGGTGGGTTGCCCGAATCCCATTCAGCACCTTCCAGTTTC  300
           |||
   Sbjct  659  TCTTATCATGTGACCACAATGGTGGGTTGCCCGAATCCCATTCAGCACCTTCCAGTTTC  718

   Query  301  TCTACACGTATATTGCCGAAGTGGATGGGGAGATCTGTGCATCACATGTCAGCAGGATGC  360
           |||
   Sbjct  719  TCTACACGTATATTGCCGAAGTGGATGGGGAGATCTGTGCATCACATGTCAGCAGGATGC  778

   Query  361  TAAACTACATTGAACAGGAAGTAATTGGTCCTGATGGTTAATCACGGTGA  411
           |||
   Sbjct  779  TAAACTACATTGAACAGGAAGTAATTGGTCCTGATGGTTAATCACGGTGA  829
```

Fig 3.6.1.3.3: BLAST result for 224191_x_at probe sequence

```

>gi|21359919|ref|NM_017578.2| UEG Homo sapiens ropporin, rhophilin associated protein 1 (ROPN1), mRNA
gi|15082272|gb|AF303889.2|AF303889 UEG Homo sapiens ropporin mRNA, complete cds
Length=1112

GENE ID: 54763 ROPN1 | ropporin, rhophilin associated protein 1 [Homo sapiens]
(10 or fewer PubMed links)

Score = 944 bits (1045), Expect = 0.0
Identities = 530/532 (99%), Gaps = 2/532 (0*)
Strand=Plus/Plus

Query 1 AACC GGGCAGAGCTAACACCTGAGCTGTTAAAGATCCTGCATTCTCAGGTTG-TGGCAGA 59
      |
Sbjct 528 AACC GGGCAGAGCTAACACCTGAGCTGTTAAAGATCCTGCATTCTCAGGTTGCTGGCAGA 587

Query 60 CTTGATCATCCGTGCAGAGGAGCTGGCCAGATGTGGAAAGTGGTGAATCTCCCAACAGA 119
      |
Sbjct 588 C-TGATCATCCGTGCAGAGGAGCTGGCCAGATGTGGAAAGTGGTGAATCTCCCAACAGA 646

Query 120 TCTGTTTAAATAGTGTGATGAATGTGGGTCGCTTCACGGAGGAGATCGAGTGGCTGAAAGTT 179
      |
Sbjct 647 TCTGTTTAAATAGTGTGATGAATGTGGGTCGCTTCACGGAGGAGATCGAGTGGCTGAAAGTT 706

Query 180 TTTAGCCCTTGCTTGCAGCGCTCTGGGAGTTACTATTACCAAAACTCTCAAGATAGTGTG 239
      |
Sbjct 707 TTTAGCCCTTGCTTGCAGCGCTCTGGGAGTTACTATTACCAAAACTCTCAAGATAGTGTG 766

Query 240 TGAGGTCCTTATCATGTGACCATAATGGTGGGTCGCCCGGATCCCCTTCAGCACCTTCCA 299
      |
Sbjct 767 TGAGGTCCTTATCATGTGACCATAATGGTGGGTCGCCCGGATCCCCTTCAGCACCTTCCA 826

Query 300 GTTCTCTACACGTATATTGCCAAAAGTGGATGGGGAGATCTCTGCATCACATGTCAGCAG 359
      |
Sbjct 827 GTTCTCTACACGTATATTGCCAAAAGTGGATGGGGAGATCTCTGCATCACATGTCAGCAG 886

Query 360 GATGCTAAACTACATGGAAACAGGAAGTAAATTGGCCCTGATGGTATAATCACAGTGAATGA 419
      |
Sbjct 887 GATGCTAAACTACATGGAAACAGGAAGTAAATTGGCCCTGATGGTATAATCACAGTGAATGA 946

Query 420 CTTTACCCAAAACCCAGGGTTCACTGGAGTAAAAGCACAAATTTGGCAATTTTAAAGG 479
      |
Sbjct 947 CTTTACCCAAAACCCAGGGTTCACTGGAGTAAAAGCACAAATTTGGCAATTTTAAAGG 1006

Query 480 AAGATACAGAGATGATTGTACTTCAGAAAGACTGAAACCCATATACCACCCA 531
      |
Sbjct 1007 AAGATACAGAGATGATTGTACTTCAGAAAGACTGAAACCCATATACCACCCA 1058

```

Fig 3.6.1.3.4: BLAST result for 224191_x_at probe sequence

Probe 220425_x_at: BLAST results indicate this probe to represent ROPN1B. A 100% match was obtained for ROPN1B reference sequence (Fig 3.6.1.3.5) and 97% match was obtained for ROPN1 reference sequence (Fig 3.6.1.3.6).

```

> gi|59891408|ref|NM\_001012337.1| UEG Homo sapiens ropporin, rhophilin associated protein 1B (ROPN1B), mRNA
   gi|12539612|gb|AF231410.1|AF231410 UEG Homo sapiens AKAP-binding sperm protein ropporin mRNA, complete cds
Length=1018

GENE ID: 152015 ROPN1B | ropporin, rhophilin associated protein 1B
[Homo sapiens] (10 or fewer PubMed links)

Score = 742 bits (822), Expect = 0.0
Identities = 411/411 (100%), Gaps = 0/411 (0%)
Strand=Plus/Plus

Query 1   GGGCAGAGCTAACACCTGAGCTGTTAAAGATCCTGCATTCTCAGGTTGCTGGCAGACTGA 60
          |||
Sbjct 419  GGGCAGAGCTAACACCTGAGCTGTTAAAGATCCTGCATTCTCAGGTTGCTGGCAGACTGA 478

Query 61  TCATCCGTGCAGAGGAGCTGGCCCAGATGTGGAAAGTGGTGAATCTCCCAACAGATCTGT 120
          |||
Sbjct 479  TCATCCGTGCAGAGGAGCTGGCCCAGATGTGGAAAGTGGTGAATCTCCCAACAGATCTGT 538

Query 121 TTAATAGTGTGATGAATGTGGGTGCGCTTCACGGAGGAGATCGAGTGGCTGAAGTTTTTAG 180
          |||
Sbjct 539  TTAATAGTGTGATGAATGTGGGTGCGCTTCACGGAGGAGATCGAGTGGCTGAAGTTTTTAG 598

Query 181 CCCTTGCTTGCAGCGCTCTGGGAGTTACTATTACAAAACCTCAAGATAGTGTGTGAGG 240
          |||
Sbjct 599  CCCTTGCTTGCAGCGCTCTGGGAGTTACTATTACAAAACCTCAAGATAGTGTGTGAGG 658

Query 241 TCTTATCATGTGACCACAATGGTGGGTTGCCCGAATCCCATTCAGCACCTTCCAGTTTC 300
          |||
Sbjct 659  TCTTATCATGTGACCACAATGGTGGGTTGCCCGAATCCCATTCAGCACCTTCCAGTTTC 718

Query 301 TCTACACGTATATTGCCGAAGTGGATGGGGAGATCTGTGCATCACATGTCAGCAGGATGC 360
          |||
Sbjct 719  TCTACACGTATATTGCCGAAGTGGATGGGGAGATCTGTGCATCACATGTCAGCAGGATGC 778

Query 361 TAAACTACATTGAACAGGAAGTAATTGGTCCTGATGGTTTAAATCACGGTGA 411
          |||
Sbjct 779  TAAACTACATTGAACAGGAAGTAATTGGTCCTGATGGTTTAAATCACGGTGA 829

```

Fig 3.6.1.3.5: BLAST result for 220425_x_at probe sequence

```

> gi|21359919|ref|NM\_017578.2| UEG Homo sapiens ropporin, rhophilin associated protein 1 (ROPN1), mRNA
gi|15082272|gb|AF303889.2|AF303889 UEG Homo sapiens ropporin mRNA, complete cds
Length=1112

GENE ID: 54763 ROPN1 | ropporin, rhophilin associated protein 1 [Homo sapiens]
(10 or fewer PubMed links)

Score = 697 bits (772), Expect = 0.0
Identities = 401/411 (97%), Gaps = 0/411 (0%)
Strand=Plus/Plus

Query 1 GGGCAGAGCTAACACCTGAGCTGTTAAAGATCCTGCATTCTCAGGTTGCTGGCAGACTGA 60
      |||
Sbjct 532 GGGCAGAGCTAACACCTGAGCTGTTAAAGATCCTGCATTCTCAGGTTGCTGGCAGACTGA 591

Query 61 TCATCCGTGCAGAGGAGCTGGCCAGATGTGGAAAAGTGGTGAATCTCCCAACAGATCTGT 120
      |||
Sbjct 592 TCATCCGTGCAGAGGAGCTGGCCAGATGTGGAAAAGTGGTGAATCTCCCAACAGATCTGT 651

Query 121 TTAATAGTGTGATGAATGTGGGTCGCTTCACGGAGGAGATCGAGTGGCTGAAGTTTTTAG 180
      |||
Sbjct 652 TTAATAGTGTGATGAATGTGGGTCGCTTCACGGAGGAGATCGAGTGGCTGAAGTTTTTAG 711

Query 181 CCCTTGCTTGCAGCGCTCTGGGAGTTACTATTACCAAACTCTCAAGATAGTGTGTGAGG 240
      |||
Sbjct 712 CCCTTGCTTGCAGCGCTCTGGGAGTTACTATTACCAAACTCTCAAGATAGTGTGTGAGG 771

Query 241 TCTTATCATGTGACCACAATGGTGGGTTGCCCGAATCCCATTCAGCACCTTCCAGTTTC 300
      |||
Sbjct 772 TCTTATCATGTGACCATAATGGTGGGTTGCCCGGATCCCGTTCAGCACCTTCCAGTTTC 831

Query 301 TCTACACGTATATTGCCGAAGTGGATGGGGAGATCTGTGCATCACATGTCAGCAGGATGC 360
      |||
Sbjct 832 TCTACACGTATATTGCCAAAGTGGATGGGGAGATCTGTGCATCACATGTCAGCAGGATGC 891

Query 361 TAAACTACATTGAACAGGAAGTAATTGGTCTGATGGTTTAATCACGGTGA 411
      |||
Sbjct 892 TAAACTACATGGAACAGGAAGTAATTGGCCCTGATGGTATAATCACAGTGA 942

```

Fig 3.6.1.3.6: BLAST result for 220425_x_at probe sequence

Therefore probe set 231535_x_at was taken to represent ROPN1 and probe set 220425_x_at was taken to represent ROPN1B. 233203_at was no longer considered to represent Ropporin.

3.6.2 Expression of ROPN1 and ROPN1B in normal and cancerous breast tissue

3.6.2.1 Expression of ROPN1 and ROPN1B in our in-house study

Ropporin (ROPN1 and ROPN1B) expression was examined in our in-house study using both microarray and qRT-PCR.

3.6.2.1.1 Results from microarray

On average, ROPN1 was 4.97-fold up-regulated in patients who relapsed and ROPN1B was 5.06-fold up-regulated in patients who relapsed compared to patients who did not relapse (overall). ROPN1 was 6.81-fold up-regulated in patients who relapsed within 5 years and ROPN1B was 7.83-fold up-regulated in patients who relapsed within 5 years compared to those who remained disease-free for 5 years.

High expression of ROPN1 and ROPN1B was observed in one of the sub-cluster enriched with ER-negative specimens. This cluster had the worst survival in comparison to other clusters (see section 3.1.3).

ROPN1B (220425_x_at) was not expressed among 54/57 (94.7%) of the patients who did not relapse, however it was expressed in 13/48 (27.1%) of the patients who did relapse based on the cut-off of 100 Affymetrix unit. Similarly ROPN1 (231535_x_at) was not expressed among 53/57 (92.9%) of the patients who did not relapse, however it was expressed in 14/48 (29.1%) of the patients who did relapse.

3.6.2.1.2 Survival analysis

Survival analysis (see section 2.2.12) was performed on ROPN1B and ROPN1 using Present/Absent Affymetrix calls. ROPN1B expression has significantly correlated with relapse-free survival (p-value =0.0340) but not with overall survival (p-value =0.3894) (Fig 3.6.2.1.2.1). Absence of ROPN1B gene expression is positively linked to relapse-free survival; however no relation was observed for overall survival. However, ROPN1 expression did not significantly correlate with relapse free survival (p-value =0.122) or with overall survival (p-value =0.37) (Fig 3.6.2.1.2.2). This analysis was performed by Dr. Lorraine O'Driscoll.

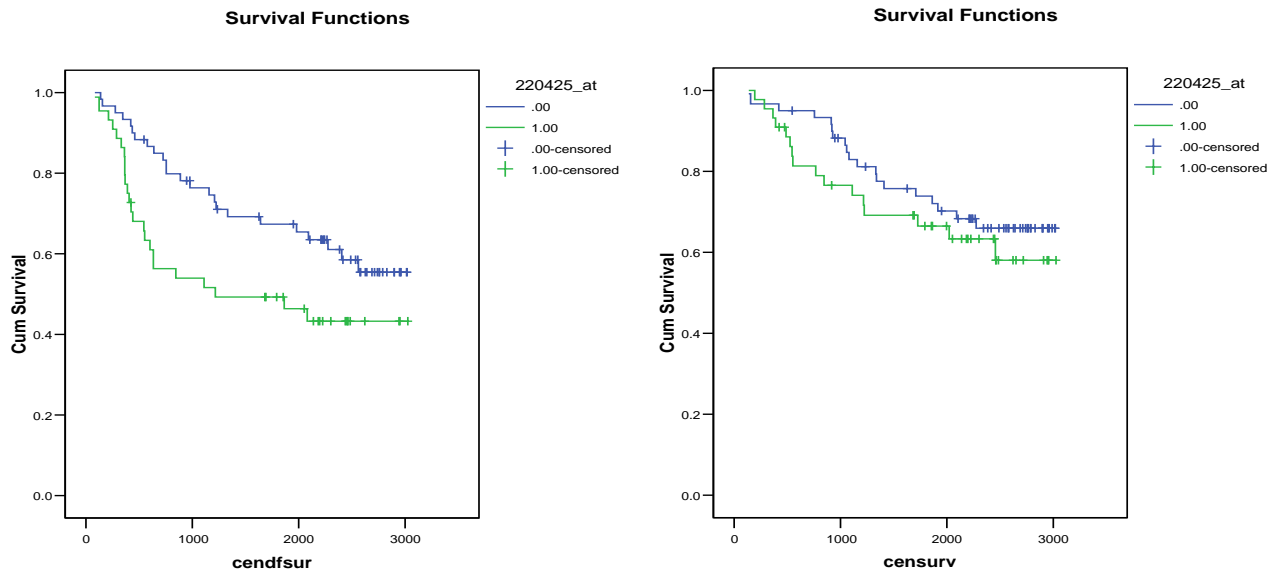


Fig 3.6.2.1.2.1: The figure on the left denotes survival curve for Relapse-free Survival for ROPN1B; the figure on the right denotes the survival curve for Overall Survival for ROPN1B. X-axis denotes the survival in days; Y-axis denotes the percent of patients still surviving.

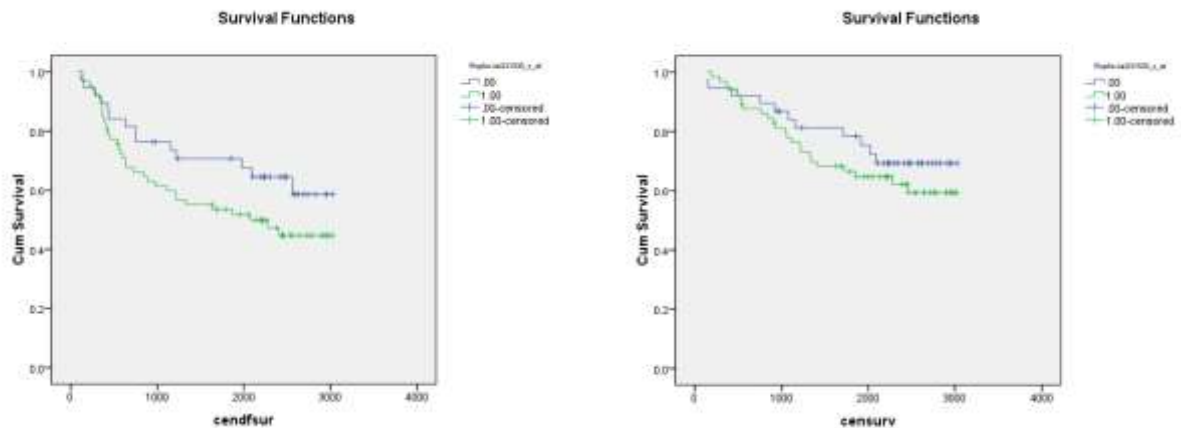


Fig 3.6.2.1.2.2: The figure on the left denotes survival curve for Relapse-free Survival for ROPN1; the figure on the right denotes the survival curve for Overall Survival for ROPN1. X-axis denotes the survival in days; Y-axis denotes the percent of patients still surviving.

3.6.2.1.3 Results from qRT-PCR

In order to confirm the microarray findings, qRT-PCR was performed on only 94 of the 104 clinical specimens from our in-house study, due to the lack of available RNA for the remaining clinical specimens. ROPN1 was found to be 2.33-fold down-regulated (baseline mean 2.39, SD 6.4; experimental mean 1.81, SD 6.4) in patients who relapsed, whereas ROPN1B was found to be 6.28-fold up-regulated (baseline mean 1.02, SD 2.4; experimental mean 11.37, SD 57.1) in patients who relapse.

3.6.2.2 Expression in normal tissue

The GSE1133 dataset (see section 2.1) was used to find the expression levels of ROPN1B in various tissues. Since the dataset is obtained from the U133A chip, information about ROPN1 was unavailable. This dataset includes gene expression data on 79 different human tissues (data shown for the top 20 organs in descending order of expression of ROPN1B represented by 220425_x_at) thereby providing ample opportunity to study distribution of ROPN1B gene expression in different human organs (see section 2.1.2). The results indicate that ROPN1B is very highly expressed in testis followed by ganglion and marginally in skin, medulla, trachea, heart and liver (Table 3.6.2.2.1).

Tissue	Expression of 220425_x_at (ROPN1B) Affymetrix units
Testis interstitial	5656.1
Testis	4401
Testis leydig cell	3411.75
Testis germ cell	3011.7
Testis seminiferous tubule	2962.4
Superior cervical ganglion	1547.65
Trigeminal ganglion	966.3
Ciliary ganglion	484.7
Skin	358.9
Medulla oblongata	310.6
Atrioventricular node	306.5
Dorsal root ganglion	283.9
Heart	263
Liver	257.9
Adrenal cortex	251.4
Prostate	245.9
Trachea	243.2
Appendix	237.6
Cingulate cortex	226.3
Cerebellum peduncles	201.4

Table 3.6.2.2.1: Expression levels of ROPN1B (220425_x_at) in various tissues.

3.6.2.3 Expression in cancer cell lines

The GSE5720 dataset (see section 2.1) was used to find the expression levels in various cancer cell lines. This dataset constitutes gene expression values of 60 cell lines of different origins, thereby giving ample opportunity to find cell lines which express gene of our interest, so that functional validation can be performed. Expression of ROPN1 and ROPN1B was observed in melanoma cell lines UACC-257, SK-MEL-28, MALME-3M, MDA-MB-435S (breast/melanoma), MDA-N, UACC-62, SK-MEL2, SK-MEL-5 and M14 (Table 3.6.2.3.1).

Cell Line	231535_x_at ROPN1	220425_x_at ROPN1B	224191_x_at ROPN1+ROPN1B	233203_at
UACC-257	376.5	910.4	392.1	21.4
SK-MEL-28	356.8	450	307.7	37.6
MALME-3M	176.9	241.8	213	22.4
UACC-62	170.8	171.4	132.3	17.7
MDA-N	112.9	216.5	150.5	9.6
SK-MEL-2	174.4	154	207.7	26.8
MDA-MB-435S	99.7	225.7	84.4	10.8
SK-MEL-5	164.5	149.2	130.3	32
M14	90.8	85.5	86.4	11.5
Hs578T	122.8	31.1	103.6	118.6

Table 3.6.2.3.1: Expression levels of ROPN1 and ROPN1B in various cancer cell lines.

3.6.2.4 Expression in melanoma and melanocyte cell lines

The GSE4570 dataset (see section 2.1) was used to estimate the expression level of Ropporin in normal melanocyte and metastatic melanoma cell lines. The dataset constitutes of 6 metastatic melanoma cell lines and 2 normal melanocyte cell line. The aim was to study the expression levels of Ropporin in normal melanocyte and metastatic melanoma. A low expression of ROPN1B was observed in normal melanocyte (NM1 and NM2) and high expression was observed in metastatic melanoma cell lines (WW165, YUCAL, YUHEIK, YUMAC, MNT1, YUSIT1) (Fig 3.6.2.4). Since the dataset was on U133A chip no information regarding the expression level of ROPN1 was available.

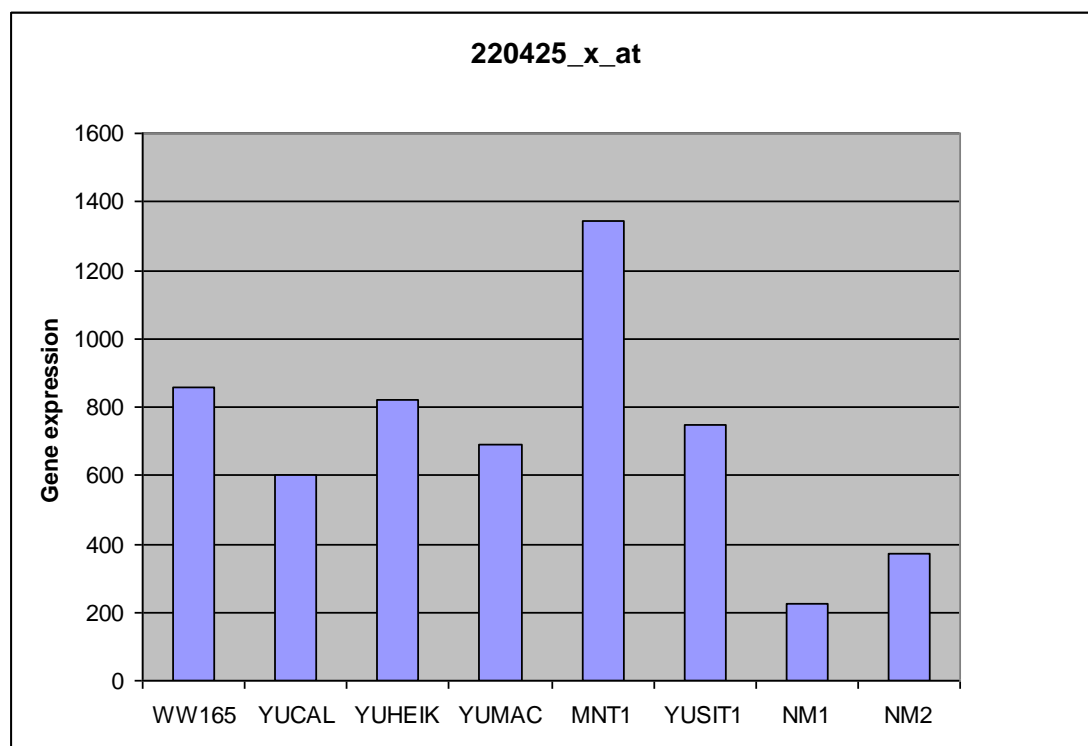


Fig 3.6.2.4: Expression level of ROPN1B (220425_x_at) in melanocyte and metastatic melanoma cell lines (Affymetrix unit). WW165, YUCAL, YUHEIK, YUMAC, MNT1 and YUSIT1 are metastatic melanoma cell lines, whereas NM1 and NM2 are melanocyte cell lines.

3.6.2.5 Melanoma clinical specimens

The GSE4587 dataset (see section 2.1) was used to estimate the expression changes of Ropporin during the different stages of melanoma progression. The results show that ROPN1 and ROPN1B expression progressively increase with disease progression and is highest in metastatic growth phase melanoma and lymph node metastasis (Fig 3.6.2.5). The results positively associates Ropporin gene to disease progression and metastasis in clinical specimens.

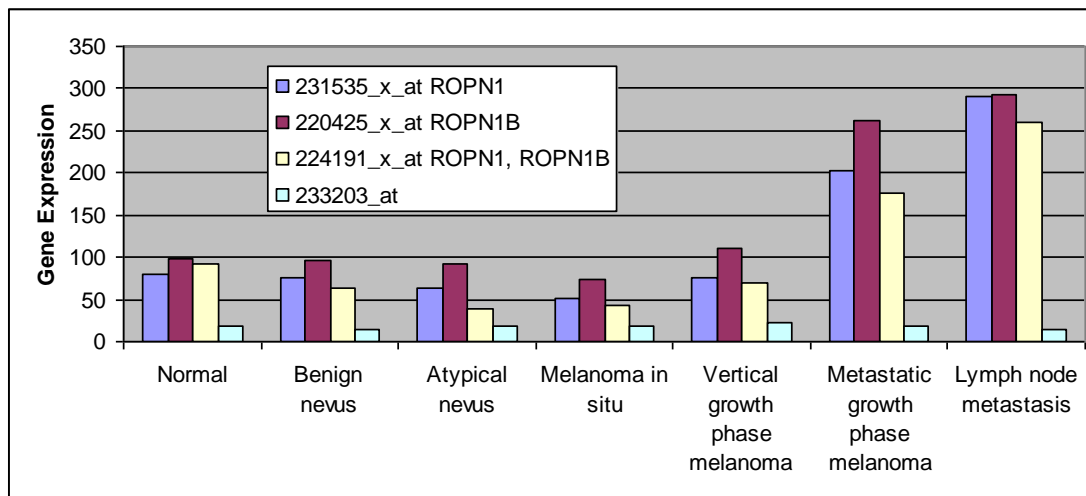


Fig 3.6.2.5: ROPN1 and ROPN1B expression in melanoma progression. The number of samples in each group are (Normal: 2; Benign nevus: 2 ; Atypical nevus: 2 ; Melanoma *in situ*: 2 ; Vertical growth phase melanoma: 2; Metastatic growth phase melanoma: 2; Lymph node metastasis: 3.

3.6.2.6 Multiple Myeloma

The GSE4581 dataset (see section 2.1) was used to estimate the expression of ROPN1 and ROPN1B in multiple myeloma. The aim was to study the expression pattern of Ropporin in multiple myeloma. This result shows that both ROPN1 and ROPN1B are expressed in most of the multiple myeloma patients (Fig 3.6.2.6). Additionally, probe set 233203_at also expresses in many of the specimens, indicating the possibility of more isoforms of this gene.

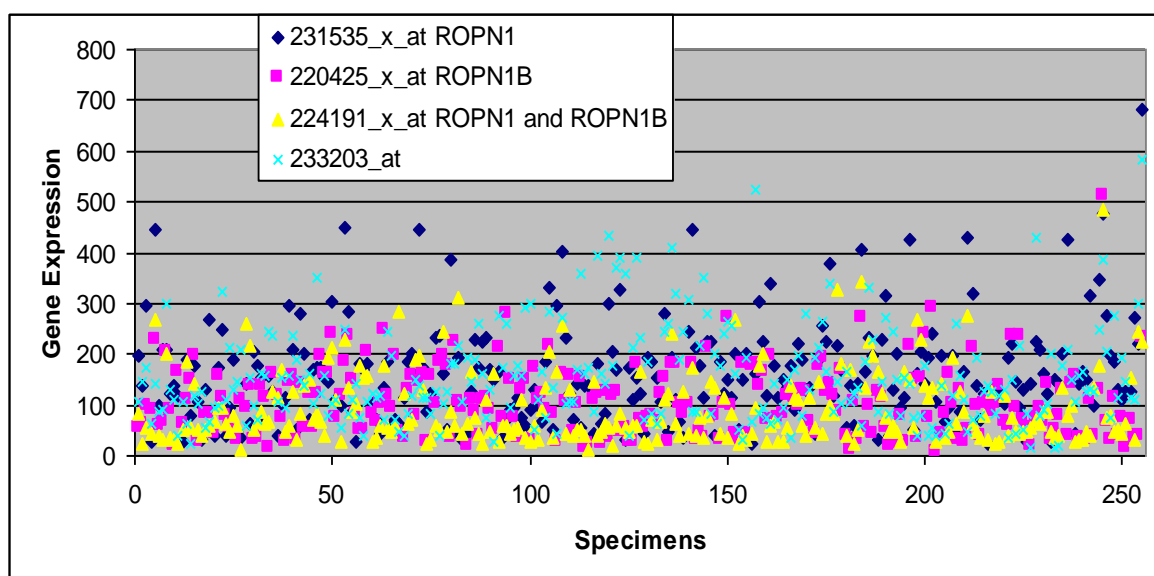


Fig 3.6.2.6: ROPN1 and ROPN1B expression in myeloma clinical specimens

3.6.2.7 qRT-PCR assessment of ROPN1B and ROPN1 expression in various melanoma cell lines

The expression of ROPN1B and ROPN1 was not found in any of the breast cancer cell lines (based on microarray and qRT-PCR), except for MDA-MB-435s. At the time, MDA-MB-435S was considered a breast cancer cell line. But recently the cell line has become controversial as far as its origin is concerned. Recent studies have reported that this cell line might be a melanoma cell line, rather than a breast cancer cell line (see section 2.4.2). In order to identify a suitable cell line model for functional validation, ROPN1 and ROPN1B expression was examined by qRT-PCR in various cell lines (Breast: HCC1954, HCC1419, HCC1937, BT-20, MCF-7, MDA-MB-231, MDA-MB-468 & SKBR3; Melanoma: MDA-MB-435S; HT144, SK-MEL-28, MEL-5, M14). ROPN1 and ROPN1B expression was detected in the

breast/melanoma cell line MDA-MB-435S and all the melanoma cell lines tested. A high amount of expression of ROPN1B relative to ROPN1 was observed in MDA-MB-435s and SK-MEL-28. ROPN1 was detected in all of the cell lines with expression of ROPN1B (Fig 3.6.2.7).

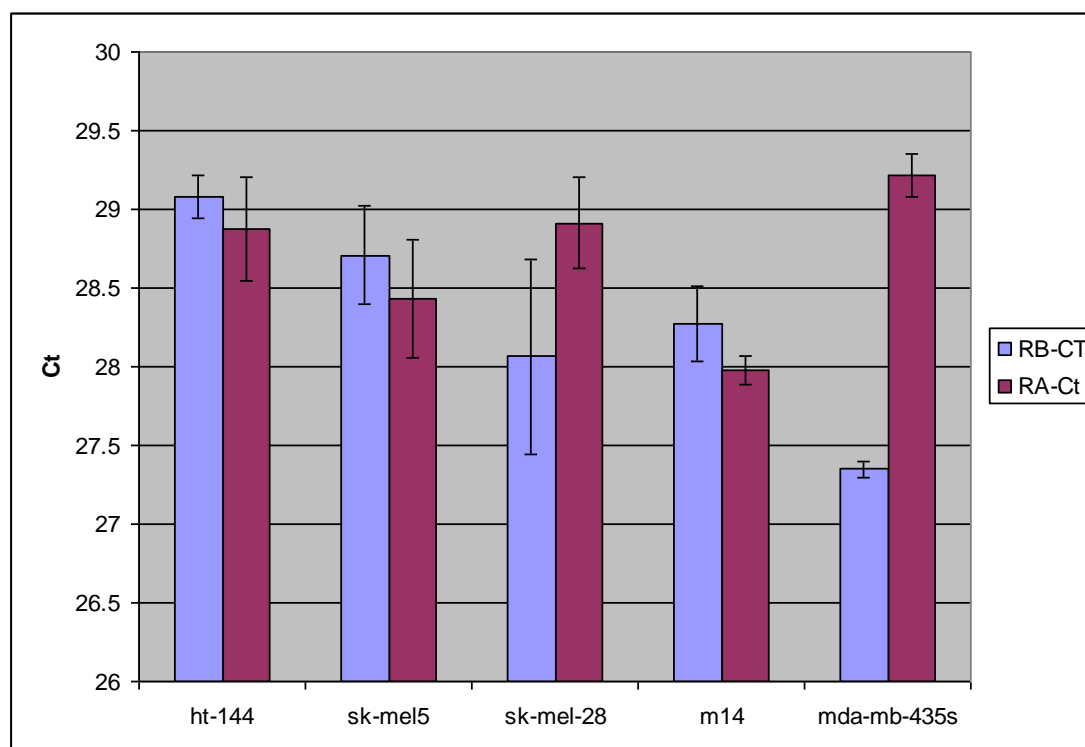


Fig 3.6.2.7: ROPN1B and ROPN1 expression in various melanoma cells. Ct is the Cycle Threshold. A lower Ct indicates higher expression. A high expression of ROPN1B was observed in MDA-MB-435s and SK-MEL-28. Nearly equal amounts of ROPN1 and ROPN1B expression were observed in M14, HT-144 and SK-Mel-5. The error represents the Standard Deviation observed among three technical replicates.

3.6.3 siRNA knockdown of ROPN1 and ROPN1B in melanoma cell lines

Since Ropporin plays an important role in sperm motility, our study aimed to investigate the possible role of Ropporin in cancer cell motility and invasion, which is a prime requirement for disease progression in melanoma, multiple myeloma and breast cancer. siRNAs targeting ROPN1 and ROPN1B were obtained from Ambion (see section 2.5.7).

3.6.3.1 ROPN1B siRNA on MDA-MB-435S

3.6.3.1.1 Gene Expression Knockdown Analysis

MDA-MB-435s cells were transfected with three different ROPN1B siRNA and qRT-PCR (see section 2.5.4) was performed after 72hrs. qRT-PCR analysis confirmed the knockdown of ROPN1B using ROPN1B-1 siRNA (75.3%), ROPN1B-2 siRNA (69.6%) and ROPN1B-3 siRNA (56.6%) compared to scrambled transfected cells (Fig 3.6.3.1.1).

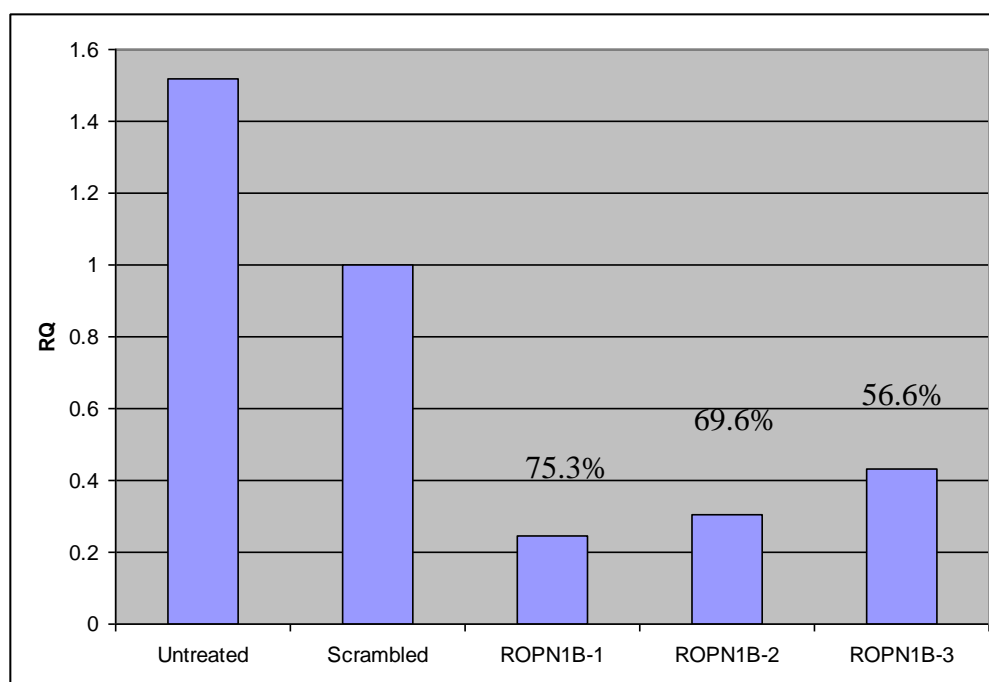


Fig 3.6.3.1.1.1: ROPN1B qRT-PCR on ROPN1B-siRNA transfected MDA-MB-435s cells. Knockdown was observed in siRNA knockdown cells. ROPN1B-1 siRNA (75.3%), ROPN1B-2 siRNA (69.6%) and ROPN1B-3 siRNA (56.6%) compared to scrambled transfected cells. RQ is relative quantification with reference to scrambled.

Expression of ROPN1 mRNA was also checked in ROPN1B-siRNA-transfected MDA-MB-435s cells using qRT-PCR. ROPN1B-2 siRNA showed 19.8% and ROPN1B-3 siRNA showed 53.7% knockdown of ROPN1 in siRNA transfected cells compared to scrambled transfected cells (Fig 3.6.3.1.2). A surge in the expression of ROPN1 mRNA (141.4%) was observed in ROPN1B-1-siRNA-transfected cells compared to scrambled transfected cells (Fig 3.6.3.1.2).

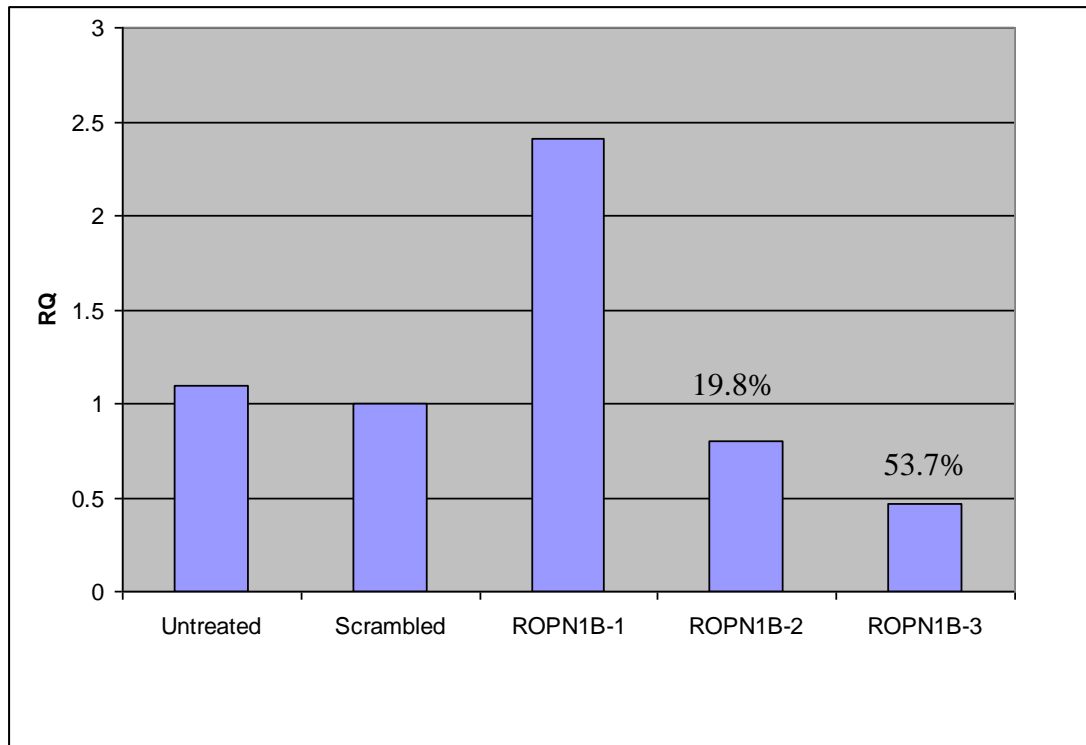


Fig 3.6.3.1.1.2: ROPN1 qRT-PCR on ROPN1B-siRNA transfected MDA-MB-435s cells. ROPN1B-1 siRNA showed 141.4% increased expression of ROPN1 whereas ROPN1B-2 siRNA showed 19.8% and ROPN1B-3 siRNA showed 53.7% knockdown of ROPN1 compared to scrambled transfected cells.

3.6.3.1.2 Western Blot Analysis

A western blot was performed to check protein expression of Ropporin in three different ROPN1B siRNA transfected MDA-MB-435s cells (Fig 3.6.3.1.3). As can be seen, ROPN1B-3 transfection resulted in the highest amount of Ropporin knockdown.

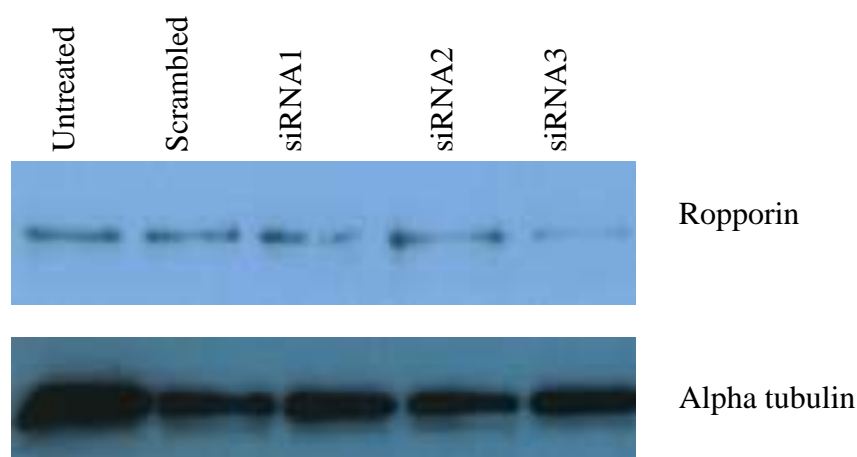


Fig 3.6.3.1.2: Ropporin western blot on ROPN1B siRNA transfected MDA-MB-435S cells. Knockdown was observed following use of ROPN1B-siRNA3 and marginal knockdown was observed in ROPN1B-siRNA1- and ROPN1B-siRNA2-transfected cells.

3.6.3.1.3 Motility Assay

MDA-MB-435s cells were transfected with three different siRNAs specific to ROPN1B and its effect on motility was observed (see section 2.5.10). The assay was performed in triplicates. There was significant loss of motility (ROPN1B-1 p-value =0.02; ROPN1B-2 p-value =0.02; ROPN1B-3 p-value =0.006) in the ROPN1B-siRNA transfected cells compared to scrambled transfected cells (Fig 3.6.3.1.3). ROPN1B-1 siRNA showed 50.8%, ROPN1B-2 siRNA showed 55.4% and ROPN1B-3 siRNA showed 60.1% reduced motility compared to scrambled transfected cells. This study indicates that the ROPN1B gene plays a role in the motility of this cancer cell line.

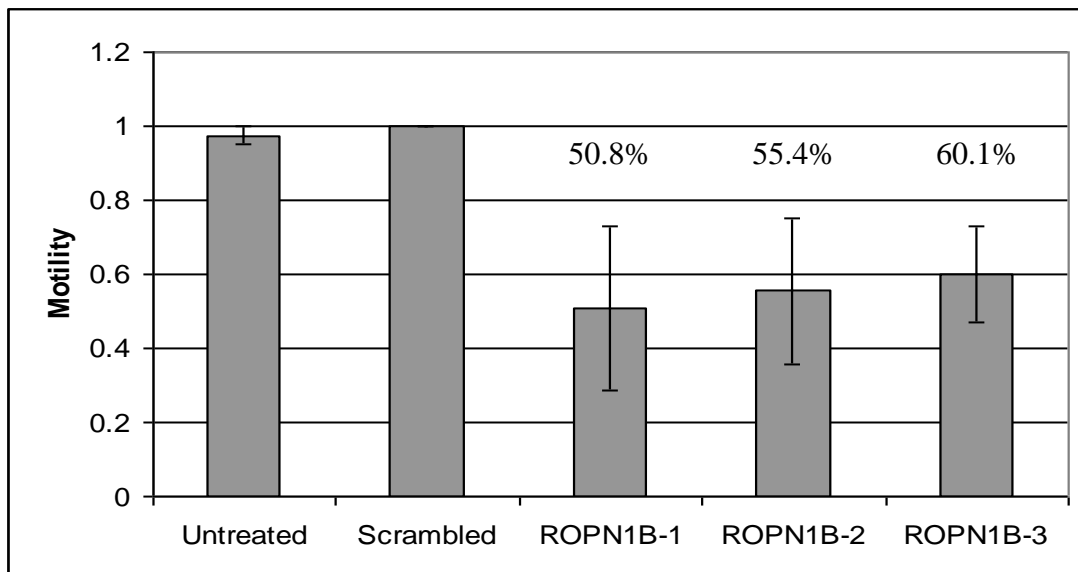
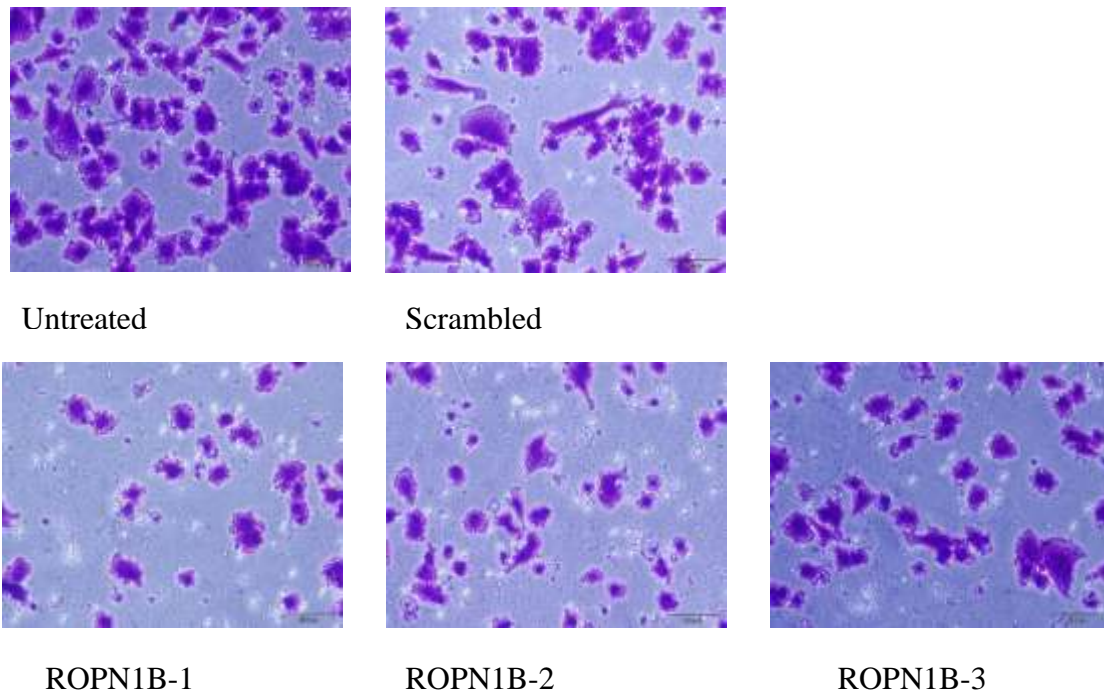


Fig 3.6.3.1.3: Motility assay on siRNA transfected MDA-MB-435S cells. All assays were performed in triplicate. There was significant reduction of motility in siRNA knockdown cells (ROPN1B-1 p-value =0.02; ROPN1B-2 p-value = 0.02; ROPN1B-3 p-value = 0.006). Y-axis denotes relative motility compared to scrambled. The error bar represents the standard deviation among three experimental repeats.

ROPN1 siRNA was not examined in MDA-MB-435s because of the low expression of ROPN1 in MDA-MB-435s. Invasion assay generated highly variable (unreproducible) results and was therefore removed from analysis. Often the cell line turns out to be non-invasive.

3.6.3.2 ROPN1B and ROPN1 siRNA in M14

The M14 melanoma cell line expressed both ROPN1 and ROPN1B in nearly equal amounts (Fig 3.6.2.7.1). siRNA knockdown was performed for ROPN1 using two ROPN1-specific siRNAs (ROPN1-1 and ROPN1-2) and ROPN1B using two ROPN1B-specific siRNAs (ROPNB-1 and ROPNB-2).

3.6.3.2.1 Gene Expression Knockdown Analysis

qRT-PCR analysis on the siRNA-transfected M14 cells confirmed the knockdown of ROPN1B (Fig 3.6.3.2.1.1) mRNA compared to scrambled transfected cells. ROPN1B-1 siRNA showed 62.4% and ROPN1B-2 siRNA showed 78.4% knockdown compared to scrambled transfected cells. ROPN1-1 siRNA showed 4% increased expression and ROPN1-2 siRNA showed 19.9% knockdown of ROPN1B in M14 transfected cells compared to scrambled transfected cells (Fig 3.6.3.2.1.2).

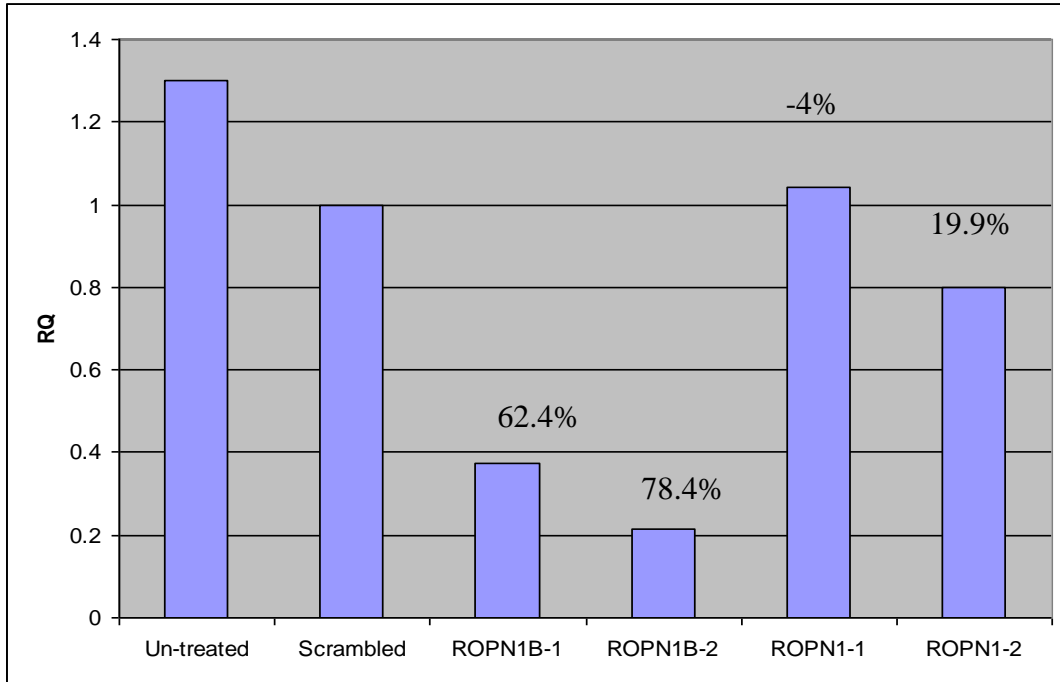


Fig 3.6.3.2.1.1: ROPN1B qRT-PCR on ROPN1B-siRNA- and ROPN1-siRNA-transfected M14 cells. The percent knockdown of each siRNA is indicated in the bar chart. 62.4% knockdown was observed for ROPN1B-1 siRNA and 78.4% knockdown was observed for the ROPN1B-2 siRNA. ROPN1-1 siRNA showed 4% increase in expression (probably a noise), whereas ROPN1-2 showed 19% decrease in expression

Expression of ROPN1 mRNA was also checked using qRT-PCR in the ROPN1B-siRNA- and ROPN1-siRNA-transfected M14 cells, 72hrs after transfection. qRT-PCR analysis confirmed the knockdown of ROPN1 mRNA (Fig 3.6.3.2.2) in siRNA transfected M14 cells compared to scrambled transfected cells. ROPN1B-1 siRNA showed 40.2%, ROPN1B-2 siRNA showed 58.5%, ROPN1-1 siRNA showed 42.7% and ROPN1-2 siRNA showed 45.5% knockdown compared to scrambled transfected cells (Fig 3.6.3.2.1.2).

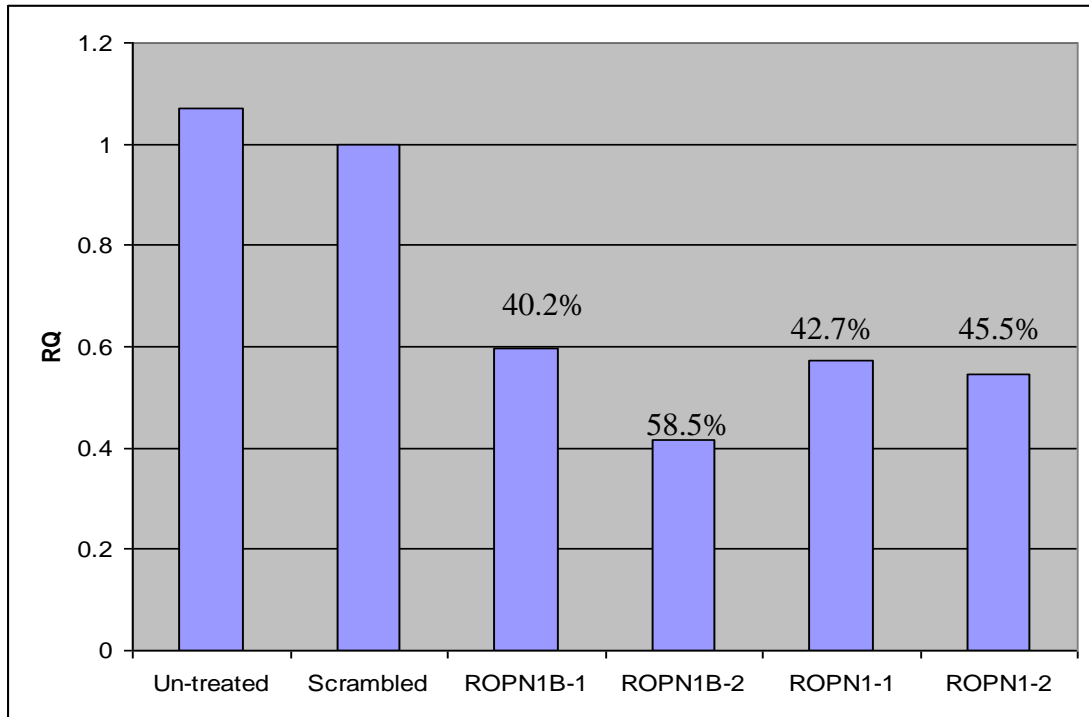


Fig 3.6.3.2.1.2: ROPN1 qRT-PCR on ROPN1B-siRNA- and ROPN1-siRNA-transfected M14 cells. Knockdown of ROPN1 by both ROPN1-siRNA and ROPN1B-siRNA was observed.

3.6.3.2.2 Western Blot Analysis

Western blot analysis was performed to analyse the expression of Ropporin protein in M14 cells 72hrs after transfection with ROPN1-siRNA and ROPN1B-siRNA. Western blot also showed knockdown of Ropporin in siRNA transfected cells (Fig 3.6.3.2.2) compared to scrambled transfected cells. Alpha tubulin was used as a loading control for the samples. As can be seen, ROPN1-siRNA transfection resulted in the highest amount of Ropporin knockdown.

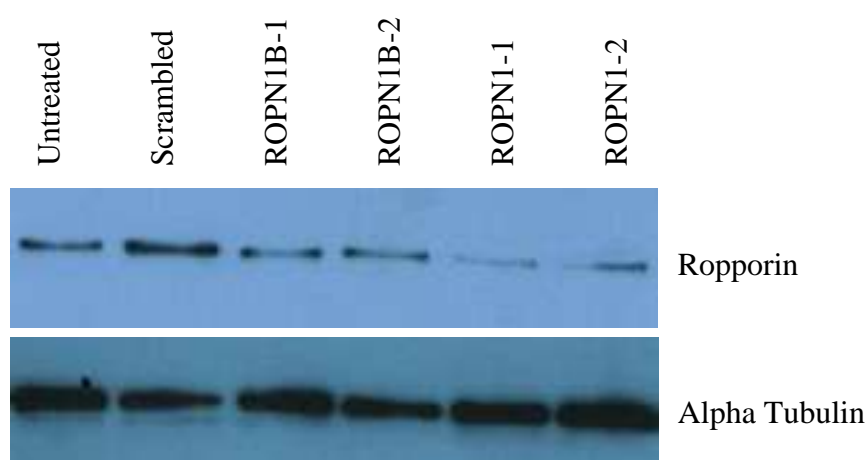


Fig 3.6.3.2.2: Ropporin western blot on siRNA transfected M14 cells. Good knockdown of Ropporin was observed after ROPN1 siRNA and marginal knockdown after ROPN1B siRNA transfection.

3.6.3.2.3 Motility Assay

ROPN1- and ROPN1B-siRNA knockdown was carried out to determine the effect of knockdown on motility in M14. As can be seen in Fig 3.6.3.2.4, there was significant loss of motility (ROPN1B-1 p-value =0.0002; ROPN1B-2 p-value =0.002; ROPN1-1 p-value =0.003; ROPN1-2 p-value =0.001) observed in ROPN1-siRNA-transfected and ROPN1B-siRNA transfected cells compared to scrambled-transfected cells. Transfection with ROPN1B-1 and ROPN1B-2 siRNA demonstrated a 31.2% and 33.2% reduction in motility, respectively, while transfection with ROPN1-1 and ROPN1-2 siRNA showed a 17.1% and 37.0% reduction in motility, respectively, compared to scrambled transfected cells (Fig 3.6.3.2.3).

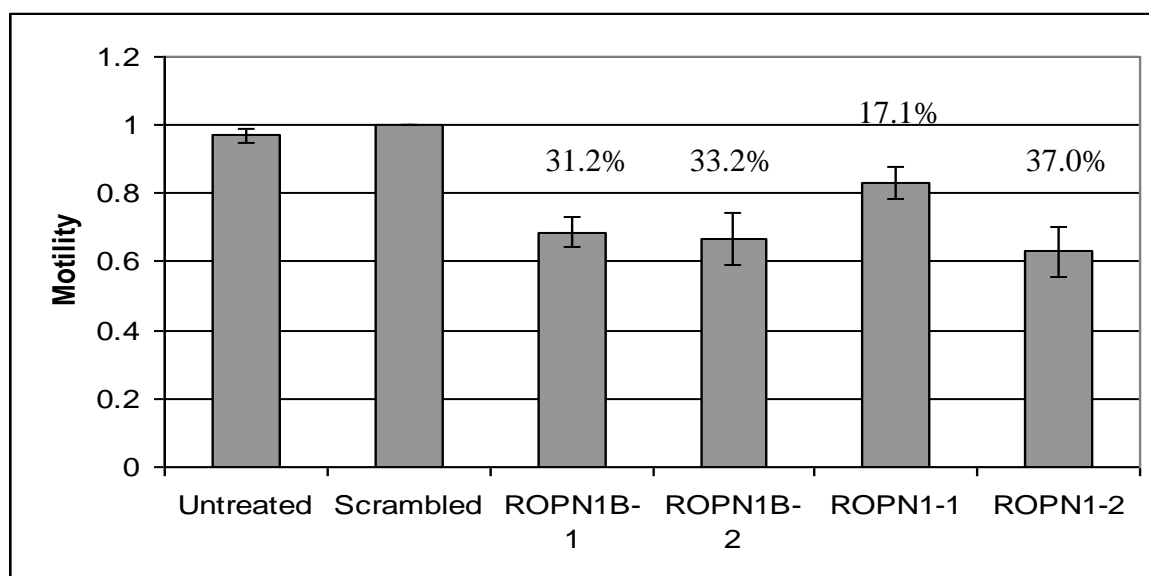
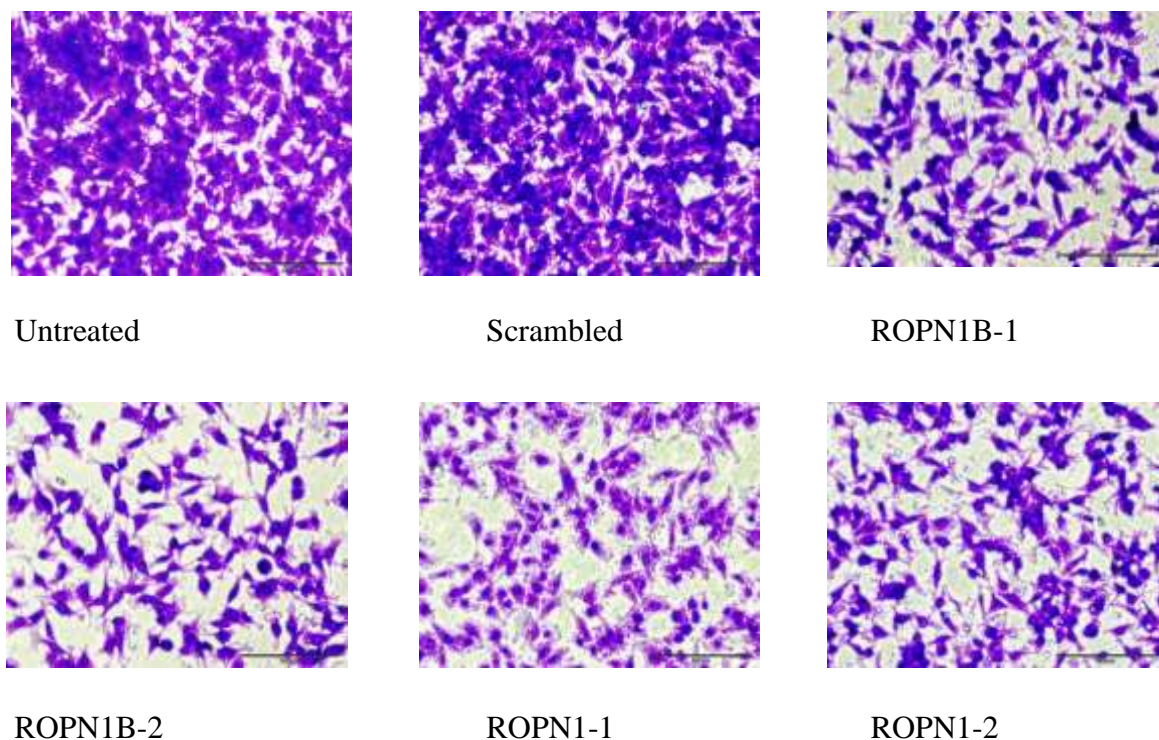


Fig 3.6.3.2.3: Motility assay on siRNA transfected M14 cells. The assay was performed in triplicate. Y-axis defines the relative motility compared to scrambled. Significant reductions in motility were observed in siRNA transfected cells. The error bar represents the standard deviation among the three experimental repeats.

3.6.3.2.4 Invasion Assay

Invasion assays were performed on M14 cells 72 hrs following transfection with ROPN1- and ROPN1B-siRNAs. There was significant loss of invasion (Fig 3.6.3.2.2) (ROPN1B-1 p-value =0.003; ROPN1B-2 p-value ~ 0; ROPN1-1 p-value ~ 0; ROPN1-2 p-value ~ 0) in ROPN1-siRNA- and ROPN1B-siRNA-transfected M14 cells compared to cells transfected with scrambled control. Transfection with ROPN1B-1 and ROPN1B-2 siRNA demonstrated a 47.7% and 56.4% reduction in invasion respectively, while transfection with ROPN1-1 and ROPN1-2 siRNA showed a 31.4% and 57.2% reduction in invasion respectively, compared to scrambled transfected cells (Fig 3.6.3.2.4).

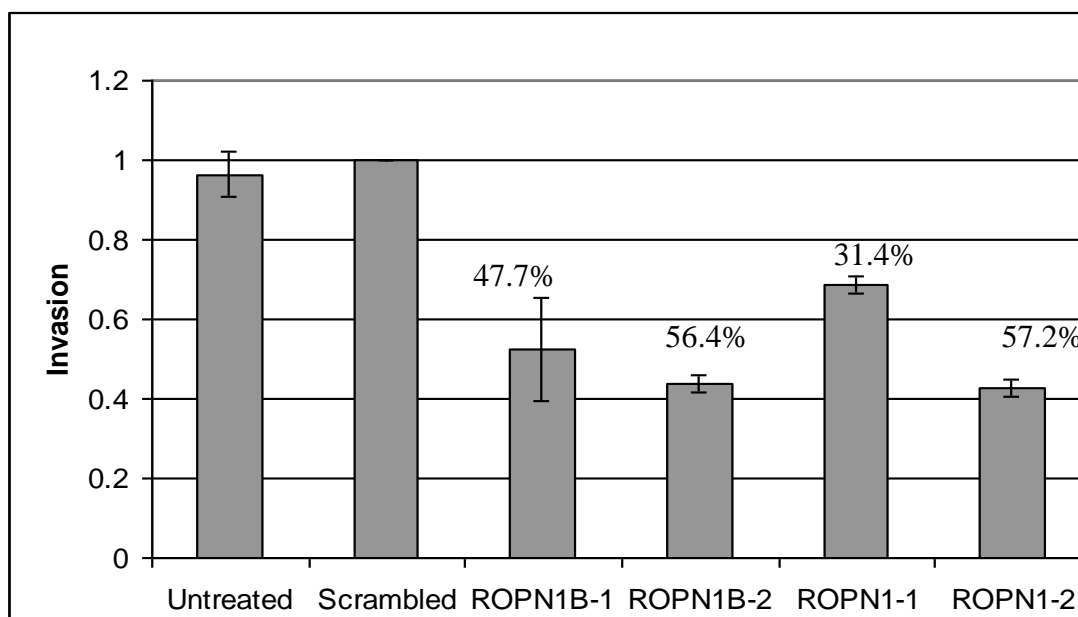
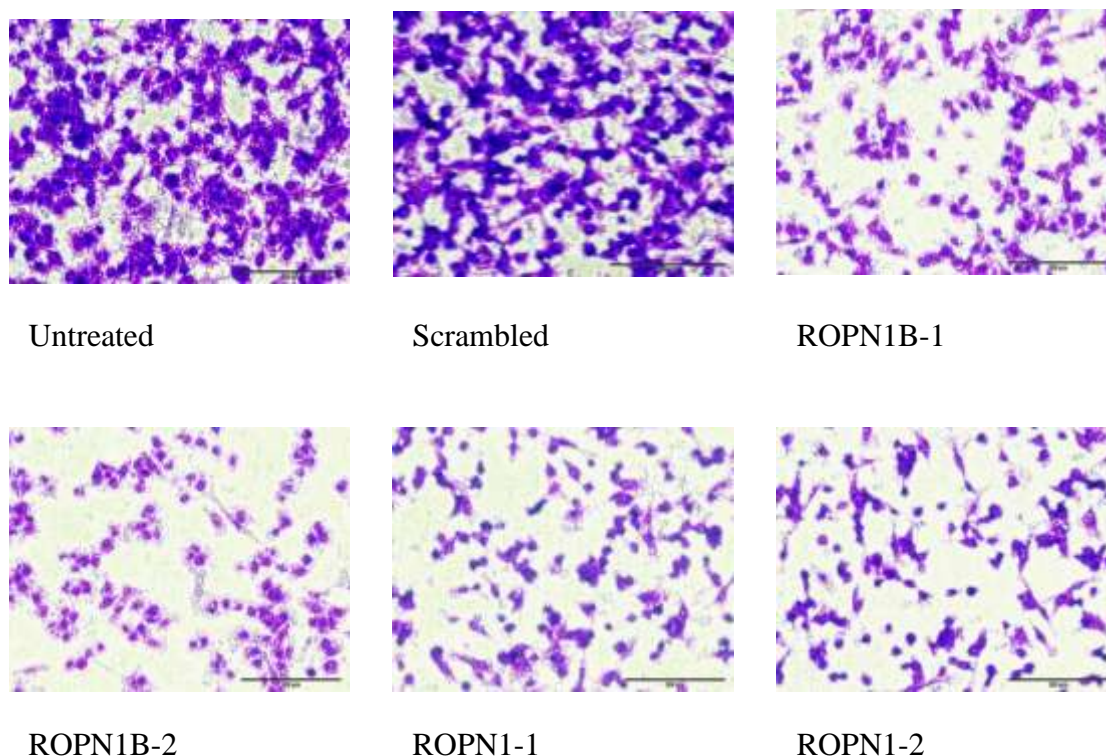


Fig 3.6.3.2.4: Invasion assays on siRNA transfected M14 cells. Assays performed in triplicate. Y-axis defines the relative invasion compared to scrambled. Significant reductions in invasion observed in ROPN1/ROPN1B-siRNA transfected cells. The error bar represents the standard deviation among the three experimental repeats.

3.6.4 ROPN1 and ROPN1B cDNA over-expression studies

ROPN1 cDNA in PCR4-TOPO plasmid (PCR4-TOPO-ROPN1) and ROPN1B cDNA in PCMV-SPORT6 plasmid (PCMV-SPORT6-ROPN1B) were obtained from Open Biosystems (see section 2.5.6). These plasmids were used to over-express these genes in the MDA-MB-231, MDA-MB-435s and M14 cell lines.

3.6.4.1 Over-expression of ROPN1B in MDA-MB-231

MDA-MB-231 cells were transfected with ROPN1B plasmid (PCMV-SPORT6-ROPN1B) and empty plasmid (PCMV-SPORT6) (see section 2.5.6).

3.6.4.1.1 qRT-PCR analysis

qRT-PCR was performed to determine expression of the gene in the transfected cells. There was 8226.1-fold up-regulation of ROPN1B RNA in the PCMV-SPORT6-ROPN1B transfected cells compared to the PCMV-SPORT6 transfected cells (Fig 3.6.4.1.1). The high fold was due to zero expression in un-transfected cells.

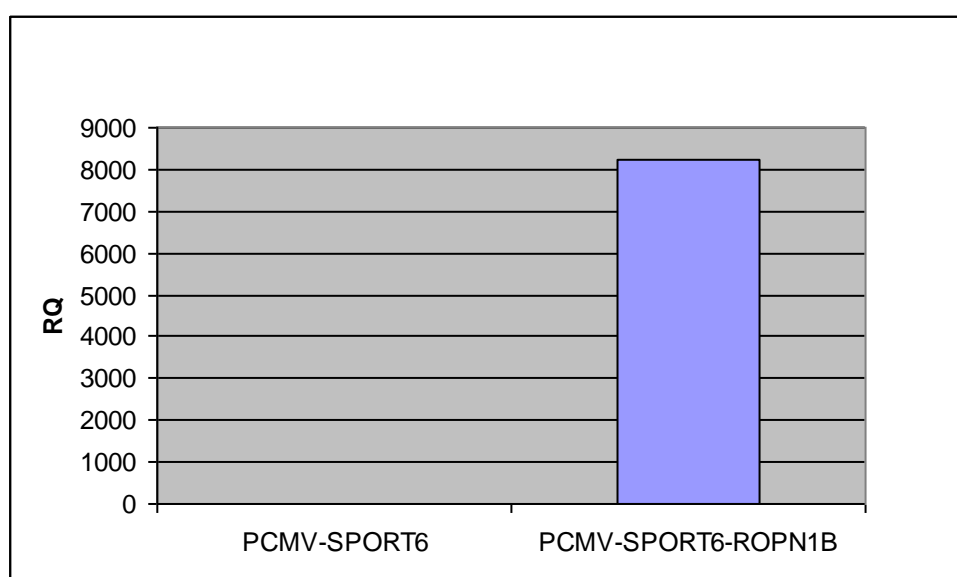


Fig 3.6.4.1.1: ROPN1B qRT-PCR. ROPN1B over-expression was observed in ROPN1B transfected MDA-MB-231 cells.

3.6.4.1.2 Invasion Assay

MDA-MB-231 was transfected with PCMV-SPORT6-ROPN1B and PCMV-SPORT6 and invasion assays were carried out on the transfected cells. Results indicate a significant (p-value =0.003) loss in invasion following PCMV-SPORT6-ROPN1B transfection (Fig 3.6.4.1.2) compared to PCMV-SPORT6-transfected cells. There was 28.1% reduced invasion in PCMV-SPORT6-ROPN1B-transfected MDA-MB-231 cells compared to PCMV-SPORT6-transfected cells.

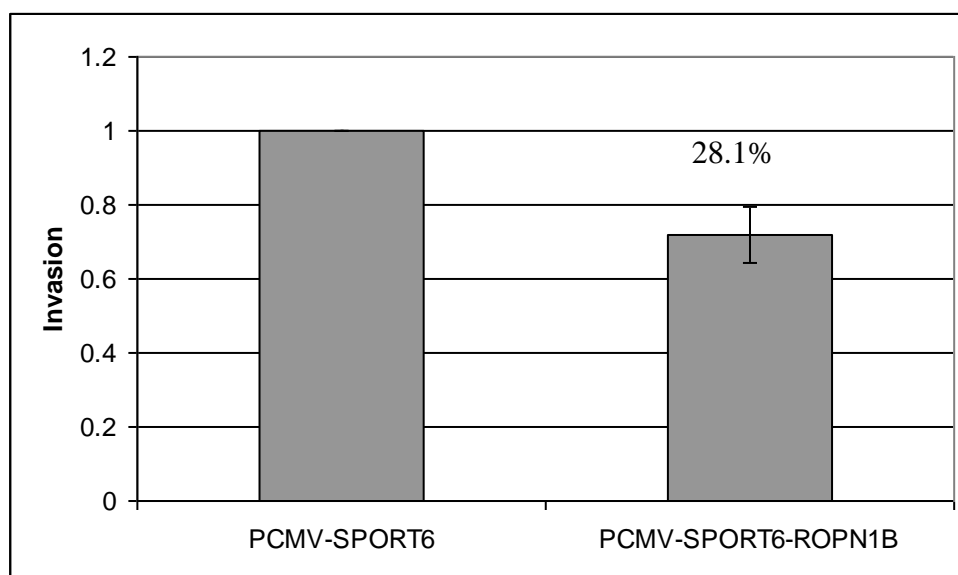
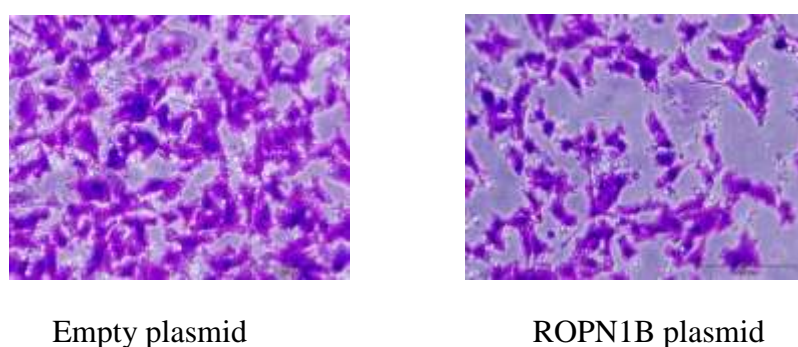


Fig 3.6.4.1.2: Invasion assay on ROPN1B transfected MDA-MB-231 cells. Assays were performed in triplicates. There was significant decrease in invasion in ROPN1B over-expressing cells. Y-axis denotes relative invasion compared to empty plasmid. The error bar represents the standard deviation among the three experimental repeats.

3.6.4.2 Over-expression of ROPN1 and ROPN1B in MDA-MB-435s

MDA-MB-435s was transfected with PCMV-SPORT6 plasmid, PCR4-TOPO-ROPN1 plasmid and PCMV-SPORT6-ROPN1B plasmid. No empty plasmid control was generated for ROPN1 over-expression studies because of time constraints.

3.6.4.2.1 qRT-PCR analysis

ROPN1 qRT-PCR was performed on these transfected cells demonstrated a 4.2-fold up-regulation of ROPN1 mRNA in ROPN1 plasmid-transfected cells compared to cells transfected with PCMV-SPORT6 empty plasmid (Fig: 3.6.4.2.1). No change in expression was observed in ROPN1 mRNA in ROPN1B plasmid transfected cells compared to empty plasmid transfected cells (Fig 3.6.4.2.1).

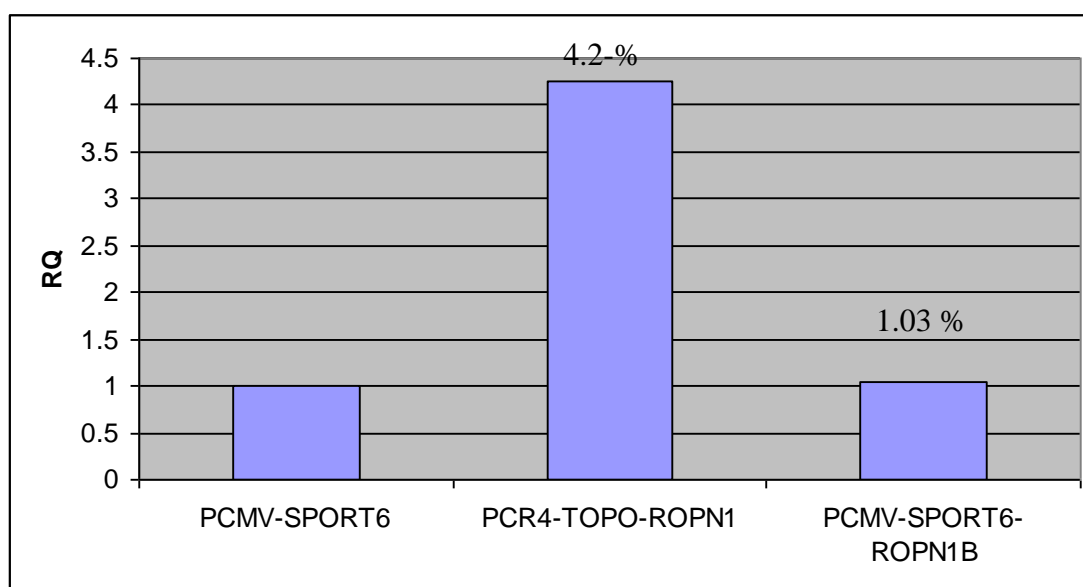


Fig 3.6.4.2.1: ROPN1 qRT-PCR. ROPN1 over-expression observed in ROPN1 transfected cells

ROPN1B qRT-PCR was also performed on these transfected cells and showed 31.5-fold up-regulation of ROPN1B mRNA in PCMV-SPORT6-ROPN1B plasmid transfected cells compared to PCMV-SPORT6 plasmid transfected cells (Fig 3.6.4.2.2). A 1.2-fold up-regulation of ROPN1B mRNA expression was observed in the PCR4-TOPO-ROPN1 plasmid transfected cells (Fig 3.6.4.2.2).

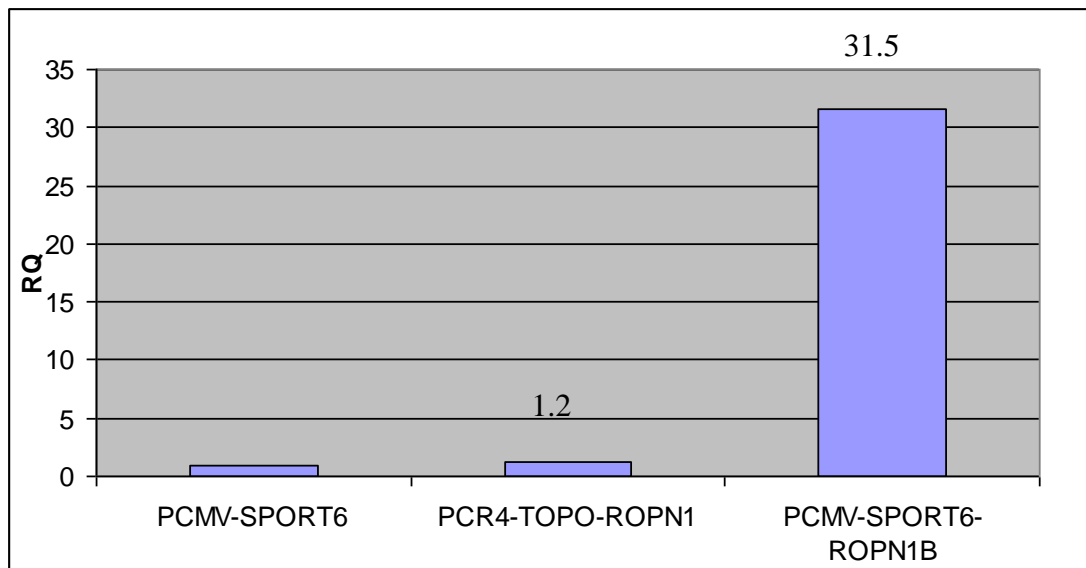


Fig 3.6.4.2.2: ROPN1B qRT-PCR. ROPN1B over expression was observed with ROPN1B over expression.

3.6.4.2.2 Western Blot Analysis

A Western blot was performed to check protein expression of Ropporin in cells transfected with ROPN1 and ROPN1B cDNA (Fig 3.6.4.2.3). As can be seen, an unexplained reduction in Ropporin protein was observed in PCVM-SPORT6-ROPN1B-transfected cells.

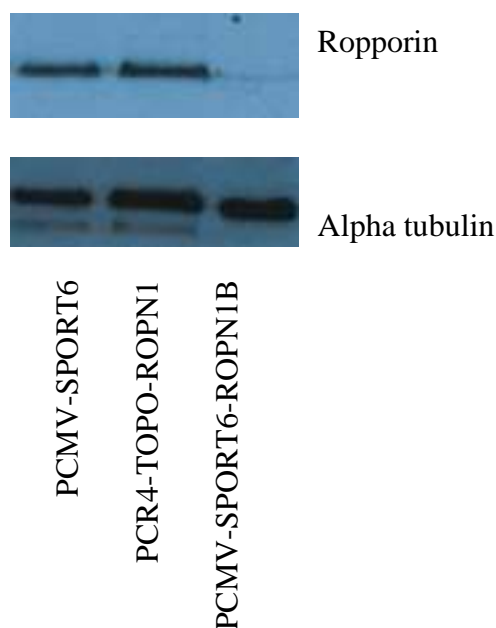
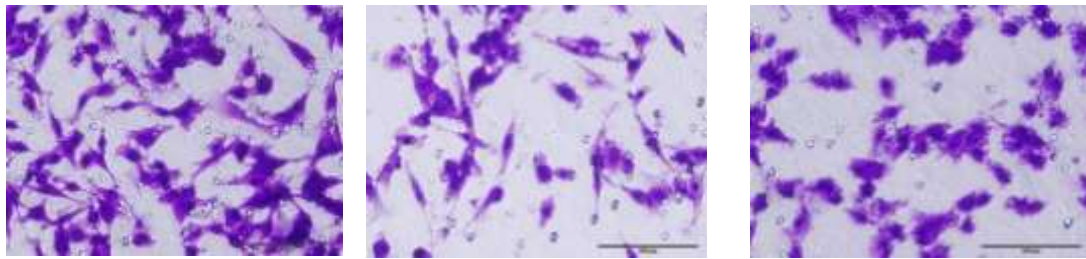


Fig 3.6.4.2.3: Western blot analysis of MDA-MB-435s cells transfected with ROPN1 and ROPN1B plasmid.

3.6.4.2.3 Motility Assay

MDA-MB-435s was transfected with PCMV-SPORT6 plasmid, PCR4-TOPO-ROPN1 plasmid and PCMV-SPORT6-ROPN1B plasmid and motility assays were performed after 72hrs. Results indicate significant (p-value =0.002) loss in invasion with ROPN1 over-expression and a marginal loss of invasion with ROPN1B over-expression (Fig 3.6.4.2.3) compared to PCMV-SPORT6 empty plasmid transfected cells. PCR4-TOPO-ROPN1 plasmid transfected cells showed 57.7% reduced invasion and PCMV-SPORT6-ROPN1B plasmid transfected cells showed 15.1% reduced motility in MDA-MB-435s cells compared to PCMV-SPORT6 empty plasmid transfected cells (Fig 3.6.4.2.4).



Empty plasmid

ROPN1 transfected

ROPN1B transfected

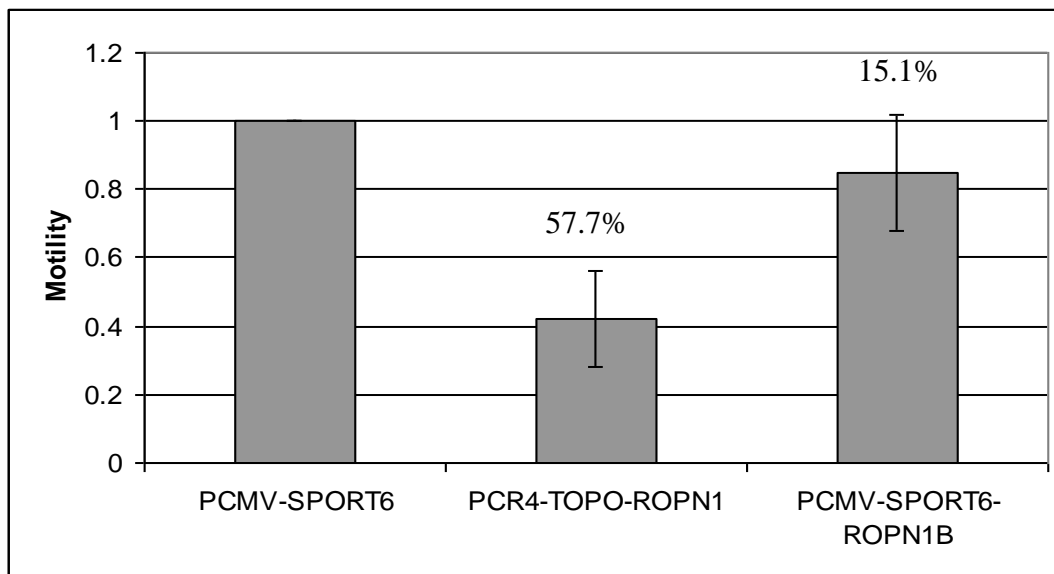


Fig 3.6.4.2.4: Motility assay on ROPN1 and ROPN1B transfected MDA-MB-435S cells. Assays were performed in triplicates. There was significant loss in motility with ROPN1 over-expression and a marginal loss of motility with ROPN1B over-expression compared to PCMV-SPORT6 empty plasmid transfected cells. The error bar represents the standard deviation among the three experimental repeats.

Invasion assay was giving highly variable (un-reproducible) results and was therefore removed from analysis.

3.6.4.3 Over-expression of ROPN1 and ROPN1B in M14

M14 cells were transfected with PCMV-SPORT6 plasmid, PCR4-TOPO-ROPN1 plasmid and PCMV-SPORT6-ROPN1B plasmid.

3.6.4.3.1 qRT-PCR analysis

ROPN1 qRT-PCR was performed on the transfected cells to check the expression of ROPN1 mRNA. qRT-PCR analysis showed 96.9-fold over-expression of ROPN1 in ROPN1 plasmid transfected cells compared to PCMV-SPORT6 empty plasmid transfected cells (Fig 3.6.4.3.1). No over-expression of ROPN1 mRNA was observed in PCMV-SPORT6-ROPN1B plasmid transfected cells compared to PCMV-SPORT6 empty plasmid transfected cells (Fig 3.6.4.3.1).

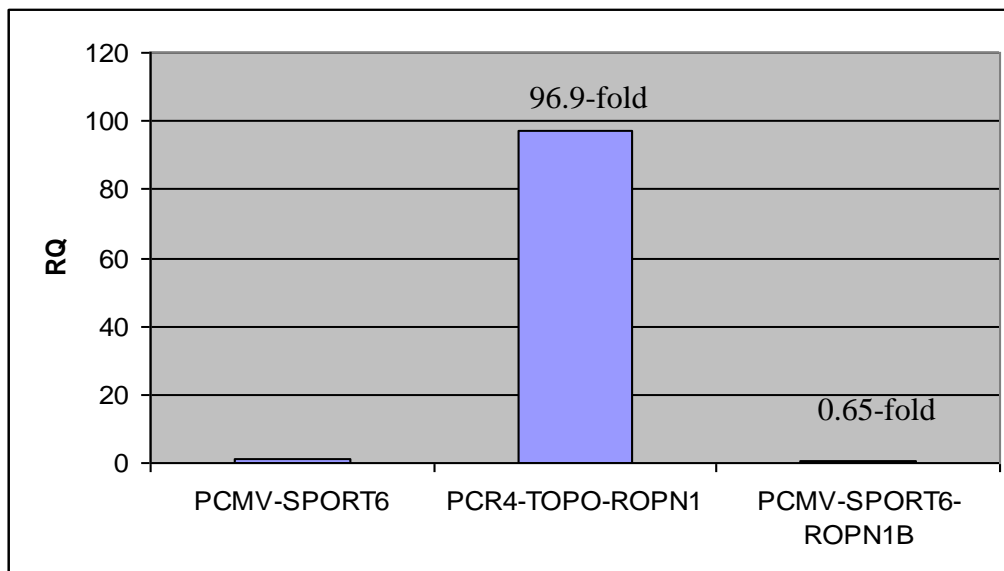


Fig 3.6.4.3.1: ROPN1 qRT-PCR. ROPN1 over expression was observed in ROPN1 transfected M14 cells.

qRT-PCR was performed to analyze the expression of ROPN1B in PCMV-SPORT6 empty plasmid, PCR4-TOPO-ROPN1 plasmid and PCMV-SPORT6-ROPN1B plasmid transfected cells. There was 50.0-fold and 6.1-fold over-expression of ROPN1B in ROPN1B plasmid transfected cells and ROPN1 plasmid transfected cells respectively, compared to PCMV-SPORT6 empty plasmid transfected cells (Fig 3.6.4.3.2).

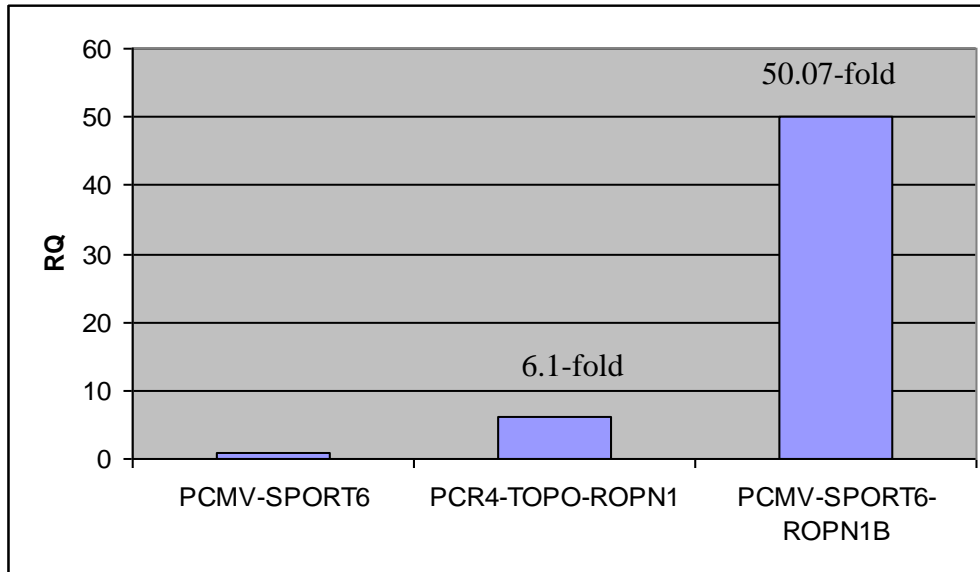


Fig 3.6.4.3.2: ROPN1B qRT-PCR. ROPN1B over expression was observed in ROPN1B transfected M14 cells.

3.6.4.3.2 Western Blot Analysis

A western blot was performed to check protein expression of Ropporin in cells transfected with ROPN1 and ROPN1B cDNA (Fig 3.6.4.3.3). As can be seen, an unexplained reduction in Ropporin protein was observed in PCR4-TOPO-ROPN1 and PCMV-SPORT6-ROPN1B transfected cells compared to PCMV-SPORT6 transfected cells.

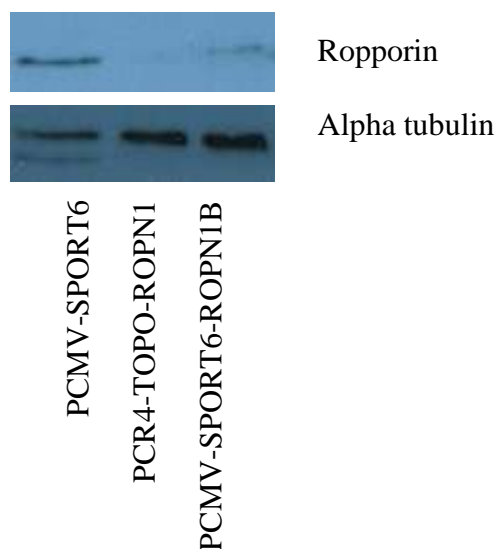
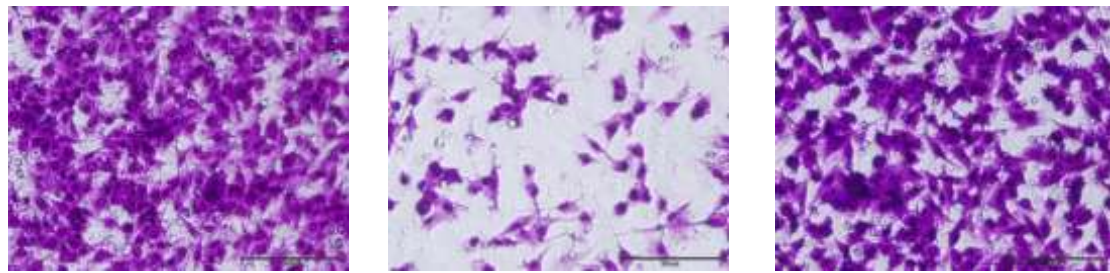


Fig 3.6.4.3.3: Western blot analysis of M14 cells transfected with ROPN1 and ROPN1B plasmid.

3.6.4.3.3 Motility Assay

M14 was transfected with PCMV-SPORT6 plasmid, PCR4-TOPO-ROPN1 plasmid and PCMV-SPORT6-ROPN1B plasmid and motility assays were performed. Results indicate significant loss in motility (p -value =0.0002) with ROPN1 over-expression and a marginal loss of motility with ROPN1B over-expression (Fig 3.6.4.3.4) compared to PCMV-SPORT6 empty plasmid transfected cells. ROPN1 plasmid transfection resulted in 65.6% reduced motility and ROPN1B plasmid transfection resulted in 25.0% reduced motility in M14 cells compared to PCMV-SPORT6 empty plasmid transfected cells (Fig 3.6.4.3.4).



Empty plasmid

ROPN1 transfected

ROPN1B transfected

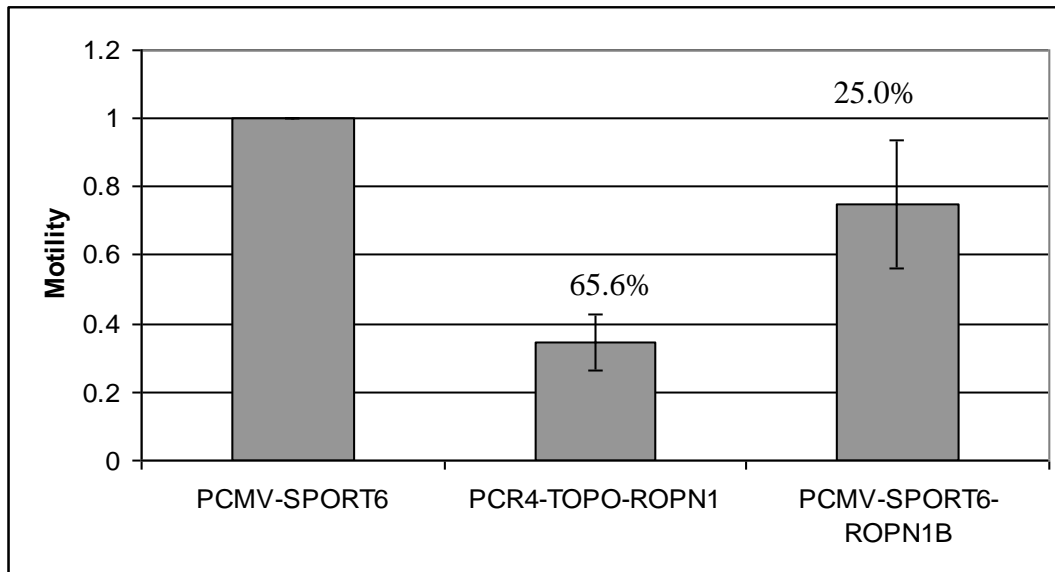


Fig 3.6.4.3.4: Motility assays on ROPN1 and ROPN1B transfected M14 cells. Assays were performed in triplicates. There was significant loss in motility with ROPN1 over-expression and a marginal loss of motility following ROPN1B over-expression. Y-axis denotes relative motility compared to empty plasmid. The error bar represents the standard deviation among the three experimental repeats.

3.6.4.3.4 Invasion Assay

Invasion assays were performed on M14 cells transfected with PCMV-SPORT6 plasmid, PCR4-TOPO-ROPN1 plasmid and PCMV-SPORT6-ROPN1B plasmid. Results indicate significant loss in invasion (p -value =0.003) following ROPN1 over-expression and a marginal loss of invasion with ROPN1B over-expression (Fig 3.6.4.3.5) compared to PCMV-SPORT6 plasmid transfected cells. PCR4-TOPO-ROPN1 plasmid transfection resulted in 48.9% and PCMV-SPORT6-ROPN1B

plasmid transfection resulted in 14.6% reduced invasion in M14 cells compared to PCMV-SPORT6 plasmid transfected cells (Fig 3.6.4.3.5).

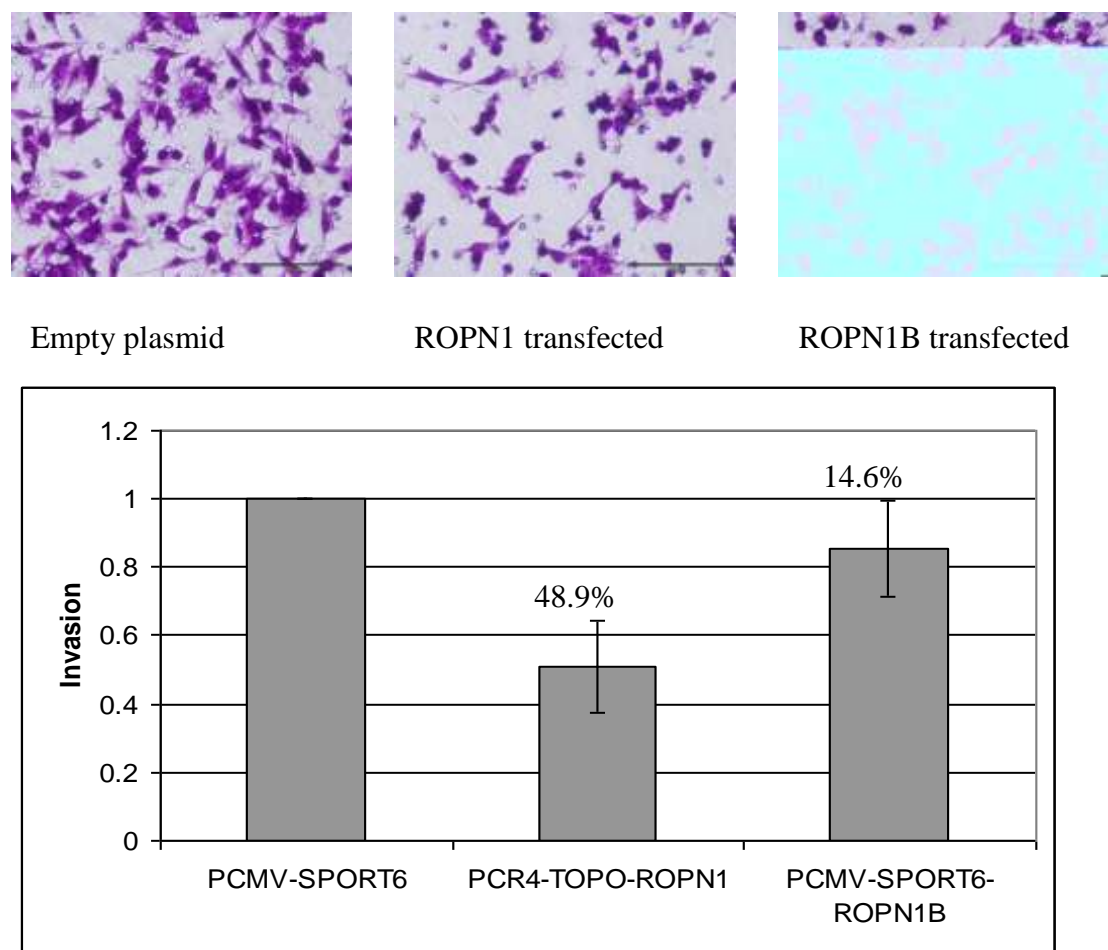


Fig 3.6.4.3.5: Invasion assay on ROPN1 and ROPN1B transfected M14 cells. Assays were performed in triplicates. There was significant loss in invasion with ROPN1 over-expression and a marginal loss of invasion following ROPN1B over-expression. Y-axis denotes relative invasion compared to empty plasmid. The error bar represents the standard deviation among the three experimental repeats.

3.6.5 Summary

ROPN1B over-expression was linked to breast cancer patients who relapsed. The gene was also linked to disease progression in melanoma. siRNA knockdown positively associated ROPN1B gene to be involved in cancer cell motility and invasion. The results from over-expression studies were inconclusive. Over-

expression of ROPN1 and ROPN1B resulted in reduction of Ropporin protein expression in M14. Over-expression of ROPN1B in M14 resulted in reduction of Ropporin protein. Over-expression studies of ROPN1 and ROPN1B resulted in reduction in protein level and reduction in invasion and motility.

3.7 How Representative are Cell line models of clinical conditions?

The aim of this section was to estimate the representative nature of breast cancer cell lines to their respective clinical specimen type using gene expression data and has been published previously (Mehta *et al.*, 2007).

Gene expression profiles of 189 breast clinical specimens (GEO accession: GSE2990) and 19 cell lines (GEO accession: GSE3156) were obtained from Gene Expression Omnibus (see section 2.1). These samples were pooled as a single experiment and normalized using the dChip algorithm (see section 2.2.1). Since the clinical specimens were analyzed using U133A and the cell lines were analyzed using U133_Plus 2.0 microarray chips, the genes not represented in U133A were removed from the cell line data, giving a total available probe set number of 22,283. However, for the ER analysis on cell lines (see section 2.1.2), all the genes on the U133_Plus 2.0 chips (54,675) were included.

3.7.1 Data filtration

Two SD filters of 0.5 and 1.0 were applied to generate gene lists for hierarchical clustering. For the pooled comparison of cell lines and clinical specimens, the total number of DE genes identified using a SD filter of 0.5 was 8,036. For the comparison of cell line and clinical clustering relative to ER status, a SD filter of 1.0 was used, giving 7,738 filtered genes for the cell lines and 6,643 genes for the clinical specimens. A lower SD for pooled experiment was used, to get the optimum representative number of genes for clustering.

3.7.2 Clustering

Hierarchical clustering and PCA were performed on these gene lists. Hierarchical clustering, as expected, using the filtered 8036 gene-list, separated the sample set into two distinct clusters (Fig 3.7.2.1), one comprising the clinical specimens and the other comprising the cell line models. To examine whether the differences in hierarchical

clustering between cell lines and tumour specimens were due to differences incorporated by sample processing at different sites, this group replicated this clustering analysis substituting a separate 104-tumour dataset for the 189-tumour dataset detailed here. In this experiment, two separate clusters of cell lines and tumour specimens were again observed (data not shown).

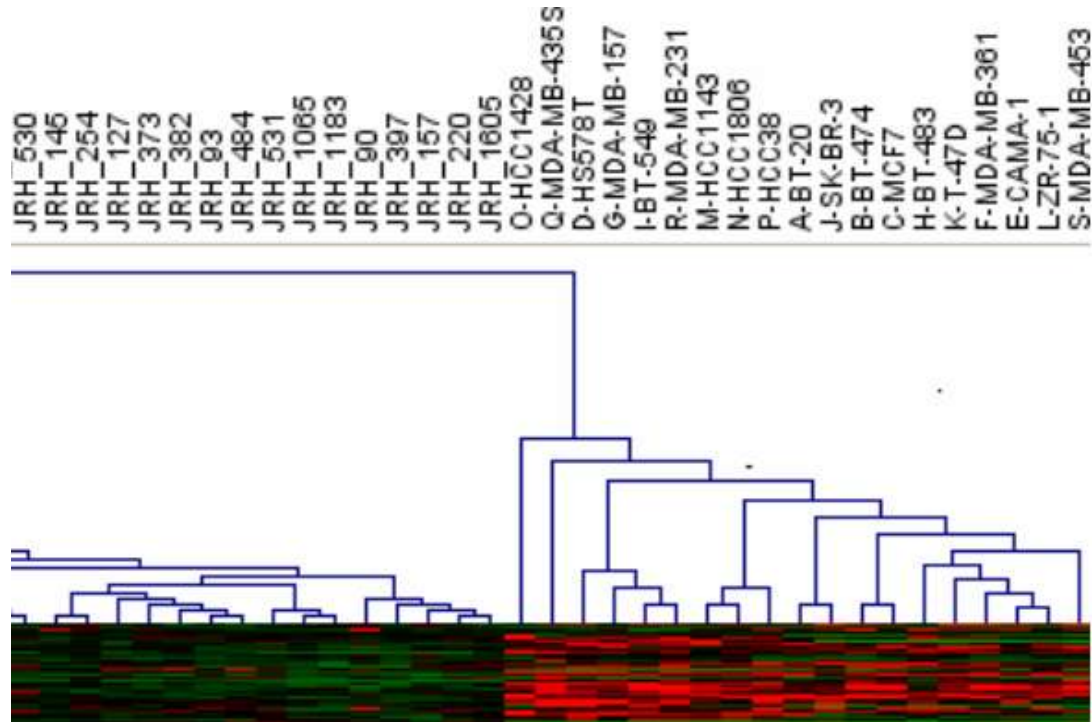


Fig 3.7.2.1: Hierarchical clustering demonstrating that cell lines and clinical specimens form two discrete groups. The right cluster is of 19 cell lines included in the study. The left cluster (incompletely shown because of large number of tumour specimens) represents 189 breast tumours.

PCA was also performed on the sample using the filtered 8036 gene-list, which also separated the clinical specimens and cell lines into two distinct groups (Fig.3.7.2.2). As can be seen on the axes, the total variance accounted for in the sample set was 27.95%. The clinical specimens also segregated into two further sub-groups, although not as distinct as that separating the cell lines and the clinical specimens.

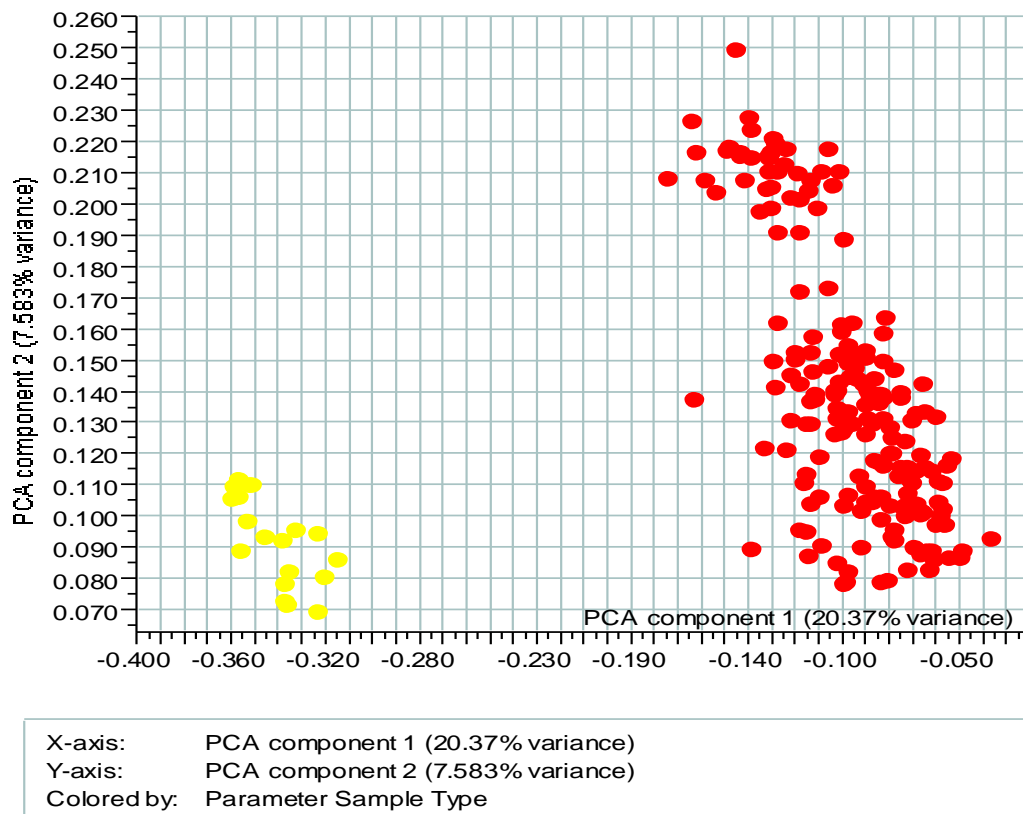


Fig 3.7.2.2: Principal component analysis was performed on the samples and the two components were plotted. Clinical specimens are highlighted in red, cell line samples are highlighted in yellow.

3.7.3 Significant genes

The clinical specimens and the breast cell lines were compared for transcripts which were significantly up- or down-regulated in the two groups (p -value <0.001 , fold change >2 and difference of 100 Affymetrix units). 2,615 genes passed the above filtration criteria, of which 1,086 were up-regulated in cell lines relative to clinical specimens and 1,529 genes were down-regulated in cell lines compared to clinical specimens.

3.7.4 Gene ontology and pathway analysis

GenMAPP Gene ontology and Pathway analysis was performed on the up- and down-regulated gene lists and the over-represented GO categories/canonical pathways are outlined in Tables 3.7.4.1 and Table 3.7.4.2. In cell lines relative to clinical specimens, many of the functions which were over-represented were related to cell

cycle functions and nucleic acid processing (Tables 3.7.4.1). Where clinical specimens were compared to cell lines, the majority of categories and pathways affected were related to the immune response and related functions (Table 3.7.4.2).

GO Name	Changed	Measured	Z Score
Mitotic cell cycle	61	281	15.236
Cell cycle	87	576	13.874
Mitosis	29	105	12.317
M phase of mitotic cell cycle	29	107	12.163
M phase	33	137	11.981
Nuclear division	31	132	11.404
Cell proliferation	94	877	10.501
DNA replication and chromosome cycle	30	154	9.791
Regulation of cell cycle	45	325	9.111
Mitotic anaphase	6	11	8.5
MAPP Name			
Cell cycle KEGG	22	84	9.081
DNA replication Reactome	11	42	6.347
G1 to S cell cycle Reactome	9	65	3.311
Translation Factors	7	48	3.069
Pentose Phosphate Pathway	2	7	2.856
mRNA processing Reactome	12	115	2.739
Cholesterol Biosynthesis	3	15	2.664

Tables 3.7.4.1: GO terms and pathways enrichment analysis for genes over-expressed in cell line models compared to clinical specimens. Higher Z score represents a stronger association of that function to genes which have over-expressed in cell lines relative to clinical specimens.

GO Name	Changed	Measured	Z Score
Immune response	107	595	12.412
Defense response	110	650	11.847
Response to biotic stimulus	115	710	11.586
MHC class II receptor activity	9	11	10.458
Extracellular matrix	48	215	9.992
Antigen processing, exogenous antigen via MHC class II	8	10	9.731
Antigen presentation, exogenous antigen	8	10	9.731
Extracellular	105	742	9.451
Antigen presentation	12	23	9.204
Antigen processing	12	23	9.204
MAPP Name			
Complement Activation Classical	7	16	5.405
Complement and Coagulation Cascades KEGG	11	49	3.897
Matrix Metalloproteinases	7	30	3.216
Smooth muscle contraction	19	143	2.574
Inflammatory Response Pathway	6	31	2.435

Tables 3.7.4.2: GO terms and pathways enrichment analysis for genes over-expressed in clinical specimens compared to cell line models. A higher Z score represents a stronger association of that function to genes which have over-expressed in clinical specimens compared to cell line models.

3.7.5 Estrogen receptor analysis

Hierarchical clustering was also performed separately on the two groups (*i.e.* cell lines and clinical specimens), to determine if either group clustered similarly when compared for ER status. This analysis segregated the cell lines into two distinct groups, which clustered largely according to their ER status (Fig.3.7.5.1). Exceptions to this rule included the ER-negative SK-BR-3 & MDA-MB-453 and the ER-positive HCC1428, which clustered with the opposite group. Hierarchical clustering performed on the 189 clinical sample dataset did not demonstrate any appreciable

clustering according to ER status, although there was a tendency of clinical specimens to cluster based on their grade (data not shown).

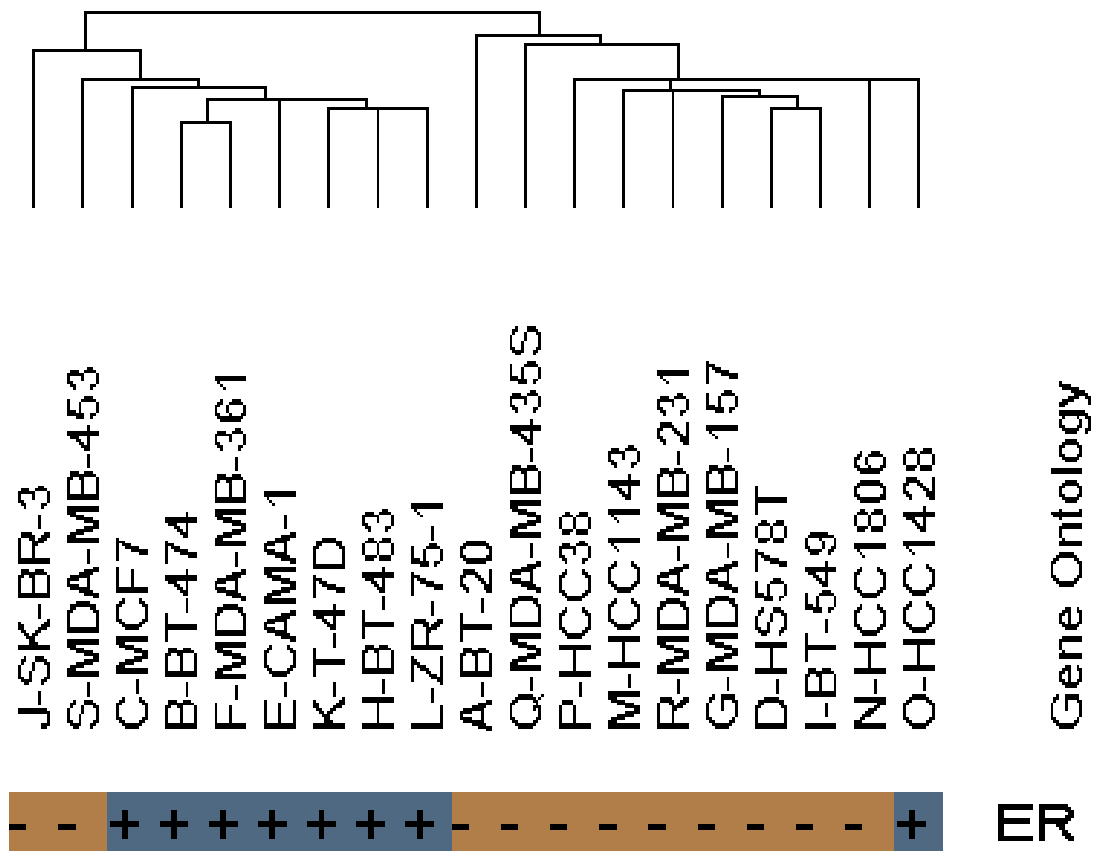


Fig 3.7.5.1: Hierarchical clustering of cell lines. The + indicates ER-positive cell lines and the ER-negative represents ER-negative cell lines. The left cluster is enriched with ER-positive cell lines and the right cluster is enriched with ER-negative cell lines.

3.7.6 Summary

The above analysis shows that there is a marked difference in gene expression in tissue compared to cell lines. This difference was consistent in Hierarchical clustering and Principal component analysis. Genes related to cell cycle, mitosis, DNA replication were highly up-regulated in cell line models. Similarly, genes related to the immune system, Complement Activation Classical pathway and Matrix Metalloproteinase were significantly up-regulated in tissue relative to cell lines. The above results indicate that cell cycle and immune response results from cell-line models of various clinical conditions may not accurately reflect their behaviour *in-*

vivo. These results should be taken into account when extrapolating the cell line results to clinically relevant conditions.

3.8 Molecular profile of basal cell carcinoma

The aim of this section was to investigate the gene expression profile of basal cell carcinoma using whole genome expression microarrays and compare these profiles with the gene expression profile of normal skin. This work is currently the only whole-genome analysis of BCC published worldwide (O'Driscoll *et al.*, 2006).

Microarray gene expression profiling of 20 basal cell carcinoma tissue specimens and 5 normal skin tissue was performed using Affymetrix U133 Plus2.0 arrays (see section 2.1.3). Microarray samples were processed by Lorraine O'Driscoll and Padraig Doolan, while my role was in the analysis of the chip data generated. Tissue specimens from twenty cases of BCC were obtained from Blackrock Clinic and the Bons Secours Hospital, Dublin, snap-frozen in liquid nitrogen, and were subsequently stored at -80°C. Five normal skin specimens (from consenting male and female volunteers of a similar age range who do not/never had skin cancer) were also included in the studies. Following this RNA was isolated and microarray was performed for each chip (see section 2.5.11).

3.8.1 Data Normalization and Quantification

The microarray raw data files were normalized and quantified using the dChip algorithm as outlined in section 2.2.1.

3.8.2 Data Filtration

Data filtration was applied on 54,675 genes present on U133 Plus2.0 chip (see section 2.2.3), to remove genes which i) did not fluctuate very highly across samples and ii) fluctuated too highly across samples to be trustworthy. Genes with a Standard deviation / Mean i) below 1 or ii) above 1000 were removed from further analysis. This set of genes was used for Hierarchical clustering. 692 genes passed this criterion and were used to carry out clustering analysis of clinical specimens.

3.8.3 Hierarchical Clustering

Hierarchical clustering (see section 2.2.4) was performed on the 692 member filtered gene list. The distance metric used was 1-correlation and the clustering algorithm used was Average linkage clustering. Prior to clustering, the individual samples were

standardised as follows: the expression of the individual genes was subtracted from their means for that sample and divided by their respective standard deviation. The results are shown in Fig 3.8.3.1.

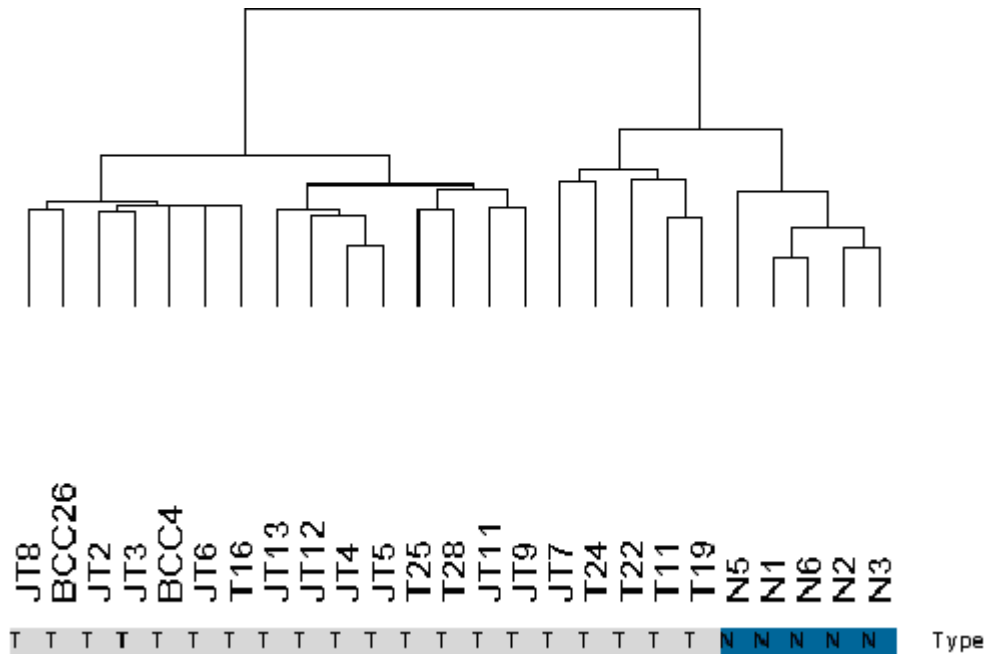


Fig 3.8.3.1: The hierarchical clustering represents the clustering pattern of the BCC and normal specimens. Type indicates whether the sample is Normal/Tumour.

The results indicate that the normal specimens clustered together and the BCC specimens clustered together. A subset of BCC samples (JT7, T24, T22, T11 & T19) clustered close to the normal specimens

3.8.4 Normal specimens vs. Basal cell carcinoma

Up-regulated gene transcripts

2,108 genes were identified as significantly up-regulated ($p \leq 0.05$, Fold Change (FC) > 1.2 and Difference > 100) in cancer specimens compared to normal specimens. Genes were ranked by fold change and, based on this criterion, the top 20 genes are listed in Table 3.8.4.1

Gene ontology analysis was performed on the up-regulated genes ($p \leq 0.05$, $FC > 2$, and $\text{Difference} > 100$). Significant functions were identified based on p-value ($p \leq 0.05$) and the 10 most significant functions represented are listed in Table 3.8.4.2.

Pathway analysis was performed using GenMAPP database on the up-regulated genes ($p \leq 0.05$, $FC > 2$, and $\text{Difference} > 100$). Significant pathways were identified based on p-value ($p \leq 0.05$) and the 10 most significant pathways are listed in Table 3.8.4.3.

Down-regulated genes

1,813 genes were identified as significantly down-regulated ($p \leq 0.05$, Fold Change (FC) < -1.2 and $\text{Difference} < -100$) in cancer specimens compared to normal specimens. Genes were ranked by fold change and, based on this criterion, the top 20 genes are listed in Table 3.8.4.4

Gene ontology analysis was performed on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and $\text{Difference} < -100$). Significant functions were identified based on p-value ($p \leq 0.05$) and 10 most significant functions represented are listed in Table 3.8.4.5.

Pathway analysis was performed using GenMAPP database on the down-regulated genes ($p \leq 0.05$, $FC < -2$, and $\text{Difference} < -100$). Significant pathways were identified based on p-value ($p \leq 0.05$) and the 10 most significant pathways are listed in Table 3.8.4.6.

probe set	Gene	baseline	experiment	fold change	P value
204697_s_at	CHGA	19.31	2516.41	130.34	0.000001
224590_at	XIST	6.27	433.45	69.08	0.001884
214218_s_at	XIST	10.05	631.37	62.79	0.00072
224588_at	XIST	42.73	2128.81	49.82	0.000552
204913_s_at	SOX11	8.48	236.67	27.92	0.000224
214913_at	ADAMTS3	26.38	632.27	23.97	0.000002
220345_at	LRRTM4	12.72	282.87	22.24	0.0061
230863_at	---	15.38	308.31	20.05	0.029359
204915_s_at	SOX11	28.16	553.08	19.64	0.000059
204424_s_at	LMO3	70.78	1358.37	19.19	0.003562
208025_s_at	HMGA2	30.2	536.39	17.76	0.000173
215311_at	---	27.85	476.31	17.1	0.000013
227671_at	XIST	24.01	407.71	16.98	0.003059
218638_s_at	SPON2	188.15	3181.08	16.91	0
208212_s_at	ALK	53.02	888.43	16.76	0.000003
226346_at	LOC92312	37.28	571.83	15.34	0
204914_s_at	SOX11	22.04	332.68	15.1	0.000124
215443_at	TSHR	17.2	257.19	14.95	0.00019
213960_at	---	40.57	575.46	14.18	0.000001
229523_at	TTMA	47.92	662.44	13.82	0.000001

Table 3.8.4.1: Genes up-regulated in cancer specimens in comparison to normal specimens

GOID	GO Name	Changed	Measured	p-value
5201	Extracellular matrix structural constituent	16	78	0
5581	Collagen	8	30	0
30199	Collagen fibril organization	3	5	0
5578	Extracellular matrix (sensu Metazoa)	34	359	0
31012	Extracellular matrix	34	365	0
7155	Cell adhesion	43	559	0
7275	Development	94	1696	0
5576	Extracellular region	63	1096	0
8201	Heparin binding	9	62	0
6817	Phosphate transport	10	81	0

Table 8.1.4.2 Functions enriched among genes up-regulated in cancer specimens in comparison to normal specimens

MAPP Name	Changed	Measured	p-value
2-Tissues-Muscle Fat and Connective	12	82	0
2-Tissues-Blood and Lymph	11	78	0
Focal adhesion KEGG	17	187	0
Wnt Signaling	9	71	0
Wnt NetPath 8	11	109	0.001
TGF Beta Signaling Pathway	6	52	0.003
1-Tissue-Embryonic Stem Cell	5	47	0.014
Apoptosis	7	82	0.017
1-Tissue-Muscle fat and connective	6	65	0.021
Chondroitin Heparan sulfate biosynthesis	3	20	0.023

Table 3.8.4.3: Pathways enriched among genes up-regulated in cancer specimens in comparison to normal specimens

probe set	Gene	baseline	experiment	fold change	P value
208962_s_at	FADS1	3354.77	166.07	-20.2	0.045399
229476_s_at	THRSP	4948.62	288.53	-17.15	0.034919
207275_s_at	ACSL1	1997.95	132.25	-15.11	0.039975
206799_at	SCGB1D2	1383.74	92.43	-14.97	0.042413
221561_at	SOAT1	1087.52	88.95	-12.23	0.008422
206714_at	ALOX15B	4264.04	372.27	-11.45	0.045644
214240_at	GAL	1747.34	157.08	-11.12	0.015802
201625_s_at	INSIG1	1773.81	182.47	-9.72	0.030044
231810_at	BRI3BP	988.19	110.99	-8.9	0.021503
211056_s_at	SRD5A1	1640.77	208.21	-7.88	0.032165
229957_at	TMEM91	1397.34	183.05	-7.63	0.025272
204675_at	SRD5A1	3279.09	446.39	-7.35	0.027894
231736_x_at	MGST1	2894.74	397.53	-7.28	0.031804
205029_s_at	FABP7	608.63	84.11	-7.24	0.021739
201627_s_at	INSIG1	728.68	100.79	-7.23	0.049971
209522_s_at	CRAT	2366.24	327.16	-7.23	0.020824
226064_s_at	DGAT2	2666.99	375.45	-7.1	0.032748
231156_at	HAO2	395.01	55.75	-7.09	0.045195
223184_s_at	AGPAT3	1603.83	228.89	-7.01	0.045149
205030_at	FABP7	2208.85	344.36	-6.41	0.021392

Table 3.8.4.4: Genes down-regulated in cancer specimens in comparison to normal specimens

GOID	GO Name	Changed	Measured	p-value
16126	Sterol biosynthesis	14	27	0
6695	Cholesterol biosynthesis	12	22	0
16125	Sterol metabolism	16	67	0
44255	Cellular lipid metabolism	44	417	0
6629	Lipid metabolism	49	535	0
8203	Cholesterol metabolism	14	61	0
42579	Microbody	15	70	0
5777	Peroxisome	15	70	0
6694	Steroid biosynthesis	14	64	0
16491	Oxidoreductase activity	48	579	0

Table 3.8.4.5: Functions enriched among genes down-regulated in cancer specimens in comparison to normal specimens

MAPP Name	Changed	Measured	p-value
Cholesterol Biosynthesis	11	15	0
Sterol biosynthesis	9	19	0
Fatty Acid Beta Oxidation Meta BiGCaT	8	32	0
Terpenoid biosynthesis	3	6	0
Fatty Acid Beta Oxidation 1 BiGCaT	6	27	0
Bile acid biosynthesis	7	37	0
Pyruvate metabolism	7	34	0.001
Citrate cycle TCA cycle	5	24	0.001
Butanoate metabolism	6	38	0.001
1-Tissue-Muscle fat and connective	7	65	0.002

Table 3.8.4.6: Pathways enriched among genes down-regulated in cancer specimens in comparison to normal specimens

3.8.5 Summary

The study was the first whole genome study on Basal cell carcinoma. The analysis identified important genes, functions and pathways which may be important in transformation of normal skin to basal cell carcinoma. Wnt signaling pathways was up-regulated in BCC vs. normal skin whereas Cholesterol Biosynthesis pathway was down-regulated in BCC patients vs. normal skin.

4.0 Discussion

Breast cancer is one of the most common malignancies among females. The heterogeneous nature of the disease coupled with the lack of robust markers for prediction, prognosis, and response to treatment has so far eluded our understanding of its complex nature. It is generally assumed that transcriptional profiling, involving multiple gene signatures would be more predictive of tumour behaviour rather than single genes. Various studies have tried to answer this question with promising results correlating gene expression profiles with prognosis, recurrence, metastatic potential, therapeutic response, as well as biological and functional aspects of the disease. The integration of genomic approaches into the clinic lies ahead, but such studies need to be validated on large datasets. The challenge also lies in getting a better understanding of the various groups and sub-groups of breast cancer and how they may correlate with various groups of prognostic mRNA.

In order to gain a better understanding of breast cancer heterogeneity and its association with clinical outcome, gene expression analysis was performed on 104 cancer specimens and 17 normal specimens. These specimens were from patients who underwent surgery during 1993–1997, and for whom follow-up clinical information was available. Additionally large datasets were downloaded from public repositories and analysed and compared to our dataset to get a holistic picture in order to help find answers to the complex questions associated with breast cancer. We aimed to find clinical heterogeneity among the breast cancer, genes, functions and pathways associated with various clinical conditions and prognostic important genes. The aim was also to perform meta-analysis on prognostic important genes and ER pathway genes.

4.1 Clinical heterogeneity in breast cancer

The heterogeneity of breast cancer and the variability in the clinical response to treatment has led to wide interest in understanding the molecular mechanisms of the different behavioural phenotypes of breast cancer. Currently the most widely used markers for breast cancer classification and treatment are ER, PR and HER2 protein. These proteins are estimated based on immunohistochemistry. ER and PR are used as indicators of endocrine-sensitive breast cancers and HER2 as indicators of breast cancer patients with

metastatic disease who may benefit from trastuzumab therapy (Duffy 2005). Gene expression profiling has been widely used to understand the molecular mechanisms involved in disease progression (Cooper 2001), metastasis (Weigelt *et al.*, 2005), drug metabolism and resistance (Brennan *et al.*, 2005).

Sample clustering techniques have been widely used to group samples with similar expression patterns and have immensely contributed to our understanding of heterogeneity associated with various types of cancers. In particular, such analysis has identified various sub-groups in breast cancer. Estrogen receptor status is the single most important criteria by which the clinical specimens tend to cluster (van 't Veer *et al.*, 2002; Sotiriou *et al.*, 2003). Cluster analysis also identified Luminal subtype A and B, ERBB2, Basal and Normal type and identified intrinsic gene expression signatures for each group which correlated with patient outcome (Sorlie *et al.*, 2001). This classification based on gene expression was later validated by many independent studies (Sorlie *et al.*, 2003; Calza *et al.*, 2006; Hu *et al.*, 2006).

The sample clustering techniques were applied to the gene expression profile of our set of 104 breast cancer and 17 normal specimens. Genes with low or extremely high variability among specimens were not used for the purpose of clustering. First, a correlation matrix using all the specimens was created followed by two-way clustering of samples. In this way, the similarity between the specimens and the homogeneity among the individual sub-groups were assessed. Additionally, this technique helped identify any further sub-clusters within a cluster.

The 104 cancer and 17 normal specimens divided into many distinct groups. Five main clusters (one of which could be sub-divided into three sub-clusters) were identified, some of them very specific to certain clinical parameters.

4.1.1 High level of correlation between Normal samples (Cluster A)

The most significant result in our study was that most of the normal specimens clustered together (Cluster A). However, three normal specimens did cluster with a group of cancer specimens over-represented by ER-positive tumors with relatively low expression of ER

partner (genes involved in ER pathway) genes (Cluster E). Also, the normal group contained one of the cancer specimens. Of the normal specimens which clustered as a group (Cluster A), a high level of correlation was observed among the normal specimens indicating that the normal breast gene expression profiles are alike (section 3.1.3).

The remaining 103 tumour specimens represented a highly heterogeneous group, (in contrast to a very similar group of normal specimens in Cluster A) and there was relatively less correlation observed among the closely related specimens on the cluster as observed from the hierarchical clustering result. These results indicate that the normal breast specimens have very similar and unique gene expression patterns, and the transformation of normal breast cells to a tumour leads to a divergent pattern of gene expression. One possible reason for this may be that uncontrolled cell division may lead to higher rates of mutations occurring in the genomic DNA leading to different combinations and a wide diversity of aberrant expression patterns (Gagos and Irminger-Finger 2005).

4.1.2 Samples closest in character to Normal samples enriched for ER- & Grade1

The group of samples that clustered closest to the normal specimen group was a set of specimens enriched for ER-negative status and Grade 1 (see section 1.3.1) tumors (Cluster B). The ESR1 gene was not over-expressed in this cluster; however ERBB2 was over-expressed in this group, when compared to the normal breast specimens. Similar results presenting gene expression profiles in groups of breast cancer specimens that are substantially “Normal-like” have been described in other studies (Sorlie *et al.*, 2001; Calza *et al.*, 2006)

The rest of the samples clustered into two big groups. One was enriched for ER-negative specimens (Cluster C) and the other was enriched for ER-positive specimens (Cluster D and Cluster E). The clusters enriched with ER-negative patients were also enriched for patients with higher grade and who relapsed (7 year relapse).

4.1.3 ER-negative samples (Cluster C) display three distinct sub-clusters

A high level of heterogeneity was observed in the ER-negative cluster (Cluster C) and three distinct sub-clusters were observed based on the hierarchical clustering and correlation among specimens. The left and right cluster had relatively worse prognosis based on KM (Kaplan-Meier), compared to the middle cluster. Specific genes of interest expressed in separate sub-clusters include high expression of the Ropporin gene, ERBB2 and genes related to the immunoglobulin family.

KM curve analysis associated the left sub-cluster of Cluster C as having the worst survival. To gain further insight, the specimens of this cluster were compared to the specimens of middle and right sub-cluster for gene expression changes. Ropporin was identified as substantially over-expressed in this left sub-cluster compared to the other two sub-clusters and a detailed study of this mRNA was performed. Our results positively associated the over-expression of this gene to an ER-negative phenotype, high relapse and shorter survival.

Our results also showed that the sub-cluster enriched with ERBB2 over-expressing patients had a higher incidence of relapse and reduced survival (right sub cluster of Cluster C). ERBB2-overexpressing clusters of patients have been observed in other similar independent studies using sample clustering (Sorlie *et al.*, 2001; Calza *et al.*, 2006) and expression of this protein is associated with a poor prognosis and poor response to chemotherapeutic drugs (Revillion, Bonnetterre and Peyrat 1998).

The third (middle) sub-cluster had patients with relatively low relapse and longer survival. This cluster of patients exhibits high expression level of genes involved in immune response. IFI6 (interferon, gamma-inducible protein 6), IL8 (Interleukin-8), LOC652128 (similar to Ig heavy chain V-II region ARH-77 precursor), IGL (immunoglobulin lambda locus), IGHM (immunoglobulin heavy constant mu) genes were found to be over-expressing in this sub-cluster compared to its neighbouring cluster. Other studies in the literature also suggest that high expression of immune response genes is associated with a favourable prognosis in ER-negative sub-groups of patients (Teschendorff *et al.*, 2007). Additionally, a previous study (Alexe *et al.*, 2007) associated

a higher expression of lymphocyte/immune response associated genes in a HER2-over-expressing cluster with a low recurrence subtype. While over-expression of immune response genes has been linked to better prognosis in ER-negative specimens, this study is the first to demonstrate the clustering of these clinical samples as a distinct group. This result positively associated the ER-negative sub-group of patients with high expression of immune response genes with a favourable clinical outcome.

4.1.4 ER-positive tumors sub-divide as two groups.

The remaining samples were all enriched for ER-positive specimens and grouped into two distinct clusters (D & E), which may be due to differences in the level of expression of ESR1 and ER partner gene expression profiles. Cluster D had a relatively higher expression of ER partner genes such as ESR1, GATA3, FOXA1, SPDEF and TFF3. This high-ER-expressing cluster contained only one specimen with an ER-negative phenotype. The lower-ER-expressing cluster (Cluster E) had 3 ER-negative specimens and the three normal specimens were clustered away from the other normal specimens clustered with this group. This low ER cluster had a marginally reduced survival and a slightly higher incidence of relapse compared to its neighbouring cluster which displayed very high ER partner gene expression. Therefore, higher expression of ER genes might be linked to better prognosis, or better response to tamoxifen, since most of the ER-positive patients were treated with this drug. Similar clustering patterns and links to prognosis has been reported in other studies (Sorlie *et al.*, 2003; Calza *et al.*, 2006).

In conclusion, our gene expression profiling results identified various groups and sub-groups of breast cancer and associated them with the clinical parameters and linked them with the clinical outcomes. Our results identified new clusters with clinical relevance.

4.2 Gene expression differences between Normal and Cancer tissue

The transition from normal tissue to cancerous tissue is an important aspect in understanding the biology of breast cancer. Gene expression profiling can help identify the differences among the normal and cancerous tissue and can help better design drugs to target the disease.

The normal and cancer specimens were compared in order to identify genes and pathways that contribute to the transition from normal breast tissue to a cancerous state.

4.2.1 Cell cycle pathway up-regulated in tumors

Gene ontology and pathways analysis identified cell cycle related pathways to be over-expressed and over-represented in tumors compared to normal specimens. This is not surprising considering the fact that uncontrolled cell division is associated with tumour development and alteration in cell cycle checkpoints may be responsible for cancer (Hartwell 1992, Kastan and Bartek 2004). Our study identified TP53 to be up-regulated in breast cancer in comparison to the normal breast tissue (section 3.1.4). Alteration in TP53 gene products is involved in bad prognosis of breast cancer (Borresen *et al.*, 1995; Overgaard *et al.*, 2000; Langerod *et al.*, 2007). A high expression of TP53 in follicular lymphoma was observed in high grade and oversized tumors and correlated with poor prognosis (Pennanen *et al.*, 2008) and also in a subset of ductal carcinomas *in situ*, with no expression observed in atypical lesions (Chitemerere *et al.*, 1996).

4.2.2 Embryonic stem cell pathway up-regulated in cancer

In our study, many genes reported to be highly expressed in embryonic stem cell pathway were up-regulated in cancer compared to normal tissue indicating that breast cancer might originate from the stem cell, or has characteristics similar to stem cells. Many other studies have implicated abnormality in stem cells to the origin of cancer. Mutations among the stem cell genes could lead to an alteration in genomic stability, resulting in immortality and onset of cancer (Ashkenazi, Gentry and Jackson 2008). Additionally, the P53 gene has roles in normal and cancer stem cell differentiation, apoptosis, self-renewal and the capacity for tumourigenesis (Zheng *et al.*, 2008).

Cyclin dependent protein and protein complex of cyclin B1 (CCNB1) induces phosphorylation of key substrates mediating cell cycle progression from G2 to M phase (Morgan 1995; Nurse 1994). Recently it has been identified as an oncogene and is over-expressed in the cells from leukemia and other tumors including breast cancer cells from patient tissues at G1 phase (Shen *et al.*, 2004). Over-expression of cyclin B1 has been

associated with tumor invasion and reduced survival and is also reported to be a prognostic marker in several tumor types (Yu, Zhan and Finn 2002; Murakami *et al.*, 1999).

Increased expression of CDC20 is reported to be a common event in various cancer including colorectal and bladder cancer tissues as well as in oral squamous cell carcinoma and gastric cancer (Kidokoro *et al.*, 2008; Mondal *et al.*, 2007; Kim *et al.*, 2005). Suppression of growth by inducing G2/M arrest was observed by reducing the expression of CDC20 using CDC20-specific siRNA suggesting it as a potential therapeutic target for various cancers (Kidokoro *et al.*, 2008).

MELK is associated with the regulation of spliceosome assembly, gene expression and cell proliferation (Vulsteke *et al.*, 2004; Saito *et al.*, 2005; Nakano *et al.*, 2008). It is also expressed in several vertebrate tissues including the blast cells of the early embryo, embryonic stem cells, adult germ cells (ovaries and spermatogonia), hematopoietic stem cells and neural stem cells (Heyer, Kochanowski and Solter 1999; Nakano *et al.*, 2005; Easterday *et al.*, 2003). MELK gene transcript controls the cell cycle and acts to regulate the self-renewal of neural stem cells (Nakano *et al.*, 2005). Marie *et al.* (2008) found progressively higher expression of MELK in a study carried out on more than 100 tumors of the central nervous system and a high level of expression in glioblastoma multiforme. It has been directly associated with proliferation and anchorage-independent growth in glioblastoma multiforme and brain tumors, identifying it as a possible therapeutic target for these types of cancer (Nakano *et al.*, 2008; Marie *et al.*, 2008).

PFS2 has been associated with ovule patterning by regulating cell proliferation of the maternal integuments and regulating the timing of cellular differentiation of the megaspore mother cell (Pillitteri *et al.*, 2007; Park *et al.*, 2005). Pillitteri *et al.*, (2007) reported that PFS2 might be responsible for properly coordinating the developmental states of the sporophytic integument tissues and gametophytic embryo sac.

PRC1 plays a functional role in regulating mitosis and the protein is highly expressed during S and G2/M phase (Jiang *et al.*, 1998). p53 is found to directly suppress PRC1

gene transcription in HCT116 p53^{+/+}, HCT116 p53^{-/-}, MCF-7, T47D, and HeLa cells (Li, Lin and Liu 2004).

4.2.3 Fatty acid biosynthesis pathway down-regulated in cancer

Several gene members of the fatty acid biosynthesis pathway were significantly down-regulated in cancer compared to normal tissue. Our results contradict other published reports where fatty acid synthesis pathway is up-regulated in cancer (Kuhajda 2000; Pizer *et al.*, 1996).

The fatty acid synthesis pathway was found to be selectively activated in a study carried out on human prostate cancer tissue using *in situ* hybridization (Swinnen *et al.*, 2000). This group also reported a relationship between increased lipogenesis and cancer progression.

A study carried out on a group of established human breast carcinoma cell lines-SKBR3, ZR-75-1, MCF-7, and MCF-7a (doxorubicin-resistant)-and normal human fibroblasts (HS-27) suggested that fatty acid synthesis was required by some of the cancers for their growth and inhibition of fatty acid synthesis can inhibit the growth of neoplastic cells (Kuhajda *et al.*, 1994).

The relationship between abnormal fatty acid synthesis and an aggressive tumor phenotype is still not fully understood. Fatty acids are reported to be involved in tumorigenesis (Cohen *et al.*, 1986), in receptor-mediated signal transduction (Tomaska and Resnick 1993), as well as modulators of tumor cell adhesion. The role of increased endogenous fatty acid biosynthesis in tumorigenesis is unknown. One of the possibilities could be that lipid mediators in the tumor cells may act in an autocrine or paracrine fashion affecting tumor behavior. It is also found that certain tumors have an apparent requirement for endogenous fatty acid biosynthesis compared to normal cells. These reported results suggest that inhibition of fatty acid biosynthesis could be a potential target for chemotherapy development (Kuhajda *et al.*, 1994).

The synthesis of fatty acids from acetyl-CoA and malonyl-CoA is carried out by an active enzyme fatty acid synthase (FASN) (Epstein, Carmichael and Partin 1995; Shurbaji *et al.*, 1992). FASN is an important enzyme of the fatty acid synthesis and are found to be highly expressed in human cancers, including carcinoma of the breast, prostate, ovary, endometrium and colon (Epstein, Carmichael and Partin 1995; Shurbaji *et al.*, 1992; Rashid *et al.*, 1997; Alo *et al.*, 1996; Milgraum *et al.*, 1997).

Stearoyl-CoA desaturase (SCD) is an enzyme which helps in the biosynthesis of monounsaturated fatty acids and also controls the regulation of metabolism in liver and skeletal muscle (Dobrzyn and Ntambi 2005). A link between SCD activity and tumor cell proliferation has been observed with increased expression of SCD in colonic and esophageal carcinoma, hepatocellular adenoma, as well as in chemically induced tumors (Thai *et al.*, 2001; Li *et al.*, 1994). SCD regulates programmed cell death and is crucial for cell survival (Scaglia and Igal, 2005). Down-regulation of SCD has also been associated with significantly decreased proliferation and invasiveness (Scaglia and Igal, 2005).

Long-chain acyl-CoA synthetases (ACSL) are necessary for fatty acid degradation, phospholipid remodeling, and production of long acyl-CoA esters that act as a regulator of various physiological processes in mammals (Soupene and Kuypers 2006).

In conclusion, our results identified genes, functions and pathways that are associated with transition of normal breast tissue to cancer. TP53 gene was found to be up-regulated in breast cancer. Cell cycle pathways and embryonic stem cells genes were up-regulated in cancer. Fatty acid biosynthesis pathway genes were down-regulated in breast cancer.

4.3 Genes up-regulated in Estrogen Receptor-positive breast patients

4.3.1 In-house study

Clinical decision making has very much relied on ER status of patients and many individual studies have tried to identify genes involved in Estrogen metabolism. In our study, 34 ER-negative breast specimens and 67 ER-positive breast specimens were used

to identify differential-expressed (DE) transcripts. Further meta-analysis was performed using five additional publicly available datasets, including one cell line data.

Our in-house data identified 855 up-regulated genes associated with ER-positive vs ER-negative specimens. Affected functions and pathways over-represented by these genes were related to cellular morphogenesis and epidermal growth factor receptor activity. Cellular morphogenesis is the key feature in the development of mammary gland and is supposed to be influenced by ESR1 gene and Estrogen (Mallepell *et al.*, 2006; Sternlicht 2006). Bi-directional regulation among EGFR and ER has been reported by other studies (Levin 2003; Britton *et al.*, 2006). Estrogen can regulate expression of EGF receptor proteins and may play a role in Estrogen-stimulated growth (Mukku and Stancel 1985). ER and EGFR signal through various kinases and influence transcriptional and non-transcriptional actions of Estrogen in breast cancer cells (Levin 2003). Increased EGFR signalling is associated with tamoxifen resistance in ER-positive breast cancer cells (Fox *et al.*, 2008).

Our in-house data identified 1145 down-regulated genes associated with ER-expression. Affected functions and pathways over-represented by these genes were related to immune response. Immune response is the main molecular process associated with prognosis in the ER and HER2 receptor-negative subgroups (Desmedt *et al.*, 2008). Despite ER-negative groups having a high proliferation and poor clinical outcome, a group exists with high expression of immune response genes with good prognosis (Schmidt *et al.*, 2008; Teschendorff *et al.*, 2007).

4.3.2 Meta analysis

Meta-analysis was performed on six independent datasets, using common criteria to identify DE genes. 62 up-regulated transcripts common to all experimental groups were identified, which will be discussed here.

ESR1 gene is very critical for ER action. ESR1 is ligand-activated transcription factor composed of several domains important for hormone binding, DNA binding, and activation of transcription. This gene was over-expressed in our study among the ER-

positive patients. This gene was also up-regulated in all cohorts among the ER-positive specimens. ESR1 gene amplification is quite frequent in breast cancer (Holst *et al.*, 2007) and endometrial cancer (Lebeau *et al.*, 2008). Genetic polymorphism in the gene is associated with transformation of benign tumors to cancer (Gallicchio *et al.*, 2006) and has been linked to breast cancer risk, tumour characteristics and survival (Einarsdottir *et al.*, 2008).

GATA3 (GATA binding protein 3) is important for mammary gland morphogenesis and luminal cell differentiation and is closely associated with ESR1 gene expression; however it has little prognostic value independent of ER (Voduc, Cheang and Nielsen 2008). GATA3 co-expresses with ESR1 gene (Tozlu *et al.*, 2006) and ER alpha pathway genes (Wilson and Giguere 2008). Genetic variability in the intronic region of GATA3 is associated with differential susceptibility to breast cancer (Garcia-Closas *et al.*, 2007). GATA3 was up-regulated in all the experimental datasets comparing ER-positive cancer to ER-negative cancer.

Forkhead box A1 (FOXA1) is a forkhead family transcription factor expressed in breast cancer cells and is associated with luminal subtype (Thorat *et al.*, 2008). It is also strongly correlated with ESR1 expression (Tozlu *et al.*, 2006; Lacroix and Leclercq 2004). FOXA1 is correlated with luminal subtype A (Badve *et al.*, 2007) and with favourable prognosis (Badve *et al.*, 2007; Habashy *et al.*, 2008; Wolf *et al.*, 2007a) and plays a growth inhibitory role in breast cancer (Wolf *et al.*, 2007a). This gene was up-regulated in all of the cohorts of ER-positive specimens vs. ER-negative specimens.

SPDEF (SAM pointed domain containing ets transcription factor) is an Ets transcription factor expressed at high levels primarily in tissues with high epithelial cell content, including prostate, colon, and breast (Seth and Watson 2005). Its protein product is reduced in human invasive breast cancer and is absent in invasive breast cancer cell lines (Feldman *et al.*, 2003). SPDEF over-expression is associated with nodal metastasis and hormone receptor positivity in invasive breast cancer (Turcotte *et al.*, 2007). High expression of this gene in tumour and peripheral blood makes it a candidate prognostic

marker (Ghadersohi and Sood 2001). SPDEF was also up-regulated in all our experimental groups of ER-positive specimens vs. ER-negative specimens.

C1ORF34 (tetratricopeptide repeat domain 39A) expression correlates with expression of ER in cell lines and was also detected in primary breast carcinomas but not in normal breast tissue (Kuang *et al.*, 1998). This gene was also up-regulated in all our experimental groups of ER-positive specimens vs. ER-negative specimens.

Other important and well known genes involved in the ER pathway which were up-regulated in ER-positive patients vs. ER-negative patients in our study were CA12, TFF1, TFF3, ERBB4, MYB, NAT1, eEF1A2, LIV-1.

Carbonic anhydrase XII (CA12) is a marker of good prognosis in invasive breast carcinoma (Watson *et al.*, 2003). CA12 expression is associated with ER-positive tumors (Barnett *et al.*, 2008; Tozlu *et al.*, 2006).

TFF1 (trefoil factor 1) is a small cysteine-rich secreted protein that is frequently expressed in breast tumors in the ER-positive patients. TFF1 is expressed in ER-positive tumors and its expression correlates with the expression of ESR1 (Tozlu *et al.*, 2006; Wilson and Giguere 2008). Estrogens can stimulate the motility of breast cancer cells via the induction of TFF1 (Prest, May and Westley 2002). Like TFF1, TFF3 (trefoil factor 3) is also frequently expressed in breast tumors. Although closely related, there are marked differences in shape, size, and surface charge of these proteins (May *et al.*, 2003). Like TFF1, TFF3 also co-expresses with ESR1 gene in breast cancer (Wilson and Giguere 2008). In our meta-analysis both of these genes were found to be up-regulated in all experimental groups among the ER-positive specimens vs. ER-negative specimens.

ERBB4 (v-erb-a erythroblastic leukemia viral oncogene homolog 4 avian) gene is a member of the tyrosine protein kinase family and the epidermal growth factor receptor subfamily. Its expression is associated with ER-positive tumors (Zhu *et al.*, 2006). The ERBB family encodes 4 proteins, ERBB (HER1), ERBB-2 (HER2, NEU), ERBB-3 (HER3), and ERBB-4 (HER4); HERs 1-3 are associated with poor survival, however HER4 is associated with better survival (Witton *et al.*, 2003; Bieche *et al.*, 2003). In our

meta-analysis ERBB4 was up-regulated in all experimental groups among the ER-positive specimens vs. ER-negative specimens.

MYB (v-myb myeloblastosis viral oncogene homolog avian) is a proto-oncogene and its expression is strongly associated with ER and PR in breast cancer (Guerin, Barrois and Riou 1988). Its expression is correlated with the expression of ESR1 gene (Tozlu *et al.*, 2006). This gene is also associated with improved prognosis (Guerin, Barrois and Riou 1988). However, proliferation of ER-positive breast cancer cell lines is inhibited when MYB expression is knocked down (Drabsch *et al.*, 2007). In our meta-analysis MYB was up-regulated in all experimental groups among the ER-positive specimens vs. ER-negative specimens.

NAT1 (N-acetyltransferase 1 arylamine N-acetyltransferase) expression correlates with expression of ESR1 (Tozlu *et al.*, 2006; Wakefield *et al.*, 2008). DNA hypomethylation in the NAT1 gene is present in cancerous breast tissue indicating that this type of methylation may significantly influence the transcriptional activation of the gene (Kim *et al.*, 2008). In our meta-analysis NAT1 was up-regulated in all experimental groups among the ER-positive specimens vs. ER-negative specimens.

Translation elongation factor eEF1A2 is a potential oncoprotein that is over expressed in two-thirds of breast tumors (Tomlinson *et al.*, 2005) and predicts favourable outcome in breast cancer (Kulkarni *et al.*, 2007). In our meta-analysis NAT1 was up-regulated in all experimental groups among the ER-positive specimens vs. ER-negative specimens.

LIV-1 (solute carrier family 39 zinc transporter, member 6) breast cancer protein belongs to a family of histidine-rich membrane proteins and controls intracellular Zn²⁺ homeostasis (Taylor 2000). LIV-1 is associated with ER-positive tumors (Tozlu *et al.*, 2006). In our meta-analysis LIV-1 was up-regulated in all experimental groups among the ER-positive specimens vs. ER-negative specimens.

Our meta-analysis also identified novel genes which may be involved in ER metabolism and disease progression e.g. MYO5C, TPBG, RGL2, MKL2, THRAP2, LASS6, INPP4B, COX6C, MCCC2, RAB17, ANXA9, THSD4, ABAT, HSPB1 for which the

available literature is limited. These genes were up-regulated in all experimental groups among the ER-positive specimens vs. ER-negative specimens.

Gene ontology and pathways analysis provided some functional insight to the mechanism of ER action. Our in-house dataset identified functions related to morphology to be significant among the ER-positive specimens. Pathway analysis identified nuclear receptors to be a significant pathway. This pathway was also found to be significant in the meta-analysis. The gene was termed as DE if it was significantly up-regulated in at least 50% of the datasets among the ER-positive specimens vs. ER-negative specimens. The DE genes were analysed (non DE as background) to identify functions and pathways enriched for ER genes (see section 3.3). Not DE genes were those which were not found differentially expressed in any of the experimental group. The functions up-regulated in ER-positive specimens vs. ER-negative specimens were Zinc ion transport, Neutral amino acid transporter activity, unconventional myosin and Insulin-like growth factor receptor activity. Estradiol regulates the expression of insulin like growth factor initiating an intracellular signal transduction pathway that activates transcription factors, including the estrogen receptor (Martin and Stoica 2002). Pathway analysis identified Tissue-Muscle fat and connective, Nuclear Receptors, Electron Transport Chain, and Androgen-Receptor NetPath 2 as up-regulated in ER-positive specimens.

4.3.2.1 Nuclear Receptor pathway

Nuclear receptors are transcription factors and become active when they detect a certain ligand in the cellular environment and have the ability to bind to DNA and activate genes (Mangelsdorf *et al.*, 1995). The most commonly studied nuclear receptors in breast cancer are ER and PR. However recently many other nuclear receptors has been found to be important in breast cancer, including those of Androgen receptor, corticosteroids, fat-soluble vitamins A and D, fatty acids and xenobiotic lipids derived from diet (Conzen 2008).

Retinoic acid receptor (RAR) alpha or RARA has found to be differentially expressed in ER+ (MDA-MB-231, MDA-MB-330, HBL100, and Hs0578T lines) and ER- (MDA-MB-361, BT 474, and BT 20) cells and is thought to be regulated by estrogen or other

steroid hormones (Fitzgerald *et al.*, 1997; Roman *et al.*, 1992). Relatively high levels of RAR alpha was observed in ER+ mammary carcinoma cells and were responsive to retinoids, whereas most undifferentiated, estrogen-independent, ER-negative (ER-) cells showed low RAR alpha expression and retinoid resistance. It has been shown to play a role in retinoid-induced growth inhibition of human breast cancer cell lines that express the estrogen receptor (ER) (Fitzgerald *et al.*, 1997; Roman *et al.*, 1992; Schneider *et al.*, 2000). All these studies suggest that RAR alpha might regulate the normal and malignant mammary epithelial cell growth, differentiation, and apoptosis (Fitzgerald *et al.*, 1997; Sheikh *et al.*, 1994; Widschwendter *et al.*, 1997).

Nuclear receptor subfamily 2, group F, member 6 (NR2F6) expressed in lymphocyte acts as a regulator of T lymphocyte activation, potentially antagonizing antigen-receptor-induced cytokine responses *in vitro* and *in vivo* (Baier 2003). A high endogenous expression of NR2F6 mRNA was observed in embryonic brain and developing liver (Warnecke *et al.*, 2005; Miyajima *et al.*, 1988). Hermann-Kleiter *et al.*, reported a potential function for NR2F6 in the immune system as it was expressed in the thymus, spleen, lymph node, and bone marrow, CD3+ T and CD19+ B lymphocytes (Hermann-Kleiter *et al.*, 2008).

Androgen receptor (AR) is the key transcription factor required for prostate cell survival and proliferation and is reported to play a critical role in the development and progression of prostate cancer (Xu *et al.*, 2009). In an immunohistochemical study carried out on 86 patients with gastric carcinoma it was observed that patients with AR-positive tumors AR-positive had worse prognosis than AR-negative patients (Koinea *et al.*, 2004).

AR is expressed in 60% of invasive breast carcinomas and almost 50% of the ER-negative tumors have been shown to be AR-positive (Agoff *et al.*, 2003; Moinfar *et al.*, 2003). The median survival after disease recurrence of patients with AR-expressing tumors was significantly longer compared to that of patients with AR-negative tumors (Schippinger *et al.*, 2006). There have been reports of prognostic advantage for patients with AR expression in early breast cancer compared to patients with AR-negative tumors (Agoff *et al.*, 2003; Bryan *et al.*, 1984; Kuenen-Boumeester *et al.*, 1996). Of the 232 breast carcinomas examined by Schippinger *et al.*, (2006), 70.7% expressed ARs

demonstrating its expression is a common characteristic in breast cancer and may be a possible therapeutic target for endocrine antitumor therapies.

VDR is expressed in the human colon, normal epithelial cells and some cancer cells (Kallay *et al.*, 2002). Increased expression of VDR is associated with a favorable prognosis in colorectal cancer (Cross *et al.*, 1996; Evans *et al.*, 1998). Some of the other studies report that VDR expression decreases in high-grade carcinomas to levels found in normal mucosa (Kallay *et al.*, 2002; Sheinin *et al.*, 2000; Cross *et al.*, 2001), while others found diminished VDR expression already in low- and intermediate-grade tumors and a decrease below normal mucosa levels in high-grade carcinomas (Palmer *et al.*, 2004).

In conclusion, our study identified various genes and pathways which are crucial for ER metabolism. Important genes identified to be up-regulated in ER-positive tumors are ESR1, FOXA1, SPDEF, TFF1, and TFF3. Nuclear receptor pathway was found to be up-regulated in ER-positive tumors.

4.4 Gene interaction network for ESR1 gene

Large scale gene expression mining can help identify and understand the intricate relationship among correlated genes. Our analysis focussed on building a gene interaction networks around the ESR1 gene, an important gene in ER metabolism.

A 5897-sample chip dataset obtained from Array Express (E-TABM-185) was used to identify genes which correlate with ESR1 gene expression, the central gene in the ER-pathway (section 3.3.3). These specimens were from diverse types of tissue and cell lines; however, all were on Affymetrix HG-U133A chips and were normalised as a group, thus making it an excellent dataset for gene correlation analysis. Correlation measure has been widely used to understand gene interaction from gene expression data (Almudevar *et al.*, 2006; Lee *et al.*, 2004). Pearson correlation coefficient as a measure of similarity between expression profiles was used to construct network graphs from gene expression data (Freeman *et al.*, 2007). The size of the graph produced is dependent on the threshold correlation value selected. At low Pearson correlation coefficient cut-offs, networks become large whereas at higher thresholds levels, the networks consist of a smaller

number of genes and tend to be more useful for most analyses (Freeman *et al.*, 2007). A Pearson correlation cut-off > 0.75 was used to identify genes that correlate in expression to that of ESR1. Expression of GATA3, FOXA1, SPDEF and C1ORF34 were found to be correlated with ESR1 expression (Spearman Correlation > 0.75 across 5897 samples). These 5 genes were also significantly over-expressed in the ER-positive group in all the six experiments when comparing ER-positive to ER-negative specimens. Interestingly, other than C1ORF34, for which function is less known, all the other genes (ESR1, GATA3, FOXA1 and SPDEF) are transcription factors and therefore these genes may be affecting the expression of large number of other genes. These findings therefore also indicate that the ER pathway may be more complex than is currently considered and more detailed study is needed to unravel the mechanism behind the ER-positive tumour progression.

Using the expression values of these 5 genes across 5897 specimens the relationships among these genes were investigated. Hierarchical clustering and PCA results indicated that FOXA1-SPDEF and ESR1-GATA3 expression are highly correlated. K-means clustering was performed to get a better understanding of the expression pattern of the genes across all specimens. The biggest cluster (3209 specimens) had little or no expression of these 5 genes in most of the specimens. The second biggest cluster (899 specimens) had good expression of all these 5 genes in most of the specimens. These results indicate that these genes are most often expressed together. It is possible that several of these genes get switched on as part of differentiation of mammary glands and are essential for the development of luminal epithelial cells of the mammary gland (Tong and Hotamisligil 2007). Since the samples in this study were of diverse origin, there is the obvious possibility of different types of interaction of these genes in different individual cancers.

The other clusters had relatively fewer numbers of genes, but they indicated that the expressions of these genes are independent of each other, except for SPDEF expression which is present only when there is high expression of FOXA1.

Correlation plots for individual combinations of genes were analysed across all 5897 specimens to get a deeper understanding on the dependency of each gene expression on another. The results conclude that expression of ESR1, GATA3, C1ORF34, SPDEF, FOXA1 (except for SPDEF-FOXA1), although highly correlated, exist independently of each other. However, it seems likely that the expression of SPDEF may be dependent on the expression of FOXA1. A very high level of expression of FOXA1 existed with a low level of expression of SPDEF; however a high level of SPDEF expression was not observed associated with low expression of FOXA1.

In conclusion, our study identified 4 genes (FOXA1, SPDEF, GATA3, and C1ORF34) to be correlated to the expression of ESR1. Additionally our results indicated the possible dependency of SPDEF expression on FOXA1.

4.5 Genes up-regulated in ER-negative breast patients

The absence of ER protein classifies the tumour as ER-negative. This phenotype is associated with a poor prognosis. The molecular biology of ER-negative tumors is poorly understood. There is lack of targeted therapies for ER-negative tumors, especially if they are triple negative (ER-negative, PR-negative, HER2-negative). Our study aimed to identify genes up-regulated in the ER-negative tumors.

Our in-house data analysis also identified the ER-negative cluster to be enriched with patients who relapsed (overall) and high grade tumors. 6 datasets were compared to identify over-expressed genes in ER-negative tumors. Common criteria was used to identify DE ($p \leq 0.05$, Fold Change (FC) > 1.2 and Difference > 100) genes. The meta-analysis on these datasets identified 20 transcripts up-regulated in ER-negative specimens in all the datasets under study (section 3.3.2).

SFRP1 (secreted frizzled-related protein 1) is a 35 kDa member of the SFRP family. It acts as a biphasic modulator of Wnt signaling, counteracting Wnt-induced effects at high concentrations and promoting them at lower concentrations (Uren *et al.*, 2000). Promoter hypermethylation is the predominant mechanism of SFRP1 gene silencing in human breast cancer and SFRP1 gene inactivation in breast cancer is associated with

unfavourable prognosis (Veeck *et al.*, 2006). SFRP1 showed the highest fold up-regulation (8.25-fold average across all experiments) among the ER-negative specimens vs. ER-positive specimens. Our results for the first time showed a strong relation of SFRP1 with ER-negative phenotype.

Other genes up-regulated by an average fold change > 2 in ER-negative specimens vs. ER-positive specimens were COTL1, SLC43A3, C10orf38, MSN and TRIM2. COTL1 (coactosin-like 1) is a human filamentous actin-binding protein and is expressed in placenta, lung, kidney and peripheral-blood leucocytes (Provost *et al.*, 2001). Coactosin-like protein has been described as a cancer antigen in pancreatic cancer (Nakatsura *et al.*, 2002). Its function in breast cancer is unknown. SLC43A3 (solute carrier family 43, member 3) is expressed in microvascular endothelium (Wallgard *et al.*, 2008) and its expression in cancer is poorly defined. The KLF9 (Krüppel-like factor 9) alters the expression of COTL1 and C10orf38 (Simmen *et al.*, 2008). Very little information is available for MSN and TRIM2 as regards their role in cancer. Interestingly, most of the genes identified here up-regulated in ER-negative tumors have not been extensively studied previously for a role in cancer. This study has for the first time identified transcripts and genes which might be important for ER-negative breast cancer, where currently the understanding and therapeutic options are very limited. ER-negative tumors are morphologically and phenotypically very distinct from ER-positive tumors and there is need for more study and development of newer promising agents for the treatment of ER-negative breast cancer (Putti *et al.*, 2005). Our study identifies newer targets which can be studied for development of targeted therapeutics for ER-negative Breast cancers.

In conclusion, our results identified important genes up-regulated in ER-negative tumors. SFRP1 (secreted frizzled-related protein 1) gene was very highly expressed in ER-negative breast tumors.

4.6 Genes up-regulated in HER2-positive breast cancers

ERBB2 (v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog avian) gene amplification is associated with a sub-set of ER-negative tumors with a very poor prognosis (Revillion, Bonnetterre and Peyrat 1998).

Two clinical and one cell line datasets (section 3.4) were analysed and compared to identify transcripts associated with ERBB2 expression. DE genelists were created for each and the genelists were compared. 13 transcripts (6 up-regulated and 7 down-regulated) were common to all three datasets. The up-regulated transcripts common to all three datasets were ERBB2, C17orf37, STARD3, ERGIC1 and C7orf24. ERBB2 (Average fold change: 9.8), C17orf37 (Average fold change: 6.7) and STARD3 (Average fold change: 5) are all located on chromosome 17q. Amplification of chromosome 17 has previously been identified in HER2-positive breast cancer (Bose *et al.*, 2001). The other two genes ERGIC1 (chromosome 5) and C7orf24 (chromosome 7) with average fold change of 1.9 and 1.7, respectively, are on different chromosomes.

ERBB2 gene amplification is associated with shorter disease-free survival and higher incidence of death due to disease (Slamon *et al.*, 1987). The gene is located on chromosome 17 and is amplified in a subset of invasive breast cancer and correlates with poor clinical outcomes (Watters *et al.*, 2003).

The C17orf34 (chromosome 17 open reading frame 34) gene is located in close proximity to ERBB2 on chromosome 17q21 (Benusiglio *et al.*, 2006) and is expressed in early and late stages of breast cancer disease, in very high amounts in metastatic patients and is absent or has very low expression in normal breast tissue (Evans *et al.*, 2006). ERBB2, C17orf34 and STARD3 are located in a close proximity of 168kb region of chromosome 17 (Maqani *et al.*, 2006). STARD3, also known as Metastatic lymph node 64 protein (MLN64) is co-amplified with ERBB2 gene in breast cancer (Vinatzer *et al.*, 2005) and is regulated by Sp/KLF transcription factors (Alpy *et al.*, 2003).

In conclusion, our results identified important genes up-regulated in HER- positive breast patients. Many of these genes were on chromosome 17.

4.7 Lymph node-negative vs. Lymph node-positive

Clinically, lymph node status is an important criterion in the treatment choice as patients with no lymph node involvement are usually spared from aggressive treatment. With the increase in the degree of lymph node involvement, which indicates the invasion of cancer

cells to nearby tissue, the survival of breast cancer patient decreases (Carter, Allen and Henson 1989). Since metastatic potential has been described as an inherent property of malignant tumors (Weigelt *et al.*, 2003; Weigelt *et al.*, 2005), the aim was to identify key transcriptional difference between lymph node-positive and lymph node-negative patients. A previously-described metagene (aggregate patterns of gene expression) study was able to predict lymph node status with an accuracy of 90% (Huang *et al.*, 2003) indicating the prognostic value of lymph node metastasis genes. Gene ontology and pathway analysis in our study did not identify promising GO functions and pathways affected, possibly due to smaller numbers of identified DE genes. Lymph node-positive associated DE genes were compared to survival/relapse gene lists generated by comparing genes involved in five year relapse, overall relapse and survival comparisons. SNIP (SNAP25-interacting protein) was significantly up-regulated (FC=2.51) in patients with lymph node-positive compared to lymph node-negative patients. The expression of SNIP was not detected in normal breast specimens but was detected in 37% breast cancer correlating with unfavourable overall survival and was published by this laboratory (Kennedy *et al.*, 2008). However, few DE genes were observed when the lymph node-positive and lymph node-negative patients were compared. In a similar study, gene expression signature was not strongly associated with lymph node status (Sotiriou *et al.*, 2003). This indicates that there might not a definite signature for Lymph node metastasis.

PHF21B (PHD finger protein 21B) gene was significantly down-regulated (FC = -6.15) in lymph node-positive specimens vs. lymph node-negative specimens. No information is known regarding its role in breast cancer.

In conclusion, our study identified a limited number of genes which are up-regulated in breast cancer patients with Lymph node involvement. SNIP was an important gene up-regulated in Lymph node-positive breast patients.

4.8 Tumour Grade

Grade is an important criterion in clinical decision making (section 1.3.1). Tumors with high Grade are more likely to undergo relapse and distant metastasis. Our results also indicate a strong association of grade with high relapse and poor survival. Hierarchical

clustering identified different groups of patients with differences in grade. The ER-negative enriched group (Cluster C) was enriched with high Grade tumors; however there was a subset of ER-negative enriched group (Cluster B) with low grade tumors (Grade 1 and Grade 2). The ER-positive enriched group (Cluster D and E) was enriched with low grade tumors (Grade 1 and Grade 2). Other independent studies have identified strong to moderate association of grade to gene expression profiles (Sotiriou *et al.*, 2003; Wang *et al.*, 2005; Calza *et al.*, 2006)

In our study, gene ontology and pathway analysis identified cytoskeleton, porin (are beta barrel proteins that cross a cellular membrane and act as a pore through which molecules can diffuse) activity, phosphatase inhibitor activity and cell division genes to be over-expressed and overrepresented in Grade 3 cancers compared to Grade 2 cancers. Genes involved in cell cycle progression and proliferation have also been identified up-regulated in high grade cancer compared to low and intermediate grade cancer in other published studies (Sotiriou *et al.*, 2003; Sotiriou *et al.*, 2006).

To find genes whose expression progressively increased with grade, comparisons were performed for genes up-regulated in Grade 2 vs. Grade 1 and Grade 3 vs. Grade 2. Twenty three transcripts were identified in which expression progressively increases with grade and two transcripts in which progression decreases with grade. IL4I1 (interleukin-four induced gene-1) expression was 5.52-fold up-regulated in Grade 2 vs. Grade 1 and 2.26-fold up-regulated in Grade 3 vs. Grade 2 (section table 3.1.8.7). Very little is known about the involvement of this gene in breast cancer although it has been found to be activated in primary mediastinal large B-cell lymphoma (Copie-Bergman *et al.*, 2003).

CCL5 (Chemokine (C-C motif) ligand 5) is an 8kDa protein classified as a chemotactic cytokine or chemokine and was found to be up-regulated with grade, with lowest expression in Grade 1 tumour and highest in Grade 3 tumour in our study. This gene is an inflammatory mediator and has pro-malignancy activities in breast cancer, and may have therapeutic potential (Soria and Ben-Baruch 2008). In stage II patients, the expression of CCL5 significantly increased the risk for disease progression (Yaal-Hahoshen *et al.*, 2006).

Our results indicate high expression of CDKN2A (cyclin-dependent kinase inhibitor) in high grade tumors, highest in Grade 3 cancers and lowest in Grade 1 cancers. CDKN2A encodes a protein that regulates 2 critical cell cycle regulatory pathways, the p53 pathway and the retinoblastoma pathway. CDKN2 germline mutations have been detected in patients with breast cancer and cutaneous melanoma (Monnerat *et al.*, 2007) and families with prevalent childhood cancer (Magnusson *et al.*, 2008).

In conclusion, our study identified important genes, functions and pathways for which the expression increases/decreases with grade. IL4I1 (interleukin-four induced gene-1) was identified as an important gene whose expression increases very highly with the progressive grade.

4.9 Tumour size

Tumour size is a significant predictor of relapse-free survival in breast cancer (Hu *et al.*, 2006). As the tumour size increases, the survival rate decreases (Carter, Allen and Henson 1989). Very few DE genes were observed in our study when tumour size was analysed; this suggests that tumour size might not be an important parameter as far as gene expression differences are concerned. Only three mRNAs RPESP (Ribulose-5-phosphate-3-epimerase-spondin), EPN3 (epsin 3) and CNTNAP2 (Contactin associated protein-like 2), were up-regulated by a fold change of greater than 2 in tumors with size greater than 2.8cm compared to tumors with size less than 2.8cm. Only seven mRNAs GRIA2 (Glutamate receptor, ionotropic, AMPA 2), PYDC1 (PYD (pyrin domain) containing 1), PTPRN2 (protein tyrosine phosphatase, receptor type, N polypeptide 2), CALML5 (Calmodulin-like 5), FOSB (FBJ murine osteosarcoma viral oncogene homolog B), SSFA2 (sperm specific antigen 2), and HSPB8 (heat shock 22kDa protein 8) were found to be down-regulated by a fold change of greater than 2 in tumors with size greater than 2.8cm compared to tumors with size less than 2.8cm. The results suggest that the gene expression profiles of large tumors are not very different from small tumors. Other studies also indicated that tumour size is not a very significant criterion as far as gene expression signature is concerned (Gieseg *et al.*, 2004; Sotiriou *et al.*, 2003) indicating that gene expression profile does not change with increase in tumour size.

In conclusion, our result identified genes up-regulated or down-regulated with increase in tumour size. RPESP, EPN3 and CNTNAP2 were up-regulated in large tumors. GRIA2, PYDC1, PTPRN2, CALML5, FOSB, SSFA2, and HSPB8 were down-regulated in large tumors.

4.10 Genes associated with relapse and survival

4.10.1 In-house study

Relapse-free survival is an important clinical parameter. Identifying genes with the prognostic importance could lead to discovery of better targeted therapies and also development of method to better classify patients for different kinds of treatment.

Relapse and survival events were studied in our in-house dataset in four distinct analyses: overall relapse, overall survival, 5 years relapse and 5 years survival. The genelists compared were patients who relapsed vs. patients who did not relapse; patients who relapsed within 5 years vs. patients who did not relapse within 5 years; patients who survived vs. patients who died of the disease and patients who survived for at least 5 years compared to patients who died within 5 years of diagnosis. The total number of common genes in all the four genelists was 384.

The up-regulated genes identified from the group of patients with a poor outcome outlined above (common to 4 genelist with average FC>2) were LCN2 (lipocalin 2), NMU (neuromedin U), SERPINB5 (serpin peptidase inhibitor, clade B ovalbumin, member 5), KCNG1 (potassium voltage-gated channel, KQT-like subfamily, member 1), SNIP (SNAP25-interacting protein), SLC4A11 (solute carrier family 4, sodium borate transporter, member 11), ST8SIA6 (ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 6), PSAT1 (phosphoserine aminotransferase 1), LOC92312, C9orf58, LOC92312, LOC92312, SOX11 (SRY sex determining region Y-box 11), PPARBP (mediator complex subunit 1), PCGF2 (polycomb group ring finger 2), ANP32E (acidic (leucine-rich) nuclear phosphoprotein 32 family, member E) and STARD3 (STAR-related lipid transfer domain containing 3).

Similarly up-regulated genes in the group of patients with a good outcome outlined above (common to 4 genelist with average FC>3) were FAM79B (tumor protein p63 regulated 1), RTN1 (reticulon 1), ZNF533 (zinc finger protein 385B), PDZK1 (PDZ domain containing 1), NOVA1 (neuro-oncological ventral antigen 1), SCUBE2 (signal peptide, CUB domain, EGF-like 2), RTN1 (reticulon 1). Additionally, ESR1 was up-regulated in the good outcome and lymph node-negative groups. A high expression of ESR1 is commonly associated with Luminal subtype A in microarray studies and correlates very well with better survival and disease-free outcomes (Hu *et al.*, 2006; Sorlie *et al.*, 2003). Our results are in agreement with other studies that show that high expression of ESR1 gene is associated with favourable outcome on breast cancer patients.

4.10.2 Meta-analysis

Four other similar datasets (section 3.2.1) which contained relapse status of individual patients were cross-compared to identify transcripts up and down-regulated among all the datasets. No transcript was found common to all the datasets. It was therefore decided to identify transcripts up-regulated or down-regulated in three or more datasets. Twenty two transcripts were found to be DE in a minimum of three out of the five datasets under study. Three of them (HSPB1, KIAA0101 and PAK3) were up-regulated in four out of the five datasets and one (FOS) was down-regulated in four out of the five datasets.

HSPB1 (heat shock 27kDa protein 1) is a 27 kDa heat shock protein and plays a role in cancer progression (Garrido *et al.*, 2006). In four of the five experimental cohorts, this gene was found to be significantly up-regulated in patients who relapsed. Down-regulation of HSPB1 in HCT116 human colon carcinoma cells caused senescence in a population of cells (O'Callaghan-Sunol, Gabai and Sherman 2007). In the 4T1 murine breast adenocarcinoma cell line, knockdown of this gene eliminates cell motility (Bausero *et al.*, 2006). HSPB1 expression in human breast cancer cells can reduce Herceptin susceptibility by increasing Her2 protein stability (Kang *et al.*, 2008).

KIAA0101 (protein-coding GC15M062444) is a proliferating cell nuclear antigen-associated (PCNA) factor and involved in cell proliferation. A high expression of this gene was found in plasma RNA of colorectal cancer (Collado *et al.*, 2007). siRNA

knockdown of KIAA0101 in pancreatic cancer cells caused a reduction in proliferation as well as a significant decrease in DNA replication (Hosokawa *et al.*, 2007). In hepatocellular carcinoma, a high expression of this gene was associated with increased stage, early tumour recurrence, and poor prognosis (Yuan *et al.*, 2007). In our meta-analysis comparing patients who relapsed vs. patients who did not relapse, this gene was found to be significantly up-regulated in patients who relapsed in four of the five experimental groups.

PAK proteins are critical effectors that link Rho GTPases to cytoskeleton reorganization and nuclear signaling. The PAK3 gene has a significant role in the nervous system and mutation of this gene is involved in many diseases of the Central Nervous System (Boda *et al.*, 2004). PAK3 (p21 protein (Cdc42/Rac)-activated kinase 3) contributes to synapse formation and plasticity in the hippocampus (Boda *et al.*, 2004). PAK3 mutations result in a specific form of X-linked mental retardation with fairly constant clinical features (Rejeb *et al.*, 2008). However, there is limited understanding of this gene's involvement in cancer. Our results show that the expression of this gene is significantly up-regulated in patients who relapsed in four of the five independent studies. PAK1 phosphorylates histone H3 and affects the Pak1-histone H3 pathway and mitotic events in breast cancer cells (Li *et al.*, 2002). PAK1 induces RAS transformation and that is essential for RAS-induced up-regulation of cyclin D1 during the G1 to S transition (Nheu *et al.*, 2004).

Members of the FOS family (c-FOS, FOSB and its smaller splice variants, Fra-1 and Fra-2) dimerise with Jun proteins to form the AP-1 transcription factor complex. Our study identified two members of AP-1 transcriptional factor (FOS and FOSB) to be down-regulated in breast cancer patients who relapsed. FOS was down-regulated in four of the five experimental groups and FOSB was down-regulated in three of the five experimental groups. FOSB expression is necessary for normal proliferation and differentiation of mammary epithelial cells, and reduced FOSB protein levels in tumors has been found to be correlated with high grading, ER-negative and PR-negative, and high HER2/neu expression (Milde-Langosch *et al.*, 2003). A previous study (Milde-Langosch *et al.*, 2004) has reported that high FOSB levels are associated with high expression of MMP1,

MMP9, PAI-1 and uPAR protein in clinical models and over-expression of the gene increased invasion in the MCF-7 cell line.

Meta-analysis for prognostic makers across various studies identify different sets of signatures and are mostly linked to proliferation This is because various mechanisms of cancer progression e.g ER+, ERBB2, have effects on increased cellular proliferation (Wirapati et al. 2008). The KIAA0101 gene was common to our meta-analysis and the meta-analysis performed by this group, with KIAA0101 up-regulated in the poor survival group in both studies.

In conclusion, our results identified many genes up and down-regulated in aggressive disease. HSPB1, KIAA0101 and PAK3 were up-regulated in patients who relapsed. AP1 transcriptional factor genes FOS and FOSB were down-regulated in patients who relapsed.

4.10.3 Comparison our in-house result with OncotypeDx

The in-house genelist comparing patients who relapsed vs. patients who did not relapse was compared with the expression of the 16 genes of OncotypeDx (Paik *et al.*, 2004), a diagnostic kit used to assay the long term survival and the possible benefit from chemotherapy. AURKA, BIRC5 ERBB2, were up-regulated in both the studies in the bad prognosis group (patients who relapsed in our study and positive association with recurrence on OncotypeDX) and ESR1, SCUBE2 were up-regulated in both the studies among the good prognosis group (patients who did not relapse in our study and negative association with recurrence). In our study, ERBB2 was also up-regulated in lymph node-positive patients and ESR1 was down-regulated in lymph node-positive patients. Of the 6 genes common to our analysis, all followed the same trend and were either associated with good or bad prognosis (section 3.2.2) indicating good agreement among the two results.

In conclusion, 37.5% of the genes were common to both the studies. Of the 6 genes common to both studies all followed the same trend of expression indicating good agreement among the two results.

4.10.4 Comparison our in-house result with MammaPrint

The relapse genelist above was also compared with that of MammaPrint (van 't Veer *et al.*, 2002). NMU (neuromedin U), GMPS (guanine monphosphate synthetase), MELK (maternal embryonic leucine zipper kinase) were up-regulated in both the studies (patients who relapsed in our study and positive association with metastasis on the MammaPrint genelist) and PEGI (peroxisomal D3, D2-enoyl-CoA isomerase), SCUBE2 (signal peptide, CUB domain, EGF-like 2) was down-regulated in both the studies (patients who did not relapse in our study and negative association with metastasis on the MammaPrint genelist) (section 3.2.3). No common genes were found when comparing with lymph node status with the genelist of MammaPrint. This may be due to differences in microarray platforms and differences in the sample and the way the specimens were selected. Also, van 't Veer *et al.*, (2002) identified genes responsible for distant metastasis, while in our study relapse and lymph node metastasis were used as prognostic criteria. Similar conclusions were also drawn by Wang *et al.*, (2005) as there was an overlap of only 3 genes among their genelist and MammaPrint genes.

In conclusion, there were not many genes common to both studies. Of the few genes common to both genelist, they followed the same trend of expression. Our study does not compare very well with the van 't Veer *et al.*, (2002) study due to the differences in the platform and clinical parameters used. In our study the clinical parameter was Relapse, whereas the clinical parameter used in the van 't Veer *et al.*, (2002) study was distant metastasis.

4.11 Relapse prediction

One of the many applications of microarrays is their potential use as prognostic/diagnostic kits. The success of these kits depends on the accuracy with which they can predict the prognosis of the patients. Many such algorithms can be used for such types of classification. K-nearest neighbour (KNN) considers all specimens in an m-dimensional space, where m is the number of variables and uses a distance metric to group them based on similarity or dissimilarity (Gregory *et al.*, 2008). Probabilistic neural networks belong to the family of Radial Basic Function (RBF) networks. The

algorithm is very similar to a feed-forward neural network with one hidden neuron. The input is directly passed to the hidden layer without weights and a Gaussian density function is used as an activation function. The interconnecting weights are optimized using a least square optimization algorithm (Haykin, 1998). Linear discriminate analysis LDA classifies the data using the linear combination of features which best separate two or more classes of objects (Geoffrey, 1992). Support vector machines (SVM) construct a hyper plane in space so as to maximally divide the margin between the different types of objects (Haykin, 1998). Ensemble classification methods works on combining different classification methods to improve the classification accuracy. The challenge in such types of classifier is to analyse the results coming from different classifier to get the optimum results.

A back propagation algorithm was implemented for accurate prediction of relapse in breast cancer patients. One hundred and sixty two differentially-regulated genes ($p \leq 0.001$) were identified among patients who relapsed and patients who did not relapse. These genes were used for training the network. To test the accuracy of the system, leave-one-out cross validation was used. The model predicted relapse to an accuracy of 97.87% (100% for patients who relapsed and 96.2% for the patients who did not relapse) with a cut-off of 0.75 for positive examples and 0.25 for negative examples i.e patients with score above 0.75 were considered to relapse and patients with score below 0.25 were considered not to relapse. Patients whose score fell between 0.25 and 0.75 were classified as undetermined. However, there were 11 specimens out of total of 104 which could not be classified. When all samples were classified, the accuracy was 93.33% (93.7% for patients who relapsed and 92.9% for the patients who did not relapse). Support vector machine (SVM) analysis was also used to classify the same data. SVM classified the relapse event with an accuracy of 93.33% which is same as that of back propagation, when all samples were classified.

A classifier using a 70 gene signature was able to predict distant metastasis with an accuracy of 83% in node-negative breast cancer (van 't Veer *et al.*, 2002). This was later developed as the MammaPrint kit for detection of distant metastasis for lymph node-negative breast cancer patients under 61 years of age with tumors of less than 5cm. An

independent classifier using back propagation and support vector machines was constructed using van 't Veer *et al.*, (2002) data. Comparing patients who developed distant metastasis and patients who did not, a total of 117 DE genes ($p \leq 0.001$) was used to generate the classification model. In leave-one-out cross validation, a back propagating algorithm was able to predict distant metastasis with an accuracy of 89.70% (82.7% for patients who developed distant metastasis within 5 years and 94.8% for the patients who did not) when a cut-off score of 0.75 for positive examples and a cut-off score of 0.25 for negative examples was taken. This resulted in 10 samples out of 78 samples as unclassified. However, when all samples were classified, the prediction accuracy was 87.17% (82.3% for patients who developed distant metastasis within 5 years and 90% for the patients who did not). Similar analysis using support vector machines had a prediction accuracy of 82.05% in leave-one-out cross validation model.

Many other studies have used gene expression data to develop prognostic models using a variety of gene selection techniques and classification algorithms. Huang *et al.*, (2003) developed a probability-based classifier using metagene to predict lymph node metastasis and recurrence with an accuracy of 90%. Oncotype DX is a PCR based diagnostic kit to calculate the recurrence score for individual patients who have no lymph node involved and are ER-positive (Paik *et al.*, 2004; Sparano and Paik 2008). This kit is based on the expression of 16 genes and 5 control genes using PCR. The output of the analysis is a recurrence score. The higher the recurrence score, the higher is the probability of recurrence of the disease. Based on these scores, patients are classified as low risk, intermediate risk and high risk with the recurrence rate of 6.8, 14.3 and 30.5% respectively.

Karlsson *et al.*, (2008) studied 46 node-negative tumors with the aim of developing a classifier to distinguish high risk and low risk breast cancer patients. A t-test was used to identify DE genes among the high and low risk patients. A total of 51 genes ($p < 0.001$) were used to build a voting feature interval classifier and correlation based classifier with a prediction accuracy of 96 and 89% respectively in leave-one-out cross validation study.

Our results based on the comparative analysis on same data indicate that back-propagation is an excellent method of developing a classification model on both Affymetrix and cDNA microarray data and can outperform Support Vector Machines based classifiers.

The gene signature along with the back propagation training algorithm has a potential to be developed as a diagnostic assay. The models performed better than the existing diagnostic kit such as MammaPrint and OncotypeDx. The analysis also indicates that the gene signature generated and used in our study is more informative because of the high accuracy obtained. This may be due to the fact that the chips utilised to generate this data were Affymetrix whole-genome U133 Plus2.0 chips, which contain vastly more transcript data than was used to identify the other two gene signatures.

In conclusion, the study presented here demonstrates the suitability of back-propagation algorithm as an efficient classifier for gene expression data with potential of its use for diagnostic/prognostic kits. Using this classifier, we were successful in predicting relapse with an accuracy of 97.87%.

4.12 Identification and functional validation of Ropporin

Our in-house microarray data indicated that Ropporin is over-expressed in patients who relapsed (overall), relapsed within 5 years and did not survive beyond 5 years. In the early stages of annotating the transcript, it was realised that there are two genes, ROPN1 and ROPN1B with a very high homology (97% on DNA sequence and 95% on protein sequence). Both genes are located on chromosome 3 (ROPN1: 3q21.1; ROPN1B: 3q21.2). Because of their close proximity and sequence homology, it is likely that one arose from the other by duplication followed by random mutations during the evolutionary process.

Ropporin expression was first detected in testis and is localized in the principal piece and the end piece of sperm flagella and is induced at late stage of spermatogenesis (Fujita *et al.*, 2000). Rhophilin protein is localised in the outer surface of the outer dense fibre of sperm. Ropporin is localized in the inner surface of fibrous sheath of sperm. Rhophilin

and Ropporin together interact with small GTPase Rho which acts as a molecular switch that regulates various cellular processes such as cell adhesion, motility, gene expression and cytokinesis (Fujita *et al.*, 2000).

Sperm motility and acrosome (process at the anterior end of a sperm cell that produces enzymes to facilitate penetration of the egg) is dependent on actin polymerization where AKAP (A-kinase anchor protein) and RHOA (Ras homolog gene family, member A) interacting proteins play an important role. The phosphorylation of AKAP3 increases its interaction with RHOA-interacting proteins and Ropporin (Fiedler, Bajpai and Carr 2008). Mutants lacking RSP11 (radial spoke protein), an ortholog of Ropporin in the flagellum of *Chlamydomonal reinhardtii* demonstrated impaired and sporadic motility (Yang and Yang 2006).

Ropporin shares sequence similarity with three other proteins, ASP, SP17 and CABYR, all of which are localised in sperm flagella. All of them contain a highly conserved dimerization/docking (R2D2) domain, suggesting that all of these proteins interact with all AKAPs. All of these proteins are also expressed in motile cilia indicating that these proteins are vital for sperm and cilia (Newell *et al.*, 2008).

However, its expression has recently been detected in multiple myeloma, chronic lymphocytic leukaemia and acute myeloid leukaemia (Li *et al.*, 2007b). Ropporin gene expression in tumour cells is associated with the high titer IgG antibodies against Ropporin. Because of its restricted expression in normal tissue and immunogenicity of the protein to the autologous hosts, this molecule may be a good target for immunotherapy (Li *et al.*, 2007a).

4.12.1 Affymetrix probe annotation for Ropporin

On the Affymetrix U133 Plus2.0 chip, there are a total of four transcripts for Ropporin. According to the Affymetrix guidelines, 233203_at is specific probes for ROPN1 while the other three (224191_x_at, 231535_x_at, 220425_x_at) are non-specific probes. The

“_x_at” probe sets contain some probes that are identical, or highly similar, to unrelated sequences. These probes may cross-hybridize with sequences other than the main target. The sequence of each probe was obtained from NetAffx and annotated using BLAST. Recently a reference sequence identifier has been provided for ROPN1 and ROPN1B. RefSeq annotation integrates information from various sources, and represents a consensus description of the sequence and its features (Pruitt, Tatusova and Maglott 2007). In the previous Genbank build there was no reference sequence available for these two genes and only early accession numbers were allocated to sequences. The Affymetrix U133 Plus2.0 chip was designed on the earlier Genbank build and therefore the probeset was designed on early submission of this gene which was poorly defined and annotated. The BLAST result of 233203_at did not have a hit with the reference sequence of either ROPN1 or ROPN1B. This indicated that the probe sequence of 233203_at may not represent Ropporin. However, it did have a hit with ROPN1 sequence gi|6599263| (not the reference sequence) indicating the possibility of more isoforms of Ropporin. Early on in our analysis, the assumption was made that only ROPN1B, and not ROPN1 is expressed in our in-house study. This was because the NetAffx-defined unique probeset for ROPN1 (233203_at) did not report any expression from the microarrays, while the non-specific ROPN1B probeset (220425_x_at) did yield a reproducible signal intensity. Following the outcome that 233203_at cannot be considered as a valid probe for ROPN1, two other probes were selected to represent ROPN1 and distinguish it from ROPN1B. A previous study, (Gautier *et al.*, 2004) analysed U133A chip annotation and documented that 64% of the Affymetrix annotation has discrepancies with current annotation due to the fact that while the probes on Affymetrix arrays remain the same for several years, the biological knowledge concerning the genomic sequences keeps changing. The BLAST result on the sequence of 231535_x_at indicated this probe to be a 100% match to the reference sequence of ROPN1 and 97% match to the reference sequence of ROPN1B. Therefore 231535_x_at was used to represent ROPN1. A BLAST result on the sequence of 224191_x_at indicated this probe to be a 100% match to the reference sequence of ROPN1B and 99% match to the reference sequence of ROPN1. Therefore, probeset 224191_x_at was not used to represent either ROPN1 or ROPN1B. The BLAST result on the sequence of 220425_x_at indicated this probe to be a 100% match to the reference

sequence of ROPN1B and 97% match to the reference sequence of ROPN1. Therefore, probeset 220425_x_at was used to represent ROPN1B. On analysing our in-house datasets, it was observed that there was no expression of 233203_at in any of the samples, however, there was a substantial amount of expression of the other probesets. The same was true for a number of the publicly available breast and melanoma datasets. However, for multiple myeloma, all the probe sets (including 233203_at) showed varying levels of expression in many of the samples indicating a possibility of more isoforms of this gene.

For the microarray analysis, probeset 231535_x_at was used to represent ROPN1 and 220425_x_at to represent ROPN1B. However, there was no way to distinguish how much each one cross-hybridises to their respective variants.

4.12.2 Ropporin expression in our in-house breast dataset.

ROPN1 was 4.97-fold and ROPN1B was 5.06-fold up-regulated in patients who relapsed compared to patients who did not relapse. ROPN1 was 6.81-fold and ROPN1B was 7.83-fold up-regulated in patients who relapsed within 5 years compared to those who remained disease free for 5 years. ROPN1B (220425_x_at) was not expressed among 54/57 (94.7%) of the patients who did not relapse, however it was expressed in 13/48 (27.1%) of the patients who did relapse based on the cut-off of 100 Affymetrix unit. Similarly, ROPN1 (231535_x_at) was not expressed among 53/57 (92.9%) of the patients who did not relapse, however, it was expressed in 14/48 (29.1%) of the patients who did relapse.

A high expression of ROPN1 and ROPN1B was observed in one of the sub-clusters enriched with ER-negative specimens. This cluster had the worst survival in comparison to other clusters.

4.12.3 Confirmation of Ropporin expression by qPCR

Using qRT-PCR on the clinical breast specimens, it was possible to confirm the high expression of Ropporin in the breast specimens used to generate our in-house dataset. Primers for both genes were designed and tested for their specificity in detecting the two

genes using plasmids with the ROPN1 and ROPN1B genes cloned in. While ROPN1 primers were very specific in detecting ROPN1 gene, ROPN1B picked up ROPN1 with 100-fold less specificity than that of ROPN1B. Therefore, it was possible to distinguish between the two genes, once both qRT-PCR reactions were run on all assayed samples. qRT-PCR was performed on 94 of the clinical specimens from our in-house study using the specific primers for ROPN1 and ROPN1B. There was no RNA available for the rest of the clinical specimens. ROPN1 was found to be 2.33-fold (baseline mean 2.39; experimental mean 1.81) down-regulated in patients who relapsed, whereas ROPN1B was found to be 6.28-fold (baseline mean 1.02; experimental mean 11.37) up-regulated in patients who relapsed. ROPN1B qRT-PCR results are in agreement to the results from microarray result, however, the results from ROPN1 qRT-PCR contradicts the findings from our microarray study. With no way to predict or estimate the amount of cross-hybridization of the Affymetrix probes, the results from qRT-PCR were considered to be a more accurate assessment of ROPN1 and ROPN1B expression in these samples.

Previous studies have not tried to distinguish between the two genes (both usually referred to as Ropporin). The studies which do mention the ROPN1 gene do not seem to be specific for ROPN1 (Carr *et al.*, 2001; Li *et al.*, 2007; Newell *et al.*, 2008). PCR primers from a study of Newell *et al.*, (2008) were blasted using primer BLAST. The primers picked up the ROPN1B gene instead of the ROPN1 gene. Performing the same analysis on primers from two other studies (Carr *et al.*, 2001; Li *et al.*, 2007) picked up both genes. The obvious reason for these differences is the constantly evolving annotation of Genbank. However, our primers were specific in picking up the differences among the two genes. The primer BLAST results were highly specific in picking up their respective genes (results based only on forward and reverse primer).

4.12.4 Functional validation using *in-vitro* cell line models

Since the gene is expressed in sperm tail and cilia, both of them involved in motility, a hypothesis was made that in cancer it might be helping the cancer cells in moving from their primary site (via invasion) to a different location (metastasis).

Cancer cell line models were used to identify the functional role of this gene in cancer. With the hypothesis that this gene might play a crucial role in cancer cell motility and invasion; siRNA and over-expression studies were performed and followed by functional assays (motility and invasion) to investigate any possible association between this gene and invasion/cell motility. siRNA knockdown of ROPN1B in MDA-MB-435s showed a reduction in motility. siRNA knockdown of ROPN1 was not performed as expression of ROPN1 was substantially less than that of ROPN1B. No invasion assay results were available as the cell line did not demonstrated reproducible invasion. Similarly siRNA knockdown of ROPN1 and ROPN1B in M14 showed a decrease in invasion and motility. However, knockdown by ROPN1 siRNA was not very specific and based on the qRT-PCR results, it knocked down the ROPN1B gene too. Thus the results could not positively associate ROPN1 to cell motility and invasion, however, as a whole (ROPN1 and ROPN1B) can be positively associated with cancer cell motility.

Results from over-expression studies were somewhat inconclusive. In M14 cells, over-expression of ROPN1 and ROPN1B led to reduction in the protein content; with a consequent loss observed in invasion and motility. In MDA-MB-435s cells over-expression of ROPN1 showed no increase/reduction in protein level, however there was reduction in motility. Cells over-expressing ROPN1B cDNA showed a reduction in protein level surprisingly and a reduction in motility.

4.12.5 Ropporin expression in cancers and normal tissues

The Ropporin gene is classified as cancer testis genes because of its expression in testis and not in other tissues, but with aberrant expression in cancers (Li *et al.*, 2007). Cancer testis (CT) genes encode a heterogeneous group of immunogenic proteins (CT antigens) expressed almost exclusively in normal testis and in a percentage of tumors of various origin. On the basis of their tissue specificity and immunogenicity to its autologous host, CT antigens are considered promising targets for development of cancer vaccines (Simpson *et al.*, 2005). Scanlan, Simpson and Old (2004), identified 44 CT gene families and studied their expression pattern in numerous cancer types. Bladder cancer, non-small cell lung cancer, and melanoma had high CT gene expression, breast and prostate cancer

had moderate CT gene expression while renal and colon cancer had low CT gene expression. CT gene expression was also observed among multiple myeloma (Condomines *et al.*, 2007) and oesophageal carcinoma (Liang *et al.*, 2005).

In normal tissue, Ropporin is expressed in testis (Fiedler, Bajpai and Carr 2008, Fujita *et al.*, 2000; Li *et al.*, 2007; Newell *et al.*, 2008), fetal liver (Li *et al.*, 2007), motile cilia, liver, brain, pancreas and prostate (Newell *et al.*, 2008). Our analysis on publicly available datasets (GSE1133) confirmed the high expression of Ropporin in testis and marginal expression in brain and liver. Additionally, a high expression of this gene was found in ganglion and marginal expression found in skin, trachea and heart. High expression of Ropporin is also found in epithelial cells with motile cilia and this may be the reason for its detection in other tissue types with presence of motile cilia (Newell *et al.*, 2008).

Ropporin is detected in tumors of multiple myeloma (Li *et al.*, 2007; Chiriva *et al.*, 2007), chronic lymphocytic leukaemia and acute myeloid leukaemia (Li *et al.*, 2007). Our analysis on a publicly available dataset confirmed that Ropporin is widely expressed in multiple myeloma. Ropporin expression was also present in normal melanocyte (from normal skin) and the expression dramatically increases with the progress of melanoma; highest in metastatic growth phase melanoma and lymph node metastasis.

4.12.6 Previous studies identifying Ropporin expression in breast cancer.

Expression of Ropporin was also found in other publicly available breast datasets with aberrant expression observed in estrogen-negative breast tumors. Ropporin expression has previously been shown to be correlated with GABA π expression which is associated with undifferentiated cell type and high grade of breast cancer (Symmans *et al.*, 2005). Ropporin expression was also observed high in patients with breast cancer which developed bone metastasis (Smid *et al.*, 2006). However, while these studies have previously linked Ropporin expression with breast cancer and metastasis, the gene was listed in these publications together with several other potentially important targets and was not specifically highlighted. This study is the first to functionally demonstrate a role for this gene in invasion in breast cancer. The identification of Ropporin as up-regulated

in these two studies has served to complement and strengthen our findings that the gene is actively involved in aggressive breast cancer.

4.13 Conclusion

Our study on clinical breast specimens and normal breast specimens has provided a deep insight to the biology of breast cancer. Our results identified groups of patients with similar expression profiles, the possible biology driving them and the subsequent clinical implications for those patients.

Two unique groups of patients, previously un-identified by other studies with significant differences in survival were identified. A “good” prognosis group with a high expression of immune response-associated genes was demonstrated among the ER-negative group of patients. A group of patients with ER-negative tumors associated with a very poor prognosis has been shown to express high levels of the Ropporin gene. Over-expression of this gene was also observed in patients who relapsed vs. not. Using cell lines models, this study positively identified the involvement of ROPN1B in breast and melanoma cancer cell motility and invasion. The results also indicate that ROPN1 has a similar function, but because of the absence of very specific siRNA, this could not be proven.

A prognostically-important genelist was used to develop a Neural Network back propagation model to predict the clinical outcomes. Using an identified set of 162 genes, the model was successful in predicting relapse with an accuracy of 97.8%.

Comparing the gene expression profiles of Normal and Cancer specimens identified genes, functions and pathway differences associated with disease. TP53, along with cell cycle genes were up-regulated in cancer compared to normal specimens. Embryonic stem cell pathway genes were up-regulated in tumors indicating the possibility of impaired stem cell as origin of cancer. The fatty acid biosynthesis pathway was down-regulated in tumour vs. normal specimens.

To get a deeper understanding of ER involvement in breast cancer and to mine genes which may play an important role in the ER metabolism, meta-analysis was performed on

an in-house dataset together with 5 public datasets. This analysis identified novel genes which had not been associated with the ER pathway. The nuclear receptor pathway was up-regulated in ER-positive tumors/cell lines. Mining for ESR1-correlated genes across a 5897-member microarray chip dataset identified FOXA1, SPDEF, C1ORF34 and GATA3 expression to be highly correlated with ESR1. Our results also indicated that most of them are expressed together; however, individual expression can occur independently (except for SPDEF expression which seems likely to be dependent on FOXA1).

4.13 Discussion of some peripheral research projects

In the course of the PhD project, I became involved in two projects which were somewhat peripheral to the main thesis work. One involved a bioinformatics analysis of publicly available data sets, to evaluate how relevant cell line models might be to human tumors *in vivo*. The second involved an opportunity to take part in bioinformatics analysis of a unique data set on microarray analysis of basal cell carcinoma vs. normal skin.

4.13.1 How Representative are cell line models of clinical conditions?

Cell lines are widely used as models of *in-vivo* systems. However, limited studies have been done to establish whether these models accurately reflect *in-vivo* scenarios. A separate study carried out in our laboratory examined gene expression differences and similarities in a representative group of breast cancer cell lines and clinical specimens to estimate their approximate level of similarity and was published previously (Mehta *et al.*, 2007).

Cell lines grow under very tight and well-optimized conditions, with enough space to grow and divide. In comparison, tumors grow in a completely different environment and are influenced by a varied range of conditions. In this study, a clear segregation of the cell lines and clinical specimens by hierarchical clustering was demonstrated. This is in agreement with other similar studies where cell lines and clinical specimens tend to cluster separately from each other (Dairkee *et al.*, 2004; Ross and Perou 2001). PCA also

demonstrated a clear separation of the two groups. A segregation of the clinical specimens into two smaller sub-groups was also observed, although the clinical/biological basis for this has not been determined here. An earlier experiment (Dairkee *et al.*, 2004) also reported considerable scatter among primary tumour cultures and cell lines compared to normal breast specimens using PCA as a comparison tool.

From the Genmapp analysis, cell cycle, mitosis, nuclear division, cell proliferation and other related functions are over-represented in cell line models in comparison to the clinical specimens, while functions related to immune response and defence response are over-represented in clinical specimens relative to cell lines. A recent study (Ertel *et al.*, 2006), also reported that genes related to proliferation and cell cycle are over-represented in cell lines relative to clinical specimens, while cell communication, cell adhesion molecules and ECM-receptor interaction are down-regulated in cell lines compared to clinical specimens. Our study also indicated a decrease in expression of genes involved in cell adhesion in the cell lines compared to clinical specimens, although this data did not make it into the top ten ontologies.

While the analysis outlined above identified the macroscopic broad-based differences between breast cancer cell lines and clinical specimens, it was considered useful to assess the similarity relationships of the cell lines and clinical specimens with regard to their ER status. It was hoped that while differences had been observed when comparing cell lines and clinical specimens directly, both cell lines and clinical specimens would cluster similarly when ER status was used as the criteria. Previous studies had demonstrated that both cell lines (Charafe-Jauffret *et al.*, 2006) and clinical specimens (Sotiriou *et al.*, 2003) cluster largely on their ER status. To this end, unsupervised clustering of the cell lines and clinical specimens separately was carried out to determine if either group clustered according to ER status. However, while the cell lines largely clustered according to ER status, the clinical samples did not. This result indicated that, even at a single parameter scale, the differences between clinical specimens and their respective cell line models may remain considerable.

In conclusion, the findings reported here indicate that significant differences in gene expression between clinical conditions and their respective cell line models exist at both the large- and small-scale levels. A previous study (Dairkee *et al.*, 2004) concluded that the results obtained from cell lines may act as good models for high-grade cancer, but may fail as useful models for most of the low- and medium-grade breast cancers. While our study does not indicate a specific clinical classification for which such cell line data may prove relevant, the data presented here demonstrate that these differences should be taken into account when extrapolating *in-vitro* cell line results to clinically-relevant *in-vivo* systems.

4.13.2 Basal cell carcinoma

Basal cell carcinoma (BCC) is the most common skin cancer in humans. It is locally aggressive, invasive but rarely metastasises (Saldanha *et al.*, 2004; Ionescu, Arida and Jukic 2006). Very few studies aimed at investigating the molecular mechanisms associated with BCCs have been published worldwide. Howell *et al.*, (2005) analyzed 50 BCC tumour specimens using cDNA microarrays and reported findings from their analysis of 1,718 transcripts. A separate study carried out in our laboratory analyzed gene expression of BCCs, compared to normal skin, using whole genome microarrays. Following extensive analysis of our data, a number of novel potential biomarkers/therapeutic targets for this disease were identified (O'Driscoll *et al.*, 2006).

In agreement with our findings, Howell *et al.*, (2005) also reported gene transcripts including collagens (type V, alpha 1 & alpha 2; type IV alpha 1 & 2; type VII alpha 1), topoisomerase II α , tumour-associated calcium signal transducer 1, profilin 2, calretinin, syndecan 2, and v-myc to be up-regulated in BCC compared to normal skin. Similarity was also observed between these two studies for transcripts down-regulated in BCCs compared to normal specimens. Examples of these include cystatin B, acetyl-Coenzyme acyltransferase 1, 3-hydroxy-3-methylglutaryl-Coenzyme A reductase, glutaredoxin, amyloid β (A4) precursor-like protein and cytochrome b-5. ADP-ribosylation factor 3 was down-regulated by 1.67-fold in our study, but was up-regulated in the study by Howell *et al.*, (2005). Glia maturation factor β was 1.42-fold up-regulated in our analysis

but Howell *et al.*, (2005) reported it as down-regulated. These conflicting results may be due to different splice variants of these transcripts being detected by cDNA compared to oligo microarrays. It was also noted that the results that differed between our study and that of Howell *et al.*, (2005) were generally transcripts <2-fold differentially-expressed between BCC and normal skin. Further comparisons between these studies cannot be performed as the fold change was not reported by Howell *et al.*, (2005) and no information is publicly available on transcripts that were present on their microarray.

Dys-regulation of the hedgehog and Wnt pathways is associated with the development of BCC (Rubin, Chen and Ratner 2005, Daya-Grosjean and Couve-Privat 2005). A tumour suppressor gene, patched homologue 1 (PTCH1) forms a part of the hedgehog signaling network (Cohen 2003) is found to be associated with the development of BCC (Boonchai *et al.*, 2000). Eleven-fold up-regulation of PTCH1, 7.39-fold up-regulation of gli2 and no significant change in shh expression levels was observed from our analysis of BCC compared to normal skin tissue. SMO (smoothed homolog) gene is associated with Hedgehog signalling heterotrimeric G proteins (Philipp and Caron 2009). GLI gene encodes a nuclear protein and binds to specific genes leading to transcriptional activity (Kinzler and Vogelstein 1990). The mechanism of action of PTCH1 is via binding to another transmembrane molecule smoothed (SMO) thereby suppressing intracellular signaling. Then sonic hedgehog (shh) binds to PTCH1 resulting in an uninterrupted signal transduction by SMO, via GLI transcription factors and subsequent activation of target genes, including members of the Wnt pathway (Yamazaki *et al.*, 2001) and PTCH1 (Cohen 2003). SMO is a protein with seven transmembrane domains that is distantly related to G-protein coupled receptors (GPCRs) (Ingham and McMahon 2001). Activated SMO stimulates transcription factors of the Cubitus interruptus (Ci) or GLI family inducing the expression of specific genes (King 2002). GLI transcription factors belongs to the Kruppel family of zinc finger proteins (Buscher and Ruther 1998).

Increased PTCH1 mRNA levels have previously been reported in nodular BCC but undetectable in superficial BCC (Tojo *et al.*, 1999); however detectable PTCH1 in both types of BCC was observed, with no significant difference in their respective expression values (t-test: $p = 0.637$).

PTCH1 is associated with tumour suppressor activity (Cohen 2003) and was found to be up-regulated in BCC compared to normal skin. The lack of tumour suppressor activity by PTCH1 may be due to lack of expression of its corresponding protein and/or lack of binding to SMO (not significantly different between BCC and normal skin). As PTCH1 is found to shuttle between the cell membrane and endocytotic vesicles in response to active hedgehog ligand, it is obvious that the expression of both mRNA and protein (at the relevant location, binding of SMO) is necessary to exert its tumour suppressor activity (Cohen 2003).

Wnt signaling may be able to regulate a number of the aspects of the biology of tumour cells and thus contribute in several ways to the tumour phenotypes including proliferation. In our study there was significantly increased expression of a number of Wnt family members including Wnt5A (3.35-fold), in agreement with a study by Saldanha *et al.*, (2004) where Wnt5A levels were increased in BCCs compared to surrounding skin; and Wnt6 (4.86-fold). Increased levels of Wnt ligand binding receptors, Frizzled D2 (8.94-fold), D7 (2.31-fold), and D8 (5.89-fold) and decreased levels of D4 (-2.78-fold), were also found.

Jun is a transcription factor involved in the Wnt pathway (Weeraratna 2005) and was found to be increased (2.34-fold) in BCCs compared to normal skin. Transcription factor associated with cancer including CHES1 (checkpoint suppressor 1) is involved in repressing expression of genes important for tumorigenesis (Scott and Plon 2005) and was differential-expressed in this study. CHES1 mRNA has been reported as down-regulated in oral squamous cell carcinoma (Chang *et al.*, 2005) and in hepatocellular carcinoma (Hong, Muller and Lai 2003). CHES1 mRNA levels were found to be significantly (-2.03-fold) down-regulated in BCC compared to normal skin. mRNAs involved in inducing apoptosis were also found to be down-regulated including CIDE and CARD15 which are 4.18-fold and 2.31-fold down-regulated in BCC compared to normal skin.

Increased levels of ChgA in serum have been associated with poor prognosis/shortened survival for prostate cancer patients (Ranno *et al.*, 2006). ChgA protein levels have been

proposed to assist in the diagnosis of Merkel cell carcinoma patients who may benefit from oncological therapy (Koljonen *et al.*, 2005; Mount and Taatjes 1994; Carlei *et al.*, 1986). In this study, ChgA levels were found to be significantly (130.3-fold) up-regulated in BCCs compared to extremely low levels in normal skin specimens.

In summary, our analysis has identified important genes, functions and pathways involved in normal skin transition to basal cell carcinoma. Wnt signaling pathway was found to be up-regulated in Basal cell carcinoma and may be potentially involved in transition of normal skin to basal cell carcinoma.

5.0 Summary and Conclusions

5.1 Hierarchical clustering analysis identified clinical heterogeneity in breast cancer

To understand the clinical heterogeneity of breast cancer, two-way clustering analysis of the samples was performed. Various groups of samples and their association with various clinical parameters were identified. The important findings are summarised below.

- The gene expression patterns of Normal specimens are very homogenous, whereas the gene expression patterns of breast tumors is highly heterogeneous.
- Group of breast cancer (mainly ER-negative) tumour specimens exists whose expression pattern is closer to normal specimens than to most of the tumors.
- Two ER-positive clusters were identified, one with low ER partner gene expression and the other with high ER partner gene expression. The cluster with high ER partner gene expression had a marginally better survival than the cluster with relatively low ER partner gene expression.
- An ER-negative enriched cluster was identified. This cluster was highly heterogeneous. There were three distinct sub-clusters in this cluster. One sub-cluster expressed high levels of the ERBB2 gene and patients in this group were linked to poor survival. The second sub-cluster of samples displayed over-expression of immune response genes and the patients in this cluster were linked to improved survival. The third sub-cluster expressed high levels of the Ropporin gene and this sub-cluster was linked to poor survival.

In conclusion, our gene expression profiling results identified various groups and sub-groups of breast cancer and associated them with defined clinical parameters and outcomes. Our results identified new clusters which may have clinical relevance.

5.2 Association of clinical parameters with genes, functions and pathways

Various clinical parameters associated with tumour specimens such as ER status, LN status, Grade and Tumour size were compared in relation to gene expression, function and pathways. The important findings are listed below.

- Cell cycle pathway genes were up-regulated in cancer specimens compared to the normal specimens. TP53, an important molecule in cell cycle regulation, was up-

regulated in cancer when compared to normal specimens. Genes associated with the embryonic stem cell pathway were also up-regulated in tumors compared to normal specimens. The fatty acid biosynthesis pathway genes were down-regulated in cancer compared to normal specimens.

- Interleukin 4-induced gene (IL4I1) and CCL5 expressions progressively increased with increase in genomic grade, higher in Grade 3 tumors and lowest in Grade 1 tumors. Both of these genes are related to immune response.
- Ropporin gene expression was found to be enriched among patients who relapsed (overall), patients who relapsed within 5 years and patients who did not survive beyond 5 years.
- SNIP and PCGF2 over-expression was linked to relapse, shorter survival and Lymph node-positive patients.

In conclusion, the analysis identified important genes and pathways up- or down-regulated when comparing various clinical conditions.

5.3 Comparing our in-house genelists with publicly available datasets

Gene expression from 4 publicly available datasets and our in-house datasets were analysed for genes which may be involved in relapse. Additionally our results were compared with genes from OncotypeDx and MammaPrint. The important findings are listed below.

- HSPB1, KIAA0101 and PAK3 were up-regulated in patients who relapsed in four out of the five cohorts.
- AP-1 transcriptional factor genes FOS and FOSB were down-regulated in patients who relapsed. FOS was down-regulated in four out of five cohorts and FOSB was down-regulated in three out of 5 cohorts.
- AURKA, BIRC5 and ERBB2 were up-regulated in patients who relapsed in our study and were also present on OncotypeDx as an indicator of bad prognosis.
- ESR1 and SCUBE2 were down-regulated in patients who relapsed in our study and were also present on OncotypeDx as an indicator of good prognosis.

- NMU, MELK and GMPS were up-regulated in patients who relapsed in our study and were also present on MammaPrint as an indicator of bad prognosis.
- Peci and SCUBE2 was down-regulated in patients who relapsed in our study and was also present on MammaPrint as an indicator of good prognosis

In conclusion, our analysis identified the important genes HSPB1, KIAA0101 and PAK3 to be up-regulated in patients who relapsed vs. those who did not relapse, and FOS and FOSB to be down-regulated in patients who relapsed vs. those who did not relapse. The NMU, GMPS, MELK, Peci and SCUBE2 genes were common to MammaPrint and our in-house study. AURKA, BIRC5, ERBB2, ESR1 and SCUBE2 were common to OncotypeDx and our in-house study.

5.4 Meta analysis for estrogen receptor pathway genes using gene expression data

Genes from 5 clinical and 1 cell line datasets were compared for differences in gene expression among ER-positive and ER-negative breast specimens and cell lines. Additionally gene expression profiles from 5897 microarray specimens were used to study the gene interaction network for ESR1 gene. The important findings are listed below:

- ANXA9, ABAT, BTG2, C10orf116, C1ORF34, C6orf211, CA12, CELSR1, COX6C, CRIP1, CSAD, EEF1A2, ERBB3, ERBB4, ESR1, FOXA1, GATA3, GREB1, HSPB1, INPP4B, KIAA1467, KRT18, KRT19, KRT8, LASS6, MAPT, MCCC2, MKL2, MLPH, MYB, MYO5C, NAT1, NME3, RAB17, RGL2, RHOB, SEMA3F, SLC19A2, SLC22A5, SLC39A6, SLC7A8, SLC9A3R1, SPDEF, TFF1, TFF3, THRAP2, THSD4, TPBG, TSPAN13, VAV3, XBP1 and ZNF552 were found to be up-regulated in all the 6 experiments comparing ER-positive specimens to ER-negative specimens
- The Nuclear Receptors pathway genes (ESR1, AR, RARA, RORC and NR2F6) were over-expressed among the ER-positive specimens.
- SFRP1, ANXA1, C10orf38, SLC43A3, PRNP, YBX1, LPIN1, TRIM2, MSN, COTL1, ODC1, TNFRSF21, YBX1, LPIN1, CEBPB, QKI, ENO1, FNDC3B and

- CREB3L2 were found to be down-regulated in all the 6 experiments comparing ER-positive specimens to ER-negative specimens
- Cell cycle and related pathways were over-expressed among the ER-negative specimens.
 - The Ropporin gene was found to be up-regulated in ER-negative patients in 3 out of 6 experiments studied, indicating Ropporin expression to be associated with ER-negativity.
 - GATA3, SPDEF, FOXA1 and C1ORF34 expression correlated with expression of ESR1 gene across 5897 specimens. All these genes were also up-regulated in ER-positive specimens in the meta-analysis.
 - k-means clustering and correlation analysis indicated that all the genes expressing together is the most obvious result from this study, however individual expression can exist independent of each other except for SPDEF and FOXA1.
 - FOXA1 expression was independent of SPDEF; however SPDEF expression appeared dependant on FOXA1 as high expression of SPDEF only existed with the high expression of FOXA1 as revealed by correlation graph and k-means clustering.

In conclusion, our study identified known and novel genes which are up- or down-regulated in ER-positive tumors compared to ER-negative tumors. The analysis also identified FOXA1, SPDEF, GATA3 and C1ORF34 expression to correlate with expression of ESR1. Furthermore, a dependency of SPDEF expression on FOXA1 expression was also identified.

5.5 Development of MLPERCEP, a software tool for predicting relapse in breast cancer

As part of the thesis work MLPERCEP (Multiple layer perceptron) implementing Back propagation Neural network algorithm was developed to predict relapse in breast cancer patients. The algorithm was implemented on our in-house dataset and publicly available datasets. The results were then compared to the results obtained from support vector machines. The important results are listed below.

- MLPERCEP is designed with user friendly graphics user interface and is available at <http://www.bioinformatics.org/mlpercep/>
- MLPERCEP can be used for gene expression arrays, or any other type of data which can be classified as two groups.
- Using 162 genes ($p < 0.001$ comparing patients who relapsed vs. patients who did not relapse), a classifier was developed to predict relapse in breast cancer patients. The classifier was able to predict relapse with an accuracy of 93.3% in a leave-one-out cross validation study. The same accuracy was obtained using support vector machines. However, with more stringent cut-off, the prediction accuracy of back propagation algorithm was 97.9%, however 10.5% of the patients could not be classified.
- Data from Van't Veer *et al.*, (2002) was used to develop a similar classifier and access the accuracy of the system. 117 genes ($p < 0.001$ comparing patients who developed distant metastasis vs. patients who remained disease-free) was used to develop the classifier. The classifier was able to predict the outcome with an accuracy of 87.2% in a leave-one-out cross validation study. Using support vector machines the prediction accuracy was 82.05%. However with more stringent cut-off, the prediction accuracy of back propagation algorithm was 89.7% and 12.8% of the patients could not be classified. This analysis indicated that back-propagation based classifier can outperform SVM classifiers.

In conclusion, a back propagation algorithm was successfully developed as a user-friendly software package which can be used to develop a prognostic model for breast cancer. The results generated were at par or better than Support Vector Machines in predicting relapse and distant metastasis in two of the datasets tested. Our classifier was better than existing diagnostic kits and has the potential to be considered for development of a diagnostic kit.

5.6 Functional analysis on Ropporin

Ropporin was over-expressed in patients who relapsed compared to patients who did not relapse. This gene was also over-expressed in patients who relapsed within 5 years, patients who did not survive beyond 5 years and ER-negative specimens. A follow up study using *in-silico* and *in-vitro* models was performed to assess the prevalence and functional role of this gene. The important findings are listed below:

- Our in-house microarray study indicated that Ropporin gene was significantly up-regulated in patients who relapsed, patients who did not survive beyond 5 years, patients who relapsed within 5 years and patients with ER-negative tumors. Of the 6 other publicly available breast cancer datasets analysed, Ropporin was found to be over-expressed in 3 datasets among the ER-negative specimens.
- Ropporin expression was found to progressively increase with melanoma progression and was highest in metastatic growth phase melanoma and lymph node metastasis. Ropporin expression was high in many of the melanoma cell lines and low in melanocytes.
- From analysis of the multiple myeloma dataset, Ropporin expression was found highly expressed in multiple myeloma patients.
- In normal cells, observation of high expression of Ropporin was limited to testes and cervical ganglion while marginal/low expression was observed in heart and liver.
- M14 was found to have nearly equal amounts of ROPN1 and ROPN1B. siRNA knockdown of ROPN1 and ROPN1B in this cell line showed reduction in invasion and motility. Over-expression of ROPN1 and ROPN1B in this cell line showed reduction in protein levels with an associated reduction in invasion and motility.
- MDA-MB-435s was found to have high expression of ROPN1B and low expression of ROPN1. siRNA knockdown of ROPN1B in this cell line showed reduction in motility. Over-expression of ROPN1 in this cell line

showed reduction in motility. Over expression of ROPN1B in this cell line showed a reduction in protein and marginal reduction in motility.

- MDA-MB-231 does not express ROPN1 or ROPN1B. Over-expression of ROPN1B showed reduction in invasion in this cell line

In conclusion, Ropporin over-expression was linked to breast cancer patients who relapsed. The gene was also linked to disease progression in melanoma. siRNA knockdown positively associated Ropporin gene to be involved in cancer cell motility and invasion. There is potential of targeting Ropporin molecule as therapeutic drug for a subset of breast cancer and melanoma and possibly multiple myeloma.

6.0 Future work

6.1 Validation of novel groups of specimens in independent studies

Our in-house study identified clusters of samples correlating with varying degrees of survival and other clinical parameters. Our study identified two novel groups of specimens; a “Poor survival group” expressing high levels of the Ropporin gene and a “Good survival group” expressing high levels of immune response genes. Future work would involve validating these results across other available datasets and to integrate this information in a prognostic model for Estrogen receptor-negative breast specimens, with the potential to develop this knowledge in a diagnostic/prognostic kit.

6.2 Diagnostic models

Our results have demonstrated the suitability of Neural Network models as predictive models for clinical outcomes for breast cancer using our in-house generated patient dataset. Future work would include validating these findings in independent studies and extending the model to predict the therapeutic options and treatment regimens that would be best for breast cancer patients. This will incorporate information from the future chemosensitivity and resistance profiles and their relation to gene expression profile. The information will be integrated in Neural Network architecture to provide personalised information for the individual patients based on tumour gene expression profiles.

6.3 Validation of gene interaction network

Our study identified a gene interaction network for ESR1 gene using a 5897-member chip dataset. GATA3, SPDEF, FOXA1 and C1ORF34 were identified to correlate in expression to ESR1. Other than C1ORF34 for which the function is not known, the GATA3, SPDEF and FOXA1 are all transcriptional factors indicating their involvement in the ER metabolism. Understanding their complex interactions may help in better understanding of the ER metabolism and lead to a deeper insight in the disease progression of ER-positive cancers. A future aim would be to identify transcriptional factors influencing expression of these genes and the way they affect other genes.

A strong dependency of SPDEF expression on FOXA1 was inferred from the microarray results. A lab validation using siRNA technology would validate the in-silico results. SPDEF expression reduction with knockdown of FOXA1 would confirm this observation and the analysis methodology for mining high throughput microarray data.

6.4 Ropporin as biomarker and targeted therapy

Our results identified the Ropporin gene to be over-expressed among the breast cancer patients who relapsed. A high expression of this gene was correlated with melanoma progression and observed in multiple myeloma. Functional analysis identified the gene to be linked to cancer cell motility. Because of the limited expression of this gene in normal tissue and its antigenic property to its autologous host, such types of cancer can be targeted for immunogenic therapy. Future aim would be to develop technology to target Ropporin-positive tumors using immunogenic therapy as the expression of this gene is localised to sperm.

Since the expression of this protein is very restricted in normal tissue with high expression in some metastatic cancers, detection of this protein or RNA in tumors could help determine the aggressive behaviour of cancer. There is also potential to look for the RNA and protein in serum and that can be developed as potential biomarker for detection of certain types of cancer (multiple myeloma, melanoma and a sub-set of breast cancer). The study by Li *et al.*, (2007) has identified the Ropporin mRNA in multiple myeloma patients. Future work would involve analysing large number of tumors and serums samples from breast, melanoma and multiple myeloma patients to establish the prognostic and predictive value of Ropporin expression.

References

Aggarwal, B. B., Shishodia, S., Takada, Y. 2006. TNF blockade: an inflammatory issue. *Ernst Schering Research Foundation workshop*. (56) (56), pp161-186.

Agoff, S. N., Swanson, P. E., Linden, H. 2003. Androgen receptor expression in estrogen receptor-negative breast cancer. Immunohistochemical, clinical, and prognostic associations. *American Journal of Clinical Pathology*. 120 (5), pp725-731.

Alexe, G., Dalgin, G. S., Scandfeld, D. 2007. High expression of lymphocyte-associated genes in node-negative HER2+ breast cancers correlates with lower recurrence rates. *Cancer research*. 67 (22), pp10669-10676.

Allred, D. C., Harvey, J. M., Berardo, M. and Clark, G. M. 1998. Prognostic and predictive factors in breast cancer by immunohistochemical analysis. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*. 11 (2), pp155-168.

Almudevar, A., Klebanov, L. B., Qiu, X. 2006. Utility of correlation measures in analysis of gene expression. *NeuroRx : the journal of the American Society for Experimental NeuroTherapeutics*. 3 (3), pp384-395.

Alo, P. L., Visca, P., Marci, A. 1996. Expression of fatty acid synthase (FAS) as a predictor of recurrence in stage I breast carcinoma patients. *Cancer*. 77 (3), pp474-482.

Alon, U., Barkai, N., Notterman, D. A. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*. 96 (12), pp6745-6750.

Alpy, F., Boulay, A., Moog-Lutz, C. 2003. Metastatic lymph node 64 (MLN64), a gene overexpressed in breast cancers, is regulated by Sp/KLF transcription factors. *Oncogene*. 22 (24), pp3770-3780.

Anders, C. and Carey, L. A. 2008. Understanding and treating triple-negative breast cancer. *Oncology (Williston Park, N.Y.)*. 22 (11), pp1233-9; discussion 1239-40, 1243.

Arts, J., Kuiper, G. G., Janssen, J. M. 1997. Differential expression of estrogen receptors alpha and beta mRNA during differentiation of human osteoblast SV-HFO cells. *Endocrinology*. 138 (11), pp5067-5070.

Ascenzi, P., Bocedi, A. and Marino, M. 2006. Structure-function relationship of estrogen receptor alpha and beta: impact on human health. *Molecular aspects of medicine*. 27 (4), pp299-402.

Ashkenazi, R., Gentry, S. N. and Jackson, T. L. 2008. Pathways to tumorigenesis-- modeling mutation acquisition in stem cells and their progeny. *Neoplasia (New York, N.Y.)*. 10 (11), pp1170-1182.

Bacac, M. and Stamenkovic, I. 2008. Metastatic cancer cell. *Annual review of pathology*. 3pp221-247.

Backvall, H., Asplund, A., Gustafsson, A. 2005. Genetic tumor archeology: microdissection and genetic heterogeneity in squamous and basal cell carcinoma. *Mutation research*. 571 (1-2), pp65-79.

Badve, S., Turbin, D., Thorat, M. A. 2007. FOXA1 expression in breast cancer-- correlation with luminal subtype A and survival. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 13 (15 Pt 1), pp4415-4421.

Baier, G. 2003. The PKC gene module: molecular biosystematics to resolve its T cell functions. *Immunological reviews*. 192pp64-79.

Barinaga, M. 1997. Designing therapies that target tumor blood vessels. *Science (New York, N.Y.)*. 275 (5299), pp482-484.

Barnett, D. H., Sheng, S., Charn, T. H. 2008. Estrogen receptor regulation of carbonic anhydrase XII through a distal enhancer in breast cancer. *Cancer research*. 68 (9), pp3505-3515.

Baum, H. P., Schmid, T., Schock, G. and Reichrath, J. 1996. Expression of CD44 isoforms in basal cell carcinomas. *The British journal of dermatology*. 134 (3), pp465-468.

Bausero, M. A., Bharti, A., Page, D. T. 2006. Silencing the hsp25 gene eliminates migration capability of the highly metastatic murine 4T1 breast adenocarcinoma cell. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine*. 27 (1), pp17-26.

Ben-Baruch, A. 2008. Organ selectivity in metastasis: regulation by chemokines and their receptors. *Clinical & experimental metastasis*. 25 (4), pp345-356.

Benusiglio, P. R., Pharoah, P. D., Smith, P. L. 2006. HapMap-based study of the 17q21 ERBB2 amplicon in susceptibility to breast cancer. *British journal of cancer*. 95 (12), pp1689-1695.

Bhargava, R., Beriwal, S., McManus, K. and Dabbs, D. J. 2008. CK5 is more sensitive than CK5/6 in identifying the "basal-like" phenotype of breast carcinoma. *American Journal of Clinical Pathology*. 130 (5), pp724-730.

Bieche, I., Onody, P., Tozlu, S. 2003. Prognostic value of ERBB family mRNA expression in breast carcinomas. *International journal of cancer. Journal international du cancer*. 106 (5), pp758-765.

Bild, A. H., Yao, G., Chang, J. T. 2006. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*. 439 (7074), pp353-357.

Blaszczyk, J., Tropea, J. E., Bubunencko, M. 2001. Crystallographic and modeling studies of RNase III suggest a mechanism for double-stranded RNA cleavage. *Structure (London, England : 1993)*. 9 (12), pp1225-1236.

BLOOM, H. J. and RICHARDSON, W. W. 1957. Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. *British journal of cancer*. 11 (3), pp359-377.

Boda, B., Alberi, S., Nikonenko, I. 2004. The mental retardation protein PAK3 contributes to synapse formation and plasticity in hippocampus. *The Journal of neuroscience : the official journal of the Society for Neuroscience*. 24 (48), pp10816-10825.

Boonchai, W., Walsh, M., Cummings, M. and Chenevix-Trench, G. 2000. Expression of beta-catenin, a key mediator of the WNT signaling pathway, in basal cell carcinoma. *Archives of Dermatology*. 136 (7), pp937-938.

Borresen, A. L., Andersen, T. I., Eyfjord, J. E. 1995. TP53 mutations and breast cancer prognosis: particularly poor survival rates for cases with mutations in the zinc-binding domains. *Genes, chromosomes & cancer*. 14 (1), pp71-75.

Bose, S., Mohammed, M., Shintaku, P. and Rao, P. N. 2001. Her-2/neu gene amplification in low to moderately expressing breast cancers: possible role of chromosome 17/Her-2/neu polysomy. *The breast journal*. 7 (5), pp337-344.

Bourguet, W., Germain, P. and Gronemeyer, H. 2000. Nuclear receptor ligand-binding domains: three-dimensional structures, molecular interactions and pharmacological implications. *Trends in pharmacological sciences*. 21 (10), pp381-388.

Brady, M. E., Ozanne, D. M., Gaughan, L. 1999. Tip60 is a nuclear hormone receptor coactivator. *The Journal of biological chemistry*. 274 (25), pp17599-17604.

Brennan, D. J., O'Brien, S. L., Fagan, A. 2005. Application of DNA microarray technology in determining breast cancer prognosis and therapeutic response. *Expert opinion on biological therapy*. 5 (8), pp1069-1083.

- Britton, D. J., Hutcheson, I. R., Knowlden, J. M. 2006. Bidirectional cross talk between ERalpha and EGFR signalling pathways regulates tamoxifen-resistant growth. *Breast cancer research and treatment*. 96 (2), pp131-146.
- Bryan, B. B., Schnitt, S. J. and Collins, L. C. 2006. Ductal carcinoma in situ with basal-like phenotype: a possible precursor to invasive basal-like breast cancer. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc.* 19 (5), pp617-621.
- Bryan, R. M., Mercer, R. J., Bennett, R. C. 1984. Androgen receptors in breast cancer. *Cancer*. 54 (11), pp2436-2440.
- Burdall, S. E., Hanby, A. M., Lansdown, M. R. and Speirs, V. 2003. Breast cancer cell lines: friend or foe? *Breast cancer research : BCR*. 5 (2), pp89-95.
- Buscher, D. and Ruther, U. 1998. Expression profile of Gli family members and Shh in normal and mutant mouse limb development. *Developmental dynamics : an official publication of the American Association of Anatomists*. 211 (1), pp88-96.
- Buyse, M., Loi, S., van't Veer, L. 2006. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *Journal of the National Cancer Institute*. 98 (17), pp1183-1192.
- Cai, Y. D., Liu, X. J. and Chou, K. C. 2003. Prediction of protein secondary structure content by artificial neural network. *Journal of computational chemistry*. 24 (6), pp727-731.
- Callagy, G. M., Pharoah, P. D., Pinder, S. E. 2006. Bcl-2 is a prognostic marker in breast cancer independently of the Nottingham Prognostic Index. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 12 (8), pp2468-2475.
- Calza, S., Hall, P., Auer, G. 2006. Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. *Breast cancer research : BCR*. 8 (4), ppR34.

Carlei, F., Lomanto, D., Chimenti, S. 1986. Immunocytochemical study of a trabecular carcinoma of the skin (Merkel cell tumor). Case report. *The Italian journal of surgical sciences / sponsored by Societa italiana di chirurgia*. 16 (1), pp55-59.

Carr, D. W., Fujita, A., Stentz, C. L. 2001. Identification of sperm-specific proteins that interact with A-kinase anchoring proteins in a manner similar to the type II regulatory subunit of PKA. *The Journal of biological chemistry*. 276 (20), pp17332-17338.

Carroll, J. S., Meyer, C. A., Song, J. 2006. Genome-wide analysis of estrogen receptor binding sites. *Nature genetics*. 38 (11), pp1289-1297.

Carter, C. L., Allen, C. and Henson, D. E. 1989. Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer*. 63 (1), pp181-187.

Chambers, A. F., Groom, A. C. and MacDonald, I. C. 2002. Dissemination and growth of cancer cells in metastatic sites. *Nature reviews.Cancer*. 2 (8), pp563-572.

Chang, J. C., Hilsenbeck, S. G. and Fuqua, S. A. 2005. The promise of microarrays in the management and treatment of breast cancer. *Breast cancer research : BCR*. 7 (3), pp100-104.

Chang, J. T., Wang, H. M., Chang, K. W. 2005. Identification of differentially expressed genes in oral squamous cell carcinoma (OSCC): overexpression of NPM, CDK1 and NDRG1 and underexpression of CHES1. *International journal of cancer.Journal international du cancer*. 114 (6), pp942-949.

Charafe-Jauffret, E., Ginestier, C., Monville, F. 2006. Gene expression profiling of breast cell lines identifies potential new basal markers. *Oncogene*. 25 (15), pp2273-2284.

Chiriva, M., Ferrari, R., Yu, Y. 2007. Ropporin is a novel cancer testis antigen for multiple myeloma. *J Immunol*. 178 (MeetingAbstracts), ppLB44-b. Available from: <<http://www.jimmunol.org>>

Chitemerere, M., Andersen, T. I., Holm, R. 1996. TP53 alterations in atypical ductal hyperplasia and ductal carcinoma in situ of the breast. *Breast cancer research and treatment*. 41 (2), pp103-109.

Chuang, T. Y., Popescu, A., Su, W. P. and Chute, C. G. 1990. Basal cell carcinoma. A population-based incidence study in Rochester, Minnesota. *Journal of the American Academy of Dermatology*. 22 (3), pp413-417.

Cohen, L. A., Thompson, D. O., Maeura, Y. 1986. Dietary fat and mammary cancer. I. Promoting effects of different dietary fats on N-nitrosomethylurea-induced rat mammary tumorigenesis. *Journal of the National Cancer Institute*. 77 (1), pp33-42.

Cohen, M. M., Jr. 2003. The hedgehog signaling network. *American journal of medical genetics. Part A*. 123A (1), pp5-28.

Collado, M., Garcia, V., Garcia, J. M. 2007. Genomic profiling of circulating plasma RNA for the analysis of cancer. *Clinical chemistry*. 53 (10), pp1860-1863.

Colleoni, M., Rotmensz, N., Peruzzotti, G. 2005. Size of breast cancer metastases in axillary lymph nodes: clinical relevance of minimal lymph node involvement. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 23 (7), pp1379-1389.

Condomines, M., Hose, D., Raynaud, P. 2007. Cancer/testis genes in multiple myeloma: expression patterns and prognosis value determined by microarray analysis. *Journal of immunology (Baltimore, Md.: 1950)*. 178 (5), pp3307-3315.

Conzen, S. D. 2008. Minireview: nuclear receptors and breast cancer. *Molecular endocrinology (Baltimore, Md.)*. 22 (10), pp2215-2228.

Cooper, C. S. 2001. Applications of microarray technology in breast cancer research. *Breast cancer research : BCR*. 3 (3), pp158-175.

Copie-Bergman, C., Boulland, M. L., Dehoule, C. 2003. Interleukin 4-induced gene 1 is activated in primary mediastinal large B-cell lymphoma. *Blood*. 101 (7), pp2756-2761.

Corona, R., Dogliotti, E., D'Errico, M. 2001. Risk factors for basal cell carcinoma in a Mediterranean population: role of recreational sun exposure early in life. *Archives of Dermatology*. 137 (9), pp1162-1168.

Couse, J. F., Lindzey, J., Grandien, K. 1997. Tissue distribution and quantitative analysis of estrogen receptor-alpha (ERalpha) and estrogen receptor-beta (ERbeta) messenger ribonucleic acid in the wild-type and ERalpha-knockout mouse. *Endocrinology*. 138 (11), pp4613-4621.

Cowley, S. M., Hoare, S., Mosselman, S. and Parker, M. G. 1997. Estrogen receptors alpha and beta form heterodimers on DNA. *The Journal of biological chemistry*. 272 (32), pp19858-19862.

Cross, H. S., Bajna, E., Bises, G. 1996. Vitamin D receptor and cytokeratin expression may be progression indicators in human colon cancer. *Anticancer Research*. 16 (4B), pp2333-2337.

Cross, H. S., Bareis, P., Hofer, H. 2001. 25-Hydroxyvitamin D(3)-1alpha-hydroxylase and vitamin D receptor gene expression in human colonic mucosa is elevated during early cancerogenesis. *Steroids*. 66 (3-5), pp287-292.

Dahlman-Wright, K., Cavailles, V., Fuqua, S. A. 2006. International Union of Pharmacology. LXIV. Estrogen receptors. *Pharmacological reviews*. 58 (4), pp773-781.

Dairkee, S. H., Ji, Y., Ben, Y. 2004. A molecular 'signature' of primary breast cancer cultures; patterns resembling tumor tissue. *BMC genomics*. 5 (1), pp47.

Daya-Grosjean, L. and Couve-Privat, S. 2005. Sonic hedgehog signaling in basal cell carcinomas. *Cancer letters*. 225 (2), pp181-192.

De Laurentiis, M., De Placido, S., Bianco, A. R. 1999. A prognostic model that makes quantitative estimates of probability of relapse for breast cancer patients. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 5 (12), pp4133-4139.

Demicheli, R. 2001. Tumour dormancy: findings and hypotheses from clinical research on breast cancer. *Seminars in cancer biology*. 11 (4), pp297-306.

Demicheli, R., Retsky, M. W., Swartzendruber, D. E. and Bonadonna, G. 1997. Proposal for a new model of breast cancer metastatic development. *Annals of Oncology : Official Journal of the European Society for Medical Oncology / ESMO*. 8 (11), pp1075-1080.

Derisi, J. 2001. Overview of nucleic acid arrays. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.,]*. Chapter 22ppUnit 22.1.

Deroo, B. J. and Korach, K. S. 2006. Estrogen receptors and human disease. *The Journal of clinical investigation*. 116 (3), pp561-570.

Desmedt, C., Haibe-Kains, B., Wirapati, P. 2008. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 14 (16), pp5158-5165.

Do, J. H. and Choi, D. K. 2008. Clustering approaches to identifying gene expression patterns from DNA microarray data. *Molecules and cells*. 25 (2), pp279-288.

Dobrzyn, A. and Ntambi, J. M. 2005. The role of stearoyl-CoA desaturase in the control of metabolism. Prostaglandins, leukotrienes, and essential fatty acids. 73 (1), pp35-41.

Dong, J. T., Lamb, P. W., Rinker-Schaeffer, C. W. 1995. KAI1, a metastasis suppressor gene for prostate cancer on human chromosome 11p11.2. *Science (New York, N.Y.)*. 268 (5212), pp884-886.

Drabsch, Y., Hugo, H., Zhang, R. 2007. Mechanism of and requirement for estrogen-regulated MYB expression in estrogen-receptor-positive breast cancer cells. *Proceedings of the National Academy of Sciences of the United States of America*. 104 (34), pp13762-13767.

Dua, R., Beetner, D. G., Stoecker, W. V. and Wunsch, D. C., 2nd. 2004. Detection of basal cell carcinoma using electrical impedance and neural networks. *IEEE transactions on bio-medical engineering*. 51 (1), pp66-71.

Duffy, M. J. 2005. Predictive markers in breast and other cancers: a review. *Clinical chemistry*. 51 (3), pp494-503.

Easterday, M. C., Dougherty, J. D., Jackson, R. L. 2003. Neural progenitor genes. Germinal zone expression and analysis of genetic overlap in stem cell populations. *Developmental biology*. 264 (2), pp309-322.

Edwards, D. P., Altmann, M., DeMarzo, A. 1995. Progesterone receptor and the mechanism of action of progesterone antagonists. *The Journal of steroid biochemistry and molecular biology*. 53 (1-6), pp449-458.

Ehmann, U. K., Stevenson, M. A., Calderwood, S. K. and DeVries, J. T. 1998. Physical connections between feeder cells and recipient normal mammary epithelial cells. *Experimental cell research*. 243 (1), pp76-86.

Einarsdottir, K., Darabi, H., Li, Y. 2008. ESR1 and EGF genetic variation in relation to breast cancer risk and survival. *Breast cancer research : BCR*. 10 (1), ppR15.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*. 95 (25), pp14863-14868.

Elbashir, S. M., Harborth, J., Lendeckel, W. 2001. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*. 411 (6836), pp494-498.

Ellison, G., Klinowska, T., Westwood, R. F. 2002. Further evidence to support the melanocytic origin of MDA-MB-435. *Molecular pathology : MP*. 55 (5), pp294-299.

Ellsworth, R. E., Ellsworth, D. L., Patney, H. L. 2008. Amplification of HER2 is a marker for global genomic instability. *BMC cancer*. 8pp297.

Epstein, J. I., Carmichael, M. and Partin, A. W. 1995. OA-519 (fatty acid synthase) as an independent predictor of pathologic state in adenocarcinoma of the prostate. *Urology*. 45 (1), pp81-86.

Ertel, A., Verghese, A., Byers, S. W. 2006. Pathway-specific differences between tumor cell lines and normal and tumor tissue cells. *Molecular cancer*. 5 (1), pp55.

Evans, E. E., Henn, A. D., Jonason, A. 2006. C35 (C17orf37) is a novel tumor biomarker abundantly expressed in breast cancer. *Molecular cancer therapeutics*. 5 (11), pp2919-2930.

Evans, S. R., Nolla, J., Hanfelt, J. 1998. Vitamin D receptor expression as a predictive marker of biological behavior in human colorectal cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 4 (7), pp1591-1595.

Fabian, C. J. and Kimler, B. F. 2005. Selective estrogen-receptor modulators for primary prevention of breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 23 (8), pp1644-1655.

Fan, M., Long, X., Bailey, J. A. 2002. The activating enzyme of NEDD8 inhibits steroid receptor function. *Molecular endocrinology (Baltimore, Md.)*. 16 (2), pp315-330.

Farmer, P., Bonnefoi, H., Becette, V. 2005. Identification of molecular apocrine breast tumors by microarray analysis. *Oncogene*. 24 (29), pp4660-4671.

Feldman, R. J., Sementchenko, V. I., Gayed, M. 2003. Pdef expression in human breast cancer is correlated with invasive potential and altered gene expression. *Cancer research*. 63 (15), pp4626-4631.

Ferrara, N. 2001. Role of vascular endothelial growth factor in regulation of physiological angiogenesis. *American journal of physiology. Cell physiology*. 280 (6), ppC1358-66.

Fidler, I. J. and Kripke, M. L. 1977. Metastasis results from preexisting variant cells within a malignant tumor. *Science (New York, N.Y.)*. 197 (4306), pp893-895.

Fiedler, S. E., Bajpai, M. and Carr, D. W. 2008. Identification and characterization of RHOA-interacting proteins in bovine spermatozoa. *Biology of reproduction*. 78 (1), pp184-192.

Fiedler, S. E., Bajpai, M. and Carr, D. W. 2008. Identification and characterization of RHOA-interacting proteins in bovine spermatozoa. *Biology of reproduction*. 78 (1), pp184-192.

Fitzgerald, P., Teng, M., Chandraratna, R. A. 1997. Retinoic acid receptor alpha expression correlates with retinoid-induced growth inhibition of human breast cancer cells regardless of estrogen receptor status. *Cancer research*. 57 (13), pp2642-2650.

Fox, E. M., Bernaciak, T. M., Wen, J. 2008. Signal transducer and activator of transcription 5b, c-Src, and epidermal growth factor receptor signaling play integral roles in estrogen-stimulated proliferation of estrogen receptor-positive breast cancer cells. *Molecular endocrinology (Baltimore, Md.)*. 22 (8), pp1781-1796.

Freeman, T. C., Goldovsky, L., Brosch, M. 2007. Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS computational biology*. 3 (10), pp2032-2042.

Fujita, A., Nakamura, K., Kato, T. 2000. Ropporin, a sperm-specific binding protein of rhophilin, that is localized in the fibrous sheath of sperm flagella. *Journal of cell science*. 113 (Pt 1) (Pt 1), pp103-112.

Fujita, A., Nakamura, K., Kato, T. 2000. Ropporin, a sperm-specific binding protein of rhophilin, that is localized in the fibrous sheath of sperm flagella. *Journal of cell science*. 113 (Pt 1) (Pt 1), pp103-112.

Gadkar-Sable, S., Shah, C., Rosario, G. 2005. Progesterone receptors: various forms and functions in reproductive tissues. *Frontiers in bioscience : a journal and virtual library*. 10pp2118-2130.

Gagos, S. and Irminger-Finger, I. 2005. Chromosome instability in neoplasia: chaotic roots to continuous growth. *The international journal of biochemistry & cell biology*. 37 (5), pp1014-1033.

Gailani, M. R., Stahle-Backdahl, M., Leffell, D. J. 1996. The role of the human homologue of Drosophila patched in sporadic basal cell carcinomas. *Nature genetics*. 14 (1), pp78-81.

Gallicchio, L., Berndt, S. I., McSorley, M. A. 2006. Polymorphisms in estrogen-metabolizing and estrogen receptor genes and the risk of developing breast cancer among a cohort of women with benign breast disease. *BMC cancer*. 6pp173.

Gao, X. and Nawaz, Z. 2002. Progesterone receptors - animal models and cell signaling in breast cancer: Role of steroid receptor coactivators and corepressors of progesterone receptors in breast cancer. *Breast cancer research : BCR*. 4 (5), pp182-186.

Garcia-Closas, M., Troester, M. A., Qi, Y. 2007. Common genetic variation in GATA-binding protein 3 and differential susceptibility to breast cancer by estrogen receptor alpha tumor status. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*. 16 (11), pp2269-2275.

Garrido, C., Brunet, M., Didelot, C. 2006. Heat shock proteins 27 and 70: anti-apoptotic proteins with tumorigenic properties. *Cell cycle (Georgetown, Tex.)*. 5 (22), pp2592-2601.

Gautier, L., Moller, M., Friis-Hansen, L. and Knudsen, S. 2004. Alternative mapping of probes to genes for Affymetrix chips. *BMC bioinformatics*. 5pp111.

Gazdar, A. F., Kurvari, V., Virmani, A. 1998. Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. *International journal of cancer. Journal international du cancer*. 78 (6), pp766-774.

Geho, D. H., Bandle, R. W., Clair, T. and Liotta, L. A. 2005. Physiological mechanisms of tumor-cell invasion and migration. *Physiology (Bethesda, Md.)*. 20pp194-200.

Ghadersohi, A. and Sood, A. K. 2001. Prostate epithelium-derived Ets transcription factor mRNA is overexpressed in human breast tumors and is a candidate breast tumor marker and a breast tumor antigen. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 7 (9), pp2731-2738.

Giangrande, P. H. and McDonnell, D. P. 1999. The A and B isoforms of the human progesterone receptor: two functionally different transcription factors encoded by a single gene. *Recent progress in hormone research*. 54pp291-313; discussion 313-4.

Giese, M. A., Man, M. Z., Gorski, N. A. 2004. The influence of tumor size and environment on gene expression in commonly used human tumor lines. *BMC cancer*. 4pp35.

Glas, A. M., Floore, A., Delahaye, L. J. 2006. Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC genomics*. 7pp278.

Greco, B., Blasberg, M. E., Kosinski, E. C. and Blaustein, J. D. 2003. Response of ERalpha-IR and ERbeta-IR cells in the forebrain of female rats to mating stimuli. *Hormones and behavior*. 43 (4), pp444-453.

Grishok, A., Pasquinelli, A. E., Conte, D. 2001. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell*. 106 (1), pp23-34.

Guerin, M., Barrois, M. and Riou, G. 1988. The expression of c-myc is strongly associated with the presence of estrogen and progesterone receptors in breast cancer. *Comptes rendus de l'Academie des sciences.Serie III, Sciences de la vie.* 307 (20), pp855-861.

Habashy, H. O., Powe, D. G., Rakha, E. A. 2008. Forkhead-box A1 (FOXA1) expression in breast cancer and its prognostic significance. *European journal of cancer (Oxford, England : 1990).* 44 (11), pp1541-1551.

Habel, L. A., Shak, S., Jacobs, M. K. 2006. A population-based study of tumor gene expression and risk of breast cancer death among lymph node-negative patients. *Breast cancer research : BCR.* 8 (3), ppR25.

Hall, J. M., Couse, J. F. and Korach, K. S. 2001. The multifaceted mechanisms of estradiol and estrogen receptor signaling. *The Journal of biological chemistry.* 276 (40), pp36869-36872.

Hamilton, A., Voinnet, O., Chappell, L. and Baulcombe, D. 2002. Two classes of short interfering RNA in RNA silencing. *The EMBO journal.* 21 (17), pp4671-4679.

Hammond, S. M., Bernstein, E., Beach, D. and Hannon, G. J. 2000. An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature.* 404 (6775), pp293-296.

Hanahan, D. and Folkman, J. 1996. Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. *Cell.* 86 (3), pp353-364.

Hartwell, L. 1992. Defects in a cell cycle checkpoint may be responsible for the genomic instability of cancer cells. *Cell.* 71 (4), pp543-546.

Haslam, S. Z. and Woodward, T. L. 2003. Host microenvironment in breast cancer development: epithelial-cell-stromal-cell interactions and steroid hormone action in normal and cancerous mammary gland. *Breast cancer research : BCR.* 5 (4), pp208-215.

Haykin, S. 1998. /Neural Networks: A Comprehensive Foundation . /2nd. Prentice Hall. ISBN 0132733501.

Healy, E., Angus, B., Lawrence, C. M. and Rees, J. L. 1995. Prognostic value of Ki67 antigen expression in basal cell carcinomas. *The British journal of dermatology*. 133 (5), pp737-741.

Hermann-Kleiter, N., Gruber, T., Lutz-Nicoladoni, C. 2008. The nuclear orphan receptor NR2F6 suppresses lymphocyte activation and T helper 17-dependent autoimmunity. *Immunity*. 29 (2), pp205-216.

Herynk, M. H. and Fuqua, S. A. 2004. Estrogen receptor mutations in human disease. *Endocrine reviews*. 25 (6), pp869-898.

Heyer, B. S., Kochanowski, H. and Solter, D. 1999. Expression of Melk, a new protein kinase, during early mouse development. *Developmental dynamics : an official publication of the American Association of Anatomists*. 215 (4), pp344-351.

Hoek, K., Rimm, D. L., Williams, K. R. 2004. Expression profiling reveals novel pathways in the transformation of melanocytes to melanomas. *Cancer research*. 64 (15), pp5270-5282.

Holst, F., Stahl, P. R., Ruiz, C. 2007. Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer. *Nature genetics*. 39 (5), pp655-660.

Hong, Y., Muller, U. R. and Lai, F. 2003. Discriminating two classes of toxicants through expression analysis of HepG2 cells with DNA arrays. *Toxicology in vitro : an international journal published in association with BIBRA*. 17 (1), pp85-92.

Horak, C. E. and Steeg, P. S. 2005. Metastasis gets site specific. *Cancer cell*. 8 (2), pp93-95.

Hornberger, J., Cosler, L. E. and Lyman, G. H. 2005. Economic analysis of targeting chemotherapy using a 21-gene RT-PCR assay in lymph-node-negative, estrogen-

receptor-positive, early-stage breast cancer. *The American Journal of Managed Care*. 11 (5), pp313-324.

Horwitz, K. B. and Alexander, P. S. 1983. In situ photolinked nuclear progesterone receptors of human breast cancer cells: subunit molecular weights after transformation and translocation. *Endocrinology*. 113 (6), pp2195-2201.

Horwitz, K. B., Koseki, Y. and McGuire, W. L. 1978. Estrogen control of progesterone receptor in human breast cancer: role of estradiol and antiestrogen. *Endocrinology*. 103 (5), pp1742-1751.

Hosokawa, M., Takehara, A., Matsuda, K. 2007. Oncogenic role of KIAA0101 interacting with proliferating cell nuclear antigen in pancreatic cancer. *Cancer research*. 67 (6), pp2568-2576.

Howell, B. G., Solish, N., Lu, C. 2005. Microarray profiles of human basal cell carcinoma: insights into tumor growth and behavior. *Journal of dermatological science*. 39 (1), pp39-51.

Htun, H., Holth, L. T., Walker, D. 1999. Direct visualization of the human estrogen receptor alpha reveals a role for ligand in the nuclear distribution of the receptor. *Molecular biology of the cell*. 10 (2), pp471-486.

Hu, Z., Fan, C., Oh, D. S. 2006. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC genomics*. 7pp96.

Huang, E., Cheng, S. H., Dressman, H. 2003. Gene expression predictors of breast cancer outcomes. *Lancet*. 361 (9369), pp1590-1596.

Hunter, T., Hunt, T., Jackson, R. J. and Robertson, H. D. 1975. The characteristics of inhibition of protein synthesis by double-stranded ribonucleic acid in reticulocyte lysates. *The Journal of biological chemistry*. 250 (2), pp409-417.

Hyun, C. L., Lee, H. E., Kim, K. S. 2008. The effect of chromosome 17 polysomy on HER-2/neu status in breast cancer. *Journal of clinical pathology*. 61 (3), pp317-321.

Ingham, P. W. and McMahon, A. P. 2001. Hedgehog signaling in animal development: paradigms and principles. *Genes & development*. 15 (23), pp3059-3087.

Ionescu, D. N., Arida, M. and Jukic, D. M. 2006. Metastatic basal cell carcinoma: four case reports, review of literature, and immunohistochemical evaluation. *Archives of Pathology & Laboratory Medicine*. 130 (1), pp45-51.

Irizarry, R. A., Bolstad, B. M., Collin, F. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research*. 31 (4), ppe15.

Ivshina, A. V., George, J., Senko, O. 2006. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer research*. 66 (21), pp10292-10301.

Jaiswal, K. and Naik, P. K. 2008. Distinguishing compounds with anticancer activity by ANN using inductive QSAR descriptors. *Bioinformation*. 2 (10), pp441-451.

Jatoi, I., Hilsenbeck, S. G., Clark, G. M. and Osborne, C. K. 1999. Significance of axillary lymph node metastasis in primary breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 17 (8), pp2334-2340.

Jiang, W., Jimenez, G., Wells, N. J. 1998. PRC1: a human mitotic spindle-associated CDK substrate protein required for cytokinesis. *Molecular cell*. 2 (6), pp877-885.

Johnson, R. L., Rothman, A. L., Xie, J. 1996. Human homolog of patched, a candidate gene for the basal cell nevus syndrome. *Science (New York, N.Y.)*. 272 (5268), pp1668-1671.

Kallay, E., Bareis, P., Bajna, E. 2002. Vitamin D receptor activity and prevention of colonic hyperproliferation and oxidative stress. *Food and chemical toxicology : an*

international journal published for the British Industrial Biological Research Association. 40 (8), pp1191-1196.

Kang, S. H., Kang, K. W., Kim, K. H. 2008. Upregulated HSP27 in human breast cancer cells reduces Herceptin susceptibility by increasing Her2 protein stability. *BMC cancer*. 8pp286.

Kaplan, E.L & Meier, P. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53: 457-481.

Kapucuoglu, N., Basak, P. Y., Bircan, S. 2009. Immunohistochemical galectin-3 expression in non-melanoma skin cancers. *Pathology, research and practice*. 205 (2), pp97-103.

Karlsson, E., Delle, U., Danielsson, A. 2008. Gene expression variation to predict 10-year survival in lymph-node-negative breast cancer. *BMC cancer*. 8pp254.

Karrison, T. G., Ferguson, D. J. and Meier, P. 1999. Dormancy of mammary carcinoma after mastectomy. *Journal of the National Cancer Institute*. 91 (1), pp80-85.

Kastan, M. B. and Bartek, J. 2004. Cell-cycle checkpoints and cancer. *Nature*. 432 (7015), pp316-323.

Kennedy, S., Clynes, M., Doolan, P. 2008. SNIP/p140Cap mRNA expression is an unfavourable prognostic factor in breast cancer and is not expressed in normal breast tissue. *British journal of cancer*. 98 (10), pp1641-1645.

Kidokoro, T., Tanikawa, C., Furukawa, Y. 2008. CDC20, a potential cancer therapeutic target, is negatively regulated by p53. *Oncogene*. 27 (11), pp1562-1571.

Kim, J. M., Sohn, H. Y., Yoon, S. Y. 2005. Identification of gastric cancer-related genes using a cDNA microarray containing novel expressed sequence tags expressed in gastric cancer cells. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 11 (2 Pt 1), pp473-482.

Kim, S. J., Kang, H. S., Chang, H. L. 2008. Promoter hypomethylation of the N-acetyltransferase 1 gene in breast cancer. *Oncology reports*. 19 (3), pp663-668.

King, R. W. 2002. Roughing up Smoothened: chemical modulators of hedgehog signaling. *Journal of biology*. 1 (2), pp8.

Kinzler, K. W. and Vogelstein, B. 1990. The GLI gene encodes a nuclear protein which binds specific sequences in the human genome. *Molecular and cellular biology*. 10 (2), pp634-642.

Ko, L., Cardona, G. R., Henrion-Caude, A. and Chin, W. W. 2002. Identification and characterization of a tissue-specific coactivator, GT198, that interacts with the DNA-binding domains of nuclear receptors. *Molecular and cellular biology*. 22 (1), pp357-369.

Koljonen, V., Haglund, C., Tukiainen, E. and Bohling, T. 2005. Neuroendocrine differentiation in primary Merkel cell carcinoma--possible prognostic significance. *Anticancer Research*. 25 (2A), pp853-858.

Kominea, A., Konstantinopoulos, P. A., Kapranos, N. 2004. Androgen receptor (AR) expression is an independent unfavorable prognostic factor in gastric cancer. *Journal of cancer research and clinical oncology*. 130 (5), pp253-258.

Korohoda, W. and Madeja, Z. 1997. Contact of sarcoma cells with aligned fibroblasts accelerates their displacement: computer-assisted analysis of tumour cell locomotion in co-culture. *Biochemistry and cell biology = Biochimie et biologie cellulaire*. 75 (3), pp263-276.

Kricker, A., Armstrong, B. K., English, D. R. and Heenan, P. J. 1995. A dose-response curve for sun exposure and basal cell carcinoma. *International journal of cancer. Journal international du cancer*. 60 (4), pp482-488.

Kuang, W. W., Thompson, D. A., Hoch, R. V. and Weigel, R. J. 1998. Differential screening and suppression subtractive hybridization identified genes differentially

expressed in an estrogen receptor-positive breast carcinoma cell line. *Nucleic acids research*. 26 (4), pp1116-1123.

Kuennen-Boumeester, V., Van der Kwast, T. H., Claassen, C. C. 1996. The clinical significance of androgen receptors in breast cancer and their relation to histological and cell biological parameters. *European journal of cancer (Oxford, England : 1990)*. 32A (9), pp1560-1565.

Kuhajda, F. P. 2000. Fatty-acid synthase and human cancer: new perspectives on its role in tumor biology. *Nutrition (Burbank, Los Angeles County, Calif.)*. 16 (3), pp202-208.

Kuhajda, F. P., Jenner, K., Wood, F. D. 1994. Fatty acid synthesis: a potential selective target for antineoplastic therapy. *Proceedings of the National Academy of Sciences of the United States of America*. 91 (14), pp6379-6383.

Kuiper, G. G., Carlsson, B., Grandien, K. 1997. Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors alpha and beta. *Endocrinology*. 138 (3), pp863-870.

Kulkarni, G., Turbin, D. A., Amiri, A. 2007. Expression of protein elongation factor eEF1A2 predicts favorable outcome in breast cancer. *Breast cancer research and treatment*. 102 (1), pp31-41.

Lacroix, M. and Leclercq, G. 2004. About GATA3, HNF3A, and XBP1, three genes co-expressed with the oestrogen receptor-alpha gene (ESR1) in breast cancer. *Molecular and cellular endocrinology*. 219 (1-2), pp1-7.

Lacroix, M., Toillon, R. A. and Leclercq, G. 2006. P53 and Breast Cancer, an Update. *Endocrine-related cancer*. 13 (2), pp293-325.

Langerod, A., Zhao, H., Borgan, O. 2007. TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. *Breast cancer research : BCR*. 9 (3), ppR30.

- Lapidus, R. G., Nass, S. J. and Davidson, N. E. 1998. The loss of estrogen and progesterone receptor gene expression in human breast cancer. *Journal of mammary gland biology and neoplasia*. 3 (1), pp85-94.
- Law, M. L., Kao, F. T., Wei, Q. 1987. The progesterone receptor gene maps to human chromosome band 11q13, the site of the mammary oncogene int-2. *Proceedings of the National Academy of Sciences of the United States of America*. 84 (9), pp2877-2881.
- Le, X. F., Pruefer, F. and Bast, R. C., Jr. 2005. HER2-targeting antibodies modulate the cyclin-dependent kinase inhibitor p27Kip1 via multiple signaling pathways. *Cell cycle (Georgetown, Tex.)*. 4 (1), pp87-95.
- Lebeau, A., Grob, T., Holst, F. 2008. Oestrogen receptor gene (ESR1) amplification is frequent in endometrial carcinoma and its precursor lesions. *The Journal of pathology*. 216 (2), pp151-157.
- Lee, H. K., Hsu, A. K., Sajdak, J. 2004. Coexpression analysis of human genes across many microarray data sets. *Genome research*. 14 (6), pp1085-1094.
- Levin, E. R. 2003. Bidirectional signaling between the estrogen receptor and the epidermal growth factor receptor. *Molecular endocrinology (Baltimore, Md.)*. 17 (3), pp309-317.
- Levin, E. R. 2005. Integration of the extranuclear and nuclear actions of estrogen. *Molecular endocrinology (Baltimore, Md.)*. 19 (8), pp1951-1959.
- Li, C. and Hung Wong, W. 2001. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome biology*. 2 (8), ppRESEARCH0032.
- Li, C. Y., Shan, S., Cao, Y. and Dewhirst, M. W. 2000. Role of incipient angiogenesis in cancer metastasis. *Cancer metastasis reviews*. 19 (1-2), pp7-11.

Li, C., Lin, M. and Liu, J. 2004. Identification of PRC1 as the p53 target gene uncovers a novel function of p53 in the regulation of cytokinesis. *Oncogene*. 23 (58), pp9336-9347.

Li, F., Adam, L., Vadlamudi, R. K. 2002. p21-activated kinase 1 interacts with and phosphorylates histone H3 in breast cancer cells. *EMBO reports*. 3 (8), pp767-773.

Li, J., Ding, S. F., Habib, N. A. 1994. Partial characterization of a cDNA for human stearyl-CoA desaturase and changes in its mRNA expression in some normal and malignant tissues. *International journal of cancer. Journal international du cancer*. 57 (3), pp348-352.

Li, X. and O'Malley, B. W. 2003. Unfolding the action of progesterone receptors. *The Journal of biological chemistry*. 278 (41), pp39261-39264.

Li, X., Huang, J., Yi, P. 2004. Single-chain estrogen receptors (ERs) reveal that the ERalpha/beta heterodimer emulates functions of the ERalpha dimer in genomic estrogen signaling pathways. *Molecular and cellular biology*. 24 (17), pp7681-7694.

Li, Z., Li, W., Meklat, F. 2007. A yeast two-hybrid system using Sp17 identified Ropporin as a novel cancer-testis antigen in hematologic malignancies. *International journal of cancer. Journal international du cancer*. 121 (7), pp1507-1511.

Li, Z., Li, W., Meklat, F. 2007a. A yeast two-hybrid system using Sp17 identified Ropporin as a novel cancer-testis antigen in hematologic malignancies. *International journal of cancer. Journal international du cancer*. 121 (7), pp1507-1511.

Li, Z., Li, W., Meklat, F. 2007b. A yeast two-hybrid system using Sp17 identified Ropporin as a novel cancer-testis antigen in hematologic malignancies. *International journal of cancer. Journal international du cancer*. 121 (7), pp1507-1511.

Liang, Z., Sun, Z. Y., Yuan, Y. H. 2005. The expression of 11 cancer/testis (CT) antigen genes in esophageal carcinoma. *Zhonghua zhong liu za zhi [Chinese journal of oncology]*. 27 (9), pp534-537.

Linn, F., Heidmann, I., Saedler, H. and Meyer, P. 1990. Epigenetic changes in the expression of the maize A1 gene in *Petunia hybrida*: role of numbers of integrated gene copies and state of methylation. *Molecular & general genetics : MGG*. 222 (2-3), pp329-336.

Lo, J. S., Snow, S. N., Reizner, G. T. 1991. Metastatic basal cell carcinoma: report of twelve cases with a review of the literature. *Journal of the American Academy of Dermatology*. 24 (5 Pt 1), pp715-719.

Ma, X. J., Salunga, R., Tuggle, J. T. 2003. Gene expression profiles of human breast cancer progression. *Proceedings of the National Academy of Sciences of the United States of America*. 100 (10), pp5974-5979.

Ma, Z. Q., Liu, Z., Ngan, E. S. and Tsai, S. Y. 2001. Cdc25B functions as a novel coactivator for the steroid receptors. *Molecular and cellular biology*. 21 (23), pp8056-8067.

MacKay, David J. C. Information theory, inference and learning algorithms. *Cambridge University Press*, 2003.

Magnusson, S., Borg, A., Kristoffersson, U. 2008. Higher occurrence of childhood cancer in families with germline mutations in BRCA2, MMR and CDKN2A genes. *Familial cancer*.

Makretsov, N. A., Huntsman, D. G., Nielsen, T. O. 2004. Hierarchical clustering analysis of tissue microarray immunostaining data identifies prognostically significant groups of breast carcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 10 (18 Pt 1), pp6143-6151.

Mallepell, S., Krust, A., Chambon, P. and Briskin, C. 2006. Paracrine signaling through the epithelial estrogen receptor alpha is required for proliferation and morphogenesis in the mammary gland. *Proceedings of the National Academy of Sciences of the United States of America*. 103 (7), pp2196-2201.

Mangelsdorf, D. J., Thummel, C., Beato, M. 1995. The nuclear receptor superfamily: the second decade. *Cell*. 83 (6), pp835-839.

Maqani, N., Belkhiri, A., Moskaluk, C. 2006. Molecular dissection of 17q12 amplicon in upper gastrointestinal adenocarcinomas. *Molecular cancer research : MCR*. 4 (7), pp449-455.

Marie, S. K., Okamoto, O. K., Uno, M. 2008. Maternal embryonic leucine zipper kinase transcript abundance correlates with malignancy grade in human astrocytomas. *International journal of cancer. Journal international du cancer*. 122 (4), pp807-815.

Martin, M. B. and Stoica, A. 2002. Insulin-Like Growth Factor-I and Estrogen Interactions in Breast Cancer. *J.Nutr.* 132 (12), pp3799S-3801. Available from: <<http://jn.nutrition.org/cgi/content/abstract/132/12/3799S>>

Martin, R. C., 2nd, Edwards, M. J., Cawte, T. G. 2000. Basosquamous carcinoma: analysis of prognostic factors influencing recurrence. *Cancer*. 88 (6), pp1365-1369.

Martinez, J., Patkaniowska, A., Urlaub, H. 2002. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell*. 110 (5), pp563-574.

Matzke, M. A., Primig, M., Trnovsky, J. and Matzke, A. J. 1989. Reversible methylation and inactivation of marker genes in sequentially transformed tobacco plants. *The EMBO journal*. 8 (3), pp643-649.

May, F. E., Church, S. T., Major, S. and Westley, B. R. 2003. The closely related estrogen-regulated trefoil proteins TFF1 and TFF3 have markedly different hydrodynamic properties, overall charge, and distribution of surface charge. *Biochemistry*. 42 (27), pp8250-8259.

McLachlan, Geoffrey J. Discriminant analysis and statistical pattern recognition / Geoffrey J. McLachlan. New York ; Chichester : Wiley, 1992.

Mehta, J. P., O'Driscoll, L., Barron, N. 2007. A microarray approach to translational medicine in breast cancer: how representative are cell line models of clinical conditions? *Anticancer Research*. 27 (3A), pp1295-1300.

Menasce, L. P., White, G. R., Harrison, C. J. and Boyle, J. M. 1993. Localization of the estrogen receptor locus (ESR) to chromosome 6q25.1 by FISH and a simple post-FISH banding technique. *Genomics*. 17 (1), pp263-265.

Milde-Langosch, K., Kappes, H., Riethdorf, S. 2003. FosB is highly expressed in normal mammary epithelia, but down-regulated in poorly differentiated breast carcinomas. *Breast cancer research and treatment*. 77 (3), pp265-275.

Milde-Langosch, K., Roder, H., Andritzky, B. 2004. The role of the AP-1 transcription factors c-Fos, FosB, Fra-1 and Fra-2 in the invasion process of mammary carcinomas. *Breast cancer research and treatment*. 86 (2), pp139-152.

Milgraum, L. Z., Witters, L. A., Pasternack, G. R. and Kuhajda, F. P. 1997. Enzymes of the fatty acid synthesis pathway are highly expressed in in situ breast carcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 3 (11), pp2115-2120.

Miyajima, N., Kadowaki, Y., Fukushige, S. 1988. Identification of two novel members of erbA superfamily by molecular cloning: the gene products of the two are highly related to each other. *Nucleic acids research*. 16 (23), pp11057-11074.

Moinfar, F., Okcu, M., Tsybrovskyy, O. 2003. Androgen receptors frequently are expressed in breast carcinomas: potential relevance to new therapeutic strategies. *Cancer*. 98 (4), pp703-711.

Mondal, G., Sengupta, S., Panda, C. K. 2007. Overexpression of Cdc20 leads to impairment of the spindle assembly checkpoint and aneuploidization in oral cancer. *Carcinogenesis*. 28 (1), pp81-92.

Monnerat, C., Chompret, A., Kannengiesser, C. 2007. BRCA1, BRCA2, TP53, and CDKN2A germline mutations in patients with breast cancer and cutaneous melanoma. *Familial cancer*. 6 (4), pp453-461.

Morgan, D. O. 1995. Principles of CDK regulation. *Nature*. 374 (6518), pp131-134.

Mount, S. L. and Taatjes, D. J. 1994. Neuroendocrine carcinoma of the skin (Merkel cell carcinoma). An immunoelectron-microscopic case study. *The American Journal of Dermatopathology*. 16 (1), pp60-65.

Mukku, V. R. and Stancel, G. M. 1985. Regulation of epidermal growth factor receptor by estrogen. *The Journal of biological chemistry*. 260 (17), pp9820-9824.

Murakami, H., Furihata, M., Ohtsuki, Y. and Ogoshi, S. 1999. Determination of the prognostic significance of cyclin B1 overexpression in patients with esophageal squamous cell carcinoma. *Virchows Archiv : an international journal of pathology*. 434 (2), pp153-158.

Nakano, I., Masterman-Smith, M., Saigusa, K. 2008. Maternal embryonic leucine zipper kinase is a key regulator of the proliferation of malignant brain tumors, including brain tumor stem cells. *Journal of neuroscience research*. 86 (1), pp48-60.

Nakano, I., Paucar, A. A., Bajpai, R. 2005. Maternal embryonic leucine zipper kinase (MELK) regulates multipotent neural progenitor proliferation. *The Journal of cell biology*. 170 (3), pp413-427.

Nakatsura, T., Senju, S., Ito, M. 2002. Cellular and humoral immune responses to a human pancreatic cancer antigen, coactosin-like protein, originally defined by the SEREX method. *European journal of immunology*. 32 (3), pp826-836.

Napoli, C., Lemieux, C. and Jorgensen, R. 1990. Introduction of a Chimeric Chalcone Synthase Gene into *Petunia* Results in Reversible Co-Suppression of Homologous Genes in trans. *The Plant Cell*. 2 (4), pp279-289.

Newell, A. E., Fiedler, S. E., Ruan, J. M. 2008. Protein kinase A RII-like (R2D2) proteins exhibit differential localization and AKAP interaction. *Cell motility and the cytoskeleton*. 65 (7), pp539-552.

Newell, A. E., Fiedler, S. E., Ruan, J. M. 2008. Protein kinase A RII-like (R2D2) proteins exhibit differential localization and AKAP interaction. *Cell motility and the cytoskeleton*. 65 (7), pp539-552.

Nheu, T., He, H., Hirokawa, Y. 2004. PAK is essential for RAS-induced upregulation of cyclin D1 during the G1 to S transition. *Cell cycle (Georgetown, Tex.)*. 3 (1), pp71-74.

Nicolson, G. L. 1988. Organ specificity of tumor metastasis: role of preferential adhesion, invasion and growth of malignant cells at specific secondary sites. *Cancer metastasis reviews*. 7 (2), pp143-188.

Nicolson, G. L. 1998. Breast cancer metastasis-associated genes: role in tumour progression to the metastatic state. *Biochemical Society symposium*. 63pp231-243.

Noel, A. and Foidart, J. M. 1998. The role of stroma in breast carcinoma growth in vivo. *Journal of mammary gland biology and neoplasia*. 3 (2), pp215-225.

Normanno, N., Di Maio, M., De Maio, E. 2005. Mechanisms of endocrine resistance and novel therapeutic strategies in breast cancer. *Endocrine-related cancer*. 12 (4), pp721-747.

Novina, C. D. and Sharp, P. A. 2004. The RNAi revolution. *Nature*. 430 (6996), pp161-164.

Nurse, P. 1994. Ordering S phase and M phase in the cell cycle. *Cell*. 79 (4), pp547-550.

O'Callaghan-Sunol, C., Gabai, V. L. and Sherman, M. Y. 2007. Hsp27 modulates p53 signaling and suppresses cellular senescence. *Cancer research*. 67 (24), pp11779-11788.

O'Driscoll, L., McMorrow, J., Doolan, P. 2006. Investigation of the molecular profile of basal cell carcinoma using whole genome microarrays. *Molecular cancer*. 5pp74.

Olayioye, M. A. 2001. Update on HER-2 as a target for cancer therapy: intracellular signaling pathways of ErbB2/HER-2 and family members. *Breast cancer research : BCR*. 3 (6), pp385-389.

Olumi, A. F., Dazin, P. and Tlsty, T. D. 1998. A novel coculture technique demonstrates that normal human prostatic fibroblasts contribute to tumor formation of LNCaP cells by retarding cell death. *Cancer research*. 58 (20), pp4525-4530.

Osborne, C. K. 1999. Aromatase inhibitors in relation to other forms of endocrine therapy for breast cancer. *Endocrine-related cancer*. 6 (2), pp271-276.

Oseni, T., Patel, R., Pyle, J. and Jordan, V. C. 2008. Selective estrogen receptor modulators and phytoestrogens. *Planta Medica*. 74 (13), pp1656-1665.

Overgaard, J., Yilmaz, M., Guldborg, P. 2000. TP53 mutation is an independent prognostic marker for poor outcome in both node-negative and node-positive breast cancer. *Acta Oncologica (Stockholm, Sweden)*. 39 (3), pp327-333.

Pace, P., Taylor, J., Suntharalingam, S. 1997. Human estrogen receptor beta binds DNA in a manner similar to and dimerizes with estrogen receptor alpha. *The Journal of biological chemistry*. 272 (41), pp25832-25838.

Paik, S., Shak, S., Tang, G. 2004. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England journal of medicine*. 351 (27), pp2817-2826.

Paik, S., Tang, G., Shak, S. 2006. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 24 (23), pp3726-3734.

- Pal-Bhadra, M., Bhadra, U. and Birchler, J. A. 2002. RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in *Drosophila*. *Molecular cell*. 9 (2), pp315-327.
- Palecek, S. P., Loftus, J. C., Ginsberg, M. H. 1997. Integrin-ligand binding properties govern cell migration speed through cell-substratum adhesiveness. *Nature*. 385 (6616), pp537-540.
- Palmer, H. G., Larriba, M. J., Garcia, J. M. 2004. The transcription factor SNAIL represses vitamin D receptor expression and responsiveness in human colon cancer. *Nature medicine*. 10 (9), pp917-919.
- Paredes, J., Lopes, N., Milanezi, F. and Schmitt, F. C. 2007. P-cadherin and cytokeratin 5: useful adjunct markers to distinguish basal-like ductal carcinomas in situ. *Virchows Archiv : an international journal of pathology*. 450 (1), pp73-80.
- Park, H. R., Min, S. K., Cho, H. D. 2004. Expression profiles of p63, p53, survivin, and hTERT in skin tumors. *Journal of cutaneous pathology*. 31 (8), pp544-549.
- Park, S. O., Zheng, Z., Oppenheimer, D. G. and Hauser, B. A. 2005. The PRETTY FEW SEEDS2 gene encodes an Arabidopsis homeodomain protein that regulates ovule development. *Development (Cambridge, England)*. 132 (4), pp841-849.
- Pawitan, Y., Bjohle, J., Amler, L. 2005. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast cancer research : BCR*. 7 (6), ppR953-64.
- Pennanen, H., Kuittinen, O., Soini, Y. and Turpeenniemi-Hujanen, T. 2008. Prognostic significance of p53 and matrix metalloproteinase-9 expression in follicular lymphoma. *European journal of haematology*. 81 (4), pp289-297.
- Pepper, S. D., Saunders, E. K., Edwards, L. E. 2007. The utility of MAS5 expression summary and detection call algorithms. *BMC bioinformatics*. 8pp273.

Pham, T. T., Selim, M. A., Burchette, J. L., Jr. 2006. CD10 expression in trichoepithelioma and basal cell carcinoma. *Journal of cutaneous pathology*. 33 (2), pp123-128.

Philipp, M. and Caron, M. G. 2009. Hedgehog signaling: is Smo a G protein-coupled receptor? *Current biology : CB*. 19 (3), ppR125-7.

Phillips, K. K., White, A. E., Hicks, D. J. 1998. Correlation between reduction of metastasis in the MDA-MB-435 model system and increased expression of the Kai-1 protein. *Molecular carcinogenesis*. 21 (2), pp111-120.

Pillitteri, L. J., Bemis, S. M., Shpak, E. D. and Torii, K. U. 2007. Haploinsufficiency after successive loss of signaling reveals a role for ERECTA-family genes in Arabidopsis ovule development. *Development (Cambridge, England)*. 134 (17), pp3099-3109.

Pizer, E. S., Jackisch, C., Wood, F. D. 1996. Inhibition of fatty acid synthesis induces programmed cell death in human breast cancer cells. *Cancer research*. 56 (12), pp2745-2747.

Poste, G. and Fidler, I. J. 1980. The pathogenesis of cancer metastasis. *Nature*. 283 (5743), pp139-146.

Prest, S. J., May, F. E. and Westley, B. R. 2002. The estrogen-regulated protein, TFF1, stimulates migration of human breast cancer cells. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology*. 16 (6), pp592-594.

Price, J. T., Bonovich, M. T. and Kohn, E. C. 1997. The biochemistry of cancer dissemination. *Critical reviews in biochemistry and molecular biology*. 32 (3), pp175-253.

Provost, P., Doucet, J., Stock, A. 2001. Coactosin-like protein, a human F-actin-binding protein: critical role of lysine-75. *The Biochemical journal*. 359 (Pt 2), pp255-263.

Pruitt, K. D., Tatusova, T. and Maglott, D. R. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*. 35 (Database issue), ppD61-5.

Putti, T. C., El-Rehim, D. M., Rakha, E. A. 2005. Estrogen receptor-negative breast carcinomas: a review of morphology and immunophenotypical analysis. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc.* 18 (1), pp26-35.

Quenel, N., Wafflart, J., Bonichon, F. 1995. The prognostic value of c-erbB2 in primary breast carcinomas: a study on 942 cases. *Breast cancer research and treatment*. 35 (3), pp283-291.

Rae, J. M., Creighton, C. J., Meck, J. M. 2007. MDA-MB-435 cells are derived from M14 melanoma cells--a loss for breast cancer, but a boon for melanoma research. *Breast cancer research and treatment*. 104 (1), pp13-19.

Ranno, S., Motta, M., Rampello, E. 2006. The chromogranin-A (CgA) in prostate cancer. *Archives of Gerontology and Geriatrics*. 43 (1), pp117-126.

Rashid, A., Pizer, E. S., Moga, M. 1997. Elevated expression of fatty acid synthase and fatty acid synthetic activity in colorectal neoplasia. *The American journal of pathology*. 150 (1), pp201-208.

Ravaioli, A., Bagli, L., Zucchini, A. and Monti, F. 1998. Prognosis and prediction of response in breast cancer: the current role of the main biological markers. *Cell proliferation*. 31 (3-4), pp113-126.

Raychaudhuri, S., Stuart, J. M. and Altman, R. B. 2000. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. pp455-466.

- Rejeb, I., Saillour, Y., Castelnau, L. 2008. A novel splice mutation in PAK3 gene underlying mental retardation with neuropsychiatric features. *European journal of human genetics : EJHG*.
- Revillion, F., Bonneterre, J. and Peyrat, J. P. 1998. ERBB2 oncogene in human breast cancer and its clinical significance. *European journal of cancer (Oxford, England : 1990)*. 34 (6), pp791-808.
- Rhee, J., Han, S. W., Oh, D. Y. 2008. The clinicopathologic characteristics and prognostic significance of triple-negativity in node-negative breast cancer. *BMC cancer*. 8pp307.
- Richardson, A. L., Wang, Z. C., De Nicolo, A. 2006. X chromosomal abnormalities in basal-like human breast cancer. *Cancer cell*. 9 (2), pp121-132.
- Roman, S. D., Clarke, C. L., Hall, R. E. 1992. Expression and regulation of retinoic acid receptors in human breast cancer cells. *Cancer research*. 52 (8), pp2236-2242.
- Rosa, F. E., Caldeira, J. R., Felipes, J. 2008. Evaluation of estrogen receptor alpha and beta and progesterone receptor expression and correlation with clinicopathologic factors and proliferative marker Ki-67 in breast cancers. *Human pathology*. 39 (5), pp720-730.
- Ross, D. T. and Perou, C. M. 2001. A comparison of gene expression signatures from breast tumors and breast tissue derived cell lines. *Disease markers*. 17 (2), pp99-109.
- Rubin, A. I., Chen, E. H. and Ratner, D. 2005. Basal-cell carcinoma. *The New England journal of medicine*. 353 (21), pp2262-2269.
- Rumelhart DE, Hinton GE and Williams RJ. 1986. Learning representations by back-propagating errors. *Nature*, 323, 533-- 536
- Saito, R., Tabata, Y., Muto, A. 2005. Melk-like kinase plays a role in hematopoiesis in the zebra fish. *Molecular and cellular biology*. 25 (15), pp6682-6693.

Saldanha, G., Fletcher, A. and Slater, D. N. 2003. Basal cell carcinoma: a dermatopathological and molecular biological update. *The British journal of dermatology*. 148 (2), pp195-202.

Saldanha, G., Ghura, V., Potter, L. and Fletcher, A. 2004. Nuclear beta-catenin in basal cell carcinoma correlates with increased proliferation. *The British journal of dermatology*. 151 (1), pp157-164.

Sato, Y., Kanno, S., Oda, N. 2000. Properties of two VEGF receptors, Flt-1 and KDR, in signal transduction. *Annals of the New York Academy of Sciences*. 902pp201-5; discussion 205-7.

Scaglia, N. and Igal, R. A. 2005. Stearoyl-CoA desaturase is involved in the control of proliferation, anchorage-independent growth, and survival in human transformed cells. *The Journal of biological chemistry*. 280 (27), pp25339-25349.

Scanlan, M. J., Simpson, A. J. and Old, L. J. 2004. The cancer/testis genes: review, standardization, and commentary. *Cancer immunity : a journal of the Academy of Cancer Immunology*. 4pp1.

Schippinger, W., Regitnig, P., Dandachi, N. 2006. Evaluation of the prognostic significance of androgen receptor expression in metastatic breast cancer. *Virchows Archiv : an international journal of pathology*. 449 (1), pp24-30.

Schmidt, M., Bohm, D., von Torne, C. 2008. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer research*. 68 (13), pp5405-5413.

Schneider, S. M., Offterdinger, M., Huber, H. and Grunt, T. W. 2000. Activation of retinoic acid receptor alpha is sufficient for full induction of retinoid responses in SK-BR-3 and T47D human breast cancer cells. *Cancer research*. 60 (19), pp5479-5487.

Scott, K. L. and Plon, S. E. 2005. CHES1/FOXN3 interacts with Ski-interacting protein and acts as a transcriptional repressor. *Gene*. 359pp119-126.

Sengupta, N. and MacDonald, T. T. 2007. The role of matrix metalloproteinases in stromal/epithelial interactions in the gut. *Physiology (Bethesda, Md.)*. 22pp401-409.

Seth, A. and Watson, D. K. 2005. ETS transcription factors and their emerging roles in human cancer. *European journal of cancer (Oxford, England : 1990)*. 41 (16), pp2462-2478.

Shakhnarovich, Gregory, Trevor Darrell, and Piotr Indyk (eds). "Chapter 2 - Nearest-Neighbor Searching and Metric Space Dimensions". *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*. The MIT Press. © 2005.

Shankavaram, U. T., Reinhold, W. C., Nishizuka, S. 2007. Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. *Molecular cancer therapeutics*. 6 (3), pp820-832.

Sheikh, M. S., Shao, Z. M., Li, X. S. 1994. Retinoid-resistant estrogen receptor-negative human breast carcinoma cells transfected with retinoic acid receptor-alpha acquire sensitivity to growth inhibition by retinoids. *The Journal of biological chemistry*. 269 (34), pp21440-21447.

Sheinin, Y., Kaserer, K., Wrba, F. 2000. In situ mRNA hybridization analysis and immunolocalization of the vitamin D receptor in normal and carcinomatous human colonic mucosa: relation to epidermal growth factor receptor expression. *Virchows Archiv : an international journal of pathology*. 437 (5), pp501-507.

Shen, M., Feng, Y., Gao, C. 2004. Detection of cyclin b1 expression in g(1)-phase cancer cell lines and cancer tissues by postsorting Western blot analysis. *Cancer research*. 64 (5), pp1607-1610.

Sherriff, A. and Ott, J. 2001. Applications of neural networks for gene finding. *Advances in Genetics*. 42pp287-297.

- Shurbaji, M. S., Kuhajda, F. P., Pasternack, G. R. and Thurmond, T. S. 1992. Expression of oncogenic antigen 519 (OA-519) in prostate cancer is a potential prognostic indicator. *American Journal of Clinical Pathology*. 97 (5), pp686-691.
- Simmen, F. A., Su, Y., Xiao, R. 2008. The Kruppel-like factor 9 (KLF9) network in HEC-1-A endometrial carcinoma cells suggests the carcinogenic potential of dys-regulated KLF9 expression. *Reproductive biology and endocrinology : RB&E*. 6pp41.
- Simpson, A. J., Caballero, O. L., Jungbluth, A. 2005. Cancer/testis antigens, gametogenesis and cancer. *Nature reviews.Cancer*. 5 (8), pp615-625.
- Slamon, D. J., Clark, G. M., Wong, S. G. 1987. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science (New York, N.Y.)*. 235 (4785), pp177-182.
- Sluysers, M., Rijkers, A. W., de Goeij, C. C. 1988. Assignment of estradiol receptor gene to mouse chromosome 10. *Journal of steroid biochemistry*. 31 (5), pp757-761.
- Smid, M., Wang, Y., Klijn, J. G. 2006. Genes associated with breast cancer metastatic to bone. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 24 (15), pp2261-2267.
- Smith, A. P., Hoek, K. and Becker, D. 2005. Whole-genome expression profiling of the melanoma progression pathway reveals marked molecular differences between nevi/melanoma in situ and advanced-stage melanomas. *Cancer biology & therapy*. 4 (9), pp1018-1029.
- Smith, C. J., Watson, C. F., Bird, C. R. 1990. Expression of a truncated tomato polygalacturonase gene inhibits expression of the endogenous gene in transgenic plants. *Molecular & general genetics : MGG*. 224 (3), pp477-481.
- Snow, S. N., Sahl, W., Lo, J. S. 1994. Metastatic basal cell carcinoma. Report of five cases. *Cancer*. 73 (2), pp328-335.

Soria, G. and Ben-Baruch, A. 2008. The inflammatory chemokines CCL2 and CCL5 in breast cancer. *Cancer letters*. 267 (2), pp271-285.

Sorlie, T., Perou, C. M., Tibshirani, R. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*. 98 (19), pp10869-10874.

Sorlie, T., Tibshirani, R., Parker, J. 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*. 100 (14), pp8418-8423.

Sotiriou, C., Neo, S. Y., McShane, L. M. 2003. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences of the United States of America*. 100 (18), pp10393-10398.

Sotiriou, C., Wirapati, P., Loi, S. 2006. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*. 98 (4), pp262-272.

Soupene, E. and Kuypers, F. A. 2006. Multiple erythroid isoforms of human long-chain acyl-CoA synthetases are produced by switch of the fatty acid gate domains. *BMC molecular biology*. 7pp21.

Sparano, J. A. and Paik, S. 2008. Development of the 21-gene assay and its application in clinical practice and clinical trials. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 26 (5), pp721-728.

Stark, G. R., Kerr, I. M., Williams, B. R. 1998. How cells respond to interferons. *Annual Review of Biochemistry*. 67pp227-264.

Stekel, D. 2003 *Microarray Bioinformatics*. Cambridge University Press.

Sternlicht, M. D. 2006. Key stages in mammary gland development: the cues that regulate ductal branching morphogenesis. *Breast cancer research : BCR*. 8 (1), pp201.

Stockmans, G., Deraedt, K., Wildiers, H. 2008. Triple-negative breast cancer. *Current opinion in oncology*. 20 (6), pp614-620.

Su, A. I., Wiltshire, T., Batalov, S. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*. 101 (16), pp6062-6067.

Swinnen, J. V., Vanderhoydonc, F., Elgamal, A. A. 2000. Selective activation of the fatty acid synthesis pathway in human prostate cancer. *International journal of cancer*. Journal international du cancer. 88 (2), pp176-179.

Symmans, W. F., Fiterman, D. J., Anderson, S. K. 2005. A single-gene biomarker identifies breast cancers associated with immature cell type and short duration of prior breastfeeding. *Endocrine-related cancer*. 12 (4), pp1059-1069.

Takimoto, G. S. and Horwitz, K. B. 1993. Progesterone receptor phosphorylation complexities in defining a functional role. *Trends in endocrinology and metabolism: TEM*. 4 (1), pp1-7.

Tamoxifen for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group. 1998. *Lancet*. 351 (9114), pp1451-1467.

Tanis, P. J., Nieweg, O. E., Valdes Olmos, R. A. 2001. History of sentinel node and validation of the technique. *Breast cancer research : BCR*. 3 (2), pp109-112.

Taylor, K. M. 2000. LIV-1 breast cancer protein belongs to new family of histidine-rich membrane proteins with potential to control intracellular Zn²⁺ homeostasis. *IUBMB life*. 49 (4), pp249-253.

Teschendorff, A. E., Miremadi, A., Pinder, S. E. 2007. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome biology*. 8 (8), ppR157.

Thai, S. F., Allen, J. W., DeAngelo, A. B. 2001. Detection of early gene expression changes by differential display in the livers of mice exposed to dichloroacetic acid. *Carcinogenesis*. 22 (8), pp1317-1322.

Thorat, M. A., Marchio, C., Morimiya, A. 2008. Forkhead box A1 expression in breast cancer is associated with luminal subtype and good prognosis. *Journal of clinical pathology*. 61 (3), pp327-332.

Tojo, M., Mori, T., Kiyosawa, H. 1999. Expression of sonic hedgehog signal transducers, patched and smoothened, in human basal cell carcinoma. *Pathology international*. 49 (8), pp687-694.

Tomaska, L. and Resnick, R. J. 1993. Suppression of platelet-derived growth factor receptor tyrosine kinase activity by unsaturated fatty acids. *The Journal of biological chemistry*. 268 (7), pp5317-5322.

Tomlinson, V. A., Newbery, H. J., Wray, N. R. 2005. Translation elongation factor eEF1A2 is a potential oncoprotein that is overexpressed in two-thirds of breast tumors. *BMC cancer*. 5pp113.

Tong, Q. and Hotamisligil, G. S. 2007. Developmental biology: cell fate in the mammary gland. *Nature*. 445 (7129), pp724-726.

Tozlu, S., Girault, I., Vacher, S. 2006. Identification of novel genes that co-cluster with estrogen receptor alpha in breast tumor biopsy specimens, using a large-scale real-time reverse transcription-PCR approach. *Endocrine-related cancer*. 13 (4), pp1109-1120.

Turcotte, S., Forget, M. A., Beauseigle, D. 2007. Prostate-derived Ets transcription factor overexpression is associated with nodal metastasis and hormone receptor positivity in invasive breast cancer. *Neoplasia (New York, N.Y.)*. 9 (10), pp788-796.

Tuschl, T., Zamore, P. D., Lehmann, R. 1999. Targeted mRNA degradation by double-stranded RNA in vitro. *Genes & development*. 13 (24), pp3191-3197.

Uren, A., Reichsman, F., Anest, V. 2000. Secreted frizzled-related protein-1 binds directly to Wingless and is a biphasic modulator of Wnt signaling. *The Journal of biological chemistry*. 275 (6), pp4374-4382.

van de Vijver, M. J., He, Y. D., van't Veer, L. J. 2002. A gene-expression signature as a predictor of survival in breast cancer. *The New England journal of medicine*. 347 (25), pp1999-2009.

van der Krol, A. R., Mur, L. A., Beld, M. 1990. Flavonoid genes in petunia: addition of a limited number of gene copies may lead to a suppression of gene expression. *The Plant Cell*. 2 (4), pp291-299.

van 't Veer, L. J., Dai, H., van de Vijver, M. J. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 415 (6871), pp530-536.

Veeck, J., Niederacher, D., An, H. 2006. Aberrant methylation of the Wnt antagonist SFRP1 in breast cancer is associated with unfavourable prognosis. *Oncogene*. 25 (24), pp3479-3488.

Verrijdt, G., Haelens, A., Schoenmakers, E. 2002. Comparative analysis of the influence of the high-mobility group box 1 protein on DNA binding and transcriptional activation by the androgen, glucocorticoid, progesterone and mineralocorticoid receptors. *The Biochemical journal*. 361 (Pt 1), pp97-103.

Vinatzer, U., Dampier, B., Streubel, B. 2005. Expression of HER2 and the coamplified genes GRB7 and MLN64 in human breast cancer: quantitative real-time reverse transcription-PCR as a diagnostic alternative to immunohistochemistry and fluorescence in situ hybridization. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 11 (23), pp8348-8357.

Voduc, D., Cheang, M. and Nielsen, T. 2008. GATA-3 expression in breast cancer has a strong association with estrogen receptor but lacks independent prognostic value. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for*

Cancer Research, cosponsored by the American Society of Preventive Oncology. 17 (2), pp365-373.

von Domarus, H. and Stevens, P. J. 1984. Metastatic basal cell carcinoma. Report of five cases and review of 170 cases in the literature. *Journal of the American Academy of Dermatology.* 10 (6), pp1043-1060.

Vulsteke, V., Beullens, M., Boudrez, A. 2004. Inhibition of spliceosome assembly by the cell cycle-regulated protein kinase MELK and involvement of splicing factor NIPP1. *The Journal of biological chemistry.* 279 (10), pp8642-8647.

Wakefield, L., Robinson, J., Long, H. 2008. Arylamine N-acetyltransferase 1 expression in breast cancer cell lines: a potential marker in estrogen receptor-positive tumors. *Genes, chromosomes & cancer.* 47 (2), pp118-126.

Wallgard, E., Larsson, E., He, L. 2008. Identification of a core set of 58 gene transcripts with broad and specific expression in the microvasculature. *Arteriosclerosis, Thrombosis, and Vascular Biology.* 28 (8), pp1469-1476.

Walling, H. W., Fosko, S. W., Geraminejad, P. A. 2004. Aggressive basal cell carcinoma: presentation, pathogenesis, and management. *Cancer metastasis reviews.* 23 (3-4), pp389-402.

Wang, H., Huang, S., Shou, J. 2006. Comparative analysis and integrative classification of NCI60 cell lines and primary tumors using gene expression profiling data. *BMC genomics.* 7pp166.

Wang, Y., Klijn, J. G., Zhang, Y. 2005. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet.* 365 (9460), pp671-679.

Warnecke, M., Oster, H., Revelli, J. P. 2005. Abnormal development of the locus coeruleus in Ear2(Nr2f6)-deficient mice impairs the functionality of the forebrain clock and affects nociception. *Genes & development.* 19 (5), pp614-625.

Watson, P. H., Chia, S. K., Wykoff, C. C. 2003. Carbonic anhydrase XII is a marker of good prognosis in invasive breast carcinoma. *British journal of cancer*. 88 (7), pp1065-1070.

Watters, A. D., Going, J. J., Cooke, T. G. and Bartlett, J. M. 2003. Chromosome 17 aneusomy is associated with poor prognostic factors in invasive breast carcinoma. *Breast cancer research and treatment*. 77 (2), pp109-114.

Weeraratna, A. T. 2005. A Wnt-er wonderland--the complexity of Wnt signaling in melanoma. *Cancer metastasis reviews*. 24 (2), pp237-250.

Weigelt, B., Glas, A. M., Wessels, L. F. 2003. Gene expression profiles of primary breast tumors maintained in distant metastases. *Proceedings of the National Academy of Sciences of the United States of America*. 100 (26), pp15901-15905.

Weigelt, B., Hu, Z., He, X. 2005. Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. *Cancer research*. 65 (20), pp9155-9158.

Widschwendter, M., Berger, J., Daxenbichler, G. 1997. Loss of retinoic acid receptor beta expression in breast cancer and morphologically normal adjacent tissue but not in the normal breast tissue distant from the cancer. *Cancer research*. 57 (19), pp4158-4161.

Wilson, B. J. and Giguere, V. 2008. Meta-analysis of human cancer microarrays reveals GATA3 is integral to the estrogen receptor alpha pathway. *Molecular cancer*. 7pp49.

Wirapati, P., Sotiriou, C., Kunkel, S. 2008. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast cancer research : BCR*. 10 (4), ppR65.

Wistuba, I. I., Behrens, C., Milchgrub, S. 1998. Comparison of features of human breast cancer cell lines and their corresponding tumors. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 4 (12), pp2931-2938.

Witton, C. J., Reeves, J. R., Going, J. J. 2003. Expression of the HER1-4 family of receptor tyrosine kinases in breast cancer. *The Journal of pathology*. 200 (3), pp290-297.

Wolf, I., Bose, S., Williamson, E. A. 2007a. FOXA1: Growth inhibitor and a favorable prognostic factor in human breast cancer. *International journal of cancer. Journal international du cancer*. 120 (5), pp1013-1022.

Wolf, K., Wu, Y. I., Liu, Y. 2007b. Multi-step pericellular proteolysis controls the transition from individual to collective cancer cell invasion. *Nature cell biology*. 9 (8), pp893-904.

Wong, C. S., Strange, R. C. and Lear, J. T. 2003. Basal cell carcinoma. *BMJ (Clinical research ed.)*. 327 (7418), pp794-798.

Woodhouse, E. C., Chuaqui, R. F. and Liotta, L. A. 1997. General mechanisms of metastasis. *Cancer*. 80 (8 Suppl), pp1529-1537.

Woodward, T. L., Xie, J. W. and Haslam, S. Z. 1998. The role of mammary stroma in modulating the proliferative response to ovarian hormones in the normal mammary gland. *Journal of mammary gland biology and neoplasia*. 3 (2), pp117-131.

Xu, K., Shimelis, H., Linn, D. E. 2009. Regulation of androgen receptor transcriptional activity and specificity by RNF6-induced ubiquitination. *Cancer cell*. 15 (4), pp270-282.

Yaal-Hahoshen, N., Shina, S., Leider-Trejo, L. 2006. The chemokine CCL5 as a potential prognostic factor predicting disease progression in stage II breast cancer patients. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 12 (15), pp4474-4480.

Yada, K., Kashima, K., Daa, T. 2004. Expression of CD10 in basal cell carcinoma. *The American Journal of Dermatopathology*. 26 (6), pp463-471.

Yamazaki, F., Aragane, Y., Kawada, A. and Tezuka, T. 2001. Immunohistochemical detection for nuclear beta-catenin in sporadic basal cell carcinoma. *The British journal of dermatology*. 145 (5), pp771-777.

Yang, C. and Yang, P. 2006. The flagellar motility of Chlamydomonas pf25 mutant lacking an AKAP-binding protein is overtly sensitive to medium conditions. *Molecular biology of the cell*. 17 (1), pp227-238.

Yu, M., Zhan, Q. and Finn, O. J. 2002. Immune recognition of cyclin B1 as a tumor antigen is a result of its overexpression in human tumors that is caused by non-functional p53. *Molecular immunology*. 38 (12-13), pp981-987.

Yuan, R. H., Jeng, Y. M., Pan, H. W. 2007. Overexpression of KIAA0101 predicts high stage, early tumor recurrence, and poor prognosis of hepatocellular carcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 13 (18 Pt 1), pp5368-5376.

Zheng, H., Ying, H., Yan, H. 2008. p53 and Pten control neural and glioma stem/progenitor cell renewal and differentiation. *Nature*. 455 (7216), pp1129-1133.

Zhou, B. P. and Hung, M. C. 2003. Dysregulation of cellular signaling by HER2/neu in breast cancer. *Seminars in oncology*. 30 (5 Suppl 16), pp38-48.

Zhou, D., Quach, K. M., Yang, C. 2000. PNRC: a proline-rich nuclear receptor coregulatory protein that modulates transcriptional activation of multiple nuclear receptors including orphan receptors SF1 (steroidogenic factor 1) and ERRA1 (estrogen related receptor alpha-1). *Molecular endocrinology (Baltimore, Md.)*. 14 (7), pp986-998.

Zhu, Y., Sullivan, L. L., Nair, S. S. 2006. Coregulation of estrogen receptor by ERBB4/HER4 establishes a growth-promoting autocrine signal in breast tumor cells. *Cancer research*. 66 (16), pp7991-7998.

Publications from this study

O'Driscoll L, Kenny E, Mehta JP, Doolan P, Joyce H, Gammell P, Hill A, O'Daly B, O'Gorman D, Clynes M. Feasibility and relevance of global expression profiling of gene transcripts in serum from breast cancer patients using whole genome microarrays and quantitative RT-PCR. *Cancer Genomics Proteomics*. 2008 Mar-Apr;5(2):94-104.

Martinez V, Kennedy S, Doolan P, Gammell P, Joyce H, Kenny E, Prakash Mehta J, Ryan E, O'Connor R, Crown J, Clynes M, O'Driscoll L. Drug metabolism-related genes as potential biomarkers: analysis of expression in normal and tumour breast tissue. *Breast Cancer Res Treat*. 2008 Aug;110(3):521-30. Epub 2007 Sep 27.

Doolan P, Clynes M, Kennedy S, Mehta JP, Crown J, O'Driscoll L. Prevalence and prognostic and predictive relevance of PRAME in breast cancer. *Breast Cancer Res Treat*. 2008 May;109(2):359-65. Epub 2007 Jul 12.

Mehta JP, O'Driscoll L, Barron N, Clynes M, Doolan P. A microarray approach to translational medicine in breast cancer: how representative are cell line models of clinical conditions? *Anticancer Res*. 2007 May-Jun;27(3A):1295-300.

O'Driscoll L, McMorrow J, Doolan P, McKiernan E, Mehta JP, Ryan E, Gammell P, Joyce H, O'Donovan N, Walsh N, Clynes M. Investigation of the molecular profile of basal cell carcinoma using whole genome microarrays. *Mol Cancer*. 2006 Dec 15;5:74.

Kennedy S, Clynes M, Doolan P, Mehta JP, Rani S, Crown J, O'Driscoll L. SNIP/p140Cap mRNA expression is an unfavourable prognostic factor in breast cancer and is not expressed in normal breast tissue. *Br J Cancer*. 2008 May 20;98(10):1641-5. Epub 2008 May 13.