UNIVERSITAT POMPEU FABRA, Barcelona, Spain

Department of Experimental and Health and Life Sciences – CEXS

International PhD Programme BASIC BIOMEDICAL RESEARCH

Pathway oriented steroid hormone-dependent transcriptome analysis

Establishment of a custom cDNA microarray to study hormone signaling in Breast Cancer

Belén Miñana Gómez

Doctoral thesis

Barcelona, 2007

International PhD Programme BASIC BIOMEDICAL RESEARCH

Department of Experimental and Health and Life Sciences (CEXS)

UNIVERSITAT POMPEU FABRA, Barcelona, Spain

Pathway oriented steroid hormone-dependent transcriptome analysis

Establishment of a custom cDNA microarray to study hormone signaling in Breast Cancer

Report presented by

Belén Miñana Gómez

to apply to the Doctoral Degree in Biomedical Research

from the Universitat Pompeu Fabra

Doctoral thesis done under the supervision of Dr. Lauro Sumoy Van Dyck and Dr. Miguel Beato del Rosal from the Department of Genomics and Bioinformatics and Gene Expression, respectively, of the Center for Genomic Regulation (CRG, Barcelona).

Dr. Lauro Sumoy Van Dyck	Dr- Miguel Beato del Rosal	Belén Miñana Gómez
Thesis director	Thesis director	the author

Barcelona, December 2007

To my parents, Rafael and Julia, for their encouragement at all times, to my beloved daughter Lucía, and Álvaro for taking care of me.

Acknowledgements

Many thanks to Dr. Lauro Sumoy and Dr. Miguel Beato for the supervision and critical reviewing of my manuscript. Thank you, Miguel, for admitting me in your weekly group meetings, getting the chance to learn about hormone signaling. Thank you Lauro, for teaching me so many statistical analysis methods and useful discussions. Thank you for your trust.

Many thanks to Dr Mª Jesús Melià and Dr Cecilia Ballaré for precious cell line RNA samples. Thank you for the time spent with me to understand hormone signaling.

Many thanks to Dr Beatriz Bellosillo, Dr Ignaci Tusquets, Dr Corominas, Dr Francesc Solé from the Hospital del Mar (Barcelona) for breast tumor biopsy samples, and many thanks to Raquel Longarón for her RNA preparations of Breast biopsy samples.

I would like also to thank the scientists who encouraged me in the past to enroll in a PhD programme, especially Dr Vladimir Benes for updating me always on bibliography, and Dr Martina Muckenthaler for her catching enthusiasm creating the first microarrays in the early times.

List of contents

Abstract

The purpose of this doctoral thesis is to improve our understanding of the biological pathways involved in breast cancer tumor progression. With this objective, a cDNA microarray platform containing 800 genes was constructed. These genes were chosen because they are in several representatives signaling pathways, namely estrogen and progesterone receptor related pathways, cell cycle, DNA repair, chromatin remodeling, cell proliferation, apoptosis, cell adhesion, cell invasion and angiogenesis. This gene expression platform was validated using, as a model, the endogenous hormone-receptor-expressing epithelial breast cancer cell line T47D with a synthetic progestin R5020 treatment in a time course experiment. The results of this validation experiment had a good correlation with previously published microarray data. Next, in order to identify the most representative signaling pathways in response to these hormones, an analysis of an extended time series of hormonal treatment (progestin and estradiol) was performed using the above mentioned model. Recent algorithms originally designed for microarray analysis in time course experiments were applied in order to investigate the dynamic hormonal response of our model and perform a functional study of the most significant gene expression profiles. Additionally, target genes induced by the action of hormones via cytoplasmatic kinase cascades or by direct genomic pathways were further classified, with the help of specific gene inhibitors or hormone antagonists. SAM (Significance Analysis of Microarrays) was employed as a statistical method for the identification of significant differentially expressed genes between conditions of specific time points of hormone response. Finally, an analysis of the gene expression profile of a group of breast tumors was carried out and good correlation with their clinical-histopatological data was obtained. We focused on those hormone dependent tumors within our set of breast tumor biopsy specimens in order to identify different tumor subclusters with distinctive phenotype characteristics within our population, which could later correlate with clinical outcome. The most significant genes able to discriminate between different tumor phenotypes of our training set of samples were determined applying a “Leave-one-out” cross-validation method of statistical analysis called PAM (Prediction analysis of microarrays). The predictor was further tested on a new incoming set of samples where we determined to which subtype of tumor new samples were allocated and predicted outcome was compared to clinical data, showing how some distinct tumor phenotypes correlate with a poor prognosis. Pathway analysis of the most significant genes belonging to each phenotype was performed to elucidate the most representative biological signaling pathways through which tumor progression might elapse. Breast tumor phenotype gene expression signatures were compared to the gene expression patterns obtained from the hormone treated breast cancer cell line model and significant resemblance with hormone dependant breast tumors was found.

We hope that, in the future, our established cDNA microarray platform or, after appropriate validation, a real time expression profiling platform constructed with a selection of the most differentially expressed genes in each subtype, could be

used as a technique to help improve the diagnosis and prognosis of the breast tumor samples of our sample population.

Keywords: Breast cancer, hormone dependent tumors, steroid hormones, T47D cell line, progestin, R5020, cDNA microarrays, breast tumor, basal subtype, luminal subtype, breast tumor phenotypes, gene expression profiling.

List of abbreviations

18S	ribosomal protein 18S
28S	ribosomal protein 28S
aRNA	antisense RNA
BCA	breast cancer array
BCC	breast cancer cells
BGA	between groups analysis
BRCA1	breast cancer associated 1
BRCA2	breast cancer associated 2
BSA	bovine serum albumine
cat.no.	catalogue number
cDNA	complementary DNA
CA	correspondence analysis
CRG	Center for Genomic Regulation, Barcelona, Spain
Cy3	fluorochrom dye Cyanine Cy3
Cy5	fluorochrom dye Cyanine Cy5
dA	desoxyadenosine
dC	desoxycytidine
DCC	dextran-coated charcoal-treated
ddH_2O	double distilled water
DEPC	diethyl pyrocarbonate
dG	desoxiguanidine
DNA	deoxyribonucleic acid
dNTPs	deoxinuclotide mix
dT	deoxythymidine
DTT	dithiothreitol
dUTP	deoxy uridinetriphosphate
E2	estradiol
EA	enrichment analysis
EB	elution buffer
EDGE	extraction of differential gene expression
EDTA	Ethylenedinitrilotetraacetic acid
EMBL	European Molecular Biology Laboratory
ER	estrogen receptor
ES	enrichment score
ERBB2	V-erb-b2 erythroblastic leukemia viral oncogene homolog 2
ESTs	expressed sequence tags
FA	factor analysis
FADA	full analysis of microarrays
FBS	Fetal bovine serum
FDR	False Discovery Rate
FISH	fluorescent in situ hybridization
FWER	family wise error rate
gDNA	genomic DNA
GO	gene ontology
GSEA	gene set enrichment analysis
HER2	protein for the ERBB2/*neu* gene
hr	hour

HPLC	high performance liquid chromatography
HRE	hormone response element
HUGO	Human Genome Organization
ICI	ICI1822780, commercial name Fulvestrant
IHC	immunohistochemistry
IMAGE	Integrated Molecular Analysis of Genomes and their Expression Consortium
IMIM	Institut Municipal d'Investigació Mèdica
IPA	Ingenuity Pathway Analysis
kb	kilobase
KEGG	Kyoto encyclopedia of genes and genomes
LOH	loss of heterozygosity
LIMMA	Bioconductor package, linear models for microarray analysis
M	molar
M	$log_2 Ratio$
M-A plot	plot of *M* versus *A*
milliQ	high quality grade purified water
min	minute
mJ	millijoules
ml	milliliter
mM	millimolar
MMARGE	microarray report generator from Genepix replicated experiments
MMTV	mouse mammary tumor virus
mRNA	messenger RNA
MsigDB	molecular signature database
NES	normalized enrichment score
NF1	nuclear factor 1
ng	nanogram
nm	nanometer
nM	nanomolar
PAM	prediction analysis of microarrays
PCA	principal component analysis
PCR	polymerase chain reaction
PD	PD98059
pmol	picomol
PMT	photomultiplier tube
PR	progesterone receptor
qPCR	quantitative PCR
RefSeq	reference sequence accession number
R5020	artificial progesterone, commercially as promegestone
RIN	RNA integrity number
RNA	ribonucleic acid
rpm	revolutions per minute
RPMI	Roswell Park Memorial Institute medium (red phenol free)
RT	reverse transcription
RT-PCR	reverse transcription polymerase chain reaction
RZPD	Deutsches Ressourcenzentrum für genomforschung GmbH
s	second
SAM	significance analysis of microarrays

SSC	salt - sodium citrate
SDS	sodium dodecyl sulfate
SH	steroid hormones
SHR	steroid hormone receptors
TIFF	tagged image file format
TIGR	The Institute for Genome Research (currently Craig Venter Institute)
TMEV	TIGR multiple experiment viewer
TNM	tumor-lymph node-metastasis
TP53	tumor protein 53
U	unit
UGRepAc	UniGene Repository Accession Number
UHRR	universal human reference RNA
UV	ultraviolet
var	variance
μg	microgram
μL	microliters
μM	micromolar
SRC	steroid receptor coactivator

1 Introduction

Breast Cancer is the most prevalent non-skin cancer in the world and the second leading cause of cancer related death in women. One in eight women is expected to be diagnosed with breast cancer sometime during their lives. Breast cancer accounts for approximately 30% of all cancers diagnosed and about 16% of all cancer deaths (Feuer EJ et al. 1993)

In Spain about 16,000 women are diagnosed with breast cancer yearly; from those, more than 6,000 will die from metastatic disease. The mortality index has decreased in the last 5 years, 1,4 % annually, possibly due to the early detection campaigns.

Mortality from breast cancer results from the ability of some tumors to metastasize to distant sites. Selecting patients with micrometastasis at diagnosis is crucial for clinicians in deciding who should receive toxic and expensive adjuvant chemotherapy to eradicate these metastatic cells. Axilliary node status, the best marker available, is still an imperfect indicator, since about 25% of node-negative patients still carry micrometastasis and are destined to recur even without adjuvant treatment after many years of follow-up.

The primary treatments of localized breast cancer are either 1) complete tumor excision and radiation, or 2) mastectomy, with or without radiotherapy. The addition to the primary treatments of localized breast cancer, systemic adjuvant therapies (chemo, endocrine or Trastuzumab), which are designed to control micrometastatic disease, has been shown to increase the chance of long-term survival (Colozza M *et al.* 2006). Thus, adjuvant systemic therapy represents a standard option for most patients with localized breast cancer.

Various clinical and pathological factors, such as age, menopausal status, tumor size, histological grade, lymph vascular invasion, estrogen receptor (ER) and ERBB2 receptor status, have been carefully evaluated as prognostic indicators of clinical course. Most of these variables are combined into prediction models such as the Nottingham Prognostic Index (NPI) and Adjuvant! Online (http://www.adjuvantonline.com) or included in algorithms used for the development of guidelines for treatment decision-making, such as the proposed by the St. Gallen consensus expert panel (Goldhirsh *et al* 2005, and 2006).

Nevertheless, despite providing valuable information about the risk of recurrence, such prognostic indicators have only limited ability to predict individual patient outcomes. Indeed, patients with the same clinical-pathological parameters can have markedly different clinical courses. In addition, these prognostic indicators are derived from the analysis of patient cohorts, which offer no information about treatment effects on outcome. In other words, models and guidelines based on prognostic factors provide too little information that is of use for determining an individual patient's needs for systemic adjuvant therapy.

Sometimes, over-treatment of many patients, in whom cure would have been achieved without chemotherapy or even endocrine treatment, sometimes have significant side effects, such as as cardiotoxicity, neurotoxicity and secondary

cancers.(http://www.cancer.org; Asociación Española contra el cancer (AECC): http://www.anticancer.org).Therefore, better tools are needed for improved diagnosis and prognosis.

Expression levels of most genes individually have not proven powerful enough for routine clinical use to predict distant metastasis over the lifetime of patient. Recent developments in applying microarray technologies to breast tumor samples suggest that these new techniques, by providing measures of expression of multiple genes at once, may allow the transfer of molecular biological discoveries to a clinical application. The hypothesis is that genomic profiles generated using microarrays could predict a more accurately long-term outcome of individual breast cancer patients.

1.1 The biology of the breast

The breast is a glandular organ. It is a network of mammary ducts, which lead to lobes that are made up of lobules. The lobules contain cells that secrete milk and are stimulated by the ovarian hormones estrogen and progesterone, required for proliferation and morphogenesis of the normal mammary gland. Estrogen drives ductal development during puberty, whereas estrogen and progesterone together mediate the proliferative and morphological changes of ductal side-branching and alveologenesis that occur at sexual maturity and during pregnancy (Wooward *et al.* 1998, Fendric *et al*. 1998). Progesterone is a mitogen in the premenopausal and postmenopausal human breast. Progestins are compounds that demonstrate progesterone-like activity and are used in oral contraception, hormone therapy, and treatment of some reproductive disorders. The greater risk of breast cancer in postmenopausal women receiving combined estrogen plus progestin hormone replacement therapy than in those receiving estrogen alone indicates a significant role for progesterone in mammary carcinogenesis (Stadel 2002).

Estrogen and progesterone dependent proliferation and morphogenesis in the epithelium depends on the surrounding stromal cell, and it is modulated by differential expression of specific growth factor and extracellular matrix proteins (Haslam and Woodward et al. 2003).

1.2 Mechanisms of action of steroid hormones

Steroid hormones (SH) including estrogens, progestins, androgens, glucocorticoids and mineralcorticoids) regulate many physiological processes in target cells which contain the corresponding steroid hormone receptors (SHR). SHR are intracellular transcription factors that can be activated in many ways, owing to the fact that they are able to bind high-affinity ligands. SHR-ligand complexes can translocate to the nucleus, bind to hormone response elements (HRE) in promoters present in chromatin DNA, activating or repressing transcription of target genes. SHR can also regulate the activity of many genes through protein-protein interactions with other sequence-specific transcription factors bound to their target sequences. Sometimes transactivation by SHR

often requires a synergistic interaction with other sequence-specific transcription factors (Beato *et al.* 1995) and finally SH response can also be cytoplasmatic through cross-talk with other signal transduction pathways activating protein kinase cascades to nuclear transcription factors which activate various target genes (Beato and Klug 2000).

Progestins act via progesterone receptors (PR), which belong to the nuclear steroid receptor ligand-activated transcription factor superfamily. There are two receptor isoforms PR-A and PR-B, transcribed by the same gene, under the regulation of two distinct promoters. PR-A contains a DNA binding domain, a ligand binding domain, and two transcription activation motifs: the C-terminal, ligand dependent AF-2 and the N-terminal, more constitutively active AF-1. The more transcriptionally potent B isoform of PR (PR-B) contains an additional N-terminal activating function AF-3. As a result, PR-A and PR-B have differing transcriptional activities (Vegeto *et al*. 1993, Tung *et al*. 1993, McDonell *et al.* 1994).

SH can act through several mechanisms (**Figure 1**). In the classical action of SHR, binding of progesterone to PR, causes receptor homodimerization, nuclear translocation, binding to HRE in promoters of target genes activating transcription (Piña *et al.* 1990). Ligand-occupied PR binds directly to DNA at progesterone response elements (Beato *et al*. 1989). Also, liganded PR can activate transcription of genes whose promoters lack HREs by acting as a bridge between transcription factors and coactivators recruited at promoters containing AP-1 and SP1 sites (Owen *et al.* 1998, Bamberger *et al.* 1996, Wardell *et al.* 2002).

In addition to direct genomic effects, progestins can activate a cytoplasmic membrane-associated PR-B isoform, via its additional N-terminal motif, and crosstalk with membrane receptor ER, activating the cytoplasmatic MAP kinase cascade. Estrogens can activate the Src/p21ras/Erk and the PI3K/Akt pathways via direct interaction of the estrogen receptor ER with c-Src and the regulatory subunit of PI3K, respectively (Castoria *et al.* 2001, Migliaccio *et al.* 1996). This results in the activation of ERK1/2, which is imported into the nucleus and phosphorylates a variety of substrates, including transcription factors such as FOS (c-fos), MYC (c-myc), JUN (c-jun), or indirectly activating transcription of different target genes such as CCND1 (Cyclin D1)/CDK4 promoting cell cycle progression (Miglaccio *et al.* 1998, Cicatello *et al.* 2004, Castoria *et al.* 1999).

SH can act by non-genomic mechanisms involving crosstalk with growth factor receptors and other cytoplasmatic signaling pathways (Lange *et al*. 1991, Lösel and Wehling 2003). PR-dependent transcriptional specificity depends on the PR isoforms and coregulators available in a target cell (Vegeto *et al.* 1993, Tung *et al.* 1993, McDonell and Goldman 1994, Richer *et al.* 2002). A functional difference between PR-A and PR-B is that PR-A can act as a dominant repressor of both PR-B and ER in a promoter- and cell type-specific manner.

Progestins can also crosstalk to kinase cascades through a direct interaction of PR with c-Src (Ballaré *et al.* 2003), which is activated in the absence of

estrogens and triggers activation of the MAP kinase cascades (Miglaccio *et al.* 1998). The ultimate targets of the activated kinase cascades are not known but likely include transcription factors and co-regulators (Bjornstrom ans Sjoberg 2005). There is also a direct connection between rapid kinase activation and gene induction by steroid hormones. The activation of Erk and Msk1 results in the recruitment of phosphorylated PR to the MMTV promoter leading to phosphoacetylation of histone H3, thus the non-genomic and genomic pathways converge on chromatin to enable gene regulation (Vincent *et al.* 2006).

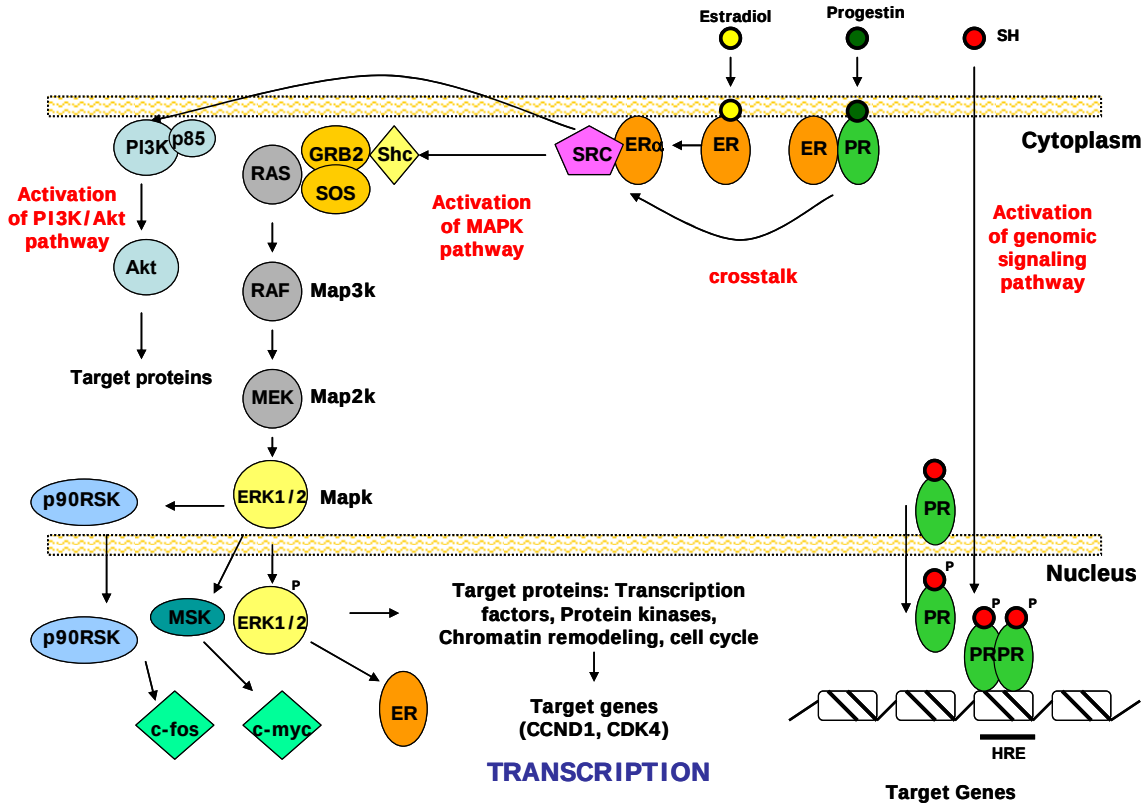


Figure 1: Genomic and ligand-mediated signaling effects of steroid hormones.

1.3 Breast cancer pathology

The most common histological types of invasive breast cancer are infiltrating ductal carcinoma (70-80%), which is caused by an abnormal proliferation in the epithelium of the ducts, and infiltrating lobular carcinoma (10%), which begins in the lobules of the mammary glands. Most cancer types can be classified by clinical stage and pathological subtype. Clinical staging is according to the TNM criteria used as indicators of breast cancer recurrence and overall survival. In the case of breast cancer, the most significant risk factors are large tumor size (T), lymph node positive status (N) and presence of distant metastases (M). These morphological classifications have only a minor prognostic significance compared to tumor grade (determined by the mitotic count or state of cell proliferation) which can be more valuable.

In most cases, death results from metastasis of breast cancer cells (BCC) which migrate to other vital organs such as bone and lungs. Elucidating the mechanisms that make BCC able to migrate and metastasize, remains a major research challenge. These cells are genetically unstable; the classical hypothesis is that they sequentially acquire genetic and phenotypic alterations in a single cell followed by clone selection and expansion, acquiring cell aggressiveness with alterations in properties such as cell proliferation, cell adhesion, angiogenic ability and loss of estrogen receptor (ER).

1.4 Molecular markers in breast cancer

The pathogenesis of this disease is thought to involve multiple genetic and epigenetic events. The molecular markers routinely used in breast cancer diagnosis, are ER, PR, TP53 (p53) and ERBB2 (also know as the Her-2/*neu* oncogene). These markers are commonly scored by immunohistochemistry (determination of the presence of the protein) in the form of a percentage (percentage of positive cells in a visual field). These molecular markers have formed the basis of the three molecular classes of breast cancer recognized in medicine: 1) hormone receptor positive tumors, 2) ERBB2 positive tumors, and 3) tumors negative for both markers. These markers have helped significantly in the diagnosis and treatment of breast cancer during the past three dacades.

Predictive markers can be defined as factors that indicate sensitivity or resistance to a specific treatment. These are often confused with prognostic markers. Both types of markers are used to provide information on the likely future behavior of a tumor, but whereas predictive markers are used to prospectively select responsiveness or avoid resistance to a specific treatment, prognostic factors provide information on outcome independently of systemic adjuvant therapy. Some markers can have both predictive and prognostic utility, such as ER, that not only predicts response to endocrine therapy but also correlates with good prognosis, at least in the short term.

The presence of Estrogen receptor (ER) or Progesterone receptor (PR) is a predictive marker that tumors are likely to respond to endocrine therapy, and predicts response to anti-estrogens (e.g., Tamoxifen), aromatase inhibitors (e.g., Anastrozole and letrozole), and luteinizing hormone-releasing hormone agonists (e.g., goserelin). About 70% of ER/PR-positive tumors will respond to Tamoxifen, whereas only 34% of ER-positive/PR-negative, and 45% of cases in ER-negative/PR-positive (Clarke *et al.* 2001).

However, the measurement of ER and PR alone is more complex, due to the impact of the two isoforms of ER (ERα and ERβ) and PR (PR-A and PR-B), as well as several variant and mutant forms, and the “cross-talk” with growth factor and other cell-signaling pathways. Therefore, analysis of additional tumor biomarkers is needed to better classify the various phenotypes of breast tumors.

Bardou *et al.* (2003) showed that the combined measurement of ER and PR is superior to ER alone in predicting benefit from adjuvant hormone therapy. They showed that it can be especially useful identifying ER+ PR- tumors which have worse prognosis than ER+ PR+ tumors. ER+ PR- tumors are less responsive to hormonal treatment (Anti-estrogens), show early Tamoxifen resistance (Arpino *et al.* 2005), and may benefit from treatment with Aromatase Inhibitors (Anastrozole), which suppress tumor and plasma estrogen levels by blocking testosterone conversion to estrogen, avoiding unnecessary costly and toxic antiestrogen treatment. (Baum *et al.* 2002, Smith *et al.* 2003, Bardou *et al.* 2003, Dowsett *et al*. 2003, Jordan *et al.* 2004, Schiff *et al.* 2004, Ross *et al.* 1998, Fuqua *et al.* 2004, Tovey *et al.* 2005, Ellis *et al.* 2005, Osborne *et al.* 2005). The contribution of PR to ER may also depend on the relative amounts of the two forms of PR present. Hopp *et al.* (2004) reported that patients with high PR-A:PR-B ratios in their breast cancer responded poorly to adjuvant therapy. Genes that are known to confer susceptibility to developing breast cancer also affect expression of PR: BRCA1 or BRCA2 mutation results in PR-A predominance. Therefore, changes in progesterone signaling may be involved in the increased risk of cancer observed in women with BRCA1 or BRCA2 mutations (for a review on the clinical significance in breast cancer of estrogen and progesterone receptor isoforms see Fuqua *et al*. 2005).

The HER-2 protein, also known as c-erbB-2 or *neu* is a member of subclass 1 of the superfamily of receptor tyrosine kinases. Other members include epidermal growth factor receptor (HER-1), HER-3, and HER-4. All these proteins possess an extracellular ligand-binding domain, a membrane spanning region and a cytoplasmatic domain with tyrosine kinase activity (Olayioye *et al*. 2000). ERBB2 can act as receptor for EGF (epidermal growth factor). It can also form a heterodimer with other HER family members. After heterodimerization, HER-2 complexes initiate intracellular signaling via the mitogen-activated protein kinase, phosphatidyl-inositol 3'-kinase, and phospholipase C pathways.

In breast cell lines and model tumor systems, overexpression of the HER-2 gene has been associated with increased mitogenesis, cell motility, invasiveness, and metastasis. ERBB2 has been found amplified and overexpressed in 20-30% in primary invasive tumors of all human breast cancers. With up to 100 copies of the gene that can lead to a larger amount of receptors per cell. Either gene amplification or increased production of HER-2 is generally associated with adverse prognosis, particularly in node-positive breast cancer patients (Winston *et al.* 2004).

The presence of ERBB2 is also a predictive marker of response to trastuzumab (Herceptin©, Genentech Inc.), a monoclonal antibody with specificity for the extracellular domain of the receptor (Slamon *et al.* 2001). Therapy with this antibody has been officially accepted in women with metastasis, with ERBB2-positive tumors, either with overexpression of HER-2 protein scored by immunohistochemistry or with fluorescence *in situ* hybridization (FISH) for gene amplification. However, the initial response of inducing tumor regression is only of about 40% when this agent is used (Albanell and Baselga 2001). Also,

cancers overexpressing HER-2 are likely to benefit from CMF (cyclophosphamide, methotrexate, and 5-fluorouracyl) or anthracycline-based adjuvant therapy.

Other cases of breast cancer can be due to inherited germline mutations in the susceptibility genes BRCA1 or BRCA2, which account for ca. 75% of autosomal dominant breast and ovarian familial cancer. There are other susceptibility genes such as PTEN, TP53, and MYC (Liao and Dickson 2000) that also act as tumor suppressors and are found to be associated with breast cancer.

However, familial breast cancer is rare, accounting for only 5% of all cases (Gayther *et al.* 1998). BRCA1/2 mutations result in a premature protein truncation; when the presence of a mutant allele is accompanied by somatic loss of the wild-type allele (loss of heterozygosity, LOH) this results in complete loss of gene function (Collins *et al.* 1997, Venkitaraman 2002). In very few cases, allele inactivation by promoter hypermethylation (gene silencing) occurs at the BRCA1 locus in somatic breast cancer, but this does not happen at the BRCA2 locus (Dobrovic and Simpfendorfer 1997). The virtual absence of BRCA1 and BRCA2 mutations in sporadic breast cancer is still unexplained.

The most commonly mutated gene in human cancer is the p53 tumor suppressor gene located on chromosome 17p. It is found mutated in a 35% of sporadic breast cancers, and 66% of BRCA1-associated breast cancers, although the mutation spectrum is distinct from that of triple-negative sporadic tumors (Crook *et al.* 1998). Most p53 mutations are missense and occur in the DNA binding domain. The consequence of these mutations is loss of the ability of p53 to bind DNA in a sequence-specific manner. The p53 protein controls the expression of multiple genes that are divided in four categories: cell cycle inhibition, induction to apoptosis, control of genome stability, and inhibition of angiogenesis (Velculescu and El-Deiry 1996).

Also, the MYC (c-myc) gene is amplified in 15-20% of human breast cancers (Bieche *et al.* 1999), however this amplification does not appear to be frequent in sporadic breast cancers (23%). But it has been reported to occur in 53% of BRCA1-mutation-related breast cancer (Grushko *et al.* 2004).

Clinical staging and routine pathology methods for identification of individuals at risk of developing metastases are still relatively basic. Current diagnostic methods to determine who should receive adjuvant chemotherapy often result in “over-treatment” of many patients, who would otherwise have survived with milder interventions, hormone alone or no treatment. whatsoever Other tools are needed to better define which patients have a poor prognosis and would benefit the most from chemotherapy and which, even though being ER-positive, would show tamoxifen resistance (Michalides *et al.* 2004, Ma *et al*. 2004), or how they would react to different treatments.

1.5 Molecular profiling in cancer

Unlike standard methodologies that rely on a few pathological and immunohistochemical markers, molecular expression profiling using microarray technology allows to define tumors by the expression pattern of thousands of genes simultaneously (Macgregor and Squire 2002, Winegarden 2003). The use of microarrays for providing diagnosis and predicting patient outcome has two major advantages: (1) microarrays permit the screening of multiple genes without a previous knowledge of which genes might be predictive, and (2) with microarrays, groups of genes rather than single genes may be a more reliable indicator of clinical response. Therefore, tumors could be better classified based on a combination of genes whose expression level can discriminate efficiently between clinically distinct subtypes of breast tumors which would require different treatment strategies.

Genome-wide expression profiling using DNA microarrays (Schena *et al.* 1995, Schena *et al*. 1996) has been widely applied to the characterization of different cancer diseases whose genetic heterogeneity is not readily resolved by standard clinical diagnostics. Initial studies demonstrated that distinct pathological features could be separated by expression microarrays. Genomic classifications from microarrays have now been developed for many diseases (Golub *et al.* 1999). At the same time gene expression profiling was used to identify systematic phenotypic variation between human cancer cell lines used to screen for anti-cancer drugs (Ross *et al*. 2000).

Early studies on the use of microarrays for predicting anticancer drug response focused on cell line (Scherf *et al.* 2000, Staunton *et al.* 2001). These studies showed that the gene expression profile of untreated cells could be used for chemosensitivity testing. However, only a few studies have been published predicting clinical response or resistance to anticancer agents. Chang *et al.* (2003) found 92 genes were differentially expressed in tumors from patients that were sensitive from those resistant to neoadjuvant (given before surgery) docetaxol therapy. Ayers *et al.* (2004) also used microarrays to identify genes predictive of response to neoadjuvant therapy in patients with breast cancer.

Technical aspects of microarray applications for genomic classification approaches have been reviewed extensively (Schulze and Downward, 2001). Subclassification of tumors by gene expression microarray analysis can be performed in two ways. Microarray data from a selection of clinical samples of tumors can be questioned for groups of samples or “clusters” that are significantly related in terms of their expression profile. Samples that share expression profile features are expected to share phenotypic features, such as, for example, the clinically relevant estrogen receptor (ER) status. This approach is referred to as unsupervised clustering analysis (Quackenbush 2001). In contrast, supervised analysis begins with the designation of the samples into a “labeled” phenotypic subcategory. A search is made to define a list of genes that are distinct in their expression between the two “labeled” groups, belonging to the “training” set, that can subsequently be used to distinguish between them.

The discriminatory accuracy of the list of genes defined in this way can be tested for its ability to separate the samples into the defined groups on an independent set of samples (called “validation set”).

Using unsupervised analysis of microarray data by hierarchical clustering, it is possible to differentiate “signatures” in breast cancer as a dominant pattern of gene expression that represent the origin and function of the predominant cell type, be it epithelial cells, infiltrating lymphocytes, adipose cells or surrounding stromal cells. Perou *et al.* (1999 and 2000) defined the “intrinsic list”: a set of 427 genes that varied significantly in tumors of different individuals but not within tumor pairs of the same individual. These genes were chosen to show that each tumor was unique and identifiable by a molecular “portrait”.

These signatures could be grouped into categories such as cell proliferation, apoptosis, cell adhesion, cell cycle, DNA repair or hormonal receptor status. They defined two major groups and 5 classes: The first group is the ER-positive family which was also named “luminal” class, with a molecular signature with resemblance to the luminal cells of the breast duct (Taylor-Papadimitriou *et al.* 1989) showing high expression of luminal epithelial endocrine specific genes, such as estrogen receptor (ER), X-box binding protein 1 (XBP1), trefoil factor 3 (TFF3), hepatocyte nuclear factor 3a (HNF3a), and estrogen-regulated LIV-1. The second major group is the ER-negative family, in turn composed by at least three classes: (1) (ER- PR-), (2) an ERBB2 (HER-2) positive group with amplification of this gene, and (3) a group of tumors with a signature of genes expressed in “basal”-contractile myoepithelial cells, including cytokeratins 5 (KRT5) and 17 (KRT17), c-kit (KIT), c-myc (MYC), a modulator of wnt signaling (SFRP1) and fatty acid binding protein 7 (FABP7).

West *et al.* (2001) made another classification by a supervised analysis of their microarray data depending on the ER status, and identified 100 genes that could discriminate between an ER+ and ER- tumor, and could classify tumors that were ambiguous in their clinical diagnostic assays by immunohistochemistry (IHC). This revealed the risk of relying only on a single prognostic marker for classification, bringing in a larger dataset that may more accurately predict the diagnosis and clinical outcome. Several other groups demonstrated that supervised data analysis can be used to derive a set of genes that can distinguish ER-positive from ER-negative tumors (Gruvberger *et al.* 2001), and how these molecular subtypes are entirely different disease entities, possibly resembling the precursor cell types.

Sorlie *et al.* (2001) and Sotiriou *et al.* (2003) expanded the classifications of the “luminal-ER-positive” to three other subtypes, one with a favorable prognosis (Luminal A), and two with less favorable prognosis (Luminal B y C). Sotiriou *et al.* (2003) found two subgroups in the “basal” type, one with genes involved in cell cycle and growth, such as PCNA, BUB1, and CDC2, and another one showing higher expression of the transcription factors c-fos, c-jun and fos B. Basal tumors have also been associated with BRCA1 inherited mutations, although it is not a necessary condition to develop this tumor type (Sorlie *et al.* 2003). By using a method called SAM (Significance Analysis of Microarrays),

identified 264 genes, involved in DNA replication, cell division, and genomic stability. They overlapped in 81 with the previous description of luminal, basal and proliferation of the “intrinsic set”. By the use of gene expression microarray profiling, the molecularly distinct subtypes of breast tumors were associated with differences in clinical outcome (Sotiriou *et al.* 2003, Sorlie *et al.* 2003) demonstrating that clinical subtypes derived from unsupervised hierarchical clustering were indeed of clinical significance.

The studies from Van’t Veer *et al.* (2002) and Van de Vijver *et al.* (2002) were the first ones to use gene expression profiling to predict survival in a multivariate analysis. They used a “leave-one-out” method to get a minimal discriminatory set of 70 genes, whose expression pattern identified a group of patients with lymph node negative sporadic tumors whom had not developed metastasis despite of systemic treatment, from the ones whom developed metastasis. They determined a signature that could identify BRCA1 carriers. In a ”poor prognosis” signature they found overexpression of genes involved in cell cycle regulation, cell invasion, metastasis, and angiogenesis.

According to these microarray studies, only a few distinct breast tumor classes seem to exist. This suggests that phenotype transition from one class to another is very unlike to occur in the same tumor during disease progression. Tumor phenotypes seem to be defined very early in development of the lesions. Besides, no “mixed classes” have been observed (Ma *et al.* 2003). This could mean that breast tumors do not really progress, as they could acquire very early the ability to invade and metastasize (Weigelt *et al*. 2003).

Other studies followed the approach to find genes correlating with disease outcome (Huang *et al.* 2003). Nevins *et al*. 2003 introduced the term “metagenes” as a multiple gene expression signature or weighted average measure of expression of defined groups of genes, capable of resolving the biological heterogeneity, together with traditional clinical factors, and achieving a more accurate prediction of outcome for individual patients. Some other studies focused on the clinical response to treatment (Michalides *et al.* 2004, Ma *et al.* 2004). Van Laere *et al.* (2005) employed microarray analysis to distinguish inflammatory breast cancer disease (IBC) from non-IBC with a set of 50 discriminant genes.

More recent studies have focused on a “wound-response signature” in a variety of epithelial tumors, and have revealed links between wound healing and cancer progression, based on the hypothesis that the molecular program of normal wound healing might play an important role in cancer metastasis. Chang *et al*. (2004) previously identified consistent features in the transcriptional response of normal fibroblasts to serum, which they called “core serum response” (CSR). The CSR genes were chosen to minimize overlap with cell cycle genes, and appeared to represent important processes in wound healing such as matrix remodeling, cell motility, and angiogenesis, processes that contribute to cancer invasion and metastasis. Subsequently, they validated the prognostic value of this gene signature and independently predicted the outcome in a large independent dataset (Chang *et al*. 2005). Also, West *et al.*

(2005) used cDNA microarrays and SAM, for differential expression statistical analysis, to determine the stromal signature in breast carcinoma in order to distinguish between two types of tumors with fibroblastic features: solitary fibrous tumor (SFT) and desmoid-type fibromatosis (DTF). They found significant differences in the patterns of expression of extracellular matrix genes and growth factors, besides that DFT group had a more favorable disease outcome.

Other studies had focused on genomic chromosomal aberrancies and changes in DNA copy number (Pollack *et al.* 1999, Pollack *et al.* 2002, Monni *et al.* 2001, Hyman *et al.* 2002). The ERRB2 tumors are seen by fluorescent *in situ* hybridization (FISH) to have an amplification of the chromosomal region that include ERBB2, a tyrosine kinase that acts as an epidermal growth factor (EGF) receptor, but which may also include other genes such as GRB7, GARP, EMSY (Hugues-Davis *et al.* 2003). There are also studies where amplifications have been seen in different chromosomal regions which may contain other familial susceptibility genes (Hedenfalk *et al.* 2001 and 2003, Albertson, 2003, Selaru *et al.* 2004).

Usary et al. (2003) investigated the different mutation variants of GATA3 in human breast ERα positive tumors, and corroborated the studies of Sorlie *et al*. (2001 and 2003) of the Luminal A subtype, which is the subtype associated with the most favorable survival outcome, where there is the highest expression of ERα and GATA3 (Hoch *et al.* 1999).

Recent scientific and technological developments from gene-array technologies enable breast cancers to be classified into prognostic categories depending on the expression of certain genes and gene panels. In February 2007, MammaPrint® (Agendia) became the first multi-gene panel test to be approved by the US Food and Drug Administration (FDA) for predicting breast cancer relapse. MammaPrint was the first customized 60-mer oligo microarray suitable for a high-throughput processing, with 1900 features or spots, containing the 70-prognosis signature from Van't Veer *et al*. (2002), where those genes were spotted in triplicate. This test is suitable for young breast cancer patients (age < 55 years) who are lymph none negative. In Glas *et al.* (2006) study, they hybridized the 162 samples from the previous study from Van't Veer, using as the common reference a breast cancer reference pool, obtaining only 7 discordant cases between MammaPrint® risk assessment and the published data.

Another gene panel, Oncotype DX® (http://www.genomichealth.com/oncotype/about/hcp.aspx), based on Real Time qPCR has been commercially available for the same use since 2004, approved by other regulatory pathway for clinical trials. The facts that both gene panels use different technologies, have a one single gene in common, and were cleared for public use in public by different regulatory agencies is indicative of the disease heterogeneity and challenges that this field must overcome. Onco*type* DX® is a diagnostic test that quantifies the likelihood of disease recurrence in women with early stage breast cancer and assesses the likely benefit from chemotherapy. Onco*type* DX®

analyzes a specific set of genes within a tumor to determine a Recurrence Score®. The Recurrence Score is a number between 0 and 100 that corresponds to a specific likelihood of breast cancer recurrence within 10 years of the initial diagnosis.

A comparison of breast cancer microarray data publications is at **Appendix A1**.

2 Objectives

The main objectives of this work are:

1. To establish a functional Breast Cancer Array cDNA microarray platform in order to investigate the expression profiles of breast cancer cell lines and breast tumor biopsies.
2. To determine new molecular markers that can be used both for diagnosis and prognosis.
3. To analyze the signaling pathways involved in tumor progression in hormonal dependent tumors.
4. To correlate hormonal response of hormone receptor-expressing breast cancer cell lines and hormone dependent breast tumors.
5. The application of different statistical algorithms and current applications for the analysis of microarray gene expression data.

3 Materials and methods

3.1 Clone selection

800 cDNA clones corresponding to genes possibly involved in breast cancer were selected to constitute our cDNA platform. In this platform there are different interlinked biological pathways represented: cell cycle, DNA repair, DNA damage, apoptosis, DNA remodeling, PR related pathways (West *et al. 2001*), endometrial and ovarian cancer related genes, and prognosis genes (Sotiriou *et al. 2002*; Van't Veer *et al. 2002*; Sorlie *et al. 2003*).

Clones were selected from the German Human Genome Resource Center (http://www.rzpd.de/) from different endometrial, ovary or breast cDNA libraries, and a clone information database was constructed using several web resources such as SOURCE (http://source.stanford.edu/cgi-bin/source/ sourceSearch, Diehn *et al. 2003*). The selection criterion was to choose the most 3' end clone, which contained a polyadenylation signal with a maximum length of 2 kb. We visualized these features and confirmed the position of the clone within its cluster using the GeneNest graphical database (http://genenest.molgen.mpg.de). All cDNA clones were confirmed by sequencing from the 5'-end, and their position within the gene was determined by BLAST (http://www.ncbi.nlm.nih.gov/BLAST/). About 10-12% of clones gave a different match after sequencing. New clones for those genes were ordered from the German Human Genome Resource Center, and were subsequently incorporated to the current version of the cDNA microarray platform. See **Appendix A2**: Breast Cancer Array complete clone list and annotations.

An overview of the steps involved in the cDNA clone selection and sequence confirmation, with all the web resources used, is detailed in **Figure 2**.

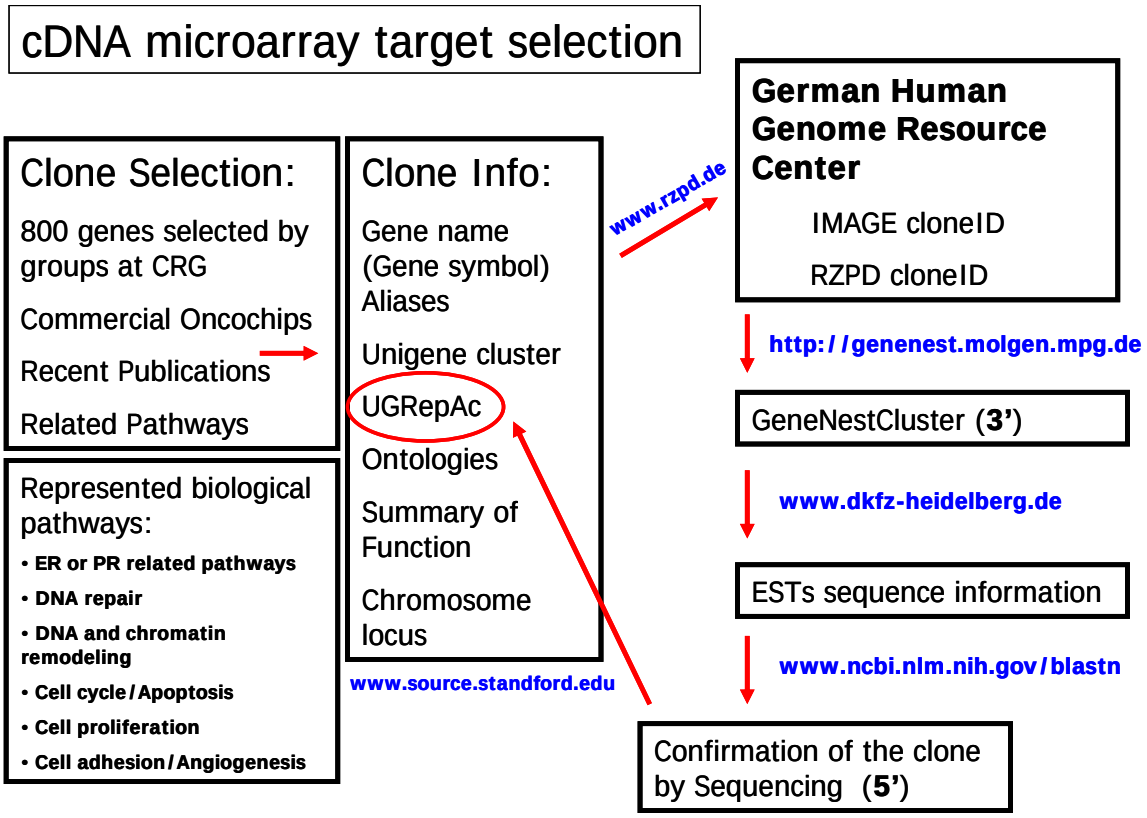


Figure 2: Overview of the steps involved in the cDNA clone selection and sequence confirmation, with all the web resources used.

3.2 Array controls

We incorporated sets of negative and positive microarray controls from EMBL (*Arabidopsis* and bacterial genes, Preiss 2001, Richter *et al*. 2002), Utrecht Microarray Controls (*B. subtilus*, Van der Peppel *et al*. 2003), and the *Amersham Universal ScoreCard* ®, in order to normalize gene expression in cases of global changes, and set a noise level cutoff.

3.3 Microarray construction

DNA plasmid preparations were generated using the Plasmid miniprep 96 kit (Millipore®; cat.no.LSKP09604). Inserts were PCR amplified in a 100 μL total volume reaction using a homemade thermostable *Thermus aquaticus* (Taq) DNA polymerase clone (Desai and Pfaffle, 1995), with vector specific primers. PCR products were purified with the PCR 96 Cleanup kit (Millipore®; cat.no.LSKC09604), eluted in 100 μL milliQ water, and 2 μL of the PCR product were visualized on 96-well format agarose gels using electrophoresis chambers (AbGene®). PCR products were normalized to 100 ng/ μL and allocated on 384-well plates. PCR products were printed in 1x Corning Pronto! ® spotting solution on CORNING UltraGAPS® II amino-modified glass slides, using a Robotic arrayer ChipWriter (Bio-Rad) and a pin-head of 16 SMP3 printing needles (Telechem®). Each spot contains millions of copies of cDNA fragments

from each gene. cDNA spots were immobilized by UV crosslinking at 2500 mJ/s with an UV oven and stored under vacuum in desiccators until they were used for hybridization.

Each gene product was printed in quadruplicates on each microarray, two of them side-by-side on the same row, and the other two in different subgrids, by inverting the spotting plates, on the top or bottom half of the array. Therefore we are using different spotting needles to control intra-slide replicate variation and the specificity of hybridization (see **Appendix A3** for microarray printing design). Replicated spots had a mirrored orientation, which made it easier to control the specificity of hybridization at once just by looking at the raw scanner images. Indeed multiple prints of each clone can be used to control within-array variability and performance owing to spatial effects due to the labeling and hybridization procedure and local artifacts (Tran *et al.* 2002) because they are printed by different spotting pins. Replicate spots in the top and bottom halves of the array are likely to be less well correlated than the side-by-side replicates.

The cDNA platform with the name "CRG Human Breast Cancer Array v4.0-0.8K" can be downloaded from GEO (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL5953) with the accession number GPL5953. GEO is a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated online resource for gene expression data browsing, query and retrieval (Barrett *et al.* 2007).

3.4 Cell cultures

The cell line used in this study was originated from the breast cancer epithelial cell line T47D-MTVL, which endogenously expresses high levels of progesterone and estrogen receptors, with a modification in its genome consisting in of stably integrated copy of the luciferase reporter gene driven by the MMTV promoter (Truss *et al*. 1995). The MMTV promoter contains five HRE (hormone response element) responsive to progestins, androgens and glucocorticoids, but not estrogens (Payvar *et al.* 1983) and an adjacent site for the ubiquitous transcription nuclear factor NF1 (Beato *et al.* 1995).

The progestin-responsive T47D breast cancer cell line has been used for examining progestin-dependent gene expression in vitro (Richer *et al.* 2002, Wan and Nordeen 2002, Bray *et al.* 2003, Bray *et al.* 2005).

Cell cultures were grown to confluence in RPMI 1640 medium (Invitrogen®) supplemented with hormone-free (charcoal-treated) 10% fetal bovine serum FBS, 2 mM L-glutamine, 100 U/ml penicillin and 100 μg/ml streptomycin. Cells were plated in RPMI medium in the absence of phenol red to prevent known estrogenic effects, supplemented with dextran-coated charcoal-treated FBS (DCC/FBS), and 48 hr later medium was replaced by fresh medium without serum. After 24 hr in serum-free conditions, cells were incubated with R5020 (10 nM) or vehicle (ethanol) during different times at 37ºC. When indicated ICI182780 (10 μM, commercial name Fulvestrant) or PD98059 (50μM) were

also added after 6 hr of hormone induction (R5020, Estradiol). Cells were collected after 30 min, and after 1, 2, 6, 12, 24 and 48 hr.

3.5 Breast biopsy samples

Frozen tumor samples come from surgical biopsies provided by the *Hospital del Mar* (Barcelona). Samples were collected from 2002 on, and all relevant clinical-histopatological data were determined: TNM criteria; grading; ER, PR, HER2 and TP53 status; age of the patient at the time of diagnosis; whether there existed treatment before surgery and what type of treatment. In 95% of cases, treatment was given after surgery, while only 3% of cases received neo-adjuvant chemotherapy before surgery, leaving a very small number of cases that can be used to study the correlation with tumor response to treatment. All follow-up clinical history is currently being collected since the possibility of recurrence is still unknown in many cases. The clinical and molecular features of these tumors are listed in **Appendix A4**.

Total RNA was prepared by using Ultraspec® RNA isolation system (Biotecx laboratories Inc.) following manufacturer's instructions. We took 111 of these samples for further analysis, since some of them were handicapped in either quantity or quality of the total RNA material. Universal human reference RNA (UHRR) was obtained from Stratagene® (cat.no.740000).

The pathological diagnose is listed in **table 1**.

Table 1: Pathological diagnose of the patients and percentages in our population**.**

Pathological diagnose	No. of patients	%
infiltrating ductal carcinoma	65	61.9
intraductal + infiltrating ductal	15	14.3
lobulillar infiltrating	9	8.6
lobulillar in situ	2	1.9
intraductal carcinoma	2	1.9
metaplasic + ductal carcinoma	2	1.9
papilar invasive	2	1.9
lobulillar *"in situ"* + infiltrating lobulillar	1	1.0
lobulillar *"in situ"*+ ductal infiltrating	1	1.0
mucinous carcinoma	1	1.0
atypical medular	1	1.0
medular infiltrating carcinoma	1	1.0
Ca Ductal with apocrine differentiation+Ca intraductal	1	1.0
tubular infiltrating	1	1.0
carcinoma celullar neuro endocrine	1	1.0

Distribution of the age of diagnosis of our population of breast tumors is shown at **Figure 3**:

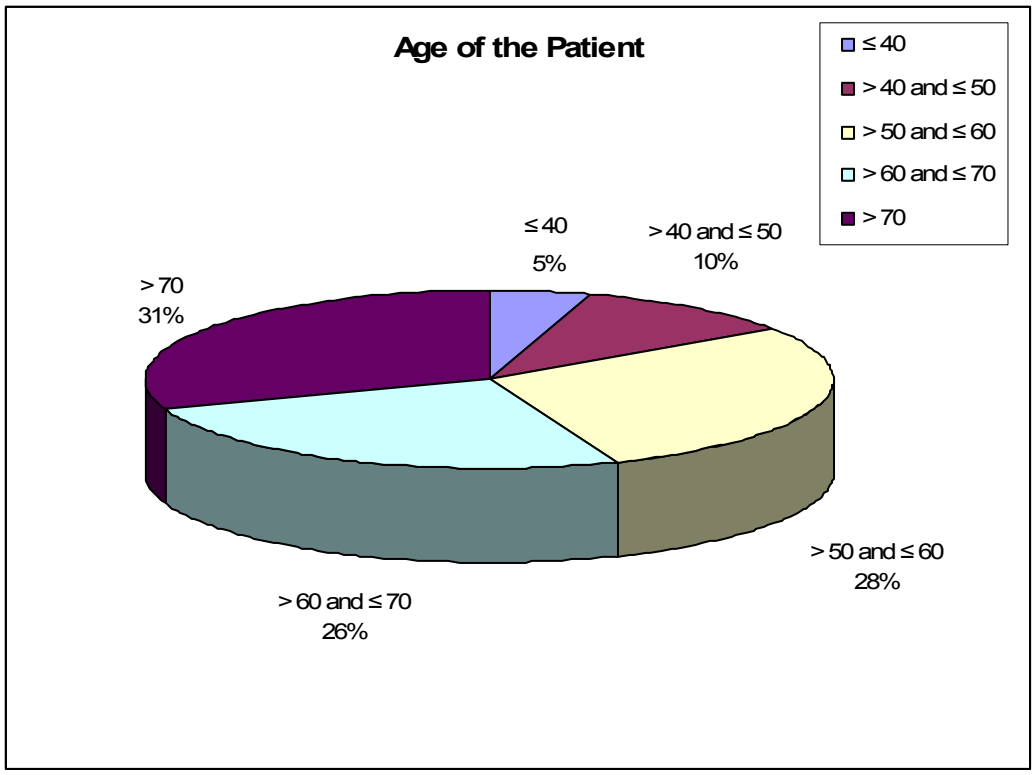


Figure 3: Distribution of our population of breast tumors based on the age of diagnosis.

3.6 RNA quality assessment

To assess the quality of the RNA and to evaluate the level of degradation, 1 μl of intact Total RNA samples were analyzed using the Agilent Bioanalyzer 2100® and the RNA 6000 LabChip Kit (Agilent®) with the Eukaryote Total RNA Nano Assay®. 28S/18S ratios, degradation factor and RIN numbers were used to quantify the state of degradation. We decided to choose specimens with RIN quality factor higher than 6.5 for microarray analysis. This is so because, after visual inspection of the graphical representation on the RNA, we thought that below the threshold of 6.5, the 28S ribosomal peak was too small in comparison to the 18S, and degradation also affected the mRNA population.

3.7 Linear T7 oligo-dT mediated mRNA amplification

Modifications of the method described by Van Gelder *et al. 1990* and Eberwine *et al. 1992* are described in detailed next.

3.7.1 First Strand cDNA synthesis

Total RNA templates were quantified with a spectrophotometer, and 3 μg were used for mRNA amplification. To each sample, 1 μl (100 pmol/ μl) of T7-oligo-dT composite primer, 5'-GGC CAG TGA ATT GTA ATA CGA CTC ACT ATA GGG AGG CGG$(T)_{24}$–3'; Sigma-Genosys) was added in 12 μl of total volume in a 0.2 ml PCR tube and RNA was denatured for 10 minutes at 70ºC in a MJ Research Thermocycler, and chilled on ice. Then 4 μl of 5X first strand buffer (Invitrogen), 2 μl of 0.1 M DTT, 1 μl 10 mM dNTPs (Roche) and 1 μl of Superscript II Reverse Transcriptase (Invitrogen) were added to the primer and RNA solution and reverse transcribed in a 20 μl reaction at 42ºC during 2 hr in a thermal block (Eppendorf ®).

3.7.2 Second Strand cDNA synthesis

After the first strand synthesis, reactions were chilled on ice, and 16.7 μl DEPC-H_2O, 10 μl 5X second strand synthesis buffer (Invitrogen), 1 μl 10 mM dNTPs, 0.35 μl *E.coli* DNA ligase (10 U/ μl, Invitrogen), 1.3 μl DNA polymerase (10 U/ μl, Invitrogen), and 0.7 μl RNaseH (2 U/μl, Invitrogen) were added to the first strand reaction, well mixed and incubated in a heating block placed in a cold room at 16ºC during 2 hr.

Then, reactions were chilled and quickly spun to bring down condensation of the sample on the lid. This was followed by addition of 100 μl ddH_2O, and 10 μl 0.5 M EDTA to stop the reaction. The cDNA was purified by addition of 160 μl of a Tris saturated Phenol:Chloroform:Isoamylalcohol solution pH 8.0, mixed by pipetting and spun at 12000 rpm for 5 minutes at room temperature. The clear aqueous phase was transferred to a clean RNase-free 1.5 ml tube. Then, 1 μl of glycogen carrier was added, mixed, followed by 80 μl 7.5 M NH_4OAc, mixing again, and adding 600 μl of ice-cold absolute Ethanol. The tube was shaken and spun down for 30 min. Supernatant was removed, and the double-stranded cDNA pellet was washed with 500 μl of 75% ice-cold ethanol. Pellets were air-dried during 5 min and dissolved in 6 μl DEPC-treated H_2O.

3.7.3 In vitro transcription with T7 RNA polymerase

RNA was *in vitro* transcribed using as a template all the double-stranded cDNA product using Megascript T7 RNA polymerase (Ambion) in a 15 μl reaction at 37ºC for 4 hr. Template DNA was digested with 0.5 μl of DNAseI (Ambion) and aRNA purified using RNeasy mini spin columns (Qiagen), and eluted from the column with 50 μl RNAse-free H_2O. *In vitro* amplified aRNA samples were quantified with a Nanodrop® spectrophotometer.

3.7.4 aRNA direct labeling method

This method was previously described and validated by Richer *et al*. (2002). Our modifications are the following: 3 μg of the *in vitro* amplified aRNA samples

were reverse transcribed with 4 μl 5X first strand buffer, 2 μl 0.1 M DTT, 0.4 μl low dT-dNTP mix (25 mM dA, dC, dG, 10 mM dT), 1.5 μl Superscript II (Invitrogen) and 2 μl 25 mM Cy3-dUTP or Cy5-dUTP (Amersham). The reaction was incubated at 42ºC for 2 hr, chilled on ice and quick spun to bring down condensation. To stop the reaction 1 μl 1 M NaOH/20 mM EDTA, was added followed by 80 μl MilliQ-H_20 and 10 μl 3M NaOAc pH 5. Labeled samples were purified with Quiaquick PCR purification columns (Qiagen) and eluted twice with 50 μl EB (10 mM Tris pH 8.5). Labeling efficiency was calculated by quantification with Nanodrop, obtaining between 50-80 pmol/ μl of Cy3 or Cy5 labeled probes. Labeled samples for the same array were combined, 1 μl 1 μg/ μl Human Cot DNA (Invitrogen) was added, and the labeled mix was desiccated by Speed-Vac centrifugation.

3.8 Design of the microarray experiment

The objective of the design is to facilitate the interpretation of the data analysis results. For this aim, this design must be simple but conclusive given the purpose of the experiment, which may be to find differentially expressed genes, to search for phenotypic class or significant dynamic time dependent changes. As many replicates as possible are needed to control for all random variation in order to have accurate enough measurements. For example, the statistical variance decreases as the number n of samples increases,

$$\mathrm{var}(\bar{A}) = \frac{\sigma_A^2}{n} \ ,$$

where $\bar{A}$ is the mean of the measurement, and σ is the standard deviation of $\bar{A}$. In the case of the experiments with breast cancer cell line samples it was feasible to have biological replicates, that is, another parallel specimen for labeling and hybridization. In the case of the experiments of breast tumor biopsy samples, where sample amount is a limiting factor, biological replicates were not possible. However, we did incorporate two biological replicates of a couple of samples and technical replicates to check correlation and hybridization specificity.

The design choice in our microarray experiments was always the same one whether the purpose was to select differentially expressed genes, to search for specific gene-expression patterns in a time-course setting, or to identify tumoral phenotypic subclasses. In every case, we used an indirect design, that is, the commercially available Universal Human Reference RNA was always used as reference sample. We found this reference very useful since it can serve as a consistent control for data set comparisons, as well as can be used in multiple experiments that need to be carried out over long periods of time. It is also the most stable, unlimited sample, where every gene is represented but no gene is

biased for overrepresentation since it is a pool of ten different tumor cell lines from different human tissues.

In this manner we would be able to correlate directly the response to progestins of a breast tumor cell line and the immortalized picture of the gene expression breast tumor biopsy. The samples under study would be always labeled with Cy5 fluorochrom and reference RNA always labeled with Cy3. Cy3- or Cy5-conjugated nucleotides are bulky, which makes their incorporation using standard enzymes very inefficient. In addition, rates of incorporation can differ between dyes, potentially resulting in dye biases (Yang *et al*. 2002). If some gene were preferentially labeled by any of these fluorochrom, it would never show up as significant at the time of contrasting two hybridizations since comparison would be between same dye channels. For that reason we did not routinely perform dye-swap experiments.

In order to infer the relative gene expression difference between two samples in the indirect design, we need to subtract two hybridizations (**Figure 4**) with the handicap that the variance of the measurement is four times the associated to the direct design.

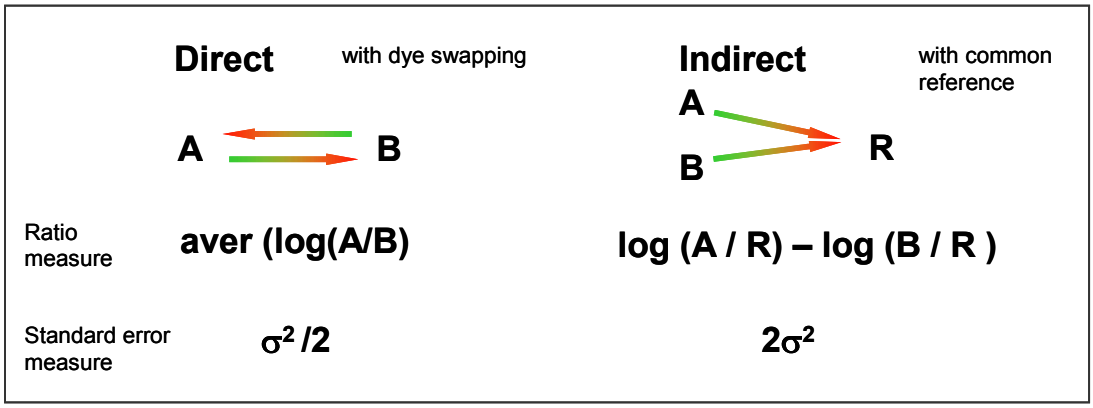


Figure 4: Standard error associated to the experimental design.

3.9 Microarray hybridization

3.9.1 Slide processing

Slides were pre-hybridized in prewarmed 5X SCC, 0.1% SDS and 0.1% BSA at 42ºC for 45 min, rinsed under milliQ water and spun dry using a centrifuge for 5 min at 1500 rpm.

3.9.2 Hybridization

Labeled samples were redissolved in 42ºC prewarmed 12 μl of Hybridization buffer A (50% formamide, 6X SSC, 0.5% SDS and 5X Denhardt's) applied to a glass coverslip, covered with the spotted glass slide. Arrays were incubated in a Corning hybridization chamber for 18 hr at 42ºC in a humid environment (In Slide Out oven, Boeckel).

3.9.3 Array post-processing

Arrays were washed at room temperature using an orbital shaker for 10 min with a high stringency wash buffer (0.1X SSC, 0.1% SDS), twice for 10 min with a low stringency wash buffer (0.1X SSC), rinsed for 5 min in milliQ water and spun dry for 5 min at 1500 rpm using a centrifuge.

3.10 Image acquisition

Fluorescent images were obtained using the G2565BA Microarray Scanner System (Agilent) with 100% laser power and 100% PMT settings and 16-bit TIFF images, one for each channel, were quantified using GenePix Pro 6.0 microarray analysis software (Molecular Devices, www.moleculardevices.com). This software discriminates between the relevant spots and the spot surrounding background by a segmentation method allowing for non-circular spots and recording every pixel intensity. First of all, mean foreground and background intensities were extracted from the red (Cy5) and green (Cy3) channels for every spot on the microarray. The background intensities are used to correct the foreground intensities for local variation on the array surface, resulting in corrected red and green intensities. A secondary purpose of the image analysis is to collect quality measurements for each spot that can be used to flag unreliable spots, which are those spots with abnormal background level due to spatial effects or “dye-dust” that interferes with spot measurement. The quality measures were the variation coefficient and the mean-median correlation of the foreground and surrounding background intensities of the spot. As a last quality control, images were also checked manually by flagging abnormal spots and the specificity of hybridization was checked by comparison of replicate spots side-by-side and at the opposite side of the array in a mirrored orientation (**Figure 5**).

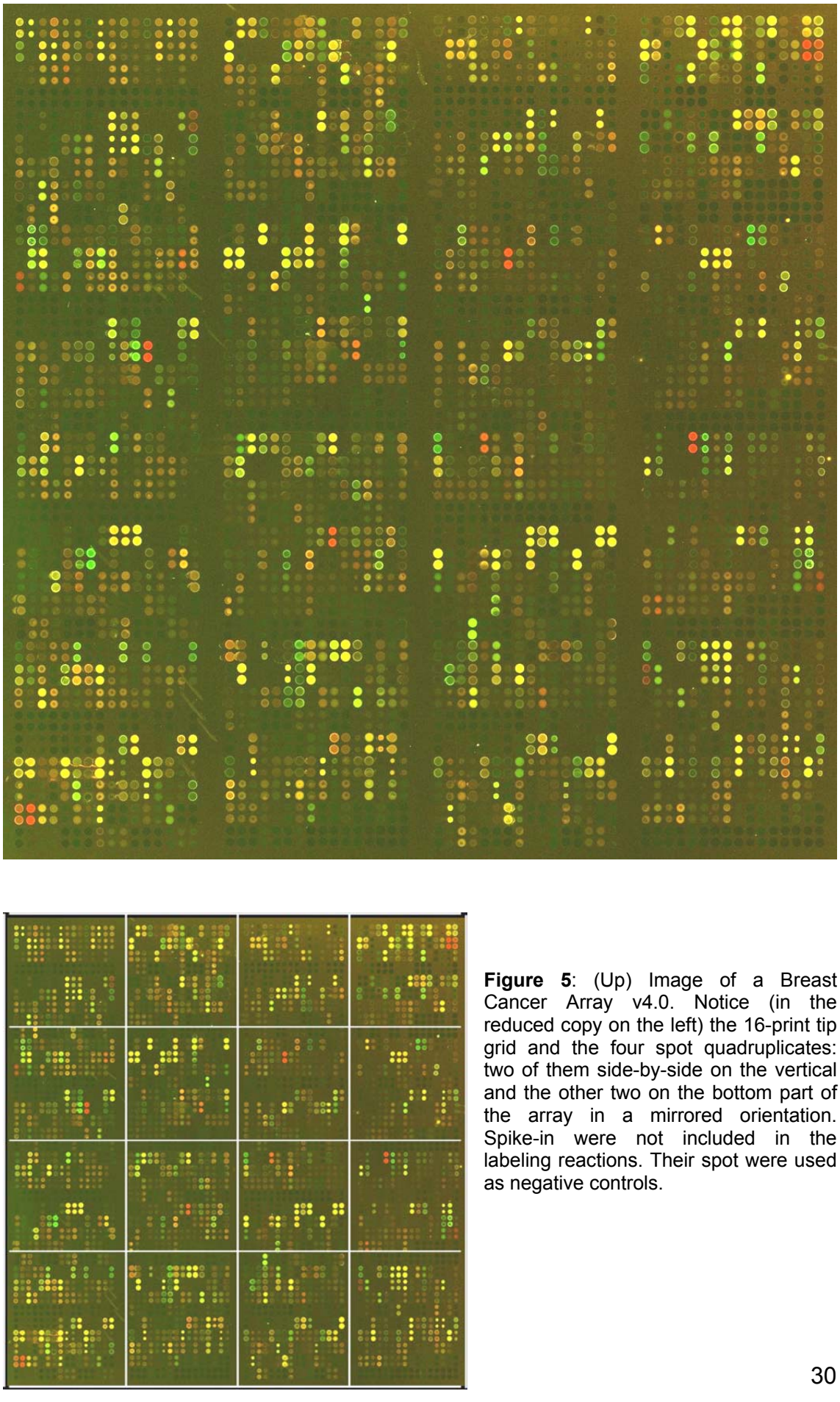


Figure 5: (Up) Image of a Breast Cancer Array v4.0. Notice (in the reduced copy on the left) the 16-print tip grid and the four spot quadruplicates: two of them side-by-side on the vertical and the other two on the bottom part of the array in a mirrored orientation. Spike-in were not included in the labeling reactions. Their spot were used as negative controls.

3.11 Array raw data normalization

Raw data were processed using MMARGE (see user's guide: http://nin.crg.es/manuals/MMarge170707.pdf) an in house developed web implementation of LIMMA, a microarray statistical analysis package of Bioconductor (http://www.bioconductor.org, Dudoit *et al.* 2003) that is run in the R programming environment (Gentleman *et al.* 2004, Wettenhall *et al.* 2004). Gene intensities were locally background subtracted using Normexp algorithm. This method adjusts the foreground adaptively for the background intensities and results in strictly positive adjusted intensities, so negative or zero corrected intensity are avoided. This results in a smooth transformation of the background subtracted intensities such that all the corrected intensities are positive. Spots with intensities smaller than two times the local background in both dye filter channels (Cy3 or Cy5), as well as control spots, were excluded from normalization, and were referred to as "non reliable".

Expression ratios *T* of each gene *i* were calculated as

$$T_i = \frac{R_i}{G_i} ,$$

where R and G are the red and green color intensities, respectively, commonly used to represent array data. We transformed ratios to logarithm base 2, to be able to treat upregulated and downregulated genes in the same manner and produce a symmetrical distribution of the ratio values.

Next, since Log_2Ratios could have a systematic dependence on intensity which most commonly appears as a deviation from zero for low-intensity spots, an intensity dependent normalization algorithm was applied (Yang IW *et al.* 2002; Yang IV *et al.* 2002). Locally weighted linear regression (Lowess) analysis was used as a normalization method to remove those intensity-dependent effects of Log_2Ratio values. The variables M and A are defined as

$$M_i = \log_2 \frac{R_i}{G_i} , \quad A_i = \frac{1}{2}\log_2 (R_i \times G_i)$$

Normalization effect can be observed on M-A plots before and after normalization, which show M, the Log_2Ratio, for each element on the array as a function of the A, the log2 geometric average intensity. Final M-A plots can be visualized at http://nin.crg.es/res/MM207610732042/index.html.

Print-tip Lowess normalization was applied to each print-tip group or subgrid to correct any systematic spatial variation on the array, between spotting needles or variability on slide surface, beside correcting for intensity-based trends, by adjusting the mean of the Log_2Ratio values in each subgrid to zero. This function relies on the assumption that most genes do not respond to experimental conditions, and so the average log ratio on the array should be

zero. The variance across all subgrids was adjusted using a smoothing factor for normalization. The smoothing factor applied was f=0.2 in order to homogenize the variance of the $\log_2$Ratio within each print-tip. The appropriate smoothing factor is chosen as the variance for a particular subgrid divided by the geometric mean of the variances for all subgrids (Huber *et al.* 2002).

Lowess normalization methods combine the least square regression with a nonlinear regression. Each M-value is normalized by subtracting from it the corresponding value of the tip-group Lowess curve, constructed applying local regression for each point in the M-A plots.

Normalized Log_2Ratios across all arrays were scaled so that every array has the same median intensity and same absolute standard deviation, in order to give the same weight to every gene on all arrays, and therefore changes of the expression ratio between arrays are not only due to the magnitude of M. Inter-array normalization adjusts the range of Log_2Ratio data.

The non-parametric empirical Bayes B-statistic was also computed for replicate hybridizations at time-course experiments, since we included hybridization replicates to determine the genes with significant regulation (Lönnsted and Speed, 2001). The B-statistic is an estimate of the posterior log-odds for each gene being differentially expressed. Values of B equal to zero correspond to a 50-50 chance that the gene is differentially expressed. The B-statistic is similar to a penalized t-statistic

$$t = \frac{\overline{M}}{\sqrt{(a+\sigma^2)/n}} \ ,$$

where the penalty a is estimated from the mean $\overline{M}$, σ is the standard deviation, and n is the number of sample replicates. With this data set, we considered genes that showed a 1.4-fold gene up or down-regulation relative to control sample with a B-rank value above 90% significant.

In the analysis of the breast tumor samples, which only had hybridization replicates of a few samples, we did not compute any other statistical criteria, and we considered as “reliable” all the genes with expression above the described background threshold, and as “non reliable” those which signal was below background level in both dye channels.

The value of fold change or relative copy number change was calculated as $2^{\text{Log}_2\text{Ratio}}$= ratio, if the ratio is positive, or $2^{-1/\text{Log}_2\text{Ratio}}$ if the ratio is negative.

3.12 Hierarchical clustering methods

Results of multiple hybridization experiments can be further analyzed to seek for similarities between gene expression profiles or sample gene expression

patterns by assembling all normalized and scaled Log_2Ratios measurement in a numerical matrix where rows correspond to genes and columns correspond to samples. In this matrix each gene or sample can be defined as a vector of Log_2Ratio value coordinates.

For all gene expression matrix analysis, we have used the open-source, freely available software package for microarray data management, visualization and analysis TM4 (Saeed *et al.* 2003) obtained from TIGR (http://www.tigr.org/software/microarray.shtm), TMEV: TIGR multiple experiment viewer) that uses hierarchical clustering analysis from Cluster and Treeview (Eisen *et al.* 1998). These packages can be also freely downloaded from Stanford software programs database (http://genome-www5.stanford.edu/).

3.12.1 Measures of similarity (or distance)

A measure is needed in order to compare the similarity or the distance between two or more genes or samples. We can regard any of these objects (rows or columns on a matrix) as points in an n-dimensional space or as n-vectors, where n is the number of genes or the number of samples.

There are two types of distance metrics extensively used in the comparison of expression profiles: the Euclidean distance and the Pearson correlation coefficient. We would like to search for genes with an identical expression profile which may represent a co-ordinate response to a stimulus, or genes with opposite profiles which may represent activation versus repression. The Euclidean distance is obtained as the absolute distance between two points in space, in this case defined by the two expression profiles or also called expression vectors. The Euclidean distance usually finds two genes or samples similar when these have the same magnitude of expression.

Although this property may be significant in some cases, it is usually biologically more relevant to search for genes expressed at different levels but with the same overall profile. The Pearson correlation coefficient is useful to identify profiles with similar shapes. It can be also be used to detect negatively correlated genes.

3.12.2 Cluster analysis

Clustering is the most widely used tool for microarray data analysis. The goal of clustering is to group together objects (genes or samples) with similar properties. It produces groups of gene expression profiles based on a distance function. Clustering can be used to find groups of co-expressed genes (Eisen *et al.* 1998), which are often functionally related, or to obtain clusters of experimental conditions (Perou *et al.* 1999). Depending on the way the data is clustered, we can distinguish between hierarchical and non-hierarchical clustering. Hierarchical clustering allows detecting higher order relationships between clusters or profiles. While the majority of non-hierarchical classification

techniques work by allocating gene expression profiles to a predefined number of clusters, without any assumption on the inter-cluster relationships. We could chose different distance functions (based on Euclidean or correlation coefficient) which can produce alternative clustering of data. Aggregative hierarchical clustering (Sneath and Sokal, 1973) is still the preferred choice for analysis of patterns of gene expression. We have always used the average linkage algorithm, which works as follows: firstly, the closest pair of genes or samples is selected and joined by a node, secondly these are averaged, and finally a new correlation matrix is created which replaces the previous pair of genes or samples by their average. The process continues until only one single element remains.

3.13 Class comparison methods

We have used SAM (Significance Analysis of Microarrays, Tusher *et al.* 2001, http://www-stat.stanford.edu/%7Etibs/SAM/index.html) in order to identify differentially expressed genes associated to a variable such as treatment or time. SAM can be used to pick out statistically significant genes based on a different expression ratio between sets of experiments by assigning gene-specific t-tests. A score is assigned to each gene on the basis of its change in gene expression relative to the standard deviation of repeated measurements for that gene with the assumption of gene-specific fluctuations. This procedure allows to dynamically change the threshold value for significance through a parameter delta Δ after looking at the distribution of the *d*-statistic. This makes the data-mining process more sensitive.

The test gives an estimate of the False Discovery Rate (FDR), which is the percentage of genes likely to have been misidentified by chance as significantly differentially expressed. To estimate the FDR, nonsense genes are identified by analyzing permutations of the measurements.

We have mainly employed a two-class unpaired design to pick out genes whose mean expression level is significantly different between two groups of samples (analogous to between subjects t-test), where under the null hypothesis of no differential expression the question we asked was whether the mean expression level of a gene in group *A* significantly different from the mean expression level in group *B*.

$$\frac{\overline{A}_{ij} - \overline{B}_{ij}}{\mathrm{var}_{ij}},$$

where i and j are the gene and experiment indices, respectively, and var is an estimate of the variance of a "gene-specific" t-test. A d-value, analogous to the t-statistic, is computed for each gene, and is called the "observed" d-value (d-statistic)

$$d_i = \frac{r_i}{\sigma_i + s_0} \ ,$$

where r is the ratio for i= 1, 2, …, p is the gene index, σ is the standard deviation, and s_0 is a small positive constant that makes that the genes with small fold-change will not be selected as significant even when the standard deviation is very small. An "expected" d-value d' is computed by order statistics permutation tests. For each permutation of the data, the test calculates the number of positive and negative significant genes for a given parameter Δ. The cutoff for significance is determined by the user by tuning Δ, based on the false discovery rate.

The median number of significant genes from these permutations divided by the median number of genes called significant is the median FDR. We set Δ so that only a 1 to 5% of falsely discovered genes show up as false positives. We can select also a fold change threshold, to ensure that "called" genes change at least a pre-specified amount. We obtain also a q-value for each gene, which is the lowest FDR at which that gene is called significant. It is like a p-value but adapted to multiple-testing situations.

The selected settings for every situation were (1) the FDR threshold to 0.05, that is, a 5% of falsely discovered genes, and (2) the fold-change threshold of 1.3, that is, the relative change of the expression ratio between two unpaired classes must be at least of 1.3.

With these conditions multiple gene-lists were generated and saved for posterior comparisons with other hybridization experiments, Ingenuity© pathway software package analysis and GSEA functional analysis (see chapter below on methods for functional analysis).

3.14 Time course microarray analysis

The identification of genes whose expression varies when biological conditions change is a frequent goal in microarray experiments. Differential expression can be analyzed from a static or a dynamic point of view. In a static experiment, gene expression is obtained instantaneously as it happens in the analysis of a breast tumor biopsy sample. In a dynamic experiment the arrays are collected as a time series, which allows us to observe the dynamic behavior of gene expression.

Up until recently, there were no established general statistical methods for identifying differentially expressed genes in a time course study, where gene expression data is collected as a function of time. None of the available clustering methods (PCA, K-means, and Bayesian model-based) are directly applicable to identify genes that show significant changes in expression over time. These previous methods require that the statistical significance be calculated under the assumption that the clustering model estimated for each

variable is true, and fail to properly use the temporal structure present in the data, leading to loss of power or incorrect calculation of significance.

EDGE, an open-source software package, applies ideas of spline-based methods. It is able to identify statistically significant genes, whose expression varies between groups of treatment or within a single group, accounting for sources of dependence over time and the untreated group. The EDGE algorithm considers two types of sampling: longitudinal or independent. We applied longitudinal sampling since, in our case, there is a dependence of the data on the cell batch (that is, the day of the experiment). Using this technique we identified statistically significant genes, whose expression varies within a single treatment (R5020, E2) or between treatments (Storey *et al.* 2005, http://faculty. Washington edu/jstorey/edge/).

Since we had two biological replicates, we first referred treated samples T_i, at a given time t_i = 30 min, 1 hr, 2, 6, 12, 24, and 48 hr, to their associated T_0, or untreated sample, of their corresponding biological replicate as follows:

$$\log_2 \frac{T_i}{T_0} = \log_2 \frac{T_i}{UHRR} - \log_2 \frac{T_0}{UHRR} \ .$$

We also decided that the time series of collected hormone treated cell line samples had a longitudinal distribution due to sampling of each biological replicate. Pre-filtered data with a threshold q-value less than 0.01 which is less than 1% FDR level was averaged and grouped into just one M value.

Following the above describe procedures, we took the significant gene list, imported the values into the TMEV (TIGR Multiple experiment viewer) microarray statistical analysis program, and finally used K-Means (a supervised hierarchical clustering method, Hartigan and Wong, 1979) with Pearson correlation coefficient as the distance metric, in order to cluster genes which follow similar trends in gene expression along time. The clustering of genes for finding co-regulated and functionally related groups had been successfully earlier carried out by DeRisi *et al.* (1997), Brazma *et al.* (1998), and Van Helden *et al.* (1998). From these clusters of genes we also generated gene-lists that can be subsequently imported into other functional analysis software (chapter 3.19: methods for functional analysis).

3.15 Class prediction methods

Conventional diagnosis of cancer has been based traditionally on the examination of the morphological appearance of stained tissue specimens under light microscopy. This method is subjective and depends on highly trained pathologists. The microarray technique could make cancer classification more objective and accurate.

Two of the most important uses for microarray data are 1) to generate gene expression profiles which can discriminate between different known cell types or conditions (like for example, to differentiate between tumor and normal tissues or different types of tumors) and 2) to identify previously unknown types or conditions (e.g. new subclasses of existing class of tumors). These two tasks have been referred to respectively as class prediction and class discovery in the work of Golub *et al.* 1999. The class prediction and discovery techniques are also known as supervised and unsupervised learning of gene expression profiles, or discrimination and class clustering, respectively. Clustering methods are appropriate if classes do not exist in advance, but if the classes are pre-existing, then discriminant analysis methods are more suitable and more efficient than clustering methods.

Initially, the FADA (Full analysis of DNA microarrays, Lozano *et al.* 2005) unsupervised clustering method was used to define new classes and assign samples to these classes, as a hierarchical clustering method. The objective was to identify and define the possible tumor classes, discriminate tumors from normal samples, find a distinct expression signature for each subcluster, find associated statistically significant GO terms, and predict the diagnosis category of a sample on the basis of its gene expression profile.

3.16 Full analysis in DNA microarrays (FADA)

FADA applies a Factor Analysis (FA, Reyment and Joreskog, 1996) a multivariate tool related to PCA, along with clustering algorithms applied to sample sets, t-test scores in gene set and data mining procedures. FA assumes that the observed gene expression levels are a result of a linear combination of an unknown number of independent underlying transcriptional programs, called factors. The contribution of each factor to the expression levels of the genes in each sample is given by the elements of the loaded data matrix. FA calculates the covariance of a data matrix. Covariance in the mRNA expression levels occurs in proteins involved in related pathways and functions, as well as co-localized proteins in the cell, and is indicative of common regulation of gene expression, and may uncover of a shared regulatory mechanism (Bar-Joseph *et al*. 2003). Specific variance of a given gene which is not associated to other genes is most probably related to artifacts, and would not represent biological significance.

Data matrix reduction is achieved by FA along with clustering algorithms to generate clusters in sample space or sample dendrograms (Hartigan *et al*. 1975). Multiple testing corrected Student t-test (q-value) is employed for the associated gene extraction of each obtained subcluster for the measurement of the differential gene expression of the gene compared with the rest of the samples. The q-value is similar to the p-value, except that it is a measure of significance in terms of the false discovery rate. Genes with q-value less than 0.05 were taken as differentially expressed for that particular cluster.

Significant genes associated to each cluster are used for the detection of pathways predominantly activated in that cluster in order to find statistically significant GO terms (Ashburner *et al.* 2000).

3.17 Between groups analysis (BGA)

BGA is a supervised classification method (Culhane *et al.* 2002). The basis of this method is to ordinate the formed groups rather than individual samples on a tri-dimensional space. BGA is a multiple discriminant analysis approach, which uses a dimension reduction technique such as principal component analysis (PCA) and correspondence analysis (COA, Fellenberg *et al*. 2001) to examine the correspondence to the most discriminant genes on each axis. Instead of dimension reduction of the individual samples performed in these classical ordination techniques, BGA ordinates the groups. It finds the eigenvectors or axes that discriminate the groups so as to maximize the between group variances. BGA, when used together with COA, ranks the genes, so that at the end of the axis the most discriminating genes are selected. In this way the genes associated with each group are determined. BGA is implemented as one of the microarray analysis tools of our laboratory group web-based server as SUCA (SUpervised ClAssification Bioconductor R-scripts) and also runs on the R programming language environment.

3.18 Prediction analysis of microarrays (PAM)

The supervised class prediction method used was PAM (Prediction Analysis of Microarray, Tibshirani *et al.* 2003). This method for class prediction was firstly applied to distinguish molecular subtypes in breast cancer and to predict overall survival (Sorlie *et al.* 2003).

PAM can be freely downloaded from http://www-stat.stanford.edu/~tibs/PAM/, as excel add-in which runs on the R environment.

Briefly, the method computes a standardized centroid for each class. This is the average gene expression for each gene in each class divided by the within-class standard deviation for that gene. Nearest centroid classification takes the gene expression profile of a new sample, and compares it to each of these class centroids. The class whose centroid it is closest to, in squared distance, is the predicted class for that new sample. Nearest shrunken centroid classification makes an additional modification to standard nearest centroid classification. It "shrinks" each of the class centroids toward the overall centroid for all classes by a constant called “threshold”. This shrinkage consists of moving the centroid towards the zero. After shrinking the centroids, the new sample is classified by the usual nearest centroid rule, but using the shrunken class centroids.

This shrinkage has two advantages: 1) it can make the classifier more accurate by reducing the effect of noisy genes; 2) it does automatic gene selection. In particular, if a gene is shrunk to zero for all classes, then it is eliminated from the prediction rule. Alternatively, it may be set to zero for all classes except one, and high or low expression for that gene characterizes that class. Typically, the user would choose the threshold value that gives the minimum cross-validated misclassification error rate.

PAM applied to breast tumor classification, shifts the mean expression level of each gene for each class towards the overall mean expression level for all classes by a fixed standardized difference (shrunken centroids). For a given shrunken centroid, only those genes for which the shrunken means still differ from the overall mean will contribute to the distance between centroids and any individual tumor sample's expression pattern. The class for which the shrunken centroids most closely reaches the observed expression pattern of a certain patient by using a Pearson correlation is then the predicted class for that patient. The standardized difference, and the number of relevant genes, is chosen by minimizing the prediction error using a 10-fold balanced, leave-10%-out cross-validation within the training set.

The same method is then used to predict classes for new samples (when they have a prior class assignment they represent a validation set): Pearson correlation coefficient is computed for each new sample to each of the five centroids and assigns each sample to the subtype with which it showed the highest correlation. Therefore, in our study, we are applying a Nearest Shrunken Centroid method to successfully find a minimal set of genes to distinguish classes with the minimal misclassification error.

3.19 Functional analysis (EASE-DAVID, EA, GSEA, Ingenuity)

3.19.1 EASE-DAVID

In order to gain biological understanding from microarray data, it is necessary to analyze functional annotations of all genes in the obtained gene lists. The Gene Ontology database (Ashburner *et al.* 2000) provides a useful catalogue to annotate and analyze the functions of a large number of genes.

With every gene lists from every produced cluster, we performed a functional analysis using EASE-DAVID (NIH, http://david.niaid.nih.gov/david/ ease.htm, Dennis *et al.* 2003). The EASE algorithm (Hosack *et al.* 2003) looks at the representation of functional classes in a significant set and compares it to the representation on the entire array taken as the reference or the background list using a Fisher's exact probability test to calculate p-values for a particular category. The resulting list of p-values is sorted. The GO terms that are more distinctive in the analyzed list of genes have lower p-values.

Since the number of GO terms for which we tested significance is large, the computed p-values were corrected in order to control the error rate associated to multiple testing (Shaffer *et al.* 1995, Dudoit *et al.* 2002). The Benjamini and Hochberg (1995) correction was selected which controls the FDR. To avoid type I errors when multiple comparison are being made, by selecting 1000 iterations for bootstrapping, and a FDR percentile threshold of 1%. Selecting a p-value below 0,01, we expect that a 1% of the selected GO terms will not be specific.

EASE can be installed and run locally on a personal computer; we have used the version 2.0 and run up locally since the number of genes (about 800) is small enough. Hyperlinks to all databases are listed in an output file.

3.19.2 Enrichment analysis (EA)

The software package GOstat (Beißbarth and Speed, 2004) is a tool that utilizes GO information to uncover which annotations are distinctive for the analyzed list of genes. GOstat automatically obtains annotations from a database and generates statistics of the annotations that are over- or under-represented in the analyzed list of genes. EA (Enrichment analysis, Falcon and Gentleman, 2007) is a Bioconductor package written in R, and implemented in our laboratory web server, that uses GOstats. EA implements GOstats, which applies a conditional hypergeometric test that use the relationships among GO terms. EA allows to test GO terms of one gene list for over or under-representation using the complete array gene list as reference or background list. For all the genes analyzed, GOstat determines the annotated GO terms and a χ^2 test is used in order to approximate the p-value.

EA differentiates with EASE in that EASE uses Fisher's exact test probability or the EASE score, the upper bound distribution of Jackknife Fischer exact test probabilities.

3.19.3 Gene set enrichment analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an *a priori* defined set of genes shows statistically significant, concordant differences in gene expression changes between two biological states (e.g. phenotypes). This method evaluates microarray data at the level of gene sets. GSEA-P is a software package which can be freely downloaded from: http://www. broad.mit.edu/gsea/msigdb/index.jsp.

GSEA compares and finds similar lists of genes in curated gene lists from a Molecular Signature Database (MsigDB), The gene sets are defined based on prior biological knowledge, as published information about biochemical pathways or coexpression in previous experiments.

The MSigDB (version 2.1) gene sets are divided into four major collections:

- c1: positional gene sets for each human chromosome and each cytogenetic band.
- c2: curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts.
- c3: motif gene sets based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat and dog genomes.
- c4: computational gene sets defined by expression neighborhoods centered on 380 cancer-associated genes.

An additional customized gene set was constructed, which we called c5, with all the obtained gene lists on microarray experiments using the Breast Cancer Array cDNA microarray platform. This includes up- or down- regulated gene lists obtained from hormonal treatments (R5020, estradiol) on a time series, gene lists obtained from K-Means cluster analysis, gene lists obtained after significance microarray contrast analysis (SAM) of the hormonal treatments with and without specific ihhibitors or hormone antagonists, and significantly over- or under-represented genes in the different classified breast tumor phenotypes.

c1, c2, c4 and c5 are mainly the gene sets used.

GSEA first calculates an enrichment score (ES) as a degree to which a set S is overrepresented at the extremes (top or bottom) of the entire ranked list L. It is calculated as the maximum deviation from zero encountered in the random walk; it corresponds to a weighted Kolmogorov-Smirnov-like statistic. On a second step, GSEA estimates the statistical significance (nominal P value) of the ES by using an empirical phenotype-based permutation test procedure that preserves the phenotype structure of the gene expression data, permutes the phenotype labels and recompute the ES of the gene set for the permuted data, which generates a null distribution for the ES. The empirical, nominal p-value of the observed ES is then calculated relative to this null distribution. The permutation of class labels preserves gene-gene correlations and, thus, provides a more biologically reasonable assessment of significance than would be obtained by permuting genes. On a third level, GSEA adjusts the estimated significance level to account for multiple hypothesis testing. First the ES for each gene is normalized to account for the size of the gene set, yielding a normalized enrichment score (NES), and determines the proportion of false positives by calculating the false discovery rate (FDR) corresponding to each NES. The FDR q-value is the estimated probability that a set with a given NES represents a false positive; it is computed by comparing the tails of the observed and null distributions for the NES. There are other corrections for multiple hypothesis testing such as the conservative family wise error rate (FWER), but nominal p-value < 0.05 were selected since the primary purpose was to generate hypothesis.

Later, the so called „Leading-edge subset“ similarity analysis“ is performed. Gene sets can be defined by using a variety of methods, but not all of the members of a gene set will typically participate in a biological process. Often it is useful to extract the core members of high scoring gene sets that contribute

to the ES. The leading-edge subset can be interpreted as the core of a gene set that accounts for the enrichment signal across diverse experimental data sets.

3.19.4 Ingenuity Pathways Analysis (IPA)

The Ingenuity Pathways Knowledge Base (Ingenuity Systems©) was used for functional analysis of genes. Briefly, the Ingenuity Pathways Knowledge Base consists of 106 individually modeled relationships into an ontology with more than 550 000 biological concepts. Relationships between genes, proteins, small molecules, complexes, cells, processes and diseases were manually extracted by scientists from more than 200 000 peer-reviewed articles.

3.20 Real time qPCR assays

Real time qPCR is a standard method for validation of microarray results, and has already been extensively reviewed (Bustin 2000, Bustin 2002, Bustin *et al.* 2005, and Stahlberg *et al*. 2005). It is currently the most sensitive method to determine the amount of a specific DNA in a complex biological sample. In real-time PCR, the amount of product is measured during an ongoing amplification using fluorescent reporters or dsDNA dyes (SYBR Green 1). Fluorescence signal is monitored each cycle during the annealing/extension phase of PCR to be quantified, the product has to accumulate enough to generate signal above background noise. The point where fluorescence rises above the background level is quantified as the second derivative maximum (Cp) or crossing threshold of the curve (Ct) and correlates with the amount of starting copies within a PCR reaction. Real time PCR is characterized by a wide quantification dynamic range (with seven to nine orders of magnitude), high sensitivity, and high reproducibility. Real time PCR is a more suitable diagnostic platform than microarrays especially for small gene sets and large numbers of samples are analyzed.

Below we describe in detail the standard procedure we followed, consisting in four steps: primer design, two step RT-PCR, determination of the reaction efficiency, and data analysis.

3.20.1 Primer-design

The primers were designed using Primer 3 (Rozen and Skaletzky 2000) (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3-www.cgi) with their distinctive RefSeq Accession number (**Appendix A5**: List of primers for Real Time qPCR Assays). The primers were designed such that their annealing temperature was 60ºC, giving a 90 - 200 base pair product. In order to minimize primer-dimer formation, the maximum self-complementary score was 4 and the maximum 3' self-complementary score was 2. The primers were designed, when possible, within the last 3 kb from the 3' end of the gene, flanking one or more introns to avoid gDNA amplification, or across exon boundaries.

The targets amplified by the primer pairs were characterized using M-fold (Santa Lucia 1998, http://bioinfo.math.rpi.edu/~mfold/dna/form1.cgi) in order to predict any secondary structures which might form at the site of primer or probe binding.

Primers were synthesized by Sigma-Genosys in a 0.2 μM scale, desalted while purified, and without any additional HPLC purification.

3.20.2 Two-step RT-PCR

First strand cDNA synthesis was carried out with a constant sample quantity of 0.5 μg total RNA using 70 pmoles of oligo-dT, and Reverse Transcriptase Superscript II (Invitrogen©, cat.no. 18064-014) in a final volume of 20 μl. Firstly, total RNA together with oligo-dT, and 1 μl of 10 mM dNTP in a total volume of 12 μl, prepared in 0.2 ml microtubes, was incubated for 5 min at 65ºC in a heat block in order to denature RNA. Reaction was chilled on ice for 5 min. A first strand reaction mix was added containing 4 μl of First strand buffer, 2 μl of 0.1 M DTT and 1 μl of RNAseOUT RNAse-inhibitor (40u/μl, Invitrogen©). Reverse transcription reaction was prewarmed at 42ºC and then 1 μl of Superscript II (200 u/μl) was added.

Reaction was further incubated at 42ºC for one hour on a thermal cycler (MJ Research), and then an additional incubation was performed at 70ºC for 15 min in order to inactivate the enzyme, followed by the chill on ice.

Before PCR amplification, every reverse transcription reaction was diluted at the proportion 1 to 20, to avoid Taq polymerase inhibition by DTT or excess of salts. No additional purification was performed.

We added, to the set of samples for reverse transcription, a duplicated RNA sample without reverse transcriptase to verify the absence of any gDNA contamination. We named it as “–RT negative control”.

PCR amplification was carried out with 1 μl of the previously diluted reverse transcription sample, as described above, with 5 μl of 2x SYBR Green Master Mix (ROCHE, cat.no. 4309155), and 3 pmol of specific gene primer pairs to a 10 μl total volume. Reactions were aliquoted in 384-well microtiter plates. The optical lid was applied and fixed to microtiter plates. Reactions were mixed by short plate vortexing, and spun for 2 min at 1600 rpm speed in a centrifuge with a rotor with microtiter plate adaptors (Eppendorf©). PCR reactions were run on a Lightcycler 480® system (ROCHE), using the following temperature cycling program:

1. Denaturing and Taq DNA polymerase heat activation step:

   95ºC for 10 min

2. Amplification step consisting of 45 cycles with the following cycle program:

   Denaturing step: 15 s at 95ºC
   Annealing step: 40 s at 60ºC
   Extension: 5 s at 72ºC

   A single fluorescence measurement at 533 nm wavelength was taken at 72ºC.

3. Dissociation step: ramp (4.8ºC/s?) from 72ºC to 95ºC with a continuous measurement of the fluorescence. This dissociation step was added in order to verify that a single amplification product and no unspecific secondary products or other products due to primer dimer formation were formed.

3.20.3 Determination of reaction efficiency

Standard dilution curves (1:4 serial dilutions) of a RNA sample were produced by dilution in nuclease-free water. A master mix was made up and aliquoted into the PCR plate prior to individual addition of the template into each reaction tube. A graph of the threshold cycle (C_t) versus the $\log_{10}[\text{copy-number}]$ of the sample from the dilution series was produced. The slope m of this graph was used to determine the reaction efficiency E as

$$E = 10^{-1/m} - 1 \ .$$

3.20.4 Data analysis

PCR reactions were always run three times, so that first of all mean Ct values and mean standard deviations were calculated for error propagation. Relative copy number (RCC) or fold-change ratio was calculated as:

$$RCC = \frac{E_{\text{target}}^{\Delta C_t}}{E_{\text{reference}}^{\Delta C_t}}, \qquad \text{where} \quad \Delta C_t \equiv C_t(\text{calibrator}) - C_t(\text{sample}),$$

E is the calculated efficiency of each gene target or reference, and ΔC_t the difference of the mean cycle threshold C_t of the calibrator sample and the experimental sample (Pfaffl 2001).

Standard error associated to the relative copy number was calculated as:

$$SD = \sqrt{SD_{TC}^2 + SD_{TS}^2 + SD_{RC}^2 + SD_{RS}^2} \ ,$$

where subscripts T and R mean target and reference genes, respectively, and subscripts C and S mean calibrator and experimental samples, respectively.

3.20.5 Determination of the normalization factor by geNORM

The geNorm VBA applet for Microsoft Excel (http://medgen.ugent.be/~jvdesomp/genorm/ determines the most stable housekeeping genes from a set of tested genes in a given cDNA sample panel. This application calculates a gene expression normalization factor for each tissue sample based on the geometric mean of a user-defined number of housekeeping genes. geNorm calculates the gene expression stability measure M for a control gene as the average pairwise variation V for that gene with all other tested control genes. Stepwise exclusion of the gene with the highest M value allows ranking of the tested genes according to their expression stability. The underlying principles and calculations are described in Vandesompele *et al*. 2002. The geometric mean of the n housekeepers (HGK) is calculated as a reliable normalization factor (NF_n):

$$NF_n = \sqrt[n]{HGK_1 \times HGK_2 \times ... \times .HGK_n}$$

4 Results

4.1 Reproducibility assays of the microarray platform

Initial microarray performance tests with the human breast cancer cell line T47D-MTVL in response to hormones were performed. Reference sample was the commercially available Universal Human Reference RNA UHRR (Stratagene®, a pool of 10 types of tissue cancer cell lines) for broad gene coverage. Methods of linear mRNA amplification and aRNA labeling have been previously validated (Van Gelder *et al*. 1990, Eberwine *et al*. 1992) from starting RNA material of 3 μg.

A random amplification method has been adapted and validated using a composite random primer T3-N_9 (Xiang *et al.* 2003), to detect differentially expressed transcripts in partially degraded RNA samples. This method was successfully applied to bladder tumor biopsies in a collaborative study with the Puigvert Foundation (Mengual *et al.* 2006). We found that transcripts were not as 3'-end biased as with the primer T7-dT, able to call more genes but, as disadvantages, there was a large amount of additional ribosomal RNA as part of the yield and amplified transcripts were smaller, as we verified using the Agilent Bioanalyzer mRNA assay, with a subsequent loss in sensitivity. Since breast tumor biopsies are often poorly preserved, we found that this method could be a good alternative to the standard T7-dT protocol.

Correlation coefficients from different mRNA amplification methods hybridizing the same sample were determined and gave good reproducibility with a linear correlation coefficient up to 0.96 when the same amplification method was compared (**Figure 6**).

Hybridization assay tests were also performed with different conditions, giving a low intra-chip variation coefficient between replicate spots (specificity) and a good hybridization signal when freshly spotted arrays were used. We learned how high ambient humidity and aging affect the quality of the hybridization signal and increase dramatically the background Cy3 intensity level of the substrate.

"Spike-in" controls had a good behavior, whenever samples similar in RNA integrity were co-hybridized. Calibration controls (with ratio $T = R/G = 1$) need to be adjusted within the concentration range of the experimental sample, since they are often found to be in the saturation region and might compete during the amplification reaction with the experimental sample owing to the fact that they can be more abundant than the sample mRNA population. Artificial "fold-ratio" controls had a good behavior too.

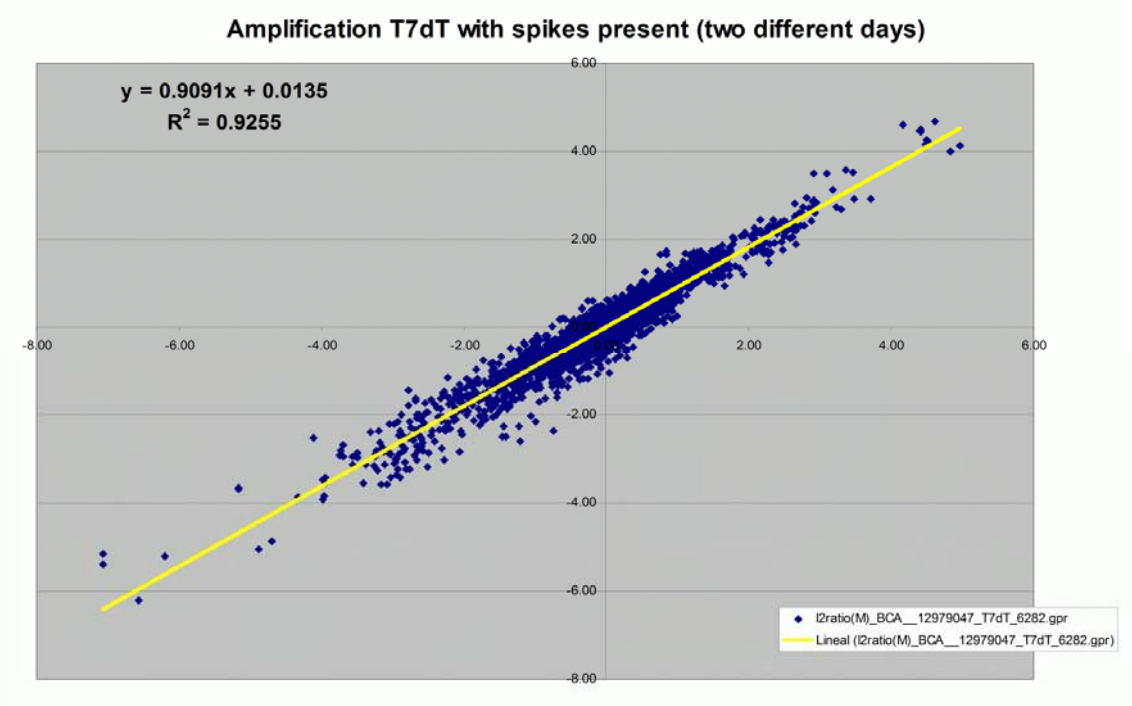


Figure 6: Correlation between two amplifications, labeling and hybridization of the same sample performed on different days giving a linear correlation coefficient of 0.96 taking only spots which were 2 times above background level threshold as reliable.

We have been also testing different methods for very low amount of total RNA increasing the number of rounds of mRNA amplification in order to apply them to laser microdissected cells (Baugh *et al.* 2001, Kenzelmann *et al.* 2004) up to one ng total RNA (in collaboration with F.X. Real, IMIM, Barcelona). After testing different protocols, and estimating linear correlation coefficients with standard conditions, we found high correlation with those methods that are able to extend and preserve the length of the transcripts.

4.2 Comparison with previously published cell line data

Firstly the microarray platform was validated with a short time course experiment, with only four time points, using a ER and PR expressing human breast cancer cell line T47D-MTVL. We hybridized every cell culture sample treated either with progestin R5020 or vehicle ethanol after 30 min, and 1, 2, and 6hr, using as common reference UHRR (in collaboration with M.J. Melià, CRG, Barcelona).

After normalization of the individual arrays, median scaling in their array intensities, and data processing as explained in material and methods (chapter 4.11), we selected the genes that were differentially expressed only due to the hormonal treatment by referring to the expression of the time-paired ethanol (hormone vehicle) treated. This was achieved by subtracting the logarithms of

the ratio from each pair of arrays corresponding to vehicle and hormone treated samples from each experimental replicates as follows

$$\log_2 \frac{R5020}{EtOH} = \log_2 \frac{R5020}{UHRR} - \log_2 \frac{EtOH}{UHRR} \ .$$

We found 147 genes that were expressed with a minimum of 1.4 fold-change in at least one of the time points, relative to the ethanol-only control treatment (**Appendix A6:** Short time course experiment), including representative GO annotations (Ashburner *et al.* 2000). Since there was no biological replicate there is not an alternative statistical parameter for thresholding.

It has been reported that progestins stimulate growth, have no effect, or even inhibit growth depending on experimental conditions or the status of the cellular hormone receptors (Sutherland *et al.* 1998, Groshong *et al.* 1997, Jeng *et al.* 1992, Lin *et al.* 1999). This disagreement reflects the insufficient understanding of progesterone biology and has hindered the effective application of progestins or antiprogestins in breast cancer treatment.

In mammary epithelial cells, ovarian hormones induce the recruitment of quiescent cells (cell cycle phase G_0) to enter the cell cycle, undergo progression G_1, and go through G_1/S transition. This is achieved in part through the direct transcriptional control of genes encoding key cell cycle regulators such as CCND1 (Cyclin D1). An immediate cell cycle arrest early in G_1, and growth inhibition, both mediated by PR have been observed after completing one round of replication (Musgrove *et al.* 1991, Skildum *et al.* 2005). These findings were corroborated in our experimental results as the most represented canonical KEGG pathways (http://david.niaid.nih.gov/david/) as shown in **Figure 7,** obtained by functional analysis using Ingenuity. Similar results were obtained using EASE. The figure shows a maximum number of expressed genes involved in ER signaling at 6 hr after hormone induction, a progressive increase in the ERK/MAPK signaling, G_1/S reach its maximum at 30 min and G_2/M reached its maximum at 6 hr.

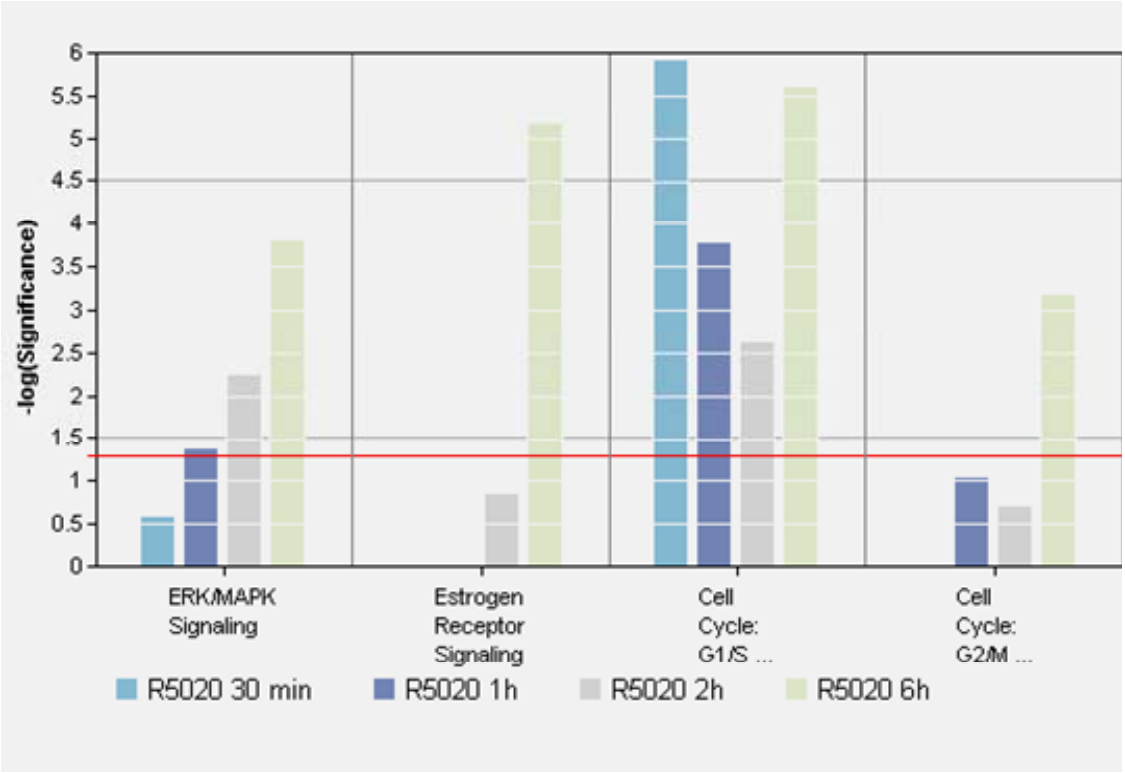


Figure 7: Representation of the most representative canonical KEGG pathways calculated from the regulated genes belonging to each pathway. Red bar denotes significance threshold (*p*-value smaller than 0.05). Significance is calculated from the number of genes present belonging to each pathway in each time point (figure from Ingenuity® pathway analysis software program).

The small subpopulation of cells that might have escaped cell cycle arrest which respond differently to hormones can dilute the measured response to hormone treatment. This occurs even though our population of cells was serum deprived 24 hr before treatment, and most cells were presumable quiescent.

We confirmed previous observations from other groups that progestins act on genes that regulate cell cycle progression, such as a high increase in EGF (epidermal growth factor) and TGFα (transforming growth factor α) after 6 hr rather than an early response (Musgrove *et al.* 1991). Also MYC (c-myc) was also previously found to be induced by progestins, suggesting that this transcription factor might participate in growth modulation (Musgrove *et al.* 1991). MYC is already activated after 30 min.

In order to visualize these events, using TMEV we applied a K-means unsupervised hierarchical grouping (Soukas *et al.* 2000) using Pearson correlation as the distance metric with complete linkage, grouping the 147 genes in 5 groups with similar patterns of expression throughout the set of experimental conditions (**Figure 8**). Since co-regulated genes are usually co-expressed, those co-regulated genes that follow a similar pattern of gene expression might be regulated by the same transcription factors, through the same binding sites.

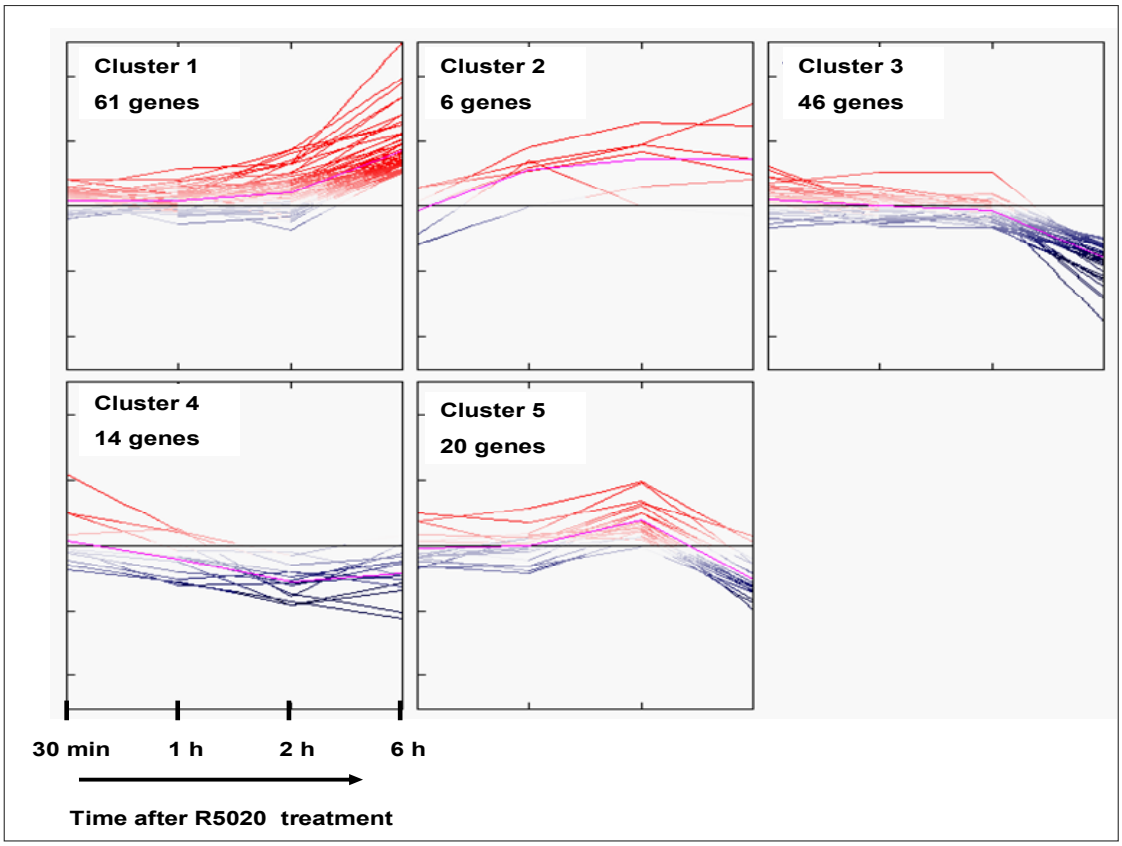


Figure 8: K-Means unsupervised grouping using Pearson correlation as the distance metric and complete linkage of the 147 regulated genes during the time course hormonal treatment with progestin R5020 of the T47D cell line in 5 groups which follow similar patterns of gene expression. The *x*-axis is time (hr) after R5020 treatment, and the *y*-axis is the Log_2Ratio.

Taking a closer look at cluster 5 (**Figure 9**), the one that shows an expression profile of a higher peak at 2 hr, and downregulation at 6 hr, we observe some of the nuclear receptors PR-Ḅ (Progesterone receptor B) and AR (Androgen Receptor) co-regulated with differentiation markers such as cytokeratins KRT5, KRT16 and KRT17. The MAP kinase pathway, by GO functional analysis, appear to be activated earlier on by 30 min -1 hr, and we see MAP3K1 (Mitogen activated protein 3 kinase 1) and MYC (c-myc) following this expression profile.

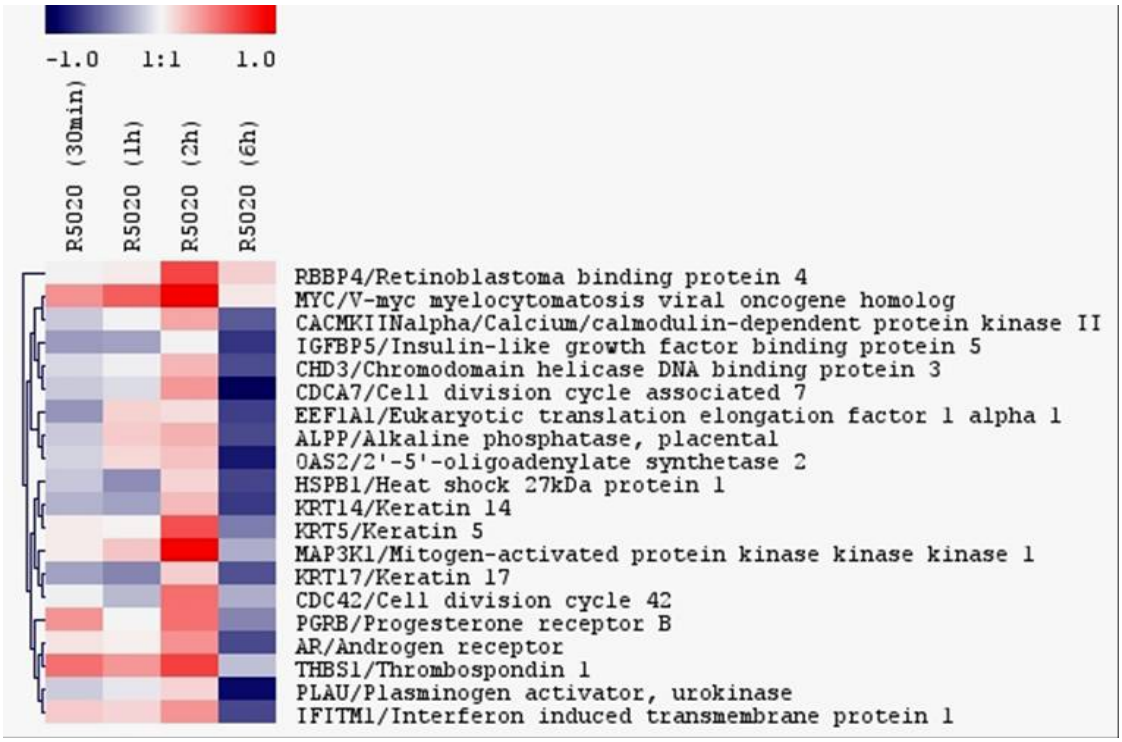


Figure 9: Cluster 5. K-Means unsupervised grouping using Pearson correlation as the distance metric with complete linkage, which shows early activation of MYC and MAP3K1. Scale-1.0 to +1.0 are $\log_2$Ratio values.

Clusters 1 and 2 (**Panel 10**) assemble all the genes whose expression increases with time of hormone treatment. The genes most highly expressed at 6 hr are, mainly

- TGFα (transforming growth factor alpha)
- DUSP1 (dual specificity phosphatase 1)
- RPS6KA5 (ribosomal protein S6 kinase, 90kDa, polypeptide 1, MSK1)
- ELL2 (elongation factor, RNA polymerase II, 2)
- HMGB3 (high-mobility group box 3)
- JUN (V-jun sarcoma virus 17 oncogene homolog, avian)
- EGF (epidermal growth factor)
- GRB2 (Growth factor receptor-bound protein 2)
- IL6ST (Interleukin 6 signal transducer)
- CCND1 (Cyclin D1).

Transcription factors that are expressed by 30 min., from our microarray results are

- TFDP1 (transcription factor dp-1)
- E2F3 (E2F transcription factor 3)
- SP1 (sp1 transcription factor)
- JUN (V-jun sarcoma virus 17 oncogene homolog)
- PR-B (progesterone receptor B)

- AR (Androgen receptor)
- ELL2 (elongation factor, RNA polymerase II, 2)
- SAP30 (sin3-associated polypeptide, 30kDa)
- GTF2H2 (general transcription factor IIH)
- PC4 (activated RNA polymerase II transcription cofactor 4).

Clusters 3 and 4 are shown in **Panel 11**.

Some of the downregulated genes after 6 hr of hormonal induction were

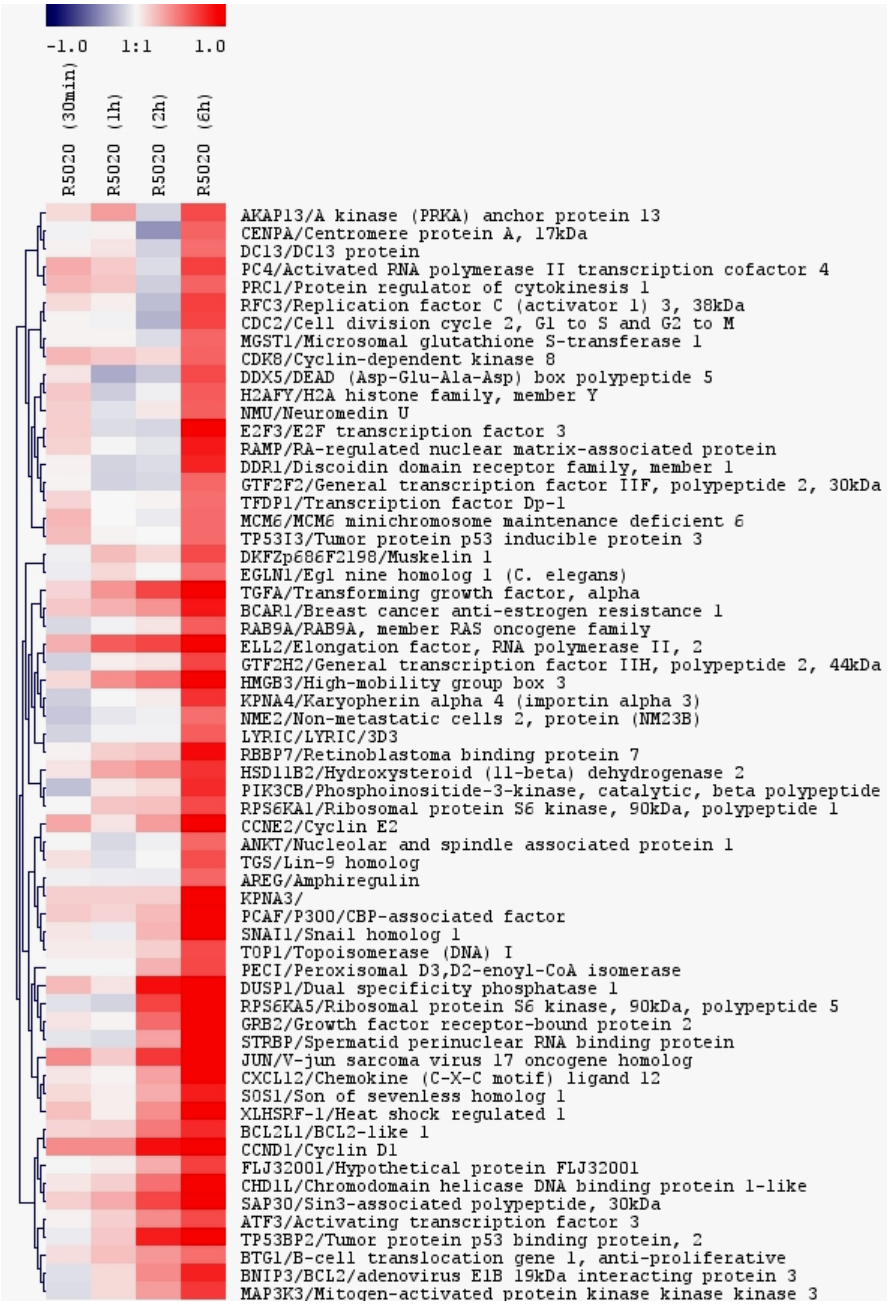
- ER-β Estrogen receptor β
- MYB (V-myb myeloblastosis viral oncogene homologue)
- SRC-2 (NCOR2 or Nuclear receptor coactivator 2)
- SRC-3 (NCOR3 or Nuclear receptor coactivator 3)

and previously reported co-regulators or direct downstream targets of ER such as

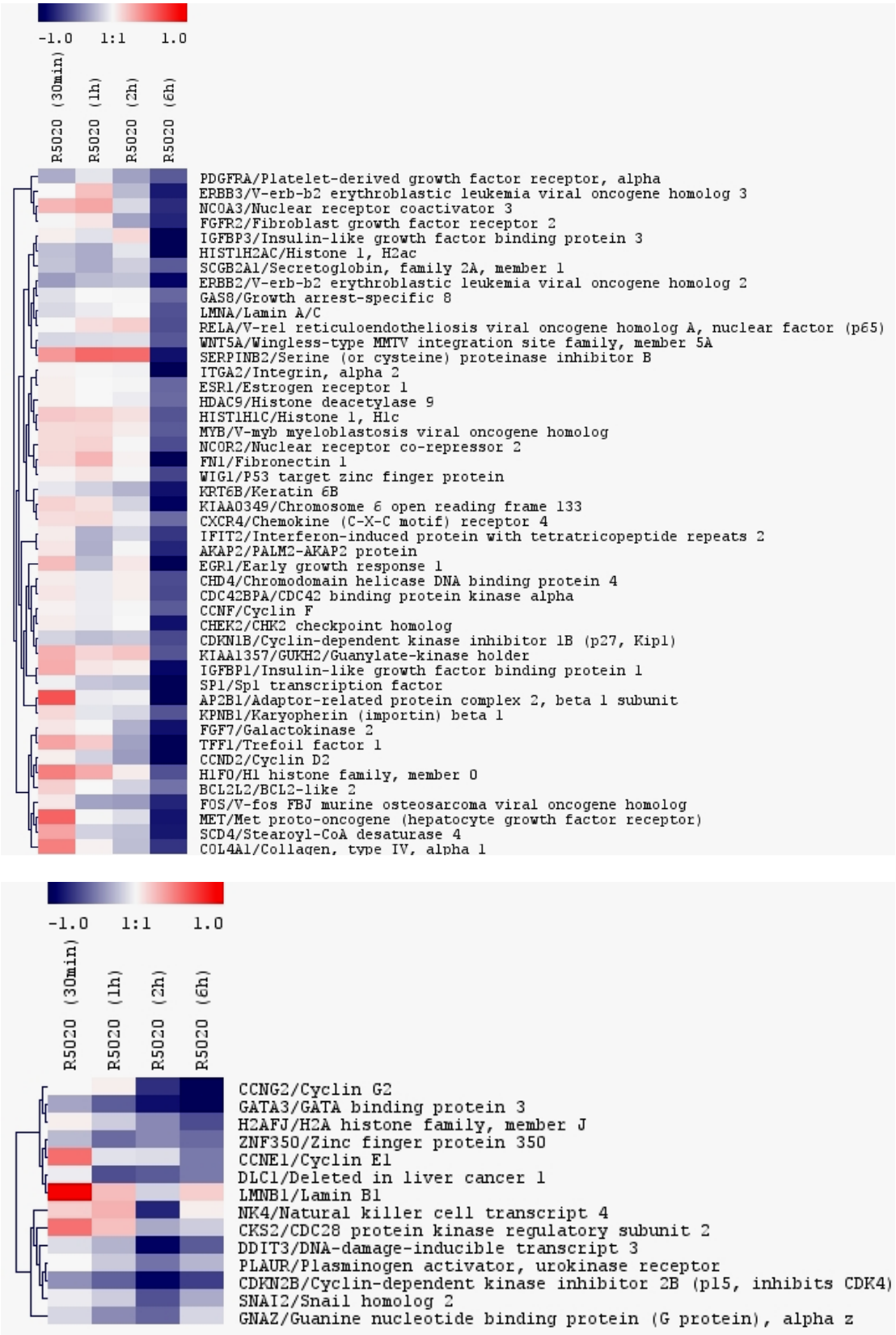
- FN1 (Fibronectin1)
- FGFR2 (Fibroblast growth factor receptor 2)
- IGFBP1 (Insulin-like growth factor binding protein 1)
- IGFBP3 (Insulin-like growth factor binding protein 3)
- GAS8 (Growth-arrest specific 8)
- TFF1 (PS2 or Trefoil factor 1)
- FOS (c-fos)
- SP1 (Sp1 transcription factor)
- PLAUR (Plasminogen activator, urokinase receptor)
- GATA3 (GATA binding protein 3)
- ZNF350 (Zinc finger protein 350)

The above list shows that the expression profile may be, in part, mediated by endogenous estrogen receptor.. Also, early response genes upregulated at 30 min, after the cell cycle progressed from G_1 to S, are rapidly downregulated such as

- CCNE1 (Cyclin E1)
- H1F0 (H1 histone family, member 0)
- LMNB (Lamin B1)
- CKS2 (CDC28 protein kinase regulatory subunit 2).



Panel 10: Cluster 1 (upper figure) and 2 (lower figure). K-Means grouping using Pearson correlation as the distance metric with complete linkage. Scale-1.0 to +1.0 are $\log_2$Ratio values.



Panel 11: Cluster 3 (upper figure) and 4 (lower figure). K-Means grouping using Pearson correlation as the distance metric and complete linkage. Scale-1.0 to +1.0 are $\log_2$Ratio values.

4.3 Comparison of the microarray cDNA platform

Independent hybridizations show high reproducibility with correlation coefficient up to 0,983 from the sample and different aRNA amplification and labeling. Some of the genes identified as being differentially expressed in our research, have been previously mentioned in earlier hormone response studies using gene expression microarrays. We confirmed the genomic and non-genomic representative pathways, in which progesterone may act, and the crosstalk with the ER ligand-activated pathway initiating the MAP kinase signaling cascade.

In order to validate the results obtained with the breast cancer cell line T47D-MTVL treated with R5020, we compared these results with those from Cunliffe *et al. 2003*. They used another cDNA platform of 13 824 sequence-verified cDNA clones (corresponding to 10 535 genes) from the National Human Genome Research Institute (NHGRI), applied different growth and differentiation regulators on three different breast cancer cell lines with different estrogen receptor status (T47D, MCF7 and MDA-MB-436), and compared the patterns of gene regulation to previously published tumor expression profiles.

We focused on the results obtained from the cell line T47D and treated with progestin R5020, in a time course of 2, 8 and 24 hr. We note the difference in the dose of the drug, which was 1000 fold higher (10^{-6} M) than the 10^{-9} M used in our test experiments, as well as the design of the microarray experiments, since they directly compared on the same array the effects of the drug against the mock treated, instead of using, as we did, a common external reference and performing indirect comparisons

After quality filtering of the data they obtained 1023 genes that responded with a fold-change larger than1.5 fold in at least 2 or more of the 42 conditions. Among these genes, we have an overlap of 108 sequences (107 genes). We note that from these 107 genes, 36 genes did not respond to progestin in the T47D cell line with fold change above the threshold of 1.5, leaving only 72 genes (**Appendix A7:** BCA overlap with Cunliffe *et al*.). From these 72 genes, there is an overlap of 35 between our regulated genes with fold change larger than 1.4. To better visualize the genes that are regulated in the same direction, we performed an unsupervised hierarchical clustering in search for similarities of the combined data using the Euclidean distance as distance metric (Figure 12).

We observe that of the 35 genes mentioned above, there are 21 genes with a high degree of overlap after 2, 6 or 8 hr, marked in figure 10 with a green bar.

Among the upregulated genes are

- ELL2 (Elongation factor, RNA polymerase II, 2)
- BCAR1 (breast cancer anti-estrogen resistance 1)
- CCND1 (Cyclin D1)
- TP53BP2 (Tumor protein p53 binding protein, 2)
- HMGB3 (High-mobility group box 3)

- RPS6KA5 (MSK1; Ribosomal protein S6 kinase, 90kDa, polypeptide 5)
- RPS6KA1 (RSK1; Ribosomal protein S6 kinase, 90kDa, polypeptide 1)
- RFC3 (Replication factor C (activator 1) 3, 38kDa)
- MYC (V-myc myelocytomatosis viral oncogene homolog, avian)
- MAP3K3 (Mitogen-activated protein 3 kinase 3), and
- THBS1 (Thrombospondin 1).

Among the downregulated genes

- IFIT2 (|Interferon-induced protein)
- CACMKIINα (Calcium/calmodulin-dependent protein kinase II)
- ERBB2 (V-erb-b2 erythroblastic leukemia viral oncogene homolog 2)
- ERBB3 (V-erb-b2 erythroblastic leukemia viral oncogene homolog 3)
- IGFBP5 (Insulin-like growth factor binding protein 5)
- FN1 (Fibronectin 1)
- HIST1H2AC (Histone 1, H2ac)
- MYB (V-myb myeloblastosis viral oncogene homolog)
- ESR1 (Estrogen receptor), and
- GATA3 (GATA binding protein 3).

We conclude that our results have good concordance with the results obtained in Cunliffe *et al*. (2003), although they obtained a greater response probably because they used 1000 fold higher progestin concentrations.

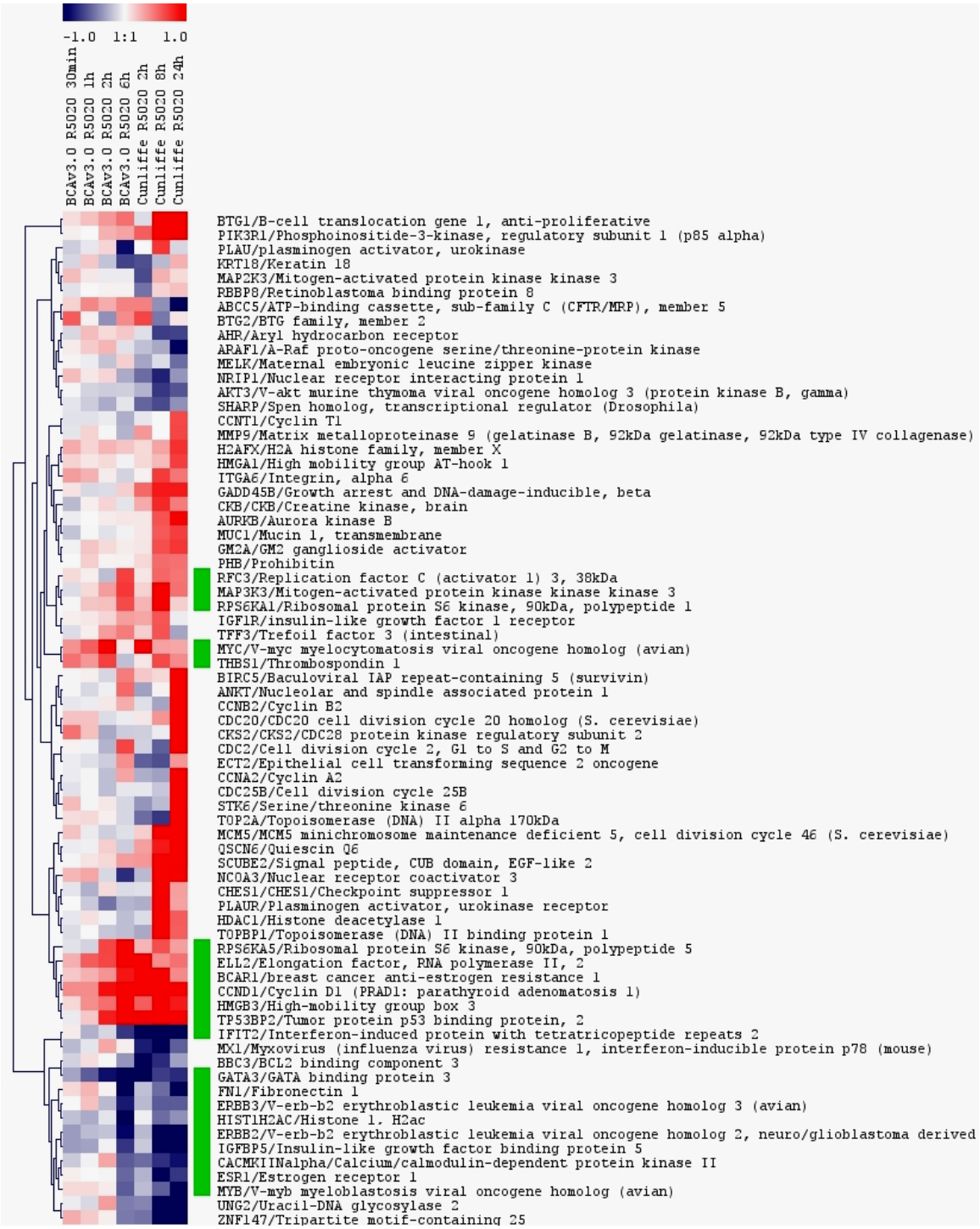


Figure 12: Comparison of the cDNA microarray data from our experiments and the published results by Cunliffe *et al.* 2003. To better visualize the results we applied a hierarchical clustering using the Euclidean distance as the similarity parameter to compare the genes that are regulated in the same orientation. Marked with a green bar are the 21 genes with a high degree of overlap after 2, 6 or 8 hr of treatment.

4.4 Application to an extended time course experiment

We performed an extended time course experiment with hormone treated breast cancer cell line samples, with either progestin R5020 or estradiol, using as a model the endogenously ER^+ PR^+ breast cancer cell line T47D-MTVL for an extended time in cell culture of up to two days of treatment. This work was done in collaboration with C. Ballaré (CRG, Barcelona).

The purpose of this experiment was to investigate the dynamic behavior of our population of cells in response to added hormones.

We collected the gene expression data using our cDNA microarray platform latest version v4.0 at times 0, 30 min, and 1, 2, 6, 12, 24 and 48 hr. We incorporated in this study two biological replicate experiments, that is, a replicate experiment was performed one week after the first experiment. By doing this we assured some statistical inference could we applied. The reference sample chosen was the universal human RNA reference (UHRR). We decided that our data had a longitudinal structure which means that samples of the same cell batch have a natural dependence between time points since they originate from the same sample.

Time course experiments allow us to study the dynamic behavior of the genes in response to hormone and to follow metabolic pathways. For this purpose, recent algorithms for analysis of microarray time-course experiments such as the software program called EDGE (Storey *et al.* 2005, http://faculty.washington.edu/jstorey/edge/ were used (see chapter 3.14 on methods for time course experiments).

The q-value is the estimate of FDR that we used to call a gene significant. These estimates are obtained by resampling of the data by bootstrapping (Efron *et al.* 1993). The resampling scheme takes into account dependence between time points since our study is structured as a longitudinal sampling. Our q-value cut-off is 0.01, a fixed FDR of 1%, which means that one percent of the genes assumed to be differentially expressed in a time series are false positives.

4.4.1 Temporal differential gene expression due to progestin R5020 hormone treatment

First of all, we carried out a "within-class" temporal differential expression analysis of the progestin treated cell line culture T47D. After referring every time point of the gene expression data to their corresponding T_0, of each biological replicate or cell batch, EDGE found 173 genes (p-value ≤ 0.01) or 226 genes (q-value ≤ 0.01). To decide which "cut-off" value would have more sense, we looked at the plot of the function of the q-value cut-off (x-axis) versus the number of significant tests (y-axis), and the plot of the function of the number of

significant genes (x-axis) versus the number of false positives (y-axis) as in Storey and Tibshirani (2003). These plots are represented in **Figure 13.** There is change in the slope of the function of the q-value cut-off versus the number of significant tests at q-value ≈ 0.01 (**Figure 13A**) while the number of expected false positives increase after the selection of 226 significant genes (**Figure 13B**).

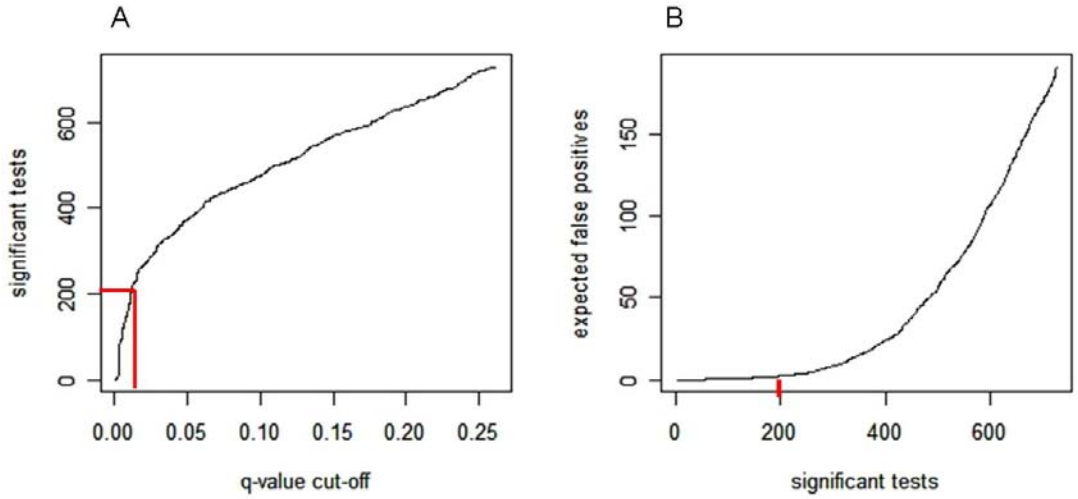


Figure 13: (A) The q-value cut-off (x-axis) versus the number of significant tests (y-axis), and (B) the number of significant genes (x-axis) versus the number of false positives (y-axis).

In addition, we looked at the significant gene list and the fold-change of both replicates to check if this decision makes sense and we concluded that the q-value cut-off of about 0.01 was the best option. Subsequently, from the obtained gene list, we averaged the M values of the two biological replicates at every time point and imported the significant gene list averaged data into TMEV for visualization and clustering analysis.

A clustering analysis compatible with the now reduced data set of time series was carried out using a supervised clustering by K-Means. The Euclidean distance was used as the similarity metric for grouping genes with a similar trend and magnitude in gene expression. This was done in order to generate a smaller gene list and to look for co-regulated genes and to perform functional analysis by looking at their overrepresented GO terms, and later on import them into pathway databases. In this case, where the data are already normalized and the data set is reduced, the Euclidean distance as the distance metric, gives results very similar to the ones given by the Pearson correlation. The Euclidean distance is the default method for this type of clustering. We decided, by inspection of the average cluster size and the gene expression vector components, that the best fitting number of groups was nine. A representation of the groups of the expression vector clusters is shown in **Figure 14.**

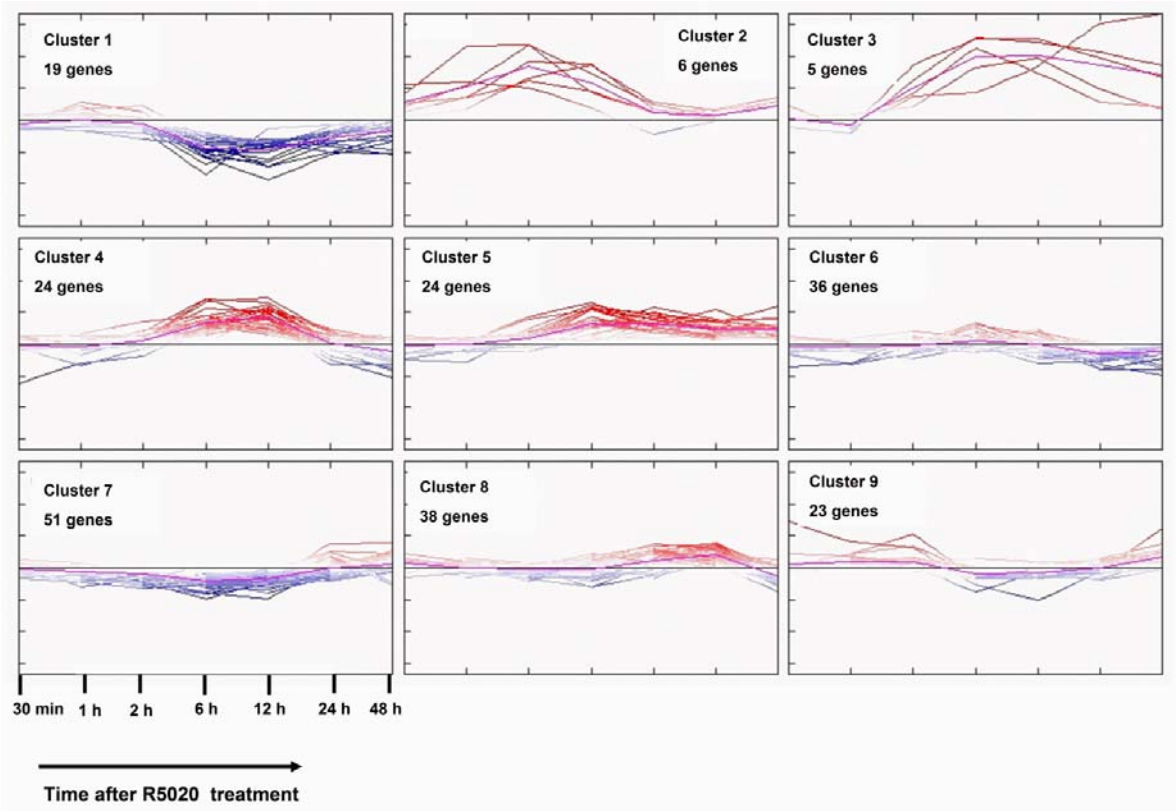
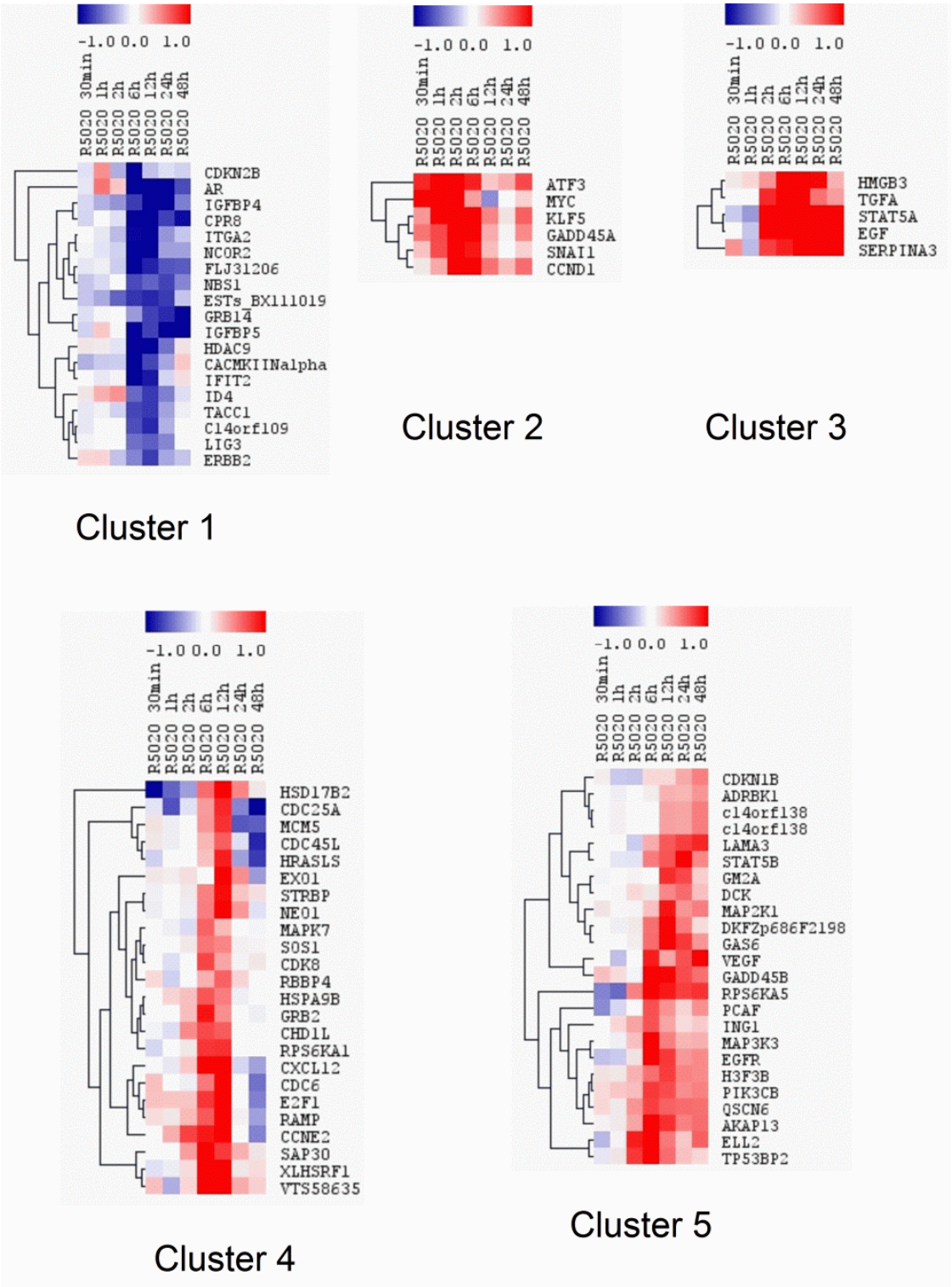
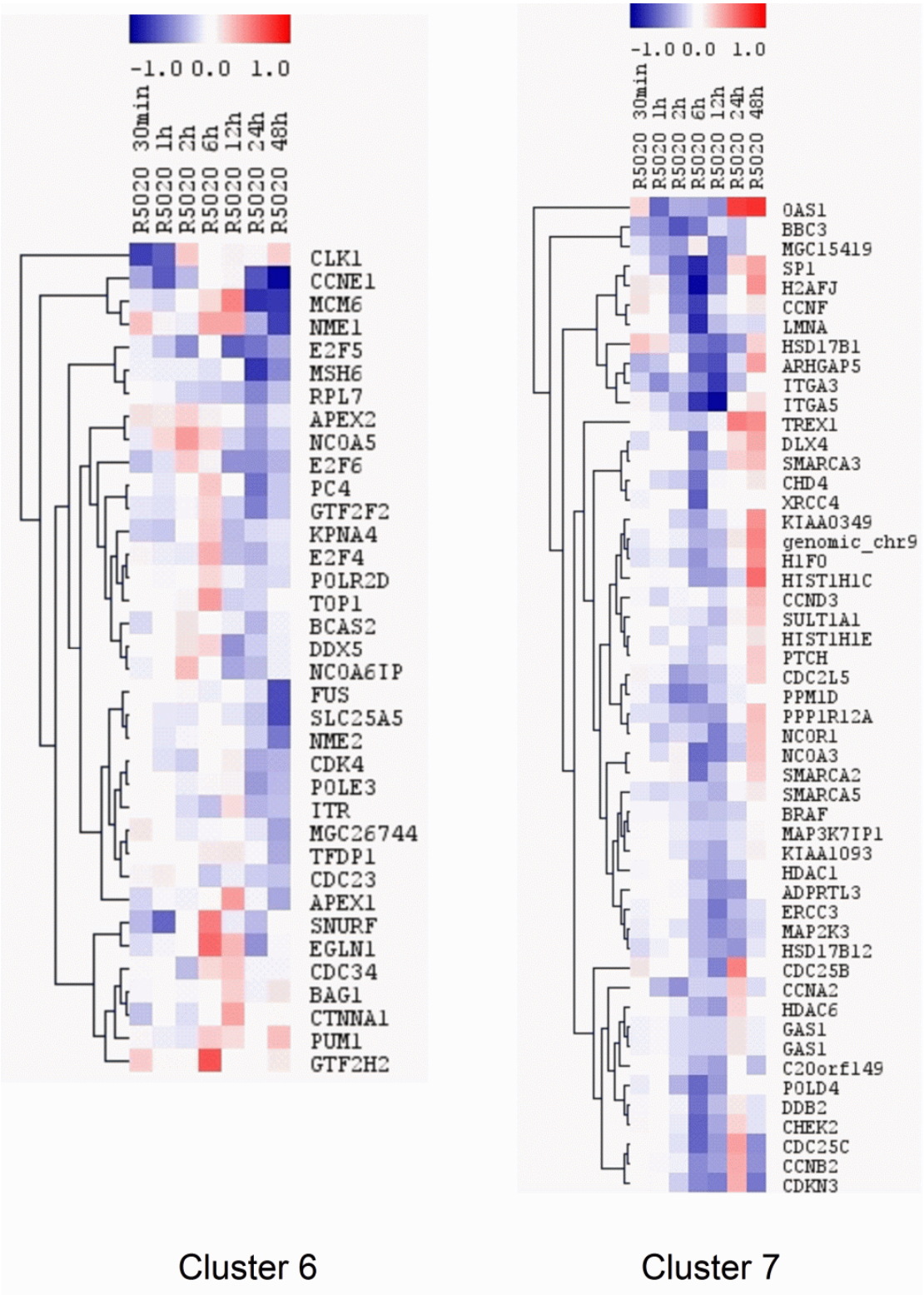


Figure 14: Time series of M (y-axis) as a function of time after R5020 treatment (x-axis) for the nine groups obtained by K-Means clustering.

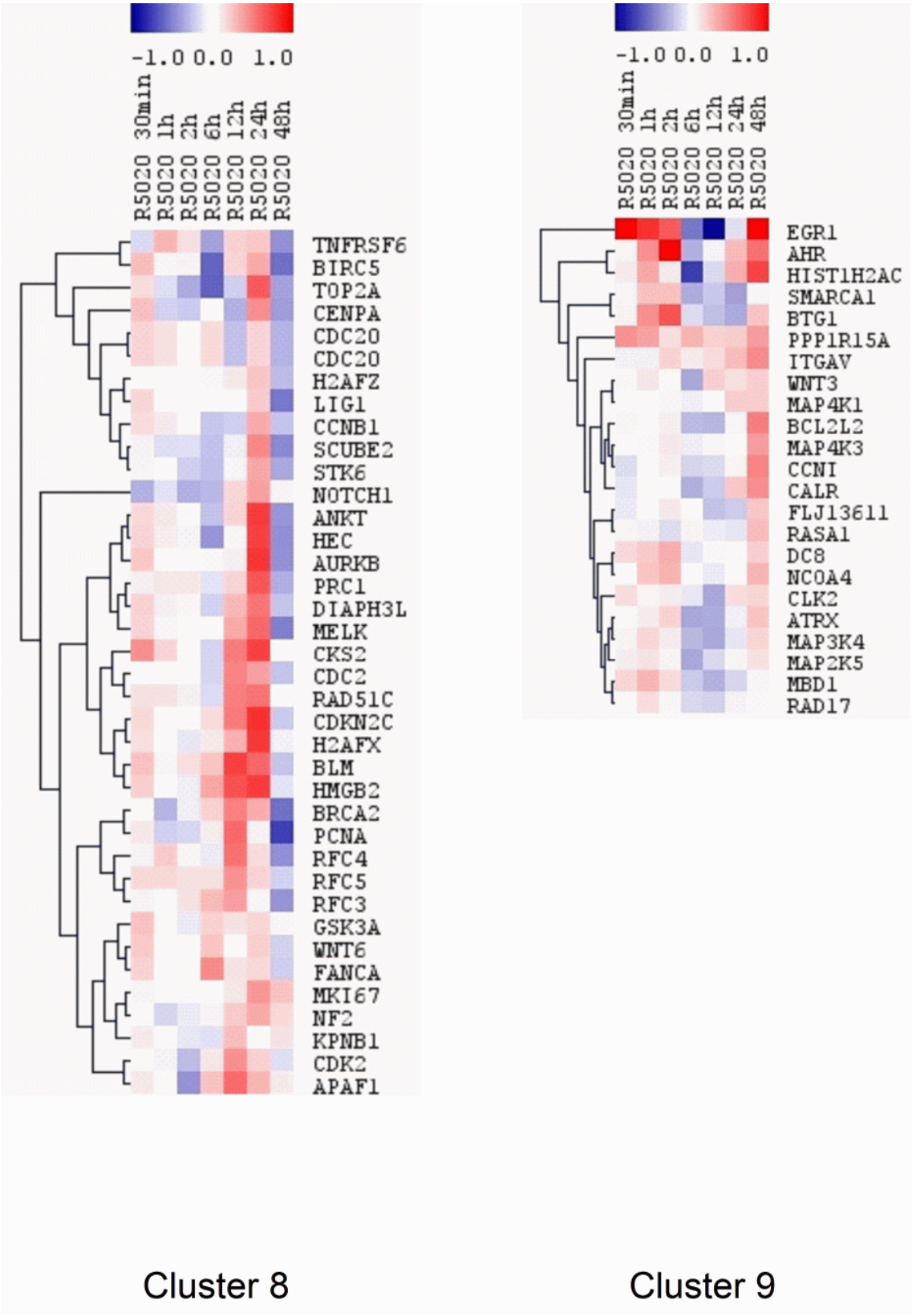
The resulting K-Means clusters grouping using the Euclidean distance as the distance metric and complete linkage of the EDGE 226 "within-temporal" significant genes (q < 0.01) in 9 groups which follow similar patterns of gene expression are shown in, **Panel 15**, **Panel 16** and **Panel 17**. Every resulting gene list from every cluster is also stored for future examination during the analysis of breast tumor samples, with the aim to identify similar patterns of gene expression in comparison with the breast cancer cell line hormonal response.



Panel 15: Resulting significant K-Means cluster (clusters number 1 to 5) grouping using the Euclidean distance as the distance metric and complete linkage of the EDGE 226 R5020 responsive significant genes (q < 0.01) in 9 groups which follow similar patterns of gene expression.



Panel 16: As in Panel 13 but for clusters 6 and 7.



Panel 17: As in **Panel 13** but for clusters 8 and 9.

A functional analysis using EASE-DAVID (NIH, http://david.niaid.nih.gov/david/ease.htm) was performed with every resulting cluster gene list to find overrepresented ontology categories among GO molecular function, GO biological process, and pathway databases such as the KEGG and the GenMAPP. The Fisher's exact probability taken as threshold was a p-value equal or less than 0.05. Top Ontology categories obtained are summarized in **Table 2.** We also obtain a Benjamini and Hochberg corrected p-value for multiple testing procedures (Benjamini and Hochberg 1995).

It was used as a background, the list of 820 genes of the collection of the Breast Cancer Array v4.0, therefore, since this list is a small set of genes, after multiple testing correction by Benjamini and Hochberg p-values are larger. Therefore functional analysis can be only considered it as an exploratory instrument unless BH p-values are less than 0.05.

Table 2: Overrepresented GO categories and GenMAPP pathways obtained with a threshold of 0.05 of the probability value of Fisher's exact test applied by the EASE program. Genes belonging to the cluster causing this overrepresentation are listed. Symbols mf and bf stand for molecular function and biological process, respectively.

K-means Cluster	Enrichment categories	Functional annotations categories	Genes	Fisher's exact test p-values	BH p.value
1	GO mf	receptor activity	IGFBP4, IGFBP5, AR, ERBB2, ITGA2	4.5E-3	0.2
	GO mf	protein binding	IGFBP4, IGFBP5, AR, NCOR2, ITGA2, ID4	1.1E-2	0.2
	GO mf	insulin-like growth factor binding	IGFBP4, IGFBP5	1.3E-2	0.2
	GO bp	regulation of cell growth	IGFBP4, IGFBP5	2.4E-2	0.2
	GO mf	signal transducer activity	IGFBP4, IGFP5, AR, ERBB2, ITGA2, GRB14	2.6E-2	0.2
	GO mf	transcription co-repressor activity	ID4, NCOR2, HDAC9	4.0E-2	0.2
2		no significant			
3	GO mf	epidermal growth factor receptor binding	EGF, TGFA	4.8E-5	1.8E-03
	GO mf	cytokine activity	EGF, TGFA	4.2E-2	8.9E-02
	GO bp	growth factor activity	EGF, TGFA	5.5E-3	8.9E-02
	GO bp	protein kinase cascade	EGF, STAT5A	1.2E-2	1.1E-01

	GO mf	receptor binding	EGF. TGFA	1.8E-2	1.2E-01
4	GO bp	cytokinesis	CDC45L, MCM5, CDC25A, CDC6, CDK8, CCNE2	8.1E-4	1.7E-01
	GO bp	mitotic cell cycle	RBBP4, SAP30, CDC45L, MCM5, CDC25A, EXO1, E2F1, CDC6, CCNE2	1.1E-3	1.7E-01
	GO bp	DNA replication checkpoint	CDC45L, CDC6	5.3E-3	2.6E-01
	GO bp	DNA replication initiation	CDC45L, MCM5	5.3E-3	2.6E-01
	GO bp	DNA dependent DNA replication	CDC45L, MCM5, EXO1, CDC6	5.7E-3	2.6E-01
	GO bp	S phase of mitotic cell cycle	RBBP4, CDC45L, MCM5, EXO1, CDC6	7.0E-3	2.6E-01
	KEGG pathway	integrin mediated cell adhesion	GRB2, HRASLS, SOS1, MAPK7	8.6E-3	2.6E-01
5	GO bp	protein kinase cascade	GADD45B, STAT5B, RPS6KA5, PIK3CB, MAP3K3	2.2E-3	2.5E-01
	GO bp	MAPKKK cascade	GADD45B, PIK3CB, MAP3K3	7.0E-3	2.5E-01
	GO bp	cell communication	ADRBK1, LAMA3, GADD45B, STAT5B, EGFR, RPS6KA5, AKAP13, VEGF, GAS6, PIK3CB, MAP2K1, MAP3K3	7.5E-3	2.5E-01
	GO bp	signal transduction	ADRBK1, GADD45B, STAT5B, EGFR, RPS6KA5, AKAP13, VEGF, GAS6, PIK3CB, MAP2K1, MAP3K3	1.0E-4	2.5E-01
	GO bp	negative regulation of cell proliferation	PCAF, QSCN6, ING1, CDKN1B	1.0E-2	2.5E-01
	GO bp	intracellular signaling cascade	GADD45B, STAT5B, RPS6KA5, AKAP13, PIK3CB, MAP3K3	1.1E-2	2.5E-01
	GO bp	regulation of cell proliferation	PCAF, VEGF, QSCN6, ING1, CDKN1B	!.2E-2	2.5E-01
	GO mf	kinase activity	ADRBK1, EGFR, RPS6KA5, AKAP13, PIK3CB, MAP2K1, DCK, MAP3K3	1.6E-2	2.5E-01
	GenMAPP pathway	Hs MAPK cascade	MAP2K1, MAP3K3	5.6E-2	2.5E-01
6	GO mf	RNA binding	RPL7, DDX5, PUM1, FUS, SNURF, BCAS2	5.1E-5	1.3E-02

	GO mf	nucleoside-diphosphate kinase activity	NME1, NME2	2.3E-3	2.8E-02
	GO mf	Nucleic acid binding	RPL7, GTF2H2, DDX5, APEX1, PUM1, FUS, POLE3, MSH6, NME2, SNURF, MCM6, TFDP1, BCAS2, E2F4, E2F5, E2F6, TOP1, GTF2F2	2.6E-3	2.8E-02
	GenMAPP pathway	Hs Cell cycle	CDC34, MCM6, TFDP1, E2F4, E2F5, E2F6, CDK4, CCNE1	2.9E-3	8.8E-02
	GO mf	pre-mRNA splicing factor activity	SNURF, BCAS2	6.9E-3	5.3E-02
	GO bp	ribonucleoside triphosphate biosynthesis	NME1, NME2	2.5E-3	2.8E-02
	GO bp	G1/S transition of mitotic cell cycle	CDK4, CCNE1, CDC34	1.1E-3	0.02
7	GO mf	phosphoprotein phosphatase activity	CDKN3, PPM1D, CDC25B, CDC25C	2.1E-3	3.0E-01
	GO mf	hydrolase activity	TREX1, CHD4, CDKN3, ERCC3, HDAC1, HDAC6, ARHGAP5, SMARCA2, SMARCA3, SMARCA5, PPM1D, CDC25B, CDC25C	4.4E-3	3.0E-01
	GO bp	establishment and/or maintenance of chromatin architecture	HIST1H1C, HIST1H1E, CHD4, H2AFJ, HDAC1, HDAC6, SMARCA5, H1F0	7.7E-3	3.3E-01
	GenMAPP pathway	Hs Cell cycle	CHEK2, HDAC1, HDAC6, CCNA2, CDC25B, CDC25C, CCNB2, CCND3	1.2E-2	3.4E-01
8	GO bp	mitotic cell cycle	LIG1, BIRC5, KNTC2, TOP2A, BRCA2, PCNA, CDC20, PRC1, STK6, CDC2, RFC3, RFC4, RFC5, CCNB1, CKS2, CDK2, BLM	2.0E-6	5.9E-04
	GO bp	M phase of mitotic cell cycle	KNTC2, BRCA2, CDC20, PRC1, STK6, CDC2, CCNB1, CDK2	1.1E-4	9.2E-03
	GO bp	cell proliferation	BLM, BRCA2, LIG1, NF2, TOP2A, BIRC5, CDC20, CDKN2C, CDC2, CDK2, CKS2, MKI67, PCNA, RFC3, RFC5, PRC1, AURKB, HEC, RFC5, CCNB1, STK6	1.7E-3	3.9E-02

9	GO mf	small GTPase regulatory / interacting protein activity	MAP4K1, MAP4K3, RASA1	6.5E-3	4.2E-01
	GO mf	transcriptional repressor activity	MBD1, PPP1R15A	1.3E-2	4.2E-01
	GO mf	transcription regulator activity	BTG1, ATRX, MBD1, NCOA4, PPP1R15A, CALR, AHR, EGR1	2.4E-2	4.2E-01
	GO mf	transcription factor activity	BTG1, ATRX, MBD1, AHR, EGR1	4.6E-2	4.2E-01

Every gene lists from the obtained K-Means cluster analysis was added to the customized gene list collection c5 for GSEA posterior analysis.

4.4.2 Temporal differential gene expression due to estradiol hormone treatments

Secondly, we carried out a "within-class" temporal differential expression analysis of the estradiol treated cell line culture T47D. After referring every time point of the gene expression data to their corresponding T_0, of each biological replicate or cell batch, we found (using EDGE) 127 genes (p-value ≤ 0.01) or 82 genes (q-value ≤ 0.01). In this case p-values behave better, and we obtained a larger number of significant genes because of a greater variability between the two estradiol cell treated replicates. The q-value as a function of the p-value is shown in **Figure 18.** There is a better behavior of the p-values with respect to the q-values, as we can note from the range and magnitude of the scale, the slope of the function of the p-value to the q-value of the "within-class" estradiol temporal differential gene expression.

Afterwards, M values were averaged from the obtained gene list of 127 genes of every experimental time of both biological replicates and imported into TMEV for visualization and clustering analysis. We performed, as with the R5020 time course experiment, a K-Means clustering analysis. We decided that, in this case, the best number of groups was 8 by looking at the average cluster size and the trend of the gene expression vector components. A representation of the obtained expression vector clusters is shown in **Figure 19.**

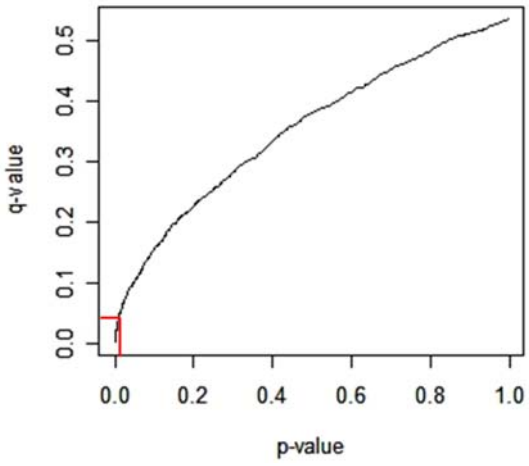


Figure 18: The q-value (y-axis) as a function of the p-value (x-axis). The red lines mark the location corresponding to p-value = 0.01.

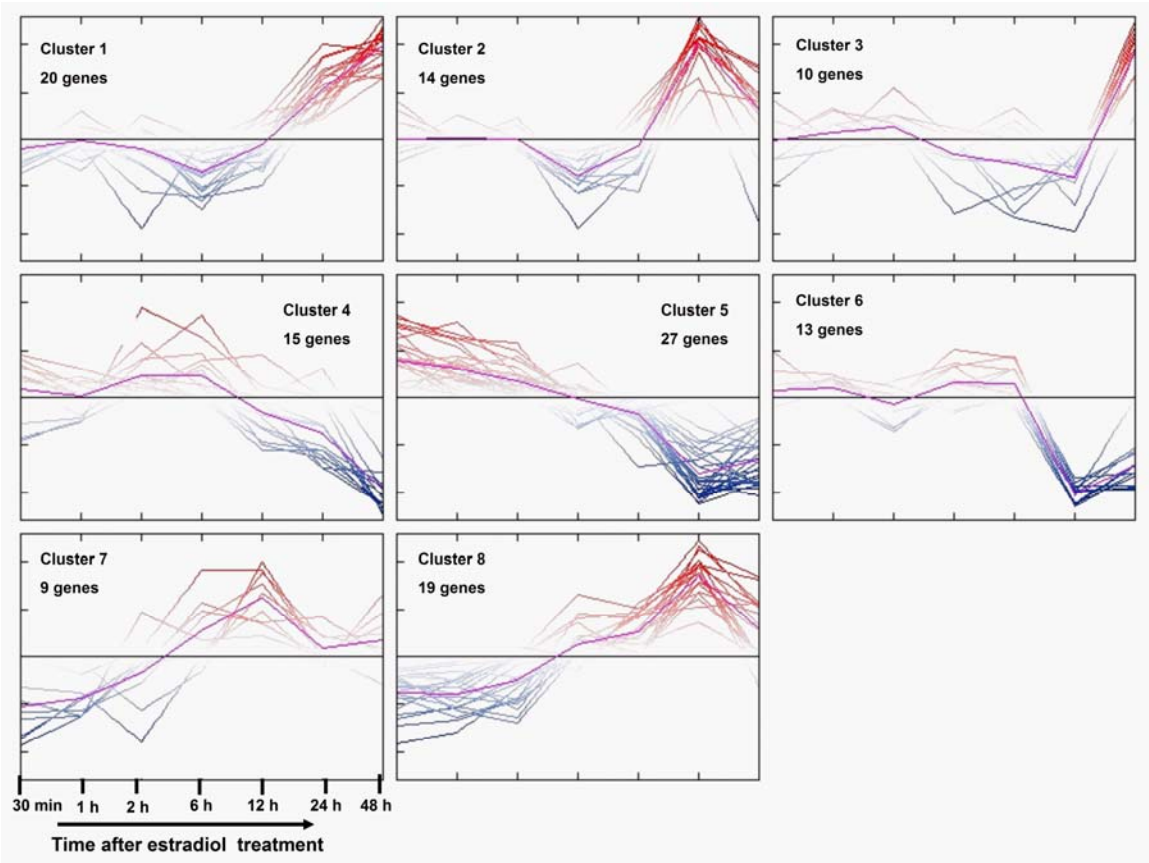


Figure 19: Time series of M (y-axis) as a function of the time after the Estradiol treatment (x-axis) for the eight groups obtained by K-Means clustering.

Smaller gene lists from each subcluster were generated, and functional analysis was performed to search their overrepresented GO term. We brought those 8 gene lists into EA (Enrichment Analysis) which uses GOstat (Falcon and Gentleman 2007), a Bioconductor package written in R, that allows to test GO terms for over or under-representation. We mainly looked into the GO

categories of biological process and molecular function. EA calculates the probability that any GO term would be overrepresented by the conditional hypergeometric test. We also obtain a Benjamini and Hochberg corrected p-value for multiple testing procedures (Benjamini and Hochberg 1995). The top GO terms are listed in **Table 3**. These values were obtained with a threshold value of 0.01 of the probability value of the conditional hypergeometric test Genes belonging to the cluster associated to the GO term are listed.

Table 3: Overrepresented GO terms from the GO biological process category.

K-Means Cluster	GO.Term	raw_p.value (p<0.01)	BH.p.value	Associated genes represented
1	chromosome condensation	6.2E-5	0.13	TOP2A, NUSAP1
	induction of apoptosis by intracellular signals	2.2E-3	0.14	CHEK2, CDKN1A, BBC3
	regulation of enzyme activity	2.2E-3	0.14	CDKN3, CKS1B, MAP4K1, CDKN1A, BBC3, CCNA2
	regulation of cyclin dependent protein kinase activity	3.3E-3	0.14	CDKN3, CKS1B, CDKN1A, CCNA2
2	DNA replication	0	8.0E-06	RFC5, RFC4, CCNE2, RBBP4, DTL, CDC6, RFC1, EXO1, CDK2
	DNA metabolism	1.0E-06	9.4E-05	RFC5, RFC4, CCNE2, HAT1, TOPBP1, RBBP4, HLTF, DTL, CDC6, RFC1, EXO1, CDK2
	DNA-dependent DNA replication	7.0E-06	3.2E-5	RFC4, CCNE2, CDC6, RFC1, EXO1, CDK2
	nucleobase, nucleoside, nucleotide and nucleic acids	2.1E-05	1.0E-2	RFC5, RFC4, E2F1, CCNE2, HAT1, TOPBP1, RBBP4, HLTF, DTL, CDC6, RFC1, EXO1, CDK2
	G1 phase of mitotic cell cycle	9.4E-6	0.03	E2F1, CDK2, CDC6
3	mitosis	2.1E-05	1.1E-3	BIRC5, CCNB1, CCNG1, CCNB2, AURKA
	M phase of mitotic cell cycle	2.1E-05	1.1E-3	BIRC5, CCNB1, CCNG1, CCNB2, AURKA

	M phase	1.5E-6	0.01	BIRC5, CCNB1, CCNG1, CCNB2, AURKA
	mitotic cell cycle	2.2E-5	0.01	BIRC5, CCNB1, CCNG1, CCNB2, AURKA
	G2/M transition of mitotic cell cycle	1.1E-3	0.03	BIRC5, CCNB1
	cell division	3.3E-3	0.06	BIRC5, CCNB1, CCNG1, CCNB2
4	no significant			
5	B cell differentiation	1.2E-3	0.24	HDAC7A, HDAC9, PIK3R1
	nucleosome spacing	0.01	0.24	HIST1H1A, HIST1H1E
	phospholipid metabolism	0.01	0.24	AYTL1, PIK3R1
6	cellular protein metabolism	0.01	0.24	UBR2, MAP3K2, RASA1, CDC2L5, SEP15, MAP4K3, PARP3
	negative regulation of cell adhesion	0.01	0.24	RASA1
	posttranslational protein folding	0.01	0.24	SEP15
7	regulation of fibroblast growth factor receptor	0.01	0.34	RUNX2
	negative regulation of smoothened signaling pathway	0.01	0.34	RUNX2
	fibroblast growth factor receptor signaling pathway	0.01	0.34	RUNX2
8	purine ribonucleotide metabolism	1.1E-3	0.07	NME1, GMPS
	purine ribonucleotide biosynthesis	1.1E-3	0.07	NME1, GMPS
	ribonucleotide biosynthesis	1.1E-3	0.07	NME1, GMPS
	ribonucleotide metabolism	1.1E-3	0.07	NME1, GMPS
	nucleotide biosynthesis	2.2E-3	0.08	NME1, GMPS

It was used as a background, the list of 820 genes of the collection of the Breast Cancer Array v4.0, therefore, since this list is a small set of genes, after multiple testing correction by Benjamini and Hochberg p-values are larger. Therefore functional analysis can be only considered it as an exploratory instrument unless BH p-values are less than 0.05.

We conclude from this experiment that gene expression changes (M values) with estradiol hormonal treatment are larger than in the case of progestin treatment. This is so because the absolute changes of the Log_2Ratio values are larger throughout the time (greater *M* values) in the case of estradiol hormonal treatment.

However, the cell cycle progression to mitosis is delayed, the transition of G_1 (the first gap phase of the cell division cycle) to the S phase (the DNA replication phase) occurs only after 24 hr, and the M phase of the mitosis takes place only once 48 hr after hormonal induction. It seems that activation of transcription on our model cell line T47D-MTVL by estradiol is not as effective as it is with progestin. This probably happens because cytoplasmatic signaling cascades through MAPK and PI3K kinases signaling pathway are to a large extent effective initiating transcription on cell cycle progression target genes.

Every gene lists from the obtained K-Means cluster analysis after estradiol treatment was also added to the customized gene list collection c5 for GSEA posterior analysis.

4.4.3 Distinctive profiles of temporal differential gene expression between progestin and estradiol hormone treatments

In order to further investigate this differential response to progestin or to estradiol in a dynamic temporal analysis, combined data from both treatments were imported into the time-dependent statistical analysis program EDGE. “Between-class” temporal differential expression was selected, being “treatment”, either progestin or estradiol the class variable selected for differential analysis of gene expression. A set of 113 significant genes were obtained with p-value ≤ 0.01, and 62 significant genes with q-value ≤ 0.01. The q-value threshold was selected this time in order to obtain a more conservative set of genes. Log_2Ratio values of both experimental biological replicates were averaged, and resulting *M* data of the 62 genes were imported into TMEV for visualization and clustering. Genes were clustered using an unsupervised hierarchical clustering algorithm, and the resulting gene tree obtained is shown in **Figure 20** with the most significant GO terms which show up as the output from the statistical program EASE-DAVID for functional analysis of each gene list.

As is observed in the pattern of the gene subclusters, genes are time delayed in the estradiol treatment in comparison with the progestin treatment, as we previously noted, so that it looks like cell cycle progression is more efficiently activated by progestins. This behavior is observed in every gene subcluster. However, it seems that genes in the group marked with a light green bar are only activated by progestins between 6 and 12 hr, and not by estradiol, or have an inferior response. These genes are the transcription factor GTF2H2 (General transcription factor IIH, polypeptide 2, 44kDa), the membrane receptor NEO1 (Neogenin chicken homolog 1), the protein tyrosine kinases MAPK7 (Mitogen-

activated protein kinase 7), ABL1 (V-abl Abelson murine leukemia viral oncogene homolog 1), and RPS6KA1 (RSK1, Ribosomal protein S6 kinase, 90kDa, polypeptide 1). These are genes involved in cytoplasmatic mitogen kinase cascades that we found in our search for functional annotations.

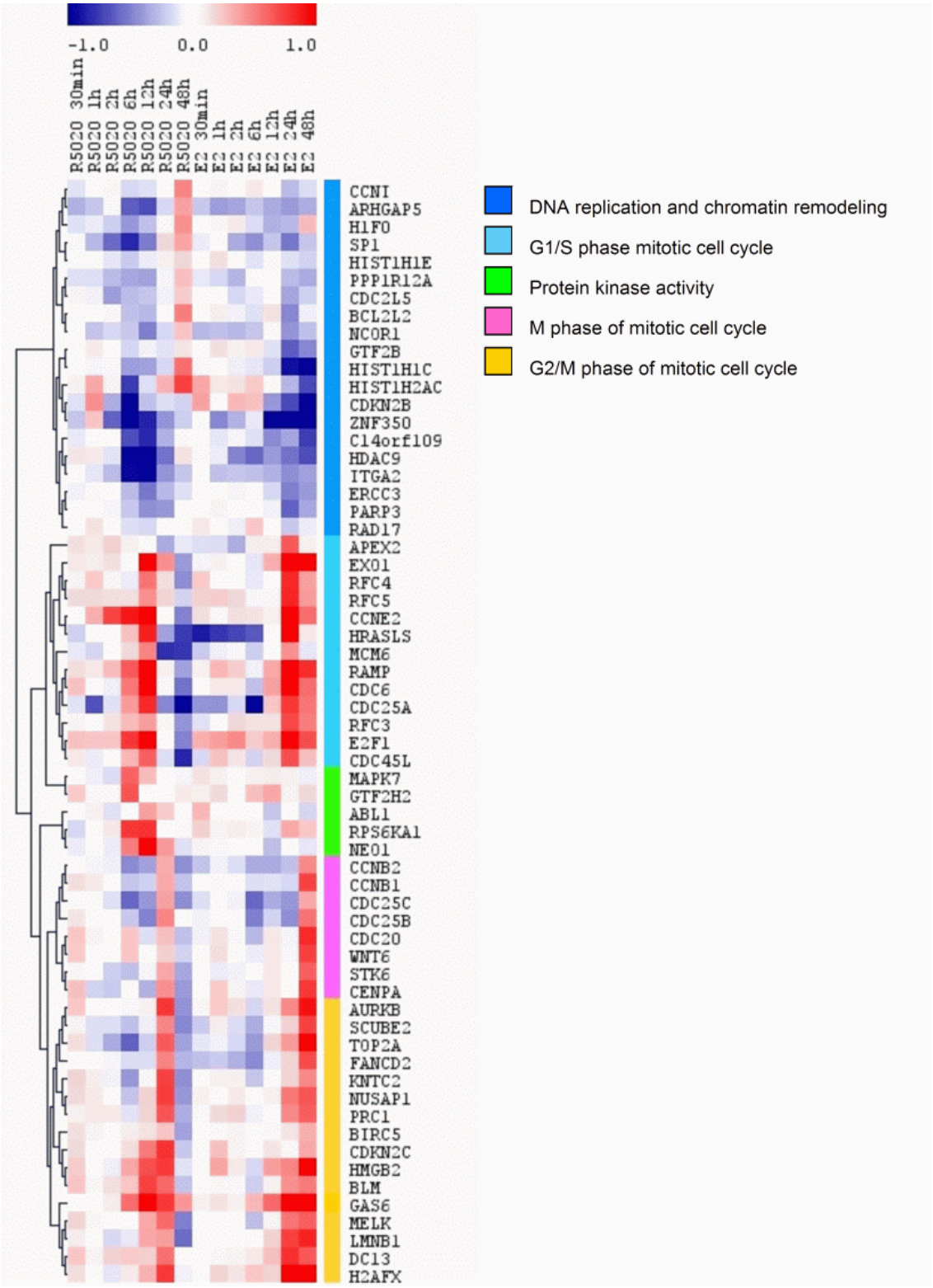


Figure 20: Unsupervised hierarchical cluster of the genes found with distinctive temporal differential expression between progestin and estradiol treatment. The different genes were clustered using the Pearson correlation coefficient as distance metric and complete linkage as the aggregative clustering algorithm. These clusters are labeled with a different color code with the significant functional analysis output from the statistical program EASE-DAVID.

4.4.4 Common profiles of temporal differential gene expression among progestin and estradiol hormone treatments

In order to investigate the genes that behave similarly on the dynamic response to progestin or to estradiol, combined data from both treatments were analyzed. The data were imported into the time-dependent statistical analysis program EDGE and a “within-class” temporal differential expression was carried out. Our aim here was to find similarities among treatments while in the previous case our aim was to search for dissimilarities.

The class variable selected for differential analysis of gene expression was “treatment”, either progestin or estradiol. A set of 97 significant genes was obtained with a threshold p-value ≤ 0.01, or a set of 24 significant genes with q-value ≤ 0.01. The p-value threshold was this time selected to obtain a larger gene list, since this time a smaller number of significant genes were found with the q-value threshold. Log_2Ratio values of both experimental replicates were then averaged, and the resulting data matrix was imported into TMEV for visualization and clustering.

Genes were clustered using an unsupervised hierarchical clustering algorithm, and the resulting gene tree obtained is shown in **Figure 19**. We observed the genes that after a different hormonal treatment, progestin or estradiol, follow a similar gene expression pattern along the time. Remarkably, there is a group of transcription factors that are activated early by both treatments, with estradiol and progestins, such as FOS (V-fos FBJ murine osteosarcoma viral oncogene homolog), ATF3 (Activating transcription factor 3), SNAI1 (Snail homolog 1, Drosophila), and MYC (V-myc myelocytomatosis viral oncogene homolog, avian). This group of genes is marked with a yellow bar in **Figure 21**.

In the next chapter we further investigate the reason why the genes in this group behave similarly, so that it is possible that different hormone treatments might activate the same transcription factors. They may have the same mechanism of transcriptional activation, or ultimately what are their target proteins as transcription factors.

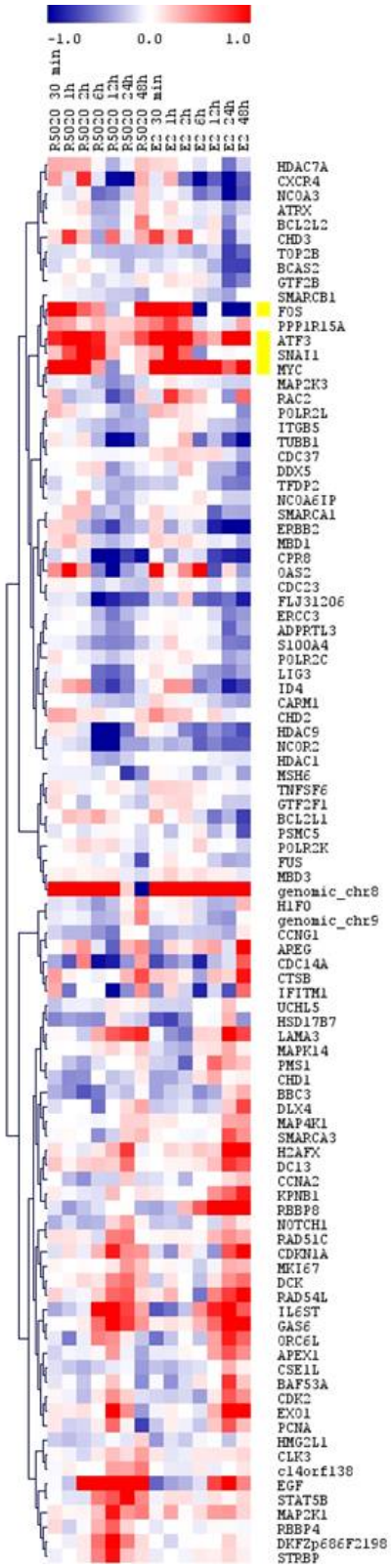


Figure 21: As in Figure 16 but for the genes found similarly expressed across time in response to both progestin and estradiol treatments.

4.5 Hormonal induction inhibitors

As described on chapter 2.2, addressing the different mechanisms of action of steroid hormones, progestins transactivate rapidly and transiently estrogen-target genes via a cross-talk between ERα and PR-B, and activate the MAP kinase Erk1/2 signaling pathway. This induction of endogenous estrogen target genes by progestins might be essential for its proliferative effects on breast cancer. It is also known that progestins repress estradiol-induced endometrial proliferation *in vivo* and repress estrogen receptor function *in vitro* (Vegeto *et al.* 1993, Tung *et al.* 1993, McDonell *et al.* 1994, Kraus *et al.* 1995).

To elucidate the contribution of the different mechanisms of action of progestins in the induction of hormone target genes, we have used PD98059 (hereafter PD) as an inhibitor of the Erk1/2 pathway, and ICI182780 (hereafter ICI) as an antagonist of ER, and investigated which hormone target genes are inhibited by PD, by ICI, or both drugs (in collaboration with C. Ballaré, CRG, Barcelona).

Genes that are induced by progestin and later on inhibited by both PD and ICI will be genes that are activated via ERα-PR-B crosstalk and are dependent on the Erk1/2 signaling pathway. On the other hand, genes that are induced by progestin and later only inhibited by PD and not by ICI will be genes dependent on the Erk1/2 signaling pathway but ER independent. Genes that are induced by progestins but inhibited only by ICI and not by PD are consequently independent of the Erk1/2 pathway and only ER dependent, and therefore are dependent on the PI3K/Akt signaling pathway or other still unknown mechanisms. And finally, genes that are induced by progestins but neither inhibited by PD nor by ICI will be independent of ER-PR crosstalk, independent of Erk1/2 and would be due to a purely genomic signaling pathway or other undefined mechanisms (**Figure 22**).

In order to investigate the effect of PD and/or ICI on the differential gene expression of hormone target proteins, we treated our model cell line T47D-MTVL with either R5020 or estradiol, and simultaneously added either PD or ICI. After 6 hr of treatment with R5020, or 1 hr after estradiol, total RNA was prepared from the corresponding cell cultures. There were also untreated cell culture samples and a sample treated with PD drug vehicle. Two biological replicates of every sample were prepared for statistical inference with one week of time difference.

Every sample was labeled and hybridized on the breast cancer array platform v3.1 taking, as usual, as reference sample the UHRR, since we wanted to contrast paired samples between two different conditions. Images were quantified, arrays were normalized, and M values were calculated (see chapters 3.9.2, 3.10, and 3.11). We carried out a paired sample contrast analysis between different conditions, using SAM “two-class unpaired” (see chapter 3.13).

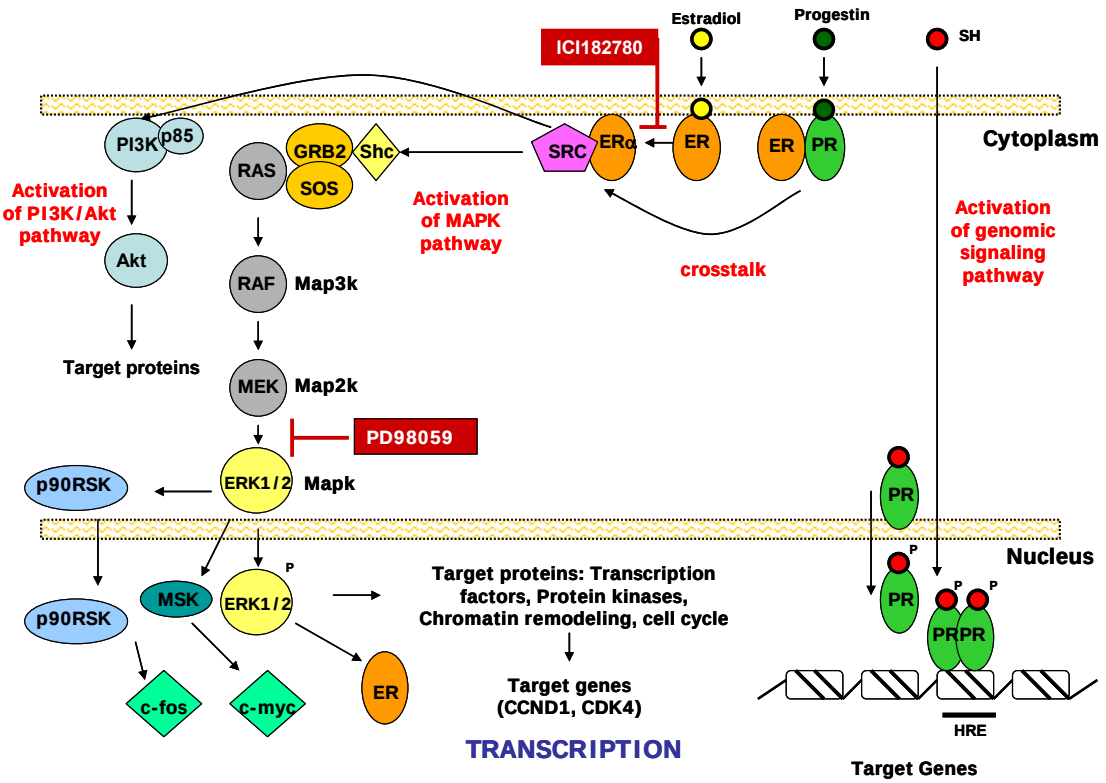


Figure 22: Effect of inhibitors PD98059 and ICI182780 on the ligand-mediated signaling pathways by steroid hormones.

The selected settings to discriminate significantly differentially expressed genes were (1) the FDR threshold to 0.05, that is, a 5% of falsely discovered genes, and (2) the fold-change threshold of 1.3, that is, the relative change of the expression ratio between two unpaired classes must be at least 1.3. Less stringent than the previous 1.4 fold-change threshold since this time we have a replicate experiment.

The SAM version used runs on the R environment and is implemented in our laboratory server. The first contrast analysis use samples induced by R5020 after 6 hr versus the untreated samples and, as a result, 120 genes were obtained with a FDR less than 5%. **Appendix A8** (inhibition by PD or ICI or both of R5020 responsive genes) represents the genes that are induced by R5020 after 6 hr, either upregulated or downregulated, and the corresponding q-value percentage. The columns show the results after PD or ICI treatment and the corresponding q-value obtained in the contrast analysis. NA means “not applicable” since the p-value is larger than our FDR threshold.

We conclude, from the results of the experiments, that some of the genes belong to the investigated signaling pathways since they were found inhibited by PD, ICI or both..

Genes that are induced by progestin and later on inhibited by both PD and ICI are genes that are activated via ERα-PR-B crosstalk and hence are target genes dependent on the Erk1/2 signaling pathway. These include, for example,

- RASL10B (RAS-like. family 10. member B)
- ELL2 (Elongation factor. RNA polymerase II. 2)
- EGFR (Epidermal growth factor receptor)
- CDKN1C (Cyclin-dependent kinase inhibitor 1C, p57. Kip2)
- EGLN1 (Egl nine homolog 1, *C. elegans*)
- H3F3B (H3 histone. family 3B)
- PTCH2 (Patched homolog 2, Drosophila)
- ORC6L (Origin recognition complex. subunit 6 homolog-like, *S. cerevisiae*)
- FANCA (Fanconi anemia. complementation group A)
- POLD4 (Polymerase (DNA-directed). delta 4)
- IGFBP5 (Insulin-like growth factor binding protein 5).

Genes that are induced by progestin and later only inhibited by PD but not by ICI are genes dependent on the Erk1/2 signaling pathway but independent of ER. Some examples are,

- EGF (epidermal growth factor)
- MUC2L (mucin 2 like)
- PPARGC1B (Peroxisome proliferative activated receptor γ coactivator 1 β)
- GRB2 (Growth factor receptor-bound protein 2)
- CCNE2 (Cyclin E2)
- PLAUR (Plasminogen activator. urokinase receptor)
- QSCN6 (Quiescin Q6)
- STAT3 (Signal transducer and activator of transcription 3)
- RAMP (Denticleless homolog, *Drosophila*)
- MXI1 (MAX interactor 1)
- MKLN1 (Muskelin 1. intracellular mediator containing kelch motifs)
- CDC6 (CDC6 cell division cycle 6 homolog, *S. cerevisiae*)
- NEO1 (Neogenin homolog 1, chicken)
- MTA1 (Metastasis associated 1
- NCOA3 (Nuclear receptor coactivator 3)
- CHEK2 (CHK2 checkpoint homolog, *S. pombe*)
- ARHGAP5 (Rho GTPase activating protein 5)
- NBS1 (Nibrin)
- IGFBP4 (Insulin-like growth factor binding orotein 4)
- ITGA5 (Integrin. alpha 5 (fibronectin receptor α polypeptide)

Genes that are induced by progestins but inhibited only by ICI and not by PD, are consequently independent of the Erk1/2 pathway and only dependent on ER, therefore dependent on the PI3K/Akt signaling pathway or other still unknown mechanisms. Examples are,

- BIRC3 (Baculoviral IAP repeat-containing 3)
- TP53BP2 (Tumor protein p53 binding protein 2)
- SERPINA3 (Serine (or cysteine) proteinase inhibitor clade A member 3)
- HSPA9B (Heat shock 70kDa protein 9B, mortalin-2)
- IGFBP3 (Insulin-like growth factor binding protein 3)
- HSD17B3 (Hydroxysteroid,17 β dehydrogenase 3)
- SOS1 (Son of sevenless homolog 1, *Drosophila*)
- E2F3 (E2F transcription factor 3)
- NFIB (Nuclear factor I)
- LMNA (Lamin A/C)
- NOTCH3 (Notch homolog 3, *Drosophila*)
- GATA3 (GATA binding prorein 3)
- AR (Androgen receptor)
- UNG2 (Uracil-DNA glycosylase 2).

And finally, genes that are induced by progestins but not inhibited by neither PD nor ICI are genes that are independent of ERα-PR-B crosstalk, independent of Erk1/2 pathway and likely to be pure genomic signaling, PI3K/Akt or JAK/STAT signaling pathway or other still unknown mechanisms. As examples, we have,

- STAT5A (Signal transducer and activator of transcription 5A)
- TGFA (Transforming growth factor α)
- DUSP1 (Dual specificity phosphatase 1)
- GADD45A (Growth arrest and DNA-damage-inducible α)
- CCND1 (Cyclin D1)
- HMGB3 (High-mobility group box 3)
- JUN (v-jun sarcoma virus 17 oncogene homolog, avian)
- IL6ST (Interleukin 6 signal transducer gp130 oncostatin M receptor)
- SAP30 (Sin3-associated polypeptide 30kDa)
- KLF5 (Kruppel-like factor 5, intestinal)
- XLHSRF-1 (Dynein axonemal. heavy polypeptide 1)
- CDC14B(CDC14 cell division cycle 14 homolog B, *S. cerevisiae*)
- RPS6KA5 (Ribosomal protein S6 kinase 90kDa polypeptide 5)
- AKAP13 (A kinase (PRKA) anchor protein 13)
- GADD45B (Growth arrest and DNA-damage-inducible β)
- MAP3K3 (Mitogen-activated protein 3 kinase 3)
- CXCL12 (Chemokine (C-X-C motif) ligand 12)
- SNAI1 (Snail homolog 1, *Drosophila*)
- VEGF (Vascular endothelial growth factor)
- RPS6KA1 (Ribosomal protein S6 kinase. 90kDa. polypeptide 1).

Secondly, a two-class unpaired contrast analysis was performed using SAM, samples induced by estradiol after 1 hr versus the untreated samples were confronted, but the FDR threshold selected could not be so stringent since, in the case of the estradiol treatment, there was a greater variation between biological replicates at the earlier time of 1 hr, and the fact that we only had two biological replicates gave us, in this case, too little statistical power. Therefore

we just selected the 96 genes induced by a fold greater than 1.3 or smaller than 1.3. (**Appendix A9**: Inhibition by PD or ICI or both of estradiol responsive genes).

On this table the genes that are induced by estradiol after 1 hr of hormonal treatment is represented, the upregulated or the downregulated ones. Also the corresponding q-value percentage is calculated by SAM. The following columns are the result after PD or ICI treatment and the corresponding q-value due to the two class contrast analysis. NA means "not applicable" since the FDR threshold came above 5%.

As a conclusion from our experiment, we could note some of the genes that are induced by estradiol and are inhibited by both PD and ICI, therefore are activated via ERα-PR-B crosstalk and thus are dependent on the Erk1/2 signaling pathway are, for example,

- TOP3B (Topoisomerase DNA, III b)
- SNAI1 Snail homolog 1, *Drosophila*)
- IGFBP3 (Insulin-like growth factor binding protein 3)
- GADD45A (Growth arrest and DNA-damage-inducible α)
- TIMP3 (Tissue inhibitor of metalloproteinase 3)
- SERPINA3 (Serine or cysteine proteinase inhibitor clade A member 3)
- CDKN1C (Cyclin-dependent kinase inhibitor 1C, p57 Kip2)
- PPP1R15A (Protein phosphatase 1 regulatory inhibitor subunit 15A)
- AKAP9 (A kinase PRKA anchor protein 9)
- STAT5A (Signal transducer and activator of transcription 5A)
- SCD4 (Stearoyl-CoA desaturase 5)
- MYB (V-myb myeloblastosis viral oncogene homolog, avian)
- KRT5 (Keratin 5)
- WISP2 (WNT1 inducible signaling pathway protein 2)
- DDB1 (Damage-specific DNA binding protein 1 127kDa)
- DNMT3B (DNA cytosine-5--methyltransferase 3 β)
- E2F1 (E2F transcription factor 1)
- GATA3 (GATA binding protein 3)
- MGST1 (Microsomal glutathione S-transferase 1)
- GSTM3 (Glutathione S-transferase M3, brain)
- RPS6KA5 (Ribosomal protein S6 kinase. 90kDa. polypeptide 5)
- TGFB3 (Transforming growth factor β 3)
- IGF1 (Insulin-like growth factor 1, somatomedin C)

Genes that are induced by estradiol and later only inhibited by PD but not by ICI would be genes dependent of the Erk1/2 signaling pathway but independent of ER. As examples we could note:

- ATF3 (Activating transcription factor 3)
- DDIT3 DNA-damage-inducible transcript 3)
- ITGA5 (Integrin. α 5)

- CDC14A (CDC14 cell division cycle 14 homolog A, *S. cerevisiae*)
- ZNF350 (Zinc finger protein 350)
- MAP3K5 (Mitogen-activated protein kinase kinase kinase 5)
- PIK3CA (Phosphoinositide-3-kinase catalytic α polypeptide)
- ADRA1B (Adrenergic α 1B receptor)
- BRCA2 (Breast cancer 2 early onset)
- SAP18 (Sin3-associated polypeptide 18kDa)
- RUNDC1 (RUN domain containing 1)
- IL6ST (Interleukin 6 signal transducer, gp130 oncostatin M receptor)
- GPR126 (G protein-coupled receptor 126)
- CTNNB1 (Catenin - cadherin-associated protein- β 1. 88 kDa)
- NCOA2 (Nuclear receptor coactivator 2)
- FLT1 (Fms-related tyrosine kinase 1 - vascular endothelial growth factor)
- ESRRB (Estrogen-related receptor β)
- C20orf46 (Chromosome 20 open reading frame 46)
- CaMKIINalpha (Calcium/calmodulin-dependent protein kinase II inhibitor 1)
- UNG2 (Uracil-DNA glycosylase 2)
- WNT5B (Wingless-type MMTV integration site family. member 5B)
- ERBB3 (V-erb-b2 erythroblastic leukemia viral oncogene homolog 3, avian)
- MAPK13 (Mitogen-activated protein kinase 13)
- ATM (Ataxia telangiectasia mutated)
- SNRPN (Small nuclear ribonucleoprotein polypeptide N)
- ADAM15 (A disintegrin and metalloproteinase domain 15 (metargidin))
- HIRA (HIR histone cell cycle regulation defective homolog A, *S. cerevisiae*)
- AR (Androgen receptor)
- DNMT3A (DNA, cytosine-5-methyltransferase 3 α)

Genes that are induced by estradiol but inhibited only by ICI and not by PD, are consequently independent of the Erk1/2 pathway and only dependent of ER, therefore dependent of the PI3K/Akt signaling pathway or other still unknown mechanisms are, for example,

- FOS (V-fos FBJ murine osteosarcoma viral oncogene homolog)
- MYC (V-myc myelocytomatosis viral oncogene homolog, avian)
- TFF1 (Trefoil factor 1)
- GADD45B (Growth arrest and DNA-damage-inducible β)
- IGFBP1 (Insulin-like growth factor binding protein 1)
- ENPP2 (Ectonucleotide pyrophosphatase phosphodiesterase 2)
- CDC42 (Cell division cycle 42, GTP binding protein 25kDa)
- HSD17B3 (Hydroxysteroid 17-β dehydrogenase 3)
- PTCH (Patched homolog, *Drosophila*)
- WT1 (Wilms tumor 1)
- HSD17B7 (Hydroxysteroid 17-β dehydrogenase 7)

And finally, genes that are induced by estradiol but not inhibited by neither PD nor ICI are genes independent of ERα-PR-B crosstalk, independent of Erk1/2 pathway, and would be due to pure genomic signaling pathway or other still unknown mechanism. Some of the genes that are strongly upregulated at 1 hr after hormone induction are:

- EGR1 (Early growth response 1)
- JUN (V-jun sarcoma virus 17 oncogene homolog, avian)
- DUSP1 (Dual specificity phosphatase 1)
- ID4 (Inhibitor of DNA binding 4)
- CKS2 (CDC28 protein kinase regulatory subunit 2)

See **Appendix A10** for the generated Venn diagrams of progestin and estradiol hormone induction and inhibition by PD and ICI.

Every gene lists from this analysis was also added to the customized gene list collection c5 for GSEA posterior analysis.

4.6 Confirmation of microarray results by Real Time qPCR

Real time quantitative PCR assays were performed to confirm the gene expression of some of the genes. This is important in order to validate the results obtained by microarray experiments. Primer design, procedures, efficiency calculation, and data analysis are described in detail in chapter 3.20.

The obtained PCR efficiencies of the assayed genes are listed in **Table 4**.

Table 4: Assay efficiency obtained for the evaluated genes.

Gen symbol (alias)	RefSeq Acc. No.	Efficiency	Efficiency (%)
GAPD	NM_002046	1.754	87.70
ACTN	NM_001102	1.775	88.75
FOS	BX647104	1.764	88.20
MYC	NM_002467	1.654	82.70
TFF1 (PS2)	NM_003225	1.733	86.65
CCND1	NM_053056	1.747	87.35
RPS6KA1 (RSK1)	BC014966	1.670	83.50
RPS6KA5 (MSK1)	AB209667	1.694	84.70
MUC2L	BG675392	1.665	83.25
CCNE2	NM_057749	1.855	92.75

First of all, we determined which gene was the best one we could use for normalization in relative quantification analysis. It is well known that there is not a universal endogenous gene, also called housekeeping, since their levels of expression are not always constant (Barber *et al*. 2005). We first carried out a bibliographic search looking for reference genes used in real time assays for our tissue type (Szabo *et al*. 2004). The reference gene of choice should be stably expressed in breast tissue, and in our breast cancer cell line model, and most importantly, with a minimal variation between treated and untreated samples or normal and tumor tissue.

A combination of genes has been used as a normalization factor as well. This is usually done by calculating a geometric mean of a combination of 3 to 5 reference genes (Vandesompele *et al.* 2002). The underlying assumption is that gene pairs showing stable expression patterns relative to each other are appropriate control genes. However, this model requires extensive practical validation to identify a combination of reference genes appropriate for an individual experiment. To avoid this extensive search, we first tried a couple of reference genes, commonly used as endogenous genes such as ACTN (Actinin, α 1) and GAPD (Glyceraldehyde-3-phosphate dehydrogenase), and carried out real time quantitative PCR assays against our treated and untreated cell line samples. Three more target genes (FOS, MYC, and TFF1) were included for comparison. Overall behavior of ACTN and GAPD with respect to the target genes was analyzed. We used for the data analysis, the excel-based program Bestkeeper©, that performs a pair-wise correlation analysis between various reference and target genes, in order to find the best reference gene with a minimal variation with the assayed samples (Pfaffl *et al.* 2004). We concluded, after this analysis, that GAPD had the minimal pair-wise correlation to the target genes, and consequently, we used it as our reference gene.

We observed, looking at the PCR amplification plots, the perfect overlap of the amplification plots of treated and untreated samples in the case of GAPD, while this did not occur with ACTN. This is because the total RNA starting amount before reverse transcription was the same one in all samples, and all these samples were equally treated throughout the qPCR procedure. This observation was our best confirmation that GAPD was the best endogenous gene to use as reference gene in relative quantification of the cell line experiments.

The results obtained with the set of analyzed genes are shown in graphical bar representation of the relative copy number of every sample normalized to the calibrator sample T_0 (**Figures 23-30**). We can compare the R5020 samples before and after PD or ICI inhibitors at 6 hr after hormone induction (orange color bars in **Figures 23-30**). For comparison we include also the estradiol treated samples before and after 1 hr of treatment with PD or ICI treatment (yellow bars).

Microarray measurements are not as accurate as those by real time qPCR assays. Fold-change ratios are larger when analyzed by real time qPCR due to the different type of complementary DNA strand hybridization. In the case of microarray hybridization, one DNA strand, the PCR product of the cDNA, is fixed to a solid support, the slide. However, in the case of Real Time qPCR, both hybridizing complementary DNA strands are in solution. This might be the reason why real time qPCR measurements are more sensitive and reproducible than microarray measurements. Besides, there is a risk of cross-hybridization at microarray experiments since it is performed under a single hybridization temperature that might not be optimal for all probes. Real time qPCR has a much broader dynamic range than microarray techniques.

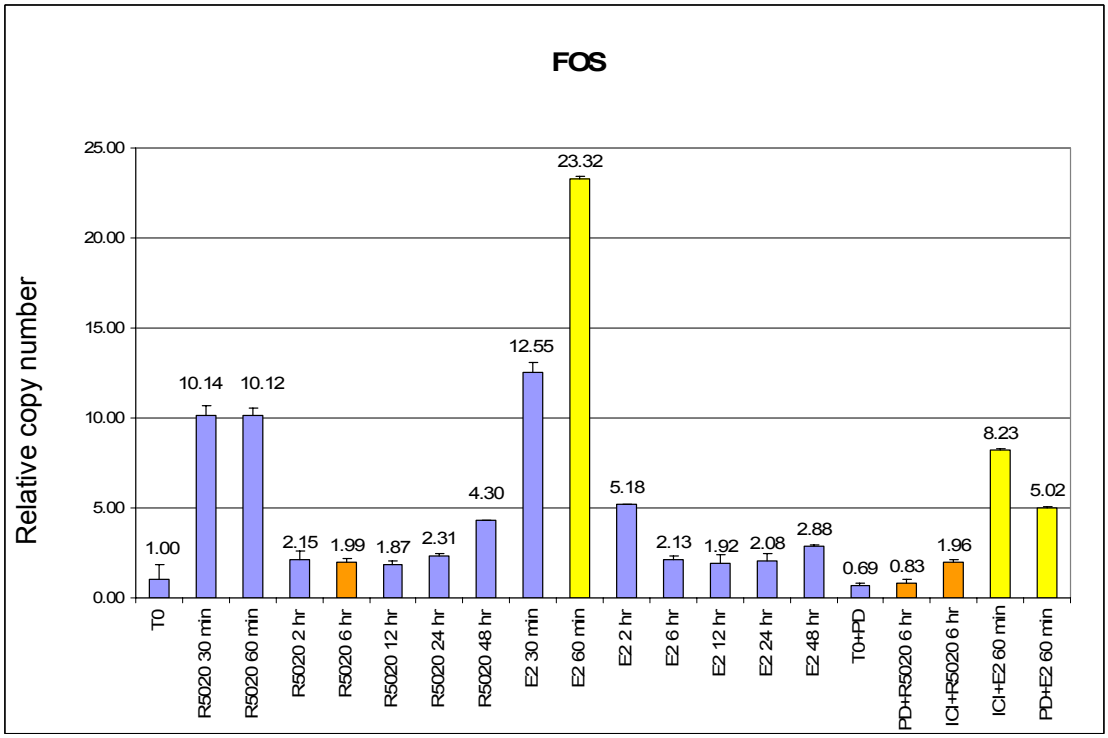


Figure 23: Relative copy number for the hormone treated time series and inhibitors by Real Time qPCR in the analysis of FOS.

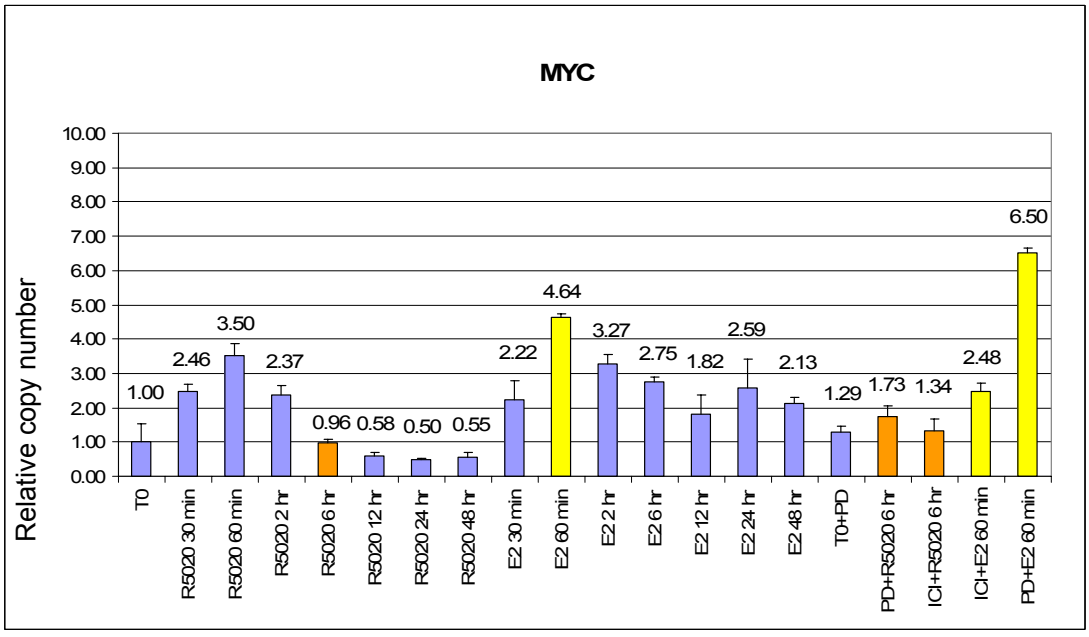


Figure 24: As in **Figure 23** but for gene MYC**.**

FOS and MYC are transcription factors that are strongly upregulated by estradiol after 1 hr of treatment, and are inhibited by the estrogen antagonist ICI. FOS is also inhibited by PD but to a minor degree as it is shown in the bar graph (**Figure 23** and **Figure 24**). FOS is induced by R5020 treatment with a 2 fold ratio at 6 hr and later inhibited by PD and not by ICI. This was not seen by

microarray analysis since fold-change ratio is below cut-off of significance, and fold-change ratios values are compressed in microarray measurements.

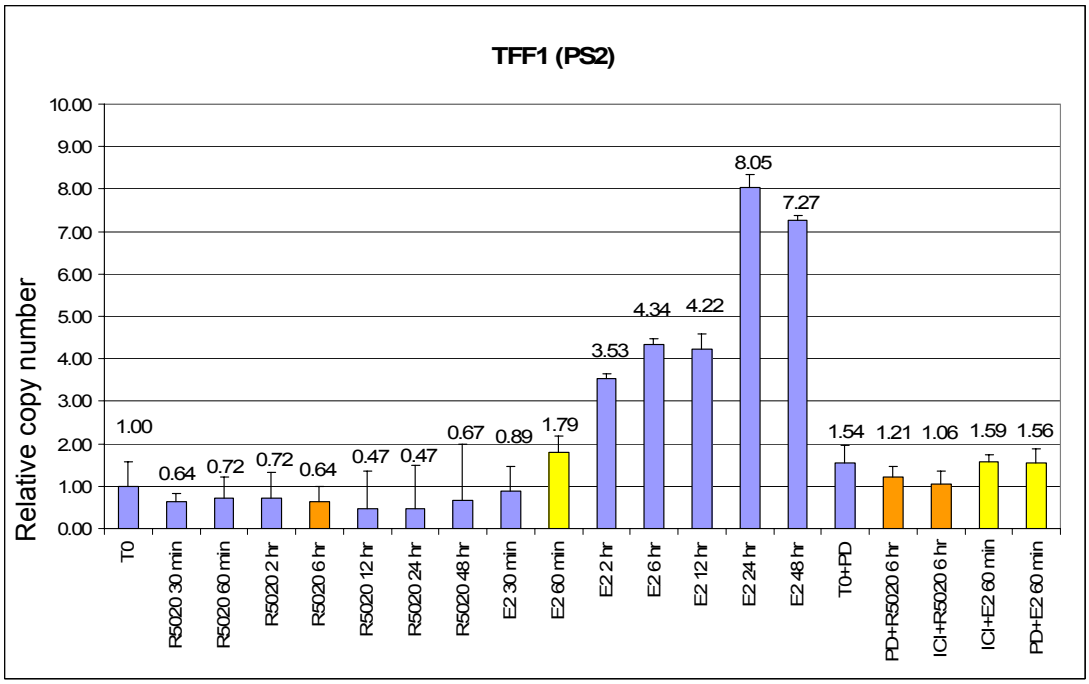


Figure 25: As in **Figure 23** but for the TFF1 (PS2) gene.

Real time PCR quantification shows that TFF1 (PS2, trefoil factor 1) is not affected by any of the two inhibitors, at least, at these times, but from microarray results seems affected by ICI after estradiol treatment (see **Table 5**: Inhibition by PD or ICI or both of estradiol responsive genes). TFF1 is strongly induced by estradiol after one hour of hormone treatment (**Figure 25**).

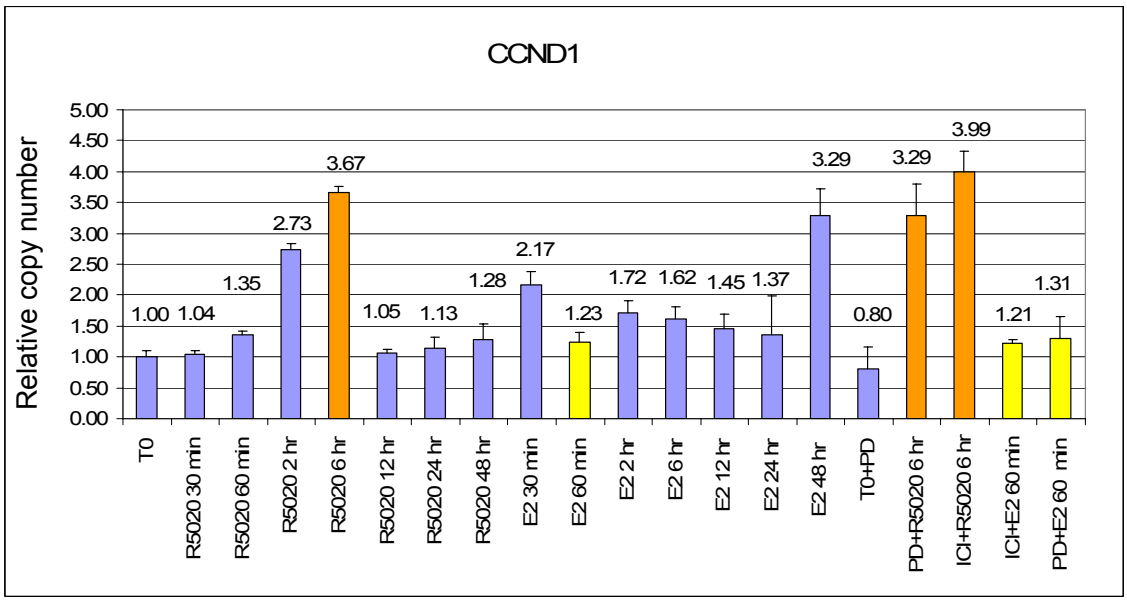


Figure 26: As in **Figure 23** but for gene CCND1 (Cyclin D1).

CCND1 (Cyclin D1) is consistently induced by progestin at 6 hr, and no inhibitor is able to affect this expression level (**Figure 26**). This fact is in agreement with microarray results. Induction by estradiol is below our microarray detection threshold, which we set at 1.3 fold-change ratio. Only at 48 hr, CCND1 has similar expression levels after estradiol hormonal treatment as it happens with progestin at 6 hr, therefore promoting cell cycle progression at earlier times in our model cell line.

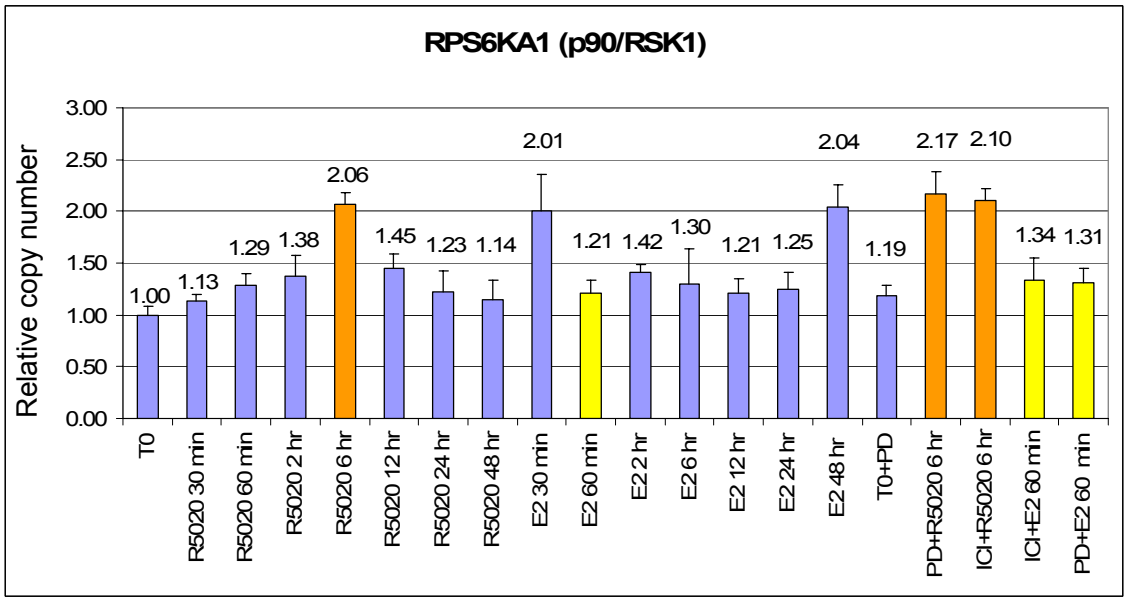


Figure 27: As in **Figure 23** but for gene RPS6KA1 (p90/RSK1).

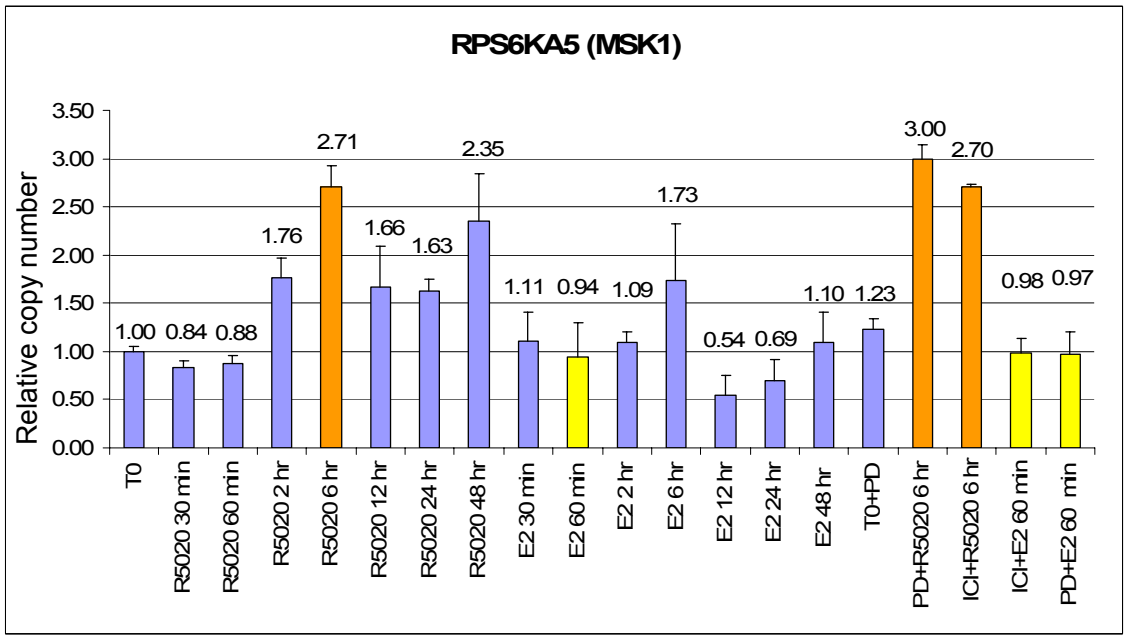


Figure 28: As in **Figure 23** but for gene RPS6KA5 (MSK1).

Neither PRS6KA1 (p90/RSK1, **Figure 27**) nor RPS6KA5 (MSK1, **Figure 28**) gene expression responses to hormones are affected by PD or ICI inhibitors, this fact agrees with microarray results. Hormone induction of both genes is independent of the ER-PR crosstalk, and independent of Erk signaling pathway, this probably means that these genes are target genes due to direct genomic signaling or other still unknown mechanisms.

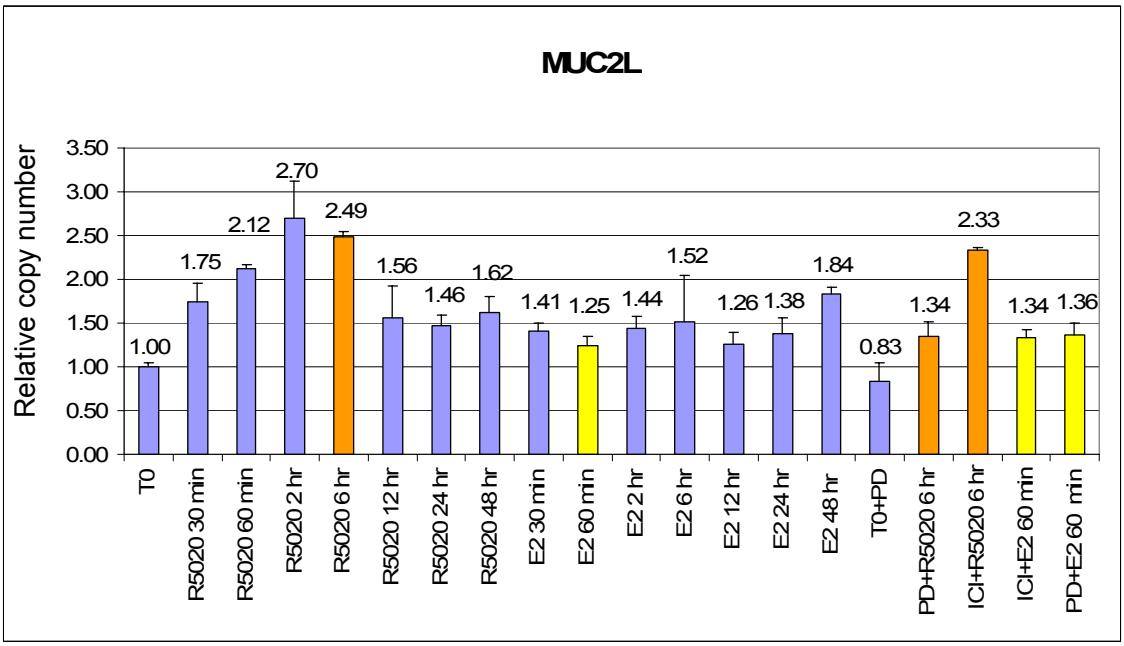


Figure 29: As in **Figure 23** but for gene MUC2L (TFCP2L3).

MUC2L (TFCP2L3, transcription factor CP2-like 3, **Figure 29**) upregulated expression in response to R5020 at 6 hr is only inhibited by PD, as it is shown from microarrays results. For that reason MUC2L is independent of ER and activated by progestins via the Erk signaling pathway.

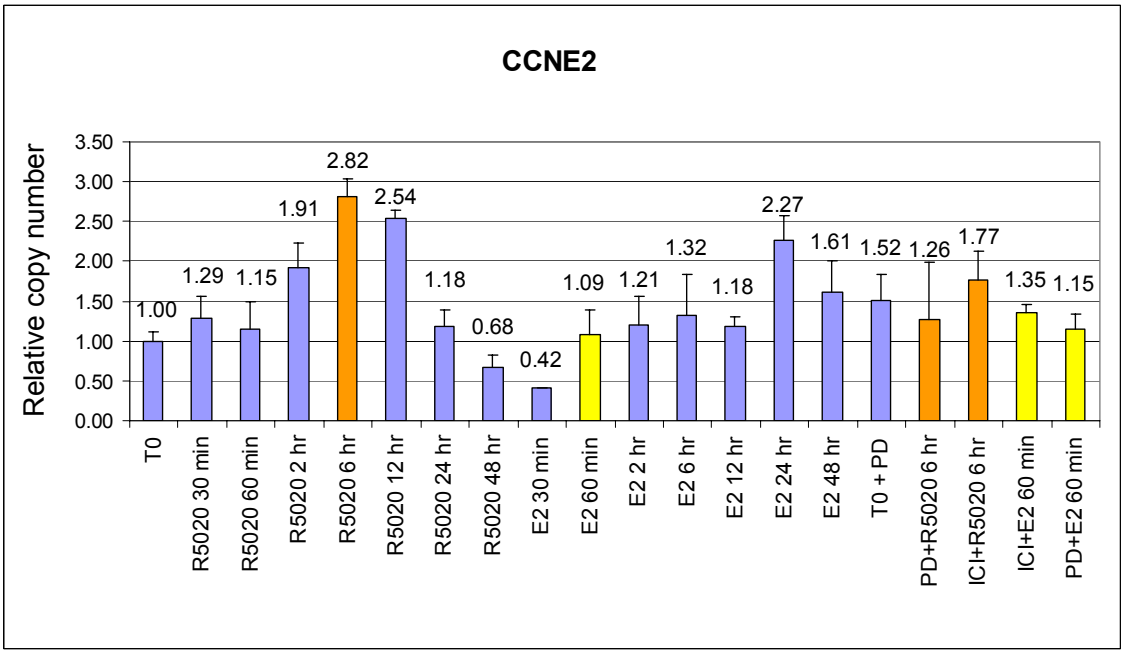


Figure 30: As in **Figure 23** but for gene CCNE2 (Cyclin E2)

CCNE2 (Cyclin E2, **Figure 30**) is strongly induced by R5020 at 6 hr. Real Time qPCR results agree with microarray measurements, though it is also observed that CCNE2 might be also affected by ICI, which means that CCNE2 is dependent of ER and the Erk signaling pathway. Induction of CCNE2 by estradiol is below the detection threshold at 60 min, and it has expression levels similar to progestin only 24 hr after estradiol hormonal induction. This is an indication of the delayed cell cycle progression in response to estradiol treatment in comparison with progestin treatment in our model cell line.

See **Appendix A11** for a comparison between Real Time qPCR values and Log_2Ratio values obtained by microarray analysis.

4.7 Pathway analysis of the time course experiment

Global functional pathway analysis was performed using a gene ontology built from experimental evidence compiled in the Ingenuity Pathways Knowledge Base (see chapter 3.19.4).

We focused on the ERK/MAPK signaling pathway to observe the genes induced by R5020 and the effect of PD as an inhibitor of the Erk1/2 pathway or the effect of ICI as the antagonist of ER on the ERK/MAPK pathway component transcript levels.

Figure 31, shows the gene expression values (M values) of the induced genes 6 hr after progestin treatment. Genes are highlighted in red or green based on their gene expression values, as up-regulated or down-regulated respectively, interpreting them as long term changes as a possible modulation on signal transduction.

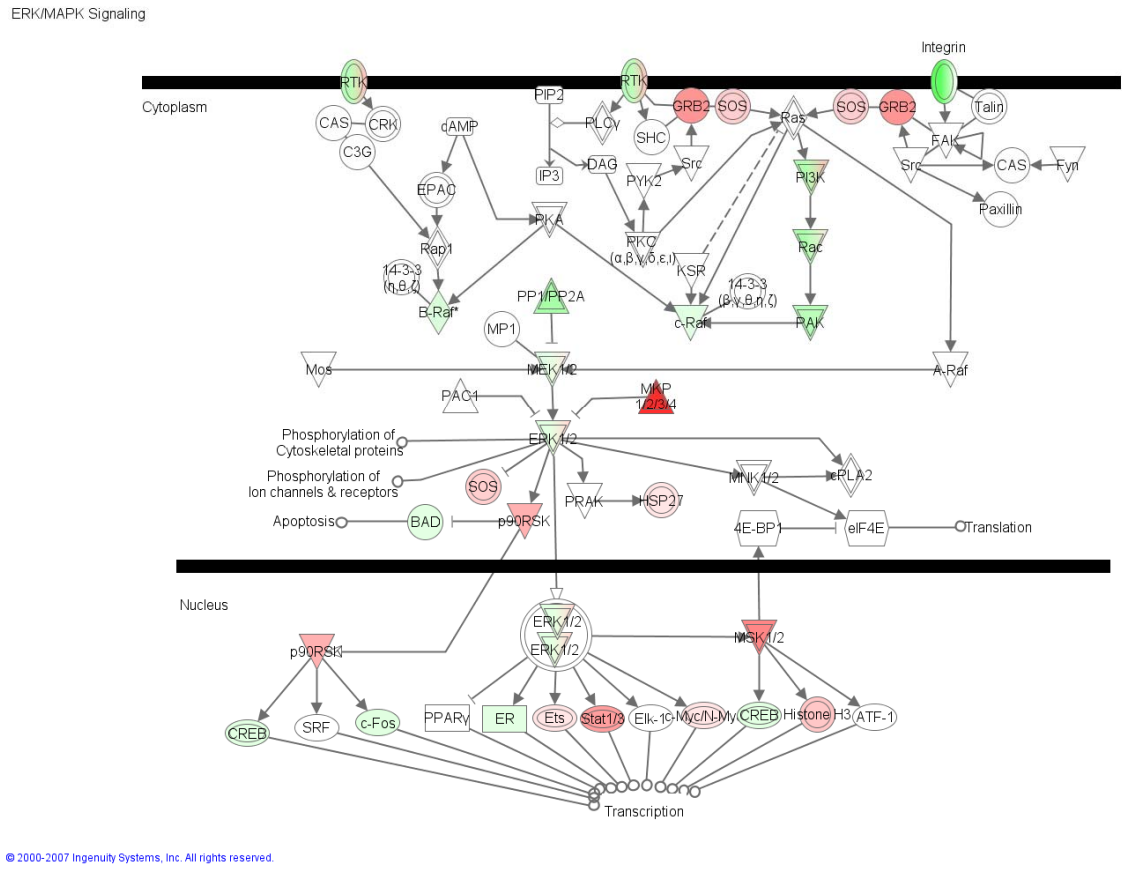


Figure 31: ERK/MAPK signaling pathway component transcript levels 6 hr after hormone induction with R5020

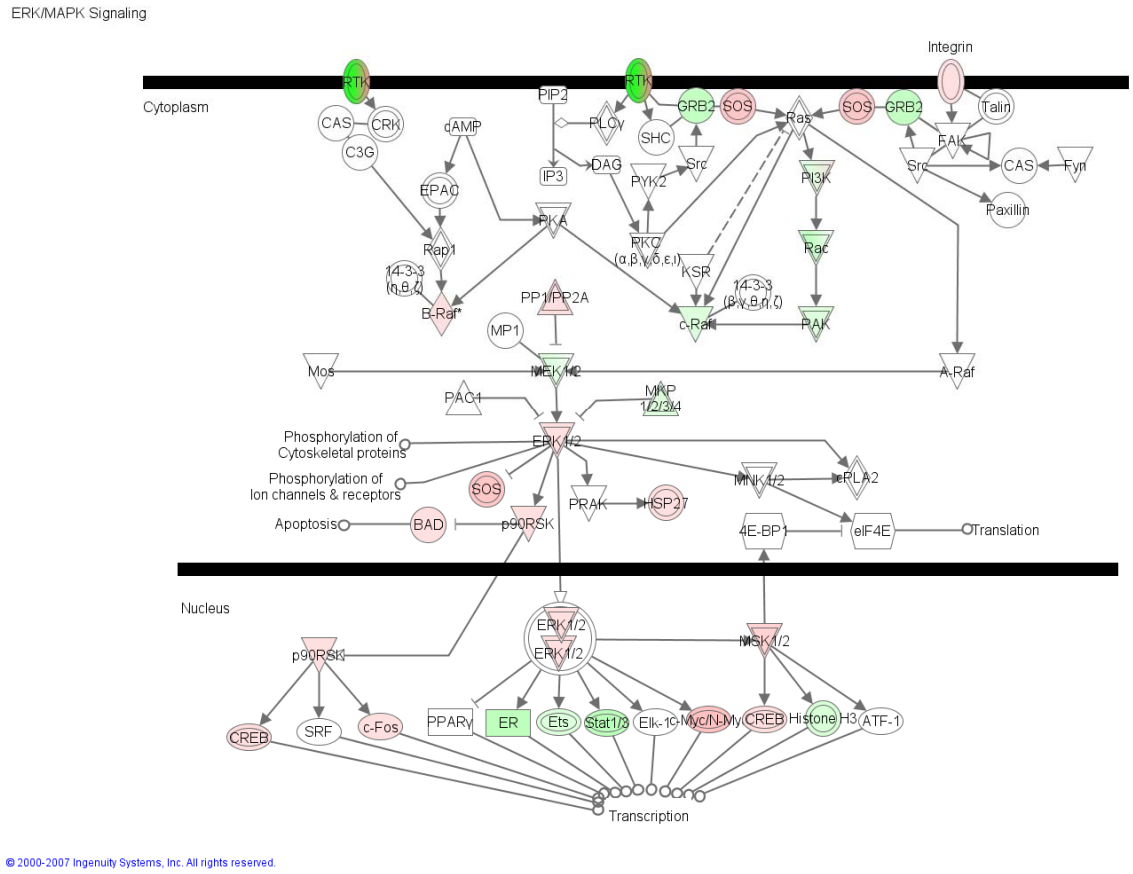


Figure 32: Effect of PD on ERK/MAPK signaling pathway component transcript levels 6 hr after hormone induction with R5020.

In order to observe the effect of PD, average M values from each of microarray experiments which were referred to average T0 were subtracted as follows:

$$\log_2(PDeffect) = \log_2 \frac{R5020 + PD}{EtOH} - \log_2 \frac{R5020}{EtOH}$$

These expression values were imported into Ingenuity Pathway analysis software. This effect is shown in **Figure 32**. Among the genes whose gene expression is reduced by the effect of PD are GRB2, and DUSP1 (MKP1) indicating that Erk signaling is not active. Gene expression of MSK1 and RSK1 are also found under expressed. Downstream Histone H3, STAT3, ER and ETS transcription regulators are down-regulated, and MYC is under-expressed.

To observe the effect of ICI, we proceeded in a similar way:

$$\log_2(ICIeffect) = \log_2 \frac{R5020 + ICI}{EtOH} - \log_2 \frac{R5020}{EtOH}$$

Erk/MAPK signaling pathway after the effect of ICI shows a different situation (**Figure 33**) where SOS1 and GRB2 are down-regulated, DUSP1 (MKP1), and HSP9B (HSP27) are over-expressed, and downstream transcription regulators such as Histone H3, STAT3, MYC, and ETS1 (Ets protein) are down-regulated and ER is up-regulated.

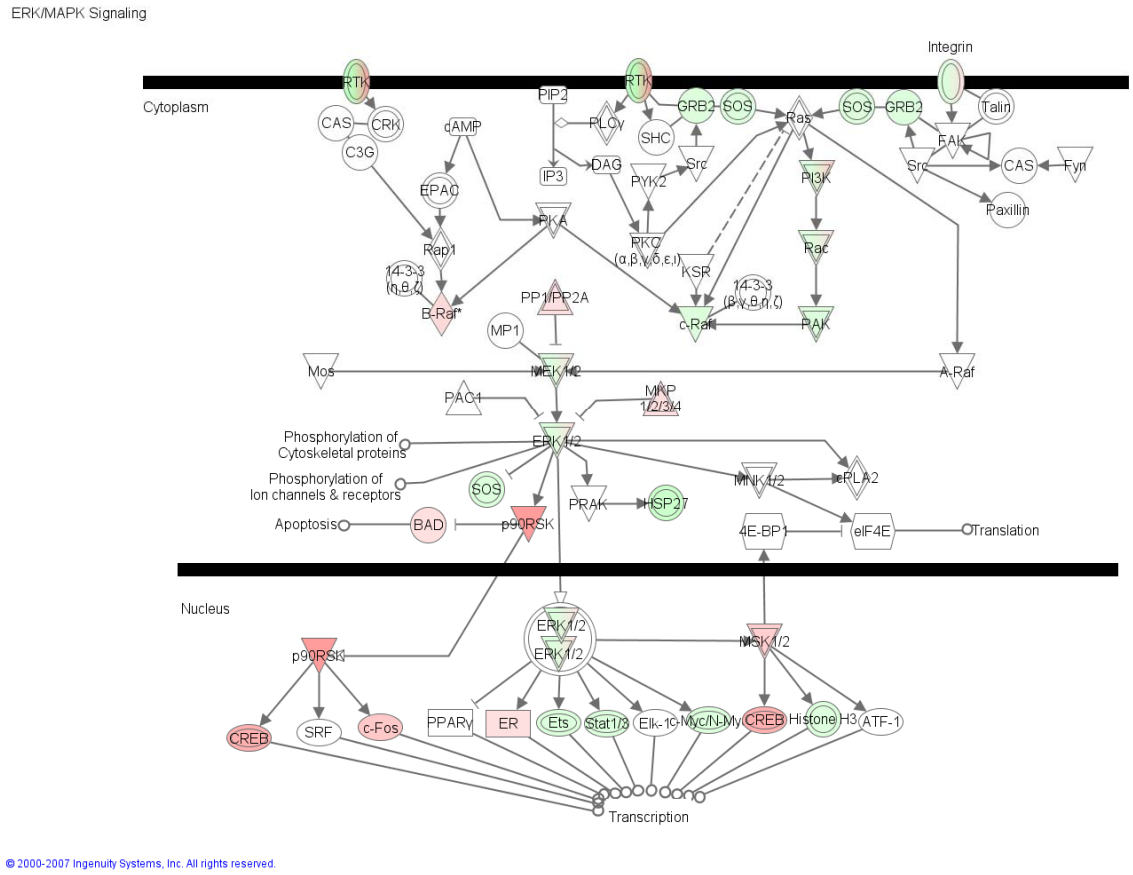


Figure 33: Effect of ICI on ERK/MAPK signaling pathway component transcript levels 6 hr after hormone induction with R5020.

Likewise, if we focused on the estradiol induced genes after 1 hr of treatment, in **Figure 34**, the gene expression values (M values) of the induced genes 1 hr after estradiol treatment are shown. Genes are also highlighted in red or green based on their gene expression values, as up-regulated or down-regulated respectively, in response to estradiol.

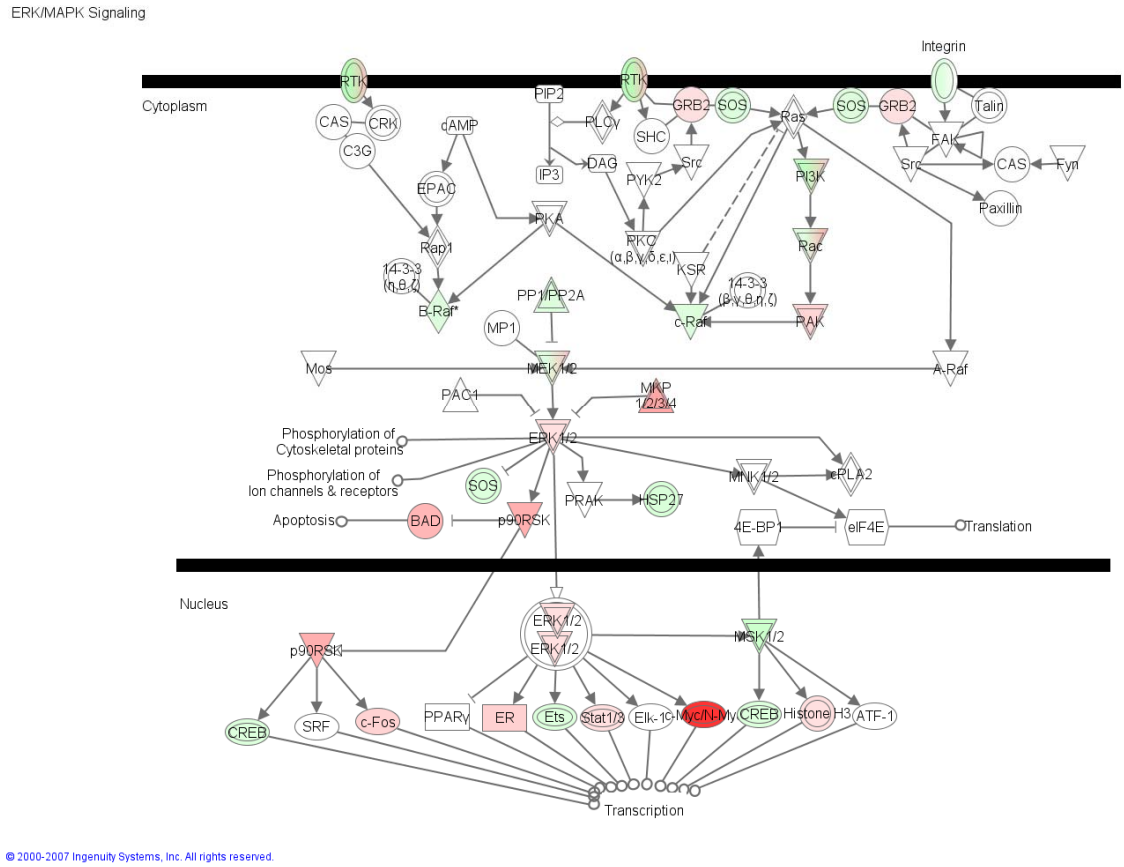


Figure 34: ERK/MAPK signaling pathway component transcript levels 1 hr after hormone induction with estradiol.

In order to observe the effect of PD after estradiol treatment, we proceed as follows:

$$\log_2(PDeffect) = \log_2 \frac{E2 + PD}{EtOH} - \log_2 \frac{E2}{EtOH}$$

In this case, the ERK/MAPK pathway is shown in **Figure 35**, where ERK1/2, p90RSK, PAK2 (p21 Activated Protein Kinase 2) transcripts are down-regulated, and SOS1 GRB2, DUSP1 (MKP1), HSP27 are being over-expressed. And downstream, transcription regulators such as FOS, ER, STAT3 are under-expressed, and MYC and Histone H3 are also affected by this treatment over-expressing them.

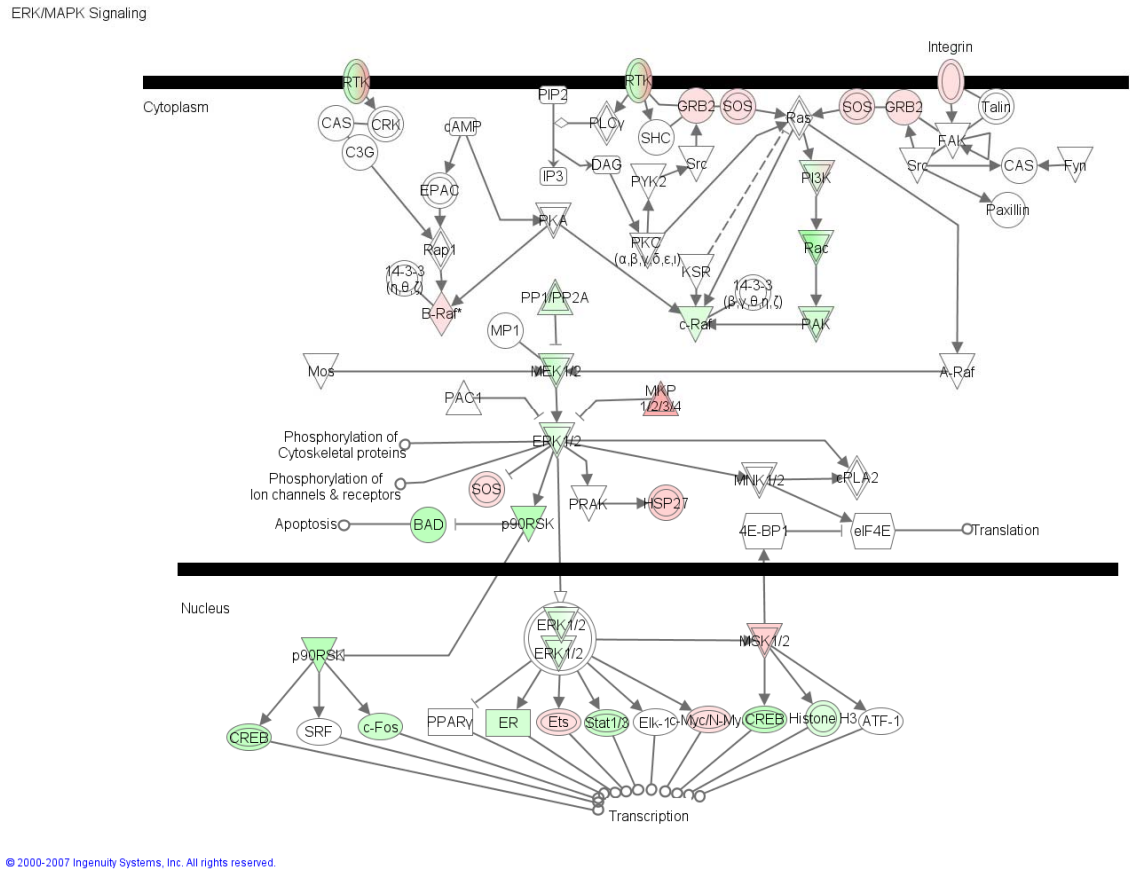


Figure 35: Effect of PD on ERK/MAPK signaling pathway component transcript levels 1 hr after hormone induction with estradiol.

Moreover, the effect of ICI 1hr after hormone induction with estradiol was calculated similarly as:

$$\log_2(ICIeffect) = \log_2 \frac{E2 + ICI}{EtOH} - \log_2 \frac{E2}{EtOH}$$

At this case, ERK/MAPK pathway looks as in **Figure 36** where this effect is more profound and SOS1, GRB2, PAK2, ERK1/2, DUSP1, and p90RSK, are under-expressed.

FOS and MYC are strongly under-expressed, as well as transcription regulators such as ER, ETS1, and STAT3.

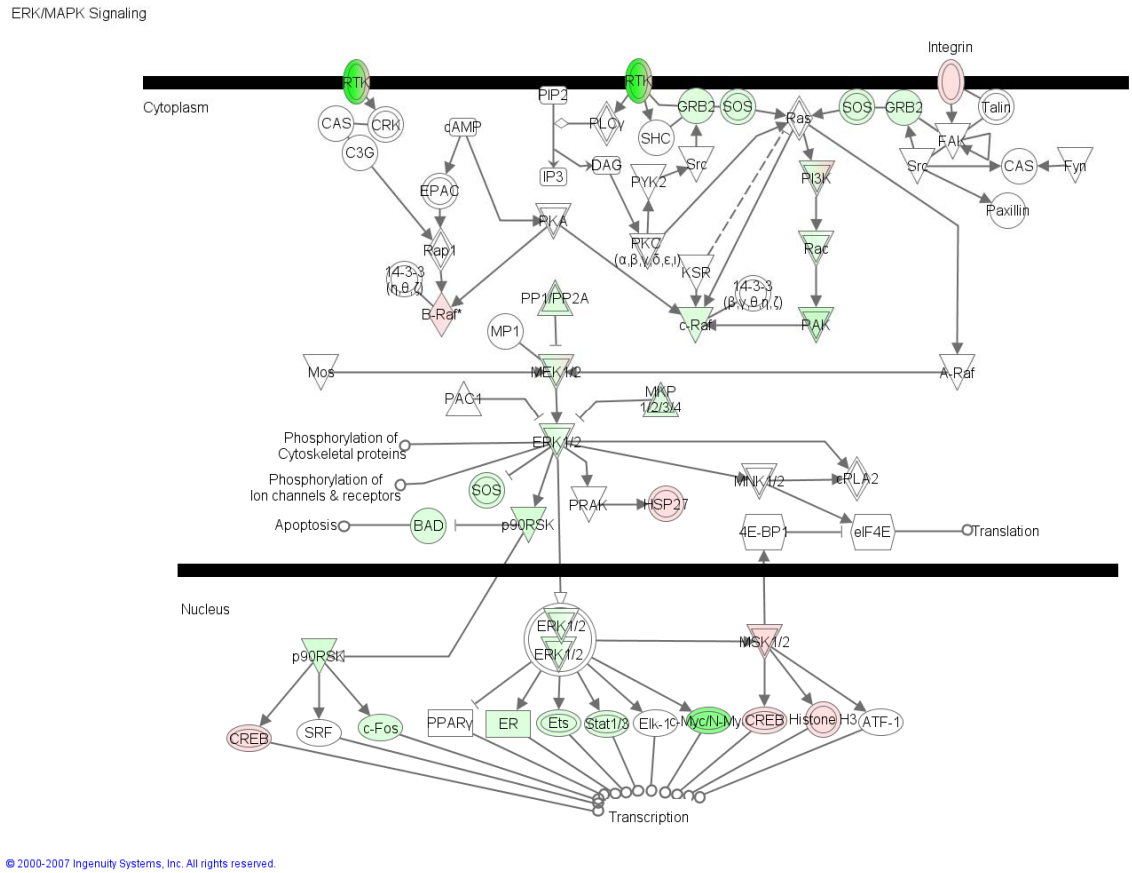


Figure 36: Effect of ICI on ERK/MAPK signaling pathway component transcript levels 1 hr after hormone induction with estradiol.

4.8 Application to the breast tumor classification

A set of 108 breast tumor tissue samples and 3 from normal breast tissue adjacent to tumors of different individuals were amplified, labeled and hybridized, as described in the methods section (chapter 3.7) to our cDNA platform, version 4. UHRR was used as reference RNA. A first batch of 77 samples, including normal breast samples was hybridized on a first batch of arrays. The second batch of 34 tumor samples was hybridized a year later with a new array printing batch. Therefore, in order to avoid any additional sources of variability due to the microarray batch, the first set of 77 samples, which have well characterized clinical-histopathological parameters, was analyzed separately and used as our training set in class prediction. The new breast sample batch was treated as a test set, and its characteristics assumed to be unknown were corroborated a posteriori. Clinical and histopathological features of the training and the test set are shown in **Table 5**.

The objectives of the experiment were (1) to classify tumor samples of our population in the various gene expression phenotypes, (2) to identify distinctive genes that can distinguish breast cancer subtypes, (3) to relate these phenotypes to specific cell-signaling pathways, and (4) identify molecular biomarkers, within our small gene collection, for breast cancer tumor progression.

The molecular biomarkers could be used in the future in clinical diagnosis, to improve the choice of treatment, to predict prognosis and identify patients at higher risk of developing metastasis, as well as for following the response to therapy.

The comparison of tumor data with the data obtained from our hormonal dependent breast cancer cell line model treated with progestin and estradiol in a time series, as well as the effect of the specific inhibitor and antagonist of the genomic and non-genomic ligand-activated pathways through which progestins act, might provide additional information about the most represented pathways influencing how hormone-dependent tumors develop.

Table 5: Clinical and histopathological characteristics of the training and the test set of breast tumor samples.

Clinical and histopathological characteristics of the patients and their breast tumors

All patients (n = 105)		Training set (n=74)	Test set (n=31)
Age (years)	Number of cases (%)	Number of cases (%)	Number of cases (%)
≤ 40	5 (4.8%)	3 (4.1%)	2 (6.5%)
> 40 and ≤ 50	11 (10.5%)	7 (9.5%)	4 (12.9%)
> 50 and ≤ 60	30 (28.6%)	22 (29.7%)	8 (25.8%)
> 60 and ≤ 70	27 (25.7%)	17 (23%)	10 (32.3%)
> 70	32 (30.5%)	23 (31.1%)	9 (29.0%)
Therapy	Number of cases (%)	Number of cases (%)	Number of cases (%)
HT only	21 (20%)	19 (25.7%)	3 (9.7%)
QT only	26 (24.8%)	24 (32.4%)	4 (9.7%)
HT + QT	38 (36.2%)	26 (35.1%)	19 (61.3%)
neoadjuvant QT	3 (2.9%)	1 (1.4%)	2 (6.5%)
HT + QT + Herceptin	1 (1%)		3 (9.7%)
none	5 (4.8%)	4 (5.4%)	1 (3.2%)
Tumor size, cm	Number of cases (%)	Number of cases (%)	Number of cases (%)
1	51 (48.6%)	32 (43.2%)	19 (61.3%)
2	42 (40%)	32 (43.2%)	10 (32.3%)
3	9 (8.6%)	7 (9.5%)	2 (6.5%)
4	2 (1.9%)	2 (2.7%)	
isquemic	1 (1%)	1 (1.4%)	
Lymph node	Number of cases (%)	Number of cases (%)	Number of cases (%)
none	60 (57.1%)	46 (62.2%)	15 (48.4%)
1	29 (27.6%)	16 (21.6%)	13 (41.9%)
2	8 (7.6%)	6 (8.1%)	2 (6.5%)
3	6 (5.7%)	5 (6.8%)	1 (3.2%)
micro	1 (1%)	1 (1.4%)	
Metastatic sites	Number of cases (%)	Number of cases (%)	Number of cases (%)
none	101 (96.2%)	70 (94.6%)	31 (100%)
one	4 (3.8%)	4 (5.4%)	
Vascular invasion	Number of cases (%)	Number of cases (%)	Number of cases (%)
yes	30 (28.6%)	18 (24.3%)	12 (38.7%)
none	74 (70.5%)	55 (74.3%)	17 (54.8%)
Histological grade	Number of cases (%)	Number of cases (%)	Number of cases (%)
1	17 (16.2%)	12 (16.2%)	5 (16.1%)
2	30 (28.6%)	21 (28.4%)	9 (29.0%)
3	33 (31.4%)	23 (31.1%)	10 (32.3%)
unknown	25 (23.8%)	18 (24.3&)	7 (22.6%)
Recurrence	Number of cases (%)	Number of cases (%)	Number of cases (%)
free of disease (VL)	58 (55.2%)	57 (77%)	
with disease (VE)	5 (4.8%)	3 (4.1%)	
exitus	10 (9.5%)	10 (13.5%)	
unknown	31 (29.5%)		31 (100%)
no follow-up	2 (1.9%)	2 (2.7%)	
p53 status	Number of cases (%)	Number of cases (%)	Number of cases (%)
negative	70 (66.7%)	47 (63.5%)	23 (74.2%)
positive	34 (32.4%)	27 (36.5%)	7 (22.6%)
unknown	1 (1%)		1 (3.2%)
Steroid receptor status	Number of cases (%)	Number of cases (%)	Number of cases (%)
ER+ PR+	56 (53.3%)	36 (48.6%)	20 (64.5%)
ER+ PR-	15 (14.3%)	11 (14.9%)	4 (12.9%)
ER- PR+	2 (1.9%)	2 (2.7%)	
ER- PR-	32 (30.5%)	25 (33.8%)	7 (22.6%)
her-2 status (IHC/FISH)	Number of cases (%)	Number of cases (%)	Number of cases (%)
negative/negative	90 (85.7%)	67 (90.5%)	23 (74.2%)
positive/amplified polysome	10 (9.5%)	5 (6.8%)	5 (16.1%)
negative/amplified polysome	4 (3.8%)	1 (1.4%)	3 (9.7%)
positive/non-amplified polysome	1 (1%)	1 (1.4%)	

HT = hormone therapy; QT = quimiotherapy: IHC = immunohistochemistry; FISH =fluorescence "*in situ*" hybridization

4.9 Classification of breast tumor samples by unsupervised hierarchical clustering

FADA was applied as described in methods (chapter 3.16) using complete linkage as the hierarchical clustering algorithm. An unsupervised cluster dendrogram was obtained where breast tumor samples were distributed based on similarity of gene expression profiles (**Figure 37**).

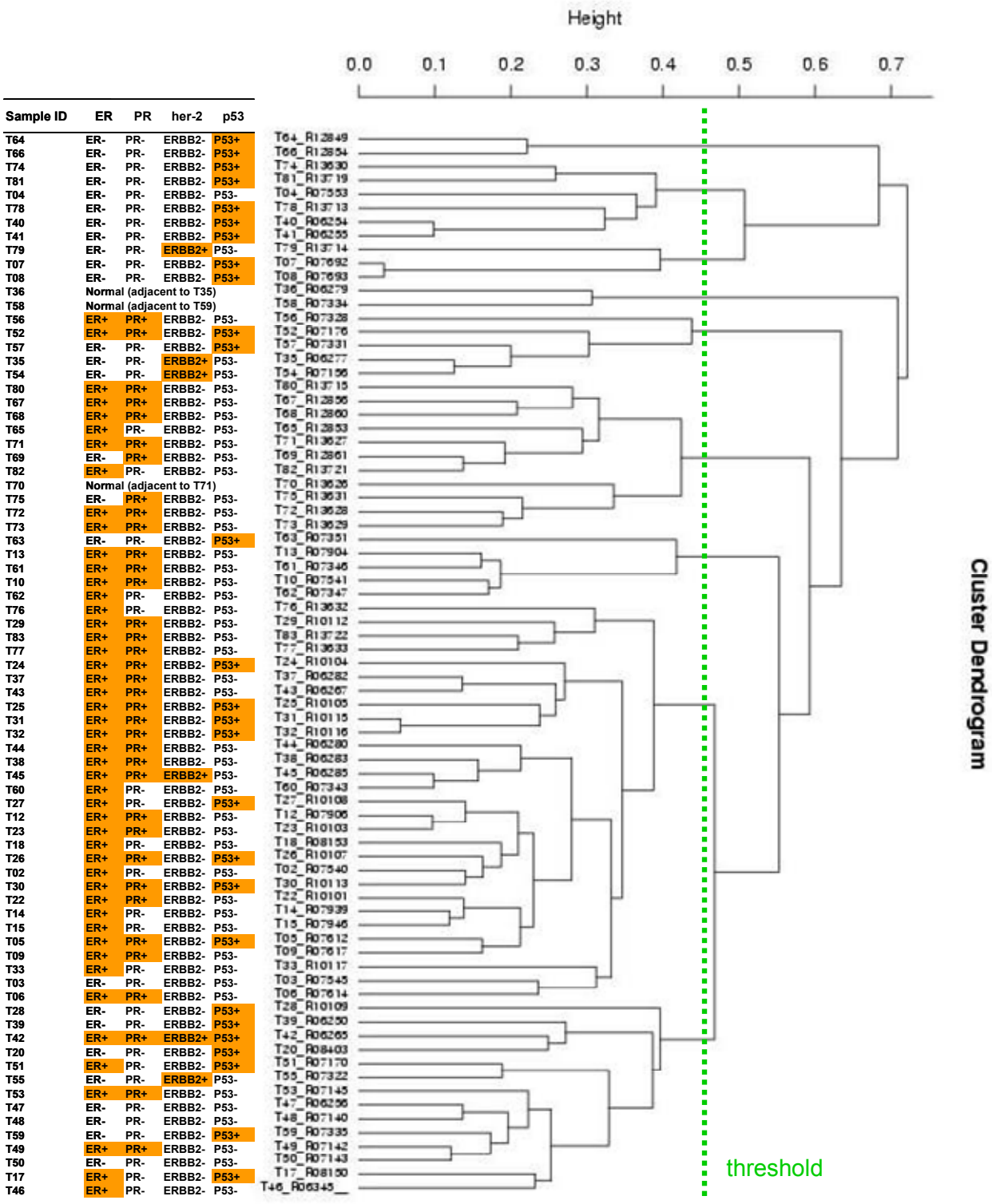


Figure 37: Unsupervised hierarchical clustering of breast samples with complete linkage.

Primarily, samples were distributed into two main branches of the tree dendrogram. These two main clusters gave two distinctive profiles with opposite differentially expressed genes (**Figure 38**). The first profile (**Figure 38 A**) brought together 11 tumor samples with significantly differential genes such as

TP53BP2 (tumor protein p53 binding protein 2), S100A2 (s100 calcium binding protein 2) , HRASLS (HRAS-like suppressor), KRT5 (cytokeratins 5), NF1B (transcription factor 1B), GPR180 (G protein-coupled receptor 180) and RUNX2 (Runt-related transcription factor 2) characteristic upregulated markers for the basal myoephitelial subtype of breast cancer, often p53+, and ER- and ERBB2- breast tumors (Rakha *et al.* 2007). Cytokeratin markers have been also shown to correlate with poor prognosis in breast cancer (Sorlie *et al*. 2003). Runt-related transcription factor RUNX2 has been associated as a breast cancer marker for developing bone metastasis (Barnes *et al.* 2005). The second profile (**Figure 38 B**) includes genes such as DDB2 (Damage-specific DNA binding protein 2), GATA3 (GATA binding protein 3), TFF1 (trefoil factor 1), CCNH (cyclin H), HSD17B4 (hydroxysteroid 17 β dehydrogenase 4, TNFRSF4 (tumor necrosis factor receptor superfamily, member 4), ITGB5 (Integrin β 5), AR (androgen receptor), and IL6ST (interleukin 6 signal transducer). Some of these genes have been reported as upregulated markers for ER+ tumors (West *et al*. 2001).

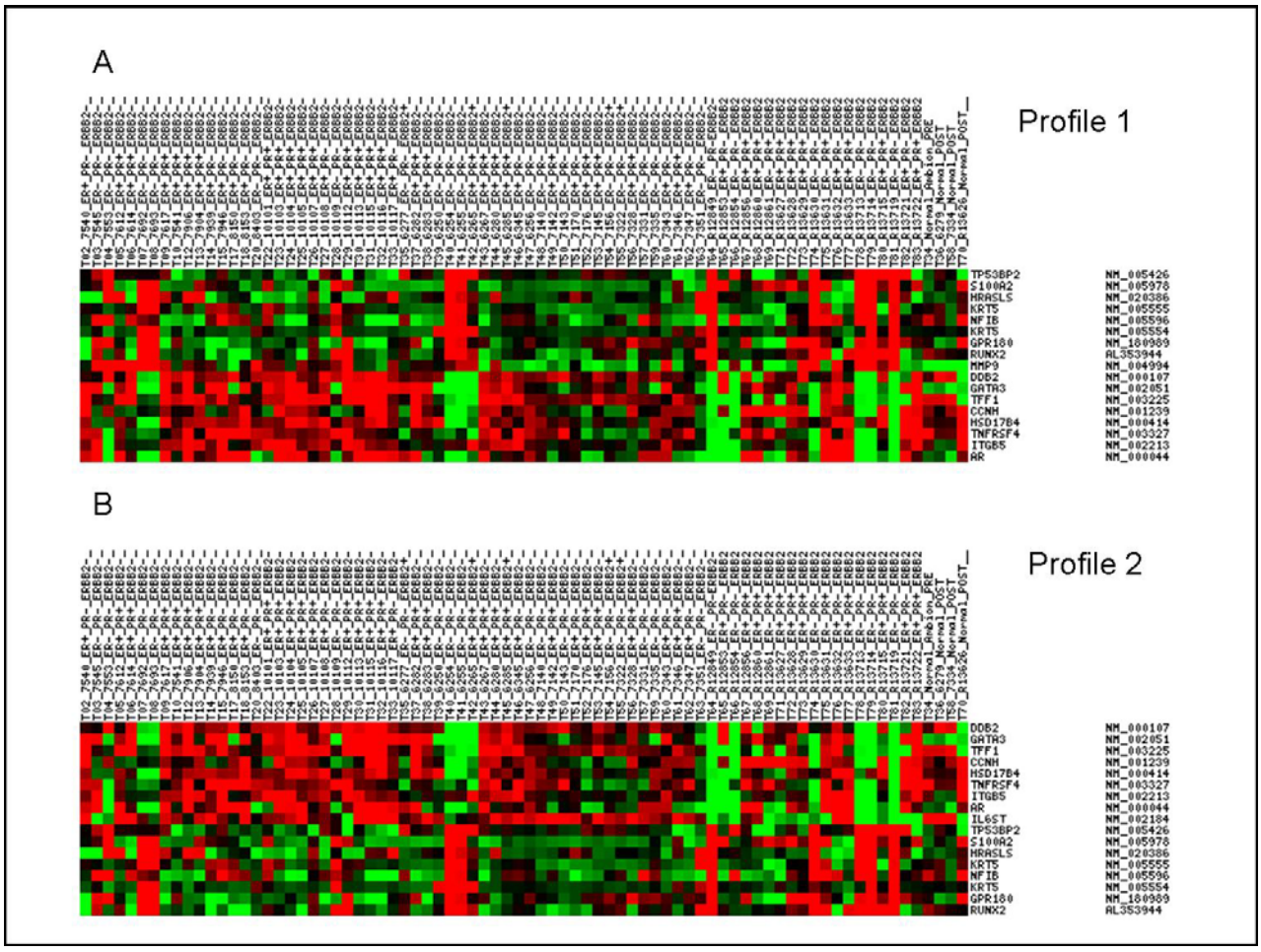


Figure 38: Expression levels of the most relevant genes selected from the two main clusters.

To reveal new distinctive tumor phenotypes within these two main branches, a threshold value (green dash line in **Figure 37**) was chosen. This threshold gives us 9 groups of samples, therefore performing a supervised analysis, with a

minimum number of 2 sample tumors per group. The most relevant gene list, up and downregulated, of each cluster was obtained with a q-value of 0.05 of being differentially expressed relative to the other cluster. The most significant gene expression profiles are represented in **Panel 39**.

Next, functional analysis was performed with the most differentially expressed genes of each group in both directions, up or downregulated, in search of enriched functional categories among GO terms in biological process (GO bp) or molecular function (GO mf). The significant hits are summarized in the following **Table 6**. Most of the groups have no relevant enriched GO term with a Fisher's exact probability less than 0.01, and a Benjamini and Hochberg (1995) adjusted p-value lower than 0.05. Background gene list used was the 820 genes from BCA v4.0.

Table 6: Significant (BH p-value<0.05) GO terms of the first three groups of gene expression clusters. ("GO mf" means GO molecular function, "GO bp" means GO biological process).

Group	Enrichment categories	Functional annotations categories	Genes	Fisher's exact p-values
1 (Down)	GO bp	Cell adhesion	ITGA5, PTEN, ITGB5, FN1	1.1E-3
	GO bp	Regulation of cell cycle dependent kinase activity	PTEN, CDKN1A, GTF2H1	2.2E-3
	GO bp	Integrin-mediated signaling pathway	ITGB5, ITGA5	1.0E-2
2 (Down)	GO bp	Cell motility	S100A2, ITGA6, ITGB2, PLAUR	5.9E-5
	GO bp	Cell migration	S100A2, ITGA6, ITGB2	1.1E-4
	GO bp	Integrin-mediated signaling pathway	ITGA6, ITGB2	1.0E-2
	GO mf	Isomerase activity	HSD17B4, PECI, TOP1	9.4E-5
	GO mf	Sterol transporter activity	HSD17B4	1.0E-2
3 (Up)	GO bp	Ectoderm development	KRT6A, KRT6B, KRT14, KRT17, KRT5	0.0
	GO bp	Tissue development	KRT6A, KRT6B, KRT14, KRT17, KRT5, RUNX2	2.0E-6
	GO mf	Structural constituent of cytoskeleton	KRT14, KRT17, KRT5, KRT6B	1.0E-6
	GO mf	Structural molecule activity	KRT6A, KRT6B, KRT14, KRT17, KRT5	1.3E-5

The profile for the groups 1, 2 and 3 clusters 11 tumor samples, 9 of which (82%) are positive for p53 and are negative for ER, PR and ERBB2 receptors. All these tumors present histological grade 3, or have developed metastatic sites. Relapse occurred in 6 cases resulting in death shortly after. These tumors would appear to match a particularly well defined class of tumors, usually called basal-cell or myoepithelial-like since they show high expression of cytokeratins 5/6, 14 and 17, are very aggressive (Nielsen *et al.* 2004). Basal-like tumor samples are included in the so-called triple-negative tumor class as defined by the absent expression of these three receptors. But not all triple-negative are basal-like tumors (for a review see Cleator *et al.* 2007). This tumor subtype do not respond to targeted therapies such as hormonal or trastuzumab treatment. This subtype of tumor (15% of all breast tumor samples) is supposed to arise from the outer (basal) layer of breast duct (myoepithelial cells). Basal-like breast

cancers usually show high p53 protein expression due to p53 mutations. Protein p53 acts as a checkpoint in the cell-cycle to trigger molecular response to DNA damage, including repair and apoptosis. A proportion (57%) of basal-like tumors expresses EGFR (epidermal growth factor receptor), c-KIT (29%) and Cyclin E (80%). Sporadic triple-negative breast cancers share similar features with BRCA1-related cancer, including ER negativity, high nuclear grade, high Ki-67 staining, and high CK5/6 and EGFR expression. Results from a report indicate that BRCA1 mRNA expression was significantly 2-fold lower in basal-like sporadic breast cancer. Additionally, ID4, a negative regulator of BRCA1, is also expressed at higher levels in basal-like breast cancers (Turner *et al.* 2006)

Group 4 clusters two samples from normal tissue adjacent to tumors.

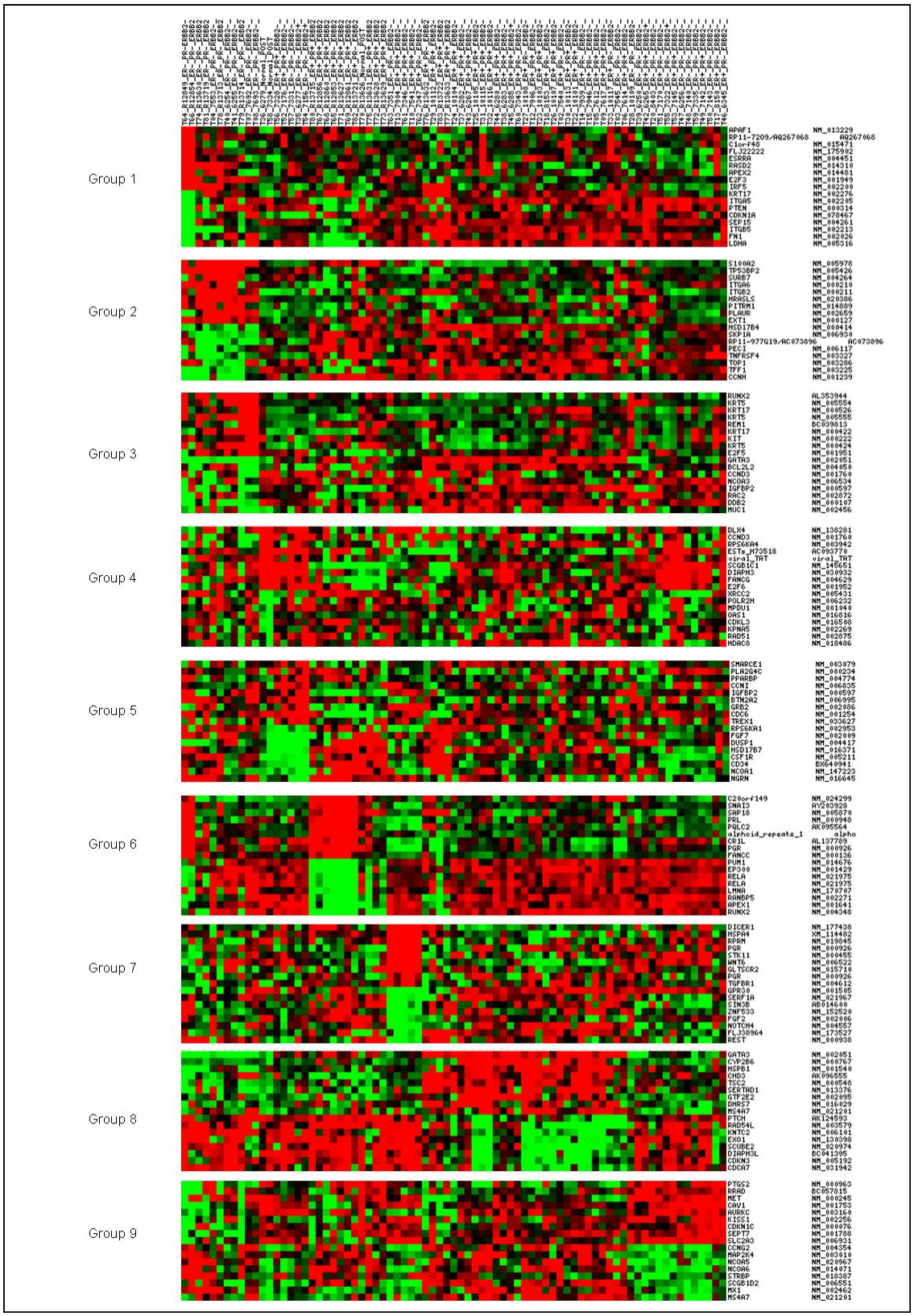
Group 5 gathers 5 samples of the ERBB2+ subtype. They present also high histological grade 2/3. Positive staining for her-2 protein was only seen in 2 cases out of 5. Gene expression profiling of this group shows high expression for the genes ERBB2, PPARBP, SMARCE1, and GRB2 among others, which has been already shown by Sorlie *et al.* 2003.

Group 6 clusters 11 samples. One of them, T70 (R13626) being a normal tissue sample adjacent to tumor sample T71 (R13628), it is assembled into the same cluster as its paired sample. We thought that it could still contain tumoral cells, but it was confirmed by pathologists that it is indeed a normal sample. Most interesting in this cluster is that 60% of these samples are positive for both estrogen and progesterone receptor proteins, and the other 4 samples are positive for either progesterone or the estrogen receptor. The gene expression profile shows high expression of PR, and some known genes found to be regulated by progestins in breast cancer cell line experiments, such as SNAI3, SAP18, PQLC2, CR1L, PRL, FANCC, and C20orf149, indicating that these samples have PR activity. These tumor samples are also all ERBB2 and p53 negative.

Group 7 assembles 5 samples, which are positive for both estrogen and progesterone receptor proteins, and negative for ERBB2 (her-2) and p53 proteins, except in one case that was found by clinicians to be p53 positive. Among the most significant genes are PR, WNT6, STK11, TGFBR1, RPRM, HSPA4, DICER1, and GLTSCR2. We hypothesize that this group of samples are hormone dependent tumors and the tumor progression is due to signaling pathways different to those of group 6.

Group 8 clusters 29 samples: 76% are ER+PR+, 20% are ER+, and only one sample (T03-R7545) is negative for both hormone receptors and negative for ERBB2 (her-2) and p53 protein. We hypothesize that this tumor sample has intact ER signaling pathway, and it is a hormonal-dependent tumor. The gene expression profile shows activated ER signaling pathways showing high levels of GATA3, CYP2B6, HSPB1, CHD3, TSC2, SERTAD1, GTF2E2, DHRS7, and MS4A7. Some of these genes have been previously reported as being regulated by ER or have been found to be co-expressed with ER.

Group 9 collects 14 samples. These breast tumor samples are more heterogeneous in their immunohistochemical clinical data: only 6 of them are either positive for one or both hormone receptors, half of them are triple-negative and p53+. Significant overexpressed genes for this cluster are PTGS2, RRAD, MET, CAV1, AURKC, KISS1, CDKN1C, SEPT7, and SLC2A3. Functional analysis gave no significant result.



Panel 39: Expression levels of the most relevant genes selected from the nine groups obtained by supervised analysis.

4.10 Principal component analysis (BGA)

Datasets were divided into training and test dataset. The training set consisted in 77 samples. At the time of the gene expression analysis by microarray technique, the clinical-histopatological characteristics of these samples (including status of the molecular markers such as ER, PR, p53, and her-2 proteins scored by immunohistochemistry) were known. On the contrary, at the time of the gene expression analysis, nothing was know about the newly hybridized batch of 34 breast tumor samples. Therefore we used this dataset as the test set, and wait for their clinical-histopatological parameters be determined. Initially a two class supervised analysis was performed. Depending on their estrogen receptor status, which was determined by IHC, samples were classified into ER positive (class 1) or ER negative (class 0). **Appendix A4** displays the clinical-histopatological patient data. Between groups analysis (BGA) was applied, and samples were distributed along one axis for visualization purposes (**Figure 40**). Since BGA used correspondence analysis (CA), the most dicriminant genes are also determined. It is observed that samples are grouped into three sets along the axis, leaving the middle group as a mixed class.

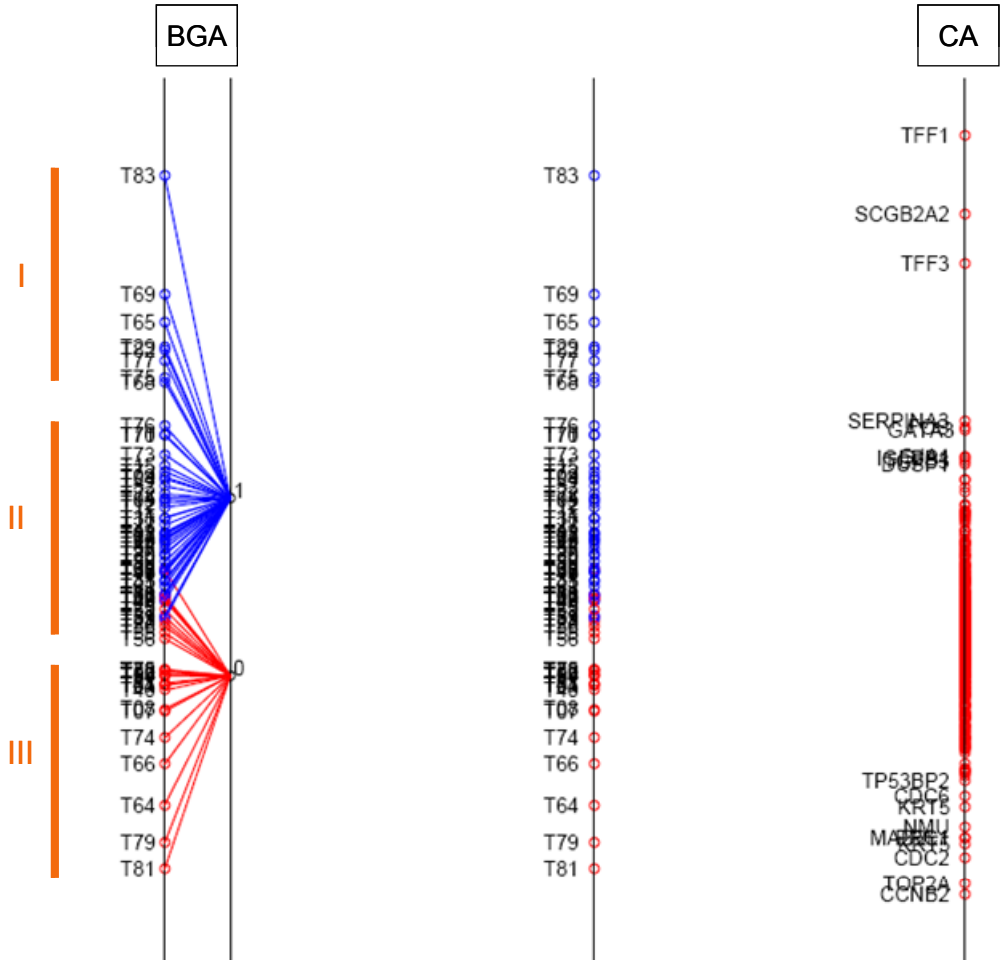


Figure 40: Supervised classification of breast samples by BGA into two groups 0 and 1 based on their ER status reveals an intermediate mixed class.

Table 7 summarizes the distribution of the samples along the axis and their clinical markers scored by immunohistochemistry. The column “IHC ER status” provides the assignment of class prior the BGA analysis, defined based on clinical status of the ER. The samples that are lying on the axis discordant with the clinical data are colored in red. Sample T70 (R13626) is a normal sample adjacent to the tumor sample T71 (R13627) and, instead of grouping with the ER- samples, close to the other normal breast samples, behaves as its paired tumor sample. This is indicative of its heterogeneity; it might contain tumoral cells. This fact was reported to the Pathology Department at the Hospital del Mar for clarification of this result and they confirm that it is indeed a normal sample.

Marked with orange color are the molecular markers ER, PR, ERBB2 (her-2) and p53 analyzed by IHC and found positive. It is observed that the classification by BGA performs well, with a few exceptions. In **Figure 40** assorted with the group 0 and labeled in red, there are several breast tumor samples that are being predicted as ER-, although this could be due to a different subtype of tumor, BGA considers that these samples are closer in distance to the ER- group as to the ER+ group. There are also a few ER+ tumor samples which are grouped with the ER- samples, such as T49 (R7142), T51 (R7170), T53 (R7145), T56 (R7328), and T52 (R7176). On the ER+ positive side, there are 2 samples ER- but PR+, which could be due to still active ER+ signaling pathway. Tumor sample T03 (R7545) is also allocated between ER+ samples, even though it has been classified by IHC as ER-. There is also a higher proportion of samples which are P53+ between the ER- samples, as it is expected from epidemiological studies.

A summary for the prediction of the ER status of the test set consisting of a new independent series of 34 breast samples is shown in **Table 8**.

Based on these results we decided to group the samples into six groups, and perform a multiple class supervised classification. The groups for the multiple class supervised clustering analysis were assigned by looking at the best fitting of the PCA model. Since this is a supervised clustering analysis, by “trial-and-error”, samples were exchanged from one group to another to get the best fitting picture of the PCA. The final class assignment is stated on the last column of **Table 7**. The best fitting picture of the PCA by multiple classes BGA is shown in **Figure 41**. BGA uses correspondence analysis (CA) to select the genes that discriminate the groups situated on the 3D space (**Figure 42**). The most distant and well-defined groups are 1 and 5. The other groups are also well-separated but the distances between them are smaller than from the group 1 to group 5. From this analysis it is corroborated that among ER+ hormone-dependent tumors there are different subtypes that can be allocated in 3D space into groups 1 and 2. Groups 2 and 3 are less well defined. The summary for the class prediction of the test set consisting of a new independent series of 34 breast samples is shown in **Table 9**.

From this analysis, we could not correlate the genes that differentiate the formed groups, but just the most significant on the four axes on the 3D space.

For this aim, Prediction Analysis of Microarrays (PAM) was employed in the next chapter.

Table 7: Distribution of tumor samples along one axis after BGA two class supervised classification upon the ER status. Samples are sorted according the BGA axis 1 value. Status of their immunohistochemical markers and predicted class into which the application would distribute the samples is stated.

Sample ID	Patient ID	IHC ER status	ER	PR	her-2	p53	BRCA1/2	Histological grade	Tumor size	Lymph node status	Metastatic sites	BGA - Two Class - Axis1	BGA predicted	multiclass
T83	R13722	1	ER+	PR+	ERBB2-	P53-		HG1	T2	N1	M0	756.2	1	1
T69	R12861	1	ER-	PR+	ERBB2-	P53-		HG2	T1	N0	M0	525.3	1	1
T65	R12853	1	ER+	PR-	ERBB2-	P53-		HG2	T2	N2	M0	469.9	1	1
T29	R10112	1	ER+	PR+	ERBB2-	P53-		HG0	T1	N0	M0	423.5	1	1
T82	R13721	1	ER+	PR-	ERBB2-	P53-		HG0	T2	N0	M0	414.9	1	1
T77	R13633	1	ER+	PR+	ERBB2-	P53-		HG1	T2	N0	M0	397.3	1	1
T75	R13631	1	ER-	PR+	ERBB2-	P53-		HG1	T1	N0	M0	362.5	1	1
T68	R12860	1	ER+	PR+	ERBB2-	P53-		HG2	T2	N3	M0	352.6	1	1
T76	R13632	1	ER+	PR-	ERBB2-	P53-		HG2	T2	N0	M0	268.2	1	2
T71	R13627	1	ER+	PR+	ERBB2-	P53-		HG2	T3	N1	M0	252.5	1	1
T70	R13626	0	Normal (adjacent to T71)									251.1	1	1
T73	R13629	1	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	212.3	1	2
T15	R7946	1	ER+	PR-	ERBB2-	P53-		HG0	T2	N1	M0	189.8	1	2
T72	R13628	1	ER+	PR+	ERBB2-	P53-		HG3	T2	N1	M0	179.6	1	2
T09	R7617	1	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	167.9	1	2
T24	R10104	1	ER+	PR+	ERBB2-	P53+		HG2	T1	N0	M0	163.8	1	2
T32	R10116	1	ER+	PR+	ERBB2-	P53+		HG2	T3	N0	M0	149.9	1	2
T22	R10101	1	ER+	PR+	ERBB2-	P53-		HG1	T2	N0	M0	134.0	1	2
T14	R7939	1	ER+	PR-	ERBB2-	P53-		HG0	T1	N0	M0	127.8	1	2
T05	R7612	1	ER+	PR+	ERBB2-	P53+		HG0	T1	N0	M0	124.3	1	2
T12	R7906	1	ER+	PR+	ERBB2-	P53-		HG0	T2	N1	M0	117.2	1	2
T17	R8150	1	ER+	PR-	ERBB2-	P53+		HG0	IS	N0	M0	108.9	1	2
T18	R8153	1	ER+	PR-	ERBB2-	P53-		HG0	T1	N0	M0	88.8	1	2
T31	R10115	1	ER+	PR+	ERBB2-	P53+		HG2	T3	N0	M0	88.3	1	2
T27	R10108	1	ER+	PR-	ERBB2-	P53+		HG2	T4	N1	M0	73.9	1	2
T42	R6265	1	ER+	PR+	ERBB2+	P53+		HG3	T1	N1	M0	59.9	1	2
T02	R7540	1	ER+	PR-	ERBB2-	P53-		HG2	T2	N3	M0	56.3	1	2
T23	R10103	1	ER+	PR+	ERBB2-	P53-		HG2	T2	N1	M0	55.5	1	2
T67	R12856	1	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	49.2	1	2
T26	R10107	1	ER+	PR+	ERBB2-	P53+		HG3	T1	N1	M0	46.6	1	2
T13	R7904	1	ER+	PR+	ERBB2-	P53-		HG0	T3	N0	M0	43.8	1	2
T44	R6280	1	ER+	PR+	ERBB2-	P53-		HG3	T2	N0	M0	36.3	1	2
T37	R6282	1	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	28.4	1	2
T80	R13715	1	ER+	PR+	ERBB2-	P53-		HG2	T2	N0	M0	17.0	1	2
T25	R10105	1	ER+	PR+	ERBB2-	P53+		HG2	T1	N0	M0	16.6	1	2
T30	R10113	1	ER+	PR+	ERBB2-	P53+		HG3	T1	N0	M0	2.1	1	2
T60	R7343	1	ER+	PR-	ERBB2-	P53-		HG2	T2	N0	M1	-9.1	0	3
T03	R7545	0	ER-	PR-	ERBB2-	P53-		HG2	T1	N2	M0	-12.4	0	3
T38	R6283	1	ER+	PR+	ERBB2-	P53-		HG0	T1	N0	M0	-14.6	0	3
T43	R6267	1	ER+	PR+	ERBB2-	P53-		HG3	T2	N1	M0	-16.8	0	3
T46	R6345	1	ER+	PR-	ERBB2-	P53-		HG1	T1	N0	M0	-19.4	0	3
T33	R10117	1	ER+	PR-	ERBB2-	P53-		HG1	T1	N0	M0	-33.3	0	3
T45	R6285	1	ER+	PR+	ERBB2+	P53-		HG0	T1	N0	M0	-36.0	0	3
T61	R7346	1	ER+	PR+	ERBB2-	P53-		HG2	T1	N0	M0	-44.7	0	3
T10	R7541	1	ER+	PR+	ERBB2-	P53-		HG0	T2	N2	M0	-59.1	0	3
T06	R7614	1	ER+	PR+	ERBB2-	P53-		HG2	T1	N0	M0	-62.6	0	3
T50	R7143	0	ER-	PR-	ERBB2-	P53-		HG3	T2	N3	M1	-63.2	0	3
T62	R7347	1	ER+	PR-	ERBB2-	P53-		HG2	T2	N2	M0	-64.5	0	3
T47	R6256	0	ER-	PR-	ERBB2-	P53-		HG3	T2	N2	M0	-68.0	0	3
T20	R8403	0	ER-	PR-	ERBB2-	P53+		HG3	T2	N3	M0	-70.4	0	3
T48	R7140	0	ER-	PR-	ERBB2-	P53-		HG3	T2	N2	M0	-71.5	0	3
T49	R7142	1	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	-73.4	0	3
T59	R7335	0	ER-	PR-	ERBB2-	P53+		HG1	T2	N0	M0	-89.9	0	3
T51	R7170	1	ER+	PR-	ERBB2-	P53+		HG2	T1	N0	M0	-103.0	0	3
T39	R6250	0	ER-	PR-	ERBB2-	P53+		HG3	T2	N0	M0	-104.9	0	3
T53	R7145	1	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	-109.4	0	3
T28	R10109	0	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M0	-115.1	0	3
T36	R6279	0	Normal (adjacent to T35)									-122.4	0	6
T55	R7322	0	ER-	PR-	ERBB2+	P53-		HG0	T3	N3	M0	-135.8	0	4
T56	R7328	0	ER+	PR+	ERBB2-	P53-		HG3	T4	N0	M0	-146.9	0	4
T78	R13713	0	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M0	-208.0	0	5
T35	R6277	0	ER-	PR-	ERBB2+	P53-		HG3	T1	N0	M0	-208.6	0	5
T52	R7176	0	ER+	PR+	ERBB2-	P53+		HG2	T1	N0	M0	-211.2	0	5
T58	R7334	0	Normal (adjacent to T59)									-214.8	0	6
T63	R7351	0	ER-	PR-	ERBB2-	P53+	BRCA2	HG2	T2	N0	M0	-217.8	0	4
T54	R7156	0	ER-	PR-	ERBB2+	P53-		HG3	T2	N0	M0	-219.8	0	4
T41	R6255	0	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M1	-234.8	0	5
T57	R7331	0	ER-	PR-	ERBB2-	P53+		HG3	T2	N0	M0	-238.3	0	4
T04	R7553	0	ER-	PR-	ERBB2-	P53-		HG3	T3	N1	M0	-238.9	0	5
T40	R6254	0	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M1	-248.7	0	5
T08	R7693	0	ER-	PR-	ERBB2-	P53+		HG0	T3	N1	M1	-285.7	0	5
T07	R7692	0	ER-	PR-	ERBB2-	P53+		HG0	T3	N1	M1	-290.6	0	5
T74	R13630	0	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M0	-341.4	0	5
T66	R12854	0	ER-	PR-	ERBB2-	P53+		HGx	T2	N1	M0	-392.0	0	5
T64	R12849	0	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M0	-471.3	0	5
T79	R13714	0	ER-	PR-	ERBB2+	P53-		HG3	T2	N0	M0	-545.6	0	5
T81	R13719	0	ER-	PR-	ERBB2-	P53+		HG3	T2	N0	M0	-598.0	0	5

Table 8: Distribution of tumor samples of the test set along one axis after two class supervised classification by BGA “two class “based on ER status determined by IHC. Samples are sorted on the table according the PCA axis 1 value.

Sample ID	BGA - Two Class - Axis1	closest.centre	BGA predicted
T88	226.0	1	1
T91	184.9	1	1
T95	170.0	1	1
T102	164.1	1	1
T99	160.2	1	1
T106	159.7	1	1
T101	158.4	1	1
T90	140.9	1	1
T115	120.7	1	1
T108	111.1	1	1
T89	110.2	1	1
T116	104.9	1	1
T100	104.5	1	1
T114	101.9	1	1
T98	97.3	1	1
T117	87.2	1	1
T105	84.5	1	1
T96	62.0	1	1
T104	55.4	1	1
T85	53.9	1	1
T111	51.9	1	1
T92	47.1	1	1
T84	45.4	1	1
T112	35.7	1	1
T109	32.1	1	1
T113	30.3	1	1
T97	25.3	1	1
T103	1.7	1	1
T93	-14.5	0	0
T94	-17.9	0	0
T87	-63.5	0	0
T86	-71.5	0	0
T107	-100.8	0	0
T110	-120.5	0	0

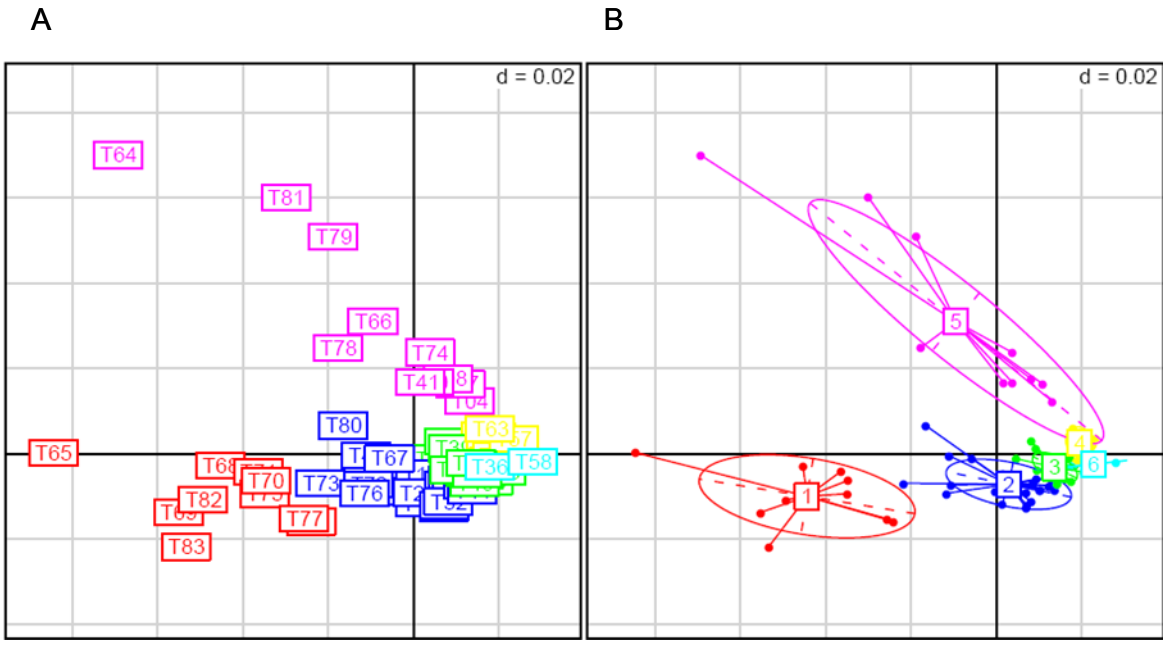


Figure 41: Supervised classification of breast samples by multiple class Between Groups Analysis (BGA) in six groups. (A) Allocation of breast samples. (B) Diagram of the spatial position of the six classes.

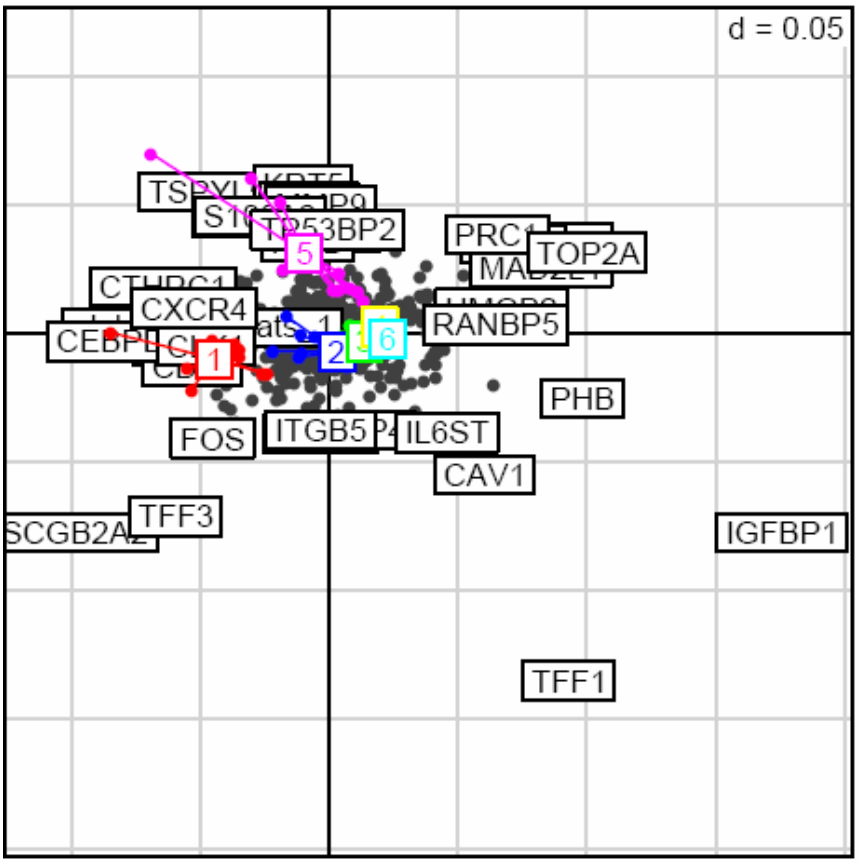


Figure 42: Most discriminant genes oriented on each axis after supervised classification of breast samples by multiple class BGA into six groups.

Table 9: Distribution of tumor samples of the test set as a result of multiple class supervised classification by BGA.

Sample ID	projected.Axis1	projected.Axis2	projected.Axis3	projected.Axis4	projected.Axis5	closest.centre	predicted
T84	64.5	-99.1	-23.2	-7.4	12.7	6	6
T85	94.6	-109.9	-3.3	30.9	53.3	6	6
T86	100.4	23.6	-28.3	-6.2	81.4	6	6
T87	61.4	38.4	3.5	-35.3	94.2	6	5
T88	-24.2	-227.3	105.0	18.8	-31.8	1	1
T89	49.3	-150.6	24.3	26.5	13.3	2	2
T90	-70.8	-114.1	84.7	-45.8	-5.8	1	1
T91	-168.0	-140.9	-11.3	-53.7	40.0	1	6
T92	-59.3	-52.7	-50.4	9.5	4.4	1	1
T93	48.2	-46.6	-112.4	-72.3	97.9	6	6
T94	-66.8	48.7	57.7	-109.2	21.4	5	5
T95	-121.6	-143.4	-6.7	-48.0	-36.6	1	1
T96	-9.0	-95.4	-70.8	-82.2	89.2	6	6
T97	49.1	-69.1	-46.0	52.3	63.0	6	6
T98	10.8	-122.9	3.9	-6.8	46.1	6	6
T99	-33.0	-146.4	117.9	-14.6	39.8	1	5
T100	25.3	-135.6	14.0	19.3	11.8	2	2
T101	-81.0	-130.1	85.3	-99.9	-27.3	1	1
T102	-54.5	-154.7	81.7	-10.7	-49.4	1	1
T103	54.9	-40.9	-13.4	-0.1	-8.3	6	3
T104	2.4	-63.3	62.9	-0.5	-64.2	2	3
T105	49.2	-118.0	36.1	-4.4	-23.7	2	3
T106	-14.8	-173.7	50.7	4.9	-32.1	1	1
T107	143.7	31.3	-17.4	4.8	8.1	6	4
T108	-3.8	-120.6	39.1	-33.8	34.5	1	5
T109	-15.0	-57.2	-93.4	-64.7	99.1	6	6
T110	52.9	100.1	-14.2	19.6	38.8	5	4
T111	117.9	-105.8	51.0	4.7	-2.8	6	2
T112	104.5	-86.7	39.7	28.4	6.9	6	2
T113	60.7	-61.7	10.2	17.6	0.1	6	2
T114	-20.9	-105.3	16.9	44.8	16.9	1	2
T115	-32.2	-115.7	56.7	-7.8	29.3	1	5
T116	-39.9	-110.1	-8.1	15.3	-6.9	1	1
T117	-2.0	-110.1	-17.4	-41.2	39.1	1	6

in addition, a BGA two class classification based on the PR status was performed, yielding a similar plot as with the previously shown for the ER status (see **Figure 43**) where all samples are distributed likewise as in **Figure 40**, and the most discriminant genes are comparable to the ones obtained for the BGA classification by means of the ER status.

On **Table 10** a summary of the distribution of samples along the axis and their clinical markers scored by immunohistochemistry. The column “IHC PR status” provides the assignment of class prior the BGA analysis, defined based on the clinical status of the PR. The samples that are lying on the axis discordant with the clinical data are colored in red.

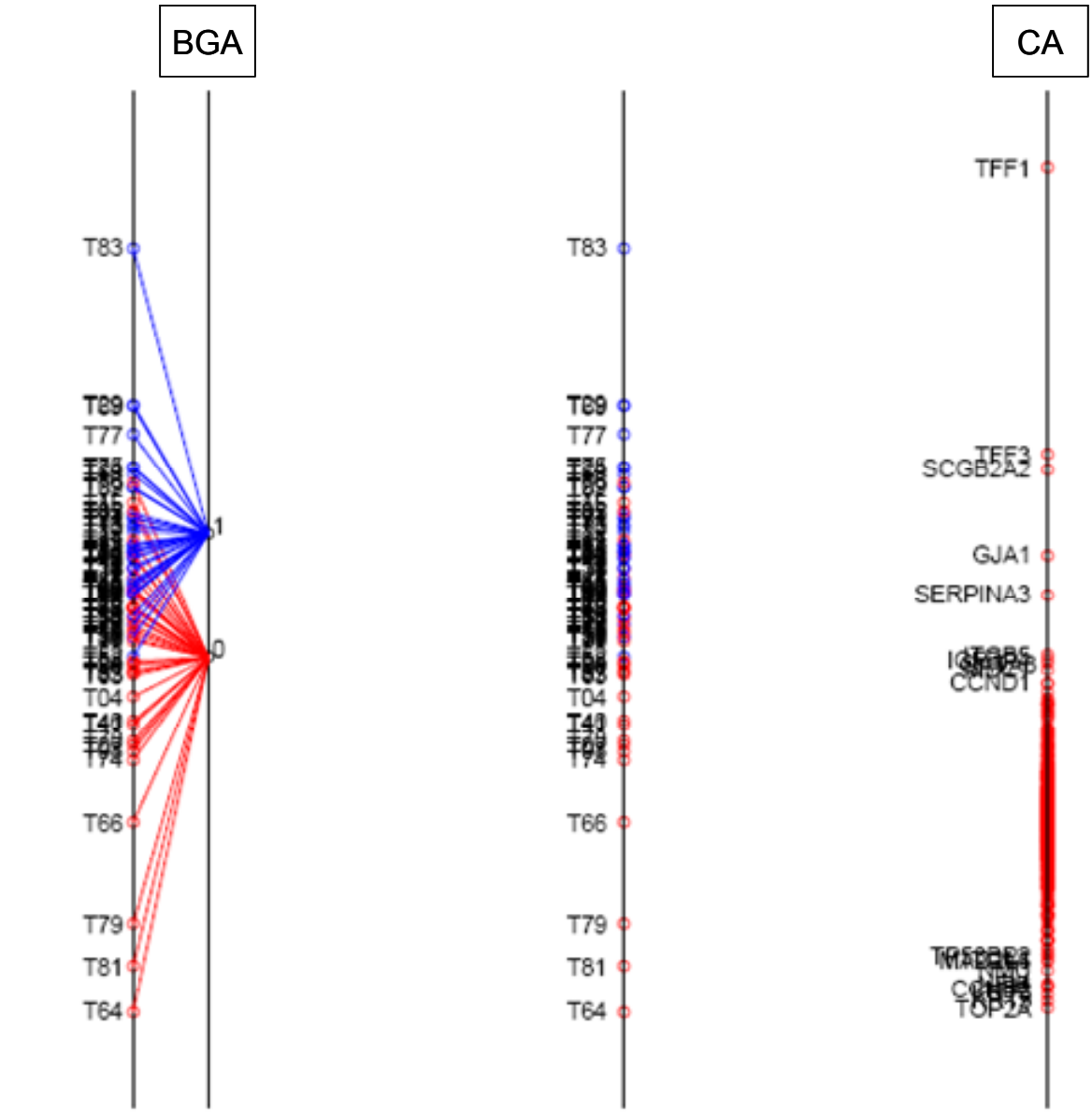


Figure 43: Supervised classification of breast samples by BGA into two groups based on their PR status.

Table 10: Distribution of tumor samples along one axis after two class supervised classification by BGA considering the PR status. Samples are sorted according the PCA axis 1 value. Status of their immunohistochemical markers and the predicted class into which the application would distribute the samples are stated.

Sample ID	Patient ID	IHC PR status	ER	PR	her-2	p53	BRCA1/2	Histological grade	Tumor size	Lymph node status	Metastatic sites	BGA - Two Class - Axis1	BGA predicted
T83	R13722	1	ER+	PR+	ERBB2-	P53-		HG1	T2	N1	M0	653.0	1
T29	R10112	1	ER+	PR+	ERBB2-	P53-		HG0	T1	N0	M0	364.5	1
T69	R12861	1	ER-	PR+	ERBB2-	P53-		HG2	T1	N0	M0	359.7	1
T77	R13633	1	ER+	PR+	ERBB2-	P53-		HG1	T2	N0	M0	307.7	1
T75	R13631	1	ER-	PR+	ERBB2-	P53-		HG1	T1	N0	M0	247.6	1
T24	R10104	1	ER+	PR+	ERBB2-	P53+		HG2	T1	N0	M0	245.6	1
T68	R12860	1	ER+	PR+	ERBB2-	P53-		HG2	T2	N3	M0	233.6	1
T76	R13632	0	ER+	PR-	ERBB2-	P53-		HG2	T2	N0	M0	219.1	1
T09	R7617	1	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	209.9	1
T32	R10116	1	ER+	PR+	ERBB2-	P53+		HG2	T3	N0	M0	208.2	1
T15	R7946	0	ER+	PR-	ERBB2-	P53-		HG0	T2	N1	M0	179.3	1
T82	R13721	0	ER+	PR-	ERBB2-	P53-		HG0	T2	N0	M0	164.6	1
T05	R7612	1	ER+	PR+	ERBB2-	P53+		HG0	T1	N0	M0	161.6	1
T72	R13628	1	ER+	PR+	ERBB2-	P53-		HG3	T2	N1	M0	153.5	1
T31	R10115	1	ER+	PR+	ERBB2-	P53+		HG2	T3	N0	M0	151.8	1
T73	R13629	1	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	141.0	1
T12	R7906	1	ER+	PR+	ERBB2-	P53-		HG0	T2	N1	M0	136.2	1
T22	R10101	1	ER+	PR+	ERBB2-	P53-		HG1	T2	N0	M0	126.8	1
T27	R10108	0	ER+	PR-	ERBB2-	P53+		HG2	T4	N1	M0	112.6	1
T23	R10103	1	ER+	PR+	ERBB2-	P53-		HG2	T2	N1	M0	103.3	1
T13	R7904	1	ER+	PR+	ERBB2-	P53-		HG0	T3	N0	M0	102.8	1
T37	R6282	1	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	98.2	1
T25	R10105	1	ER+	PR+	ERBB2-	P53+		HG2	T1	N0	M0	91.6	1
T71	R13627	1	ER+	PR+	ERBB2-	P53-		HG2	T3	N1	M0	90.0	1
T44	R6280	1	ER+	PR+	ERBB2-	P53-		HG3	T2	N0	M0	88.6	1
T14	R7939	0	ER+	PR-	ERBB2-	P53-		HG0	T1	N0	M0	84.9	1
T02	R7540	0	ER+	PR-	ERBB2-	P53-		HG2	T2	N3	M0	82.1	1
T43	R6267	1	ER+	PR+	ERBB2-	P53-		HG3	T2	N1	M0	78.6	1
T26	R10107	1	ER+	PR+	ERBB2-	P53+		HG3	T1	N1	M0	78.5	1
T30	R10113	1	ER+	PR+	ERBB2-	P53+		HG3	T1	N0	M0	62.5	1
T18	R8153	0	ER+	PR-	ERBB2-	P53-		HG0	T1	N0	M0	58.2	1
T38	R6283	1	ER+	PR+	ERBB2-	P53-		HG0	T1	N0	M0	56.9	1
T17	R8150	0	ER+	PR-	ERBB2-	P53+		HG0	IS	N0	M0	40.8	1
T61	R7346	1	ER+	PR+	ERBB2-	P53-		HG2	T1	N0	M0	35.1	1
T45	R6285	1	ER+	PR+	ERBB2+	P53-		HG0	T1	N0	M0	29.0	1
T10	R7541	1	ER+	PR+	ERBB2-	P53-		HG0	T2	N2	M0	28.3	1
T67	R12856	1	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	21.9	1
T46	R6345	0	ER+	PR-	ERBB2-	P53-		HG1	T1	N0	M0	21.2	1
T03	R7545	0	ER-	PR-	ERBB2-	P53-		HG2	T1	N2	M0	20.4	1
T60	R7343	0	ER+	PR-	ERBB2-	P53-		HG2	T2	N0	M1	17.0	1
T80	R13715	1	ER+	PR+	ERBB2-	P53-		HG2	T2	N0	M0	11.8	1
T42	R6265	1	ER+	PR+	ERBB2+	P53+		HG3	T1	N1	M0	10.5	1
T33	R10117	0	ER+	PR-	ERBB2-	P53-		HG1	T1	N0	M0	10.5	1
T06	R7614	1	ER+	PR+	ERBB2-	P53-		HG2	T1	N0	M0	8.3	1
T62	R7347	0	ER+	PR-	ERBB2-	P53-		HG2	T2	N2	M0	-13.0	0
T65	R12853	0	ER+	PR-	ERBB2-	P53-		HG2	T2	N2	M0	-14.6	0
T48	R7140	0	ER-	PR-	ERBB2-	P53-		HG3	T2	N2	M0	-16.5	0
T47	R6256	0	ER-	PR-	ERBB2-	P53-		HG3	T2	N2	M0	-16.7	0
T70	R13626	0	Normal (adjacent to T71)									-18.4	0
T53	R7145	1	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	-29.4	0
T50	R7143	0	ER-	PR-	ERBB2-	P53-		HG3	T2	N3	M1	-31.2	0
T51	R7170	0	ER+	PR-	ERBB2-	P53+		HG2	T1	N0	M0	-35.3	0
T49	R7142	1	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	-46.0	0
T20	R8403	0	ER-	PR-	ERBB2-	P53+		HG3	T2	N3	M0	-48.7	0
T55	R7322	0	ER-	PR-	ERBB2+	P53-		HG0	T3	N3	M0	-56.1	0
T59	R7335	0	ER-	PR-	ERBB2-	P53+		HG1	T2	N0	M0	-62.5	0
T56	R7328	1	ER+	PR+	ERBB2-	P53-		HG3	T4	N0	M0	-68.6	0
T39	R6250	0	ER-	PR-	ERBB2-	P53+		HG3	T2	N0	M0	-72.0	0
T36	R6279	0	Normal (adjacent to T35)									-76.9	0
T52	R7176	1	ER+	PR+	ERBB2-	P53+		HG2	T1	N0	M0	-106.3	0
T58	R7334	0	Normal (adjacent to T59)									-114.1	0
T54	R7156	0	ER-	PR-	ERBB2+	P53-		HG3	T2	N0	M0	-121.2	0
T28	R10109	0	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M0	-121.8	0
T57	R7331	0	ER-	PR-	ERBB2-	P53+		HG3	T2	N0	M0	-132.6	0
T35	R6277	0	ER-	PR-	ERBB2+	P53-		HG3	T1	N0	M0	-135.6	0
T63	R7351	0	ER-	PR-	ERBB2-	P53+	BRCA2	HG2	T2	N0	M0	-141.2	0
T04	R7553	0	ER-	PR-	ERBB2-	P53-		HG3	T3	N1	M0	-181.5	0
T40	R6254	0	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M1	-227.2	0
T41	R6255	0	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M1	-233.8	0
T78	R13713	0	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M0	-262.5	0
T07	R7692	0	ER-	PR-	ERBB2-	P53+		HG0	T3	N1	M1	-273.2	0
T08	R7693	0	ER-	PR-	ERBB2-	P53+		HG0	T3	N1	M1	-283.0	0
T74	R13630	0	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M0	-299.0	0
T66	R12854	0	ER-	PR-	ERBB2-	P53+		HGx	T2	N1	M0	-414.5	0
T79	R13714	0	ER-	PR-	ERBB2+	P53-		HG3	T2	N0	M0	-604.5	0
T81	R13719	0	ER-	PR-	ERBB2-	P53+		HG3	T2	N0	M0	-683.5	0
T64	R12849	0	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M0	-765.9	0

4.11 Prediction analysis of microarrays (PAM)

Variation in expression patterns of human breast tumors analyzed by cDNA microarrays and hierarchical clustering provides a "molecular portrait" of each tumor, and tumors can be classified into subtypes based solely on differences in these patterns. Sorlie *et al.* 2003, performed an analysis of a large number of tumor samples and explored the clinical value of the found subtypes by searching for correlations between gene expression patterns and clinically relevant parameters. They found that classification of tumors based on gene expression patterns could be used as a prognostic marker with respect to overall and relapse-free survival. Local recurrence is associated with an increased risk of developing distant metastasis and subsequent death from breast cancer. Their approach was to correlate the PAM predicted tumor subtypes with already known local recurrence, presence of metastatic sites or death. The supervised class prediction method employed by Sorlie was PAM (Prediction analysis of microarrays; Tibshirani *et al.* 2003).

Sorlie's training set was a pre-selected set of samples of invasive ductal carcinoma (IDC), the ones that showed best fitting in their model (see **Figure 44**). IDCs constitute 80% of all breast cancers. Invasive lobular carcinoma, constitute an additional 10-15% of breast cancers. Besides, ten additional types of breast cancer have also been described, although they account for less than 10% (Vargo-Gogola and Rosen, 2007).

Sorlie "training" set of tumors, a total of 81, classified in five groups:

- Luminal A (28),
- Luminal B (11),
- Basal (19),
- ERBB2+ (11), and
- Normal (10).

Firstly, we tried to follow the classification of our set of tumors based on Sorlie's set to see if we could find similar patterns. We tried to use the overlapping genes within the "intrinsic list" (552 spots that correspond to 526 genes; Sorlie *et al.* 2003, hereafter Sorlie). Genes from the Sorlie "intrinsic list" that overlap with BCA v4.0 were determined using Matchminer (http://discover.nci.nih.gov/matchminer/index.jsp). There were only 50 spots corresponding to 42 genes. Gene expression data from Sorlie were downloaded from the Stanford database (http://genome-www5.stanford.edu/). We observed that from these 50 spots there were several gene products or ESTs included for the same RefSeq accession number but with a significantly different behavior.

We have also noticed in their array data that they had many spots with missing values, which could handicap the data and give misleading results, even though PAM checks for missing values and imputes them, using the K-nearest neighbor average expression for that gene (by default k=10).

We tried to use this overlapping set of genes to classify samples of our dataset, but in this way the misclassification error was too large, and even the training error for her training set of tumors was larger than 10%.

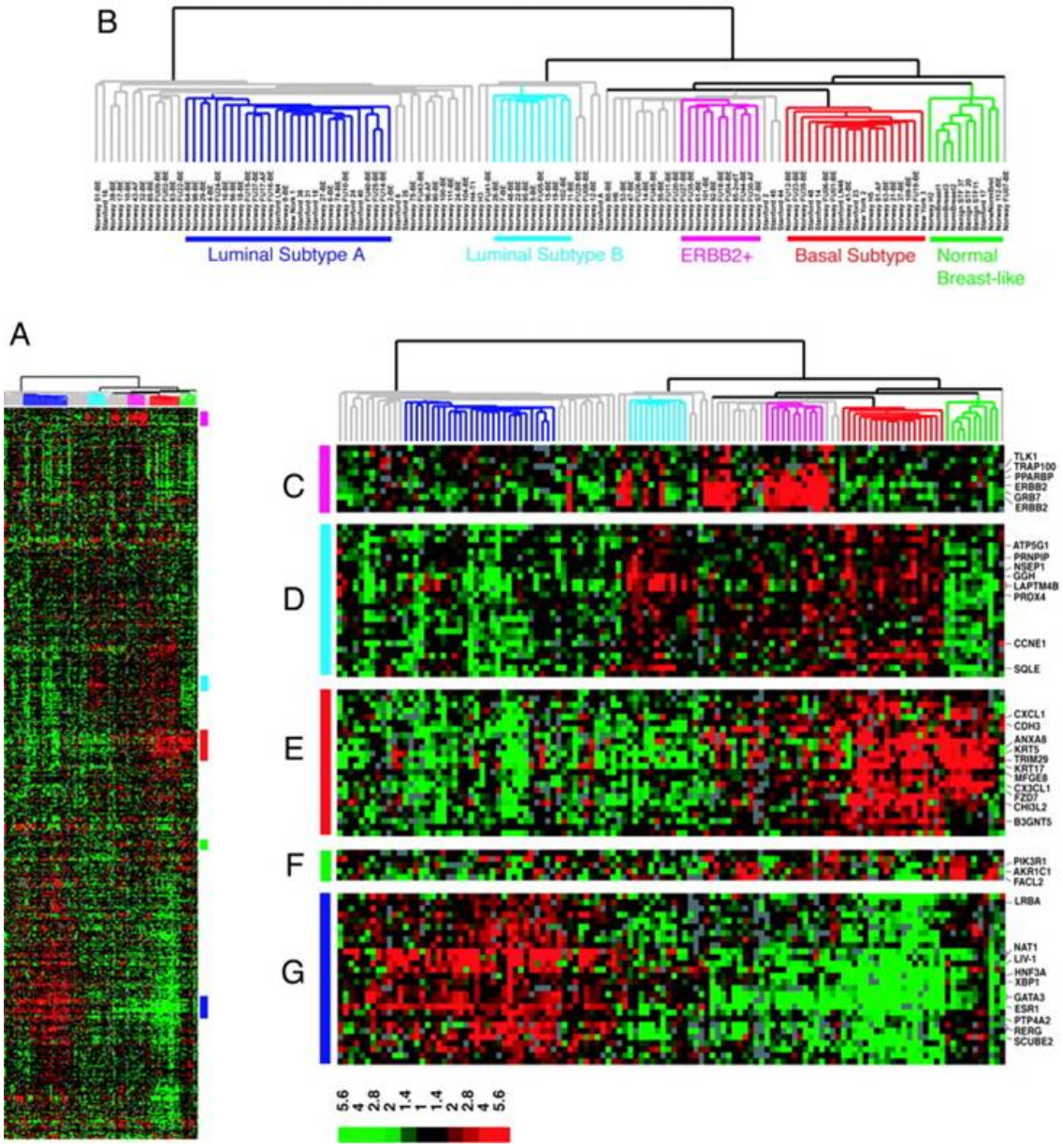


Figure 44: Hierarchical clustering of 115 tumor tissues and 7 nonmalignant tissues using the "intrinsic" gene set. Dendrogram shows the clustering of the samples into five subtypes of IDC. Figure from Sorlie *et al.* 2003.

Besides, Sorlie's "intrinsic list" has also failed for class prediction when applied to other sets of breast tumor microarray data, possibly due to work with a small number of overlapping genes, change of nature of microarray platform (oligos/cDNA), microarray technology, sample used as reference, selection of genes represented on the array, experimental conditions, patient selection or breast tumor sample origin. This is being reported in various studies (Ein-Dor *et al.* 2005; Michiels *et al*. 2007; Wang *et al.* 2005).

Therefore after this attempt, we concluded that it was better to use our own dataset of well-characterized samples by clinicians and our complete gene collection to find distinctive breast tumor subtypes. Later, our approach was, once the discriminant genes from our phenotypes were selected, to extract the data from Sorlie, and see if with our gene set we could find other patterns in their expression and clinical data.

4.11.1 Determination of the training set of samples

The first set of 74 well characterized tumor samples tumors was used as a training set. Starting from the BGA supervised classification into 6 groups, we considered 5 subtypes labeled 1-5 after removing the "normal" group due to its small size (only 3 samples). Besides, there is no need for a prediction of normal samples since these samples are extracted from tissue adjacent to tumor. Scheme for this analysis is shown in **Figure 45**.

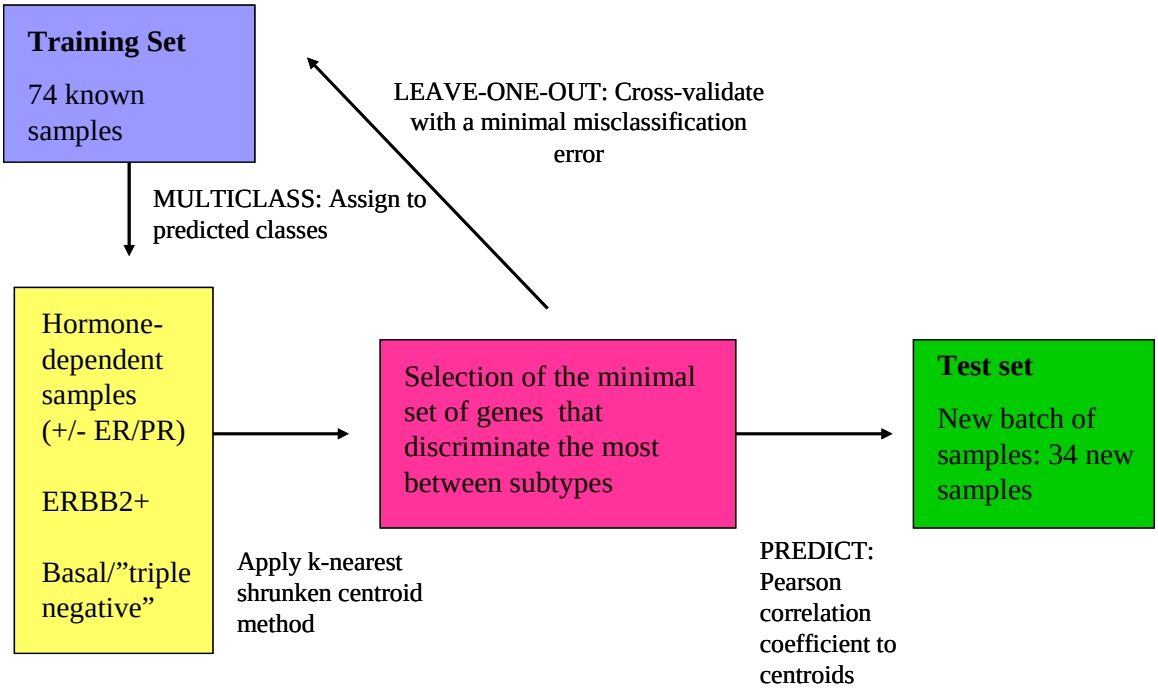


Figure 45: Scheme of the PAM procedure applied to our set of breast tumor samples.

The K-nearest neighbor shrunken centroid method was used as classification engine, employing 10 neighbors. Columns, that is, tumor samples, were centered to the mean and scaled, for correction of batch effects. Training error was plotted and sample classification was observed. Tumor samples were exchanged among the classes, in order to obtain the lowest training error. The training error plot is represented in **Figure 46**.

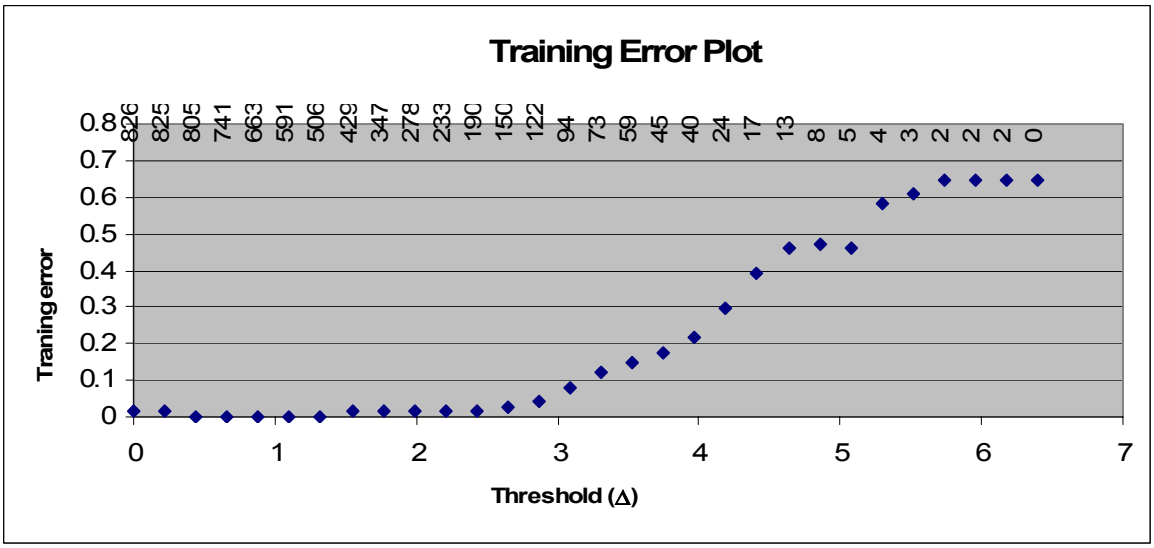


Figure 46: Training error plot for PAM classification. The minimum training error is found at a threshold Δ of 1.32, using a set of 506 genes as the minimal classification set. Using only 94 genes the training error is still below 10%.

Classification of tumor samples was distributed as shown in **Table 11.**

Table 11: Tumor sample prediction among the 5 predicted classes.

Training Confusion Matrix (Threshold=1.32)

True\Predicted	1	2	3	4	5	Class Error rate
1	12	0	0	0	0	0
2	0	26	0	0	0	0
3	0	0	19	0	0	0
4	0	0	0	5	0	0
5	0	0	0	0	12	0

During the cross-validation the parameter Δ, which sets the amount of shrinkage, is iteratively increased to balance prediction accuracy and the selection of a minimal set of genes. Using as threshold a Δ of 1.33, which includes 506 genes as significant, just two samples were differently assign: sample T77 which would change from class 1 to class 3, and sample T60 classified as belonging to class 2 (**Figure 47**).

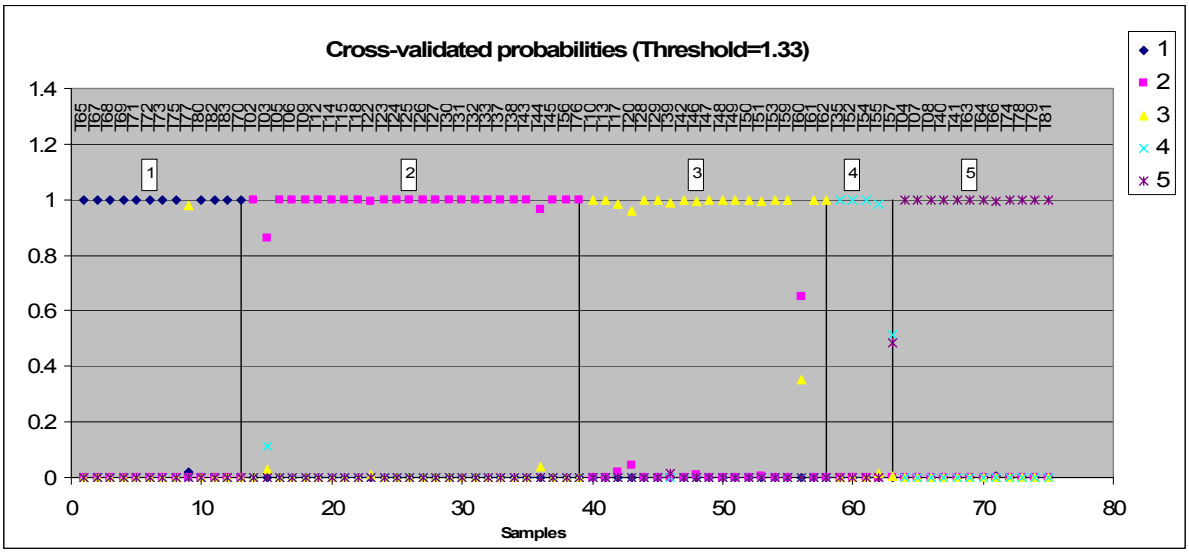


Figure 47: Cross-validated probabilities of the training set of tumor samples.

From those 506 genes, our "intrinsic list", a set of top discriminating genes between classes can be selected, which have greater classification weight due to their absolute change in gene-expression amongst the classes. The parameter Δ is further used to minimize the gene list during the prediction. During a ten-fold cross-validation, Δ was increased so that a lesser misclassification error could be chosen for the final model. Using just 150 significant genes between classes, that is, selecting a threshold Δ of 2.65, misclassification error was less than 1%, and only two samples were misclassified after cross-validation: sample T60 firstly assign to class 3 was predicted to belong to class 2, and sample T77 was predicted to be in class 3. The final class prediction model with their nearest shrunken centroids is shown in **Figure 48** showing their cross-validated probabilities.

From the clinical histopatological records, the 12 tumor samples from the training set which belonged to PAM class 1 are ER+ (10/12) and PR+ (10/12), all are p53- and have low histological grade. From these 12 samples, 11 are free of disease and only one, tumor sample ID T65, had recurrence in liver and bone. From these findings, this subtype of tumor consists of hormone-dependent tumors, often luminal or endocrine-like, which may reflect the origin of the cancer cell, with overall good prognosis.

PAM Class 2 contains 26 samples which are mostly ER+ (25/26), PR+ (18/26), a few samples were positive for p53+ (8/26), and also a few have high histological grade (5/26). From these 26 samples, only 2 had recurrence of visceral or systemic type. The other 24 samples are disease-free. This class has high expression of ER co-regulated genes such as GATA3, genes found regulated by estrogens as TFF1, TFF3, SERPINA3, or products of estradiol metabolism as CYP2B6 or HSD17B4. This tumor subtype is hormone-dependent, luminal or endocrine-like, and has good prognosis with a 92% of

disease-free-survival. This class is similarly termed in Sorlie's as luminal subtype A and shares genes in common with Sorlie's luminal subtype A.

Class 3 contains 19 samples, including ER+ (12/19), PR+ (7/19), p53+ (7/19) and high grade is found in 7 samples. There has not been follow-up in three cases, and in three other cases patient had recurrence locally, liver and bone.

Class 4 is the ERBB2+ subtype characterized by high expression of several genes in the ERBB2 amplicon at 17q12-q24 (Bergamaschi *et al*. 2006) including the ERBB2 gene, TOP2A, CDC6, PPARBP, GRB2, and SMARCE1. One out if five samples is ER+. These patients are now, after treatment, disease-free.

Class 5 is the characteristic basal-like subtype characterized by high expression of keratins 5A and 5B and showed also high expression of other basal epithelial genes as KIT1 and ID4. These are also termed as "triple-negative tumors" as they are found negative for ER, PR and ERBB2. 9 out of 12 samples are p53+ and have high histological grade (10/12). This tumor subtype has a lower disease-free survival time and 7 patients presented metastasis in lung, viscerae, central nervous system, systematically or locally, and died from the disease.

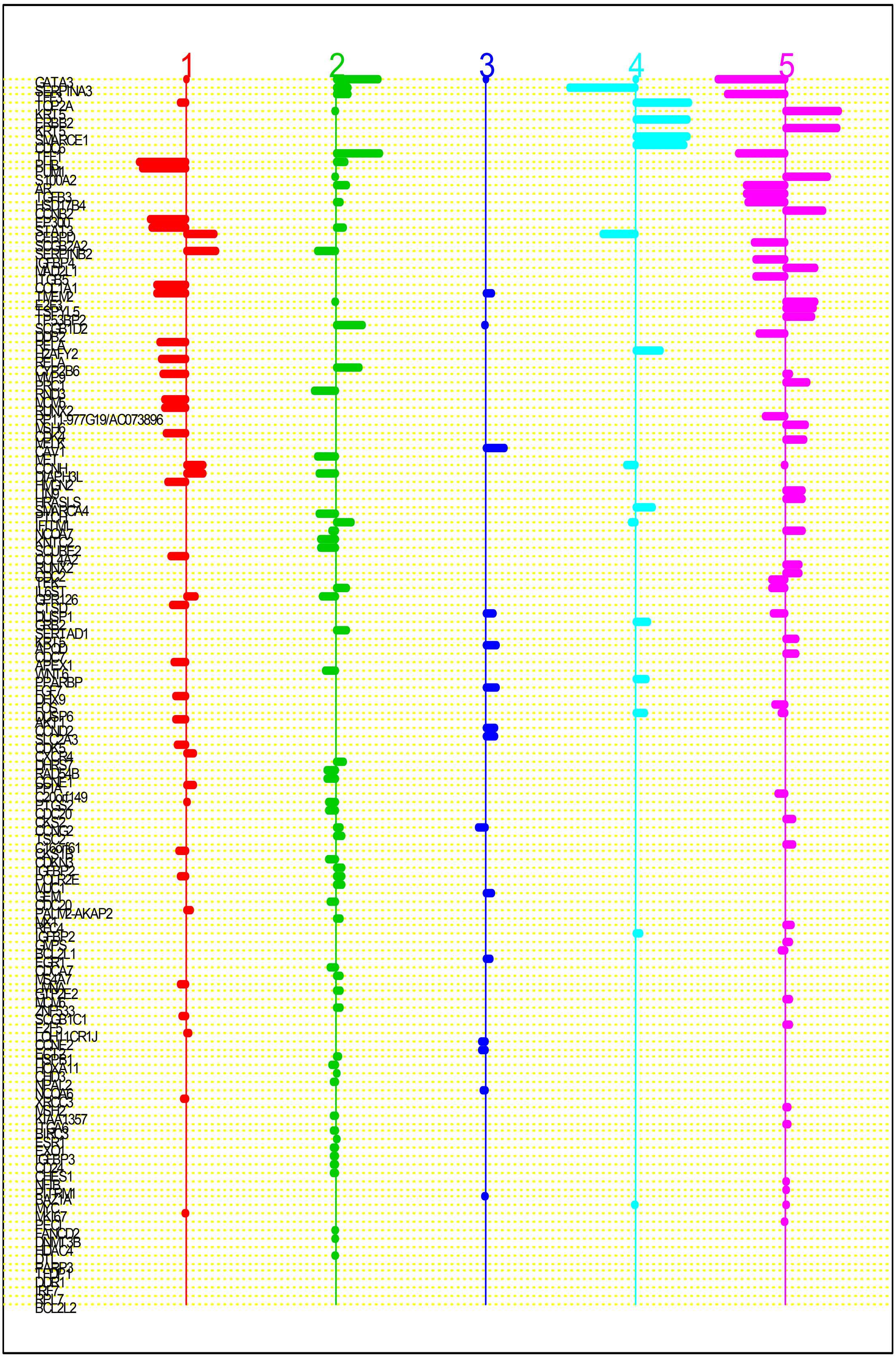


Figure 48: Expression shrunken centroids of each of the five subclasses showing the most discriminating genes, with higher scores, between predicted breast tumor subtypes.

A summary of the PAM predicted classification of the training set together with the clinical histopathological features is shown at **Table 12**. BGA two classes and BGA multiple class predicted classes is shown in parallel to PAM.

Table 12: Class prediction of tumor samples by PAM.

Sample ID	Patient ID	ER	PR	her-2	p53	BRCA1/2	Histological grade	Tumor size	Lymph node status	Metastatic sites	BGA - Two Class - Axis1	BGA ER two class	BGA multiclass	PAM multiclass
T02	R7540	ER+	PR-	ERBB2-	P53-		HG2	T2	N3	M0	56.3	1	2	2
T03	R7545	ER-	PR-	ERBB2-	P53-		HG2	T1	N2	M0	-12.4	0	3	2
T04	R7553	ER-	PR-	ERBB2-	P53-		HG3	T3	N1	M0	-238.9	0	5	5
T05	R7612	ER+	PR+	ERBB2-	P53+		HG0	T1	N0	M0	124.3	1	2	2
T06	R7614	ER+	PR+	ERBB2-	P53-		HG2	T1	N0	M0	-62.6	0	3	2
T07	R7692	ER-	PR-	ERBB2-	P53+		HG0	T3	N1	M1	-290.6	0	5	5
T08	R7693	ER-	PR-	ERBB2-	P53+		HG0	T3	N1	M1	-285.7	0	5	5
T09	R7617	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	167.9	1	2	2
T10	R7541	ER+	PR+	ERBB2-	P53-		HG0	T2	N2	M0	-59.1	0	3	3
T12	R7906	ER+	PR+	ERBB2-	P53-		HG0	T2	N1	M0	117.2	1	2	2
T13	R7904	ER+	PR+	ERBB2-	P53-		HG0	T3	N0	M0	43.8	1	2	3
T14	R7939	ER+	PR-	ERBB2-	P53-		HG0	T1	N0	M0	127.8	1	2	2
T15	R7946	ER+	PR-	ERBB2-	P53-		HG0	T2	N1	M0	189.8	1	2	2
T17	R8150	ER+	PR-	ERBB2-	P53+		HG0	IS	N0	M0	108.9	1	2	3
T18	R8153	ER+	PR-	ERBB2-	P53-		HG0	T1	N0	M0	88.8	1	2	2
T20	R8403	ER-	PR-	ERBB2-	P53+		HG3	T2	N3	M0	-70.4	0	3	3
T22	R10101	ER+	PR+	ERBB2-	P53-		HG1	T2	N0	M0	134.0	1	2	2
T23	R10103	ER+	PR+	ERBB2-	P53-		HG2	T2	N1	M0	55.5	1	2	2
T24	R10104	ER+	PR+	ERBB2-	P53+		HG2	T1	N0	M0	163.8	1	2	2
T25	R10105	ER+	PR+	ERBB2-	P53+		HG2	T1	N0	M0	16.6	1	2	2
T26	R10107	ER+	PR+	ERBB2-	P53+		HG3	T1	N1	M0	46.6	1	2	2
T27	R10108	ER+	PR-	ERBB2-	P53+		HG2	T4	N1	M0	73.9	1	2	2
T28	R10109	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M0	-115.1	0	3	3
T29	R10112	ER+	PR+	ERBB2-	P53-		HG0	T1	N0	M0	423.5	1	1	3
T30	R10113	ER+	PR+	ERBB2-	P53+		HG3	T1	N0	M0	2.1	1	2	2
T31	R10115	ER+	PR+	ERBB2-	P53+		HG2	T3	N0	M0	88.3	1	2	2
T32	R10116	ER+	PR+	ERBB2-	P53+		HG2	T3	N0	M0	149.9	1	2	2
T33	R10117	ER+	PR-	ERBB2-	P53-		HG1	T1	N0	M0	-33.3	0	3	2
T35	R6277	ER-	PR-	ERBB2+	P53-		HG3	T1	N0	M0	-208.6	0	5	4
T37	R6282	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	28.4	1	2	2
T38	R6283	ER+	PR+	ERBB2-	P53-		HG0	T1	N0	M0	-14.6	0	3	2
T39	R6250	ER-	PR-	ERBB2-	P53+		HG3	T2	N0	M0	-104.9	0	3	3
T40	R6254	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M1	-248.7	0	5	5
T41	R6255	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M1	-234.8	0	5	5
T42	R6265	ER+	PR+	ERBB2+	P53+		HG3	T1	N1	M0	59.9	1	2	3
T43	R6267	ER+	PR+	ERBB2-	P53-		HG3	T2	N1	M0	-16.8	0	3	2
T44	R6280	ER+	PR+	ERBB2-	P53-		HG3	T2	N0	M0	36.3	1	2	2
T45	R6285	ER+	PR+	ERBB2+	P53-		HG0	T1	N0	M0	-36.0	0	3	2
T46	R6345	ER+	PR-	ERBB2-	P53-		HG1	T1	N0	M0	-19.4	0	3	3
T47	R6256	ER-	PR-	ERBB2-	P53-		HG3	T2	N2	M0	-68.0	0	3	3
T48	R7140	ER-	PR-	ERBB2-	P53-		HG3	T2	N2	M0	-71.5	0	3	3
T49	R7142	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	-73.4	0	3	3
T50	R7143	ER-	PR-	ERBB2-	P53-		HG3	T2	N3	M1	-63.2	0	3	3
T51	R7170	ER+	PR-	ERBB2-	P53+		HG2	T1	N0	M0	-103.0	0	3	3
T52	R7176	ER+	PR+	ERBB2-	P53+		HG2	T1	N0	M0	-211.2	0	5	4
T53	R7145	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	-109.4	0	3	3
T54	R7156	ER-	PR-	ERBB2+	P53-		HG3	T2	N0	M0	-219.8	0	4	4
T55	R7322	ER-	PR-	ERBB2+	P53-		HG0	T3	N3	M0	-135.8	0	4	4
T56	R7328	ER+	PR+	ERBB2-	P53-		HG3	T4	N0	M0	-146.9	0	4	2
T57	R7331	ER-	PR-	ERBB2-	P53+		HG3	T2	N0	M0	-238.3	0	4	4
T59	R7335	ER-	PR-	ERBB2-	P53+		HG1	T2	N0	M0	-89.9	0	3	3
T60	R7343	ER+	PR-	ERBB2-	P53-		HG2	T2	N0	M1	-9.1	0	3	3
T61	R7346	ER+	PR+	ERBB2-	P53-		HG2	T1	N0	M0	-44.7	0	3	3
T62	R7347	ER+	PR-	ERBB2-	P53-		HG2	T2	N2	M0	-64.5	0	3	3
T63	R7351	ER-	PR-	ERBB2-	P53+	BRCA2	HG2	T2	N0	M0	-217.8	0	4	5
T64	R12849	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M0	-471.3	0	5	5
T65	R12853	ER+	PR-	ERBB2-	P53-		HG2	T2	N2	M0	469.9	1	1	1
T66	R12854	ER-	PR-	ERBB2-	P53+		HGx	T2	N1	M0	-392.0	0	5	5
T67	R12856	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	49.2	1	2	1
T68	R12860	ER+	PR+	ERBB2-	P53-		HG2	T2	N3	M0	352.6	1	1	1
T69	R12861	ER-	PR+	ERBB2-	P53-		HG2	T1	N0	M0	525.3	1	1	1
T71	R13627	ER+	PR+	ERBB2-	P53-		HG2	T3	N1	M0	252.5	1	1	1
T72	R13628	ER+	PR+	ERBB2-	P53-		HG3	T2	N1	M0	179.6	1	2	1
T73	R13629	ER+	PR+	ERBB2-	P53-		HG1	T1	N0	M0	212.3	1	2	1
T74	R13630	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M0	-341.4	0	5	5
T75	R13631	ER-	PR+	ERBB2-	P53-		HG1	T1	N0	M0	362.5	1	1	1
T76	R13632	ER+	PR-	ERBB2-	P53-		HG2	T2	N0	M0	268.2	1	2	2
T77	R13633	ER+	PR+	ERBB2-	P53-		HG1	T2	N0	M0	397.3	1	1	1
T78	R13713	ER-	PR-	ERBB2-	P53+		HG3	T1	N0	M0	-208.0	0	5	5
T79	R13714	ER-	PR-	ERBB2+	P53-		HG3	T2	N0	M0	-545.6	0	5	5
T80	R13715	ER+	PR+	ERBB2-	P53-		HG2	T2	N0	M0	17.0	1	2	1
T81	R13719	ER-	PR-	ERBB2-	P53+		HG3	T2	N0	M0	-598.0	0	5	5
T82	R13721	ER+	PR-	ERBB2-	P53-		HG0	T2	N0	M0	414.9	1	1	1
T83	R13722	ER+	PR+	ERBB2-	P53-		HG1	T2	N1	M0	756.2	1	1	1

4.11.2 Selection of the most significant molecular markers

PAM can identify the minimal subsets of the genes that distinctively characterize each cluster. The effect of selecting a lower threshold Δ gives a higher weight to those genes whose expression is more stable within samples of the same gene expression signature class. The 150 most dicriminant genes for each identified class and their scores is shown in **Table 13**.

Table 13: List of most significant genes for subtype discrimination at a shrinkage parameter Δ of 2.65. A positive score means up-regulated gene expression, and a negative score means down-regulated gene expression.

GENES UP-REGULATED

CLASS 1			CLASS 2			CLASS 3			CLASS 4			CLASS 5		
Symbol	GenBank	Score	Symbol	GenBank	Score	Symbol	GenBank	Score	Symbol	GenBank	Score	Symbol	GenBank	Score
SERPINB2	NM_002575	0.4208	TFF1	NM_003225	0.5936	CAV1	NM_001753	0.2563	TOP2A	NM_001067	0.717	KRT5	NM_005554	0.7143
CEBPD	BM924801	0.3854	GATA3	NM_002051	0.582	APOD	NM_001647	0.1576	ERBB2	NM_004448	0.709	KRT5	NM_005555	0.7006
CCNH	NM_001239	0.2498	SCGB1D2	NM_006551	0.3562	FGF7	NM_002009	0.1458	SMARCE1	NM_003079	0.7011	S100A2	NM_005978	0.579
DIAPH3L	BC041395	0.24	CYP2B6	NM_000767	0.3195	CCND2	NM_001759	0.1321	CDC6	NM_001254	0.6574	CCNB2	NM_004701	0.5079
GPR126	NM_020455	0.1314	IFITM1	NM_003641	0.227	SLC2A3	NM_006931	0.1318	H2AFY2	NM_018649	0.3356	MAD2L1	NM_002358	0.401
CXCR4	NM_003467	0.1244	SERPINA3	NM_001085	0.1779	DUSP1	NM_004417	0.1211	SMARCA4	NM_003072	0.236	E2F3	NM_001949	0.392
PPIA	NM_021130	0.1105	TFF3	NM_003226	0.1744	TMEM2	NM_013390	0.0989	GRB2	NM_002086	0.1721	TSPYL5	NM_033512	0.3796
PALM2-AKAP2	NM_007203	0.0807	AR	NM_000044	0.1631	GEM	NM_005261	0.0878	PPARBP	NM_004774	0.1495	TP53BP2	NM_005426	0.3581
LOH11CR1J	AB096249	0.0615	SERTAD1	NM_013376	0.1608	EGR1	NM_001964	0.0725	DUSP6	BC037236	0.1385	PRC1	NM_003981	0.288
PTGS2	NM_000963	0.0402	IL6ST	NM_002184	0.1492	HDAC4	NM_006037	0.013	IGFBP2	NM_000597	0.0773	MSH6	NM_000179	0.2681
EGR1	NM_001964	0.0213	PHB	NM_002634	0.1247							MELK	NM_014791	0.2591
SCUBE2	NM_020974	0.0158	DHRS7	NM_016029	0.1203							LIN9	NM_173083	0.2369
DTL	NM_016448	0.0067	STAT3	NM_139276	0.1094							HRASLS	NM_020386	0.2362
GADD45A	NM_001924	0.0015	TSC2	NM_000548	0.1022							NCOA7	AL834442	0.2252
			IGFBP2	NM_000597	0.0923							RUNX2	AL353944	0.2001
			POLR2E	NM_002695	0.0894							CDC2	NM_001786	0.1966
			MUC1	NM_002456	0.0893							KRT5	NM_000424	0.1588
			MX1	NM_002462	0.078							CDC7	NM_003503	0.1552
			HSD17B4	NM_000414	0.076							CKS2	NM_001827	0.1052
			CCNG2	NM_004354	0.0727							C16orf61	NM_020188	0.0983
			MS4A7	NM_021201	0.0706							RFC4	NM_002916	0.0775
			GTF2E2	NM_002095	0.067							GMPS	NM_003875	0.075
			ZNF533	NM_152520	0.0625							MMP9	NM_004994	0.0673
			HSPB1	NM_001540	0.0516							MCM6	NM_005915	0.0668
			CHD3	AK096555	0.0376							E2F5	NM_001951	0.0627
			ESR1	NM_000125	0.0326							MSH2	NM_000251	0.0356
			IGFBP4	NM_001552	0.0192							ITGA6	NM_000210	0.0345
			DDR1	NM_013994	0.0057							NFIB	NM_005596	0.0271
			IRF7	NM_004031	0.0042							PITRM1	NM_014889	0.0253
			BCL2L2	NM_004050	0.0029							MYC	NM_002467	0.0222
												TFDP1	NM_007111	0.009
												KNTC2	NM_006101	0.0085
												RPL7	NM_000971	0.0052
												TOP2A	NM_001067	0.0022

GENES DOWN-REGULATED

CLASS 1			CLASS 2			CLASS 3			CLASS 4			CLASS 5		
Symbol	GenBank	Score	Symbol	GenBank	Score	Symbol	GenBank	Score	Symbol	GenBank	Score	Symbol	GenBank	Score
PHB	NM_002634	-0.6152	RND3	X97758	-0.2853	CCNG2	NM_004354	-0.104	SERPINA3	NM_001085	-0.8794	GATA3	NM_002051	-0.9088
PUM1	NM_014676	-0.5897	MET	NM_000245	-0.2512	CCNE2	NM_057749	-0.054	CEBPD	BM924801	-0.4391	KRT5	NM_005555	0.7006
EP300	NM_001429	-0.473	SERPINB2	NM_002575	-0.2457	ECT2	NM_018098	-0.0539	CCNH	NM_001239	-0.1246	TFF3	NM_003226	-0.7792
STAT3	NM_139276	-0.4536	PTCH	AK124593	-0.2293	NCOA6	NM_014071	-0.0359	IFITM1	NM_003641	-0.0619	TFF1	NM_003225	-0.6242
COL1A1	NM_000088	-0.3977	DIAPH3L	BC041395	-0.2186	BAZ1A	NM_013448	-0.0234	MYC	NM_002467	-0.0252	AR	NM_000044	-0.5377
TMEM2	NM_013390	-0.3928	KNTC2	NM_006101	-0.2141	SCGB1D2	NM_006551	-0.0158				TGFB3	NM_003239	-0.5344
RELA	NM_021975	-0.3469	SCUBE2	NM_020974	-0.2127							HSD17B4	NM_000414	-0.5131
RELA	NM_021975	-0.3337	GPR126	NM_020455	-0.1812							SCGB2A2	NM_002411	-0.4286
MMP9	NM_004994	-0.3009	WNT6	NM_006522	-0.1488							IGFBP4	NM_001552	-0.4012
MCM5	NM_006739	-0.285	RAD54B	NM_012415	-0.1173							ITGB5	NM_002213	-0.4007
RUNX2	NM_004348	-0.2787	CCNE1	NM_001238	-0.1167							DDB2	NM_000107	-0.3569
CDK4	NM_000075	-0.2643	PTGS2	NM_000963	-0.108							RP11-977G19	AC073896	-0.2732
HMGN2	NM_005517	-0.237	CDC20	NM_001255	-0.1061							TEK	NM_000459	-0.1911
COL4A2	NM_001846	-0.2097	CDKN3	NM_005192	-0.0933							IL6ST	NM_002184	-0.1832
CTSD	NM_001909	-0.178	CDC20	NM_001255	-0.0828							DUSP1	NM_004417	-0.1745
APEX1	NM_001641	-0.1533	CDCA7	NM_031942	-0.0717							FOS	NM_005252	-0.1443
DHX9	NM_001357	-0.1453	NCOA7	AL834442	-0.0674							C20orf149	NM_024299	-0.1096
AKT1	NM_005163	-0.1342	HOXA11	NM_005523	-0.0482							BCL2L1	NM_138578	-0.0748
CDK5	NM_004935	-0.1271	NPAL2	NM_024759	-0.0368							DUSP6	BC037236	-0.0594
CKS1B	NM_001826	-0.095	KIAA1357	XM_050421	-0.0341							PECI	NM_006117	-0.0194
TOP2A	NM_001067	-0.0846	BIRC3	NM_001165	-0.0327							CCNH	NM_001239	-0.018
POLR2E	NM_002695	-0.0732	EXO1	NM_130398	-0.0296							PARP3	NM_005485	-0.0098
LMNA	NM_170707	-0.0687	IGFBP3	NM_000598	-0.0285									
SCGB1C1	NM_145651	-0.0631	CD24	AK125531	-0.0284									
XRCC3	NM_005432	-0.0366	CHES1	NM_005197	-0.0262									
MKI67	NM_002417	-0.0222	KRT5	NM_005554	-0.0253									
RPL7	NM_000971	-0.0032	S100A2	NM_005978	-0.0225									
			E2F3	NM_001949	-0.0205									
			FANCD2	NM_033084	-0.0172									
			DNMT3B	NM_006892	-0.0137									
			DTL	NM_016448	-0.0113									

4.11.3 Analysis of PAM predicted subtypes by Gene Set Enrichment Analysis (GSEA) and Ingenuity Pathway Analysis.

GSEA was supplied with our data files: a chip annotation file, an expression dataset file, a phenotype label file, and gene sets files. Firstly the Breast Cancer Chip v4.0 chip annotation file (file extension .chip) with HUGO gene symbols was loaded in GSEA in order to employ it as background for gene set enrichment analysis.

The expression dataset containing features (genes), samples and an expression value of all 74 tumor samples (format .gct) and the PAM predicted phenotype labels (format .cls) were also imported into GSEA following (see chapter 3.19.3).

Gene sets of most significant genes of each of the PAM predicted breast tumor phenotypes on our set of tumor samples were added to the customized collection of gene sets c5.

GSEA was used in order to perform gene expression signature analysis with already known gene set collections previously described (c1, c2, c3, and c4), and the newly created customized gene-set collection c5, to identify pathways in which the genes could be associated to the predicted breast tumor phenotype. Enrichment test were run using the default settings of the program but changing the minimal size of the gene set to 10 since some of them are smaller than the default minimum of 15.

We selected gene sets with nominal p-value less than 1%, since our intention was to discover similarities between gene sets without being too stringent, since and our gene sets were sometimes small. Nominal p-value is an unadjusted p-value which estimates the statistical significance of a gene-set without adjusting for gene-set size or multiple hypothesis testing. The FDR statistic adjusts for both. Typically, an FDR of less than 25% is most likely to generate more consistent hypothesis but is infrequent to achieve this statistical power with small gene sets.

Analysis of PAM subtype 1

PAM class 1 has 2 gene sets form the c5 dataset collection which are significantly enriched at nominal p-value < 1% and 6 gene sets are significantly enriched at nominal p-value < 5%. In this case no gene sets were found enriched with a FDR of less than 25%.

Leading-edge analysis was performed to find the similarities among different significant gene lists and to determine which genes would have more weight. Mainly, the enriched gene sets correspond to genes which were typically induced by progestins after 6 h in studies using T47D model cell line, such as

AKAP13, CXCR4, BIRC3, GADD45A, ARID1A (SMARCF1), EGR1, DUSP1, CHES1, TGFB3, and FOS (**Figure 49**).

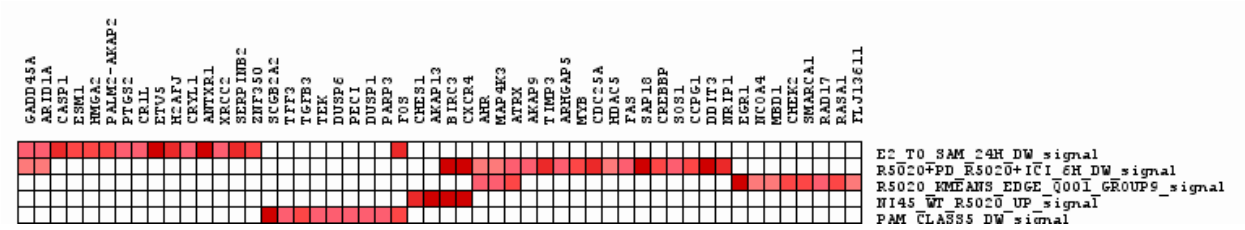


Figure 49: Leading edge analysis of PAM 1 tumor phenotype versus de rest of tumors.

Ingenuity pathway analysis (IPA, see chapter 3.19.4) was used to analyze the most significant signaling pathways and functions of this phenotype. Lists of the most significant genes of each phenotype were imported into IPA. The most relevant functions of this tumor phenotype were cancer (with 26 molecules involved, p-value = 6.97E-26) showing 3 activated molecules involved in cell tumorigenesis such as EGR1, PTGS2 and CXCR4, and others such as CEPDB, CCNH, SERPINB2 and GADD45A, are associated to tumor growth. The most represented molecular and cellular functions are cell growth and proliferation with 25 molecules taking part, and cell signaling with 20 genes of the list. PAM class 1 gene expression signature yielded three main networks. In **Figure 50** is shown the most represented network with the discriminant genes of this subtype.

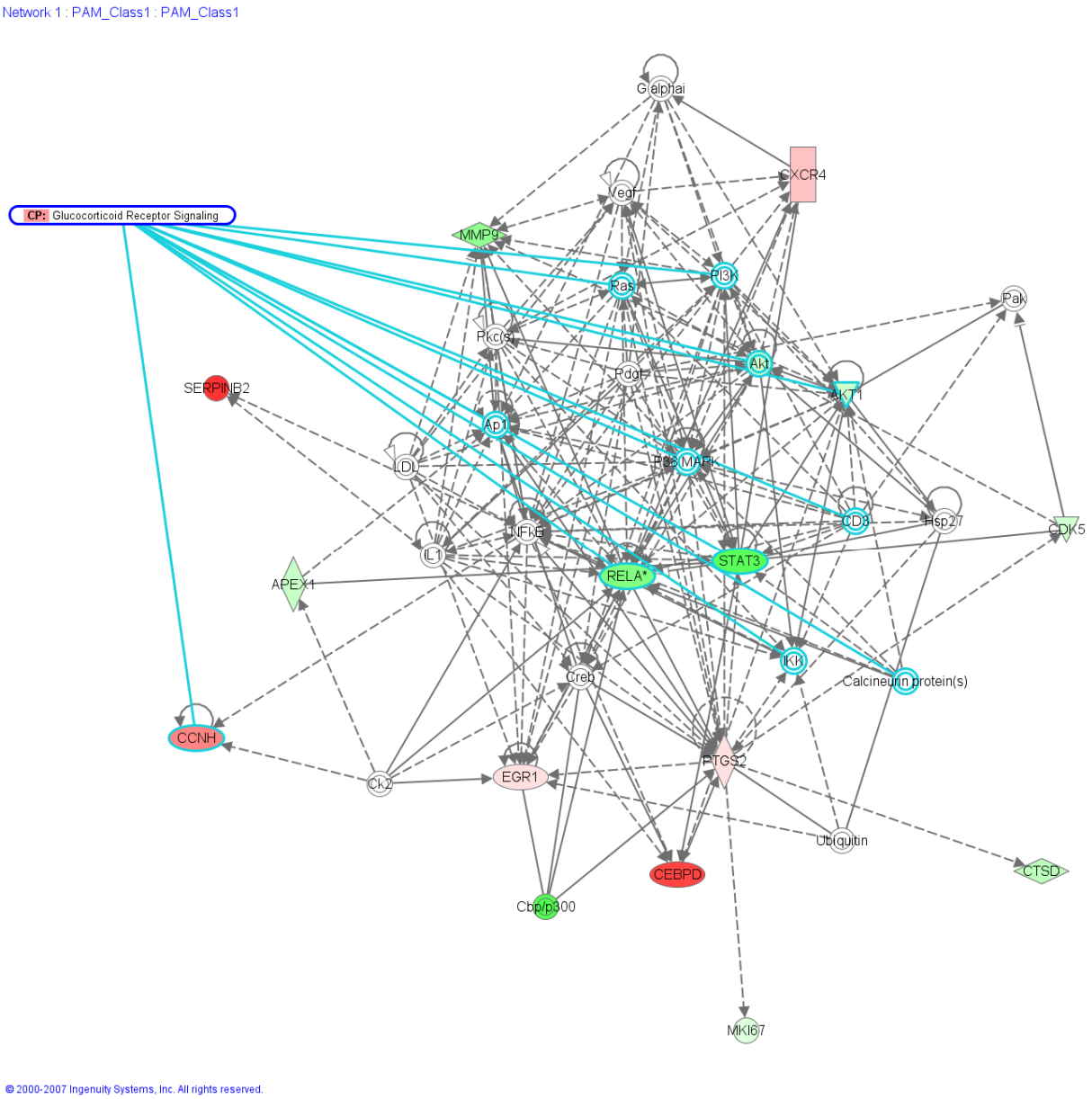


Figure 50: Network 1 of phenotype 1, in which the most significant signaling pathway is the activated glucocorticoid receptor signaling pathway. Molecules are shown in red/green relative to the PAM centroid expression value, red represented a positive score or up-regulated gene expression and green represented a negative score or down-regulated gene expression.

In network 1, there are 14 molecules of our intrinsic list involved, and there appear to be an activated glucocorticoid receptor pathway characteristic of a hormone-dependant tumor type. In network 2, there are 12 molecules of our set involved, related with ongoing inflammatory disease (**Figure 51**).

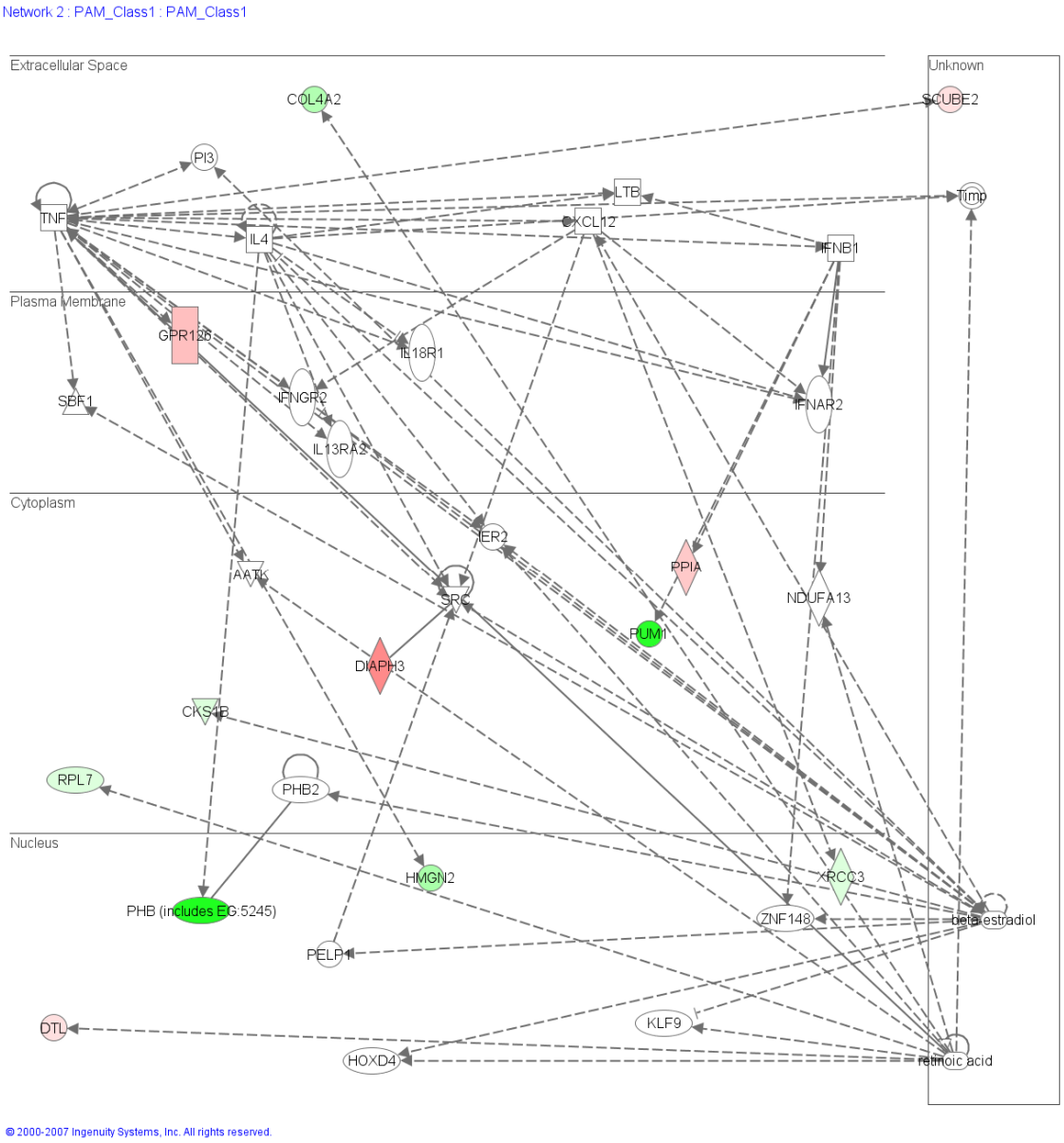


Figure 51: Network 2 of phenotype 1 in a subcellular view, in which to most significant function is related to an inflammatory disease state.

If these two networks are merged, a gene network is obtained in which activated glucocorticoid receptor signaling is connected with inflammation and cell proliferation (**Figure 52**).

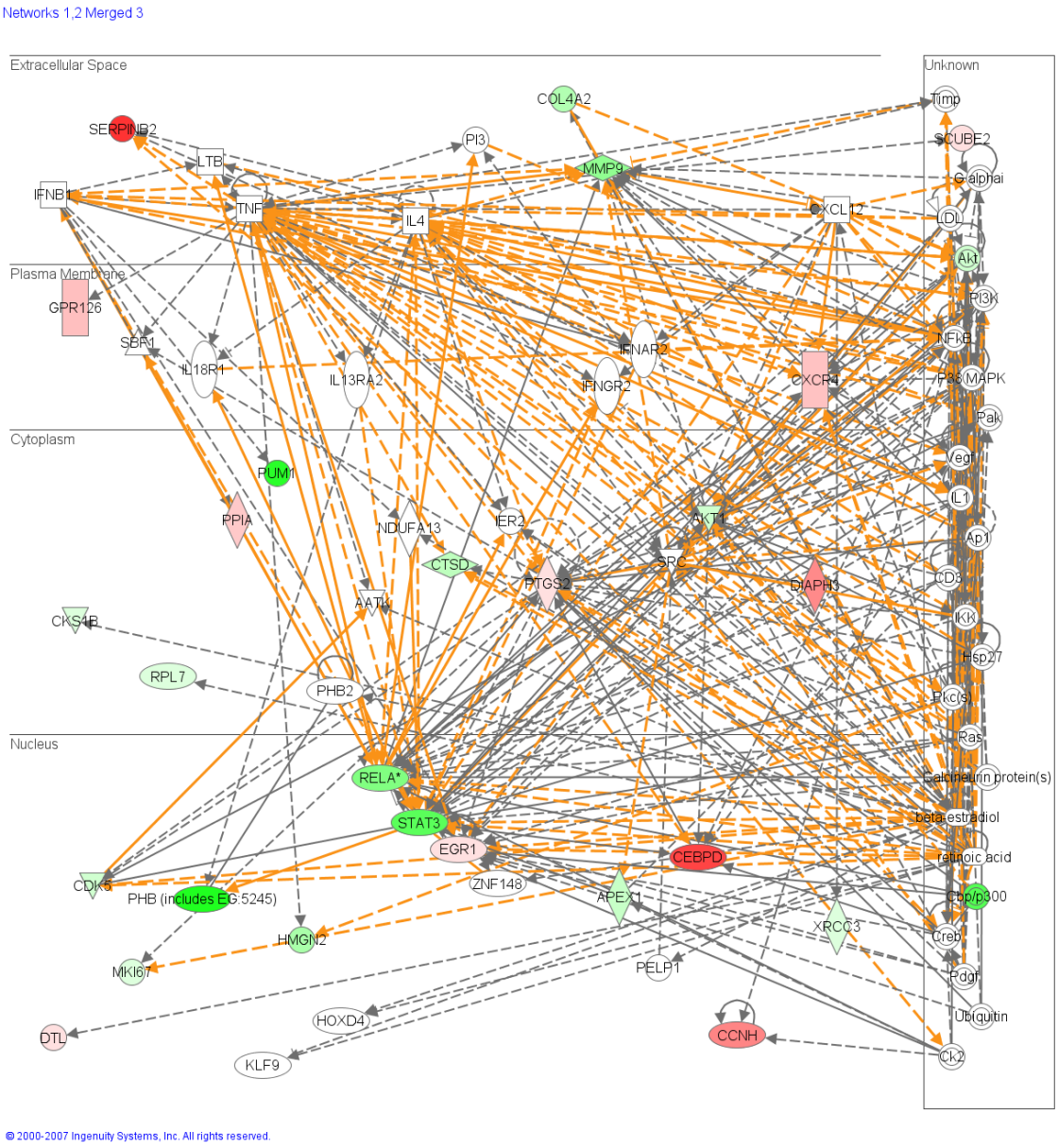


Figure 52: Merged network in a subcellular view of most significant genes of predicted subtype 1.

Analysis of PAM subtype 2

Gene set analysis of this tumor phenotype overlaps significantly with a gene list of estrogen signaling (nominal p-value of 0.017), the Sorlie's Luminal A phenotype list (nominal p-value of 0.040 and the Van't Veer good prognosis signature, a gene set which is called "BRCA_ER_positive_signal" (**Figure 53**).

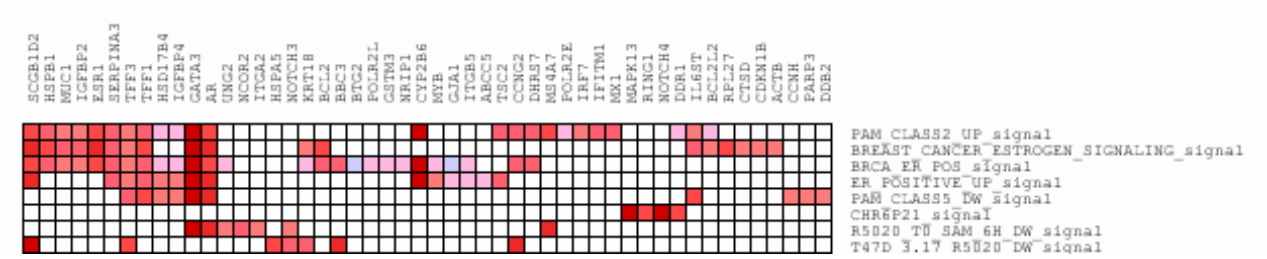


Figure 53: Leading-edge Analysis of the GSEA results of the PAM phenotype 2. Genes found distinctive from PAM subtype 2 are aligned on the horizontal axis. Found similar gene lists from other studies are on the vertical axis. Red boxes are marked the overlaps between the different gene sets showing genes that are found over-expressed on those studies.

The analysis with IPA gave to their distinctive gene signature of that tumor phenotype a top scoring functional molecular categories which includes 37 significant molecules which were previously found in cancer (p-value of 9.11E-28). The top molecular and cellular functions are cell growth and proliferation with 31 molecules being involved, and cell signaling with 28 molecules being represented. Top canonical pathways found are aryl hydrocarbon receptor signaling, Erk/MAPK pathway, IL-6 signaling, and IGF-1 signaling.

Pathway analysis gave 3 most relevant networks. The most significant with 21 molecules (**Figure 54**) gave canonical signaling pathways strongly activated as the Erk/MAPK pathway with genes up-regulated such as ESR1, DDR1, HSPB1 and STAT3; IGF-1 signaling pathway with genes being up-regulated such as IGFBP1 and IGFBP4, but IGFBP2 down-regulated; IL-6 signaling with genes such as HSPB1, IL6ST, and STAT3; and glucocorticoid receptor signaling (PI3K, RAS, STAT3, and TGFB) and JAK/STAT signaling (PI3K, RAS, STAT, and STAT3).

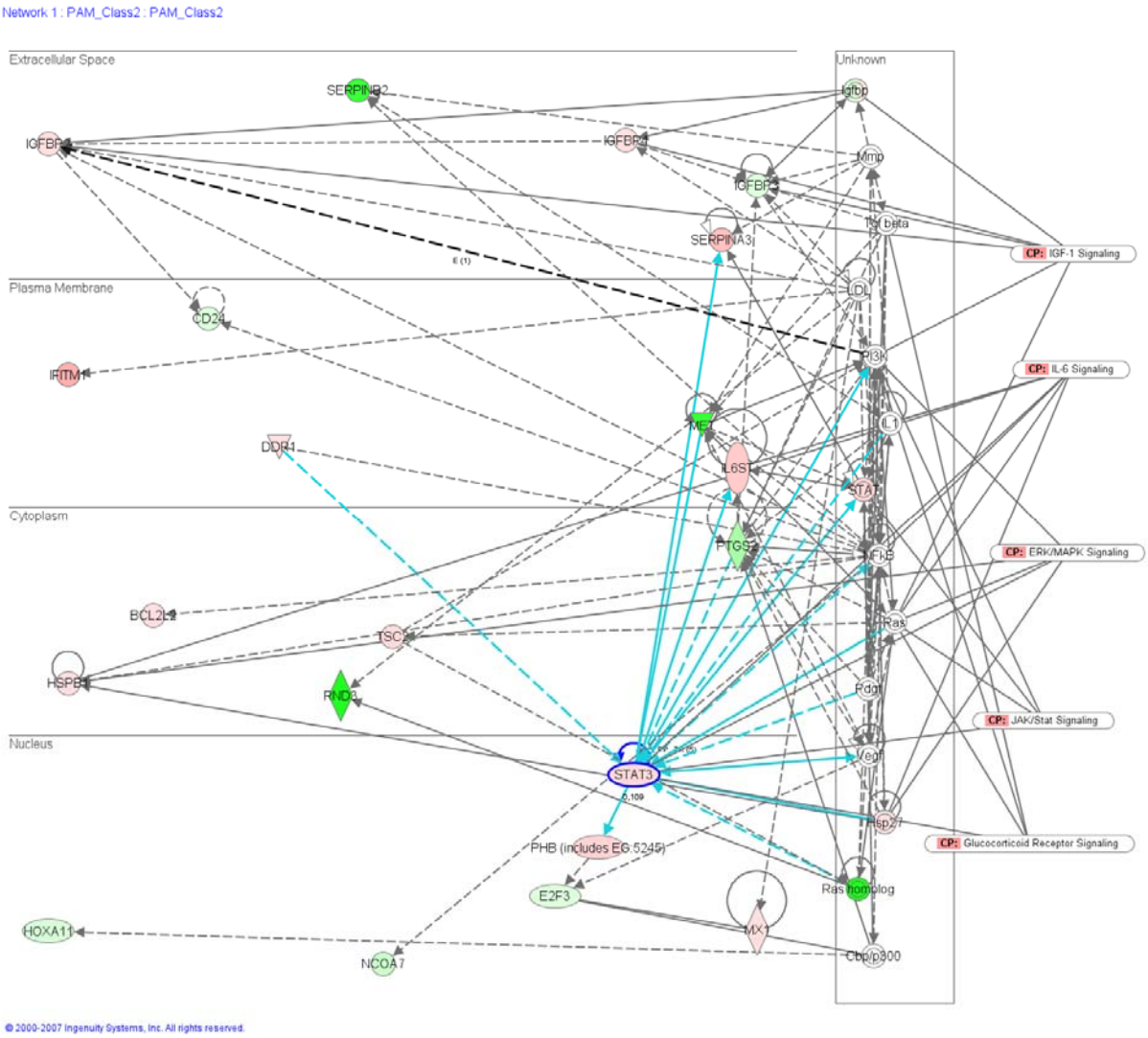


Figure 54: Network 1 of phenotype 2 in a subcellular view, in which the most significant canonical pathways are marked**.**

A second relevant network with a high score since it contains 30 genes from our gene list as it is shown in **Figure 55**. The score is based on a p-value calculation, which calculates the likelihood that the Network Eligible Molecules that are part of a network are found therein by random chance alone. A merged network 1 and 2 of phenotype 2 is shown in **Figure 56** in a subcellular view.

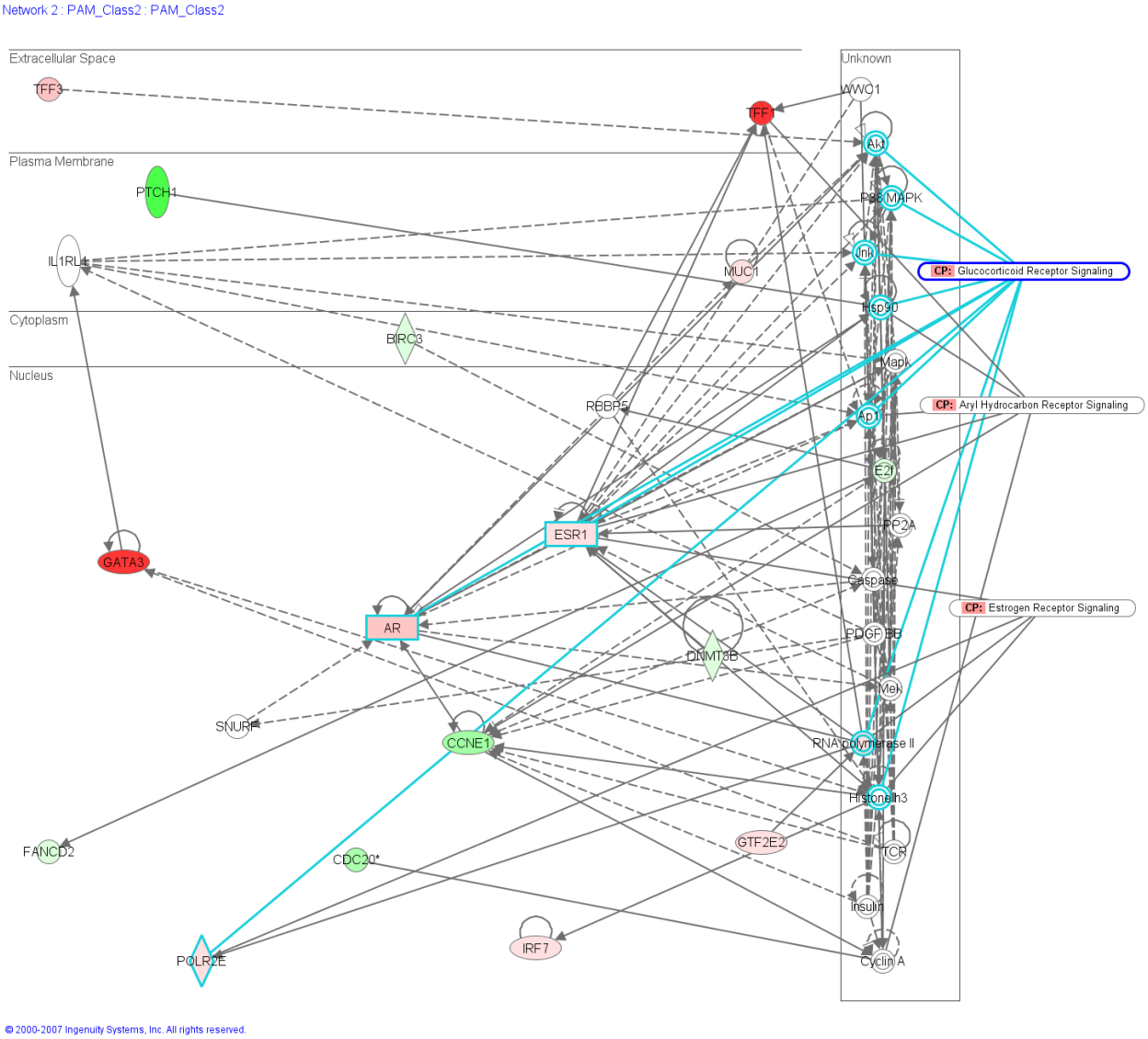


Figure 55: Network 2 of phenotype 2 in a subcellular view, in which to most significant canonical pathways are added, such as estrogen receptor signaling, glucocorticoid receptor signaling and the aryl hydrocarbon receptor signaling**.**

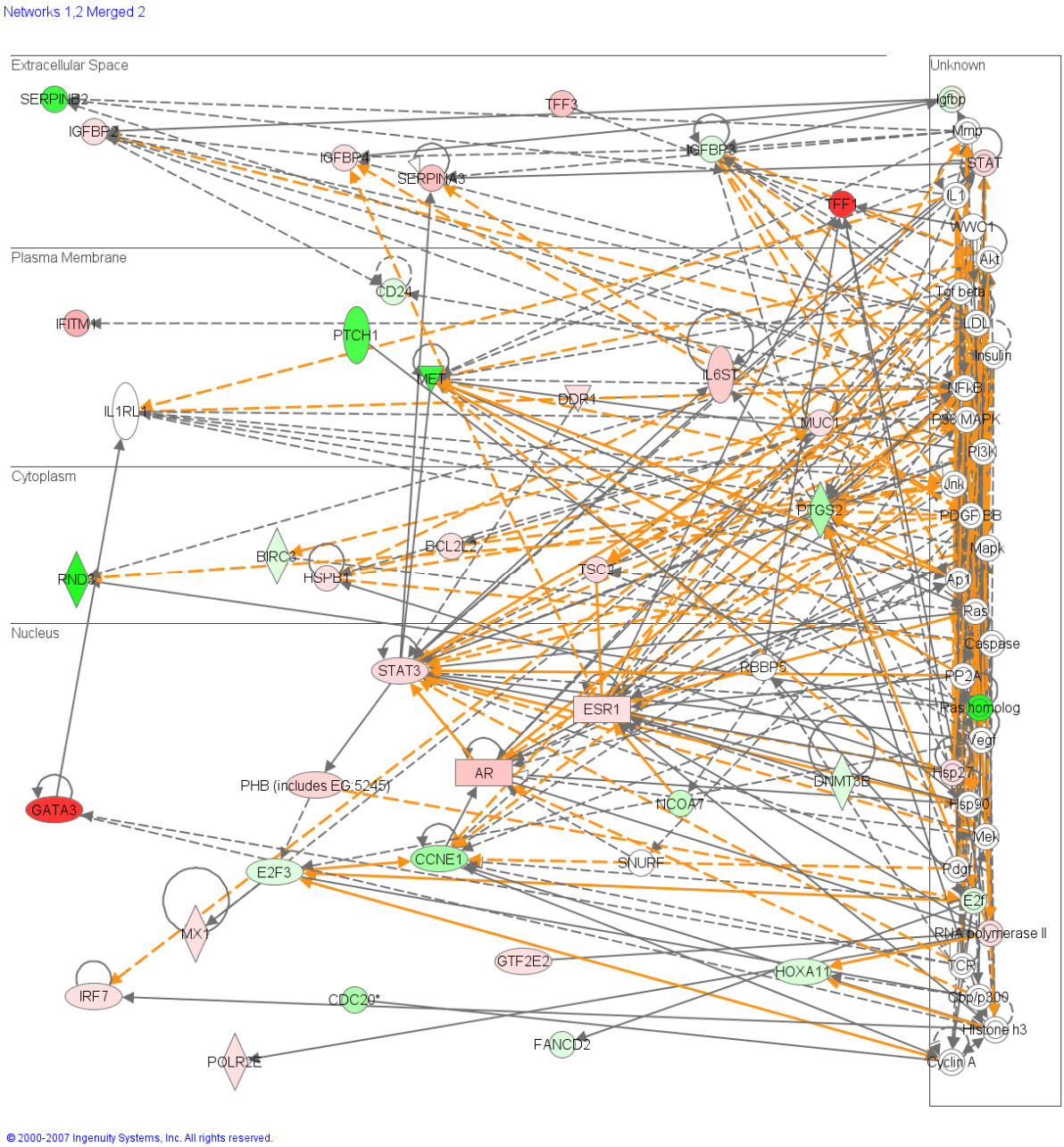


Figure 56: Merged networks 1 and 2 of phenotype 2 in a subcellular view**.**

Analysis of PAM subtype 3

Ingenuity pathway analysis shows that this tumor subtype gene signature contains 12 significant molecules which were previously found in cancer (*p*-value of 1.95E-11). The top molecular and cellular functions are cell cycle (12 genes involved), gene expression (9), cell growth and proliferation (9), cellular development (8), and cell-cell signaling (8). Top canonical signaling pathways found significant are G1/S cell cycle checkpoint regulation with 3 genes involved: a positive expression of CCND2 and HDAC4, and negative expression of CCNE2; aryl hydrocarbon receptor signaling (CCND2 and CCNE2); PDGF signaling (CAV1, caveolin); and FGF signaling (FGF7). This tumor subtype has

two high scoring networks which overlap in two molecules (DUSP1 and CAV1). The first network (**Figure 57**) is driven to cellular growth and proliferation containing 10 molecules of our gene list of this tumor subtype. It is observed at the diagram that cell proliferation could be driven extracellularly by the fibroblast growth factor FGF7, an EGF-like molecule, interacting downstream with the transcription regulators CCND2 and EGR1. DUSP1 is known to take part in the glucocorticoid receptor pathway and act as negative feedback regulator of JNK/STAT and p38 MAPK signaling pathway (Amit *et al.* 2007). This phenotype of tumors partially resembles the “core serum response” described by Chang *et al*. (2004). Plasma membrane associated ER interacts with CAV1 which plays an important role in E2 induced signal transduction. Phosphorylation of caveolin-1 forces *caveolae* to leave the plasma membrane, thereby decreasing the amount of plasma membrane-associated caveolin-1. This loss of caveolin/*caveolae* activates the signal cascade that triggers cell proliferation (Kiss *et al*. 2005).

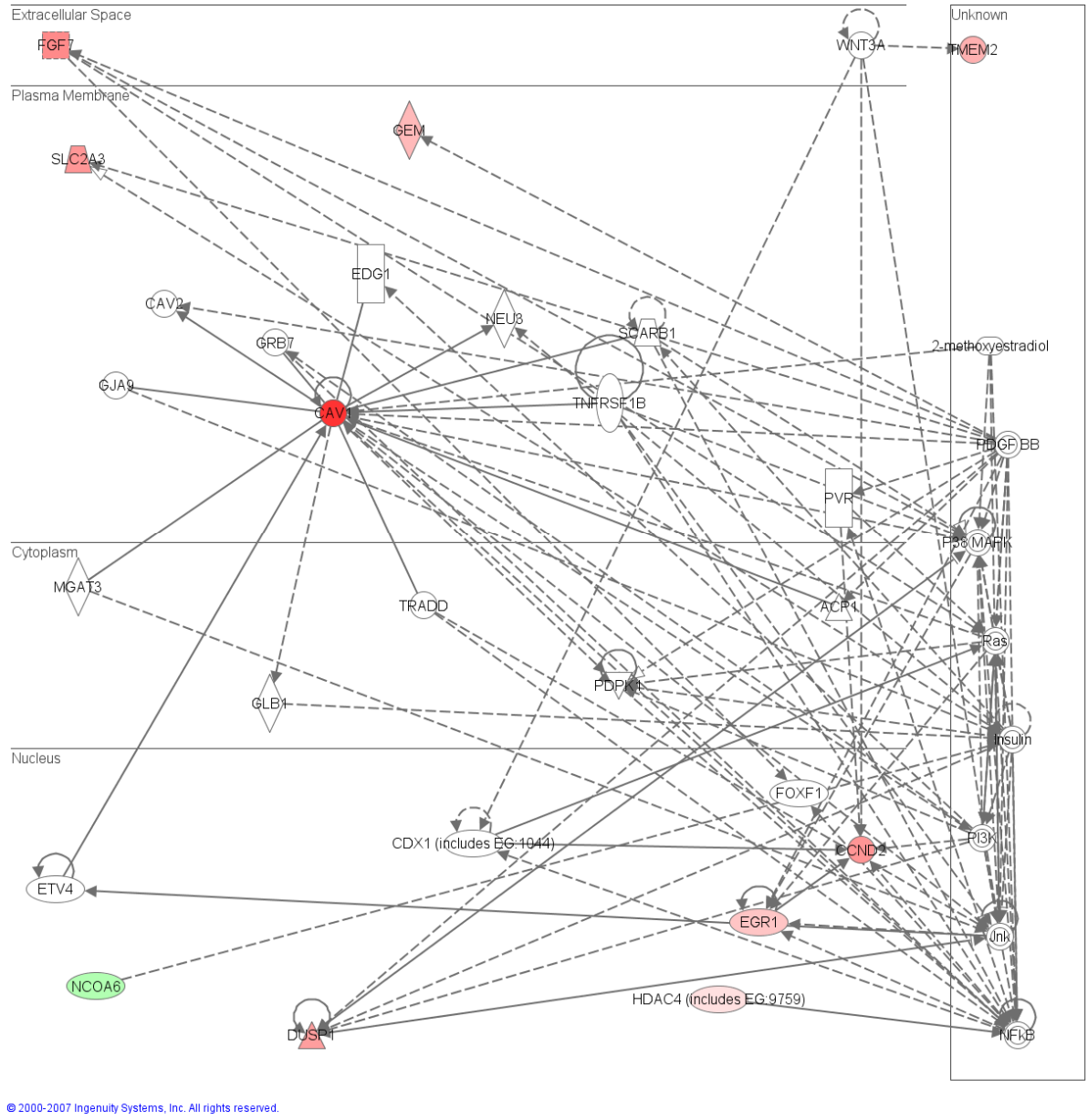


Figure 57: Network 1 of phenotype 3 in a subcellular view, of the cell growth and proliferation signaling is driven extracellular by FGF7.

GSEA analysis shows an overlap with a gene list of down-regulated genes after hormone induction by means of progestin on breast cancer model cell line T47D, and with early induced genes 1 hr after E2 treatment such as MYC, EGR1, WISP2, STAT5A, CDKN1C, ITGA5, IGFBP1 and IGFP3 (NOM p-value = 0.045). Also there is an overlap with a gene set of MYC target genes such as CCND2, CCNE1, DUSP1, CDKN1A, CDK4, CDKN2B, CCNA2, FN1, APEX1, HSPA4, and MYC itself (NOM p-value = 0.045). Leading edge analysis of this tumor subtype is shown in **Figure 58**.

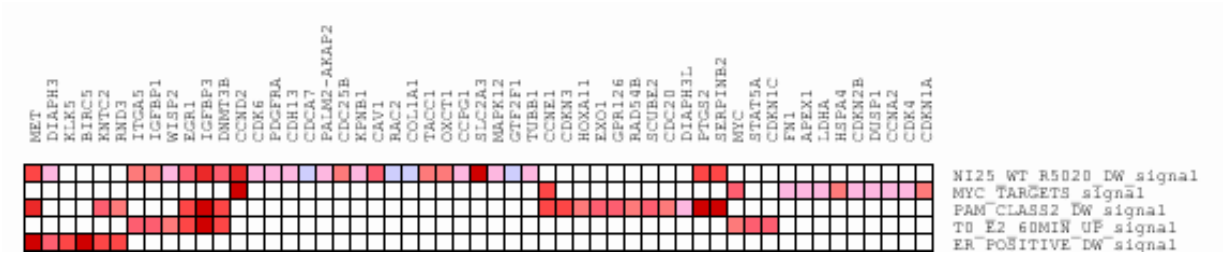


Figure 58: Leading edge analysis of tumor subtype 3**.**

If the two networks are merged, the obtained figure is shown in **Figure 59**.

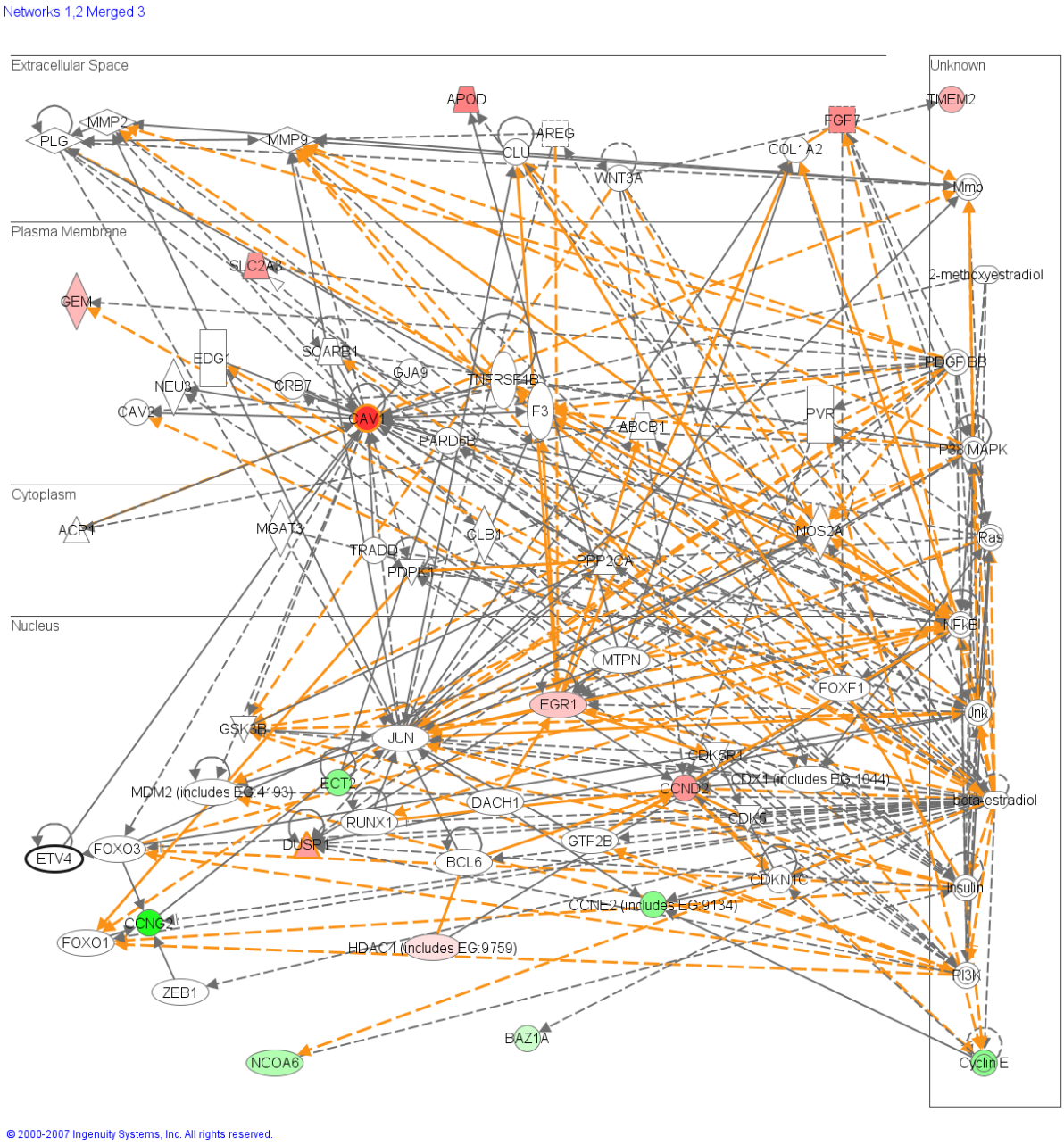


Figure 59: Merged network 1 and 2 of phenotype 3 in a subcellular view**.**

Analysis of PAM subtype 4

This tumor subtype consists of the know ERBB2+ like tumors. 3 of the 5 cases in this class had HER-2 gene amplification by fluorescence *in situ* hybridization analysis. The most discriminating genes of this subtype are genes mainly involved in cell cycle progression and proliferation in breast cancer such as ERBB2, SMARCE1, TOP2A, SMARCA4, IGFBP2, and PPARBP (MED1), involved in cell signaling and activating Erk/MAPK signaling through GRB2 and the transcription modulator DUSP6 which acts as a repressor of Erk. The network obtained has a score of 40, and connects 14 genes of our list of 15 discriminant genes of the PAM predicted ERBB2+ tumor subtype (**Figure 60**). The most represented pathways are the estrogen receptor pathway, the glucocorticoid receptor pathway and the Erk/MAPK pathway. There is some evidence for the interaction of growth factor signaling (EGF, ERBB2) and steroid hormone pathway in controlling the growth of breast cancer cells. It seems that there is an apparent cross-talk between these growth regulatory systems (Wilson and Slamon 2005). ERBB2 overexpression is associated to resistance to Tamoxifen and could directly modulate ER levels (Arpino *et al.* 2005). As it is shown in the diagram, ERBB2 directly interacts with DUSP6 which act as a specific transcriptional repressor of Erk (Amit *et al*. 2007). All these findings have led to the hypothesis that involves peptide growth factor pathways as possible mediators of the steroid hormone-independent phenotype in some human breast cancers, where peptide hormone pathways are replacing, in part, the steroid hormone pathways in regulating growth for these tumors.

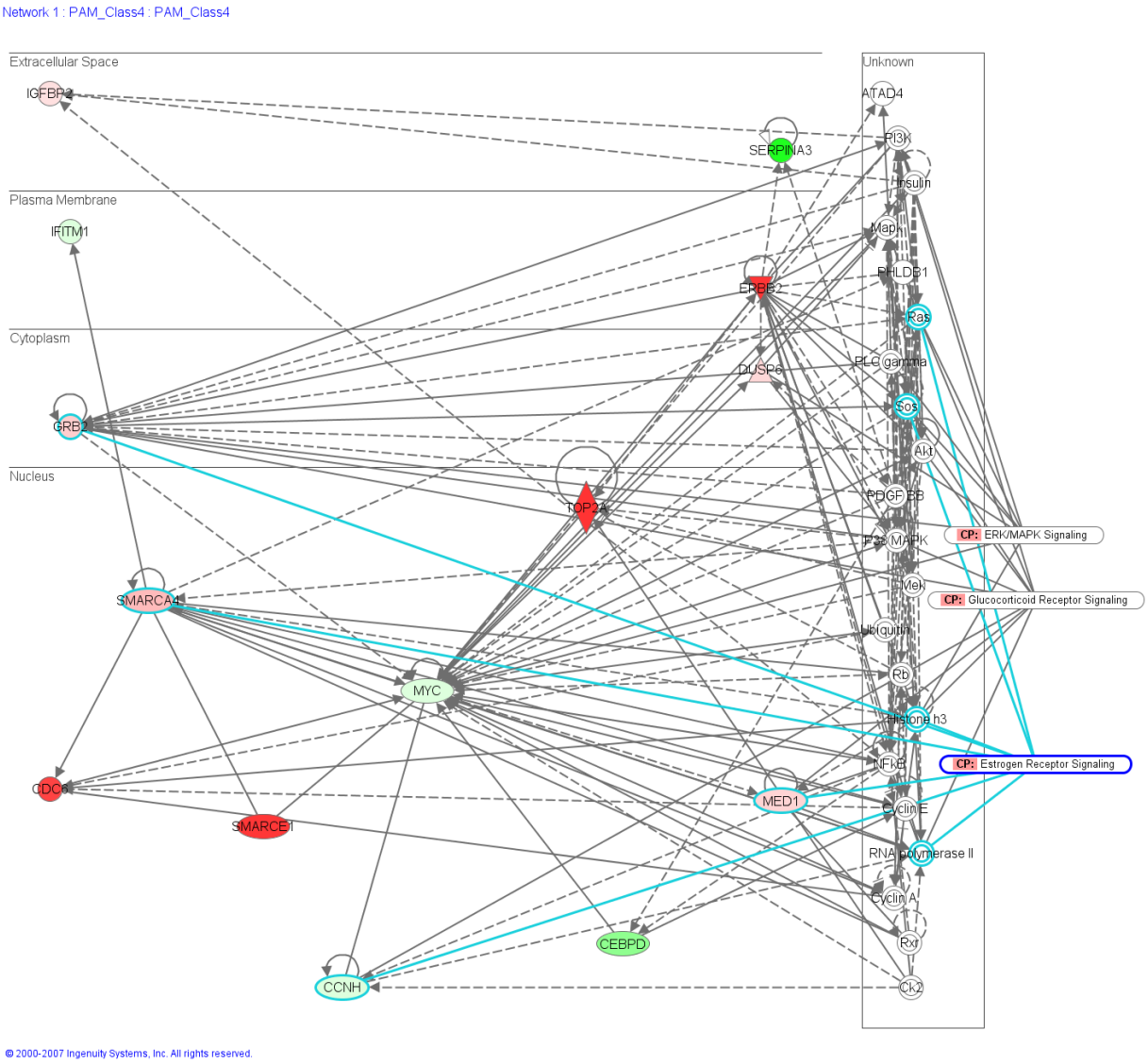


Figure 60: Network 1 phenotype 4, ERBB2+-like, which shows how most of the discriminating genes (14/15) are connected**.**

GSEA analysis shows an overlap with Van't Veer negative prognosis gene list (called "BRCA_PROGNOSIS_NEG_signal", nominal p-value = 0.006) with 18 genes (**Figure 61**), being highly expressed genes such as AURKA, MCM6, ECT2, NMU, DIAPH3, MAD2L1, CCNB2, and MELK. Other gene lists are genes of a serum fibroblast signature (NOM p-value = 0.011), and a set of down-regulated genes after hormonal induction by progestins using another model cell line T47y which does not express PR endogenously but was transfected with a vector containing PR carrying a point mutation on the ERID domain, the interaction domain with ERα, and therefore has lost its ability to crosstalk and activate Erk1/2 signaling pathway (Ignacio Quiles, pH Doctoral Thesis, University of Pompeu Fabra, CRG, Barcelona).This indicates that genes that are usually induced by progestins are by introducing a mutation on the ERID domain down-regulated, and overlap in part with the discriminant genes of the PAM subtype 4.

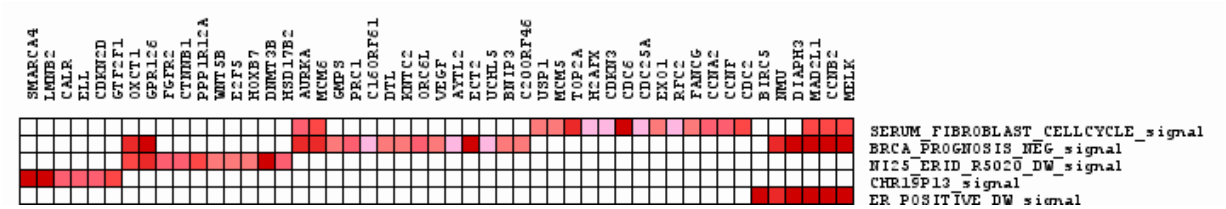


Figure 61: Leading edge analysis of ERBB2+ tumor phenotype.

Analysis of PAM subtype 5

This tumor subtype is the characteristic basal-like subtype characterize with high expression of keratins 5A and 5B and showed also high expression of other basal epithelial genes as KIT1 and ID4. These are also termed as “triple-negative tumors” as they are found negative for ER, PR and ERBB2. As we mentioned earlier, 9 out of 12 samples are p53+ and have high histological grade (10/12).

Ingenuity analysis shows that the genes most discriminating for this tumor subtype are mainly involved in cancer (36 of them, obtaining a p-value of 2.17E-30. The molecular and cellular functions of this subtype are cell death, cell cycle, cellular growth and proliferation, and cell signaling. Pathway analysis gives three significant networks.

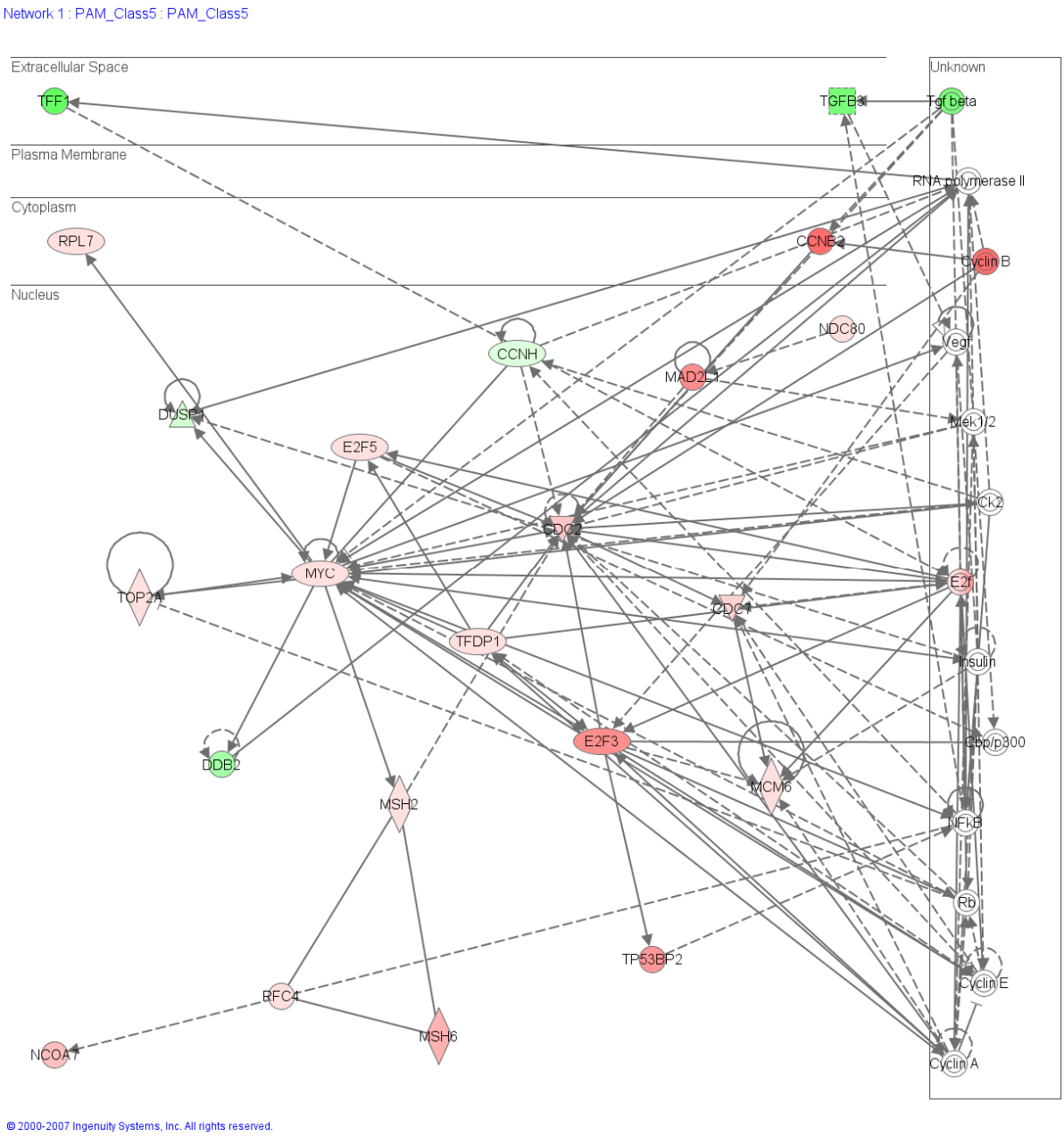


Figure 62: Network 1 phenotype 5, basal-like-like**.**

A first network involves 21 molecules of our list, is shown at **Figure 62.** ER signaling pathway is inactive, ER regulated genes TFF1, FOS, TGFB3 and CCNH are down-regulated, and now, there is high expression of genes involved in G1/S mitotic cell phase with transcription factors being up-regulated such as E2F3, E2F5, MYC, TFDP1, promoting cell cycle progression to mitosis, indicative of high proliferative tumors. Erk/MAPK pathway is also inactive with dual specificity phosphatases DUSP1 and DUSP6, TEK tyrosine kinase and FOS down-regulated in this tumor subtype.

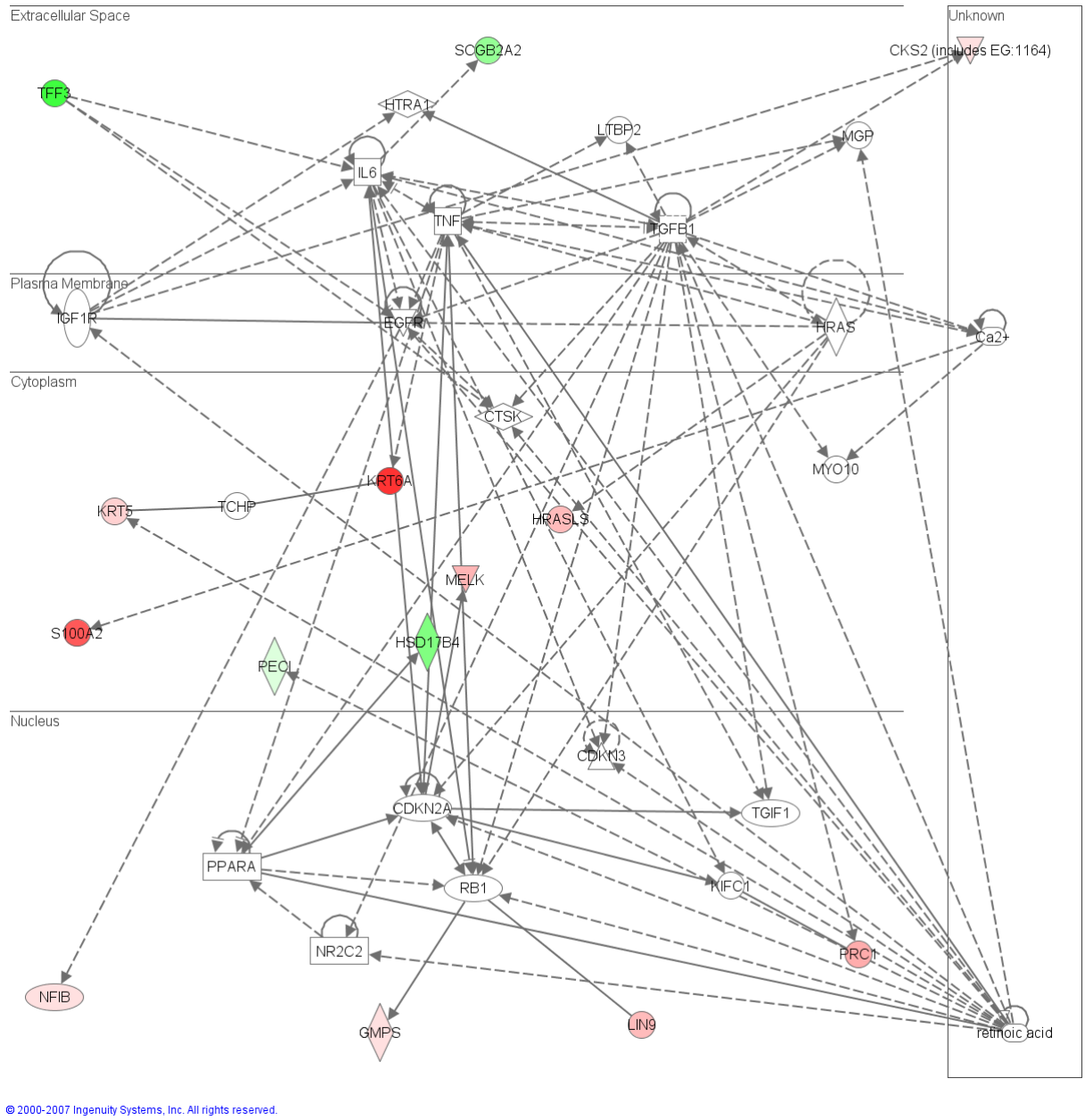


Figure 63: Network 2 phenotype 5, basal-like-like.

On a second significant network there is high expression of basal/myoepithelial genes such as KRT5/6 and calcium binding protein such as S100A (**Figure 63**).

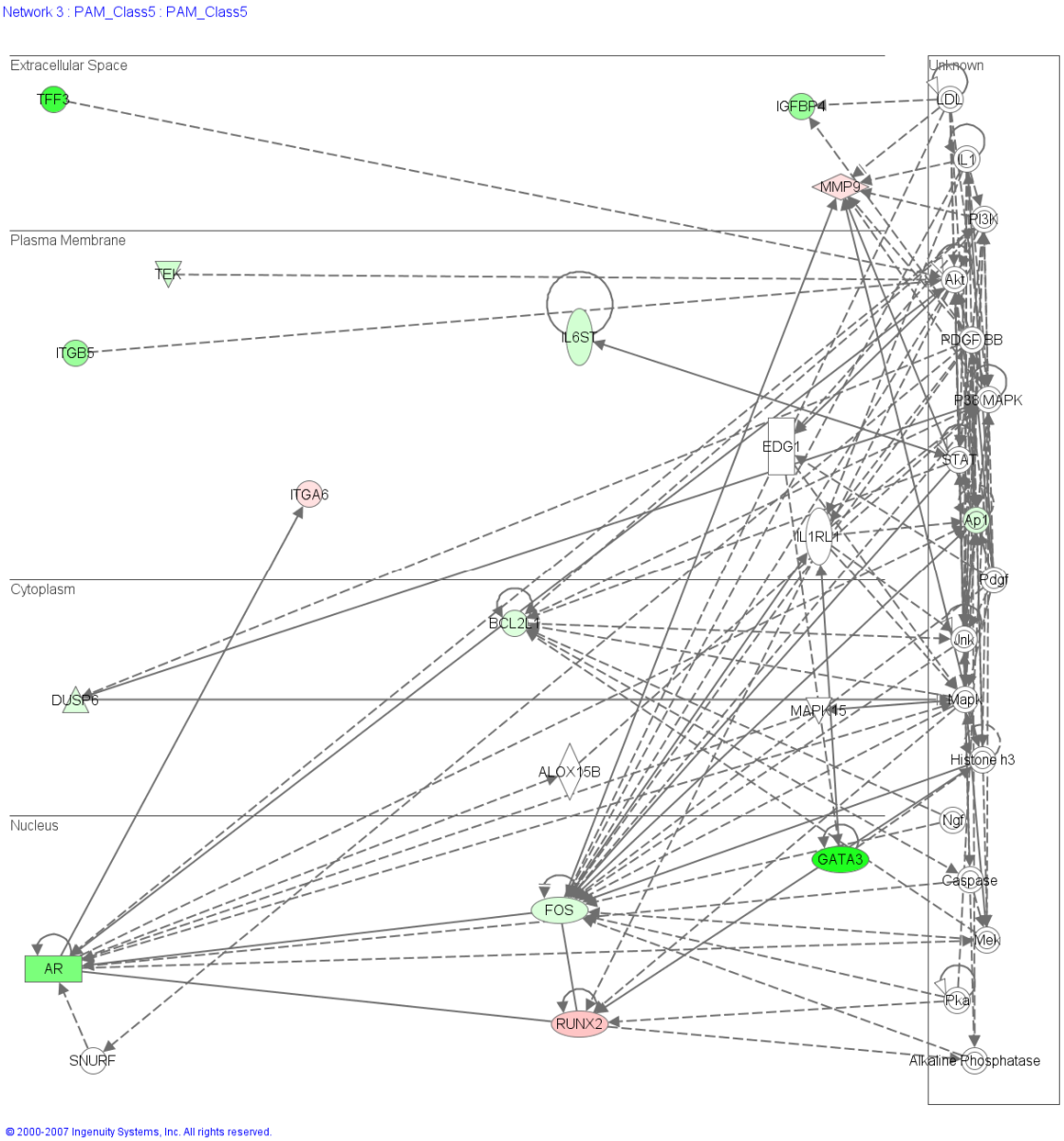


Figure 64: Network 3 phenotype 5, basal-like-like.

On the third significant network by IPA, ER signaling pathway is also shown inactive in this tumor phenotype, as many genes typically regulated by ER or co-expressed by ER are down-regulated such as TFF3, ILST6, GATA3, FOS, and AR. (**Figure 64**).

Figure 65: Merged networks of PAM predicted phenotype 5, basal-like breast tumors.

Merging the three networks (**Figure 65**), a gene expression picture of this tumor subtype is obtained where the ER signaling pathway is shut off, and a set of transcription factors such as MYC, RUNX2, TFDP1, E2F3 activate CCD2 and CCNB2, inducing the breast cancer cell to proliferate.

GSEA analysis found significant the list of Sorlie's basal-like subtype (called "BRCA_ER_negative_signal") with an FDR q-value of 0.101, nominal p-value of 0.006), Van't Veer poor prognosis (FDR q-value of 0.209, nominal p-value of 0.004), and Van't Veer breast cancer outcome good versus poor (FDR q-value of 0.464, nominal p-value of 0.064). Leading edge analysis of most relevant results of GSEA is shown in **Figure 66**.

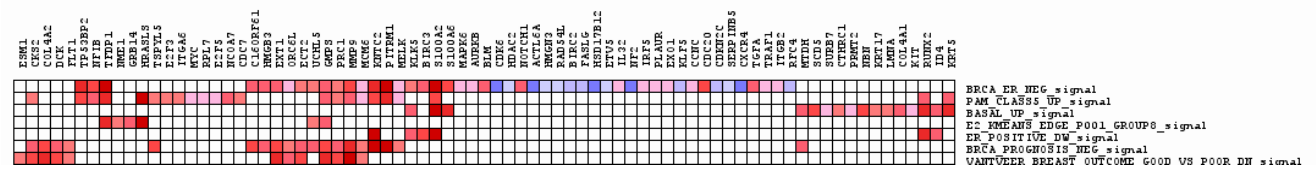


Figure 66: Leading edge analysis of GSEA result of PAM predicted subtype 5**.**

4.11.4 Analysis of the test or validation set

Using the 150 centroid genes used to classify into five breast cancer subtypes described above, a classification was performed of a new incoming batch of breast biopsy samples using Pearson correlation coefficient to centroids to assign each tumor sample into a predicted phenotype.

Since in this new batch of breast biopsies there were also included 3 samples of normal breast tissue taken from a tissue adjacent to a tumor, it was also used a classifier adding the centroid resulting from normal breast samples. The training set had only 3 normal samples, but one of them was not consistent by unsupervised clustering and was aligned with tumor samples (T70).

Selecting a threshold Δ of 1.31, breast samples were assigning to a subtype centroid by means of Pearson correlation coefficient as the probability to belong to that class. Probabilities of the test set of new incoming samples are shown in **Figure 67**.

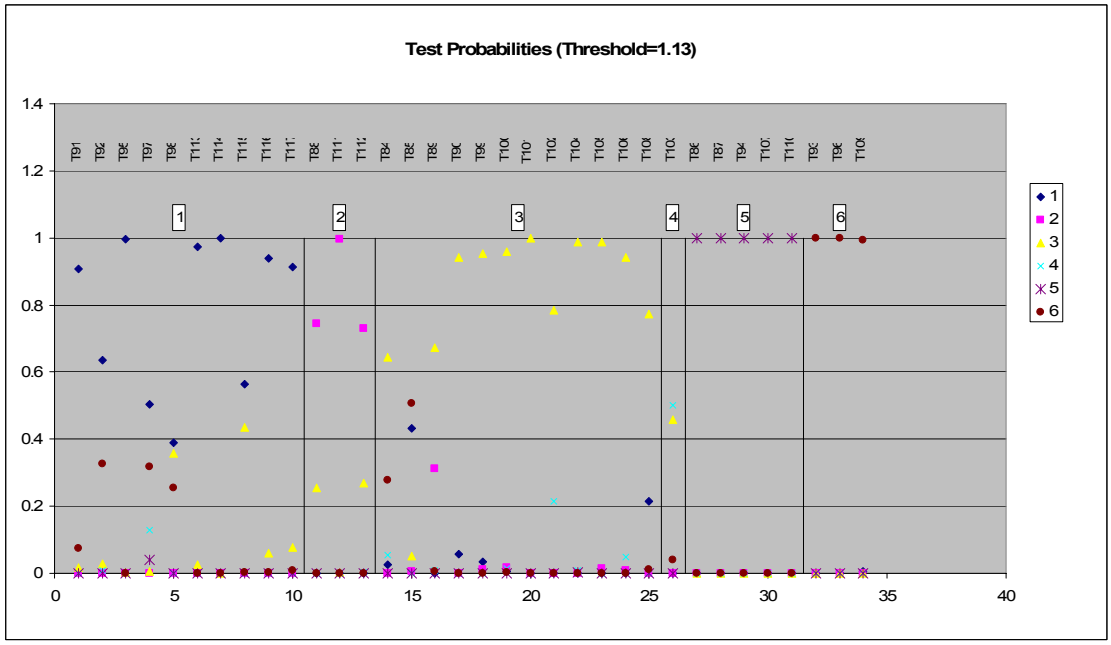


Figure 67: Class assignment and probabilities of the predicted test set**.**

In **Table 14** the predicted probability is shown together with the clinical and histopathological patient data showing good correlation with molecular markers. Normal breast samples are correctly assigned. Triple-negative samples, those samples which are negative for ER, PR, and HER-2, are assigned with high probability to the poor prognosis basal-like subtype 5. Only one sample is left unclassified, since probability of being assigned to a tumor subtype did not reach a 50%. Grey cells means data not yet available.

Table 14: Test set probabilities and clinical histopathological data.

		Predicted Probabilities						Clinical and Histopathological data								
Array ID	Predicted Class	1	2	3	4	5	6	Patient ID	ER	PR	HER-2	p53	Histological Grade	Tumor size	Lymph node status	Metastatic sites
T84	3	0.02	0.00	0.64	0.05	0.00	0.28	R15301	POS	NEG	3+++	NEG	HG2	T1	N0	M0
T85	3	0.24	0.00	0.76	0.00	0.00	0.00	R15302	POS	POS	NEG- POLISOMICO	NEG	HG2	T2	N0	M0
T86	5	0.00	0.00	0.00	0.00	1.00	0.00	R15303	NEG	NEG	NEG	POS	HG3	T1	N0	M0
T87	5	0.00	0.00	0.00	0.00	1.00	0.00	R15304	NEG	NEG	NEG	POS	HG3	T1	N0	M0
T88	2	0.00	0.75	0.25	0.00	0.00	0.00	R15305	POS	POS	NEG	NEG	HG2	T1	N1	M0
T89	3	0.00	0.31	0.67	0.01	0.00	0.01	R15306	POS	POS	NEG	NEG	HG1	T1	N0	M0
T90	3	0.06	0.00	0.94	0.00	0.00	0.00	R15307	POS	POS	NEG	NEG	HG1	T1	N0	M0
T91	1	0.91	0.00	0.02	0.00	0.00	0.07	R15308	POS	NEG	NEG	NEG	-	-	-	-
T92	1	0.64	0.00	0.03	0.01	0.00	0.33	R15309	NEG	NEG	3+++	NEG	-	-	-	-
T93	6	0.00	0.00	0.00	0.00	0.00	1.00	R15326	NORMAL BREAST							
T94	5	0.00	0.00	0.00	0.00	1.00	0.00	R15328	NEG	NEG	NEG	NEG	HG3	T1	N1	M0
T95	1	1.00	0.00	0.00	0.00	0.00	0.00	R15330	POS	POS	NEG	NEG	-	T2	N0	M0
T96	6	0.00	0.00	0.00	0.00	0.00	1.00	R15331	NORMAL BREAST							
T97	1	0.50	0.00	0.01	0.13	0.04	0.32	R15332	POS	POS	3+AMPLIFICADO FISH	NEG	HG3	T3	N1	M0
T98	Unclass	0.39	0.00	0.36	0.00	0.00	0.25	R15333	POS	POS	NEG	NEG	HG2	T2	N2	M0
T99	3	0.03	0.01	0.95	0.00	0.00	0.00	R15334	POS	POS	NEG	NEG	HG1	T1	N1	M0
T100	3	0.01	0.02	0.96	0.02	0.00	0.00	R15335	POS	POS	NEG	NEG	-	T2	N1	M0
T101	3	0.00	0.00	1.00	0.00	0.00	0.00	R15336	POS	POS	NEG	NEG	HG1	-	-	-
T102	3	0.00	0.00	0.79	0.21	0.00	0.00	R15371	POS	POS	2+AMPLIFICADO	NEG	HG3	-	-	-
T103	4	0.00	0.00	0.46	0.50	0.00	0.04	R15372	POS	NEG	NEG 1+	POS	HG3	T1	N0	M0
T104	3	0.00	0.00	0.99	0.01	0.00	0.00	R15374	NEG	NEG	3+AMPLIFICADO FISH	NEG	HG3	-	-	-
T105	3	0.00	0.01	0.99	0.00	0.00	0.00	R15375	POS	POS	NEG	POS	-	-	-	-
T106	3	0.00	0.01	0.94	0.05	0.00	0.00	R15376	POS	POS	NEG	NEG	-	-	-	-
T107	5	0.00	0.00	0.00	0.00	1.00	0.00	R15378	NEG	NEG	NEG	POS	HG3	-	-	-
T108	3	0.22	0.00	0.77	0.00	0.00	0.01	R15398	POS	POS	NEG	NEG	HG2	-	-	-
T109	6	0.01	0.00	0.00	0.00	0.00	0.99	R15399	NORMAL BREAST							
T110	5	0.00	0.00	0.00	0.00	1.00	0.00	R15400	NEG	NEG	NEG	POS	HG3	-	-	-
T111	2	0.00	1.00	0.00	0.00	0.00	0.00	R15401	POS	POS	NEG	-	HG2	-	-	-
T112	2	0.00	0.73	0.27	0.00	0.00	0.00	R15402	POS	POS	NEG	NEG	-	-	-	-
T113	1	0.97	0.00	0.03	0.00	0.00	0.00	R15404	POS	NEG	NEG	NEG	HG2	-	-	-
T114	1	1.00	0.00	0.00	0.00	0.00	0.00	R15405	POS	POS	NEG 1+	NEG	HG2	-	-	-
T115	1	0.56	0.00	0.44	0.00	0.00	0.00	R15406	POS	POS	NEG	NEG	HG2	-	-	-
T116	1	0.94	0.00	0.06	0.00	0.00	0.00	R15407	POS	POS	NEG	POS	HG3	-	-	-
T117	1	0.91	0.00	0.08	0.00	0.00	0.01	R15408	POS	POS	NEG	NEG	HG1	-	-	-

Clinical and histopathological characteristics of the all patients and their tumors of the PAM predicted groups in both training and test set are listed in **Table 15**.

Table 15: Clinical and histopathological characteristics of the patients and their tumors of the PAM predicted groups.

Clinical and histopathological characteristics of the patients of the PAM predicted groups

All patients (n = 105)		PAM 1 (n=21)	PAM 2 (n=29)	PAM 3 (n=31)	PAM 4 (n=6)	PAM 5 (n=17)
Age (years)	Number of cases (%)	Number of cases	Number of cases	Number of cases	Number of cases	Number of cases
≤ 40	5 (4.8%)	1 (4.8%)	2 (6.9%)	1 (3.2%)		
> 40 and ≤ 50	11 (10.5%)	2 (9.5%)	4 (13.8%)	3 (9.7%)		
> 50 and ≤ 60	30 (28.6%)	6 (28.6%)	4 (13.8%)	9 (29.0%)	2 (33.3%)	6 (35.3%)
> 60 and ≤ 70	27 (25.7%)	6 (28.6%)	7 (24.1%)	5 (16.1%)	3 (50%)	1 (5.9%)
> 70	32 (30.5%)	4 (19.0%)	11 (37.9%)	6 (19.4%)	1 (16.7%)	5 (29.4%)
Therapy	Number of cases (%)	Number of cases	Number of cases	Number of cases	Number of cases	Number of cases
HT only	21 (20%)	5 (23.8%)	9 (31.0%)	8 (25.8%)		
QT only	26 (24.8%)		3 (10.3%)	8 (25.8%)	2 (33.3%)	13 (76.5%)
HT + QT	38 (36.2%)	15 (71.4%)	17 (56.6%)	10 (32.3%)	3 (50%)	
neoadjuvant QT	3 (2.9%)					3 (17,6%)
HT + QT + Herceptin	1 (1%)	1 (4.8%)		2 (6.5%)		
none	5 (4.8%)			2 (6.5%)	1 (16.7%)	1 (5.9%)
Tumor size, cm	Number of cases (%)	Number of cases	Number of cases	Number of cases	Number of cases	Number of cases
1	51 (48.6%)	9 (42.9%)	17 (56.6%)	14 (45.2%)	2 (33.3%)	9 (52.9%)
2	42 (40%)	10 (47.6%)	8 (27.6%)	15 (48.4%)	4 (66.7%)	4 (24.5%)
3	9 (8.6%)	2 (9.5%)	2 (6.9%)	1 (3.2%)		4 (24.5%)
4	2 (1.9%)		2 (6.9%)			
isquemic	1 (1%)			1 (3.2%)		
Lymph node	Number of cases (%)	Number of cases	Number of cases	Number of cases	Number of cases	Number of cases
none	60 (57.1%)	11 (52.4%)	18 (62.1%)	20 (64.5%)	4 (66.7%)	8 (47.1%)
1	29 (27.6%)	7 (33.3%)	9 (31.0%)	4 (12.9%)		9 (52.9%)
2	8 (7.6%)	1 (4.8%)	1 (3.4%)	5 (16.1%)		
3	6 (5.7%)	2 (9.5%)	1 (3.4%)	2 (6.5%)	1 (16.7%)	
micro	1 (1%)				1 (16.7%)	
Metastatic sites	Number of cases (%)	Number of cases	Number of cases	Number of cases	Number of cases	Number of cases
none	101 (96.2%)	21 (100%)	29 (100%)	29 (93.5%)	6	15 (88.2%)
one	4 (3.8%)			2 (6.5%)		2 (11.8%)
Vascular invasion	Number of cases (%)	Number of cases	Number of cases	Number of cases	Number of cases	Number of cases
yes	30 (28.6%)	7 (33.3%)	10 (34.5%)	7 (22.6%)	5	4 (23.5%)
none	74 (70.5%)	14 (66.7%)	19 (65.5%)	23 (74.2%)	1 (16.7%)	11 (64.7%)
Histological grade	Number of cases (%)	Number of cases	Number of cases	Number of cases	Number of cases	Number of cases
1	17 (16.2%)	6 (28.6%)	3 (10.3%)	8 (25.8%)		
2	30 (28.6%)	8 (38.1%)	12 (41.4%)	7 (22.6%)	1 (16.7%)	1 (5.9%)
3	33 (31.4%)	3 (14.3%)	4 (13.8%)	9 (29.0%)	4	13 (76.5%)
unknown	25 (23.8%)	4 (19.0%)	10 (34.5%)	7 (22.6%)	1 (16.7%)	3 (17.6%)
Recurrence	Number of cases (%)	Number of cases	Number of cases	Number of cases	Number of cases	Number of cases
free of disease (VL)	58 (55.2%)	11 (52.4%)	24 (82.8%)	13 (41.9%)	4 (66.7%)	5 (29.4%)
with disease (VE)	5 (4.8%)	2 (9.5%)	1 (3.4%)	3 (9.7%)		1 (5.9%)
exitus	10 (9.5%)	1 (4.8%)	1 (3.4%)	2 (6.5%)		6 (35.3%)
unknown	31 (29.5%)		3 (10.3%)	14 (45.2%)	1 (16.7%)	5 (29.4%)
no follow-up	2 (1.9%)			1 (3.2%)	1 (16.7%)	
p53 status	Number of cases (%)	Number of cases	Number of cases	Number of cases	Number of cases	Number of cases
negative	70 (66.7%)	20 (95,2%)	20 (69.0%)	23 (74.2%)	3 (50%)	3 (17,6%)
positive	34 (32.4%)	1 (4.8%)	8 (27.6%)	8 (25.8%)	3 (50%)	14 (82.4%)
unknown	1 (1%)		1 (3.4%)			
Steroid receptor status	Number of cases (%)	Number of cases	Number of cases	Number of cases	Number of cases	Number of cases
ER+ PR+	56 (53.3%)	14 (66,7%)	23 (79.3%)	17 (54.8%)	1 (16.7%)	
ER+ PR-	15 (14.3%)	4 (19.0%)	5 (17.2%)	5 (16.1%)	1 (16.7%)	
ER- PR+	2 (1.9%)	2 (9.5%)				
ER- PR-	32 (30.5%)	1 (4.8%)	1 (3.4%)	9 (29.0%)	4 (66.7%)	17 (100%)
her-2 status (IHC/FISH)	Number of cases (%)	Number of cases	Number of cases	Number of cases	Number of cases	Number of cases
negative/negative	90 (85.7%)		28 (96.6%)	26 (83.9%)		16 (94.1%)
positive/amplified polysome	10 (9.5%)	2 (9.5%)	1 (3.4%)	4 (12.9%)	3 (50%)	1 (5.9%)
negative/amplified polysome	4 (3.8%)	2 (9.5%)		1 (3.2%)	1 (16.7%)	
positive/non-amplified polysome	1 (1%)				1 (16.7%)	

HT = hormone therapy; QT = chemotherapy: IHC = immunohistochemistry; FISH =fluorescence "*in situ*" hybridization

The more discriminating parameters of the basal-like subtype, PAM class 5, are the higher percentage of high grade tumors (76.5%) higher percentage of recurrence with *exitus* cases (35.55), higher percentage of p53 positive tumors (82.4%), all are ER-PR- (100%), and ERBB2- (94.1%).

4.12 Real Time qPCR assay for confirming results of breast tumor samples

To validate the class prediction of the test set, breast samples T84 to T117, it was performed real time qPCR assays with a set of genes selected from the top discriminatory genes from each predicted class. It was also included DUSP1 and DUSP6, for their role as negative feedback in growth factor signaling. DUSP1 preferentially inactivates JNK/p38 signaling, and DUSP6 is a repressor of Erk signaling pathway (Amit *et al*. 2007).

Primer pairs were design as explained in chapter 3.20.1. Reverse transcription of Total RNA samples were performed as stated in chapter 3.20.2, with a difference, which is that it was employed a mix of oligo-dT and random primers (hexamers) in a 1 to 1 proportion, that is, 50 pmoles of oligo dT and 100 ng/μl per transcription reaction, since our breast tumor samples are not intact but partially degraded.

List of primers for real time qPCR assays are listed at **Appendix A4**.

Assay efficiencies for each primer pair were calculated as in chapter 3.20.3, but using for these assays a pool of breast tumor samples, for taking in consideration the quality of the RNA besides of the primer pair efficiency (**Table 16**).

6 genes were included as endogenous controls to determine whether a single gene or a combination of various genes were suitable for normalization of the whole set. These genes were selected since they were previously selected for primary breast tumor samples (Szabo *et al*. 2004). Each sample was assayed by duplicate. It was included a “minus RT” control for checking any genomic DNA carry-over, and a “non-template” control for each primer pair for check for the specificity of the assay. As a result, the breast tumor “gene panel” consisted in 21 genes, and looked as in **Figure 68.**

Ct values from duplicates were averaged, and referred to the calibrator sample, which is UHRR, as it was used at microarray gene profiling analysis.

Table 16: Breast tumor samples gene panel.

Function	Symbol (alias)	RefSeq	Efficiency	Efficiency (%)
Endogenous	ACTN	NM_001102	1.855	92.75
Endogenous	GAPD	NM_002046	1.943	97.15
Endogenous	MRPL19	NM_014763	1.831	91.55
Endogenous	SF3A1	NM_005877	1.904	95.20
Endogenous	PUM1	NM_014676	1.98	99.00
Endogenous	PSMC4	NM_006503	1.97	98.50
Up-regulated in PAM Class 1	SERPINB2	NM_002575	1.806	90.30
Up-regulated in PAM Class 1	CEBPD	NM_005195	1.852	92.60
Up-regulated in PAM Class 2	TFF1 (pS2)	NM_003225	1.894	94.70
Up-regulated in PAM Class 2	GATA3	NM_002051	1.942	97.10
Up-regulated in PAM Class 2	SERPINA3	NM_001085	1.873	93.65
Up-regulated in PAM Class 3	CAV1	NM_001753	1.916	95.80
Up-regulated in PAM Class 3	APOD	NM_001647	1.989	99.45
Up-regulated in PAM Class 3	DUSP1	NM_004417	1.864	93.20
Up-regulated in PAM Class 4	TOP2A	NM_001067	1.97	98.50
Up-regulated in PAM Class 4	ERBB2	NM_004448	1.517	75.85
Up-regulated in PAM Class 4	SMARCE1	NM_003079	1.788	89.40
Up-regulated in PAM Class 4	SMARCA4	NM_003072	1.866	93.30
Up-regulated in PAM Class 4	DUSP6	BC037236	1.982	99.10
Up-regulated in PAM Class 5	KRT5	NM_000424	1.855	92.75
Up-regulated in PAM Class 5	S100A2	NM_005978	1.845	92.25

		ACTN	GAPD	MRPL19	SF3A1	PUM1	PSMC4	SERPINB2	CEBPD	TFF1	GATA3	SERPINA3	CAV1	APOD	DUSP1	TOP2A	ERBB2	SMARCE1	SMARCA4	DUSP6	KRT5	S100A2			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Sample 1	A																								
	B																								
Sample 2	C																								
	D																								
Sample 3	E																								
	F																								
Sample 4	G																								
	H																								
Sample 5	I																								
	J																								
Sample 6	K																								
	L																								
Sample 7	M																								
	N																								
Sample 8	O																								
	P																								

Figure 68: Gene panel the breast tumor samples for microarray data validation.

Three of the assayed tumor samples failed, T91, T92, and T93, since there was not enough Total RNA sample for reverse transcription.

Real time qPCR data was normalized applying geNORM (see chapter 3.20.5) to determine the more suitable set of genes for normalization of real time qPCR data. This algorithm gives as the most stable genes across all tumor samples, were MRPL19 and PSMC4, and the least stable was GAPD and ACTN1 (**Figure 69**).

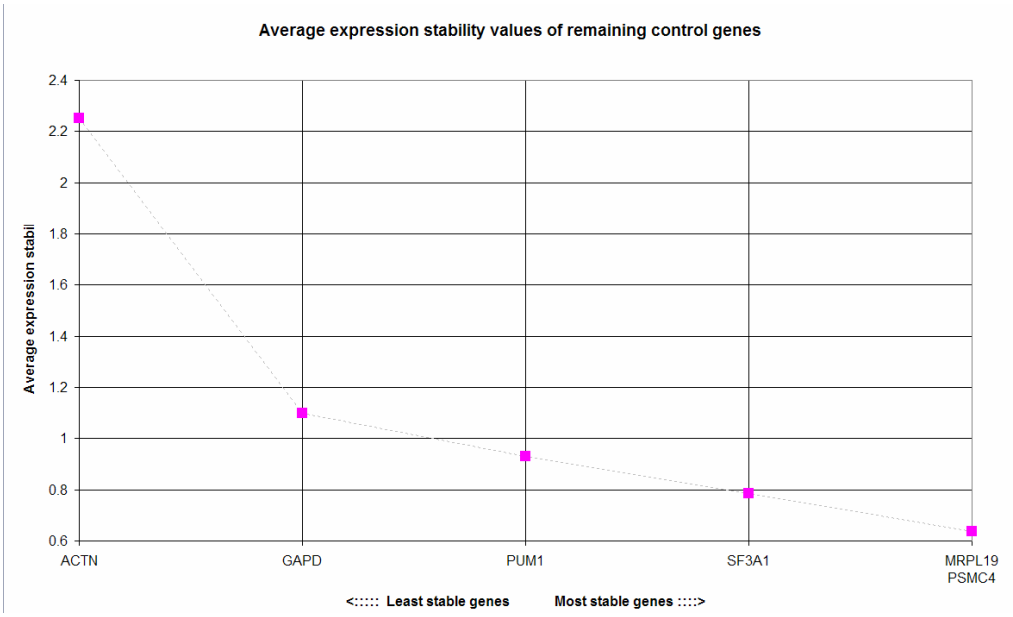


Figure 69: Average expression stability of reference genes by geNORM.

GeNORM calculates the pairwise variation V between two sequential normalization factors containing an increasing number of genes. A large variation means that the added gene has a significant effect and should preferably be included for calculation of a reliable normalization. The lowest cut-off value for the pairwise variation was 0.247 using 4 genes (excluding ACTN1 and GAPD) (**Figure 70**).

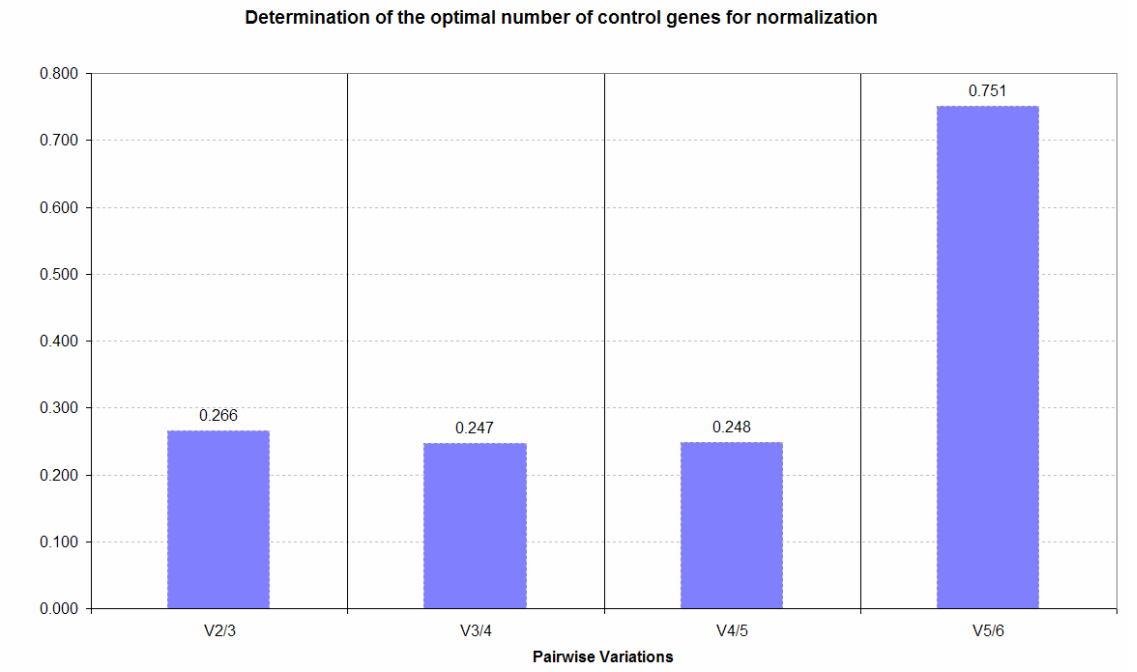


Figure 70: Pairwise variation is the lowest by using 4 control genes (V3/4) to compute the normalization factor for each gene.

The obtained Cp values are transformed to quantities by referring to the calibrator sample, in our case the UHRR sample, which was our reference sample at microarrays. A normalization factor for each gene is calculated from the geometric mean of all control genes (n=4). Normalized quantities of each gene are obtained by dividing raw quantity values by the normalization factor.

Results of the screen of the later test set of 34 tumor samples are resumed at **Appendix A12**. Microarray $\log_2$Ratio value is also stated in parallel. It is observed how the expression values by real time qPCR are similar to those obtained by means of microarray analysis.

Microarray data has good correlation with Real Time qPCR; expression values as it is noted by the coefficient of correlation between both data values of each tumor sample. As expected, values are larger by means of Real Time qPCR than with microarray technique.

As a conclusion from this experiment, gene panel could be useful, it is able discriminate between distinct phenotypic subtypes by its distinctive genes in a context of a group of samples.

But, in my opinion, a gene panel with only 15 genes, it is a too reduced approximation to the reality. Gene expression profiling by means of microarray gives a much better characterization of the breast tumor samples.

5 Discussion

5.1 Establishment of the custom cDNA breast cancer microarray platform

Gene expression profiling using DNA microarrays allows the simultaneous determination of gene expression status of hundreds to thousands of genes, thus it permits to obtain an instant picture of a tissue, being either a cell line or a more heterogeneous sample such as a breast tumor specimen.

The establishment of the customized breast cancer cDNA microarray has been a major success at the CRG. Due to its replicate sample reproducibility, the highest signal to noise at microarray facility, and the intrinsic hybridization specificity due to DNA long probes, it has been successfully used, so far in 476 samples at various gene expression studies at the CRG. Since it has a reduced format, having only 820 unique genes, and the actual lower cost of whole genome commercial platforms, it has been slowly replaced.

Another interesting feature of this platform is that in every new print were included new biomarkers that were found interesting at references and potentially involved in breast cancer. This made the BCA platform to be always evolving.

5.2 *In vitro* studies of the dynamic hormonal response.

To study the signaling pathways that play a major role in a breast tumor environment and lead to tumor progression, an *in vitro* model system was first approached to understand the underlying mechanisms of gene expression in response to hormones.

Breast cancer model cell line T47D-MTVL was investigated under the effect of progestins and estradiol, with the objective to elucidate the signaling pathways in which the endogenously expressing hormone receptors ER and PR respond to artificially added hormones over time. Recent mathematical algorithms specially indicated for this type of experiment, were applied to investigate the dynamic behavior of our population of cells. These algorithms account for the longitudinal sampling of each replicate, and lead us to the identification of the molecules involved in hormonal response, assigning them to various cell biological and molecular functions. Statistical significance analysis was performed to detect genes whose expression changed with a single hormone treatment (E2, R5020) and to identify genes responding differently or similarly between treatments. Genes were further classified based upon their behavior throughout the experimental course, grouping genes that have the same trend, following the premise by which genes that are found to be co-expressed at a certain time and follow similar patterns of expression are functionally related, and often activated through the same transcription factors. Early response

transcription factor genes whose transcription was activated by hormonal induction were also identified, which will be later involved in the rapid activation of various target genes initiating transcription of target proteins. A group of transcription factors were activated early by both hormone treatment which included FOS, ATF3, MYC and SNAI1. This fact indicated that both hormone treatments share similar mechanisms of cell signaling. On the contrary, another set of genes were found to be only activated by progestins and not by estradiol or with a minor response. These genes are GTFH2, NEO1, MAPK7, ABL1, and RPS6KA1, which are involved in cytoplasmatic mitogenic signaling cascades. As a global result, cell cycle progression is time delayed with estradiol in comparison with progestin.

By means of the Erk1/2 pathway specific kinase inhibitor PD98059 (PD) or ER antagonist ICI182780 (ICI), the contribution of the different mechanisms of action of progestins or estradiol in the induction of hormone target genes were further elucidated. Genes which are inhibited by PD, by ICI, or by both drugs at fixed time intervals after hormone induction were also identified. The genes that are induced by hormones and later are inhibited by both PD and ICI are genes that are activated via ERα-PR-B crosstalk mechanism and are dependent on the activation of the Erk1/2 signaling pathway. On the other hand, genes that are induced by hormones, and later only inhibited by PD and not by ICI, are genes dependent on the Erk1/2 signaling pathway but ER independent. Genes that are induced by hormones, but inhibited only by ICI and not by PD, are consequently independent of the Erk1/2 pathway and only ER dependent, and therefore are dependent on the PI3K/Akt signaling pathway, the JAK/STAT signaling pathway or other still unknown mechanisms. And finally, genes that are induced by hormones but neither inhibited by PD nor by ICI will be independent of ERα-PR-B crosstalk, independent of Erk1/2 and would be due to a purely genomic signaling pathway or other still unknown mechanisms (see **Appendix 10** for Venn diagrams).

From the analysis of 120 genes found to be induced by progestin, 75 of them are neither inhibited by PD nor ICI after 6 h of hormonal induction. Therefore it can be confirmed that the majority of genes in our analysis are induced independently of the ERα-PR-B crosstalk, independent of the Erk1/2 signaling pathway, and are most probably due to a direct genomic signaling pathway, PI3K/Akt signaling pathway, JAK/STAT signaling pathway or other still unknown mechanisms.

In the case of E2, from the 95 genes induced by E2 in the T47D model cell line, a third of them are unaffected by either PD or ICI one hour after induction. Again, global gene response to progestins, in our model cell line, is more mitogenic than with estradiol. In the case of estradiol, the cell cycle progression to mitosis is delayed. This probably occurs since cytoplasmatic signaling cascades through MAPK and PI3K kinase signaling pathways are to a large extent more effective at initiating transcription of target genes. It has been observed that other model cell lines, like for example MCF7, have a stronger response to estradiol, possibly due to a higher content of ER.

5.3 Breast tumor gene expression signatures

Gene expression profiling using microarray technology allows the definition of a tumor phenotype from the expression pattern of several genes simultaneously, in contrast to standard methodologies which rely on a few pathological and immunohistochemical markers. Tumors can be more finely classified based on a combination of genes whose expression level is able to discriminate efficiently between clinically different phenotypes of breast tumor. This in turn could be used to establish if they require different treatment strategies.

Due to the genetic heterogeneity of the sample, breast tumors can not be treated individually. Probably, expression profiling variation due to the heterogeneity of the biopsy specimen (proportion of tumoral cells, blood vessels, stromal cells, ductal or lobular tissue), can be observed on the microarray result of different samples from the same tumor, and can explain variability among tumors of similar clinical type.

Our first objective was to classify tumor samples in our population of breast tumor biopsies into various gene expression phenotypes. This was firstly approached employing an unsupervised hierarchical clustering algorithm, where samples were distributed into two groups with maximal differential gene expression. The two main branches of the tree dendrogram, segregated samples based upon their hormone receptor status into ER negative or positive. The first profile of 11 breast tumor samples had characteristic overexpression of genes such as TP53BP2, S100A2, KRT5, NF1B, HRASLS, GPR180, MMP9 and RUNX2, which are distinctive basal/myoepithelial-like molecular markers. These breast tumors are hormone independent and often p53+, and have a poor prognosis (Sorlie *et al.* 2003, van't Veer *et al.* 2002, and van de Vijver *et al*. 2002). To discover new subtypes a threshold line was drawn and 7 sets of tumors were obtained. These were investigated at the functional ontology level to see whether they showed any characteristic cell signaling property, but since the gene lists were small, significant hits were found in three groups of seven.

Unsupervised clustering can be a first exploratory approach as a method to distribute samples on a tree dendrogram, grouping samples upon their similarity metric as the distance, and ultimately for class discovery. Unsupervised clustering can be used only when the discovered classes are clearly divergent. In our dataset of tumor samples, the basal/myoepithelial-like subtype showed this behavior and yielded significant functional annotation traits.

A supervised approach was employed applying principal components analysis by BGA, dividing the tumor collection into two datasets, the training and the test set, since at the time of the analysis the clinical histopathological data were unknown. Unexpectedly, the use of this test set as a validation set awaits follow up of patients' disease progress. The distribution of the samples versus their ER status resulted in more than two groups. The samples that were determined to be basal-like by FADA were again distributed together in a separate group, indicating that this set of tumors is a coherent class with a distinct signature in

agreement with other studies in different patient populations. The limit between ER+ or ER- is unclear, since there is mix of samples, it can not be set from this sample classification.

However the prediction for the test set yields that 6 samples are predicted to be ER- (group 0). We later learned that all these samples (T94, T87, T86, T107, T110) but one (T93) would be also predicted to be basal-like by PAM.

Since the collection comprehended more than two classes, a multiple-class analysis by PCA was performed. This analysis gave that the most distant and coherent groups were the basal-like subtype (group 5) which also included a few ERBB2+ samples, and a group of 10 tumor samples positive for ER, PR, or both receptors (group 1). The other groups were also distant in the 3D PCA space but not enough compared to the previous groups.

Finally, applying prediction analysis of microarrays 6 subtypes (5 breast tumor subtypes and one normal breast subtype) were successfully predicted with a miss-classification error of less than 1%. However, the “normal breast” subtype has to be excluded since it had only three members, and one of them showed dubious aggregative results.

Discriminant genes for each of the 5 predicted pathological subtypes were determined. PAM class 1 and 2 were found to have the lowest number of recurrence cases. PAM class 5 was found to be the class associated with the lowest disease-free survival, and the majority presented recurrence in a period of time of less than 5 years.

5.4 Analysis of PAM predicted subtypes by Gene Set Enrichment Analysis (GSEA) and Ingenuity Pathway Analysis.

To relate these phenotypes to specific cell-signaling pathways and compare tumor data with the data obtained from the hormone dependent breast cancer cell line model treated with progestin and estradiol in a time series, as well as the data from the experiments with inhibitors, we employed recently developed specific applications such as GSEA and Ingenuity pathway analysis.

PAM class 1 over-expresses genes commonly found induced by progestins after 6 hr such as AKAP13, CXCR4, GADD45A, ARID1A (SMARCF1), EGR1, DUSP1, CHES1, TGFB3, and FOS, as well as genes found down-regulated by ICI but not by PD after the effect of progestin treatment. Therefore R5020 induced but ER dependent, such as BIRC3 and SOS1. PAM class 1 overexpressed genes overlap with the gene list of the K-Means cluster 9 after EDGE analysis of the progestin treatment time series with genes such as EGR1, RASA1, and MBD1 functionally related with transcription regulator activity.

Pathway analysis of the discriminating genes of PAM class 1 showed a connection between the glucocorticoid receptor signaling pathway, inflammation and cell proliferation. My hypothesis is that once patients are treated with a hormonal therapy, such an ER antagonist, tumor growth and inflammation could be diminished. All these patients received hormone therapy, and some received adjuvant chemotherapy and the majority are disease-free (11/12).

PAM class 2 matches ER expressing phenotype described by Sorlie, with good overall prognosis. GSEA analysis gave significant overlaps with the good prognosis signature of Van't Veer and Sorlie. It also overlaps with gene lists of T47D cell line found down-regulated after 6 hr of R5020 treatment, such as GATA3, AR, SCGB1D2, HSPA5, NCOR2, ITGA2, UNG2, and NOTCH3.

Top canonical pathways represented are ER signaling, with up-regulated genes such as ESR1, AR, GTF2E2, POLR2A, GATA3, and TFF1; Erk/MAPK signaling, with up-regulated genes such as ESR1, DDR1, HSPB1 and STAT3; glucocorticoid receptor signaling with genes involved such as HSP27, STAT3, and RAS; IL-6 signaling with genes involved such as PI3K, RAS, STAT3 and TGFB3; IGF-1 signaling (IGFBP1 and IGFBP4), and JAK/STAT signaling with PI3K, RAS, STAT, and STAT3.

PAM class 3 analyses by GSEA finds significant overlaps with genes down-regulated after 6 hr of progestin treatment induced early by estradiol after 1 hr of hormonal induction, showing that these tumor subtype is of the luminal or endocrine type with still a hormonal dependence. Cellular growth and proliferation is driven by FGF7, an EGF-like molecule, interacting with CCND2 and EGR1. DUSP1, a gene with is known to act as a negative feedback regulator of JAK/STAT and p38 signaling pathway is activated in this subtype.

PAM class 4 matches the ERBB2+ subtype. In this tumor subtype there is a genomic amplification typically found by fluorescence *in situ* hybridization localized on chromosome region 17q12-q21, which amplifies genes such as SMARCE1 (17q21.2), GRB2 (17q24-26), PPARBP (17q12-21), CDC8 (17q21.3), ERBB2 (17q11.3-q12-17q21.1), TOP2A (17q21-22), and deletes NME1 (17q21.3). Some of these genes are involved in cell cycle progression and proliferation such as ERBB2, SMARCE1, TOP2A, SMARCA4, IGFBP2, and PPARBP. GRB2 is involved in cell signaling by the Erk/MAPK pathway, which is later repressed by DUSP6.

Ingenuity pathway analysis shows that the most represented canonical pathways are the Erk/MAPK pathway, the ER pathway and the glucocorticoid receptor pathway. There is evidence for the interaction of growth factor signaling by EGF and ERBB2, and steroid hormone signaling in controlling the growth of breast cancer cells, in the form of an apparent cross-talk between these growth regulatory systems (Wilson and Slamon, 2005). ERBB2 directly interacts with DUSP6 which is a specific inhibitor of Erk which acts desphosphorylating MAPK and p38 (Amit *et al*. 2007). ERBB2 overexpression is associated to resistance to Tamoxifen and could directly modulate ER levels (Arpino *et al*. 2005). This finding has lead to the hypothesis which involves

peptide growth factor pathways as possible mediators of the steroid hormone-independent phenotype in some human breast cancers, suggesting that the peptide hormone pathways are replacing, in part, the steroid hormone pathways in regulating growth for these tumors.

GSEA analysis of PAM class 4 shows an overlap with the Van't Veer negative prognosis gene set, with highly expressed genes such as AURKA, MCM6, ECT2, DIAPH3·, MAD2L1, CCNB2, and MELK. It also overlaps with a fibroblast serum response gene set. Another significant overlap is with a set of down-regulated genes after hormonal induction by progestins using another model cell line T47y which does not express PR endogenously but was transfected with a vector containing PR carrying a point mutation on the ERID domain, the interaction domain with ERα, where PR has lost its ability to crosstalk and activate Erk1/2 signaling pathway (Ignacio Quiles, Doctoral Thesis, University of Pompeu Fabra, CRG, Barcelona). Thus, some genes that are usually induced by progestins are down-regulated by introducing a mutation on the ERID domain, and overlap in part with the discriminant genes of the PAM subtype 4.

PAM class 5 is the characteristic basal-like subtype characterized by high expression of keratins 5A and 5B and showing high expression of other basal epithelial genes such as KIT1 and ID4. They are also termed as triple-negative tumors as they are found negative for ER, PR and ERBB2. 9 out of 12 samples are p53+ and have high histological grade (10/12).

Ingenuity analysis shows that the significant molecular and cellular functions of this subtype are cell death, cell cycle, cellular growth and proliferation, and cell signaling. The ER signaling pathway is inactive, ER regulated genes TFF, FOS, TGFB3 and CCNH are down-regulated, and now, there is high expression of genes involved in G1/S mitotic cell phase with transcription factors being up-regulated such as E2F3, E2F5, MYC, TFDP1, promoters of cell cycle progression to mitosis, indicative of high proliferative tumors. The Erk/MAPK pathway is also inactive.

High expression of basal/myoepithelial genes such as KRT5/6 and the calcium binding protein S100A, shutt off the ER signaling pathway, and up-regulation of an alternative set of transcription factors such as MYC, RUNX2, TFDP1, and E2F3 which in turn activate cell cycle genes CCND2 and CCNB2 induces the breast cancer cell to proliferate.

Employing GSEA a significant overlap was found with gene lists from Van't Veer's poor prognosis and Sorlie's basal-like phenotype. Poor prognosis of this set of tumors has been also observed by clinicians where tumor samples have high histological grade (HG3 in 10 of 12 samples), most of them are p53+ (9/12), and 7 patients presented early recurrence and died from metastatic disease.

Our PAM predicted phenotypes partially corroborate the tumor subtypes found by Sorlie: the basal-like subtype, which is predominantly ER negative, PR negative and ERBB2 negative (often referred as triple-negative); the ERBB2-

like subtype characterized by the increased expression of several genes at the ERBB2 amplicon, and at least 3 luminal-like subtypes, predominantly hormone receptor positive. Besides identifying the Sorlie's Luminal A, we find 2 additional hormone-dependent tumor phenotypes: PAM class 1 and 3. These defined molecular subtypes have distinct molecular outcomes and responses to therapy. The low grade and low proliferation of the ER positive Luminal A, in our classification PAM class 2, are sensitive to endocrine therapy, and have more favorable prognosis than the ER negative and high grade tumors as in our case the PAM class 5, that are unresponsive to endocrine therapy and respond better to chemotherapy.

Another interesting finding from all these studies was that the identification of extremely distinct expression patterns that differentiated breast tumors beyond the expression of ER, ERBB2 and tumor grade may reflect distinct cell types of origin. However, other factors like menopausal status, tumor size, and nodal status were not associated to dissimilar gene expression patterns

5.5 Prediction of the test set

Normal breast samples which were also included in the test set were correctly assigned considering the 6 PAM predicted subtypes, being PAM class 6 the normal-like. Triple-negative samples, which are negative for ER, PR, and HER-2, are assigned with high probability to the poor prognosis basal-like subtype 5: T86, T87, T94, T107, and T110. Only one tumor sample was left unclassified, since the predicted probability of being assigned to a specific tumor subtype did not reach a 50%. Up to now there is no prognosis data available for this test set of breast tumors samples to corroborate our prediction.

It could be interesting for this type of triple-negative tumors to discover of a functional androgen receptor (AR) pathway in this subgroup of ER and PR negative patients reported by two independent groups. These investigators identified a subgroup with an expression pattern that suggests an active hormone-regulated transcriptional program involving the AR, suggesting to the potential of this pathway for therapeutic targeting in breast cancer (Farmer *et al.* 2005, Doane *et al*. 2006).

5.6 Gene expression patterns as a tool for risk assessment

The potential advantages of improving tumor classification by expression profiling has been central for several-large scale breast cancer studies that have reported identification of signature gene lists with potential for prediction of clinical outcome over the past few years. The ability to identify patients who have a favorable prognosis could, after independent confirmation, allow clinicians to avoid systemic therapy or to choose less aggressive therapeutic options. About 60-70% of patients with lymph node negative breast cancer are cured by local or regional treatment alone. St Gallen (Goldhirsh *et al.* 2003) and the US National Institutes of Health' consensus criteria (Eifel *et al.* 2000)

recommend adjuvant systemic therapy for 85-90% of lymph-node negative patients (Early Breast Cancer Trialists, 1998). There is a need for a definition of an individual patient's risk of disease recurrence to ensure that she receives appropriate therapy.

In the work of van't Veer *et al*. (2002) focused on younger patients (44± 8 years) who were lymph node negative (See **Appendix A1** for more details) they was determined a 70-gene signature (the "Amsterdam signature") was able to predict distant metastasis in less than five years after diagnosis. Van't Veer randomly selected a set of 78 patients as training set, which was used to measure for the correlation between each gene expression and disease outcome. The genes were ranked according to this correlation, and the 70 most-correlated genes were used to construct a classifier discriminating between patients with good and poor prognosis. The remaining 19 patients served as the test set to validate their prognosis classifier. They developed the "MammaPrint" in collaboration with Agilent for patients below 55 years of age, and lymph node negative, as a predictor for distant metastasis-free survival. Following they validated this prognostic gene signature on a cohort of 295 young patients, including lymph node negative and positive breast tumors.

Sorlie *et al.* (2003) concentrated on the classification of breast cancer subtypes and survival-related feature of each of these subtypes was demonstrated on two independent breast cancer datasets (Van't Veer *et al*. and West *et al*.). Only 17 genes appeared in both lists of the "intrinsic list" of 456 genes and the 231 predictor genes of Van't Veer. A follow-up study (van de Vijver et al. 2002) proved the efficiency of Sorlie's classifier as a survival predictor on a large set of 295 tumor specimens.

On a third study, Ramaswamy *et al.* (2003) identified a set of 128 genes separating metastasis from primary tumors. A set of 17 metastasis associated genes were tested on a large diverse set of primary tumors, and were found to successfully distinguish patients with good versus poor prognosis, with only a 2 gene overlap with the 456 intrinsic list of Sorlie's.

Wang *et al.* (2005), in collaboration with a company called Veridex LLC, identified a 76-gene signature consisting in 60 genes for ER+ patients and 16 genes in ER- patients (the "Rotterdam signature"). This gene profile could identify patients who develop distant metastasis within 5 years in lymph node negative patients. Cohort size of ER+ patients was 80, and for ER- patients were 35. Test set consisted in 171 patients. Patients came from 25 different hospitals. There were no differences among the ER+ and ER- groups in age or menopausal status. The test set also did not differ from the training set in any of the characteristics of patients and tumors. Comparison of their results with those of Van der Vijver is difficult because of differences in patients, techniques, and materials used. Their study included node-negative and node positive patients, who had not received adjuvant systemic therapy, and only women younger than 53 years. Microarray platforms differ, Affymetrix and Agilent. Of the 70 genes in the study of Van't Veer and co-workers, 48 genes are present on the Affymetrix, whereas only 38 of the 76-gene signature are present on the

Agilent array. The most striking finding when comparing the signature lists is the virtually complete lack of agreement in the included genes, only a 3 gene overlap between the two signatures. However, both signatures included genes that identified several common pathways that might be involved in tumor recurrence. This finding supports the idea that effective signatures could be required to include representation of specific pathways. The importance of Wang' 76-gene prognostic signature is that, only 30-40% of untreated lymph node negative patients would develop tumor recurrence, and this signature could provide a powerful tool to identify patients at low risk, preventing overtreatment and thus reducing the amount of systemic treatment used in early breast cancer.

In their study, Huang *et al.* (2003) also identified aggregate patterns of gene expression that associate with lymph node status and disease recurrence, capable of predicting outcome in individual cancer patients. Genes found implicated in recurrence prediction were found associated to distinct biological processes as cell proliferation control, specific to cell-cycle, cell signaling activities, growth factor receptors, and G-protein coupled receptors. None of the 70-gene "Amsterdam" Van't Veer signature genes appears to be present in key metagenes in their recurrence study. They believe that the integration of genomic data with clinical risk factors determines the strategy for treating patients as individuals with distinct genomic disease feature, and that genomic data can not replace traditional clinical risk factors but can add substantial detail to this clinical data, especially in a disease such as breast cancer in which multiple, interacting biological and environmental processes define a physiological state.

In our opinion, only the change in microarray technology from Van't Veer, Wang and Huang' works can not satisfactorily explain the disagreements in gene signature, since the different platforms have thousands of genes in common. Also neither age nor the microarray analysis method can be relevant factors in this disagreement. One possible cause is the observation made by Ein-Dor *et al*. 2005 that many genes might correlate with survival, and it is possible to combine genes in many ways to produce signatures with similar predictive power, even from the same data set. In their study, they commented how several microarray studies yielded gene sets whose expression profiles successfully predicted survival, nevertheless with an overlap of the gene sets of almost zero. They focused on data from a single experiment (van't Veer *et al.* 2002) using a bootstrapping method selected 10 training sets of samples randomly, and obtained 10 different top 70 genes correlating with survival with a minimal overlap between them. They confirmed that the dataset is characterized by three main properties: (1) many genes are correlated with survival, (2) the differences between these correlations are small, and (3) the correlation-based rankings of the genes depend strongly on the training set. These properties indicate that the top 70 genes are not superior to others in predicting disease outcome and raises doubts about the reliability and robustness of the reported predictive gene lists. This has been also observed in other complex diseases, how many variables can account for the found differences, and how it is strongly influenced by the subset of patients used for

gene selection, and the small number of samples that were used to generate the gene lists. The same author (Ein-Dor *et al.* 2006) introduce a mathematical model called probably approximately correct (PAC), for evaluating how many samples are needed to generate a robust gene list for predicting outcome in cancer and calculate this number for several breast cancer published studies. He states that thousand of samples are needed to achieve an overlap of 50% between two predictive lists.

Also Michiels *et al.* 2007 investigated the stability of 7 predictive gene lists from 7 large microarray studies and showed that the prediction performances were overoptimistic in comparison with results obtained by reanalysis of the same data performed using different training sets. They listed the potential limitations of microarray for the prediction of cancer outcome in the context of a disease characterized by complex heterogeneous mechanisms.

However, despite these differences most classifiers show a high degree of concordance in predicting the outcome of independent patient populations, suggesting that all these signatures contain very similar information with regards to the outcome. This has been investigated by Fan *et al*. (2006), in which they applied 5 different gene expression signatures with a very small gene overlap to the same data set and found that four of the five predictors showed similar prognostic values.

We can conclude that the ability to use appropriate profiles of gene expression, in clinically homogeneous group of patients, as, for example, dividing the patients into smaller subgroups or phenotypes as in the work of Sorlie's, or to correlate tumor characteristics such as the S-phase fraction, tumor histological grade, ERBB2 overexpression, vascular invasion, presence of lymph node metastasis, hormone receptor status, could add great information towards the a best characterization of a breast tumor, besides the few known clinical histopathological attributes.

Conclusions

The main conclusions from this work are:

- We have establishment a cDNA breast cancer array platform, which has been successfully applied as a useful tool for exploring gene expression profiles of various studies of hormone signaling.
- Tumor samples from our population were classified into various gene expression phenotypes.
- Predicted breast tumor subtypes can be identified by distinctive genes, which discriminate efficiently between them.
- Basal-like phenotype is efficiently discriminated from the other subtypes, and correlates with tumors with high histological grade, p53 negative, triple-negative tumors, and with a large number of recurrence cases which ends in a big percentage in death of the patients.
- An ERBB2 phenotype which typically over-expresses genes of the ERBB2 amplicon region is efficiently discriminated from the other subtypes.
- Sorlie's Luminal subtype A, an ER+ expressing phenotype which correlates with good prognosis, can be discriminated from the other subtypes efficiently. Discriminant genes of this subtype overlap partially with genes inhibited by ER antagonist ICI after hormone induction by R5020.
- Pathway analyses of the predicted breast tumor phenotypes have predominantly characteristic cell-signaling pathways mostly related to cell cycle progression.
- We have been able to detect similar breast tumor subtypes associated with differential survival outcome found in other populations in a heterogeneous local population.
- Different molecular biomarkers were identified for breast cancer tumor progression for each tumor subtype. This set of molecular markers could be used in the future in clinical diagnosis, to improve the choice of treatment, to predict prognosis and identify patients at higher risk of developing metastasis, as well as for following the response to therapy.
- We have accomplished the purpose of giving an alternative tool to clinicians for the complementary validation of their standard diagnostic methods.

6 Future work

Real Time qPCR of multiple marker sets is an alternative emerging technique to obtain expression profiles of a manageable number of genes which could be faster, cheaper, give more reproducible results, and it is a more suitable routine diagnostic platform especially for samples whose RNA integrity is not optimal for gene expression arrays.

To develop appropriate profiles of gene expression, in clinically homogeneous group of patients, as to correlate tumor characteristics such as the S-phase fraction, tumor histological grade, ERBB2 overexpression, vascular invasion, presence of lymph node metastasis, hormone receptor status, could add greater detail in tumor characterization besides the few known biological attributes

It is desirable to use gene expression profiling in diagnosis, prognosis and prediction to treatment, since this technique can potentially be a more precise diagnostic tool and a fine predictor of poor prognosis than any standard diagnostic parameter.

Once all the clinical history is collected, the long term goal of this study will be to carry out prognosis studies with the help of a statistician.

List of references

1. Albanell J, Baselga J. 2001 Unraveling resistance to trastuzamab (Herceptin): insulin-like growth factor-I receptor, a new suspect. J Natl Cancer Inst 93(24):1830-32.
2. Albertson D. 2003 Profiling breast cancer by array CGH. Breast Cancer Research and Treatment, 78:289-298.
3. Arpino G, Weiss H, Lee AV, Schiff R, De Placido S, Osborne CK, and Elledge RM. 2005 Estrogen receptor-positive, progesterone receptor-negative breast cancer: association with growth factor receptor expresión and tamoxifen resistance. J Natl Cancer Inst 97: 1254-61.
4. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G. 2000 Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25-29.
5. Ayers M, Symmans WF, Stec J, Damokosh AI, Clark E, Hess K, *et al*. 2004 Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluoracil, doxorubicin, and cyclophophamide chemotherapy in breast cancer. J. Clin. Oncol.22:2284-93.
6. Bamberger AM, Bamberger CM, Gellersen B, and Schulte HM. 1996 Modulation of AP-1 activity by the human progesterone receptor in endometrial adenocarcinoma cells. Proc Natl Acad Sci U S A 93(12):6169-74.
7. Barber RD, Harmer DW, Coleman RA, and Clark BJ. 2005 GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. Physiol Genomics 21: 389–395.
8. Bardou VJ, Arpino G, Elledge RM, Osborne CK, Clark, GM. 2003 Progesterone receptor status significantly improves outcome prediction over estrogen receptor status alone for adjuvant endocrine therapy in two large breast cancer databases. J Clin Oncol 21:1973-1979.
9. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, and Gifford DK. 2003 Computational discovery of gene modules and regulatory networks. Nat Biotechnol 21:1337-1342.
10. Barnes GL, Herbert KE, Kamal M, Javed A, Einhorn TA, Lian JB, Stein GS, and Gerstenfeld LC. 2004 Fidelity of Runx2 activity in breast cancer cells is required for the generation of metastases-associated osteolytic disease. Cancer research 64, 4506–4513.
11. Barrett T, Dennis B. Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M and Edgar R. 2007 NCBI GEO: mining tens of millions of expression profiles—database and tools update. Nucleic acid Res (35) D760-D765.
12. Baum M, Buzdar AU, Cuzick J, Forbes J, Houghton J, Howell A, and Sahmoud T. 2002 The ATAC (Arimidex, Tamoxifen Alone or in Combination) Trialists' Group. Anastrozole alone or in combination with tamoxifen vesus tamoxifen alone for adjuvant treatment of postmenopausal women with early breast cancer : first results of the ATAC randomized trial. Lancet 359:2131-2139.

13. Beato M, and Klug J. 2000 Steroid hormone receptors: an update. Human reproduction update 6(3):225-236.
14. Beato M, Herrlich P, and Schutz G. 1995 steroid hormone receptors: many actors in serach of a plot. Cell 83:851-7.
15. Beato M. 1989 Gene regulation by steroid hormones. Cell 56:335-344.
16. Beißbarth, T and Speed, TP. 2004 GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 20(9): 1464-1465.
17. Benjamini Y, and Hochberg Y. 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Statistical Society series B57, 289-300.
18. Bergamashi A, Kim YH, Wang P, Sørlie T, Hernandez-Boussard T, Lonning PE, Tibshirani R, Børresen-Dale AL, and Pollack JR. 2004 Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. Genes Chromosomes Cancer 45(11):1033-40.
19. Bieche I, Laurendeau I, Tozlu S, Olivi M, Vidaud D, Lidereau R, and Vidaud M. 1999 Quantitation of MYC Gene Expression in Sporadic Breast Tumors with a Real-Time Reverse Transcription-PCR Assay. Cancer Research 59, 2759-2765.
20. Bonferroni CE. 1936. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8:3-62.
21. Bray JD, Jelinsky S, Ghatge R, Bray JA, Tunkey C, Saraf K, Jacobsen BM, Richer JK, Brown EL, Winneker RC, Horwitz KB, and Lyttle CR. 2005 Quantitative analysis of gene regulation by seven clinically relevant progestins suggests a highly similar mechanism of action through progesterone receptors in T47D breast cancer cells. J Steroid Biochem Mol Biol. 97(4):328-41.
22. Bray JD, Zhang Z, Winneker RC, and Lyttle CR. 2003 Regulation of gene expression by PRA-910, a novel progesterone receptor modulator, in T47D cells. Steroids 68:995-1003.
23. Brazma A, Jonassen I, Vilo J and Ukkonen E. 1998 Predicting gene regulatory elements *in silico* on a genomic scale. Genome Res. 8:1202-1215.
24. Bustin SA, Benes V, Nolan T, and Pfaffl. 2005 Quantitative real-time RT-PCR – a perspective. J Mol. Endocrinol. 34: 597-601.
25. Bustin SA. 2001 Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): Trends and problems. J Mol Endocrinol. 29:23–39.
26. Bustin, S.A. 2000 Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. J. Mol. Endocrinol. 25:169-193.
27. Castoria G, Barone MV, Di Domenico M, Bilancio A, Ametrano D, Migliaccio A, and Auricchio F. 1999 Non-transcriptional action of oestradiol and progestin triggers DNA synthesis EMBO J 18(9):2500-10
28. Castoria G, Miglaccio A, Bilancio A, Di Domenico M, de Falco A, Lombardt M, Gong W, Beato M, and Auricchio F. 2001 PI3-kinase in

concert with Src promotes the S-phase entry of oestradiol-stimulated MCF-7 cells. EMBO J 20:6050-9.
29. Chang HY, Nuyten DSA, Sneddon JB, Hastie T, Tibshirani R, Sorlie T, Dai H, He YD, van't Veer LJ, Bartelink H, van de Rijn M, Brown PO, and van de Vijver MJ. 2005 Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc. Natl. Acad. Sci. USA 102(10):3738-3743.
30. Chang HY, Sneddon JB, Alizadeh AA, Sood R, west RB, Montgomery K, Chi JT, van de Rijn M, Botstein D, and Brown PO. 2004 Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. PLOS biology 2(2):206-214.
31. Chang JC, Wooten EC, Tsimeizon A, Hilsenbeck SG, Gutierrez MC, Elledge R, *et al.* 2003 Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. Lancet 362:362-9.
32. Cicatello L, Addeo R, Sasso A, Altucci L, Petrizzi VB, Borgo R, Cancemi M, Caporali S, Caristi S, Scafoglio C, Teti D, Bresciani F, Perillo B, and Weisz A. 2004 Estrogens and progesterone promote persistent CCND1 gene activation during G1 by inducing transcriptional derepression via c-jun/c-fos/Estrogen receptor (Progesterone receptor) complex assembly to a distal regulatory element and recruitment of Cyclin D1 to its own gene promoter. Mol and Cell Bio 24(16):7260-7274.
33. Clarke R, Leonessa F, Welch JN, and Skaar TC. 2001 Cellular and molecular pharmacology of antiestrogen action and resistance. Pharmacol. Rev. 53(1):25-71.
34. Cleator s, Heller W, and Coombes RC. 2007 Triple-negative breast cancer:therapeutic options. The Lancet 8:235-244.
35. Collins N, Wooster R, and Stratton MR. 1997 Absence of methylation of CpG dinucleotides within the promoter of the breast cancer susceptibility gene BRCA2 in normal tissues and in breast and ovarian cancers. Br J Cancer 76(9):1150-6.
36. Crook T, Brooks LA, Crossland S, *et al.* 1998 p53 mutation with frequent novel condons but not a mutator phenotype in BRCA1- and BRCA2-associated breast tumors. Oncogene 17:1681-89.
37. Culhane AC, Perrière G, Considine EC, Cotter TG, and Higgins DG. 2002 Between-group analysis of microarray data. Bioinformatics 18(12):1600-8.
38. Cunliffe HE, Ringner M, Bilke S, Walker RL, Cheung JM, Chen Y, and Meltzer PS. 2003 The gene expression response of breast cancer to growth regulators: Patterns and correlation with tumor expression profiles.Cancer Research 63:7158-7166.
39. Dennis GJ, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, and Lempicki RA. 2003 DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 4(5):P3.
40. DeRisi JL, Lyer VR, and Brown PO. 1997 Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, 278: 680-686.

41. Desai UJ, and Pfaffle PK. 1995 Single-step purification of a thermostable DNA polymerase expressed in *Escherichia coli.* Biotechniques 19:780-784.
42. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO,and Alizadeh AA. 2003 SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Res 31(1):219-223.
43. Doane, A. S. *et al.* 2006 An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. Oncogene 25, 3994–4008.
44. Dobrovic A, and Simpfendorfer D. 1997 Methylation of the BRCA1 gene in sporadic breast cancer. Cancer Res 15;57(16):3347-50.
45. Dowsett M, Ebbs SR, Dixon JM, Skene A, Griffith C, Boeddinghouse I, Salter J, Detre S, Hills M, Ashley S, Francis S, Walsh G, and Smith IE. 2003 Biomarker changes during neoadjuvant Anastrozole, Tamoxifen, or the combination: influence of hormonal status and HER-2 in breast cancer – A study from the IMPACT Trialists. J Clin Oncol 23:2477-2492.
46. Dudoit S, Gentleman RC, and Quackenbush J. 2003 Open source software for the analysis of microarray data. Biotechniques Mar Suppl:45-51.
47. Dudoit S, Yang YH, Speed TP, and Callow MJ. 2002. statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica 12: 111-140.
48. Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of randomised trials. 1998 Lancet 351: 1451–67, and 352: 930-42.
49. Eberwine J, Teh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M, and Coleman P. 1992 Analysis of gene expression in single live neurons. Proc Natl Acad Sci 89:3010-3114.
50. Efron, B and Tibshirani, RJ. 1993 An introduction to the Bootstrap (Chapman & Hall, Boca ratón, FL.).
51. Eifel P, Axelson JA, Costa J, *et al*. 2000 National Institutes of Health Consensus Development Conference Statement: adjuvant therapy for breast cancer. J Natl Cancer Inst 2001; 93: 979–89.
52. Ein-Dor I, Kela I, Getz G, Givol D, and Dormany E. 2005 Outcome signature genes in breast cancer: is there a unique set? Bioinformatics, 21:171-178.
53. Ein-Dor L, Zuk O, and Domany E. 2006 Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci 103:5923-5928.
54. Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein. 1998 Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *95*:14863-14868.
55. Ellis MJ. 2005 Neoadjuvant endocrine therapy for breast cancer: more questions than answers. J Clin. Oncol. 2005 Aug 1;23(22):4842-4.
56. Falcon and Gentleman. 2007 Using GOstats to test gene lists for GO term association. Bioinformatics 23(2):257-8.

57. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, van't Veer LJ, and Perou CM. 2006 Concordance among gene-expression-based predictors for breast cancer. N Engl J Med. 355(6):560-9.
58. Farmer, P. et al. 2005 Identification of molecular apocrine breast tumours by microarray analysis. Oncogene 24:4660–4671.
59. Fellenberg K, Hauser NC, Brors B, Neutzer A, Hoheisel JD, and Vingron M. 2001 Correspondance analysis applied to microarray data. Proc. Natl. Acad. Sci. 98 (19):10781-10786.
60. Fendrick JL, Raafat AM, and Haslam SZ. 1998 Mammary gland growth and development from the postnatal period to menopause: ovarian steroid receptor ontogeny and regulation in the mouse. J Mammary Gland Biol Neopla 3:7-22.
61. Fuqua SAW and Cui Y. 2004 Estrogen and progesterone receptor isoforms: clinical significance in breast cancer. Breast cancer Research and treatment 87:S3-S10.
62. Fuqua SAW, Cui Y, Lee AV, and Osborne CK. 2005 Insights into the role of progesterone receptors in breast cancer. J Cli Oncol 5:931-932.
63. Gayther SA, Pharoah PD, and Ponder BA. 1998 The genetics of inherited breast cancer. J Mammary Gland Biol Neoplasia 3(4):365-76.
64. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, and Zhang J. 2004 Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5(10):R80.
65. Glas AM, Floore A, Delahaye LJ, Witteveen AT, Pover RC, Baks N, Lahti-Domenici JS, Bruinsma TJ, Warmoes MO, Bernards R, Wessels LFA, and Van't Veer LJ. 2006 Converting a breast cancer microarray signature into a high-throughput diagnostic test. BMC Genomics 7: 278.
66. Goldhirsch A, Wood C, Gelber RD, Coates AS, Thürlimann B, Senn HJ. 2003 Meeting highlights: updated international expert consensus on the primary therapy of early breast cancer. J Clin Oncol. 21: 3357–65.
67. Golub TR, Slonim DK, Tamayo P, Huard C, Gasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfeld CD, and Lander ES. 1999 Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286:531-537.
68. Graham JD, Yeates C, Balleine RL, Harvey SS, Milliken JS, Bilous AM, and Clarke CL. 1995 Characterization of progesterone receptor A and B expression in human breast cancer. Cancer Res 55(21):5063-8.
69. Groshong SD, Owen GI, Grimison B, Schauer IE, Todd MC, Langan TA, Sclafani RA, Lange CA, and Horwitz KB. 1997 Biphasic regulation of breast cancer cell growth by progesterone: role of the cyclin-dependent kinase inhibitors, p21 and p27(Kip1). Mol Endocrinol 11:1593-1607.
70. Grushko TA, Dignam JJ, Das S, et al. 2004 MYC is amplified in BRCA1-associated breast cancer. Clin. Cancer Res. 10:499-507.
71. Gruvberber S, Ringner M, Chen Y, Panavally S, Saal LH, Borg A, Ferno M, Peterson C, and Meltzer PS. 2001 Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. Cancer Res. 61:5979-5984.

72. Hartigan JA and Wong MA. 1979 A k-means clustering algorithm. Appl. Stat. 28:100-108.
73. Hartigan JA. 1975 Clustering algorithms. New York, NY, Wiley & Sons.
74. Haslam SZ and Woodward TL. 2003 Epithelial-cell-stromal-cell interactions and steroid hormone action in normal and cancerous mammary gland. Breast Cancer Res. 5(4):208-215.
75. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallionemi OP, Wilfond B, Borg A, and Trent J. 2001 Gene-expression profiles in hereditary breast cancer. N Engl J Med 344:539-548.
76. Hedenfalk I, Ringner M, Ben-Dor A, Yakhini Z, Chen Y, Chebil G, Ach R, Loman N, Olsson H, Meltzer P, Borg A, and Trent J. 2003 Gene expression profiles in hereditary breast cancer. Proc Natl Acad Sci 100:2532-2537.
77. Hoch RV, Thompson DA, Baker RJ, and Weigel RJ. 1999 GATA-3 is expressed in association with estrogen receptor in breast cancer. Int J Cancer 84;122-128.
78. Hopp TA, Weiss HL, Hilsenneck SG, Cui Y, Allred DC, Horwitz KB, *et al*. 2004 Breast cancer patients with progesterone receptor PR-A rich tumors have poorer desease-free survival rates. Clin. Cancer Res. 10:2751-2760.
79. Hosack DA, Dennis G, Sherman BT, Lane HC, and Lempicki RA. 2003 Identifying biological themes within lists of genes with EASE. Genome Biol. 4(10):R70.
80. Hovland AR, Powell RL, Takimoto GS, Tung L, and Horwitz KB. 1998 An N-terminal inhibitory function, IF, suppresses transcription by the A-isoform but not the B-isoform of human progesterone receptors. J Biol Chem 273(10):5455-60.
81. Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, West M, Nevins JR, Huang AT. 2003 Gene expression predictors of breast cancer outcomes. Lancet 362:95-102.
82. Huber W, von Heydebreck A, Sültman H, Poustka A and Vingron M. 2002 Variace stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18: S96-S104.
83. Hugues-Davies L, Huntsman D, Ruas M, Fuks F, Bye J, Chin SF, Milner J, Brown LA, Hsu F, Gilks B, Nielsen T, Schulzer M, Chia S, Ragaz J, Cahn A, Linger L, Ozdag H, Cattaneo E, Jordanova ES, Schuuring E, Yu DS, Venkitaraman A, Ponder B, Doherty A, Aparicio S, Bentley D, Theillet C, Ponting CP, Caldas C, and Kouzarides T. 2003 EMSY links the BRCA2 pathway to sporadic breast and ovarian cancer. Cell 115:523-535.
84. Huse B, Verca SB, Matthey P, and Rusconi S. 1998 Definition of a negative modulation domain in the human progesterone receptor. Mol Endocrinol 12(9):1334-42.
85. Hyman E, Kauraniemi P, Hautaniemi S, Wolf M, Mousses S, Rozenblum E, Ringner M, Sauter G, Monni O, Elkahloun A, Kallioniemi OP, and Kallioniemi A. 2002 Impact of DNA amplification on Gene expression patterns in breast cancer. Cancer Research 62:6240-6245.

86. Jeng MH, Parker CJ, and Jordan VC. 1992 Estrogenic potential of progestins in oral contraceptives to stimulate human breast cancer cell proliferation. Cancer Res 52:6539-6546.
87. Jordan VC. 2004 Selective estrogen receptor modulation: concept and consequences in cancer. Cancer Cell 5:207-213.
88. Kiss AL, Turi A, Müllner N, Kovács E, Botos E, and Greger A. 2005 Oestrogen-mediated tyrosine phosphorylation of caveolin-1 and its effect on the oestrogen receptor localization: an in vivo study. Mol Cel Endocrinol. 245(1-2):128-37.
89. Kraus WL, Weis KE, and Katzenellenbogen BS. 1995 Inhibitory crosstalk between steroid hormone receptors: differential targeting of estrogen receptor in the repression of its transcriptional activity by agonist- and antagonist-occupied progestin receptors. Mol. Cell. Biol. 15:1847-1857.
90. Lange CA, Richter Jk, Horwitz KB. 1999 hypothesis: Progesterone primes breast cancer cells for cross talk with proliferative or antiproliferative signals. Mol. Endocrinol. 13:829-36.
91. Leo JCL, Wang SM, Guo CH, Aw SE, Zhao Y, Li JM, Hui KM, and Lin VCL. 2005 Gene regulation profile reveals consistent anticancer properties of progesterone in hormone-independent breast cancer cells transfected with progesterone receptor. Int. J. Cancer. 117(4):561-8.
92. Liao DJ and Dickson RB. 2000 c-Myc in breast cancer. Endocr. Relat. Cancer 7(3):143-163.
93. Lin VC, Ng EH, Aw SE, Tan MG, Ng EH, Chan VS, and Ho GH. 1999 Progestins inhibit the growth of MDA-MB-231 cells transfected with progesterone receptor complementary DNA. Clin Cancer Res 5:395-404.
94. Lönnsted L and Speed T. 2001 Replicated microarray data. Statistica Sinica 12:31.
95. Lösel R, Wehling M. 2003 Nongenomic actions of steroid hormones. Nat Rev Mol Cell Biol 4:46-56.
96. Lozano JJ, Soler M, Bermudo R, Abia D, Fernandez PL, Thompson TM, and Ortiz AR. 2005 Dual activation of pathways regulated by steroid receptors and peptide growth factors in primary prostate cancer revealed by Factor Analysis of microarray data. BMC Genomics 6:109.
97. Lozano JJ, Soler M, Bermudo R, Abia D, Fernandez PL, Thomson TM, Ortiz AR. 2005 Dual activation of pathways regulated by steroid receptors and peptide growth factors in primary prostate cancer revealed by Factor Analysis of microarray data. BMC Genomics. Aug 17;6(1):109.
98. Ma XJ, Salunga R, Tuggle JT, Gaudet J, Enright E, McQuary P, Payette T, Pistone M, Stecker K, Zhang BM, Zhou YX, Varnholt H, Smith B, Gadd M, Chatfield E, Kessler J, Baer TM, Erlander MG, and Sgroi DC. 2003 Gene expression profiles of human breast cancer progression. Proc Natl Acad Sci USA 100:5974-5979.
99. Ma XJ, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, Fuller A, muir B, Mohapatra G, Salunga R, Tuggle JT, Tran Y, Tran D, Tassin A, Amon P, Wang Wilson , Wang Wei, Enright E, Stecker K, Estepa-Sabal E, Smith B, Younger J, Balis U, Michaelson J, Bhan A, Habib K, Baer TM, Brugge J, Haber DA, Erlander MG, and Sgroi DC. 2004 A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with Tamoxifen. Cancer Cell 5:607-616.

100. Macgregor PF and Squire JA. 2002 Application of microarray to the analysis of gene expression in cancer. Clin. Chem. 48:1170-7.
101. McDonell DP, and Goldman ME. 1994 RU486 exerts antiestrogenic activities through a novel progesterone receptor A form-mediated mechanisms. J. Biol. Chem. 269:11945-11949.
102. McDonell DP, Shahbaz MM, Vegeto E, and Goldman ME. 1994 The human progesterone receptor A-form functions as a transcriptional modulator of minaralcorticoid receptor transcriptional activity. J. Steroid Biochem. Mol. Biol. 48:425-432.
103. Michalides R, Griekspoor A, Balkenende A, Verwoerd D, Janssen L, Jalink K, Floore A, Velds A, van't Veer, and Neefjes J. 2004 Tamoxifen resistance by a conformational arrest of the estrogen receptor a after PKA activation in breast cancer. Cancer Cell 5:597-605.
104. Michalides R, Griekspoor A, Balkenende A, Verwoerd D, Janssen L, Jalink K, Floore A, Velds A, van't Veer L, and Neefjes J. 2004 Cancer Cell 5:597-604.
105. Michiels S, Koscielny S, and Hill C. 2007 Interpretation of microarray data in cancer. British J Cancer 96:1155-1158.
106. Migliaccio A, Di Domenico M, Castoria G, de Falco A, Bontempo P, Nola E, Auricchio F. 1996 Tyrosine kinase/p21ras/MAP-kinase pathway activation by estradiol-receptor complex in MCF-7 cells. EMBO J 15:1292-1300.
107. Migliaccio A, Piccolo D, Castoria G, Di Domenico M, Bilancio A, Lombardi M, Gong W, Beato M, Auricchio F. 1998 Activation of the src/p21ras/erk pathway by progesterone receptor via a crosstalk with estrogen receptor. EMBO J 17:2008-2018.
108. Monni O, Hyman E, Mousses S, Barlund M, Kallioniemi A, and Kallioniemi OP. From chromosomal alterations to target genes for therapy: integrating cytogenetic and functional genomic views of the breast cancer genome. 2001 Cancer Biology, 11:395-401.
109. Mote PA, Bartow S, Tran N, and Clarke CL. 2002 Loss of co-ordinate expression of progesterone receptors A and B is an early event in breast carcinogenesis. Breast Cancer Res Treat. 72(2):163-72.
110. Muckenthaler M, Richter A, Gunkel N, Riedel D, Polycarpou-Schwarz M, Hentze S, Falkenhahn M, Stremmel W, Ansorge W, and Hentze MW. 2003 Relationships and distinctions in iron regulatory networks responding to interrelated signals. Blood 101(9):3690-8.
111. Musgrove EA, Lee CS, and Sutherland RL. 1991 Progestins both stimulate and inhibit breast cancer cell cycle progression while increasing expression of transforming growth factor alpha, epidermal growth factor receptor, c-fos, and c-myc genes. Mol Cell Biol 11(10):5032-43.
112. Nevins JR, Huang ES, Dressman H, Pittman J, Huang AT, and West M. 2003 Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. Hum. Mol. Genet. (15):12.
113. Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, Hu Z, Hernandez-Boussard T, Livasy C, Cowan D, Dressler L, Akslen LA, Ragaz J, Gown AM, Gilks CB, van de Rijn M, Perou CM. 2004

Immunohistochemical and clinical characterization of the basal-subtype of invasive breast carcinoma. Clin. Cancer Res. 10:5367-74.

114. Olayioye MA, Neve RM, Lane HA, and Hynes NE. 2000 The erbB signaling network: heterodimerization in development and cancer. EMBO J. 19:3159-3167.

115. Osborne CK and Schiff R. 2005 Aromatase inhibitors: Future directions. J Steroid Biochem Mol Biol 95(1-5):183-7.

116. Owen GI, Richer JK, Tung L, Takimoto G, and Horwitz KB. 1998 Progesterone regulates transcription of the p21(WAF1) cyclin- dependent kinase inhibitor gene through Sp1 and CBP/p300. J Biol Chem 273(17):10696-701.

117. Payvar F, DeFranco D, Firestone GL, Edgar B, Wrange O, Okret S, Gustafsson JA, and Yamamoto KR. 1983 Sequence-specific binding of glucocorticoid receptor to MTV DNA at sites within and upstream of the transcribed region. Cell 35:381-392.

118. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenshikov A, Williams CF, Zhu SX, Lee JC, Lashkari D, Shalon D, Brown PO, and Botstein D.1999 Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci USA 96:9212-9217.

119. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, ,Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenshikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO and Botstein D. 2000 Molecular portraits of human breast cancer tumors. Nature, 406:747-752.

120. Pfaffl et al. 2004 Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: Excel-based tool using pair-wise correlations. Biotechnology Letters 26:509-512.

121. Pfaffl MW. 2001 A new mathematic model for relative quantification in real-time RT-PCR. Nucleic Acid Res. 29:2002-2007.

122. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamanshikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO. 1999 Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet, 23:41-46.

123. Pollack JR, Sorlie T, Perou CM, Rees CM, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borrelsen-Dale AL, Brown PO. 2002 Microarray analysis reveals a major direct role of DNA copy number alterations in the transcriptional program of human breast tumors. Proc Natl Acad Sci USA 99:12963-12968.

124. Preiss T. 2001 Analysis of transcriptome changes in response to rapamycin using DNA microarrays. Habilitation report. University of Heidelberg, Germany.

125. Quanckenbush J. 2001 Computational analysis of microarray data. Nat Rev Genet 2: 418-427.

126. Rakha EA, El-Sayed ME, Green AR, Lee AH, Robertson JF, and Ellis IO. 2007 Prognostic markers in triple-negative breast cancer. Cancer 109:25-32.

127. Ramaswamy S, Ross KN, Lander ES, and Golub TR. 2003 A molecular signature of metastasis in primary solid tumors. Nat Genet 33:49-54.
128. Reyment RJ, and Joreskog KG.1996 Applied Factor Analysis in the Natural Sciences. Cambridge, Cambridge University Press.
129. Richer JK, Jacobsen BM, Manning NG, Abel MG, Wolf DM, and Horwitz KB. 2002 Differential gene regulation by the two progesterone receptor isoforms in human breast cancer cells. J. Biol. Chem. 277:5209-5218.
130. Richter A, Schwager C, Hentze S, Ansorge W, Hentze MW, and Muckenthaler M. 2002 Comparison of fluorescent tag DNA labeling methods used for expression analysis by DNA microarrays. Biotechniques 33(3):620-8, 630.
131. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, van der Rijn M, Waltham M, Pergamenschikov A, Lee JCF, Lashkari D, Salón D, Myers TG, Weinstein JN, Botstein D, and Brown PO. 2000 Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 24:227-235.
132. Ross JS and Fletcher JA. 1998 The HER-2/*neu* oncogene in breast cancer: prognostic factor, predictive factor, and target for therapy. Stem Cells 16:413-428.
133. Rozen S and Skaletzky H. 2000 Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. (132):365-86.
134. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky L, Liu Z, Vinsavich A, Trush V, and Quackenbush J. 2003 TM4: A Free, Open-Source System for Microarray Data Management and Analysis, BioTechniques 34:374-378.
135. Schena M, Shalon S, Davis RW, and Brown PO. 1995 Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270:467-470.
136. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, *et al*. 2000 A gene expression database for the molecular pharmacology of cancer. Nat. Genet. 24:236-44.
137. Schiff R, Massarweh SA, Shou J, Bharwani L, Mohsin SK, and Osborne SK. 2004 Cross-talk between estrogen receptor and growth factor pathways as a molecular target for overcoming endocrine resistance. Clin Cancer Res 10:331-6S.
138. Schulze A, and Downward J. 2001 Navigating gene expression using microarrays – a technology review. Nat Cell Biol 3: E190-E195.
139. Selaru FM, Yin J, Olaru A, Mori Y, Xu Y, Epstein SH, Sato F, Deacu E, Wang s, Sterian A, Fulton A, Abraham JM, Shibata D, Baquet C, Stass SA, and Meltzer SJ. 2004 An unsupervised approach to identify molecular phenotypic components influencing breast cancer features. Cancer Res 64:1584-1588.
140. Shaffer, JP. 1995 Multiple hypothesis testing. Annu. Rev. Psychol. 46, 56 1—576.

141. Shena M, Shalon D, Heller R, Chai A, Brown PO, and Davis RW. 1996 Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci USA 93:10614-10619.
142. Skildum A, Faivre E, and Lange CA. 2005 Progesterone receptors induce cell cycle progression via activation of mitogen activated protein kinases. Mol Endocrinol 19(2):327-39.
143. Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, Fleming T, Eiermann W, Wolter J, Pegram M, Baselga J, and Norton L. 2001 Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med 344(11):783-92.
144. Smith IE, and Dowsett M. 2003 Aromatase inhibitors in breast cancer. N Engl J Med 348:2431-2442.
145. Sneath PHA and Sokal RR. 1973 numerical Taxonomy. WH Freeeman, San Francisco.
146. Sorlie T, Perou CM, Tibshirani R, Aas T, Gelder S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, and Borresen-Dale AL. 2001 Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA, 98 :10869-10874.
147. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL, Botstein D. 2003 Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA, 100:8418-8423.
148. Sotiriou C, Neo SY, NcShane LM, korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, and Liu ET. 2003 Breast cancer classification and prognosis based on gene expresión profiles from a population-based study. Proc Natl Acad Sci USA 100(18):10393-10398.
149. Soukas A, Cohen P, Socci ND, and Friedman JM. 2000 Leptin-specific patterns of gene expression in white adipose tissue. Genes Dev. *14*:963-980.
150. Szabo A, Perou CM, Karaca M, Perreard L, and Quackenbush JF, Bernard PS. 2004 Statistical modeling for selecting housekeeper genes. Genome Biol. 5(8): R59.
151. Stadel BV. 2002 Hormone replacement therapy and risk of breast cancer. JAMA 287:2360-61.
152. Stahlberg A Zoric N, Aman P, and Kubista M. 2004 Quantitative real-time PCR for cancer detección: the lymphoma case. Expert Rev. Mol. Dign. 5(2):221-230.
153. Staunton JE, SlonimColler HA, Tamayo P, Angelo MJ, Park J, *et al*. 2001 Chemosensitivity prediction by transcriptional profiling. Proc. Natl. Acad. Sci. USA 98:10787-92.
154. Storey JD and Tibshirani, R 2005 Genome wide studiesProc Natl Acad Sci USA 98:31-36.
155. Storey JD, Xiao W, Leek JT, Tompkins RG, and Davis RW. 2005 Significance analysis of time course microarray experiments. Proc Natl Acad Sci USA 102:12837-12842.

156. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gilette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, and Mesirov JP. 2005 Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102:15545-15550.
157. Sumida T, Itahana Y, Hamacaba H, and Desprez PY. 2004 Reduction of human metastatic breast cancer cell aggressiveness on introduction of Esther form A or B of the progesterona receptor and then treatment with progestins. Cancer Research 64:7886-7892.
158. Sutherland RL, Hall RE, Pang GYN, Musgrove EA, and Clarke CL. 1998 Effect of medroxyprogesterone acetate on proliferation and cell cycle kinetics of human mammary carcinoma cells. Cancer Res 48:5084-5091.
159. Szabo A, Perou CM, Karaca M, Perreard L, and Quackenbush JF, and Bernard PS. 2004 Statistical modeling for selecting housekeeper genes. Genome Biol. 5(8): R59.
160. Taylor-Papadimitriou J, Stampfer M, Bartek J, Lewis A, Boshell M, Lane EB, and Leigh IM. 1989 Keratin expression in human mammary epithelial cells cultured from normal and malignant tissue: relation to in vivo phenotypes and influence of medium. J Cell Sci 94 (Pt 3):403-13.
161. Tibshirani R, Hastie T, Narasimhan B, and Chu G. 2003 Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl acad Sci USA 99:6567-6572.
162. Tovey S, Dunne B, Witton CJ, Forsyth A, Cooke TG, and Bartlett JMS. 2005 Can molecular markers predict when to implement treatment with aromatase inhibitors in invasive breast cancer?. Clin Cancer Res 11(13):4835-4842.
163. Tran PH, Peiffer DA, Shin Y, Meek LM, Brody JP and Cho KW. 2002 microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. Nucleic Acid Res. 30(12):e54.
164. Tung L, Mohamed JP, Hoeffler JP, Takimoto GS, and Horwitz KB. 1993 Antagonist-occupied human progesterone B-receptors activate transcription without binding to progesterone response elements and are dominantly inhibited by A-receptors. Mol. Endocrinol. 7:1256-1265.
165. Turner NC, Reis-Filho JS, Russell AM, Springall RJ, Ryder K, Steele D, Savage K, Gillett CE, Schmitt FC, Ashworth A, and Tutt AN. 2006 BRCA1 dysfunction in sporadic basal-like breast cancer. Oncogene 26(14):2126-32.
166. Tusher VG, Tibshirani R, Chu G. 2001 Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci 98:5115-5121.
167. Usary J, Llaca V, Karaca G, Presswala S, Karaca M, He X, Langerod A, Karesen R, Oh DS, Dressler Lg, Lonning Pe, Strausberg RL, Chanock S, Borresen-Dale AL, and Perou CM. 2004 Mutation of GATA3 in human breast tumors. Oncogene 23(46):7669-78.
168. van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Llenen D, and Holstege FCP. 2003 Monitoring global messenger RNA changes in externally controlled microarray experiments. EMBO reports 4:387-393.

169. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, and Bernards R. 2002 A gene expression signature as a predictor of survival in breast cancer. N. Engl.J.Med, 347:1999-2009.
170. Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, and Eberwine JM. 1990 Amplified RNA synthesized from limited quantities of heterogenous cDNA. Proc Natl Acad Sci USA 87:1663-1667.
171. Van Helden J, Andre B, and Collado-vides J. 1998 Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol. 281(5):827-42.
172. Van Laere S, Van der Auwera I, Van den Eynden GG, Fox SB, Bianchi F, Harris AL, van Dam P, Van Marck EA, Vermeulen PB, and Dirix Ly. 2005 Distinct molecular signature of inflammatory breast cancer by cDNA microarray analysis. Breast Cancer Res.Treat. 93(3):237-46.
173. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy, K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, and Friend SH. 2002 Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530-536.
174. Vandesompele J, De Preter K, Pattyn F, et al. Accurate normalization of real-time geometric averaging of multiple multiple internal control genes. 2002 Genome Biol. 0034.1–0034.11.
175. Vegeto E, Shahbaz MM, Wen DX, Goldman ME, O'Malley BW, and McDonell DP. 1993 Human progesterone receptor A form is a cell- and promoter-specific repressor of human progesterone receptor B function. Mol. Endocrinol. 7:1244-1255-
176. Velculescu V, and El-Deiry WS. 1996 Biological and clinical importance of the p53 tumor suppressor gene. Clin. Chem. 42: 858-68.
177. Venkitaraman AR. 2002 Cancer Susceptibility and the Functions of BRCA1 and BRCA2. Cell 108:171-182.
178. Wan Y and Nordeen SK. 2002 Overlapping but distinct gene regulation profiles by glucocorticoids and progestins in human breast cancer cells. Mol.Endocrinol. 16:1204-1214.
179. Wang Y, Klijn JGM, Zhang Y, Siuwerts AM, Look MP, Yang F, Talantov D, timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EMJJ, Atkins D, and Foekens JA. 2005 Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet 365:671-679.
180. Wardell SE, Boonyaratanakornkit V, Adelman JS, Aronheim A, and Edwards DP. 2002 .Jun dimerization protein 2 functions as a progesterone receptor N-terminal domain coactivator. Mol Cell Biol. 22(15):5451-66.
181. Weigelt B, Glas AN, Wessels LFA, Witteveen AT, Peterse JL, and van't Veer LJ. 2003 Gene expression profiles of primary breast tumors

maintained in distant metastases. Proc Natl Acad Sci USA 100:15902-15905.

182. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA, Marks JR, and Nevins JR. 2001 Predicting the clinical status of human breast cancer by using gene expression profiles. Proc. Natl, Acad. Sci. USA 98:11462-11467.

183. West RB, Nuyten DSA, Subramanian S, Nielsen TO, Corless CL, Rubin BP, Montgomery K, Zhu S, Patel R, Hernandez-Boussard T, Goldblum JR, Brown PO, van de Vijver M, and van de Rijn M. 2005 Determination of stromal signatures in breast carcinoma. PLOS biology 3(6):e187.

184. Wettenhall JM, Smyth GK. 2004 limmaGUI: a graphical user interface for linear modeling of microarray data. Bioinformatics 20(18):3705-6.

185. Wilson CA and Slamon DJ. 2005 Evolving understanding of growth regulation in human breast cancer: interaction of the steroid and peptide growth regulatory pathways. J Natl Cancer Inst 97:1238-1239. Amit I, Citri A, Shay T, Lu S, Katz M, Zhang F, Tarcic G, Siwak D, Lahad J, Jacob-Hirsch J, Amariglio N, Abisman N, Segal E, Rechavi G, Alon U, Mills GB, Domany E, and Yarden Y. 2007 A module of negative feedback regulators defines growth factor signaling. Nat Genet 39: 503-512.

186. Winegarden N. 2003 Microarrays in cancer: moving from hype to clinical reality. Lancet 362:1428.

187. Winston JS, Ramanaryanan J, and Levine E. 2004 HER-2/neu evaluation in breast cancer. Am. J. Clin. Pathol. 121(Supl.):S33-49.

188. Woodward TL, Xie JW, and Haslam SZ. 1998 Role of mammary stroma in modulating the proliferative response to ovarian hormones in the normal mammary gland. J Mammary Gland Biol Neopla 3: 117-131.

189. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, et al. Within the fold: assessing differential expression measures and reproducibility in microarray assays. 2002 Genome Biol.3, 0062.1-0062.12.

190. Yang YW, Dudoit S, Liu P, Lin DM, Peng V, Ngai J, and Speed TP. 2002 Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 30(4)e15.

191. Zheng ZY, Bay BH, Aw SE, and Lin VCL. 2005 A novel anti-estrogenic mechanism in progesterone receptor-transfected breast cancer cells. J Biol Chem 280(17):17480-7.

192. Zuker M. 2003 Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acid Res. 31(13):3406-15.

List of tables

List of figures

Appendices

Appendix_A1: Breast cancer studies using microarrays

17/12/2007

Publication	Title	Array Platform	Patients (invasive tumor size)	Reference sample used	Cohort sze	Follow-up	Minimal discriminatory element
Pollack, Nature Genetics, 1999	Genome-wide analysis of DNA copy-number changes using cDNA microarrays	cDNA microarray (5,240), CGH analysis		BT-474 tumor cell line			
Perou et al, Nature 1999	Distinctive gene expression patterns in human mammary epithelial cells and breast cancers	cDNA/Stanford 8,102 genes	T3-T4: >5 cm	HMEC cell line	38 invasive	-------	496 elements. 4 biological classes of invasive
Perou et al, Nature 2000	Molecular portraits of Human Breast Tumors	cDNA/Stanford 8,102 genes		Pool of mRNAs isolated from 11 different cultured cell lines	42 breast tumors		
Hedenfalk, N Engl J Med,, 2001	Gene-Expression Profiles in Hereditary Breast Cancer			MCF-10A			
Gruvberger, Cancer Res, 2001	Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns	cDNA microarray (6,728)		BT-474 tumor cell line			
West, PNAS, 2001	Predicting the clinical sstatus of human breast cancer by using gene expression profiles	Affymetrix 7,129 25-mer oligos	T1-T2: 1.5-5 cm		49	------	100 genes. Differentiate ER+ from ER-
Sorlie, PNAS,2001	Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications	cDNA microarray 8,102 features		Four normal breast tissue samples from different individuals, three of which were pooled, plus normal breast samples from multiple individuals (Clontech), gave three different batches of common reference, each differently produced	78 breast tumor samples (49 samples forsurvival analysis)		
Dan Cancer Res, 2002	An Integrated Database of Chemosensitivity to 55 Anticancer Drugs and Gene Expression Profiles of 39 Human Cancer Cell Lines	cDNA microarray (9,216)		mRNA pool from all 39 human cancer cell lines			
Ellis et al, Clin Cancer Research, 2002	Development and Validation of a Method for Using Breast Core Needle Biopsies for Gene Expression Microarray Analyses	nylon filter arrays (4032)		housekeeping genes, normalizing factor			top 30 breast cancer genes
Hyman, Kallioniemi, Cancer Research, 2002	Impact of DNA amplification on GE patterns in Breast Cancer	cDNA microarray, CGH analysis		Universal Human RNA reference (Stratagene)			
Van´t Veer, Nature, 2002	Gene Expression profiling predicts clinical outcome of breast cancer	Agilent 24,479 60-mer oligos	T1-T2: <5 cm	Reference cRNA made by pooling equal amounts of cRNA from each of the sporadic carcinomas	97 young patients (prediction set: 78 patients, test set: 19 patients) 44± 8 years	5 years minimum	70 genes. Risk of distant metastasis in LN- patients
Van de Vijver, N Engl J Med, 2002	A gene expression signature as a predictor of survival in breast cancer	Agilent 24,479 60-mer oligos	T1-T2: <5 cm	Reference cRNA made by pooling equal amounts of cRNA from each of the sporadic carcinomas	295	5 years minimum	70 genes. Risk of distant metastasis in LN- and LN+ patients
Pollack, PNAS 2002	Microarray analysis reveals a major direct role of DNA copy number alterations in the transcriptional program of human breast tumors	cDNA array (6,691), CGH analysis		Normal female leukocyte DNA from a single donor			

Appendix_A1: Breast cancer studies using microarrays

17/12/2007

Sotiriou, PNAS, 2003	Breast cancer classification and prognosis based on gene expression profiles from a population-based study	cDNA microarray (7,650 features, NCI)	T1-T2: <5 cm	Universal Human RNA reference (Stratagene)	99	6 yeras minimum	706 elements. 6 biological classes of invasive
Hedenfalk, PNAS, 2003	Molecular classification of familial non-BRCA1/BRCA2 breast cancer	cDNA microarray (Agilent), CGH analysis		Breast Cancer cell likne BT-474 (ATCC)			
Sorlie, PNAS,2003	Repeated obsevation of breast tumor samples subtypes in independent gene expression data sets	cDNA/Stanford 8,102 genes	T3-T4: >5 cm	Four normal breast tissue samples from different individuals, three of which were pooled, plus normal breast samples from multiple individuals (Clontech), gave three different batches of common reference, each differently produced	115 tumor samples (81 samples for training set)	8 years	
Ramaswamy et al. Nat Genet 2003	A molecular signature of metastasis in primary solid tumors	Rosetta inkjet (24,479 genes; breast adenocarcinoma) oligonucleotide microarrays.	78 small stage I primary breast adenocarcinomas		279 primary tumors of diverse types (lung, breast, prostate)		Identified a set of 128 genes separating metastasis from primary tumors able to distinguish patuients with good versus poor prognosis.Selection of the 17-gene signature associated with metastasis.
Huang et al, Lancet, 2003	Gene expression predictors of breast cancer outcomes	Hu U95Av2 Affymetrix			89 tumor samples		Identification of aggregates of gene expression (metagenes) that associate with lymph node status and recurrence, predicting outcome.
Selaru, Cancer Research, 2004	An unsupervised approach to identify molecular phenotypic components influencing breast cancer features	cDNA microarray (8,064 features)		Mixture containing aRNAs from eight human cancer cell lines			
Ma et al. Cancer Cell, 2004	A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen	22K oligonucleotide array	103 ER+ early stage cases	Universal Human RNA reference (Stratagene)	103 ER+ early stage cases	minimum 5 years	2 gene expression ratio to predict tumor recurrence in the setting of adjuvant tamoxifen therapy
Nielsen TO, et al. Clin Cancer Res, 2004	Immunohistochemical and clinical characterization of the Basal-like subtype of invasive breast carcinoma.	22K oligonucleotide array	T3-T4: >5 cm	Universal Human RNA reference (Stratagene) plus 1/10 of RNA from MCF7 cells plus 1/10 ME16C	115 breast primary tumors	median 17.4 years	
Nevins et al. 2003	Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction.	multiple gene expression signatures "metagenes"	86 LN+ breast cancer samples		86 LN+ breast cancer samples	5 years minimum	Determination of two metagenes of 117 genes and 31 genes, to predict clinical outcome
Zhao et al. ,Mol Biol Cell, 2004	Different gene expression patterns in invasive lobular and ductal carcinomas of the breast	cDNA microarray > 42,000 features (Stanford)	samples were IDC (38) or ILC (21)	Universal Human RNA reference (Stratagene)	59 primary breast cancer	——	78 clones selected to discriminate between IDC and ILC
Usary et al . Oncogene, 2004	Mutation of GATA3 in human breast cancer	Agilent Human 1A plus 3000 custom oligos		(SAM analysis multiclass (tumor grade i, II, III) breast cancer cell lines and 122 tumors from Sorlie	breast cancer cell line samples		identification of 74 GATA3 regulated genes
Michalides 2004	Tamoxifen resistance by a conformational arrest of the ERα after PKA activation in breast cancer	NKI 18K Human cDNA array	ER+ tumors stageI-III Tamoxifen versus no-Tamoxifen	Reference pool mada with 56 tumors (16 recurrence, 40 non-recurrence	56 tumors samples	median 132 months	phosporylation of serine-305 of ERα by protein kinase A (PKA) induces resistance to Tamoxifen
Van Laere et al. Breast Cancer Res Treat, 2005	Distinct molecular signature of inflammatory breast cancer by cDNA microarray analysis	cDNA microarray (Sanger center) 10,750 features	breast adenocarcinoma	Universal Human Reference RNA (Stratagene)	34 breast cancer samples		distinguish inflammatory breast cancer (IBC) and non-inflammatory breast cancer (non-IBC)

Appendix_A1: Breast cancer studies using microarrays

17/12/2007

West et al. PLOS Biology, 2005	Determination of the stromal signatures in breast carcinoma	42,000 cDNA microarray	early breast cancer (stage I and II)		58 tumors	7.8 years	SAM analysis between solitary fibrous tumor (SFT) and desmoid-type fibromatosis (DTF) with 700 discriminatory set
Chang et al 2004	Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds.	36,000 cDNA microarray	50 fibroblast cultures from 10 anatomic sites	Pool of mRNAs isolated from 11 different cultured cell lines			459 genes of fibroblast core serum response, the CSR gene set
Chang et al 2005	Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival.	36,000 cDNA microarray	T1-T2: <5 cm	Pool of cRNA from all samples	295	5 years minimum	459 genes of fibroblast core serum response, the CSR gene set
Wang et al. The Lancet 2005	Gene-expression profiles to predict distant metastasis of lymph-node negative primary breast cancer	Affymetrix U133A 25-mer oligos	LN- and LN+ patients with invasive breast cancer		286 lymph-node-negative patients (Training set 80 patients ER+, 35 patients ER-, test set of 171	median 101 months	76-gene signature to distinguish LN- primary breast cancer to develop distant metastasis within 5 years
Rouzier et al. Clin Cancer Res. 2005	Breast cancer molecular subtypes respond differently to preoperative chemotherapy	Affymetrix U133A 25-mer oligos	Stage I to III breast cancer		83 invasive breast tumors	-----	298 genes (SAM) to discriminate between four subtypes
Bertucci *et al.* Cancer Res. 2005	Gene expression profiling identifies molecular subtypes of inflamatory breast cancer	nylon membranes microarrays with 8016 spots			81 invasive breast cancer		120 commomn genes with Sorlie *et al.*
Naderi et al. Oncogene, 2006.	A gene-expression signature to predict survival in breast cancer across independent data sets	Agilent 22,575 60-mer oligos	No criteria for selection	reference cRNA pool generated by mixing the labelled targets from 50 samples chosen at random.	135 tumor samples (only quantity of RNA as criteria)	10-15 yeras	70-gen prognostic signature
Nuyten et al. Breast Cancer Res) 2006	Predicting a local recurrence after breast-conserving therapy by gene expression profiling	Agilent 24,479 60-mer oligos	Stage I and II breast cancer with age < 53 years	Reference cRNA made by pooling equal amounts of cRNA from each of the sporadic carcinomas	81 patients (training set), 80 patients validation set	6.2 years median	70-gen predictor of local recurrence
Feng *et al.* Breast Cancer Res Treat. 2006	Differentially expressed genes between primary cancer and paired lymph node metastases predict clinical outcome of node-positive breast cancer patients	Operon 70-mer two-color 21,239 probes	primary tumor and Lymph node metastasis paired samples	not stated	35 patients	43 months average	79 differentially expressed genes between primary cancers and metastasis samples
Sorlie et al. BMC Genomics 2006	Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across threee different platforms	Three different platforms: Agilent 44,000K, Applied Biosystems 31,700 60-mer; TaqMan array-based real time qPCR	early breast carcinomas (T1/T2)	Human reference RNA (Stratagene)	20 tumor biopsies	-----	54 discriminatory genes to discriminate luminal subtype and basal-like subtype
Teschendorff et al. Genome Biology 2007	An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer	Agilent/Affymetrix	240 breast tumor samples from different datasets		186 ER- samples integrated from different datase4ts		immune response related 7-gene module for identifying higher risk of distant metastasis
Kapp et al. BMC Genomics, 2007	Discovery and validation of breast cancer subtypes	datasets from Sorlie et al. 2003	datasets from Sorlie et al. 2003	datasets from Sorlie et al. 2003	98 breast samples		classification of ESR1+/ERBB2-, ESR1-/ERBB2-, and ERBB2 subtypes
Glas et al. BMC Genomics 2006	Converting a breast cancer microarray signature into a high-throughput diagnostic test	Agilent customized 1,900 60-mer oligos "MammaPrint"	young breast cancer patients < 55 years (LN-)	Pooled and amplified RNA from 105 primary breast tumors selected from patients of the clinical validation series	162 patients	minimum 5 years	232-gen gene signature to predict disease outcome
Yau et al. Breast Cancer Res 2007	Aging impacts transcriptme but not genome of hormone-dependent breast cancers	expression Affymetrix 13,000 and array-CGH analysis	two cohorts of 66 and 71 breast tumor mixed samples		two cohorts of 66 and 71 breast tumor mixed samples		two-age independent phenotypes

Appendix A2: CRG Human Breast Cancer Array v4.0-0.8K

GEO accesion number GPL5953

Symbol	Name	GenBank AccNo	UGRepAcc	UGCluster	Cytoband
28SrRNA	28SrRNA	28rRNA			
BC029276	BC029276	BC029276			
ABCC5	ATP-binding cassette, sub-family C (CFTR/MRP), member 5	NM_005688	AF146074	Hs.368563	3q27
ABL1	V-abl Abelson murine leukemia viral oncogene homolog 1	NM_005157	NM_007313	Hs.431048	9q34.1
ACTB	Actin, beta	NM_001101	AK125561	Hs.520640	7p15-p12
ADAM15	ADAM metallopeptidase domain 15 (metargidin)	NM_003815	AB209157	Hs.312098	1q21.3
PARP1	Poly (ADP-ribose) polymerase family, member 1	NM_001618	NM_001618	Hs.177766	1q41-q42
PARP3	Poly (ADP-ribose) polymerase family, member 3	NM_005485	NM_005485	Hs.271742	3p21.31-p21.1
ADRA1B	Adrenergic, alpha-1B-, receptor	NM_000679	NM_000679	Hs.368632	5q23-q32
ADRA1B	Adrenergic, alpha-1B-, receptor	NM_000679	NM_000679	Hs.368632	5q23-q32
ADRBK1	Adrenergic, beta, receptor kinase 1	NM_001619	AB209588	Hs.83636	11q13
AHR	Aryl hydrocarbon receptor	NM_001621	NM_001621	Hs.171189	7p15
AKAP13	A kinase (PRKA) anchor protein 13	NM_007200	NM_006738	Hs.459211	15q24-q25
PALM2-AKAP2	PALM2-AKAP2 protein	NM_007203	NM_053016	Hs.591908	9q31-q33
AKAP9	A kinase (PRKA) anchor protein (yotiao) 9	NM_147171	NM_147171	Hs.527348	7q21-q22
AKT1	V-akt murine thymoma viral oncogene homolog 1	NM_005163	NM_005163	Hs.525622	14q32.32|14q32.32
AKT2	V-akt murine thymoma viral oncogene homolog 2	NM_001626	AK122839	Hs.541273	19q13.1-q13.2
AKT3	V-akt murine thymoma viral oncogene homolog 3 (protein kinase B, gamma	NM_005465	NM_005465	Hs.498292	1q43-q44
ALDH4A1	Aldehyde dehydrogenase 4 family, member A1	NM_003748	NM_003748	Hs.77448	1p36
ALPP	Alkaline phosphatase, placental (Regan isozyme)	NM_001632	BC094743	Hs.284255	2q37
NUSAP1	Nucleolar and spindle associated protein 1	NM_016359	BC037888	Hs.406234	15q15.1
ANTXR1	Anthrax toxin receptor 1	NM_032208	AF279145	Hs.165859	2p13.1
AP2B1	Adaptor-related protein complex 2, beta 1 subunit	NM_001282	NM_001030006	Hs.514819	17q11.2-q12
APAF1	Apoptotic peptidase activating factor	NM_013229	NM_181861	Hs.552567	12q23
APC	Adenomatosis polyposis coli	NM_000038	NM_000038	Hs.158932	5q21-q22
APEX1	APEX nuclease (multifunctional DNA repair enzyme) 1	BC002338	CR611116	Hs.73722	14q11.2-q12
APEX1	APEX nuclease (multifunctional DNA repair enzyme) 1	NM_001641	CR611116	Hs.73722	14q11.2-q12
APEX2	APEX nuclease (apurinic/apyrimidinic endonuclease) 2	NM_014481	NM_014481	Hs.555936	Xp11.21
APOD	Apolipoprotein D	NM_001647	BF790155	Hs.522555	3q26.2-qter
AR	Androgen receptor	NM_000044	NM_000044	Hs.496240	Xq11.2-q12
ARAF	V-raf murine sarcoma 3611 viral oncogene homolog	NM_001654	AB208831	Hs.446641	Xp11.4-p11.2
AREG	Amphiregulin (schwannoma-derived growth factor)	NM_001657	BC009799	Hs.270833	4q13-q21
RND3	Rho family GTPase 3	X97758	X97758	Hs.6838	2q23.3
ARHGAP5	Rho GTPase activating protein 5	NM_001173	NM_001030055	Hs.592313	14q12
ATE1	Arginyltransferase 1	NM_007041	BC022026	Hs.501239	10q26.13
ATF2	Activating transcription factor 2	NM_001880	BC107698	Hs.591614	2q32
ATF3	Activating transcription factor 3	NM_004024	AB209032	Hs.460	1q32.3
ATF4	Activating transcription factor 4 (tax-responsive enhancer element B67)	NM_001675	NM_001675	Hs.496487	22q13.1
ATM	Ataxia telangiectasia mutated (includes complementation groups A, C and D	NM_000051	NM_000051	Hs.435561	11q22-q23
ATR	Ataxia telangiectasia and Rad3 related	NM_001184	NM_001184	Hs.271791	3q22-q24
ATRX	Alpha thalassemia/mental retardation syndrome X-linked (RAD54 homolog,	U09820	NM_138271	Hs.533526	Xq13.1-q21.1
AURKB	Aurora kinase B	NM_004217	CD049340	Hs.442658	17p13.1
AURKC	Aurora kinase C	NM_003160	AF059681	Hs.98338	19q13.43
BAD	BCL2-antagonist of cell death	NM_004322	AK023420	Hs.370254	11q13.1
ACTL6A	Actin-like 6A	NM_178042	NM_178042	Hs.435326	3q26.33
BAG1	BCL2-associated athanogene	NM_004323	NM_004323	Hs.377484	9p12
BARD1	BRCA1 associated RING domain 1	NM_000465	AK223409	Hs.591642	2q34-q35
BAX	BCL2-associated X protein	NM_138764	AK001361	Hs.433670	19q13.3-q13.4
BAZ1A	Bromodomain adjacent to zinc finger domain, 1A	NM_013448	NM_013448	Hs.592311	14q12-q13
BBC3	BCL2 binding component 3	NM_014417	AF332558	Hs.467020	19q13.3-q13.4
BCAR1	Breast cancer anti-estrogen resistance 1	NM_014567	AK124526	Hs.479747	16q22-q23
BCAS2	Breast carcinoma amplified sequence 2	NM_005872	BC022880	Hs.22960	1p21-p13.3
BCCIP	BRCA2 and CDKN1A interacting protein	NM_078469	AK092054	Hs.370292	10q26.1
BCL2	B-cell CLL/lymphoma 2	NM_000633	NM_000633	Hs.592350	18q21.33|18q21.3
BCL2L1	BCL2-like 1	NM_138578	CR936637	Hs.516966	20q11.21
BCL2L2	BCL2-like 2	NM_004050	AK024489	Hs.410026	14q11.2-q12
ACSBG2	Acyl-CoA synthetase bubblegum family member 2	NM_030924	AY358766	Hs.465720	19p13.3
BIRC2	Baculoviral IAP repeat-containing 2	NM_001166	BC028578	Hs.503704	11q22
BIRC3	Baculoviral IAP repeat-containing 3	NM_001165	NM_001165	Hs.127799	11q22
BIRC5	Baculoviral IAP repeat-containing 5 (survivin)	NM_001168	NM_001012271	Hs.514527	17q25
BIRC5	Baculoviral IAP repeat-containing 5 (survivin)	NM_001168	NM_001012271	Hs.514527	17q25
BLM	Bloom syndrome	NM_000057	BC034480	Hs.169348	15q26.1
BNIP3	BCL2/adenovirus E1B 19kDa interacting protein 3	NM_004052	BX647339	Hs.144873	10q26.3
BRAF	V-raf murine sarcoma viral oncogene homolog B1	NM_004333	NM_004333	Hs.550061	7q34
BRAF	V-raf murine sarcoma viral oncogene homolog B1	NM_004333	NM_004333	Hs.550061	7q34
BRCA1	Breast cancer 1, early onset	NM_007295	NM_007295	Hs.194143	17q21
BRCA2	Breast cancer 2, early onset	NM_000059	NM_000059	Hs.34012	13q12.3
BTG1	B-cell translocation gene 1, anti-proliferative	NM_001731	BC009050	Hs.255935	12q22
BTG2	BTG family, member 2	NM_006763	NM_006763	Hs.519162	1q32
BTN2A2	Butyrophilin, subfamily 2, member A2	NM_006995	U90550	Hs.373938	6p22.1
C14orf130	Chromosome 14 open reading frame 130	BU739864	NM_018108	Hs.275352	14q32.12
C14orf138	Chromosome 14 open reading frame 138	NM_024558	BX247997	Hs.558541	14q22.1
C20orf149	Chromosome 20 open reading frame 149	NM_024299	BG168849	Hs.79625	20q13.33
C20orf46	Chromosome 20 open reading frame 46	NM_018354	AK126837	Hs.516834	20p13
VHL	Von Hippel-Lindau tumor suppressor	NM_018462	NM_000551	Hs.517792	3p26-p25
C9orf3	Chromosome 9 open reading frame 3	BX372918	AF043897	Hs.434253	9q22.32
CA2	Carbonic anhydrase II	NM_000067	AK123309	Hs.155097	8q22
CAMK2N1	Calcium/calmodulin-dependent protein kinase II inhibitor 1	NM_018584	CR604926	Hs.197922	1p36.12
CALR	Calreticulin	NM_004343	M84739	Hs.515162	19p13.3-p13.2
CARM1	Coactivator-associated arginine methyltransferase 1	XM_032719	NM_199141	Hs.371416	19p13.2
CASP1	Caspase 1, apoptosis-related cysteine peptidase	NM_033292	AK223503	Hs.2490	11q23
CAV1	Caveolin 1, caveolae protein, 22kDa	NM_001753	NM_001753	Hs.74034	7q31.1
CBFB	Core-binding factor, beta subunit	NM_001755	NM_001755	Hs.460988	16q22.1
CCNA2	Cyclin A2	NM_001237	CR604810	Hs.58974	4q25-q31
CCNB1	Cyclin B1	NM_031966	BX537394	Hs.23960	5q12
CCNB2	Cyclin B2	NM_004701	AL080146	Hs.194698	15q22.2
CCNC	Cyclin C	NM_005190	BC041123	Hs.430646	6q21

CCND1	Cyclin D1	NM_053056	NM_053056	Hs.523852	11q13
CCND2	Cyclin D2	NM_001759	NM_001759	Hs.376071	12p13
CCND3	Cyclin D3	NM_001760	AK096276	Hs.534307	6p21
CCNE1	Cyclin E1	NM_001238	BC035498	Hs.244723	19q12
CCNE2	Cyclin E2	NM_057749	NM_057749	Hs.567387	8q22.1
CCNF	Cyclin F	NM_001761	NM_001761	Hs.1973	16p13.3
CCNG1	Cyclin G1	NM_004060	NM_004060	Hs.79101	5q32-q34
CCNG2	Cyclin G2	NM_004354	BC032518	Hs.13291	4q21.1
CCNH	Cyclin H	NM_001239	AB209342	Hs.292524	5q13.3-q14
CCNI	Cyclin I	NM_006835	NM_006835	Hs.591702	4q21.1
CCNK	Cyclin K	XM_085179	AB209373	Hs.510409	14q32
CCNT1	Cyclin T1	NM_001240	NM_001240	Hs.279906	12pter-qter
CCNT2	Cyclin T2	NM_058241	BX648174	Hs.591241	2q21.3
CCRK	Cell cycle related kinase	NM_178432	AF113130	Hs.522274	9q22.1
CD24	CD24 molecule	AK125531	AK057112	Hs.375108	6q21
CD34	CD34 molecule	BX640941	BX640941	Hs.374990	1q32
SEPT7	Septin 7	NM_001788	AB209677	Hs.191346	7p14.3-p14.1
CDC14A	CDC14 cell division cycle 14 homolog A (S. cerevisiae)	NM_003672	BC071578	Hs.127411	1p21
CDC14B	CDC14 cell division cycle 14 homolog B (S. cerevisiae)	NM_033332	AF064105	Hs.40582	9q22.33
CDC16	CDC16 cell division cycle 16 homolog (S. cerevisiae)	NM_003903	AB209850	Hs.374127	13q34
CDC2	Cell division cycle 2, G1 to S and G2 to M	NM_001786	CR933728	Hs.334562	10q21.1
CDC20	CDC20 cell division cycle 20 homolog (S. cerevisiae)	NM_001255	BG256659	Hs.524947	1p34.1
CDC20	CDC20 cell division cycle 20 homolog (S. cerevisiae)	NM_001255	BG256659	Hs.524947	1p34.1
CDC23	CDC23 (cell division cycle 23, yeast, homolog)	NM_004661	NM_004661	Hs.153546	5q31
CDC25A	Cell division cycle 25A	NM_001789	NM_001789	Hs.437705	3p21
CDC25B	Cell division cycle 25B	NM_021874	NM_021873	Hs.153752	20p13
CDC25C	Cell division cycle 25C	NM_001790	BC039100	Hs.656	5q31
CDC27	Cell division cycle 27	NM_001256	S78234	Hs.463295	17q12-17q23.2
CDC2L5	Cell division cycle 2-like 5 (cholinesterase-related cell division controller)	NM_003718	NM_003718	Hs.233552	7p13
CDC34	Cell division cycle 34	NM_004359	BM906315	Hs.514997	19p13.3
CDC37	CDC37 cell division cycle 37 homolog (S. cerevisiae)	NM_007065	NM_007065	Hs.160958	19p13.2
CDC42	Cell division cycle 42 (GTP binding protein, 25kDa)	NM_001791	BC018266	Hs.487266	1p36.1
CDC42BPA	CDC42 binding protein kinase alpha (DMPK-like)	NM_003607	NM_003607	Hs.35433	1q42.11
CDC42BPB	CDC42 binding protein kinase beta (DMPK-like)	NM_006035	NM_006035	Hs.569310	14q32.3
CDC45L	CDC45 cell division cycle 45-like (S. cerevisiae)	NM_003504	NM_003504	Hs.474217	22q11.21
CDC6	CDC6 cell division cycle 6 homolog (S. cerevisiae)	NM_001254	NM_001254	Hs.405958	17q21.3
CDC7	CDC7 cell division cycle 7 (S. cerevisiae)	NM_003503	AB209337	Hs.533573	1p22
CDCA7	Cell division cycle associated 7	NM_031942	AL834186	Hs.470654	2q31
CDH1	Cadherin 1, type 1, E-cadherin (epithelial)	NM_004360	NM_004360	Hs.461086	16q22.1
CDH13	Cadherin 13, H-cadherin (heart)	NM_001257	NM_001257	Hs.436040	16q24.2-q24.3
CDK10	Cyclin-dependent kinase (CDC2-like) 10	NM_003674	BC045670	Hs.109	16q24
CDK2	Cyclin-dependent kinase 2	NM_001798	NM_001798	Hs.19192	12q13
CDK3	Cyclin-dependent kinase 3	NM_001258	BX647274	Hs.584745	17q22-qter
CDK4	Cyclin-dependent kinase 4	NM_000075	BM467999	Hs.95577	12q14
CDK5	Cyclin-dependent kinase 5	NM_004935	BG577212	Hs.166071	7q36
CDK6	Cyclin-dependent kinase 6	NM_001259	NM_001259	Hs.119882	7q21-q22
CDK7	Cyclin-dependent kinase 7	NM_001799	NM_001799	Hs.184298	5q12.1
CDK8	Cyclin-dependent kinase 8	NM_001260	BC047364	Hs.382306	13q12
CDK9	Cyclin-dependent kinase 9 (CDC2-related kinase)	NM_001261	NM_001261	Hs.557646	9q34.1
CDK9	Cyclin-dependent kinase 9 (CDC2-related kinase)	NM_001261	NM_001261	Hs.557646	9q34.1
CDKL1	Cyclin-dependent kinase-like 1 (CDC2-related kinase)	NM_004196	AF390028	Hs.280881	14q22.1
CDKL2	Cyclin-dependent kinase-like 2 (CDC2-related kinase)	NM_003948	NM_003948	Hs.591698	4q21.1
CDKL3	Cyclin-dependent kinase-like 3	NM_016508	BC041799	Hs.105818	5q31
CDKN1A	Cyclin-dependent kinase inhibitor 1A (p21, Cip1)	NM_078467	NM_078467	Hs.370771	6p21.2
CDKN1B	Cyclin-dependent kinase inhibitor 1B (p27, Kip1)	NM_004064	NM_004064	Hs.238990	12p13.1-p12
CDKN1C	Cyclin-dependent kinase inhibitor 1C (p57, Kip2)	NM_000076	BC067842	Hs.106070	11p15.5
CDKN2A	Cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)	NM_058197	BQ945397	Hs.512599	9p21
CDKN2B	Cyclin-dependent kinase inhibitor 2B (p15, inhibits CDK4)	NM_078487	NM_078487	Hs.72901	9p21
CDKN2C	Cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)	NM_001262	AK091170	Hs.525324	1p32
CDKN2D	Cyclin-dependent kinase inhibitor 2D (p19, inhibits CDK4)	NM_001800	NM_001800	Hs.435051	19p13
CDKN3	Cyclin-dependent kinase inhibitor 3 (CDK2-associated dual specificity phosp	NM_005192	BQ056337	Hs.84113	14q22
CEBPA	CCAAT/enhancer binding protein (C/EBP), alpha	NM_004364	NM_004364	Hs.590973	19q13.1
CEBPB	CCAAT/enhancer binding protein (C/EBP), beta	BC021931	BC021931	Hs.592138	20q13.1
CEBPB	CCAAT/enhancer binding protein (C/EBP), beta	BC021931	BC021931	Hs.592138	20q13.1
CEBPD	CCAAT/enhancer binding protein (C/EBP), delta	BM924801	BM924801	Hs.440829	8p11.2-p11.1
CEBPE	CCAAT/enhancer binding protein (C/EBP), epsilon	CR594219	CR594219	Hs.558308	14q11.2
CENPA	Centromere protein A, 17kDa	NM_001809	BM911202	Hs.1594	2p24-p21
CENPA	Centromere protein A, 17kDa	NM_001809	BM911202	Hs.1594	2p24-p21
CFTR	Cystic fibrosis transmembrane conductance regulator, ATP-binding cassette	NM_000492	NM_000492	Hs.489786	7q31.2
CHD1	Chromodomain helicase DNA binding protein 1	NM_001270	NM_001270	Hs.121098	5q15-q21
CHD1L	Chromodomain helicase DNA binding protein 1-like	NM_024568	AF537213	Hs.191164	1q12
CHD2	Chromodomain helicase DNA binding protein 2	NM_001271	NM_001271	Hs.220864	15q26
CHD2	Chromodomain helicase DNA binding protein 2	NM_001271	NM_001271	Hs.220864	15q26
CHD3	Chromodomain helicase DNA binding protein 3	AK096555	AK125928	Hs.596899	17p13.1
CHD1	Chromodomain helicase DNA binding protein 1	AK096553	AK125926	Hs.596897	17p13.1
CHD4	Chromodomain helicase DNA binding protein 4	NM_001273	BC038596	Hs.162233	12p13
CHEK1	CHK1 checkpoint homolog (S. pombe)	NM_001274	NM_001274	Hs.24529	11q24-q24
CHEK2	CHK2 checkpoint homolog (S. pombe)	NM_007194	AF217975	Hs.291363	22q11|22q12.1
CHES1	Checkpoint suppressor 1	NM_005197	NM_005197	Hs.434286	14q31.3
chr6p21.3_HistoneClu	chr6p21.3_HistoneCluster	chr6p21.3_HistoneCluster			
chr6p22.2_HistoneClu	chr6p22.2_HistoneCluster	chr6p22.2_HistoneCluster			
CKB	Creatine kinase, brain	NM_001823	BC040666	Hs.173724	14q32
CKS1B	CDC28 protein kinase regulatory subunit 1B	NM_001826	BQ278454	Hs.374378	1q21.2
CKS2	CDC28 protein kinase regulatory subunit 2	NM_001827	BQ898943	Hs.83758	9q22
CLK1	CDC-like kinase 1	NM_004071	NM_004071	Hs.433732	2q33
CLK2	CDC-like kinase 2	NM_003993	AK091036	Hs.73986	1q21
CLK3	CDC-like kinase 3	NM_003992	CR933693	Hs.584748	15q24
COL1A1	Collagen, type I, alpha 1	NM_000088	Z74615	Hs.591172	17q21.33
COL4A1	Collagen, type IV, alpha 1	NM_001845	NM_001845	Hs.17441	13q34
COL4A2	Collagen, type IV, alpha 2	NM_001846	NM_001846	Hs.508716	13q34
CPNE4	Copine IV	NM_130808	AK128117	Hs.199877	3q22.1

Appendix A2: CRG Human Breast Cancer Array v4.0-0.8K

GEO accesion number GPL5953

CCPG1	Cell cycle progression 1	NM_004748	NM_020739	Hs.126115	15q21.1
CR1L	Complement component (3b/4b) receptor 1-like	AL137789	XM_114735	Hs.632488	1q32.1
CREBBP	CREB binding protein (Rubinstein-Taybi syndrome)	NM_004380	NM_004380	Hs.459759	16p13.3
CSE1L	CSE1 chromosome segregation 1-like (yeast)	NM_001316	NM_001316	Hs.90073	20q13
CSF1R	Colony stimulating factor 1 receptor, formerly McDonough feline sarcoma vi	NM_005211	X03663	Hs.483829	5q33-q35
CSH1	Chorionic somatomammotropin hormone 1 (placental lactogen)	NM_022640	CR610932	Hs.406754	17q24.2
CTHRC1	Collagen triple helix repeat containing 1	BC021025	BC021025	Hs.405614	8q22.3
CTNNA1	Catenin (cadherin-associated protein), alpha 1, 102kDa	NM_001903	NM_001903	Hs.534797	5q31
CTNNB1	Catenin (cadherin-associated protein), beta 1, 88kDa	NM_001904	BC058926	Hs.476018	3p21
CTSB	Cathepsin B	NM_147780	NM_147780	Hs.520898	8p22
CTSD	Cathepsin D (lysosomal aspartyl peptidase)	NM_001909	AK022293	Hs.121575	11p15.5
CTSL	Cathepsin L	NM_001912	AK055599	Hs.418123	9q21-q22
CXCL12	Chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)	NM_000609	BX647204	Hs.522891	10q11.1
CXCR4	Chemokine (C-X-C motif) receptor 4	NM_003467	AF147204	Hs.421986	2q21
BRCC3	BRCA1/BRCA2-containing complex, subunit 3	NM_024332	NM_024332	Hs.558537	Xq28
CYP19A1	Cytochrome P450, family 19, subfamily A, polypeptide 1	NM_000103	NM_031226	Hs.511367	15q21.1
CYP2B6	Cytochrome P450, family 2, subfamily B, polypeptide 6	NM_000767	NM_000767	Hs.1360	19q13.2
CYR61	Cysteine-rich, angiogenic inducer, 61	Y11307	Y11307	Hs.8867	1p31-p22
C16orf61	Chromosome 16 open reading frame 61	NM_020188	BM463756	Hs.388255	16q23.2
C1orf48	Chromosome 1 open reading frame 48	NM_015471	AF255793	Hs.497692	1q41
DCK	Deoxycytidine kinase	NM_000788	CD014015	Hs.709	4q13.3-q21.1
DCLRE1A	DNA cross-link repair 1A (PSO2 homolog, S. cerevisiae)	NM_014881	D42045	Hs.1560	10q25.1
DDB1	Damage-specific DNA binding protein 1, 127kDa	NM_001923	BC050530	Hs.290758	11q12-q13
DDB2	Damage-specific DNA binding protein 2, 48kDa	NM_000107	BC050455	Hs.446564	11p12-p11
DDIT3	DNA-damage-inducible transcript 3	NM_004083	BC107859	Hs.505777	12q13.1-q13.2
DDR1	Discoidin domain receptor family, member 1	NM_013994	AB067472	Hs.485070	6p21.3
DDX5	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5	NM_004396	BX571764	Hs.279806	17q21
DDX6	DEAD (Asp-Glu-Ala-Asp) box polypeptide 6	NM_004397	NM_004397	Hs.408461	11q23.3
DHX9	DEAH (Asp-Glu-Ala-His) box polypeptide 9	NM_001357	L13848	Hs.191518	1q25
DES	Desmin	NM_001927	BC032116	Hs.471419	2q35
DHRS7	Dehydrogenase/reductase (SDR family) member 7	NM_016029	BU541074	Hs.59719	14q23.1
DIAPH3	Diaphanous homolog 3 (Drosophila)	NM_030932	AY750055	Hs.283127	13q21.2
DIAPH3L	DIAPH3L	BC041395	BC041395		
DICER1	Dicer1, Dcr-1 homolog (Drosophila)	NM_177438	NM_177438	Hs.87889	14q32.13
MKLN1	Muskelin 1, intracellular mediator containing kelch motifs	BX648653	NM_013255	Hs.44693	7q32
DKFZp686P03110	DKFZp686P03110	AC097632	AC097632		
RPL27	Ribosomal protein L27	NM_173079	BC039247	Hs.514196	17q21.1-q21.2
RUNX2	Runt-related transcription factor 2	AL353944	NM_004348	Hs.535845	6p21
DLC1	Deleted in liver cancer 1	NM_006094	NM_182643	Hs.134296	8p22
DLX4	Distal-less homeobox 4	NM_138281	BC005812	Hs.591167	17q21.33
DNA2L	DNA2 DNA replication helicase 2-like (yeast)	XM_166103	D42046	Hs.532446	10q21.3-q22.1
DNMT1	DNA (cytosine-5-)-methyltransferase 1	NM_001379	NM_001379	Hs.202672	19p13.2
DNMT3A	DNA (cytosine-5-)-methyltransferase 3 alpha	NM_175629	AB208833	Hs.515840	2p23
DNMT3B	DNA (cytosine-5-)-methyltransferase 3 beta	NM_006892	DQ321787	Hs.570374	20q11.2
DUSP1	Dual specificity phosphatase 1	NM_004417	AK127679	Hs.171695	5q34
DUSP6	Dual specificity phosphatase 6	BC037236	BC037236	Hs.298654	12q22-q23
APBA2BP	Amyloid beta (A4) precursor protein-binding, family A, member 2 binding pro	NM_005225	BC050369	Hs.516986	20q11.22
E2F2	E2F transcription factor 2	NM_004091	NM_004091	Hs.194333	1p36
E2F3	E2F transcription factor 3	NM_001949	NM_001949	Hs.269408	6p22
E2F4	E2F transcription factor 4, p107/p130-binding	NM_001950	NM_001950	Hs.108371	16q21-q22
E2F5	E2F transcription factor 5, p130-binding	NM_001951	AB209185	Hs.445758	8q21.2
E2F6	E2F transcription factor 6	NM_001952	NM_198258	Hs.135465	2p25.1
ECT2	Epithelial cell transforming sequence 2 oncogene	NM_018098	AY376439	Hs.518299	3q26.1-q26.2
EEF1A1	Eukaryotic translation elongation factor 1 alpha 1	NM_001402	NM_001402	Hs.439552	6q14.1
EGF	Epidermal growth factor (beta-urogastrone)	NM_001963	NM_001963	Hs.419815	4q25
EGFR	Epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) onc	NM_005228	NM_005228	Hs.488293	7p12
EGLN1	Egl nine homolog 1 (C. elegans)	NM_022051	AF229245	Hs.444450	1q42.1
EGR1	Early growth response 1	NM_001964	NM_001964	Hs.326035	5q31.1
ELL	Elongation factor RNA polymerase II	NM_006532	NM_006532	Hs.515260	19p13.1
ELL2	Elongation factor, RNA polymerase II, 2	NM_012081	BX538289	Hs.192221	5q15
CTTN	Cortactin	NM_005231	NM_003626	Hs.530749	11q13
C11orf30	Chromosome 11 open reading frame 30	NM_020193	NM_020193	Hs.352588	11q13.5
ENPP2	Ectonucleotide pyrophosphatase/phosphodiesterase 2 (autotaxin)	NM_006209	NM_006209	Hs.190977	8q24.1
EP300	E1A binding protein p300	NM_001429	U01877	Hs.517517	22q13.2
ERBB2	V-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblasto	NM_004448	NM_001005862	Hs.446352	17q11.2-q12|17q21.1
PA2G4	Proliferation-associated 2G4, 38kDa	NM_001982	NM_001982	Hs.524498	12q13
ERBB4	V-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)	NM_005235	BX537810	Hs.390729	2q33.3-q34
ERCC2	Excision repair cross-complementing rodent repair deficiency, complementa	NM_000400	AK092872	Hs.487294	19q13.3
ERCC3	Excision repair cross-complementing rodent repair deficiency, complementa	NM_000122	AK127469	Hs.469872	2q21
PGBD3	PiggyBac transposable element derived 3	NM_000124	NM_000124	Hs.133444	10q11
ESM1	Endothelial cell-specific molecule 1	NM_007036	X89426	Hs.129944	5q11.2
ESR1	Estrogen receptor 1	NM_000125	NM_000125	Hs.208124	6q25.1
ESR2	Estrogen receptor 2 (ER beta)	NM_001437	AB209620	Hs.443150	14q23.2
ESRRA	Estrogen-related receptor alpha	NM_004451	BC092470	Hs.110849	11q13
ESRRB	Estrogen-related receptor beta	NM_004452	NM_004452	Hs.435845	14q24.3
ESRRG	Estrogen-related receptor gamma	NM_001438	BC064700	Hs.444225	1q41
IGSF4	Immunoglobulin superfamily, member 4	AV728294	BX641042	Hs.370510	11q23.2
KREMEN1	Kringle containing transmembrane protein 1	BG741722	NM_001039570	Hs.229335	22q12.1
ESTs_BM701446_ok	Transcribed locus	BM701446	BG677838	Hs.446388	
CRYL1	Crystallin, lambda 1	BX092299	BC071810	Hs.370703	13q12.11
TACC2	Transforming, acidic coiled-coil containing protein 2	BX111019	AF528099	Hs.501252	10q26
ESTs_H73518	ESTs_H73518	AC093770	AC093770		
ETS1	V-ets erythroblastosis virus E26 oncogene homolog 1 (avian)	NM_005238	NM_005238	Hs.369438	11q23.3
ETV5	Ets variant gene 5 (ets-related molecule)	NM_004454	NM_004454	Hs.43697	3q28
EXO1	Exonuclease 1	NM_130398	NM_130398	Hs.498248	1q42-q43
EXT1	Exostoses (multiple) 1	NM_000127	BX537744	Hs.492618	8q24.11-q24.13
FANCA	Fanconi anemia, complementation group A	NM_000135	X99226	Hs.567267	16q24.3
FANCC	Fanconi anemia, complementation group C	NM_000136	NM_000136	Hs.494529	9q22.3
FANCD2	Fanconi anemia, complementation group D2	NM_033084	BC038666	Hs.208388	3p26
FANCF	Fanconi anemia, complementation group F	NM_022725	NM_022725	Hs.523543	11p15
FANCG	Fanconi anemia, complementation group G	NM_004629	AJ007669	Hs.591084	9p13

FGF1	Fibroblast growth factor 1 (acidic)	NM_000800	NM_000800	Hs.483635	5q31
FGF18	Fibroblast growth factor 18	NM_003862	AF075292	Hs.87191	5q34
FGF2	Fibroblast growth factor 2 (basic)	NM_002006	NM_002006	Hs.284244	4q26-q27
FGF7	Fibroblast growth factor 7 (keratinocyte growth factor)	NM_002009	NM_002009	Hs.567268	15q15-q21.1
FGFR2	Fibroblast growth factor receptor 2 (bacteria-expressed kinase, keratinocyte	NM_023028	NM_023028	Hs.533683	10q26
FHIT	Fragile histidine triad gene	NM_002012	BX641016	Hs.196981	3p14.2
AYTL2	Acyltransferase like 2	NM_024830	AK090444	Hs.368853	5p15.33
FLJ13611	Hypothetical protein FLJ13611	NM_024941	BX641157	Hs.591760	5q12.3
NPAL2	NIPA-like domain containing 2	NM_024759	AK025015	Hs.309489	8q22.2
FLJ22222	Hypothetical protein FLJ22222	NM_175902	BM923625	Hs.567578	17q25.3
ZNF533	Zinc finger protein 533	NM_152520	BC092423	Hs.580076	2q31.2-q31.3
RECQL5	RecQ protein-like 5	AL832643	AL832643	Hs.528605	17q25.2-q25.3
C1orf71	Chromosome 1 open reading frame 71	NM_152609	AL834246	Hs.368353	1q44
PQLC2	PQ loop repeat containing 2	AK095564	BC021040	Hs.523036	1p36.13
FLJ38964	Hypothetical protein FLJ38964	NM_173527	NM_173527	Hs.444911	14q11.2
SNRPD1	Small nuclear ribonucleoprotein D1 polypeptide 16kDa	NM_173484	NM_173484	Hs.567674	18q11.2
PTCH	Patched homolog (Drosophila)	AK124593	AB209495	Hs.494538	9q22.3
FLT1	Fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular p	NM_002019	NM_002019	Hs.507621	13q12
FN1	Fibronectin 1	NM_002026	NM_212482	Hs.203717	2q34
FOS	V-fos FBJ murine osteosarcoma viral oncogene homolog	NM_005252	BX647104	Hs.25647	14q24.3
FOSL1	FOS-like antigen 1	NM_005438	AL833733	Hs.283565	11q13
FOXO3A	Forkhead box O3A	NM_001455	BX649171	Hs.591328	6q21
FUS	Fusion (involved in t(12;16) in malignant liposarcoma)	NM_004960	AB208902	Hs.513522	16p11.2
GADD45A	Growth arrest and DNA-damage-inducible, alpha	NM_001924	CR612719	Hs.80409	1p31.2-p31.1
GADD45B	Growth arrest and DNA-damage-inducible, beta	NM_015675	AF087853	Hs.110571	19p13.3
GADD45G	Growth arrest and DNA-damage-inducible, gamma	NM_006705	CR613579	Hs.9701	9q22.1-q22.2
GAK	Cyclin G associated kinase	NM_005255	NM_005255	Hs.369607	4p16
GAS1	Growth arrest-specific 1	NM_002048	NM_002048	Hs.65029	9q21.3-q22
GAS1	Growth arrest-specific 1	NM_002048	NM_002048	Hs.65029	9q21.3-q22
GAS6	Growth arrest-specific 6	NM_000820	AL833761	Hs.369201	13q34
GAS8	Growth arrest-specific 8	NM_001481	CR749499	Hs.431792	16q24.3
GATA3	GATA binding protein 3	NM_002051	NM_001002295	Hs.524134	10p15
GEM	GTP binding protein overexpressed in skeletal muscle	NM_005261	NM_005261	Hs.345139	8q13-q21
genomic_chr9_ok/AL5	genomic_chr9_ok/AL513503.2	AL513503.2			
GJA1	Gap junction protein, alpha 1, 43kDa (connexin 43)	NM_000165	NM_000165	Hs.74471	6q21-q23.2
GLTSCR2	Glioma tumor suppressor candidate region gene 2	NM_015710	NM_000554	Hs.421907	19q13.3
GM2A	GM2 ganglioside activator	NM_000405	NM_000405	Hs.483873	5q31.3-q33.1
GMPS	Guanine monphosphate synthetase	NM_003875	NM_003875	Hs.591314	3q24
GNAZ	Guanine nucleotide binding protein (G protein), alpha z polypeptide	NM_002073	BC037333	Hs.584760	22q11.22
GPR126	G protein-coupled receptor 126	NM_020455	NM_020455	Hs.318894	6q24.1
GPR30	G protein-coupled receptor 30	NM_001505	NM_001505	Hs.592167	7p22.3
GRB14	Growth factor receptor-bound protein 14	NM_004490	NM_004490	Hs.411881	2q22-q24
GRB2	Growth factor receptor-bound protein 2	NM_002086	NM_002086	Hs.444356	17q24-q25
GSK3A	Glycogen synthase kinase 3 alpha	NM_019884	BC027984	Hs.466828	19q13.2
GSK3B	Glycogen synthase kinase 3 beta	NM_002093	BC012760	Hs.445733	3q13.3
GSTM3	Glutathione S-transferase M3 (brain)	NM_000849	NM_000849	Hs.2006	1p13.3
GTF2B	General transcription factor IIB	NM_001514	BC021000	Hs.481852	1p22-p21
GTF2E1	General transcription factor IIE, polypeptide 1, alpha 56kDa	NM_005513	NM_005513	Hs.445272	3q21-q24
GTF2E2	General transcription factor IIE, polypeptide 2, beta 34kDa	NM_002095	CR606311	Hs.77100	8p21-p12
GTF2E2	General transcription factor IIE, polypeptide 2, beta 34kDa	NM_002095	CR606311	Hs.77100	8p21-p12
GTF2F1	General transcription factor IIF, polypeptide 1, 74kDa	NM_002096	NM_002096	Hs.68257	19p13.3
GTF2F2	General transcription factor IIF, polypeptide 2, 30kDa	NM_004128	CR609785	Hs.58593	13q14
LDHA	Lactate dehydrogenase A	NM_005316	NM_005316	Hs.2795	11p15.4
GTF2H2	General transcription factor IIH, polypeptide 2, 44kDa	NM_001515	AF078847	Hs.398348	5q12.2-q13.3
GTF2H3	General transcription factor IIH, polypeptide 3, 34kDa	NM_001516	BU956289	Hs.355348	12q24.31
DDR1	Discoidin domain receptor family, member 1	NM_001517	AB067472	Hs.485070	6p21.3
H1F0	H1 histone family, member 0	NM_005318	NM_005318	Hs.226117	22q13.1
H1FX	H1 histone family, member X	NM_006026	BC000426	Hs.75307	3q21.3
F8	Coagulation factor VIII, procoagulant component (hemophilia A)	BM722299	NM_000132	Hs.6917	Xq28
H2AFJ	H2A histone family, member J	NM_018267	AL133626	Hs.524280	12p12
H2AFX	H2A histone family, member X	NM_002105	BM917453	Hs.477879	11q23.2-q23.3
H2AFY	H2A histone family, member Y	NM_004893	AB209490	Hs.586218	5q31.3-q32
H2AFY2	H2A histone family, member Y2	NM_018649	AL359572	Hs.499953	10q22
H2AFZ	H2A histone family, member Z	NM_002106	AK056803	Hs.119192	4q24
H2AFV	H2A histone family, member V	NM_138635	AL110212	Hs.488189	7p13
H3F3B	H3 histone, family 3B (H3.3B)	NM_005324	BX537379	Hs.180877	17q25
HAT1	Histone acetyltransferase 1	NM_003642	AK127840	Hs.298250	2q31.2-q33.1
HDAC1	Histone deacetylase 1	NM_004964	BX648055	Hs.88556	1p34
HDAC2	Histone deacetylase 2	NM_001527	AB209190	Hs.3352	6q21
HDAC3	Histone deacetylase 3	NM_003883	AF005482	Hs.519632	5q31
HDAC4	Histone deacetylase 4	NM_006037	NM_006037	Hs.20516	2q37.2
HDAC4	Histone deacetylase 4	NM_006037	NM_006037	Hs.20516	2q37.2
HDAC5	Histone deacetylase 5	NM_139205	NM_001015053	Hs.438782	17q21
HDAC6	Histone deacetylase 6	NM_006044	BC069243	Hs.6764	Xp11.23
HDAC7A	Histone deacetylase 7A	NM_015401	AK122588	Hs.200063	12q13.1
HDAC8	Histone deacetylase 8	NM_018486	AK074326	Hs.310536	Xq13
HDAC9	Histone deacetylase 9	NM_178423	AJ459808	Hs.196054	7p21.1
KNTC2	Kinetochore associated 2	NM_006101	NM_006101	Hs.414407	18p11.32
HGF	Hepatocyte growth factor (hepapoietin A; scatter factor)	NM_000601	NM_000601	Hs.396530	7q21.1
HIRA	HIR histone cell cycle regulation defective homolog A (S. cerevisiae)	NM_003325	BC039835	Hs.474206	22q11.2|22q11.21
HIST1H1B	Histone 1, H1b	NM_005322	BG292093	Hs.131956	6p22-p21.3
HIST1H1C	Histone 1, H1c	NM_005319	BQ940876	Hs.7644	6p21.3
HIST1H1D	Histone 1, H1d	NM_005320	NM_005320	Hs.136857	6p21.3
HIST1H1E	Histone 1, H1e	NM_005321	BU603483	Hs.248133	6p21.3
HIST1H1E	Histone 1, H1e	NM_005321	BU603483	Hs.248133	6p21.3
HIST1H2AC	Histone 1, H2ac	NM_003512	CR608156	Hs.484950	6p21.3
HMG20A	High-mobility group 20A	NM_018200	NM_018200	Hs.69594	15q24
HMG2L1	High-mobility group protein 2-like 1	NM_005487	NM_005487	Hs.588815	22q13.1
HMGA1	High mobility group AT-hook 1	NM_145904	BC078664	Hs.518805	6p21
HMGA2	High mobility group AT-hook 2	NM_003483	AB209853	Hs.505924	12q15
HMGB1	High-mobility group box 1	NM_002128	AK122825	Hs.434102	13q12

HMGB2	High-mobility group box 2	NM_002129	CR600021	Hs.434953	4q31
HMGB3	High-mobility group box 3	NM_005342	NM_005342	Hs.19114	Xq28
HMGN1	High-mobility group nucleosome binding domain 1	NM_004965	AB209245	Hs.356285	21q22.3|21q22.2
HMGN2	High-mobility group nucleosomal binding domain 2	NM_005517	BG034748	Hs.181163	1p36.1
HMGN3	High mobility group nucleosomal binding domain 3	NM_004242	BX648085	Hs.77558	6q14.1
HMGN4	High mobility group nucleosomal binding domain 4	NM_006353	NM_006353	Hs.236774	6p21.3
HOXA10	Homeobox A10	NM_018951	NM_018951	Hs.592166	7p15-p14
HOXA11	Homeobox A11	NM_005523	NM_005523	Hs.249171	7p15-p14
HOXB7	Homeobox B7	NM_004502	AK223249	Hs.436181	17q21.3
HRASLS	HRAS-like suppressor	NM_020386	BC005856	Hs.36761	3q29
PRMT2	Protein arginine methyltransferase 2	AL109794	AK123352	Hs.154163	21q22.3
STK32B	Serine/threonine kinase 32B	NM_018401	AY358353	Hs.133062	4p16.2-p16.1
HSD11B1	Hydroxysteroid (11-beta) dehydrogenase 1	NM_005525	BF698867	Hs.195040	1q32-q41
HSD11B2	Hydroxysteroid (11-beta) dehydrogenase 2	NM_000196	AF370400	Hs.1376	16q22
HSD17B1	Hydroxysteroid (17-beta) dehydrogenase 1	NM_000413	AK127832	Hs.50727	17q11-q21
HSD17B12	Hydroxysteroid (17-beta) dehydrogenase 12	NM_016142	BX537496	Hs.132513	11p11.2
HSD17B2	Hydroxysteroid (17-beta) dehydrogenase 2	NM_002153	BG261436	Hs.162795	16q24.1-q24.2
HSD17B3	Hydroxysteroid (17-beta) dehydrogenase 3	NM_000197	BC034281	Hs.477	9q22
HSD17B4	Hydroxysteroid (17-beta) dehydrogenase 4	NM_000414	AB208932	Hs.406861	5q21
HSD17B7	Hydroxysteroid (17-beta) dehydrogenase 7	NM_016371	AK022929	Hs.492925	1q23
RING1	Ring finger protein 1	NM_014234	NM_006979	Hs.415058	6p21.3
HSF1	Heat shock transcription factor 1	NM_005526	AK125467	Hs.530227	8q24.3
HSPA4	Heat shock 70kDa protein 4	XM_114482	NM_002154	Hs.90093	5q31.1-q31.2
HSPA5	Heat shock 70kDa protein 5 (glucose-regulated protein, 78kDa)	NM_005347	NM_005347	Hs.522392	9q33-q34.1
HSPA9B	Heat shock 70kDa protein 9B (mortalin-2)	NM_004134	NM_004134	Hs.184233	5q31.1
HSPB1	Heat shock 27kDa protein 1	NM_001540	BM907768	Hs.520973	7q11.23
HSPB2	Heat shock 27kDa protein 2	NM_001541	AB096250	Hs.78846	11q22-q23
HSP90AA1	Heat shock protein 90kDa alpha (cytosolic), class A member 1	NM_005348	AJ890082	Hs.525600	14q32.33
HSP90AB1	Heat shock protein 90kDa alpha (cytosolic), class B member 1	NM_007355	AY359878	Hs.509736	6p12
HUS1	HUS1 checkpoint homolog (S. pombe)	NM_004507	CR619988	Hs.152983	7p13-p12
ID4	Inhibitor of DNA binding 4, dominant negative helix-loop-helix protein	NM_001546	NM_001546	Hs.519601	6p22-p21
IFIT2	Interferon-induced protein with tetratricopeptide repeats 2	NM_001547	BC032839	Hs.437609	10q23-q25
IFITM1	Interferon induced transmembrane protein 1 (9-27)	NM_003641	BF210063	Hs.458414	11p15.5
IGF1	Insulin-like growth factor 1 (somatomedin C)	NM_000618	NM_000618	Hs.160562	12q22-q23
IGF1R	Insulin-like growth factor 1 receptor	NM_000875	BX640783	Hs.592020	15q26.3
IGFBP1	Insulin-like growth factor binding protein 1	NM_000596	NM_000596	Hs.401316	7p13-p12
IGFBP2	Insulin-like growth factor binding protein 2, 36kDa	NM_000597	AB209509	Hs.438102	2q33-q34
IGFBP2	Insulin-like growth factor binding protein 2, 36kDa	NM_000597	AB209509	Hs.438102	2q33-q34
IGFBP3	Insulin-like growth factor binding protein 3	NM_000598	NM_001013398	Hs.450230	7p13-p12
IGFBP4	Insulin-like growth factor binding protein 4	NM_001552	NM_001552	Hs.462998	17q12-q21.1
IGFBP5	Insulin-like growth factor binding protein 5	NM_000599	NM_000599	Hs.369982	2q33-q36
IL18	Interleukin 18 (interferon-gamma-inducing factor)	NM_001562	NM_001562	Hs.83077	11q22.2-q22.3
IL2RA	Interleukin 2 receptor, alpha	NM_000417	X01057	Hs.231367	10p15-p14
IL6	Interleukin 6 (interferon, beta 2)	NM_000600	BM906445	Hs.512234	7p21
IL6R	Interleukin 6 receptor	NM_000565	NM_000565	Hs.591492	1q21
IL6ST	Interleukin 6 signal transducer (gp130, oncostatin M receptor)	NM_002184	BC071555	Hs.532082	5q11
ING1	Inhibitor of growth family, member 1	NM_005537	NM_198219	Hs.46700	13q34
IRF1	Interferon regulatory factor 1	NM_002198	AB209624	Hs.436061	5q31.1
IRF1	Interferon regulatory factor 1	NM_002198	AB209624	Hs.436061	5q31.1
IRF5	Interferon regulatory factor 5	NM_002200	NM_002200	Hs.521181	7q32
IRF7	Interferon regulatory factor 7	NM_004031	AF076494	Hs.166120	11p15.5
ITGA2	Integrin, alpha 2 (CD49B, alpha 2 subunit of VLA-2 receptor)	NM_002203	X17033	Hs.591770	5q23-q31
ITGA3	Integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)	NM_002204	AB209658	Hs.265829	17q21.33
ITGA5	Integrin, alpha 5 (fibronectin receptor, alpha polypeptide)	NM_002205	NM_002205	Hs.505654	12q11-q13
ITGA6	Integrin, alpha 6	NM_000210	X53586	Hs.133397	2q31.1
ITGAV	Integrin, alpha V (vitronectin receptor, alpha polypeptide, antigen CD51)	NM_002210	NM_002210	Hs.436873	2q31-q32
ITGB1	Integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 include	NM_002211	NM_002211	Hs.558072	10p11.2
ITGB2	Integrin, beta 2 (complement component 3 receptor 3 and 4 subunit)	NM_000211	NM_000211	Hs.375957	21q22.3
ITGB3	Integrin, beta 3 (platelet glycoprotein IIIa, antigen CD61)	NM_000212	NM_000212	Hs.218040	17q21.32
ITGB3BP	Integrin beta 3 binding protein (beta3-endonexin)	NM_014288	CR596268	Hs.166539	1p31.3
ITGB4	Integrin, beta 4	NM_000213	X53587	Hs.592106	17q25
ITGB5	Integrin, beta 5	NM_002213	X53002	Hs.536663	3q21.2
ITGB8	Integrin, beta 8	NM_002214	AB209429	Hs.592171	7p15.3
GPR180	G protein-coupled receptor 180	NM_180989	NM_180989	Hs.439363	13q32.1
JUN	V-jun sarcoma virus 17 oncogene homolog (avian)	NM_002228	NM_002228	Hs.525704	1p32-p31
CD82	CD82 molecule	NM_002231	NM_002231	Hs.527778	11p11.2
UBR2	Ubiquitin protein ligase E3 component n-recognin 2	NM_015255	BX647467	Hs.529925	6p21.1
PALLD	Palladin, cytoskeletal associated protein	NM_016081	NM_016081	Hs.151220	4q32.3
PALLD	Palladin, cytoskeletal associated protein	NM_016081	NM_016081	Hs.151220	4q32.3
KIAA1093	KIAA1093	XM_039385	XM_039385		
KIAA1357	KIAA1357	XM_050421	XM_050421		
RP5-860F19.3	KIAA1442 protein	XM_044921	XM_044921	Hs.471955	20p13
TSPYL5	TSPY-like 5	NM_033512	NM_033512	Hs.173094	8q22.1
KISS1	KiSS-1 metastasis-suppressor	NM_002256	BM807845	Hs.95008	1q32
KISS1	KiSS-1 metastasis-suppressor	NM_002256	BM807845	Hs.95008	1q32
KIT	V-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog	NM_000222	BC071593	Hs.479754	4q11-q12
KLF5	Kruppel-like factor 5 (intestinal)	NM_001730	AF132818	Hs.508234	13q22.1
KLK5	Kallikrein 5	NM_012427	AY359010	Hs.50915	19q13.3-q13.4
KPNA3	Karyopherin alpha 3 (importin alpha 4)	NM_002267	BC035090	Hs.527919	13q14.3
KPNA4	Karyopherin alpha 4 (importin alpha 3)	NM_002268	NM_002268	Hs.288193	3q25.33
KPNA5	Karyopherin alpha 5 (importin alpha 6)	NM_002269	NM_002269	Hs.182971	6q22.2
KPNA6	Karyopherin alpha 6 (importin alpha 7)	NM_012316	NM_012316	Hs.591500	1p35.1-p34.3
KPNB1	Karyopherin (importin) beta 1	NM_002265	L38951	Hs.532793	17q21.32
TNPO1	Transportin 1	NM_002270	U70322	Hs.482497	5q13.2
RANBP5	RAN binding protein 5	NM_002271	NM_002271	Hs.588179	13q32.2
KRT17	Keratin 17	NM_000526	AK122864	Hs.2785	17q12-q21
KRT17	Keratin 17	NM_000422	AK122864	Hs.2785	17q12-q21
KRT18	Keratin 18	NM_000224	CR616919	Hs.406013	12q13
KRT17	Keratin 17	NM_002276	AK122864	Hs.2785	17q12-q21
KRT5	Keratin 5 (epidermolysis bullosa simplex, Dowling-Meara/Kobner/Weber-Co	NM_000424	AJ508777	Hs.433845	12q12-q13
KRT5	Keratin 5 (epidermolysis bullosa simplex, Dowling-Meara/Kobner/Weber-Co	NM_005554	AJ508777	Hs.433845	12q12-q13

KRT5	Keratin 5 (epidermolysis bullosa simplex, Dowling-Meara/Kobner/Weber-Co	NM_005555	AJ508777	Hs.433845	12q12-q13
LAMA3	Laminin, alpha 3	NM_000227	NM_198129	Hs.436367	18q11.2
LCMT2	Leucine carboxyl methyltransferase 2	NM_014793	NM_014793	Hs.200596	15q15.3
TSEN34	TRNA splicing endonuclease 34 homolog (S. cerevisiae)	NM_024075	AK054944	Hs.15580	19q13.4
LGP2	Likely ortholog of mouse D11lgp2	NM_024119	AK021416	Hs.55918	17q21.2
LIF	Leukemia inhibitory factor (cholinergic differentiation factor)	NM_002309	NM_002309	Hs.2250	22q12.2
PLA2G4C	Phospholipase A2, group IVC (cytosolic, calcium-independent)	NM_000234	AB208791	Hs.1770	19q13.3
LIG3	Ligase III, DNA, ATP-dependent	NM_013975	NM_013975	Hs.100299	17q11.2-q12
LIG4	Ligase IV, DNA, ATP-dependent	NM_002312	NM_002312	Hs.166091	13q33-q34
LMNA	Lamin A/C	NM_170707	NM_170707	Hs.491359	1q21.2-q21.3
LMNB1	Lamin B1	NM_005573	BC052951	Hs.89497	5q23.3-q31.1
LMNB2	Lamin B2	NM_032737	NM_032737	Hs.538286	19p13.3
LOC153770	Transcribed locus	CA430603	CA430603	Hs.623375	
LOC221143	Hypothetical protein LOC221143	NM_174928	BG403594	Hs.26674	13q12.11
LOC286052	Hypothetical protein LOC286052	AK095104	AK095104	Hs.588365	8q24.13
LOC286478	Transcribed locus	BX089115	BX089115	Hs.449499	
LOH11CR1J	CDNA: FLJ21561 fis, clone COL06415	AB096249	AK025214	Hs.96918	
LPXN	Leupaxin	NM_004811	BC034230	Hs.125474	11q12.1
MTDH	Metadherin	NM_178812	BC045642	Hs.377155	8q22.1
MAD1L1	MAD1 mitotic arrest deficient-like 1 (yeast)	NM_003550	NM_003550	Hs.209128	7p22
MAD2L1	MAD2 mitotic arrest deficient-like 1 (yeast)	NM_002358	U65410	Hs.591697	4q27
MAD2L2	MAD2 mitotic arrest deficient-like 2 (yeast)	NM_006341	AK094316	Hs.19400	1p36
MAP2K1	Mitogen-activated protein kinase kinase 1	NM_002755	NM_002755	Hs.145442	15q22.1-q22.33
MAP2K2	Mitogen-activated protein kinase kinase 2	NM_030662	BM809871	Hs.465627	19p13.3
MAP2K3	Mitogen-activated protein kinase kinase 3	NM_145110	AK093838	Hs.514012	17q11.2
MAP2K4	Mitogen-activated protein kinase kinase 4	NM_003010	AK131544	Hs.514681	17p11.2
MAP2K5	Mitogen-activated protein kinase kinase 5	NM_145160	AK025177	Hs.114198	15q23
MAP2K6	Mitogen-activated protein kinase kinase 6	NM_002758	BX641121	Hs.463978	17q24.3
MAP2K7	Mitogen-activated protein kinase kinase 7	Hs.531754	19p13.3-p13.2	Mitogen-activated protein kinase kin	
MAP3K1	Mitogen-activated protein kinase kinase kinase 1	AF042838	AF042838	Hs.508461	5q11.2
MAP3K10	Mitogen-activated protein kinase kinase kinase 10	NM_002446	X90846	Hs.466743	19q13.2
MAP3K2	Mitogen-activated protein kinase kinase kinase 2	NM_006609	NM_006609	Hs.145605	2q14.3
MAP3K3	Mitogen-activated protein kinase kinase kinase 3	NM_002401	NM_203351	Hs.29282	17q23.3
MAP3K4	Mitogen-activated protein kinase kinase kinase 4	NM_005922	NM_005922	Hs.390428	6q26
MAP3K5	Mitogen-activated protein kinase kinase kinase 5	NM_005923	U67156	Hs.186486	6q22.33
MAP3K6	Mitogen-activated protein kinase kinase kinase 6	NM_004672	NM_004672	Hs.194694	1p36.11
MAP3K7IP1	Mitogen-activated protein kinase kinase kinase 7 interacting protein 1	NM_006116	AB209372	Hs.507681	22q13.1
MAP4K1	Mitogen-activated protein kinase kinase kinase kinase 1	NM_007181	NM_007181	Hs.95424	19q13.1-q13.4
MAP4K2	Mitogen-activated protein kinase kinase kinase kinase 2	NM_004579	NM_004579	Hs.534341	11q13
MAP4K3	Mitogen-activated protein kinase kinase kinase kinase 3	NM_003618	BC071579	Hs.468239	2p22.1
MAP4K4	Mitogen-activated protein kinase kinase kinase kinase 4	NM_145686	NM_145686	Hs.431550	2q11.2-q12
MAP4K5	Mitogen-activated protein kinase kinase kinase kinase 5	NM_006575	NM_198794	Hs.130491	14q11.2-q21
MAPK1	Mitogen-activated protein kinase 1	NM_002745	AL157438	Hs.431850	22q11.2|22q11.21
MAPK1	Mitogen-activated protein kinase 1	NM_002745	AL157438	Hs.431850	22q11.2|22q11.21
MAPK10	Mitogen-activated protein kinase 10	NM_002753	AK124791	Hs.125503	4q22.1-q23
MAPK11	Mitogen-activated protein kinase 11	NM_002751	BC027933	Hs.57732	22q13.33
MAPK12	Mitogen-activated protein kinase 12	NM_002969	CR620424	Hs.432642	22q13.33
MAPK13	Mitogen-activated protein kinase 13	NM_002754	AB209586	Hs.178695	6p21.31
MAPK14	Mitogen-activated protein kinase 14	NM_139012	NM_001315	Hs.588289	6p21.3-p21.2
MAPK3	Mitogen-activated protein kinase 3	AK091009	BX537897	Hs.861	16p12-p11.2
MAPK4	Mitogen-activated protein kinase 4	NM_002747	BC050299	Hs.433728	18q12-q21
MAPK6	Mitogen-activated protein kinase 6	NM_002748	NM_002748	Hs.411847	15q21
MAPK7	Mitogen-activated protein kinase 7	NM_139033	AB209611	Hs.150136	17p11.2
MAPK8	Mitogen-activated protein kinase 8	NM_139046	CR614448	Hs.138211	10q11.22
MBD1	Methyl-CpG binding domain protein 1	NM_015846	NM_015846	Hs.405610	18q21
MBD3	Methyl-CpG binding domain protein 3	NM_003926	NM_003926	Hs.178728	19p13.3
ENSA	Endosulfine alpha	NM_021960	NM_021960	Hs.163936	1q21.2
MCM5	MCM5 minichromosome maintenance deficient 5, cell division cycle 46 (S. c	NM_006739	AB209612	Hs.517582	22q13.1
MCM6	MCM6 minichromosome maintenance deficient 6 (MIS5 homolog, S. pombe	NM_005915	NM_005915	Hs.444118	2q21
MDH2	Malate dehydrogenase 2, NAD (mitochondrial)	NM_005918	AK095803	Hs.520967	7p12.3-q11.2
MDM2	Mdm2, transformed 3T3 cell double minute 2, p53 binding protein (mouse)	NM_002392	M92424	Hs.567303	12q14.3-q15
MED6	Mediator of RNA polymerase II transcription, subunit 6 homolog (yeast)	NM_005466	AK222587	Hs.497353	14q24.2
MELK	Maternal embryonic leucine zipper kinase	NM_014791	NM_014791	Hs.184339	9p13.2
MET	Met proto-oncogene (hepatocyte growth factor receptor)	NM_000245	NM_000245	Hs.132966	7q31
FBXO31	F-box protein 31	NM_024735	AF318348	Hs.567582	16q24.2
C9orf30	Chromosome 9 open reading frame 30	NM_080655	AK092292	Hs.530272	9q31.1
CYBASC3	Cytochrome b, ascorbate dependent 3	NM_153611	BC004391	Hs.22546	11q12.2
MGC26744	Hypothetical protein MGC26744	NM_144645	AK057662	Hs.339646	4q21.3
MGMT	O-6-methylguanine-DNA methyltransferase	NM_002412	CR618411	Hs.501522	10q26
MGST1	Microsomal glutathione S-transferase 1	NM_020300	AK058030	Hs.389700	12p12.3-p12.1
MKI67	Antigen identified by monoclonal antibody Ki-67	NM_002417	NM_002417	Hs.80976	10q25-qter
MLH1	MutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)	NM_000249	BX648844	Hs.195364	3p21.3
MMP11	Matrix metallopeptidase 11 (stromelysin 3)	NM_005940	NM_005940	Hs.143751	22q11.2|22q11.23
MMP3	Matrix metallopeptidase 3 (stromelysin 1, progelatinase)	NM_002422	AK223291	Hs.375129	11q22.3
MMP9	Matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV c	NM_004994	NM_004994	Hs.297413	20q11.2-q13.1
MRE11A	MRE11 meiotic recombination 11 homolog A (S. cerevisiae)	NM_005591	NM_005590	Hs.192649	11q21
MS4A7	Membrane-spanning 4-domains, subfamily A, member 7	NM_021201	NM_021201	Hs.530735	11q12
MSH2	MutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)	NM_000251	AK223284	Hs.156519	2p22-p21
MSH6	MutS homolog 6 (E. coli)	NM_000179	BC071594	Hs.445052	2p16
MSX1	Msh homeobox homolog 1 (Drosophila)	NM_002448	NM_002448	Hs.424414	4p16.3-p16.1
MTA1	Metastasis associated 1	NM_004689	NM_004689	Hs.525629	14q32.3
MTA2	Metastasis associated 1 family, member 2	NM_004739	NM_004739	Hs.173043	11q12-q13.1
MTA3	Metastasis associated 1 family, member 3	AB033092	AB033092	Hs.435413	2p21
MUC1	Mucin 1, cell surface associated	NM_002456	X52228	Hs.89603	1q21
GRHL2	Grainyhead-like 2 (Drosophila)	BG675392	AK023844	Hs.561796	8q22.3
MX1	Myxovirus (influenza virus) resistance 1, interferon-inducible protein p78 (m	NM_002462	AK096355	Hs.517307	21q22.3
MXI1	MAX interactor 1	NM_005962	NM_130439	Hs.501023	10q24-q25
MYB	V-myb myeloblastosis viral oncogene homolog (avian)	NM_005375	AJ606319	Hs.591337	6q22-q23
MYC	V-myc myelocytomatosis viral oncogene homolog (avian)	NM_002467	NM_002467	Hs.202453	8q24.12-q24.13
NBN	Nibrin	NM_002485	BX640816	Hs.492208	8q21
NCOA1	Nuclear receptor coactivator 1	NM_147223	NM_147223	Hs.412293	2p23

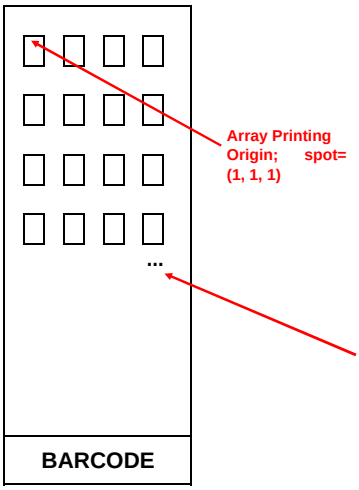
NCOA2	Nuclear receptor coactivator 2	NM_006540	AL832112	Hs.446678	8q13.3
NCOA3	Nuclear receptor coactivator 3	NM_006534	NM_181659	Hs.592142	20q12
NCOA4	Nuclear receptor coactivator 4	NM_005437	AL162047	Hs.591356	10q11.2
NCOA5	Nuclear receptor coactivator 5	NM_020967	NM_020967	Hs.25669	20q12-q13.12
NCOA6	Nuclear receptor coactivator 6	NM_014071	AF208227	Hs.368971	20q11
NCOA6IP	Nuclear receptor coactivator 6 interacting protein	NM_024831	NM_024831	Hs.335068	8q11
NCOA7	Nuclear receptor coactivator 7	AL834442	AL834442	Hs.171426	6q22.31-q22.32
NCOR1	Nuclear receptor co-repressor 1	NM_006311	AF087856	Hs.462323	17p11.2
NCOR2	Nuclear receptor co-repressor 2	NM_006312	NM_006312	Hs.137510	12q24
NDUFS8	NADH dehydrogenase (ubiquinone) Fe-S protein 8, 23kDa (NADH-coenzym	NM_002496	AK002110	Hs.90443	11q13
NEO1	Neogenin homolog 1 (chicken)	NM_002499	AB209412	Hs.536488	15q22.3-q23
NGRN	Neugrin, neurite outgrowth associated	NM_016645	NM_016645	Hs.513145	15q26.1
NFIB	Nuclear factor I/B	NM_005596	BX537698	Hs.370359	9p24.1
NF2	Neurofibromin 2 (bilateral acoustic neuroma)	NM_000268	NM_181826	Hs.187898	22q12.2
NFYB	Nuclear transcription factor Y, beta	NM_006166	NM_006166	Hs.84928	12q22-q23
RTN4RL1	Reticulon 4 receptor-like 1	NM_178568	NM_178568	Hs.22917	17p13.3
IL32	Interleukin 32	NM_004221	BF569086	Hs.943	16p13.3
NME1	Non-metastatic cells 1, protein (NM23A) expressed in	NM_000269	BG114681	Hs.463456	17q21.3
NME1	Non-metastatic cells 1, protein (NM23A) expressed in	NM_002512	BG114681	Hs.463456	17q21.3
NMU	Neuromedin U	NM_006681	BE735948	Hs.418367	4q12
NOS3	Nitric oxide synthase 3 (endothelial cell)	NM_000603	NM_000603	Hs.585117	7q36
NOTCH1	Notch homolog 1, translocation-associated (Drosophila)	NM_017617	NM_017617	Hs.495473	9q34.3
NOTCH3	Notch homolog 3 (Drosophila)	NM_000435	NM_000435	Hs.8546	19p13.2-p13.1
NOTCH4	Notch homolog 4 (Drosophila)	NM_004557	NM_004557	Hs.436100	6p21.3
NR1I2	Nuclear receptor subfamily 1, group I, member 2	NM_003889	AJ009936	Hs.7303	3q12-q13.3
NRG1	Neuregulin 1	NM_013957	NM_013957	Hs.453951	8p21-p12
NRIP1	Nuclear receptor interacting protein 1	NM_003489	NM_003489	Hs.155017	21q11.2
NSD1	Nuclear receptor binding SET domain protein 1	NM_022455	NM_022455	Hs.106861	5q35.2-q35.3
OAS1	2',5'-oligoadenylate synthetase 1, 40/46kDa	NM_016816	NM_016816	Hs.524760	12q24.1
OAS2	2'-5'-oligoadenylate synthetase 2, 69/71kDa	NM_002535	NM_002535	Hs.414332	12q24.2
OGG1	8-oxoguanine DNA glycosylase	NM_016819	NM_016819	Hs.380271	3p26.2
ORC6L	Origin recognition complex, subunit 6 homolog-like (yeast)	NM_014321	NM_014321	Hs.49760	16q12
OSM	Oncostatin M	NM_020530	NM_020530	Hs.248156	22q12.2
OXCT1	3-oxoacid CoA transferase 1	NM_000436	NM_000436	Hs.278277	5p13.1
P53AIP1	P53-regulated apoptosis-inducing protein 1	NM_022112	AB045832	Hs.160953	11q24
PAK2	P21 (CDKN1A)-activated kinase 2	BC063539	NM_002577	Hs.518530	3q29
PAK2	P21 (CDKN1A)-activated kinase 2	NM_002577	NM_002577	Hs.518530	3q29
PAPPA	Pregnancy-associated plasma protein A, pappalysin 1	NM_002581	NM_002581	Hs.494928	9q33.2
PARVA	Parvin, alpha	NM_018222	AL832682	Hs.436319	11p15.3
PAX2	Paired box gene 2	NM_003988	NM_003988	Hs.155644	10q24
SUB1	SUB1 homolog (S. cerevisiae)	NM_006713	BX537584	Hs.229641	5p13.3
PCAF	P300/CBP-associated factor	NM_003884	NM_003884	Hs.533055	3p24
PCDH17	Protocadherin 17	NM_014459	NM_014459	Hs.106511	13q21.1
PCNA	Proliferating cell nuclear antigen	NM_002592	BM462208	Hs.147433	20pter-p12
PDGFRA	Platelet-derived growth factor receptor, alpha polypeptide	NM_006206	NM_006206	Hs.74615	4q11-q13
PECI	Peroxisomal D3,D2-enoyl-CoA isomerase	NM_006117	AB209917	Hs.15250	6p24.3
PELP1	Proline, glutamic acid and leucine rich protein 1	NM_014389	BC069058	Hs.513883	17p13.2
PPARGC1B	Peroxisome proliferative activated receptor, gamma, coactivator 1, beta	NM_133263	AY188950	Hs.591261	5q33.1
PGR	Progesterone receptor	NM_000926	NM_000926	Hs.368072	11q22-q23
PGR	Progesterone receptor	NM_000926	NM_000926	Hs.368072	11q22-q23
PGR	Progesterone receptor	NM_000926	NM_000926	Hs.368072	11q22-q23
PHB	Prohibitin	NM_002634	BF676086	Hs.514303	17q21
PIK3C2A	Phosphoinositide-3-kinase, class 2, alpha polypeptide	NM_002645	NM_002645	Hs.175343	11p15.5-p14
PIK3C2B	Phosphoinositide-3-kinase, class 2, beta polypeptide	NM_002646	Y11312	Hs.497487	1q32
PIK3C2G	Phosphoinositide-3-kinase, class 2, gamma polypeptide	NM_004570	AJ000008	Hs.22500	12p12
PIK3CA	Phosphoinositide-3-kinase, catalytic, alpha polypeptide	NM_006218	BX640788	Hs.478376	3q26.3
PIK3CB	Phosphoinositide-3-kinase, catalytic, beta polypeptide	NM_006219	CR749357	Hs.239818	3q22.3
PIK3CD	Phosphoinositide-3-kinase, catalytic, delta polypeptide	NM_005026	NM_005026	Hs.518451	1p36.2
PIK3CG	Phosphoinositide-3-kinase, catalytic, gamma polypeptide	NM_002649	X83368	Hs.32942	7q22.3
PIK3R1	Phosphoinositide-3-kinase, regulatory subunit 1 (p85 alpha)	NM_181504	NM_181523	Hs.132225	5q13.1
PIK3R2	Phosphoinositide-3-kinase, regulatory subunit 2 (p85 beta)	NM_005027	NM_005027	Hs.371344	19q13.2-q13.4
PIN1	Protein (peptidylprolyl cis/trans isomerase) NIMA-interacting 1	NM_006221	AK092970	Hs.465849	19p13
PITRM1	Pitrilysin metallopeptidase 1	NM_014889	CR749279	Hs.528300	10p15.2
PLAU	Plasminogen activator, urokinase	NM_002658	NM_002658	Hs.77274	10q24
PLAUR	Plasminogen activator, urokinase receptor	NM_002659	CR601067	Hs.466871	19q13
PMS1	PMS1 postmeiotic segregation increased 1 (S. cerevisiae)	NM_000534	CR749432	Hs.111749	2q31-q33|2q31.1
PMS2	PMS2 postmeiotic segregation increased 2 (S. cerevisiae)	NM_000535	AB037790	Hs.520205	7p22.2
POLD4	Polymerase (DNA-directed), delta 4	NM_021173	AB209274	Hs.523829	11q13
POLE3	Polymerase (DNA directed), epsilon 3 (p17 subunit)	NM_017443	AK092840	Hs.108112	9q33
POLR2A	Polymerase (RNA) II (DNA directed) polypeptide A, 220kDa	NM_000937	NM_000937	Hs.270017	17p13.1
REST	RE1-silencing transcription factor	NM_000938	BC023503	Hs.307836	4q12
POLR2C	Polymerase (RNA) II (DNA directed) polypeptide C, 33kDa	NM_032940	NM_032940	Hs.79402	16q13-q21
WDR33	WD repeat domain 33	NM_004805	NM_018383	Hs.554831	2q14.3
POLR2E	Polymerase (RNA) II (DNA directed) polypeptide E, 25kDa	NM_002695	AK122813	Hs.24301	19p13.3
POLR2F	Polymerase (RNA) II (DNA directed) polypeptide F	NM_021974	AL832562	Hs.436578	22q13.1
POLR2H	Polymerase (RNA) II (DNA directed) polypeptide H	NM_006232	CR590527	Hs.432574	3q28
POLR2I	Polymerase (RNA) II (DNA directed) polypeptide I, 14.5kDa	NM_006233	BU598062	Hs.47062	19q12
POLR2J2	DNA directed RNA polymerase II polypeptide J-related gene	NM_006234	NM_006989	Hs.530089	7q11.22
POLR2J2	DNA directed RNA polymerase II polypeptide J-related gene	BC050405	NM_006989	Hs.530089	7q11.22
POLR2K	Polymerase (RNA) II (DNA directed) polypeptide K, 7.0kDa	NM_005034	BI758413	Hs.351475	8q22.2
POLR2L	Polymerase (RNA) II (DNA directed) polypeptide L, 7.6kDa	NM_021128	BM919305	Hs.441072	11p15
POU2F2	POU domain, class 2, transcription factor 2	NM_002698	M36542	Hs.118990	19q13.2
PPARBP	PPAR binding protein	NM_004774	Y13467	Hs.462956	17q12-q21.1
PPARG	Peroxisome proliferative activated receptor, gamma	NM_015869	NM_138711	Hs.162646	3p25
PPIA	Peptidylprolyl isomerase A (cyclophilin A)	NM_021130	AK130101	Hs.356331	7p13-p11.2
PPM1D	Protein phosphatase 1D magnesium-dependent, delta isoform	NM_003620	NM_003620	Hs.591184	17q23.2
PPP1R12A	Protein phosphatase 1, regulatory (inhibitor) subunit 12A	NM_002480	AF458589	Hs.49582	12q15-q21
BAX	BCL2-associated X protein	NM_014330	AK001361	Hs.433670	19q13.3-q13.4
BAX	BCL2-associated X protein	NM_014330	AK001361	Hs.433670	19q13.3-q13.4
PRC1	Protein regulator of cytokinesis 1	NM_003981	NM_003981	Hs.567385	15q26.1
PRL	Prolactin	NM_000948	CD512992	Hs.1905	6p22.2-p21.3

PSMC5	Proteasome (prosome, macropain) 26S subunit, ATPase, 5	NM_002805	CR595677	Hs.79387	17q23-q25
PTCH	Patched homolog (Drosophila)	NM_000264	AB209495	Hs.494538	9q22.3
PTCH2	Patched homolog 2 (Drosophila)	NM_003738	AY359016	Hs.591497	1p33-p34
PTEN	Phosphatase and tensin homolog (mutated in multiple advanced cancers 1)	NM_000314	NM_000314	Hs.500466	10q23.3
PTGS2	Prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cy	NM_000963	NM_000963	Hs.196384	1q25.2-q25.3
PUM1	Pumilio homolog 1 (Drosophila)	NM_014676	NM_001020658	Hs.281707	1p35.2
QSCN6	Quiescin Q6	NM_002826	NM_002826	Hs.518374	1q24
RAB6A	RAB6A, member RAS oncogene family	NM_016577	NM_016577	Hs.12152	11q13.3
RAB9A	RAB9A, member RAS oncogene family	NM_004251	BM926730	Hs.495704	Xp22.2
RAC2	Ras-related C3 botulinum toxin substrate 2 (rho family, small GTP binding p	NM_002872	BC001485	Hs.517601	22q13.1
RAC3	Ras-related C3 botulinum toxin substrate 3 (rho family, small GTP binding p	NM_005052	BM561442	Hs.45002	17q25.3
RAD1	RAD1 homolog (S. pombe)	NM_133377	NM_133377	Hs.531879	5p13.2
RAD17	RAD17 homolog (S. pombe)	NM_002873	AF076838	Hs.16184	5q13
RAD50	RAD50 homolog (S. cerevisiae)	NM_005732	U63139	Hs.128904	5q31
RAD51	RAD51 homolog (RecA homolog, E. coli) (S. cerevisiae)	NM_002875	NM_002875	Hs.511067	15q15.1
RAD51C	RAD51 homolog C (S. cerevisiae)	NM_058216	BC073161	Hs.412587	17q22-q23
RAD52	RAD52 homolog (S. cerevisiae)	NM_134423	NM_134423	Hs.410355	12p13-p12.2
RAD54B	RAD54 homolog B (S. cerevisiae)	NM_012415	NM_012415	Hs.30561	8q21.3-q22
RAD54L	RAD54-like (S. cerevisiae)	NM_003579	NM_003579	Hs.523220	1p32
RAD9A	RAD9 homolog A (S. pombe)	NM_004584	NM_004584	Hs.240457	11q13.1-q13.2
RAF1	V-raf-1 murine leukemia viral oncogene homolog 1	NM_002880	NM_002880	Hs.159130	3p25
DTL	Denticleless homolog (Drosophila)	NM_016448	NM_016448	Hs.126774	
RASA1	RAS p21 protein activator (GTPase activating protein) 1	NM_002890	CR749722	Hs.553501	5q13.3
RASD2	RASD family, member 2	NM_014310	BC013419	Hs.474711	22q13.1
RB1	Retinoblastoma 1 (including osteosarcoma)	NM_000321	L41870	Hs.408528	13q14.2
RBBP4	Retinoblastoma binding protein 4	NM_005610	AK056550	Hs.16003	1p35.1
RBBP4	Retinoblastoma binding protein 4	NM_005610	AK056550	Hs.16003	1p35.1
RBBP7	Retinoblastoma binding protein 7	NM_002893	AK127332	Hs.495755	Xp22.2
RBBP8	Retinoblastoma binding protein 8	NM_002894	NM_002894	Hs.546282	18q11.2
RECQL	RecQ protein-like (DNA helicase Q1-like)	NM_002907	L36140	Hs.235069	12p12
RELA	V-rel reticuloendotheliosis viral oncogene homolog A, nuclear factor of kapp	NM_021975	BC110830	Hs.502875	11q13
RELA	V-rel reticuloendotheliosis viral oncogene homolog A, nuclear factor of kapp	NM_021975	BC110830	Hs.502875	11q13
REM1	RAS (RAD and GEM)-like GTP-binding 1	BC039813	BC039813	Hs.247729	20q11.21
RPRM	Reprimo, TP53 dependent G2 arrest mediator candidate	NM_019845	AB043585	Hs.100890	2q23.3
RFC1	Replication factor C (activator 1) 1, 145kDa	NM_002913	NM_002913	Hs.507475	4p14-p13
RFC2	Replication factor C (activator 1) 2, 40kDa	NM_002914	NM_181471	Hs.139226	7q11.23
RFC3	Replication factor C (activator 1) 3, 38kDa	NM_002915	NM_002915	Hs.115474	13q12.3-q13
RFC4	Replication factor C (activator 1) 4, 37kDa	NM_002916	NM_002916	Hs.591322	3q27
RFC5	Replication factor C (activator 1) 5, 36.5kDa	NM_007370	NM_181578	Hs.506989	12q24.2-q24.3
RP11-1334A24/AC145	RP11-1334A24/AC145098.2	AC145098.2			
RP11-137L15 /AC023	RP11-137L15 /AC023991.9	AC023991.9			
RP11-264I13/AL35907	RP11-264I13/AL359076	AL359076			
RP11-567N19/AC0167	RP11-567N19/AC016772.9	AC016772.9			
RP11-62E9/012533	RP11-62E9/012533	AC012533			
RP11-72O9/AQ26706	RP11-72O9/AQ267068	AQ267068			
RP11-977G19/AC073	RP11-977G19/AC073896	AC073896			
RP11-99I9/AC099818.	RP11-99I9/AC099818.2	AC099818.2			
RPL7	Ribosomal protein L7	NM_000971	BQ057523	Hs.571841	8q21.11
RPS6KA1	Ribosomal protein S6 kinase, 90kDa, polypeptide 1	NM_002953	BC014966	Hs.149957	1p
RPS6KA4	Ribosomal protein S6 kinase, 90kDa, polypeptide 4	NM_003942	AK223561	Hs.105584	11q11-q13
RPS6KA5	Ribosomal protein S6 kinase, 90kDa, polypeptide 5	NM_004755	AB209667	Hs.510225	14q31-q32.1
RRAD	Ras-related associated with diabetes	BC057815	BC057815	Hs.1027	16q22
RUNX2	Runt-related transcription factor 2	NM_004348	NM_004348	Hs.535845	6p21
SCGB1C1	Secretoglobin, family 1C, member 1	NM_145651	BX098294	Hs.127059	11p15.5
S100A2	S100 calcium binding protein A2	NM_005978	BU589956	Hs.516484	1q21
S100A4	S100 calcium binding protein A4 (calcium protein, calvasculin, metastasin, n	NM_002961	CF619147	Hs.557609	1q21
S100A6	S100 calcium binding protein A6 (calcyclin)	NM_014624	BM904612	Hs.275243	1q21
SAP18	Sin3A-associated protein, 18kDa	NM_005870	AK126385	Hs.524899	13q12.11
SAP30	Sin3A-associated protein, 30kDa	NM_003864	BC016757	Hs.591715	4q34.1
SART3	Squamous cell carcinoma antigen recognised by T cells 3	NM_014706	CR933631	Hs.584842	12q24.1
SCD5	Stearoyl-CoA desaturase 5	NM_024906	AF389338	Hs.379191	4q21.22
SCGB1A1	Secretoglobin, family 1A, member 1 (uteroglobin)	NM_003357	BI819219	Hs.523732	11q12.3-q13.1
SCGB1D2	Secretoglobin, family 1D, member 2	NM_006551	BP314377	Hs.204096	11q13
SCGB2A1	Secretoglobin, family 2A, member 1	NM_002407	CB957406	Hs.97644	11q13
SCGB2A2	Secretoglobin, family 2A, member 2	NM_002411	BC067220	Hs.46452	11q13
SCN9A	Sodium channel, voltage-gated, type IX, alpha	NM_002977	NM_002977	Hs.2319	2q24
SCUBE2	Signal peptide, CUB domain, EGF-like 2	NM_020974	NM_020974	Hs.523468	11p15.3
SEH1L	SEH1-like (S. cerevisiae)	NM_031216	NM_031216	Hs.301048	18p11.21
SERTAD1	SERTA domain containing 1	NM_013376	AK074652	Hs.269898	19q13.1-q13.2
SEP15	15 kDa selenoprotein	NM_004261	NM_004261	Hs.362728	1p31
SERF1A	Small EDRK-rich factor 1A (telomeric)	NM_021967	AF073519	Hs.32567	5q12.2-q13.3
SERPINA3	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), mem	NM_001085	NM_001085	Hs.534293	14q32.1
SERPINB2	Serpin peptidase inhibitor, clade B (ovalbumin), member 2	NM_002575	BC012609	Hs.514913	18q21.3
SERPINB5	Serpin peptidase inhibitor, clade B (ovalbumin), member 5	NM_002639	BX640597	Hs.55279	18q21.3
SET	SET translocation (myeloid leukemia-associated)	NM_003011	NM_003011	Hs.436687	9q34
SETDB1	SET domain, bifurcated 1	NM_012432	NM_012432	Hs.591479	1q21
SPEN	Spen homolog, transcriptional regulator (Drosophila)	NM_015001	NM_015001	Hs.558463	1p36.33-p36.11
MPDU1	Mannose-P-dolichol utilization defect 1	NM_001040	NM_001678	Hs.78854	17p13.1-p12
SHMT2	Serine hydroxymethyltransferase 2 (mitochondrial)	NM_005412	AK055053	Hs.75069	12q12-q14
SIN3B	SIN3 homolog B, transcription regulator (yeast)	AB014600	NM_015260	Hs.13999	19p13.11
SKP1A	S-phase kinase-associated protein 1A (p19A)	NM_006930	NM_006930	Hs.171626	5q31
SLC25A5	Solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocat	NM_001152	AK092094	Hs.496658	Xq24-q26
SLC2A3	Solute carrier family 2 (facilitated glucose transporter), member 3	NM_006931	AB209607	Hs.419240	12p13.3
SMARCA1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin	NM_003069	NM_003069	Hs.152292	Xq25
SMARCA2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin	NM_003070	X72889	Hs.298990	9p22.3
SMARCA3	SWI/SNF related, matrix associated, actin dependent regulator of chromatin	NM_003071	BC044659	Hs.3068	3q25.1-q26.1
SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin	NM_003072	NM_003072	Hs.327527	19p13.2
SMARCA5	SWI/SNF related, matrix associated, actin dependent regulator of chromatin	NM_003601	NM_003601	Hs.589489	4q31.1-q31.2
SMARCB1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin	NM_003073	NM_003073	Hs.534350	22q11.23|22q11
SMARCC1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin	NM_003074	NM_003074	Hs.476179	3p23-p21
SMARCC2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin	NM_003075	AB209006	Hs.236030	12q13-q14

SMARCD1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin	NM_003076	NM_003076	Hs.79335	12q13-q14
SMARCD2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin	NM_003077	NM_003077	Hs.250581	17q23-q24
SMARCD3	SWI/SNF related, matrix associated, actin dependent regulator of chromatin	NM_003078	BX648385	Hs.444445	7q35-q36
SMARCE1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin	NM_003079	BC069196	Hs.547509	17q21.2
ARID1A	AT rich interactive domain 1A (SWI- like)	NM_006015	NM_006015	Hs.468972	1p35.3
SNAI1	Snail homolog 1 (Drosophila)	NM_005985	NM_005985	Hs.48029	20q13.1-q13.2
SNAI2	Snail homolog 2 (Drosophila)	NM_003068	NM_003068	Hs.360174	8q11
SNAI3	Snail homolog 3 (Drosophila)	AY203928	BX640980	Hs.499548	16q24.3
SNRPN	Small nuclear ribonucleoprotein polypeptide N	NM_005678	U81001	Hs.564847	15q11.2
SNW1	SNW domain containing 1	NM_012245	AF045184	Hs.546550	14q24.3
SOS1	Son of sevenless homolog 1 (Drosophila)	NM_005633	L13857	Hs.278733	2p22-p21
C14orf138	Chromosome 14 open reading frame 138	NM_024558	BX247997	Hs.558541	14q22.1
QSCN6L1	Quiescin Q6-like 1	NM_181701	NM_181701	Hs.144073	9q34.3
SP1	Sp1 transcription factor	XM_028606	BQ774060	Hs.620754	12q13.1
SPRR2C	Small proline-rich protein 2C	NM_006518	M21539	Hs.592363	1q21-q22
STAT3	Signal transducer and activator of transcription 3 (acute-phase response fac	NM_139276	NM_012448	Hs.463059	17q21.31
STAT5A	Signal transducer and activator of transcription 5A	NM_003152	NM_003152	Hs.437058	17q11.2
STAT3	Signal transducer and activator of transcription 3 (acute-phase response fac	NM_012448	NM_012448	Hs.463059	17q21.31
STK11	Serine/threonine kinase 11	NM_000455	AB209553	Hs.515005	19p13.3
STK11	Serine/threonine kinase 11	NM_000455	AB209553	Hs.515005	19p13.3
AURKA	Aurora kinase A	NM_198433	NM_198433	Hs.250822	20q13.2-q13.3
STRBP	Spermatid perinuclear RNA binding protein	NM_018387	NM_018387	Hs.287659	9q33.3
STXBP3	Syntaxin binding protein 3	NM_007269	NM_007269	Hs.530436	1p13.3
SULT1A1	Sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1	NM_177534	AB209149	Hs.567342	16p12.1
SURB7	SRB7 suppressor of RNA polymerase B homolog (yeast)	NM_004264	NM_004264	Hs.286145	12p11.23
TACC1	Transforming, acidic coiled-coil containing protein 1	NM_006283	CR933618	Hs.279245	8p11
TAF3	TAF3 RNA polymerase II, TATA box binding protein (TBP)-associated facto	BC028077	BC062352	Hs.527688	10p15.1
TCL6	T-cell leukemia/lymphoma 6	NM_012468	AB035333	Hs.510368	14q32.1
TEK	TEK tyrosine kinase, endothelial (venous malformations, multiple cutaneous	NM_000459	NM_000459	Hs.89640	9p21
CD248	CD248 molecule, endosialin	NM_020404	BC051340	Hs.195727	11q13
TFDP1	Transcription factor Dp-1	NM_007111	NM_007111	Hs.79353	13q34
TFDP2	Transcription factor Dp-2 (E2F dimerization partner 2)	NM_006286	NM_006286	Hs.379018	3q23
TFF1	Trefoil factor 1 (breast cancer, estrogen-inducible sequence expressed in)	NM_003225	BM923753	Hs.162807	21q22.3
TFF3	Trefoil factor 3 (intestinal)	NM_003226	BU536516	Hs.82961	21q22.3
TGFA	Transforming growth factor, alpha	NM_003236	NM_003236	Hs.170009	2p13
TGFB3	Transforming growth factor, beta 3	NM_003239	AK122902	Hs.592317	14q24
TGFBR1	Transforming growth factor, beta receptor I (activin A receptor type II-like kir	NM_004612	NM_004612	Hs.494622	9q22
TGFBR2	Transforming growth factor, beta receptor II (70/80kDa)	NM_003242	NM_001024847	Hs.82028	3p22
LIN9	Lin-9 homolog (C. elegans)	NM_173083	BC045625	Hs.120817	1q42.12
THBS1	Thrombospondin 1	NM_003246	NM_003246	Hs.164226	15q15
THOC1	THO complex 1	NM_005131	AK055354	Hs.592342	18p11.32
TIE1	Tyrosine kinase with immunoglobulin-like and EGF-like domains 1	NM_005424	NM_005424	Hs.78824	1p34-p33
TIMP3	TIMP metallopeptidase inhibitor 3 (Sorsby fundus dystrophy, pseudoinflamn	NM_000362	AB051444	Hs.297324	22q12.1-q13.2|22q12
TIMP3	TIMP metallopeptidase inhibitor 3 (Sorsby fundus dystrophy, pseudoinflamn	NM_000362	AB051444	Hs.297324	22q12.1-q13.2|22q12
TMEM2	Transmembrane protein 2	NM_013390	AF137030	Hs.494146	9q13-q21
TNF	Tumor necrosis factor (TNF superfamily, member 2)	NM_000594	BC028148	Hs.241570	6p21.3
TNFRSF10B	Tumor necrosis factor receptor superfamily, member 10b	NM_003842	NM_003842	Hs.521456	8p22-p21
TNFRSF4	Tumor necrosis factor receptor superfamily, member 4	NM_003327	BC040257	Hs.129780	1p36
FAS	Fas (TNF receptor superfamily, member 6)	NM_000043	AB209361	Hs.244139	10q24.1
TNFSF10	Tumor necrosis factor (ligand) superfamily, member 10	NM_003810	NM_003810	Hs.478275	3q26
TNFSF11	Tumor necrosis factor (ligand) superfamily, member 11	NM_003701	AF053712	Hs.333791	13q14
FASLG	Fas ligand (TNF superfamily, member 6)	NM_000639	NM_000639	Hs.2007	1q23
TNRC4	Trinucleotide repeat containing 4	NM_007185	NM_007185	Hs.26047	1q21
TOP1	Topoisomerase (DNA) I	NM_003286	NM_003286	Hs.592136	20q12-q13.1
TOP2A	Topoisomerase (DNA) II alpha 170kDa	NM_001067	NM_001067	Hs.156346	17q21-q22
TOP2B	Topoisomerase (DNA) II beta 180kDa	NM_001068	NM_001068	Hs.475733	3p24
TOP3B	Topoisomerase (DNA) III beta	NM_003935	AL833505	Hs.436401	22q11.22
TOPBP1	Topoisomerase (DNA) II binding protein 1	NM_007027	D87448	Hs.53454	3q22.1
TP53	Tumor protein p53 (Li-Fraumeni syndrome)	NM_000546	DQ186648	Hs.408312	17p13.1
TP53BP1	Tumor protein p53 binding protein, 1	NM_005657	AF078776	Hs.440968	15q15-q21
TP53BP2	Tumor protein p53 binding protein, 2	NM_005426	NM_005426	Hs.523968	1q42.1
TP53I3	Tumor protein p53 inducible protein 3	AF010309	AK223382	Hs.50649	2p23.3
TRAF1	TNF receptor-associated factor 1	NM_005658	AL832989	Hs.531251	9q33-q34
TREX1	Three prime repair exonuclease 1	NM_033627	NM_033627	Hs.344812	3p21.3-p21.2
TSC2	Tuberous sclerosis 2	NM_000548	NM_000548	Hs.90303	16p13.3
TUBB1	Tubulin, beta 1	NM_030773	NM_030773	Hs.592143	20q13.32
UBE3A	Ubiquitin protein ligase E3A (human papilloma virus E6-associated protein,	NM_130839	AF400501	Hs.22543	15q11-q13
UCHL5	Ubiquitin carboxyl-terminal hydrolase L5	NM_015984	BC015381	Hs.591458	1q32
UNC5B	Unc-5 homolog B (C. elegans)	NM_170744	AB096256	Hs.585457	10q22.1
UNG2	Uracil-DNA glycosylase 2	NM_021147	NM_001024592	Hs.3041	5p15.2-p13.1
USP1	Ubiquitin specific peptidase 1	NM_003368	NM_003368	Hs.35086	1p31.3
VEGF	Vascular endothelial growth factor	NM_003376	AB209485	Hs.73793	6p12
RASL10B	RAS-like, family 10, member B	NM_033315	AK122652	Hs.437035	17q12
WEE1	WEE1 homolog (S. pombe)	NM_003390	BX641032	Hs.249441	11p15.3-p15.1
WIG1	P53 target zinc finger protein	NM_022470	AK122768	Hs.386299	3q26.3-q27
WISP1	WNT1 inducible signaling pathway protein 1	NM_003882	AF100779	Hs.492974	8q24.1-q24.3
WISP2	WNT1 inducible signaling pathway protein 2	NM_003881	AK074695	Hs.592145	20q12-q13.1
WNT1	Wingless-type MMTV integration site family, member 1	NM_005430	NM_005430	Hs.248164	12q13
WNT10B	Wingless-type MMTV integration site family, member 10B	NM_003394	U81787	Hs.91985	12q13
WNT2B	Wingless-type MMTV integration site family, member 2B	NM_004185	AK127449	Hs.258575	1p13
WNT3	Wingless-type MMTV integration site family, member 3	NM_030753	NM_030753	Hs.591180	17q21
WNT3A	Wingless-type MMTV integration site family, member 3A	NM_033131	NM_033131	Hs.336930	1q42
WNT4	Wingless-type MMTV integration site family, member 4	NM_030761	AY358947	Hs.591521	1p36.23-p35.1
WNT5A	Wingless-type MMTV integration site family, member 5A	NM_003392	NM_003392	Hs.561260	3p21-p14
WNT5B	Wingless-type MMTV integration site family, member 5B	NM_032642	BC001749	Hs.306051	12p13.3
WNT6	Wingless-type MMTV integration site family, member 6	NM_006522	AY009401	Hs.29764	2q35
WNT7A	Wingless-type MMTV integration site family, member 7A	NM_004625	NM_004625	Hs.72290	3p25
WNT8A	Wingless-type MMTV integration site family, member 8A	NM_058244	AB057725	Hs.591274	5q31
WT1	Wilms tumor 1	NM_024426	BC046461	Hs.591980	11p13
DNAH1	Dynein, axonemal, heavy polypeptide 1	AB037831	NM_015512	Hs.209786	3p21.1
XPC	Xeroderma pigmentosum, complementation group C	NM_004628	NM_004628	Hs.475538	3p25

XRCC1	X-ray repair complementing defective repair in Chinese hamster cells 1	NM_006297	CR591751	Hs.98493	19q13.2
XRCC2	X-ray repair complementing defective repair in Chinese hamster cells 2	NM_005431	CR749256	Hs.591828	7q36.1
XRCC3	X-ray repair complementing defective repair in Chinese hamster cells 3	NM_005432	AK126706	Hs.592325	14q32.3
XRCC4	X-ray repair complementing defective repair in Chinese hamster cells 4	NM_022550	NM_022550	Hs.567359	5q13-q14
XRCC5	X-ray repair complementing defective repair in Chinese hamster cells 5 (dou	NM_021141	NM_021141	Hs.388739	2q35
YWHAE	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein	NM_006761	NM_006761	Hs.513851	17p13.3
TRIM25	Tripartite motif-containing 25	NM_005082	NM_005082	Hs.528952	17q23.2
ZNF350	Zinc finger protein 350	NM_021632	NM_021632	Hs.407694	19q13.33

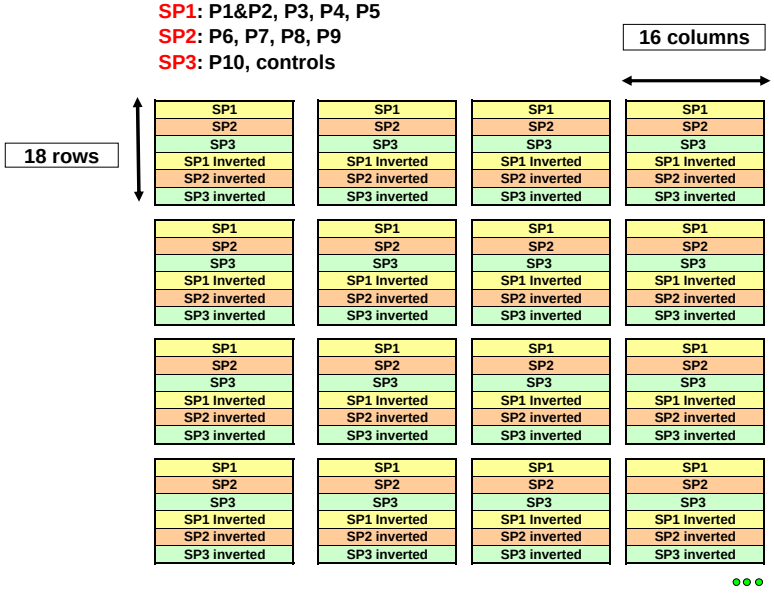
Breast Cancer Array v4.0



16 subgrids
each subgrid consisting in 16 columns x 18 rows
Subgrid total number of spots = 288
Array total number of spots = 4608
Number of 384-multiwell plates = 6
duplicated spots of each sample (vertical orientation)
Number of control spots = 48
3 plates 384-well plates spotted twice (inverted)

Spotting Plate Order:

SP1: P1&P2, P3, P4, P5
SP2: P6, P7, P8, P9
SP3: P10, controls



Belen Miñana
Laboratorio de Microarrays
CRG

Appendix A4: Clinical and histopathological patient data

17/12/2007

Array ID	Date	Sample ID	Prediction Analysis Microarrays	No Biopsia	nº clinical history	Age at diagnosis	Diagnose	POST-QUIMIO NEOADJUVANT	MENOPAUSAL STATUS	T (Tumor size)	N (Axilliary Lymph node status) + Number	M (metastasis) + sites	VASCULAR INVASION	TREATMENT (QT=chemotherapy, HT=hormone therapy)	DATE OF RECURRENCE	RECURRENCE	FINAL STATUS (VL=DISEASE-FREE, VE=WITH DISEASE)	p53	her2	ER	PR	BRCA1/2	Histological grade
2	01/02/2005	R7540	PAM 2	03B9690-T	928837	38	Infiltrating ductal carcinoma (IDC)	NO	PRE	T2	N3	M0	NO	QT	01/09/2003	SYSTEMICAL	VE	NEG	NEG	POS	NEG		G2
3	01/02/2005	R7545	PAM 2	03B3920	107120	65	Infiltrating ductal carcinoma (IDC) + Intraductal carcinoma	NO	POST	T1	N2	M0	YES	QT	NO		VL	NEG	NEG	NEG	NEG		G2
4	01/02/2005	R7553	PAM 5 (Basal-like)	03B12106-T	1107453	51	Infiltrating ductal carcinoma (IDC)	NO	PRE	T3	N1	M0	NO	QT NEOADJUVANT	01/06/2005	SYSTEMICAL	VE	NEG	NEG	NEG	NEG		G3
5	04/02/2005	R7612	PAM 2	02B8927	1061065	49	Infiltrating lobulillarl carcinoma (ILC)	NO	PRE	T1	N0	M0	NO	QT	NO		VL	POS	NEG	POS	POS		-
6	04/02/2005	R7614	PAM 2	03B13019	175284	67	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	NO	QT+HT	NO		VL	NEG	NEG	POS	POS		G2
7	14/02/2005	R7692	PAM 5 (Basal-like)	03B9980-T	330884	84	Metaplasic carcinoma + Intraductal carcinoma	NO	POST	T3	N1	M1	NO	QT	01/10/2003	LUNG	01/07/2005-EXITUS	POS	NEG	NEG	NEG		_
8	14/02/2005	R7693	PAM 5 (Basal-like)	03B9980-T	330884	84	Metaplasic carcinoma + Intraductal carcinoma	NO	POST	T3	N1	M1	NO	QT	01/10/2003	LUNG	01/07/2005-EXITUS	POS	NEG	NEG	NEG		_
9	23/02/2005	R7617	PAM 2	03B1059	421567	73	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	NO	HT	NO		VL	NEG	NEG	POS	POS		-
10	01/03/2005	R7541	PAM 3	04B986	1077524	57	Infiltrating ductal carcinoma (IDC) + Intraductal carcinoma	NO	POST	T2	N2	M0	NO	QT+HT	NO		VL	NEG	NEG	POS	POS		-
12	21/03/2005	R7906	PAM 2	03B8737-T	442621	83	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N1	M0	NO	HT	NO		VL	NEG	NEG	POS	POS		-
13	08/04/2005	R7904	PAM 3	02B8830-T/N	610750	81	Infiltrating ductal carcinoma (IDC)	NO	POST	T3	N0	M0	NO	HT	NO		VL	NEG	NEG	POS	POS		-
14	08/04/2005	R7939	PAM 2	03B11058	476285	63	Intraductal carcinoma	NO	POST	T1	N0	M0	NO	HT	NO		VL	NEG	NEG	POS	NEG		-
15	08/04/2005	R7946	PAM 2	03B213-A	539670	54	Infiltrating ductal carcinoma (IDC) + Intraductal carcinoma	NO	POST	T2	N1	M0	NO	QT+HT	NO		VL	NEG	NEG	POS	POS		-
17	22/04/2005	R8150	PAM 3	02B8678	168150	58	Intraductal carcinoma	NO	POST	IS	N0	M0	NO	-	NO		VL	POS	NEG	POS	POS		-
18	22/04/2005	R8153	PAM 2	02B8725-T	991556	61	Infiltrating ductal carcinomarcinoma (IDC)	NO	POST	T1	N0	M0	NO	QT+HT	NO		VL	NEG	NEG	POS	POS		-
20	20/05/2005	R8403	PAM 3	03B3358	1087199	52	Infiltrante medular carcinoma (IMC)	NO	POST	T2	N3	M0	NO	QT	NO		VL	POS	NEG	NEG	NEG		G3
22	10/01/2006	R10101	PAM 2	04B4839-A	225014	54	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N0	M0	NO	QT+HT	NO		VL	NEG	NEG	POS	POS		G1
23	10/01/2006	R10103	PAM 2	04B5128-3	107791	50	Infiltrating ductal carcinoma (IDC)	NO	PRE	T2	N1	M0	YES	QT+HT	NO		VL	NEG	NEG	POS	POS		G2
24	10/01/2006	R10104	PAM 2	04B989	422591	82	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	NO	HT	NO		VL	POS	NEG	POS	POS		G2
25	10/01/2006	R10105	PAM 2	04B3507	463332	55	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	NO	QT+HT	NO		VL	POS	NEG	POS	POS		G2
26	10/01/2006	R10107	PAM 2	04B1646-T	1137827	40	Infiltrating ductal carcinoma (IDC)	NO	PRE	T1	N1	M0	YES	QT+HT	NO		VL	POS	NEG	POS	POS		G3
27	10/01/2006	R10108	PAM 2	04B3805	850976	84	Infiltrating ductal carcinoma (IDC)	NO	POST	T4	N1	M0	YES	HT	NO		VL	POS	NEG	POS	NEG		G2
28	10/01/2006	R10109	PAM 3	04B1945	47838	62	Medular atypical carcinoma	NO	POST	T1	N0	M0	NO	QT	NO		VL	POS	NEG	NEG	NEG		G3
29	10/01/2006	R10112	PAM 3	04B5421-2A	254821	61	Lobulillar carcinoma	NO	POST	T1	N0	M0	NO	QT+HT	NO		VL	NEG	NEG	POS	POS		-
30	10/01/2006	R10113	PAM 2	04B3820	138535	48	Infiltrating ductal carcinoma (IDC)	NO	PRE	T1	N0	M0	YES	QT+HT	NO		VL	POS	NEG	POS	POS		G3
31	10/01/2006	R10115	PAM 2	04B1313-4TA	502904	79	Papilar carcinoma invasive (IPC)	NO	POST	T3	N0	M0	NO	HT	NO		VL	POS	NEG	POS	POS		G2
32	10/01/2006	R10116	PAM 2	04B1313-4TB	502904	79	Papilar carcinoma invasive (IPC)	NO	POST	T3	N0	M0	NO	HT	NO		VL	POS	NEG	POS	POS		G2
33	10/01/2006	R10117	PAM 2	04B2874	758269	42	Infiltrating ductal carcinoma (IDC)	NO	PRE	T1	N0	M0	NO	QT+HT	NO		VL	NEG	NEG	POS	NEG		G1

(-) unknown data

VL = disease-free
VE = with disease

QT = chemotherapy
HT = hormone therapy

Appendix A4: Clinical and histopathological patient data

17/12/2007

35	08/07/2004	R6277	PAM 4 (ERBB2+)	02B11538-T	70334	66	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N0	M0	NO	QT	NO		VL	NEG	3+/POLISOMICO AMPLIFICADO	NEG	NEG		G3
36	08/07/2004	R6279	NORMAL BREAST	02B12222-N	70334	66	NORMAL BREAST																
37	08/07/2004	R6282	PAM 2	02B11437-T	747740	69	Infiltrating lobulillarl carcinoma (ILC) + *"in situ"* lobulillar carcinoma	NO	POST	T1	N0	M0	NO	QT+HT	NO		VL	NEG	NEG	POS	POS		G1
38	08/07/2004	R6283	PAM 2	02B10486	1064204	89	Infiltrating ductal carcinoma (IDC) + intraductal solid type	NO	POST	T1	N0	M0	NO	HT	NO		VL	NEG	NEG	POS	POS		-
39	08/07/2004	R6250	PAM 3	02B1020-1	331909	68	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N0	M0	NO	QT	NO		VL	POS	NEG	NEG	NEG		G3
40	08/07/2004	R6254	PAM 5 (Basal-like)	02B478-TA	774806	77	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N1	M0	YES	QT	01/05/2004	LUNG	EXITUS JULIO/2005	POS	NEG	NEG	NEG		G3
41	08/07/2004	R6255	PAM 5 (Basal-like)	02B478-TB	774806	77	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N1	M0	YES	QT	01/05/2004	LUNG	EXITUS JULIO/2005	POS	NEG	NEG	NEG		G3
42	08/07/2004	R6265	PAM 3	02B5853-T	169883	53	Infiltrating ductal carcinoma (IDC) + "in situ" intraductal	NO	POST	T1	N1	M0	YES	QT	01/07/2007	LOCAL-CID	VE	POS	POS 3+	NEG	NEG		G3
43	08/07/2004	R6267	PAM 2	02B1757-A	1027181	78	Infiltrating ductal carcinoma (IDC)+ *"in situ"* intraductal	NO	POST	T2	N1	M0	YES	QT+HT	NO		VL	NEG	NEG	POS	POS		G3
44	08/07/2004	R6280	PAM 2	02B12222-T	1075673	82	Infiltrating lobulillar carcinoma (ILC) + *"in situ"* ductal carcinoma (IDC)	NO	POST	T2	N0	M0	NO	HT	NO		VL	NEG	NEG	POS	POS		G3
45	08/07/2004	R6285	PAM 2	02B10780	938536	62	Infiltrating ductal carcinoma (IDC)+Ca intraductal	NO	POST	T1	N0	M0	NO	QT+HT	NO		VL	NEG	2+/NO AMPLIFICADO	POS	POS		-
46	08/10/2004	R6345	PAM 3	02B10550-A	1018280	46	Infiltrating ductal carcinoma (IDC) + Intraductal carcinoma	NO	PRE	T1	N0	M0	NO	HT	NO		VL	NEG	NEG	POS	NEG		G1
47	08/07/2004	R6256	PAM 3	02B8601-T	640770	81	Infiltrating lobulillarl carcinoma (ILC)	NO	POST	T2	N2	M0	YES	QT	SI		VE	NEG	NEG	NEG	NEG		G3
48	08/07/2004	R7140	PAM 3	02B8601-T	640770	81	Infiltrating lobulillarl carcinoma (ILC)	NO	POST	T2	N2	M0	YES	QT	SI		VE	NEG	NEG	NEG	NEG		G3
49	08/07/2004	R7142	PAM 3	02B5536-T	113450	57	Infiltrating ductal carcinoma (IDC)	POST-QUIMIO	POST	T1	N0	M0	NO	HT	NO		VL	NEG	NEG	POS	POS		G1
50	08/07/2004	R7143	PAM 3	02B11630-T	133240	66	Infiltrating lobulillarl carcinoma (ILC)	NO	POST	T2	N3	M1	YES	QT	07/06/2005	LIVER	EXITUS NOV 2005	NEG	NEG	NEG	NEG		G3
51	08/07/2004	R7170	PAM 3	02B11964-T	39012	69	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	NO	QT	NO		VL	POS	NEG	POS	NEG		G2
52	03/12/2004	R7176	PAM 4 (ERBB2+)	02B3203-T	924333	58	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	MICRO	M0	NO	QT+HT	NO		VL	POS 20%	NEG	POS	POS		G2
53	03/12/2004	R7145	PAM 3	02B11349-T	958835	56	Infiltrante tubular carcinoma (ITC)	NO	POST	T1	N0	M0	NO	HT	NO		VL	NEG	NEG	POS	POS		G1
54	03/12/2004	R7156	PAM 4 (ERBB2+)	02B11538-T	70334	66	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N0	M0	NO	QT	NO		VL	NEG	3+/POLISOMICO AMPLIFICADO	NEG	NEG		G3
55	13/01/2005	R7322	PAM 4 (ERBB2+)	03B3918	12424	82	*"in situ"* lobulilar carcinoma	NO	POST	T2	N3	M0	YES	NO	NO FOLLOW-UP		-	NEG	POS	NEG	NEG		-
56	13/01/2005	R7328	PAM 2	03B8192-T	499276	57	Infiltrating ductal carcinoma (IDC)	POST-QUIMIO	POST	T4	N0	M0	YES	QT+HT	01/09/2006	VISCERAL	FEB 2007 EXITUS	NEG	NEG	POS	POS		-
57	13/01/2005	R7331	PAM 4 (ERBB2+)	03B6280-T	84794	58	Infiltrating ductal carcinoma (IDC) + *"in situ"* intraductal carcinoma	NO	POST	T2	N0	M0	NO	QT	NO		VL	POS	NEG	NEG	NEG		G3
58	13/01/2005	R7334	NORMAL BREAST	03B13956-N	1131458	60	NORMAL BREAST																
59	13/01/2005	R7335	PAM 3	03B13956-T	1131458	60	Ductal carcinoma with apocrine differentiation + intraductal carcinoma	NO	POST	T2	N0	M0	NO	-	NO FOLLOW-UP		NO FOLLOW-UP	POS (10%)	NEG	NEG	NEG		G1
60	13/01/2005	R7343	PAM 3	03B1070-T	1077899	84	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N0	M1	NO	HT	01/10/2004	BONE	EXITUS AGOSTO 2007	NEG	NEG	POS	NEG		G2
61	13/01/2005	R7346	PAM 3	03B2680-T	300752	40	Infiltrating ductal carcinoma (IDC)+ *"in situ"* intraductal carcinoma	NO	PRE	T1	N0	M0	NO	QT+HT	NO		VL	NEG	NEG	POS (LOW)	POS		G2
62	13/01/2005	R7347	PAM 3	03B13006	93530	56	Infiltrating ductal carcinoma (IDC) + "in situ" intraductal carcinoma	NO	POST	T2	N2	M0	NO	QT+HT	NO		VL	NEG	NEG	POS	NEG		G2

(-) unknown data

VL = disease-free
VE = with disease

QT = chemotherapy
HT = hormone therapy

Appendix A4: Clinical and histopathological patient data

17/12/2007

63	13/01/2005	R7351	PAM 5 (Basal-like)	03B6680-T	1048947	53	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N0	M0	NO	QT	NO		VL	POS	NEG	NEG	NEG	BRCA2	G2
64	15/03/2007	R12849	PAM 5 (Basal-like)	04B6078	852308	57	Infiltrating ductal carcinoma (IDC)	NO	0	T1	N0	M0	NO	QT	NO		VL	POS	NEG	NEG	NEG		G3
65	15/03/2007	R12853	PAM 1	04B7412	1155409	60	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N2	M0	NO	QT+HT	JUNE-2005	BONE and LIVER	EXITUS JUNE-2007	NEG	NEG	POS	NEG		G2
66	15/03/2007	R12854	PAM 5 (Basal-like)	04B7426	698695	79	celullar neuroendocrine carcinoma	NO	POST	T2	N1	M0	_	QT	FEBRUARY/2005	LOCAL, GANGLIONAR AND PLEURAL	EXITUS DEC-2006	POS	NEG	NEG	NEG		_
67	15/03/2007	R12856	PAM 1	04B7479	900870	78	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	NO	HT	NO		VL	NEG	NEG	POS	POS		G1
68	15/03/2007	R12860	PAM 1	04B8335	1161192	72	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N3	M0	NO	QT+HT	NO		VL	NEG	NEG	POS	POS		G2
69	15/03/2007	R12861	PAM 1	04B8509	533215	66	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	YES	HT	NO		VL	NEG	NEG	NEG	POS		G2
70	15/03/2007	R13626	NORMAL BREAST	04B8755-N	1146211	66	NORMAL BREAST																
71	15/03/2007	R13627	PAM 1	04B8755-T	1146211	66	Infiltrating ductal carcinoma (IDC)	POST-QUIMIO NEOADJUVANT	POST	T3	N1	M0	YES	QT+HT	NO		VL	NEG	NEG	POS	POS		G2
72	15/03/2007	R13628	PAM 1	04B8846	403695	41	Infiltrating ductal carcinoma (IDC)	NO	PRE	T2	N1	M0	NO	QT+HT	NO		VL	NEG	NEG	POS	POS		G3
73	15/03/2007	R13629	PAM 1	04B8940-3	787895	67	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	NO	HT	NO		VL	NEG	NEG	POS	POS		G1
74	15/03/2007	R13630	PAM 5 (Basal-like)	04B9287	1052080	59	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	NO	_	NO		VL	POS	NEG	NEG	NEG		G3
75	15/03/2007	R13631	PAM 1	04B9474	510624	54	Infiltrating ductal carcinoma (IDC)	NO	_	T1	N0	M0	NO	HT	NO		VL	NEG	NEG/1+	NEG	POS		G1
76	15/03/2007	R13632	PAM 2	04B9489	91985	66	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N0	M0	NO	QT+HT	NO		VL	NEG	NEG	POS	NEG		G2
77	15/03/2007	R13633	PAM 1	04B9669	116521	72	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N0	M0	NO	HT	NO		VL	NEG	NEG	POS	POS		G1
78	15/03/2007	R13713	PAM 5 (Basal-like)	05B11983	787007	73	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	NO	QT	NO		VL	POS	NEG	NEG	NEG		G3
79	15/03/2007	R13714	PAM 5 (Basal-like)	05B12298	1295730	51	Infiltrating ductal carcinoma (IDC)	NO	_	T2	N0	M0	YES	QT	19/09/2006	VISCERAL & CNS	EXITUS APRIL/2007	NEG	2+	NEG	NEG		G3
80	15/03/2007	R13715	PAM 1	05B12684-1	902995	61	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N0	M0	YES	QT+HT	NO		VL	NEG	NEG	POS	POS		G2
81	15/03/2007	R13719	PAM 5 (Basal-like)	05B1274-T	319722	56	Infiltrating ductal carcinoma (IDC)	NO	_	T2	N0	M0	NO	QT	NO		VL	POS	NEG	NEG	NEG		G3
82	15/03/2007	R13721	PAM 1	05B1287-T	785120	60	Infiltrating lobulillar carcinoma (ILC)	NO	POST	T2	N0	M0	NO	QT+HT	NO		VL	NEG	NEG	POS	NEG		-
83	15/03/2007	R13722	PAM 1	05B13766-1	730344	41	Infiltrating ductal carcinoma (IDC)	NO	PRE	T2	N1	M0	NO	QT+HT	NO		VL	NEG	NEG	POS	POS		G1
84	21/06/2007	R15301	PAM 3 (64%)	05B1879-T-B	229461	42	Infiltrating ductal carcinoma (IDC)+Ca intraductal no infilt	NO	PRE	T1	N0	M0	YES	QT+HT+HERCEPTIN	-		-	NEG	3+++	POS	NEG		G2
85	21/06/2007	R15302	PAM 3 (76%)	05B2482	529783	71	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N0	M0	NO	QT+HT	-		-	NEG	NEG-POLISOMICO	POS	POS		G2
86	21/06/2007	R15303	PAM 5 (100%)	05B3194-T2	804492	66	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	NO	QT NEOADJUVANT	_		_	POS	NEG	NEG	NEG		G3
87	21/06/2007	R15304	PAM 5 (100%)	05B3194-T1	804492	66	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	NO	QT NEOADJUVANT	_		_	POS	NEG	NEG	NEG		G3
88	21/06/2007	R15305	PAM 2 (75%)	05B3232	1089948	64	Infiltrating ductal carcinoma (IDC) + *"in situ"* intraductal carcinoma	NO	POST	T1	N1	M0	YES	QT+HT				NEG	NEG	POS	POS		G2
89	21/06/2007	R15306	PAM 3 (67%)	05B3262	1269742	73	Infiltrating ductal carcinoma (IDC) + *"in situ"* intraductal carcinoma	NO	POST	T1	N0	M0	NO	HT	-		-	NEG	NEG	POS	POS		G1
90	21/06/2007	R15307	PAM 3 (94%)	05B4217	1038941	59	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	NO	QT+HT	-		-	NEG	NEG	POS	POS		G1

(-) unknown data

VL = disease-free
VE = with disease

QT = chemotherapy
HT = hormone therapy

Appendix A4: Clinical and histopathological patient data

17/12/2007

91	21/06/2007	R15308	PAM 1 (91%)	04B6939-T1	1029905	58	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	NO	QT+HT	01/04/2007	LUNG and GANGLIONAR	VE	NEG	NEG	POS	NEG		-
92	21/06/2007	R15309	PAM 1 (64%)	04B6939-T2	1029905	58	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	NO	QT+HT	01/04/2007	LUNG and GANGLIONAR	VE	NEG	3+++	NEG	NEG		-
93	21/06/2007	R15326	PAM 6 (100%)	05B4526-1N	834974		NORMAL BREAST																
94	21/06/2007	R15328	PAM 5 (100%)	05B5491	756733	62	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N1	M0	NO	QT	_		_	NEG	NEG	NEG	NEG		G3
95	21/06/2007	R15330	PAM 1 (100%)	05B5735-1T	789865	59	Infiltrating lobulillar carcinoma (ILC)	NO	POST	T2	N0	M0	NO	QT+HT				NEG	NEG	POS	POS		-
96	21/06/2007	R15331	PAM 6 (100%)	05B5750-N	799616	34	NORMAL BREAST																
97	21/06/2007	R15332	PAM 1 (50%)	05B5750-T	799616	34	Infiltrating ductal carcinoma (IDC)	NO	PRE	T3	N1	M0	NO	QT+HERCEPTIN+HT				NEG	3+AMPLIFICADO FISH	POS	POS		G3
98	21/06/2007	R15333	UNCLASSIFIED	05B6615-1	136483	74	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N2	M0	YES	QT + HT				NEG	NEG	POS	POS		G2
99	21/06/2007	R15334	PAM 3 (95%)	05B14082	1069288	46	Infiltrating ductal carcinoma (IDC)	NO	PRE	T1	N1	M0	NO	QT+HT	-		-	NEG	NEG	POS	POS		G1
100	21/06/2007	R15335	PAM 3 (96%)	05B14109-2	858200	60	Infiltrating lobulillarl carcinoma (ILC)	NO	POST	T2	N1	M0	NO	QT+HT	-		-	NEG	NEG	POS	POS		-
101	21/06/2007	R15336	PAM 3 (100%)	05B8816-3	653988	60	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	NO	HT	NO		-	NEG	NEG	POS	POS		G1
102	21/06/2007	R15371	PAM 3 (79%)	05B7791-T	878199	67	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	-	-	NO		-	NEG	2+AMPLIFICADO	POS	POS		G3
103	21/06/2007	R15372	PAM 4 (50%)	05B6910	1284355	63	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N0	M0	NO	QT+ HT	-		-	POS	NEG1+	POS	NEG		G3
104	21/06/2007	R15374	PAM 3 (99%)	05B8038	1293585	50	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N1	M0	YES	QT+HT+HERCEPTIN	NO		-	NEG	3+AMPLIFICADO FISH	NEG	NEG		G3
105	21/06/2007	R15375	PAM 3 (99%)	05B8169	1101319	47	Infiltrating lobulillarl carcinoma (ILC)	NO	PRE	T2	N2	M0	YES	QT+HT	NO		-	POS	NEG	POS	POS		-
106	21/06/2007	R15376	PAM 3 (94%)	05B6946	427209	76	Infiltrating lobulillarl carcinoma (ILC)	NO	POST	T2	N0	M0	NO	HT	NO		-	NEG	NEG	POS	POS		-
107	21/06/2007	R15378	PAM 5 (100%)	05B7801-T	1289294	61	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N1	M0	_	QT	NO		_	POS	NEG	NEG	NEG		G3
108	21/06/2007	R15398	PAM 3 (77%)	05B9232-1T	294780	72	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N0	M0	NO	QT+HT	NO		-	NEG	NEG	POS	POS		G2
109	21/06/2007	R15399	PAM 6 (99%)	05B9085-1N	202854	75	NORMAL BREAST																
110	21/06/2007	R15400	PAM 5 (100%)	05B9085-1T	202854	75	Infiltrating ductal carcinoma (IDC)	NO	POST	T3	N1	M0	YES	QT	NO		_	POS	NEG	NEG	NEG		G3
111	21/06/2007	R15401	PAM 2 (100%)	05B9712-T	144369	74	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N1	M0	YES	QT+HT	NO			-	NEG	POS	POS		G2
112	21/06/2007	R15402	PAM 2 (73%)	05B9882	1301901	74	Mucinous carcinoma	NO	POST	T1	N1	M0	YES	QT+HT	NO			NEG	NEG	POS	POS		-
113	21/06/2007	R15404	PAM 1 (97%)	05B966	39236	68	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N1	M0	YES	QT+HT	NO			NEG	NEG	POS	NEG		G2
114	21/06/2007	R15405	PAM 1 (100%)	05B981	1249812	72	Infiltrating ductal carcinoma (IDC)	NO	POST	T1	N3	M0	YES	QT+HT	NO			NEG	NEG 1+	POS	POS		G2
115	21/06/2007	R15406	PAM 1 (56%)	05B8909-1	1263134	36	Infiltrating ductal carcinoma (IDC)	NO	PRE	T1	N0	M0	YES	QT+HT	NO			NEG	NEG	POS	POS		G2
116	21/06/2007	R15407	PAM 1 (94%)	06B10736-2	633021	68	Infiltrating ductal carcinoma (IDC)	NO	POST	T2	N1	M0	YES	QT+HT	NO			POS	NEG	POS	POS		G3
117	21/06/2007	R15408	PAM 1 (91%)	06B10729-2	442435	52	Infiltrating ductal carcinoma (IDC)	NO	PRE	T1	N1	M0	NO	QT+HT	NO			NEG	NEG	POS	POS		G1

(-) unknown data

VL = disease-free
VE = with disease

QT = chemotherapy
HT = hormone therapy

Appendix A5: List of Real Time qPCR primers

17/12/2007

Symbol	Name	Aliases	UGCluster	LLID	UGRepAcc	Cytoband	RefSeq	transcript size (bp)	primer	start (bp)	engh	tm (°C)	gc%	any cc	3' cor	repeat	oligo sequence	Oligo Name	product size (bp)	localization (CDS	intron size
ACTN1	Actinin, alpha 1	||ACTN1||ACTININ, ALPHA-1||alpha-actinin 1||Actinin, alpha 1||	Hs.509765	87	BX641076	14q24.1-q24.2	NM_001102	3398												231		1112
									left	1738	20	59.97	50	3.00	2.00	11.00	gcacccttcaacaactggat	ACTN1-left		exon 14	
									right	1968	20	60.09	50	2.00	0.00	11.00	gatggttgtgtaggggtgg	ACTN1-right		exon 15	
GAPD	Glyceraldehyde-3-phosphate dehydrogenase	||G3PD||MGC88685||1.2.1.12||GAPD||GAPDH||Glyceraldehyde-3-phosphate dehydrogenase||OCAS, p38 COMPONENT||OCT1 COACTIVATOR IN S PHASE, 38-KD COMPONENT||	Hs.544577	2597	BF983396	12p13	NM_002046	1310												238		1632+90+129
									left	122	20	59.97	50	3.00	0.00	10.00	gagtcaacggatttggtcgt	GAPD-left		exon 2-exon3	
									right	359	20	60.01	45	2.00	2.00	11.00	ttgattttggagggatctcg	GAPD-right		exon 5	
MRPL19	Mitochondrial ribosomal protein L19	||MRPL19||mitochondrial ribosomal protein L19||	Hs.44024	9801	BF242952	2q11.1-q11.2	NM_014763	1347												197		2078
									left	416	20	59.93	45	4.00	2.00	12.00	cagtttctggggatttgcat	MRPL19-left		exon 4	
									right	612	20	60.17	50	4.00	2.00	12.00	tattcaggaagggcatctcg	MRPL19-right		exon 5	
SF3A1	Splicing factor 3a, subunit 1, 120kDa	||SAP114||SF3A1||SF3A120||SF3A, 120-KD SUBUNIT||SPLICEOSOME-ASSOCIATED PROTEIN, 114-KD||PRP21, S. CEREVISIAE, HOMOLOG OF||splicing factor 3a, subunit 1, 120kDa||	Hs.406277	10291	NM_005877	22q12.2	NM_005877	2462												104		1272
									left	2072	20	60.35	55	4.00	2.00	11.00	tcatgccagaggaggagttc	SF3A1-left		exon 13	
									right	2175	20	59.93	45	3.00	1.00	11.00	ctgcccattcagtttccatt	SF3A1-right		exon 14	
PSMC4	Proteasome (prosome, macropain) 26S subunit, ATPase, 4	||S6||TBP7||PSMC4||TAT-BINDING PROTEIN 7||PROTEASE 26S, SUBUNIT 6||PROTEASOME 26S SUBUNIT, ATPase, 4||proteasome (prosome, macropain) 26S subunit, ATPase, 4||	Hs.211594	5704	CR611800	19q13.11-q13.13	NM_006503	1257												144		474
									left	1095	20	59.66	50	4.00	2.00	11.00	ccggccagataagatttcag	PSMC4-left		exon 10	
									right	1238	20	59.95	50	2.00	2.00	10.00	tcctgctcgtccttcttgat	PSMC4-right		exon 11	
PUM1	Pumilio homolog 1 (Drosophila)	||||PUM1||Pumilio homolog 1 (Drosophila)||	Hs.281707	9698; 28997	NM_001020658	1p35.2	NM_014676	3561												121		4315
									left	3156	20	60.07	50	7.00	1.00	11.00	cggtcgtcctgaggataaaa	PUM1-left		exon 20	
									right	3276	20	59.93	55	6.00	2.00	9.00	cgtacgtgaggcgtgagtaa	PUM1-right		exon 21	
TFF1	trefoil factor 1 (breast cancer, estrogen-inducible sequence expressed in)	||HP1.A||D21S21||HPS2||TFF1||BCEI||pNR-2||pS2 protein||GASTROINTESTINAL TREFOIL PROTEIN pS2||BREAST CANCER ESTROGEN-INDUCIBLE SEQUENCE||Breast cancer estrogen-inducible protein||Trefoil factor 1 precursor||trefoil factor 1 (breast cancer, estrogen-inducible protein)||	Hs.350470	7031	BM923753	21q22.3	NM_003225	492												209		759
									left	160	20	59.98	45	3.00	3.00	11.00	ttgtggttttcctggtgtca	TFF1-left		exon 2	
									right	368	20	60.35	55	6.00	2.00	10.00	ccgagctctgggactaatca	TFF1-right		exon 1	
EGFR	Epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)	||S7||EGFR||2.7.1.112||mENA||ONCOGENE ERBB||ERBB1 SPECIES ANTIGEN 7||truncated epidermal growth factor receptor||epidermal growth factor receptor isoform c||epidermal growth factor receptor isoform d||epidermal growth factor receptor isoform allepidermal growth factor receptor isoform b||V	Hs.488293	1956	NM_005228	7p12	NM_005228	5114												143		NA
									left	4728	20	60.01	40	4.00	0.00	11.00	tggattcatcagcatttgga	EGFR-left		exon28	
									right	4870	20	60.03	50	4.00	1.00	9.00	ctgatctatgccccaaagga	EGFR-right		exon28	
EGF	Epidermal growth factor (beta-urogastrone)	||URG||EGF||Epidermal growth factor (beta-urogastrone)||	Hs.419815	1950	NM_001963	4q25	NM_001963	4874												193		4788
									left	3316	20	60.05	55	5.00	1.00	10.00	tcacctcagggaagatgacc	EGF-left		exon 20	
									right	3508	20	60	55	3.00	3.00	11.00	cagttcccaccacttcaggt	EGF-right		exon 21	
DUSP1	Dual specificity phosphatase 1	||PTPN10||3.1.3.48||CL100||MKP1||MKP-1||DUSP1||3.1.3.16||HVH1||DUAL-SPECIFICITY PHOSPHATASE 1||Dual specificity phosphatase 1||PROTEIN-TYROSINE PHOSPHATASE, NONRECEPTOR-TYPE, 10||MAP KINASE PHOSPHATASE 1||serine/threonine specific protein phosphatase	Hs.171695	1843	AK127679	5q34	NM_004417	2448												183		NA
									left	1590	20	59.93	55	7.00	3.00	12.00	cagctgctgcagtttgagtc	DUSP1-left		exon4	
									right	1772	20	60.1	55	4.00	1.00	10.00	agaggtcgtaatggggctct	DUSP1-right		exon4	
TGFA	Transforming growth factor, alpha	||TGFA||Transforming growth factor, alpha||	Hs.170009	7039	NM_003236	2p13	NM_003236	4119												144		NA
									left	3423	20	59.99	50	5.00	2.00	9.00	cttcaagccaggttttcgag	TGFA-left		exon6	
									right	3566	20	60.2	50	4.00	0.00	10.00	ggcaggttggaagagatcaa	TGFA-right		exon6	
STAT5A	Signal transducer and activator of transcription 5A	||MGF||STAT5A||MAMMARY GLAND FACTOR||signal transducer and activator of transcription 5A||	Hs.437058	6776	NM_003152	17q11.2	NM_003152	4298												143		1022
									left	2832	20	60.17	55	4.00	2.00	11.00	ccccaggctccctataacat	STAT5A-left		exon19	
									right	2974	20	60.11	55	6.00	2.00	10.00	cgggagtcaagactgtccat	STAT5A-right		exon20	
STAT5B	Signal transducer and activator of transcription 5B	||STAT5B||STAT5B STAT5B/RARA FUSION GENE||signal transducer and activator of transcription 5B||	Hs.132864	5171	NM_012448	17q11.2	NM_012448	5171												151		NA
									left	3867	20	60.1	55	4.00	2.00	ggctacac	cctggaattgccactcagtt	STAT5B-left		exon19	
									right	4017	20	60.02	45	4.00	0.00	gttttggg	aacagcagcttcctgggtaa	STAT5B-right		exon19	
AURKC	aurora kinase C	||AURKC||AIE2||STK13||AURORA/IPL1/EG2-2||AIK3||aurora kinase C||SERINE/THREONINE PROTEIN KINASE 13||serine/threonine kinase 13 (aurora/IPL1-like)||	Hs.98338	6795	NM_003160	19q13.43	AF059681	1321 bps												125		317
									left	250	20	59.93	45.00	4.00	2.00	11.00	atttcattgtggccctgaag	AURKC-left		exon3	
									right	374	20	59.94	50.00	4.00	0.00	12.00	atacaggcgcaggatattgg	AURKC-right		exon4	
RPS6KA1	Ribosomal protein S6 kinase, 90kDa, polypeptide 1	||RSK1||RPS6KA1||RIBOSOMAL S6 KINASE 1||RIBOSOMAL PROTEIN S6 KINASE, 90-KD, 1||ribosomal protein S6 kinase, 90kDa, polypeptide 1||	Hs.149957	6195	NM_002953	1p36-p35	BC014966	3186 bp												179		166+1667
									left	538	20	59.98	55	4.00	0.00	9.00	gaagctgcactatgccttcc	RPS6KA1-left		exon5-exon6	
									right	716	20	59.96	50	5.00	3.00	12.00	aaatgatacccaggctgtgc	RPS6KA1-right		exon7	
RPS6KA5	Ribosomal protein S6 kinase, 90kDa, polypeptide 5	||MSK1||RPS6KA5||MITOGEN- AND STRESS-ACTIVATED PROTEIN KINASE 1||RIBOSOMAL PROTEIN S6 KINASE, 90-KD, 5||ribosomal protein S6 kinase, 90kDa, polypeptide 5||	Hs.510225	9252	NM_004755	14q31-q32.1	AB209667	3883 bp												133		15126
									left	1957	20	60.07	50	4.00	2.00	10.00	ctccatgcttcacccttcat	RPS6KA5-left		exon14	
									right	2089	20	60.04	50	3.00	1.00	12.00	tgagattggaagggaacctg	RPS6KA5-right		exon15	
CCND1	Cyclin D1 (PRAD1: parathyroid adenomatosis 1)	||BCL1||CCND1||D11S287E||U21B31||B-cell CLL/lymphoma 1||G1/S-specific cyclin D1||cyclin D1 (PRAD1: parathyroid adenomatosis 1)||B-CELL LEUKEMIA 1 CCND1/IGHG1 FUSION GENE||	Hs.523852	595	NM_053056	11q13	NM_053056	4304 bp												149		1519 bp
									left	327	20	60.1	55	3.00	3.00	10.00	ccctcggtgtcctacttcaa	CCND1-left		exon1	
									right	475	20	60.13	55	3.00	0.00	10.00	aggaagcggtccaggtagtt	CCND1-right		exon2	
CCNE2	Cyclin E2	||||CCNE2||Cyclin E2||	Hs.408658	9134; 55656	NM_057749	8q22.1	NM_057749	2748 bp												199		491 + 386
									left	1050	20	60	55	3.00	3.00	11.00	caggtttggagtgggacagt	CCNE2-left		exon10-exon11	
									right	1248	20	60.33	50	5.00	0.00	11.00	ccccctttctgaaggtgtt	CCNE2-right		exon12	
CDKN1C	Cyclin-dependent kinase inhibitor 1C (p57, Kip2)	||WBS||p57(KIP2)||BWCR||CDKN1C||BWS||Beckwith-Wiedemann syndrome||cyclin-dependent kinase inhibitor 1C (p57, Kip2)||	Hs.106070	1028	NM_000076	11p15.5	BG714000	1584 bp												147		NA
									left	1309	20	59.87	50	4.00	3.00	12.00	gaccgttcatgtagcagcaa	CDKN1C-left		exon3	
									right	1455	20	60.28	60	4.00	3.00	10.75	caccttgggaccagtgtacc	CDKN1C-right		exon3	
MAP3K6	Mitogen-activated protein kinase kinase kinase 6	||MAPKKK6||MAP3K6||Mitogen-activated protein kinase kinase kinase 6||	Hs.194694	9064	NM_004672	1p36.11	NM_004672	4333 bp												129		280 + 199
									left	3878	20	60.13	50	3.00	1.00	10.00	agcctccaactgctctttca	MAP3K6-left		exon26-exon27	
									right	4006	20	60.01	55	3.00	1.00	12.00	tggcataggtgagcagagtg	MAP3K6-right		exon28	
MUC2L	Homo sapiens transcribed sequence with weak similarity to protein sp:Q02817 (H.sapiens)	IMAGE:4754978	Hs.33106		BG675392			1230 bp										122		
									left	72	20	59.96	50	4.00	0.00	11.00	gatctcaagctgcgttttcc	MUC2L-left			

Appendix A5: List of Real Time qPCR primers

17/12/2007

	MUC2_HUMAN Mucin 2 precursor									right	193	20	59.84	55	4.00	2.00	10.00	cgacccacaatccctagtgt	MUC2L-right			
FOS	V-fos FBJ murine osteosarcoma viral oncogene homolog	||FOS||cellular oncogene c-fos||FBJ OSTEOSARCOMA VIRUS||v-fos FBJ murine osteosarcoma viral oncogene homolog||FBJ murine osteosarcoma viral (v-fos) oncogene homolog (oncogene FOS)||	Hs.25647	2353	NM_005252	14q24.3	BX647104	2121 bp											121		753
										left	223	20	60.05	60	5.00	0.00	11.00	ccggggatagcctctctcttac	FOS-left		exon1	
										right	343	20	59.97	50	3.00	3.00	11.00	gtgggaatgaagttggcact	FOS-right		exon2	
MYC	V-myc myelocytomatosis viral oncogene homolog (avian)	||MYC||ONCOGENE MYC||PROTOONCOGENE HOMOLOGOUS TO MYELOCYTOMATOSIS VIRUS||V-myc myelocytomatosis viral oncogene homolog (avian)||V-MYC AVIAN MYELOCYTOMATOSIS VIRAL ONCOGENE HOMOLOG||	Hs.202453	4609	NM_002467	8q24.12-q24.13	NM_002467	2377 bp											157		1376
										left	1227	20	59.64	55	3.00	0.00	11.00	tcggattctctgctctcctc	MYC-left		exon2	
										right	1383	20	60.37	50	3.00	2.00	11.00	cctgcctcttttccacagaa	MYC-right		exon3	
CDKN1A	Cyclin-dependent kinase inhibitor 1A (p21, Cip1)	||SDI1||WAF1||CDKN1A||MDA-6||p21CIP1||CAP20||CDK-interaction protein 1||DNA synthesis inhibitor||CDK-INTERACTING PROTEIN 1||wild-type p53-activated fragment 1||WILDTYPE p53-ACTIVATED FRAGMENT 1||melanoma differentiation associated protein 6||cyclin-dependent kinase inhibitor 1A (p21, Cip1)||	Hs.370771	1026	NM_078467	6p21.2	NM_078467	2281 bp											93		1204
										left	641	20	59.84	55	7.00	2.00	9.00	gactctcagggtcgaaaacg	CDKN1A-left		exon 2	
										right	733	20	60.03	55	3.00	0.00	12.00	ggattagggcttcctcttgg	CDKN1A-right		exon 3	
CDKN1B	Cyclin-dependent kinase inhibitor 1B (p27, Kip1)	||CDKN1B||P27KIP1||p27(KIP1)||CDKN4||cyclin-dependent kinase inhibitor 1B (p27, Kip1)||	Hs.238990	1027	NM_004064	12p13.1-p12	NM_004064	2422 bp											147		2081
										left	1039	20	59.85	55	6.00	1.00	12.00	ggcctcagaagacgtcaaac	CDKN1B-left		exon 2	
										right	1185	20	58.72	45	4.00	2.00	11.00	acaggatgtccattccatga	CDKN1B-right		exon 3	
CDC25A	Cell division cycle 25A	||CDC25A2||3.1.3.48||CDC25A||cell division cycle 25A||dual specificity phosphatase CDC25A||M-phase inducer phosphatase 1||cell division cycle 25A isoform a||cell division cycle 25A isoform b||	Hs.1634	993	NM_001789	3p21	NM_001789	3717 bp											165		258
										left	1707	20	60.04	50	4.00	0.00	11.00	gagatcgcctgggtaatgaa	CDC25A-left		exon 14	
										right	1871	20	59.99	50	5.00	2.00	11.00	tgcggaacttcttcaggtct	CDC25A-right		exon 15	
PLAUR	Plasminogen activator, urokinase receptor	||UPAR||PLAUR||CD87 ANTIGEN||UPA RECEPTOR||Plasminogen activator, urokinase receptor||PLASMINOGEN ACTIVATOR RECEPTOR, UROKINASE-TYPE||	Hs.466871	5329	CR601067	19q13	NM_002659	1548 bp											96		3067
										left	764	20	59.86	45	3.00	2.00	12.00	caatggtttccacaacaacg	PLAUR-left		exon 5	
										right	859	20	59.96	50	4.00	2.00	11.00	ggcagattttcaagctccag	PLAUR-right		exon 6	
SNAI1	Snail homolog 1 (Drosophila)	||SNAI1||snail homolog 1 (Drosophila)||SNAIL, DROSOPHILA, HOMOLOG OF, 1||	Hs.48029	6615	NM_005985	20q13.1-q13.2	NM_005985	1708 bp											134		3520
										left	637	20	61.43	65	5.00	1.00	11.00	ctctaggccctggctgctac	SNAI1-left		exon 2	
										right	770	20	59.66	55	3.00	0.00	11.00	tgacatctgagtgggtctgg	SNAI1-right		exon 3	
ATF3	activating transcription factor 3	||ATF3||Cyclic-AMP-dependent transcription factor ATF-3||Activating transcription factor 3||activating transcription factor 3 long isoform||activating transcription factor 3 delta Zip isoform||	Hs.460	467	NM_001040619	1q32.3	NM_004024	2033 bp											131		1123
										left	547	20	59.57	45	3.00	0.00	10.00	gaaaaagaggcgacgagaaa	ATF3-left		exon 3	
										right	677	20	59.86	50	5.00	2.00	11.00	tctgagccttcagttcagca	ATF3-rifgt		exon 4	
SERPINB2	Serpin peptidase inhibitor, clade B (ovalbumin), member 2	||PLANH2||SERPINB2||PAI2||UROKINASE INHIBITOR||MONOCYTE ARGININE-SERPIN||MONOCYTE-DERIVED PLASMINOGEN ACTIVATOR INHIBITOR||PLASMINOGEN ACTIVATOR INHIBITOR, TYPE 2||Serpin peptidase inhibitor, clade B (ovalbumin), member 2||serine (or cysteine) proteinase inhibitor, clade B (ovalbumin)	Hs.514913	5055	BC012609	18q21.3	NM_002575	1900 bp											97		
										left	878	20	60.07	45	4.00	2.00	10.00	cagatgaaattgccgatgtg	SERPINB2-left		exon 7	332
										right	974	20	60.12	45	3.00	0.00	12.00	gccattttgtctttgctggt	SERPINB2-right		exon 8	
CEBPD	CCAAT/enhancer binding protein (C/EBP), delta	||CEBPD||NF-IL6-beta||CRP3||C/EBP-delta||CELF||CCAAT/ENHANCER-BINDING PROTEIN, DELTA||CCAAT/enhancer binding protein delta||CCAAT/enhancer binding protein (C/EBP), delta||	Hs.440829	1052	NM_005195	8p11.2-p11.1	NM_005195	1269											101		
									BM924801	left	219	20	59.87	50	4.00	2.00	10.00	atcgacttcagcgcctacat	CEBPD-left		exon 1	NA
										right	319	20	59.85	45	3.00	3.00	12.00	gccttgtgattgctgttgaa	CEBPD-right		exon 1	
GATA3	GATA binding protein 3	||IMGC2346||MGC5445||GATA3||MGC5199||HDR||GATA-binding protein 3||ENHANCER-BINDING PROTEIN GATA3||GATA binding protein 3||GATA binding protein 3 isoform 2||GATA binding protein 3 isoform 1||	Hs.524134	2625	NM_001002295	10p15	NM_002051	3067											158		
										left	1502	20	60.03	55	5.00	3.00	10.00	gtcctgtgcgaactgtcaga	GATA3-left		exon 5	4140
										right	1659	20	60.08	45	2.00	0.00	11.00	tttttcggtttctggtctgg	GATA3-right		exon 6	
SERPINA3	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3	||alpha-1-antichymotrypsin||MGC88254||AACT||SERPINA3||ANTICHYMOTRYPSIN, ALPHA-1||serine (or cysteine) proteinase inhibitor, clade A member 3 precursor||Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3||serine (or cysteine)	Hs.534293	12	NM_001085	14q32.1	NM_001085	1629											134		1119
										left	1101	20	59.66	55	8.00	2.00	11.00	ctgacctgtcagggatcaca	SERPINA3-left		exon 4	
										right	1234	20	59.67	45	4.00	1.00	12.00	tgcagaaaggagggtgattt	SERPINA3-right		exon 5	
CAV1	Caveolin 1, caveolae protein, 22kDa	||MSTP085||VIP21||CAV1||caveolae protein, 22-kD||caveolin 1, beta isoform||caveolin 1, alpha isoform||caveolin 1 caveolae protein, 22kD||caveolin 1, caveolae protein, 22kDa||	Hs.74034	857	NM_001753	7q31.1	NM_001753	2704											152		32256
										left	381	20	60.04	55	5.00	1.00	11.00	gagctgagcgagaagcaagt	CAV1-left		exon 2	
										right	532	20	60.15	45	4.00	2.00	10.00	caaatgccgtcaaaactgtg	CAV1-right		exon 3	
APOD	Apolipoprotein D	||APOD||apolipoprotein D||apolipoprotein D precursor||	Hs.522555	347	BF790155	3q26.2-qter	NM_001647	1061											139		2141
										left	553	20	60.36	55	6.00	2.00	9.00	aagccaccccagttaacctc	APOD-left		exon 4	
										right	691	20	60.07	50	5.00	1.00	11.00	tggatgatgcaggtacagga	APOD-right		exon 5	
TOP2A	Topoisomerase (DNA) II alpha 170kDa	||TOP2A||DNA TOPOISOMERASE II||TOPOISOMERASE, DNA, II, ALPHA||topoisomerase (DNA) II alpha 170kDa||	Hs.156346	7153	NM_001067	17q21-q22	NM_001067	5698											175		317
										left	4492	20	60.05	45	4.00	0.00	11.00	gatcctgccaaaaccaagaa	TOP2A-left		exon 34	
										right	4666	20	60.07	50	4.00	2.00	10.00	gtacagattttgcccgagga	TOP2A-right		exon 35	
ERBB2	V-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	||HER-2/neu||ERBB2||2.7.1.112||TKR1 HERSTATIN||ONCOGENE ERBB2||erbB-2 isoform a||c-erb B2/neu protein||erbB-2 isoform b||neuroblastoma/glioblastoma derived oncogene homolog||ONCOGENE NGL, NEUROBLASTOMA- OR GLIOBLASTOMA-DERIVED)TYROSINE KINASE-TYPE CELL	Hs.446352	2064	NM_001005862	q11.2-q12|17q2	NM_004448	4624											181		141
										left	3494	20	60.17	45	5.00	2.00	11.00	tttgatggtgacctgggaat	ERBB2-left		exon 26	
										right	3674	20	59.68	50	4.00	0.00	10.00	gaacatctggctggttcaca	ERBB2-right		exon 27	
SMARCE1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1	||BAF57||SMARCE1||BRG1-ASSOCIATED FACTOR, 57-KD||SWI/SNF-RELATED, MATRIX-ASSOCIATED, ACTIN-DEPENDENT REGULATOR OF CHROMATIN, SUBFAMILY E, MEMBER 1||SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1||	Hs.547509	6605	BC069196	17q21.2	NM_003079	2425											170		1720
										left	1128	20	60.07	50	4.00	2.00	9.00	acgagaacattccgatggag	SMARCE1-left		exon 10	
										right	1297	20	60.02	60	5.00	3.00	9.00	gctctccgagccagtgttac	SMARCE1-right		exon 11	
SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4	||SNF2-BETA||BRG1||SMARCA4||BRM/SWI2-RELATED GENE 1||SWI/SNF-RELATED, MATRIX-ASSOCIATED, ACTIN-DEPENDENT REGULATOR OF CHROMATIN, SUBFAMILY A, MEMBER 4||SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4||	Hs.327527	6597	NM_003072	19p13.2	NM_003072	5681											171		1596
										left	5047	20	60.11	50	5.00	2.00	11.00	cggtccgtcaaagtgaagat	SMARCA4-left		exon 35	
										right	5217	20	60.14	55	3.00	0.00	10.00	gtcttcttcgctgccacttc	SMARCA4-right		exon 36	
DUSP6	Dual specificity phosphatase 6	||3.1.3.48||PYST1||MKP3||3.1.3.16||DUSP6||DUAL-SPECIFICITY PHOSPHATASE 6||Dual specificity phosphatase 6||MAP kinase phosphatase 3||serine/threonine specific protein phosphatase||dual specificity phosphatase 6 isoform a||dual specificity phosphatase 6 isoform b||	Hs.298654	1848	BC037236	12q22-q23	BC037236	2874											151		1026
										left	1256	20	59.94	45	5.00	1.00	11.00	agtttttccctgaggccatt	DUSP6-left		exon 2	
										right	1406	20	59.82	50	5.00	1.00	10.00	taggcatcgttcatcgacag	DUSP6-right		exon 3	
KRT5	Keratin 5 (epidermolysis bullosa simplex, Dowling-Meara/Kobner/Weber-Cockayne types)	||KRT5||K5||keratin 5 (epidermolysis bullosa simplex, Dowling-Meara/Kobner/Weber-Cockayne types)||	Hs.433845	3852	AJ508777	12q12-q13	NM_000424	2320											149		557+804
										left	1525	20	60.81	55	8.00	2.00	11.00	ggagctcatgaacaccaagc	KRT5-left		exon 7	
										right	1673	20	59.08	45	4.00	0.00	10.00	tgccatatccagaggaaaca	KRT5-right		exon 9	
S100A2	S100 calcium binding protein A2	||S100L||S100A2||S100 CALCIUM-BINDING PROTEIN A2||S100 calcium binding protein A2||	Hs.516484	6273	BU589956	1q21	NM_005978	970											182		2142
										left	365	20	60.33	60	3.00	1.00	11.00	ggctgtgctggtcactacct	S100A2-left		exon 2	
										right	546	20	60.02	55	4.00	1.00	11.00	cctgctggtcactgttctca	S100A2-right		exon 3	

Appendix A6: Genes regulated in T47D by the R5020 in a time course experiment

17/12/2007

Symbol	GenBank AcNo	Symbol/Name	Symbol/GO terms	Fold-Change R5020 (30 min)	Fold-Change R5020 (1 h)	Fold-Change R5020 (2 h)	Fold-Change R5020 (6 h)
AKAP13	NM_007200	AKAP13/A kinase (PRKA) anchor protein 13	AKAP13/cAMP-dependent protein kinase activity|intracellular signaling cascade|kinase activity|receptor activity|signal transducer activity	1.08	1.27	-1.09	1.56
AKAP2	NM_007203	AKAP2/PALM2-AKAP2 protein	AKAP2/A-kinase anchor protein 2	1.05	-1.21	-1.01	-1.72
ALPP	NM_001632	ALPP/Alkaline phosphatase, placental	ALPP/alkaline phosphatase activity|hydrolase activity|integral to membrane|magnesium ion binding|metabolism	-1.13	1.13	1.20	-1.56
ANKT	NM_016359	ANKT/Nucleolar and spindle associated protein 1	ANKT/NUSAP1/Nucleolar and spindle associated protein 1	-1.01	-1.08	-1.02	1.44
AP2B1	NM_001282	AP2B1/Adaptor-related protein complex 2, beta 1 subunit	AP2B1/intracellular protein |protein complex assembly	1.53	-1.04	-1.02	-3.43
AR	NM_000044	AR/Androgen receptor	AR/androgen receptor activity|cell proliferation|cell-cell signaling|nucleus|nucleus|regulation of transcription, DNA-dependent|signal transduction|steroid binding|transcription factor activity	1.06	1.03	1.30	-1.56
AREG	NM_001657	AREG/Amphiregulin	AREG/cell proliferation|cell-cell signaling|cytokine activity|growth factor activity|integral to membrane	-1.02	-1.03	-1.04	1.45
ATF3	NM_004024	ATF3/Activating transcription factor 3	ATF3/DNA binding|nucleus|regulation of transcription, DNA-dependent|transcription corepressor activity|transcription factor activity	1.02	1.11	1.32	1.56
BCAR1	NM_014567	BCAR1/Breast cancer anti-estrogen resistance 1	BCAR1/cell adhesion|cell proliferation|cytoplasm|protein binding|signal transducer activity	1.13	1.20	1.29	1.79
BCL2L1	NM_138578	BCL2L1/BCL2-like 1	BCL2L1/anti-apoptosis|apoptotic mitochondrial changes|integral to membrane|mitochondrial outer membrane|mitochondrion|negative regulation of survival gene product activity|regulation of apoptosis	1.09	1.11	1.39	1.69
BCL2L2	NM_004050	BCL2L2/BCL2-like 2	BCL2L2/anti-apoptosis|cytoplasm|membrane|regulation of apoptosis|spermatogenesis	1.11	-1.00	-1.10	-1.40
BTG1	NM_001731	BTG1/B-cell translocation gene 1, anti-proliferative	BTG1/B-cell translocation protein 1|B-cell translocation gene 1, anti-proliferative|	1.07	1.16	1.28	1.42
CACMKIINalpha	NM_018584	CACMKIINalpha/Calcium/calmodulin-dependent protein kinase II	CACMKIINalpha/kinase activity	-1.13	-1.01	1.23	-1.49
CCND1	NM_053056	CCND1/Cyclin D1	CCND1/G1/S transition of mitotic cell cycle|cytokinesis|nucleus|regulation of cell cycle	1.32	1.32	1.83	2.41
CCND2	NM_001759	CCND2/Cyclin D2	CCND2/cytokinesis|nucleus|regulation of cell cycle	1.03	-1.10	-1.27	-2.11
CCNE1	NM_001238	CCNE1/Cyclin E1	CCNE1/G1/S transition of mitotic cell cycle|cytokinesis|nucleus|nucleus|regulation of cell cycle	1.42	-1.06	-1.07	-1.38
CCNE2	NM_057749	CCNE2/Cyclin E2	CCNE2/cell cycle checkpoint|cytokinesis|nucleus|regulation of cell cycle|regulation of cyclin dependent protein kinase activity	1.23	1.06	1.27	2.19
CCNF	NM_001761	CCNF/Cyclin F	CCNF/cytokinesis|mitosis|nucleus|regulation of cell cycle	1.02	-1.03	1.01	-1.49
CCNG2	NM_004354	CCNG2/Cyclin G2	CCNG2/cell cycle|cell cycle checkpoint|cytokinesis|mitosis	-1.00	1.03	-1.67	-2.03
CDC2	NM_001786	CDC2/Cell division cycle 2, G1 to S and G2 to M	CDC2/ATP binding|cyclin-dependent protein kinase activity|cytokinesis|mitosis|nucleus|protein amino acid phosphorylation|protein serine/threonine kinase activity|protein tyrosine kinase activity|transferase activity|traversing start control point of mitotic cell cycle	1.01	-1.01	-1.19	1.59
CDC42	NM_001791	CDC42/Cell division cycle 42	CDC42/GTP binding|GTPase activity|actin filament organization|cytoplasm|establishment and/or maintenance of cell polarity|filopodium|macrophage cell differentiation|protein binding|protein |small GTPase mediated signal transduction	-1.02	-1.17	1.42	-1.21
CDC42BPA	NM_003607	CDC42BPA/CDC42 binding protein kinase alpha	CDC42BPA/ATP binding|intracellular signaling cascade|membrane|protein amino acid phosphorylation|protein amino acid phosphorylation|protein serine/threonine kinase activity|protein serine/threonine kinase activity|protein-tyrosine kinase activity|small GTPase regulatory/interacting protein activity	1.05	-1.04	1.03	-1.56
CDCA7	NM_031942	CDCA7/Cell division cycle associated 7	CDCA7/cytokinesis	-1.13	-1.07	1.28	-1.97
CDK8	NM_001260	CDK8/Cyclin-dependent kinase 8	CDK8/ATP binding|cytokinesis|protein amino acid phosphorylation|protein serine/threonine kinase activity|regulation of cell cycle|regulation of transcription, DNA-dependent|transferase activity	1.18	1.13	1.09	1.46
CDKN1B	NM_004064	CDKN1B/Cyclin-dependent kinase inhibitor 1B (p27, Kip1)	CDKN1B/cell cycle arrest|cyclin-dependent protein kinase inhibitor activity|cytoplasm|negative regulation of cell proliferation|nucleus|protein binding|regulation of cyclin dependent protein kinase activity|transforming growth factor beta receptor, cytoplasmic mediator activity	-1.09	-1.16	-1.13	-1.56
CDKN2B	NM_078487	CDKN2B/Cyclin-dependent kinase inhibitor 2B (p15, inhibits CDK4)	CDKN2B/cell cycle|cell cycle arrest|cyclin-dependent protein kinase inhibitor activity|cyclin-dependent protein kinase inhibitor activity|cytoplasm|kinase activity|negative regulation of cell cycle|negative regulation of cell proliferation|nucleus|regulation of cyclin dependent protein kinase activity	-1.30	-1.46	-1.88	-1.60
CENPA	NM_001809	CENPA/Centromere protein A, 17kDa	CENPA/DNA binding|chromatin binding|chromosome organization and biogenesis |nucleosome assembly|nucleus	-1.01	1.02	-1.29	1.46
CHD1L	NM_024568	CHD1L/Chromodomain helicase DNA binding protein 1-like	CHD1L/ATP binding|DNA binding|helicase activity|nucleus	1.05	1.11	1.41	1.92
CHD3	U91543	CHD3/Chromodomain helicase DNA binding protein 3	CHD1L/ATP binding|DNA binding|helicase activity|nucleus	-1.08	-1.02	1.18	-1.55
CHD4	NM_001273	CHD4/Chromodomain helicase DNA binding protein 4	CHD1L/ATP binding|DNA binding|helicase activity|nucleus	1.04	-1.04	1.03	-1.55
CHEK2	NM_007194	CHEK2/CHK2 checkpoint homolog	CHEK2/ATP binding|DNA damage checkpoint|cell cycle|kinase activity|membrane|nucleus|protein amino acid phosphorylation|protein serine/threonine kinase activity	1.04	-1.04	-1.00	-1.80
CKS2	NM_001827	CKS2/CDC28 protein kinase regulatory subunit 2	CKS2/cell cycle|cyclin-dependent protein kinase activity|cytokinesis|regulation of cyclin dependent protein kinase activity	1.41	1.16	-1.22	-1.12
COL4A1	NM_001845	COL4A1/Collagen, type IV, alpha 1	COL4A1/collagen|collagen type IV|cytoplasm|extracellular matrix structural constituent|phosphate	1.37	1.02	-1.16	-1.65
CXCL12	NM_000609	CXCL12/Chemokine (C-X-C motif) ligand 12	CXCL12/G-protein coupled receptor protein signaling pathway|calcium ion homeostasis|cell adhesion|cell-cell signaling|chemokine activity|chemotaxis|growth factor activity|inflammatory response/signal transduction	1.05	1.01	1.25	1.96
CXCR4	NM_003467	CXCR4/Chemokine (C-X-C motif) receptor 4	CXCR4/C-C chemokine receptor activity|C-X-C chemokine receptor activity|G-protein coupled receptor protein signaling pathway|integral to membrane|rhodopsin-like receptor activity	1.07	1.08	-1.03	-1.42
DC13	NM_020188	DC13/DC13 protein	DC13	1.02	1.06	-1.10	1.42

DDIT3	NM_004083	DDIT3/DNA-damage-inducible transcript 3	DDIT3/cell cycle arrest|nucleus|regulation of transcription, DNA-dependent|response to DNA damage stimulus|transcription corepressor activity|transcription factor activity	-1.07	-1.19	**-1.89**	**-1.49**
DDR1	NM_013994	DDR1/Discoidin domain receptor family, member 1	DDR1/ATP binding|cell adhesion|integral to plasma membrane|protein amino acid phosphorylation|protein kinase activity|protein serine/threonine kinase activity|protein-tyrosine kinase activity|receptor activity|transferase activity|transmembrane receptor protein tyrosine kinase activity|transmembrane receptor protein tyrosine kinase signaling pathway	1.02	-1.09	-1.07	**1.73**
DDX5	NM_004396	DDX5/DEAD (Asp-Glu-Ala-Asp) box polypeptide 5	DDX5/ATP binding|ATP-dependent helicase activity|RNA binding|RNA helicase activity|cell growth|nucleus	1.06	-1.21	-1.13	**1.56**
DKFZp686F2198	BX648653	DKFZp686F2198/Muskelin 1	DKFZp686F2198/MKLN1/Muskelin 1, intracellular mediator containing kelch motifs	-1.02	1.16	1.09	**1.56**
DLC1	NM_006094	DLC1/Deleted in liver cancer 1	DLC1/Rho GTPase activator activity|cytoplasm|cytoskeleton organization and biogenesis|extracellular region|negative regulation of cell growth|protein binding|regulation of cell adhesion	-1.03	**-1.54**	**-1.49**	-1.38
DUSP1	NM_004417	DUSP1/Dual specificity phosphatase 1	DUSP1/MAP kinase phosphatase activity|cell cycle|hydrolase activity|non-membrane spanning protein tyrosine phosphatase activity|protein amino acid dephosphorylation|response to oxidative stress	1.17	1.06	**1.83**	**3.97**
E2F3	NM_001949	E2F3/E2F transcription factor 3	E2F3/nucleus|protein binding|regulation of cell cycle|regulation of transcription, DNA-dependent|transcription factor activity|transcription factor complex|transcription initiation from Pol II promoter	1.12	-1.07	-1.09	**1.96**
EEF1A1	NM_001402	EEF1A1/Eukaryotic translation elongation factor 1 alpha 1	EEF1A1	-1.28	1.10	1.07	**-1.60**
EGF	NM_001963	EGF/Epidermal growth factor (beta-urogastrone)	EGF/DNA replication|activation of MAPK|calcium ion binding|chromosome organization and biogenesis (sensu Eukaryota)|epidermal growth factor receptor activating ligand activity|epidermal growth factor receptor signaling pathway|extracellular region|growth factor activity|integral to membrane|nucleus|plasma membrane|positive regulation of cell proliferation	-1.03	**1.51**	**1.93**	**2.99**
EGLN1	NM_022051	EGLN1/Egl nine homolog 1 (C. elegans)	EGLN1/cytosol|oxidoreductase activity|oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors|oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen|protein metabolism	-1.03	1.09	1.01	**1.41**
EGR1	NM_001964	EGR1/Early growth response 1	EGR1/nucleus|regulation of transcription, DNA-dependent|transcription factor activity|zinc ion binding	1.16	-1.16	1.04	**-2.64**
ELL2	NM_012081	ELL2/Elongation factor, RNA polymerase II, 2	ELL2/RNA elongation from Pol II promoter|RNA polymerase II transcription factor activity|nucleus|regulation of transcription, DNA-dependent|transcription elongation factor complex	1.21	**1.49**	**1.58**	**3.25**
ERBB2	NM_004448	ERBB2/V-erb-b2 erythroblastic leukemia viral oncogene homolog 2	ERBB2/ATP binding|ErbB-3 class receptor binding|cell proliferation|electron |electron er activity|epidermal growth factor receptor activity|extracellular region|integral to membrane|iron ion binding|membrane|non-membrane spanning protein tyrosine kinase activity|protein amino acid phosphorylation|protein amino acid phosphorylation|protein serine/threonine kinase activity|protein-tyrosine kinase activity|receptor activity|receptor signaling protein tyrosine kinase activity|transferase activity|transmembrane receptor protein tyrosine kinase signaling pathway|transmembrane receptor protein tyrosine kinase signaling pathway	-1.27	-1.16	-1.14	**-1.87**
ERBB3	NM_001982	ERBB3/V-erb-b2 erythroblastic leukemia viral oncogene homolog 3	ERBB3/ATP binding|epidermal growth factor receptor activity|integral to plasma membrane|protein amino acid phosphorylation|receptor activity|transferase activity|transmembrane receptor protein tyrosine kinase signaling pathway	-1.00	1.16	-1.17	**-1.77**
ESR1	NM_000125	ESR1/Estrogen receptor 1	ESR1/DNA binding|cell growth|chromatin remodeling complex|estrogen receptor activity||estrogen receptor signaling pathway|membrane|negative regulation of mitosis|nucleus|receptor activity|regulation of transcription, DNA-dependent|signal transduction|steroid binding|steroid hormone receptor activity|transcription factor activity	1.03	-1.00	-1.01	**-1.44**
FGF7	NM_002009	FGF7/Galactokinase 2	FGF7/cell proliferation|cell-cell signaling|epidermis development|extracellular region|growth factor activity|positive regulation of cell proliferation|regulation of cell cycle|response to wounding|signal transduction	1.05	1.01	-1.20	**-1.80**
FGFR2	NM_023028	FGFR2/Fibroblast growth factor receptor 2	FGFR2/ATP binding|fibroblast growth factor receptor activity|heparin binding|integral to membrane|protein amino acid phosphorylation|protein amino acid phosphorylation|protein serine/threonine kinase activity|protein tyrosine kinase activity|protein-tyrosine kinase activity|receptor activity|transferase activity	-1.01	1.05	-1.24	**-1.72**
FLJ32001	NM_152609	FLJ32001/Hypothetical protein FLJ32001	FLJ32001	-1.01	1.04	1.21	**1.60**
FN1	NM_002026	FN1/Fibronectin 1	FN1/acute-phase response|cell adhesion|cell migration|collagen binding|extracellular matrix structural constituent|extracellular region|extracellular region|heparin binding|metabolism|oxidoreductase activity|response to wounding	1.09	1.19	1.02	**-2.71**
FOS	NM_005252	FOS/V-fos FBJ murine osteosarcoma viral oncogene homolog	FOS/DNA binding|DNA methylation|inflammatory response|nucleus|regulation of transcription from Pol II promoter|specific RNA polymerase II transcription factor activity	1.04	-1.24	-1.27	**-1.72**
GADD45A	NM_001924	GADD45A/Growth arrest and DNA-damage-inducible, alpha	GADD45A/DNA repair|apoptosis|cell cycle arrest|nucleus|regulation of cyclin dependent protein kinase activity	1.21	**1.59**	**1.93**	**1.65**
GAS8	NM_001481	GAS8/Growth arrest-specific 8	GAS8/cytoskeleton|flagellum (sensu Eukaryota)|microtubule|molecular_function unknown|negative regulation of cell proliferation|sperm motility	-1.06	-1.00	1.01	**-1.44**
GATA3	NM_002051	GATA3/GATA binding protein 3	GATA3/defense response|morphogenesis|nucleus|perception of sound|regulation of transcription, DNA-dependent|transcription factor activity|transcription from Pol II promoter	-1.23	**-1.48**	**-1.82**	**-2.16**
GNAZ	NM_002073	GNAZ/Guanine nucleotide binding protein (G protein), alpha z	GNAZ/G-protein coupled receptor protein signaling pathway|GTP binding|GTPase activity|endoplasmic reticulum|nuclear membrane|plasma membrane|receptor signaling protein activity|signal transduction	-1.09	-1.32	**-1.45**	-1.09

GRB2	NM_002086	GRB2/Growth factor receptor-bound protein 2	GRB2/Ras protein signal transduction|SH3/SH2 adaptor protein activity|cell-cell signaling|epidermal growth factor receptor binding|epidermal growth factor receptor signaling pathway|intracellular signaling cascade	1.06	1.01	1.43	2.50
GTF2F2	NM_004128	GTF2F2/General transcription factor IIF, polypeptide 2, 30kDa	GTF2F2/ATP binding|DNA binding|RNA elongation from Pol II promoter|general RNA polymerase II transcription factor activity|helicase activity|nucleus|regulation of transcription, DNA-dependent|transcription factor TFIIF complex|transcription initiation from Pol II promoter	1.01	-1.10	-1.08	1.44
GTF2H2	NM_001515	GTF2H2/General transcription factor IIH, polypeptide 2, 44kDa	GTF2H2/transcription factor activity/DNA repair/Regulation of transcription, DNA-dependent/Nucleus	-1.10	1.03	1.06	1.58
H1F0	NM_005318	H1F0/H1 histone family, member 0	H1F0/DNA binding|DNA binding|chromosome|chromosome organization and biogenesis (sensu Eukaryota)|nucleosome|nucleosome assembly|nucleus|nucleus	1.37	1.22	1.04	-1.53
H2AFJ	NM_018267	H2AFJ/H2A histone family, member J	H2AFJ/DNA binding|chromosome organization and biogenesis (sensu Eukaryota)|nucleosome|nucleosome assembly|nucleus	1.03	-1.12	-1.33	-1.55
H2AFY	NM_004893	H2AFY/H2A histone family, member Y	H2AFY/DNA binding|DNA binding|chromosome|chromosome organization and biogenesis (sensu Eukaryota)|nucleosome|nucleosome|nucleosome assembly|nucleosome assembly|nucleus	1.13	-1.12	-1.02	1.49
HDAC9	NM_178423	HDAC9/Histone deacetylase 9	HDAC9/histone deacetylase activity|histone deacetylase complex|histone deacetylation|hydrolase activity|inflammatory response|negative regulation of myogenesis|nucleus|nucleus|regulation of cell cycle|regulation of transcription, DNA-dependent|specific transcriptional repressor activity|transcription factor binding	1.03	-1.00	-1.03	-1.43
HIST1H1C	NM_005319	HIST1H1C/Histone 1, H1c	HIST1H1C/DNA binding|chromosome|chromosome organization and biogenesis (sensu Eukaryota)|nucleosome|nucleosome assembly|nucleus	1.13	1.12	1.06	-1.52
HIST1H2AC	NM_003512	HIST1H2AC/Histone 1, H2ac	HIST1H2AC/DNA binding|chromosome|chromosome organization and biogenesis (sensu Eukaryota)|nucleosome|nucleosome assembly|nucleus	-1.15	-1.21	-1.05	-2.27
HMGB3	NM_005342	HMGB3/High-mobility group box 3	HMGB3/DNA binding activity|DNA binding|chromatin|development|nucleus|regulation of transcription, DNA-dependent	1.09	1.31	1.42	3.23
HSD11B2	NM_000196	HSD11B2/Hydroxysteroid (11-beta) dehydrogenase 2	HSD11B2/cell-cell signaling|glucocorticoid biosynthesis|metabolism|microsome|oxidoreductase activity	1.06	1.23	1.29	1.66
HSPB1	NM_001540	HSPB1/Heat shock 27kDa protein 1	HSPB1/cytoplasm|protein folding|regulation of translational initiation|response to unfolded protein|response to unfolded protein	-1.13	-1.31	1.09	-1.58
IFIT2	NM_001547	IFIT2/Interferon-induced protein with tetratricopeptide repeats 2	IFIT2/binding|cellular_component unknown|immune response	1.04	-1.20	-1.09	-1.64
IFITM1	NM_003641	IFITM1/Interferon induced transmembrane protein 1	IFITM1/cell surface receptor linked signal transduction|immune response|integral to membrane|negative regulation of cell proliferation|plasma membrane|receptor signaling protein activity|regulation of cell cycle|response to biotic stimulus	1.13	1.09	1.29	-1.57
IGFBP1	NM_000596	IGFBP1/Insulin-like growth factor binding protein 1	IGFBP1/regulation of cell growth|signal transduction	1.21	1.05	1.03	-1.85
IGFBP3	NM_000598	IGFBP3/Insulin-like growth factor binding protein 3	IGFBP3/negative regulation of signal transduction|positive regulation of apoptosis|positive regulation of myoblast differentiation|protein tyrosine phosphatase activator activity|regulation of cell growth	1.03	-1.06	1.09	-2.03
IGFBP5	NM_000599	IGFBP5/Insulin-like growth factor binding protein 5	IGFBP5/Signal Transduction/Regulation of cell growth	-1.27	-1.24	-1.01	-1.65
IL6ST	NM_002184	IL6ST/Interleukin 6 signal transducer	IL6ST/cell surface receptor linked signal transduction|immune response|integral to plasma membrane|interleukin-6 receptor activity|oncostatin-M receptor activity|receptor activity	1.04	1.89	2.46	2.38
ITGA2	NM_002203	ITGA2/Integrin, alpha 2	ITGA2/cell-matrix adhesion|collagen binding|integral to membrane|integrin complex|integrin-mediated signaling pathway|	1.02	1.01	-1.01	-2.38
JUN	NM_002228	JUN/V-jun sarcoma virus 17 oncogene homolog	JUN/RNA polymerase II transcription factor activity|nuclear chromosome|regulation of transcription, DNA-dependent|transcription factor activity	1.33	1.13	1.64	2.68
KIAA0349	NM_015255	KIAA0349/Chromosome 6 open reading frame 133	KIAA0349/Protein ubiquitination	1.10	1.06	-1.09	-1.91
KIAA1357	XM_050421_3	KIAA1357/GUKH2/Guanylate-kinase holder	KIAA1357	1.21	1.09	1.14	-1.52
KPNA3	NM_002267	KPNA3/	KPNA3/intracellular protein |nuclear localization sequence binding|nuclear pore|nucleus|protein complex assembly|protein er activity	1.11	1.11	1.12	1.95
KPNA4	NM_002268	KPNA4/Karyopherin alpha 4 (importin alpha 3)	KPNA4/intracellular protein |nucleus|protein er activity	-1.11	1.01	1.04	1.66
KPNB1	NM_002265	KPNB1/Karyopherin (importin) beta 1	KPNB1/cytoplasm|nuclear localization sequence binding|nuclear pore|nucleus|protein |protein er activity|protein-nucleus import, docking|protein-nucleus import, translocation|zinc ion binding	1.09	-1.05	-1.09	-1.52
KRT14	NM_000526	KRT14/Keratin 14	KRT14/intermediate filament|structural constituent of epidermis	-1.19	-1.24	1.17	-1.62
KRT17	NM_000422	KRT17/Keratin 17	KRT17/epidermis development|intermediate filament|intermediate filament|structural constituent of cytoskeleton|structural molecule activity	-1.24	-1.34	1.11	-1.53
KRT5	NM_000424	KRT5/Keratin 5	KRT5/epidermis development|intermediate filament|structural constituent of cytoskeleton	1.04	1.01	1.54	-1.37
KRT6B	NM_005555	KRT6B/Keratin 6B	KRT6B/cytoskeleton organization and biogenesis|ectoderm development|intermediate filament|intermediate filament|structural constituent of cytoskeleton|structural molecule activity	-1.04	-1.10	-1.19	-1.80
LMNA	NM_170707	LMNA/Lamin A/C	LMNA/lamin filament|muscle development|nucleus|protein binding|structural molecule activity|structural molecule activity	-1.08	-1.04	-1.00	-1.53
LMNB1	NM_005573	LMNB1/Lamin B1	LMNB1/lamin filament|nucleus|structural molecule activity	2.14	1.16	-1.10	1.12
LYRIC	NM_178812	LYRIC/LYRIC/3D3	LYRIC/integral to membrane|nucleus	-1.09	-1.01	-1.01	1.48
MAP3K1	AF042838	MAP3K1/Mitogen-activated protein kinase kinase kinase 1	MAP3K1/ATP binding|MAP kinase kinase kinase activity|magnesium ion binding|protein amino acid phosphorylation|protein serine/threonine kinase activity|protein ubiquitination|transferase activity|ubiquitin ligase complex|ubiquitin-protein ligase activity|zinc ion binding	1.04	1.14	1.96	-1.21
MAP3K3	NM_002401	MAP3K3/Mitogen-activated protein kinase kinase kinase 3	MAP3K3/ATP binding|MAP kinase kinase kinase activity|MAPKKK cascade|magnesium ion binding|positive regulation of I-kappaB kinase/NF-kappaB cascade|protein amino acid phosphorylation|protein serine/threonine kinase activity|signal transducer activity|transferase activity	-1.07	1.08	1.27	1.64

MCM6	NM_005915	MCM6/MCM6 minichromosome maintenance deficient 6	MCM6/ATP binding|DNA binding|DNA replication|DNA replication initiation|DNA-dependent ATPase activity|cell cycle|nucleus|regulation of transcription, DNA-dependent	1.18	-1.00	-1.04	**1.43**
MET	NM_000245	MET/Met proto-oncogene (hepatocyte growth factor receptor)	MET/ATP binding|cell proliferation|hepatocyte growth factor receptor activity|integral to plasma membrane|protein amino acid phosphorylation|protein binding|protein-tyrosine kinase activity|receptor activity|signal transduction|transferase activity	**1.46**	-1.01	-1.07	**-1.79**
MGST1	NM_020300	MGST1/Microsomal glutathione S-transferase 1	MGST1/glutathione transferase activity|membrane|microsome|mitochondrion|transferase activity	1.01	1.01	-1.07	**1.45**
MUC2L	BG675392	MUC2L/Mucin 2	MUC2L	1.21	**1.45**	**1.79**	**1.40**
MYB	NM_005375	MYB/V-myb myeloblastosis viral oncogene homolog	MYB/DNA binding|nuclear matrix|regulation of transcription, DNA-dependent|transcriptional activator activity	1.08	1.09	1.04	**-1.49**
MYC	NM_002467	MYC/V-myc myelocytomatosis viral oncogene homolog	MYC/cell cycle arrest|cell proliferation|iron ion homeostasis|nucleus|protein binding|regulation of transcription from Pol II promoter|transcription factor activity	1.30	**1.48**	**1.97**	1.04
NCOA3	NM_006534	NCOA3/Nuclear receptor coactivator 3	NCOA3/acyltransferase activity|histone acetyltransferase activity|nucleus|regulation of transcription, DNA-dependent|signal transducer activity|signal transduction|thyroid hormone receptor binding|transcription|transcription coactivator activity|transferase activity	1.18	1.23	-1.08	**-1.68**
NCOR2	NM_006312	NCOR2/Nuclear receptor co-repressor 2	NCOR2/DNA binding|nucleus|regulation of transcription, DNA-dependent|transcription corepressor activity	1.08	1.10	1.01	**-1.55**
NK4	NM_004221	NK4/Natural killer cell transcript 4	NK4/cell adhesion|immune response	1.12	1.20	**-1.72**	1.03
NME2	NM_002512	NME2/Non-metastatic cells 2, protein (NM23B)	NME2/ATP binding|CTP biosynthesis|GTP biosynthesis|UTP biosynthesis|kinase activity|magnesium ion binding|negative regulation of cell cycle|negative regulation of cell proliferation|nucleoside triphosphate biosynthesis|nucleoside-diphosphate kinase activity|nucleotide binding|nucleotide metabolism|nucleus|regulation of transcription, DNA-dependent|transcription factor activity|transferase activity	-1.13	-1.04	-1.02	**1.42**
NMU	NM_006681	NMU/Neuromedin U	NMU/digestion|neuropeptide signaling pathway|receptor binding|signal transduction	1.11	-1.06	1.04	**1.47**
OAS2	NM_002535	OAS2/2'-5'-oligoadenylate synthetase 2	OAS2/ATP binding|RNA binding|immune response|membrane|microsome|nucleobase, nucleoside, nucleotide and nucleic acid metabolism|nucleotidyltransferase activity|transferase activity	-1.09	1.09	1.15	**-1.78**
PC4	NM_006713	PC4/Activated RNA polymerase II transcription cofactor 4	PC4/nucleus|regulation of transcription from Pol II promoter|single-stranded DNA binding|transcription|transcription coactivator activity|transcription factor complex	1.22	1.13	-1.07	**1.60**
PCAF	NM_003884	PCAF/P300/CBP-associated factor	PCAF/N-acetyltransferase activity|cell cycle arrest|chromatin remodeling|histone acetyltransferase activity|histone deacetylase binding|negative regulation of cell proliferation|nucleus|protein amino acid acetylation|regulation of transcription, DNA-dependent|transcription cofactor activity|transferase activity	1.13	1.09	1.17	**2.06**
PDGFRA	NM_006206	PDGFRA/Platelet-derived growth factor receptor, alpha	PDGFRA/ATP binding|cell proliferation|cell surface receptor linked signal transduction|transmembrane receptor protein tyrosine kinase signaling pathway|vascular endothelial growth factor receptor activity	-1.21	-1.05	-1.25	**-1.49**
PECI	NM_006117	PECI/Peroxisomal D3,D2-enoyl-CoA isomerase	PECI/acyl-CoA binding|dodecenoyl-CoA delta-isomerase activity|fatty acid metabolism|isomerase activity|metabolism|peroxisome	-1.01	-1.01	1.20	**1.57**
PGRB	NM_000926	PGRB/Progesterone receptor B	PGRB/Steroid hormone receptor activity/Transcription factor activity/Regulation of transcription/Cell-cell signaling	1.29	-1.01	**1.41**	-1.34
PIK3CB	NM_006219	PIK3CB/Phosphoinositide-3-kinase, catalytic, beta polypeptide	PIK3CB/G-protein coupled receptor protein signaling pathway|activation of MAPK|regulation of cell cycle|signal transduction|transferase activity	-1.15	1.04	1.08	**1.69**
PLAU	NM_002658	PLAU/Plasminogen activator, urokinase	PLAU/blood coagulation|chemotaxis|chymotrypsin activity|hydrolase activity|kinase activity|negative regulation of blood coagulation|plasminogen activator activity|proteolysis and peptidolysis|signal transduction|trypsin activity	-1.12	-1.04	1.09	**-1.84**
PLAUR	NM_002659	PLAUR/Plasminogen activator, urokinase receptor	PLAUR/U-plasminogen activator receptor activity|blood coagulation|cell surface receptor linked signal transduction|chemotaxis|extrinsic to membrane|plasma membrane|protein binding	-1.00	-1.13	**-1.40**	-1.18
PMS2	NM_000535	PMS2/PMS2 postmeiotic segregation increased 2	PMS2/ATP binding|DNA binding|mismatch repair|negative regulation of cell cycle|nucleus	**-1.51**	-1.01	1.22	1.32
PRC1	NM_003981	PRC1/Protein regulator of cytokinesis 1	PRC1/cytokinesis|mitotic spindle elongation|nucleus|spindle microtubule	1.18	1.14	-1.11	**1.46**
RAB9A	NM_004251	RAB9A/RAB9A, member RAS oncogene family	RAB9A/GTP binding|GTPase activity |small GTPase mediated signal transduction	-1.08	-1.01	1.06	**1.48**
RAMP	NM_016448	RAMP/RA-regulated nuclear matrix-associated protein	RAMP/associates with the spliceosome late in the splicing pathway	1.10	-1.01	-1.05	**1.78**
RBBP4	NM_005610	RBBP4/Retinoblastoma binding protein 4	RBBP4/DNA repair|DNA replication|cell cycle|negative regulation of cell proliferation|nucleus|regulation of transcription, DNA-dependent	-1.01	1.04	**1.58**	1.11
RBBP7	NM_002893	RBBP7/Retinoblastoma binding protein 7	RBBP7/cell proliferation|development|nucleus	1.02	1.12	1.14	**1.85**
RELA	NM_021975	RELA/V-rel reticuloendotheliosis viral oncogene homolog A, nuclear factor (p65)	RELA/anti-apoptosis|nucleus|nucleus|positive regulation of I-kappaB kinase/NF-kappaB cascade|protein binding|regulation of transcription, DNA-dependent|response to toxin|signal transducer activity|transcription factor activity|transcription factor activity|transcription factor complex|transcription from Pol II promoter	1.01	1.07	1.12	**-1.54**
RFC3	NM_002915	RFC3/Replication factor C (activator 1) 3, 38kDa	RFC3/DNA binding|DNA replication|DNA replication factor C complex|DNA replication factor C complex|DNA strand elongation|DNA-directed DNA polymerase activity|delta-DNA polymerase cofactor complex|enzyme activator activity|nucleoside-triphosphatase activity|nucleotide binding|nucleus	1.08	1.03	-1.16	**1.60**
RPS6KA1	NM_002953	RPS6KA1/Ribosomal protein S6 kinase, 90kDa, polypeptide 1	RPS6KA1/ATP binding|protein amino acid phosphorylation|protein serine/threonine kinase activity|protein serine/threonine kinase activity|protein-tyrosine kinase activity|signal transduction|transferase activity	1.01	1.14	1.16	**1.56**

RPS6KA5	NM_004755	RPS6KA5/Ribosomal protein S6 kinase, 90kDa, polypeptide 5	RPS6KA5/ATP binding|epidermal growth factor receptor signaling pathway|histone phosphorylation|metallopeptidase activity|nucleus|nucleus|protein amino acid phosphorylation|protein kinase cascade|protein serine/threonine kinase activity|protein serine/threonine kinase activity|protein-tyrosine kinase activity|proteolysis and peptidolysis|regulation of transcription, DNA-dependent|response to chemical substance|response to external stimulus|response to stress|transferase activity|zinc ion binding	-1.06	-1.09	**1.59**	**3.76**
SAP30	NM_003864	SAP30/Sin3-associated polypeptide, 30kDa	SAP30/histone deacetylase complex|transcription corepressor activity	1.12	1.22	**1.58**	**2.20**
SCD4	NM_024906	SCD4/Stearoyl-CoA desaturase 4	SCD4/endoplasmic reticulum|fatty acid biosynthesis|iron ion binding|membrane|oxidoreductase activity|	1.24	-1.10	-1.15	**-1.78**
SCGB2A1	NM_002407	SCGB2A1/Secretoglobin, family 2A, member 1	SCGB2A1/androgen binding	-1.14	-1.21	-1.11	**-1.49**
SERPINA3	NM_001085	SERPINA3/Serine (or cysteine) proteinase inhibitor A	SERPINA3/DNA binding|acute-phase response|chymotrypsin inhibitor activity|extracellular region|inflammatory response|intracellular|protein binding|regulation of lipid metabolism|serine-type endopeptidase inhibitor activity	-1.37	**1.62**	-1.01	-1.10
SERPINB2	NM_002575	SERPINB2/Serine (or cysteine) proteinase inhibitor B	SERPINB2/anti-apoptosis|serine-type endopeptidase inhibitor activity	1.28	**1.44**	**1.43**	**-1.80**
SNAI1	NM_005985	SNAI1/Snail homolog 1	SNAI1/DNA binding|nucleus|zinc ion binding	1.05	-1.03	1.19	**2.69**
SNAI2	NM_003068	SNAI2/Snail homolog 2	SNAI2/DNA binding|development|ectoderm and mesoderm interaction|negative regulation of transcription from Pol II promoter|nucleus|regulation of transcription, DNA-dependent|zinc ion binding	-1.04	-1.12	**-1.53**	-1.21
SOS1	NM_005633	SOS1/Son of sevenless homolog 1	SOS1/DNA binding|Ras guanyl-nucleotide exchange factor activity|Ras protein signal transduction|Rho GTPase activator activity|Rho guanyl-nucleotide exchange factor activity	1.08	1.03	1.21	**1.75**
SP1	XM_028606	SP1/Sp1 transcription factor	SP1/DNA binding|RNA polymerase II transcription factor activity|nucleus|regulation of transcription, DNA-dependent|transcriptional activator activity|zinc ion binding	-1.02	-1.13	-1.15	**-2.20**
STRBP	NM_018387	STRBP/Spermatid perinuclear RNA binding protein	STRBP/DNA binding|double-stranded RNA binding|intracellular|nucleus|regulation of transcription, DNA-dependent	-1.05	-1.07	1.25	**2.04**
TFDP1	NM_007111	TFDP1/Transcription factor Dp-1	TFDP1/nucleus|regulation of cell cycle|regulation of transcription from Pol II promoter|transcription coactivator activity|transcription factor activity|transcription factor complex	1.09	-1.00	1.01	**1.42**
TFF1	NM_003225	TFF1/Trefoil factor 1	TFF1/carbohydrate metabolism|defense response|digestion|growth factor activity	1.24	1.13	-1.25	**-2.17**
TGFA	NM_003236	TGFA/Transforming growth factor, alpha	TGFA/cell proliferation|cell-cell signaling|epidermal growth factor receptor activating ligand activity|growth factor activity|integral to plasma membrane|protein binding|protein-tyrosine kinase activity|regulation of cell cycle|signal transducer activity|soluble fraction	1.09	1.29	**1.58**	**5.78**
TGS	NM_173083	TGS/Lin-9 homolog	TGS/LIN9/Receptor activity	1.06	-1.06	1.01	**1.55**
THBS1	NM_003246	THBS1/Thrombospondin 1	THBS1/blood coagulation|calcium ion binding|cell adhesion|cell motility|development|endopeptidase inhibitor activity|extracellular region|heparin binding|neurogenesis|protein binding|signal transducer activity|structural molecule activity	**1.41**	1.28	**1.60**	-1.15
TOP1	NM_003286	TOP1/Topoisomerase (DNA) I	TOP1/DNA topoisomerase type I activity|DNA topological change|DNA unwinding	1.04	1.04	1.11	**1.56**
TP53BP2	NM_005426	TP53BP2/Tumor protein p53 binding protein, 2	TP53BP2/SH3/SH2 adaptor protein activity|apoptosis|cytoplasm|regulation of cell cycle|signal transduction	-1.04	1.13	**1.75**	**2.39**
TP53I3	AF010309	TP53I3/Tumor protein p53 inducible protein 3	TP53I3/alcohol dehydrogenase activity, zinc-dependent|induction of apoptosis by oxidative stress|zinc ion binding	1.16	1.01	-1.00	**1.43**
WIG1	NM_022470	WIG1/P53 target zinc finger protein	WIG1/nucleic acid binding|nucleus|zinc ion binding	1.02	1.06	-1.01	**-1.58**
WNT5A	NM_003392	WNT5A/Wingless-type MMTV integration site family, member 5A	WNT5A/cell-cell signaling|frizzled-2 signaling pathway|receptor binding|signal transduction|	-1.09	-1.07	-1.06	**-1.51**
XLHSRF1	AB037831	XLHSRF-1/Heat shock regulated 1	XLHSRF-1/Heat shock regulated 1	1.16	1.03	1.31	**2.17**
ZNF350	NM_021632	ZNF350/Zinc finger protein 350	ZNF350/DNA binding|nucleus|regulation of transcription, DNA-dependent|zinc ion binding	-1.17	**-1.43**	-1.33	**-1.43**

Appendix A7: Overlap regulated genes at Cunliffe et al and our study

17/12/2007

GenBank Acc No	Symbol / Name	Fold Change (R5020 30min)	Fold Change (R5020 1h)	Fold Change (R5020 2h)	Fold Change (R5020 6h)	Fold Change (Cunliffe R5020 2h)	Fold Change (Cunliffe R5020 8h)	Fold Change (Cunliffe R5020 24h)	Overlap Regulated
NM_005688	ABCC5/ATP-binding cassette, sub-family C (CFTR/MRP), member 5	1.10	1.34	1.21	1.36	1.36	-1.30	-1.96	None
NM_001621	AHR/aryl hydrocarbon receptor	-1.08	1.16	1.08	1.13	-1.06	-1.61	-1.59	None
NM_005465	SDCCAG8/serologically defined colon cancer antigen 8	-1.01	-1.09	-1.11	-1.11	-1.19	-1.64	-1.52	None
NM_016359	ANKT/nucleolar protein ANKT	-1.01	-1.08	-1.02	1.44	-1.25	1.02	2.45	None
NM_001654	TIMP1/tissue inhibitor of metalloproteinase 1 (erythroid potentiating activity, collagenase inhibitor)	1.04	1.12	1.19	-1.05	-1.15	-1.22	-1.92	None
NM_004217	STK12/serine/threonine kinase 12	-1.03	-1.00	1.03	1.05	1.04	1.53	1.88	None
NM_014417	BBC3/BCL2 binding component 3	-1.10	1.07	1.01	-1.28	-2.63	-1.69	-1.25	None
NM_014567	BCAR1/breast cancer anti-estrogen resistance 1	1.13	1.19	1.29	1.79	2.20	2.19	1.33	Overlap
NM_001168	BIRC5/baculoviral IAP repeat-containing 5 (survivin)	-1.01	1.03	-1.08	1.38	1.13	1.09	2.50	None
NM_001731	BTG1/B-cell translocation gene 1, anti-proliferative	1.07	1.16	1.28	1.43	-1.08	9.88	10.49	None
NM_006763	BTG2/BTG family, member 2	1.49	1.02	-1.30	1.32	1.58	-1.39	1.06	None
NM_018584	PRO1489/hypothetical protein PRO1489	-1.12	-1.01	1.23	-1.50	-1.43	-1.67	-1.82	Overlap
NM_001237	CCNA2/cyclin A2	-1.03	-1.00	-1.15	1.24	-1.20	-1.25	2.12	None
NM_004701	CCNB2/cyclin B2	-1.05	-1.03	1.04	1.22	-1.06	-1.23	2.83	None
NM_053056	CCND1/cyclin D1 (PRAD1: parathyroid adenomatosis 1)	1.32	1.32	1.83	2.41	3.12	2.49	1.91	Overlap
NM_001240	CCNT1/cyclin T1	-1.06	-1.05	-1.03	-1.20	-1.04	-1.00	1.57	None
NM_001786	CDC2/cell division cycle 2, G1 to S and G2 to M	1.01	-1.01	-1.19	1.59	-1.10	-1.47	2.01	None
NM_001255	CDC20/CDC20 cell division cycle 20 homolog (S. cerevisiae)	1.15	1.15	-1.07	-1.03	1.04	1.29	2.97	None
NM_021874	CDC25B/cell division cycle 25B	-1.06	-1.00	-1.15	-1.04	-1.04	-1.14	2.13	None
NM_005197	CHES1/checkpoint suppressor 1	-1.05	-1.20	1.08	-1.07	-1.06	2.04	1.26	None
NM_001823	CKB/creatine kinase, brain	-1.17	1.03	1.10	-1.05	1.22	1.71	1.38	None
NM_001827	CKS2/CDC28 protein kinase 2	1.42	1.16	-1.22	-1.12	-1.11	-1.03	3.03	None
NM_018098	ECT2/epithelial cell transforming sequence 2 oncogene	-1.03	-1.06	-1.12	1.31	-1.47	-1.54	1.28	None
NM_012081	ELL2/ELL-related RNA polymerase II, elongation factor	1.21	1.49	1.58	3.26	2.95	1.31	1.41	Overlap
NM_004448	ERBB2/v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	-1.27	-1.17	-1.14	-1.86	-1.18	-2.44	-2.17	Overlap
NM_001982	ERBB3/v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)	1.00	1.16	-1.18	-1.76	-1.18	-1.49	-1.49	Overlap
NM_000125	ESR1/estrogen receptor 1	1.03	1.00	-1.01	-1.44	-1.19	-1.96	-1.64	Overlap
NM_002026	FN1/fibronectin 1	1.09	1.19	1.02	-2.72	-1.15	-1.43	-2.22	Overlap
NM_015675	GADD45B/growth arrest and DNA-damage-inducible, beta	1.06	-1.08	-1.07	1.11	1.44	1.78	1.78	None
NM_002051	GATA3/GATA binding protein 3	-1.23	-1.48	-1.82	-2.16	-2.78	-1.61	-1.69	Overlap
NM_000405	GM2A/GM2 ganglioside activator protein	-1.01	1.14	1.08	1.03	1.05	1.55	1.65	None
NM_002105	H2AFX/H2A histone family, member X	1.22	1.10	1.19	1.14	1.08	1.24	1.67	None

Appendix A7: Overlap regulated genes at Cunliffe et al and our study

17/12/2007

NM_004964	HDAC1/histone deacetylase 1	-1.04	1.07	-1.02	-1.16	-1.08	2.56	1.50	None
NM_003512	H2AFL/H2A histone family, member L	-1.14	-1.21	-1.05	-2.27	-1.02	-1.43	-1.30	Overlap
NM_145904	HMGA1/high mobility group AT-hook 1	1.10	1.05	1.08	-1.01	1.09	1.17	1.61	None
NM_005342	HMGB3/high-mobility group box 3	1.09	1.31	1.42	3.23	1.48	2.11	1.76	Overlap
NM_001547	EST/ESTs, Highly similar to IFT2_HUMAN Interferon-induced protein with tetratricopeptide repeats 2 (IFIT-2) (Interferon-induced 54 kDa protein) (IFI-54K) (ISG-54 K) [H.sapiens]	1.03	-1.20	-1.09	-1.64	-3.45	-12.50	-3.45	Overlap
NM_000875	IGF1R/insulin-like growth factor 1 receptor	1.03	1.04	1.09	1.25	1.23	1.52	-1.03	None
NM_000599	ESTs/Homo sapiens, clone IMAGE:4183312, mRNA, partial cds	-1.27	-1.24	-1.01	-1.64	-1.12	-2.04	-2.44	Overlap
NM_000210	ITGA6/integrin, alpha 6	1.27	1.20	1.03	-1.09	1.02	1.62	1.39	None
NM_000224	KRT18/keratin 18	-1.08	1.13	-1.14	-1.59	-1.54	-1.18	1.20	None
NM_145110	MAP2K3/mitogen-activated protein kinase kinase 3	1.16	1.04	-1.04	-1.04	-1.56	1.20	1.08	None
NM_002401	MAP3K3/mitogen-activated protein kinase kinase kinase 3	-1.07	1.08	1.26	1.63	1.10	2.39	1.41	Overlap
NM_006739	MCM5/MCM5 minichromosome maintenance deficient 5, cell division cycle 46 (S. cerevisiae)	1.15	-1.03	-1.11	-1.06	1.12	4.15	2.95	None
NM_014791	MELK/maternal embryonic leucine zipper kinase	1.01	-1.03	-1.14	1.11	-1.23	-1.08	-1.43	None
NM_004994	MMP9/matrix metalloproteinase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)	-1.05	1.10	1.01	-1.19	1.28	-1.00	1.58	None
NM_002456	MUC1/mucin 1, transmembrane	-1.14	1.00	-1.06	-1.01	1.04	1.51	1.61	None
NM_002462	MX1/myxovirus (influenza virus) resistance 1, interferon-inducible protein p78 (mouse)	-1.17	-1.05	1.23	1.03	-1.75	-5.26	-1.47	None
NM_005375	MYB/v-myb myeloblastosis viral oncogene homolog (avian)	1.08	1.09	1.03	-1.49	-1.16	-1.47	-2.22	Overlap
NM_002467	MYC/v-myc myelocytomatosis viral oncogene homolog (avian)	1.30	1.48	1.98	1.04	1.97	1.25	1.23	Overlap
NM_006534	NCOA3/nuclear receptor coactivator 3	1.18	1.23	-1.08	-1.68	-1.18	2.24	1.95	None
NM_003489	NRIP1/nuclear receptor interacting protein 1	1.17	1.04	1.09	-1.21	-1.43	-1.72	-1.28	None
NM_002634	PHB/prohibitin	1.00	1.11	1.01	1.02	1.09	1.43	1.41	None
M61906	PIK3R1/phosphoinositide-3-kinase, regulatory subunit, polypeptide 1 (p85 alpha)	1.03	-1.00	1.20	1.24	1.48	4.74	6.68	None
NM_002658	PLAU/plasminogen activator, urokinase	-1.11	-1.05	1.09	-1.85	-1.00	1.65	-1.09	None
NM_002659	PLAUR/plasminogen activator, urokinase receptor	1.00	-1.13	-1.41	-1.18	-1.20	1.98	1.25	None
NM_002826	QSCN6/quiescin Q6	-1.03	-1.02	1.12	-1.12	1.26	1.78	2.58	None
NM_002894	RBBP8/retinoblastoma binding protein 8	-1.06	-1.00	1.02	1.02	-1.41	1.10	1.16	None
NM_002915	RFC3/replication factor C (activator 1) 3, 38kDa	1.08	1.03	-1.15	1.61	1.05	1.45	1.40	Overlap
NM_002953	RPS6KA1/ribosomal protein S6 kinase, 90kDa, polypeptide 1	1.01	1.14	1.16	1.56	1.19	2.16	1.11	Overlap
NM_004755	RPS6KA5/ribosomal protein S6 kinase, 90kDa, polypeptide 5	-1.06	-1.10	1.59	3.77	1.19	1.55	1.26	Overlap
NM_020974	CEGP1/CEGP1 protein	-1.01	1.06	1.09	1.26	1.28	2.27	2.64	None
NM_015001	SHARP/SMART/HDAC1 associated repressor protein	-1.05	-1.12	-1.22	-1.05	-1.45	-1.56	-1.30	None
NM_198433	STK6/serine/threonine kinase 6	1.17	-1.01	1.04	-1.09	-1.28	-1.16	1.97	None
NM_003226	TFF3/trefoil factor 3 (intestinal)	-1.07	1.03	1.27	1.37	1.11	1.53	-1.23	None
NM_003246	THBS1/thrombospondin 1	1.41	1.28	1.60	-1.15	1.04	1.57	1.33	Overlap
NM_001067	TOP2A/topoisomerase (DNA) II alpha 170kDa	1.07	1.09	1.03	-1.08	-1.43	-1.64	2.56	None
NM_007027	TOPBP1/topoisomerase (DNA) II binding protein	-1.07	1.00	-1.13	-1.16	-1.22	3.03	1.48	None
NM_005426	TP53BP2/tumor protein p53 binding protein, 2	-1.03	1.13	1.76	2.40	1.93	2.23	1.92	Overlap
NM_021147	UNG2/uracil-DNA glycosylase 2	-1.03	-1.08	1.26	-1.37	-1.33	-2.50	-2.00	None
NM_005082	EST/Homo sapiens cDNA: FLJ20944 fis, clone ADSE01780	1.21	1.13	1.03	-1.38	-1.41	-2.86	-2.38	None

Appendix A8: Inhibition by PD and/or ICI of R5020 responsive genes

17/12/2007

			T0 and R5020 6 hr	T0 and R5020 6 hr	T0+PD and R5020+PD 6h	T0+PD and R5020+PD 6h	T0 and R5020+ICI 6 hr	T0 and R5020+ICI 6 hr	
Gene name	GenBank Acc. No.	Gene symbol	Fold	q-value	Fold	q-value	Fold	q-value	Inhibited by
Epidermal growth factor (beta-urogastrone)	NM_001963	EGF	5.77	0.0000	1.70	2.6945	5.42	0.0000	PD
Signal transducer and activator of transcription 5A	NM_003152	STAT5A	5.41	0.7161	5.69	0.0000	5.09	0.0000	
Transforming growth factor α	NM_003236	TGFA	4.81	0.0000	6.59	0.0000	7.93	0.0000	
Dual specificity phosphatase 1	NM_004417	DUSP1	4.73	0.7161	3.46	0.0000	3.63	0.0000	
Growth arrest and DNA-damage-inducible α	NM_001924	GADD45A	3.43	0.0000	2.47	1.5753	2.56	0.0000	
Cyclin D1	NM_053056	CCND1	3.29	0.0000	6.68	0.0000	4.00	0.0000	
High-mobility group box 3	Y10043	HMGB3	3.20	0.0000	3.10	0.0000	2.84	0.0000	
V-jun sarcoma virus 17 oncogene homolog (avian)	NM_002228	JUN	3.10	0.0000	2.53	0.0000	2.90	0.0000	
Interleukin 6 signal transducer (gp130. oncostatin M receptor)	NM_002184	IL6ST	2.89	0.7161	2.61	0.0000	3.14	0.0000	
Sin3-associated polypeptide 30kDa	NM_003864	SAP30	2.71	0.0000	2.07	1.5753	2.58	0.0000	
Kruppel-like factor 5 (intestinal)	NM_001730	KLF5	2.71	0.7161	3.08	0.0000	1.75	0.0732	
Dynein. axonemal. heavy polypeptide 1	AB037831	XLHSRF-1	2.70	0.0000	2.83	0.0000	3.09	0.0000	
RAS-like. family 10. member B	NM_033315	RASL10B	2.49	0.7161	1.33	NA	1.69	0.0000	PD & ICI
Elongation factor. RNA polymerase II. 2	NM_012081	ELL2	2.42	0.7161	1.81	1.1377	1.56	0.0732	PD & ICI
CDC14 cell division cycle 14 homolog B (S. cerevisiae)	NM_033332	CDC14B	2.35	0.0000	2.39	0.0000	1.87	0.0732	
Ribosomal protein S6 kinase 90kDa polypeptide 5	NM_004755	RPS6KA5	2.34	0.0000	2.66	0.0000	3.09	0.0000	
A kinase (PRKA) anchor protein 13	NM_007200	AKAP13	2.21	0.7161	1.88	1.1377	2.61	0.0000	
Growth arrest and DNA-damage-inducible β	NM_015675	GADD45B	2.19	0.7161	2.05	2.6945	2.71	0.0000	
Mitogen-activated protein kinase kinase kinase 3	NM_002401	MAP3K3	2.19	0.7161	2.86	0.0000	2.67	0.0000	
mucin 2 like	BG675392	MUC2L	2.19	0.7161	1.63	1.5753	1.99	0.0000	PD
Chemokine (C-X-C motif) ligand 12	NM_000609	CXCL12	2.18	0.7161	3.05	0.0000	2.43	0.0000	
Baculoviral IAP repeat-containing 3	NM_001165	BIRC3	2.16	0.7161	2.90	0.0000	1.49	0.0000	ICI
Peroxisome proliferative activated receptor γ coactivator 1 β	NM_133263	PPARGC1B	2.14	0.7161	1.33	NA	1.74	0.0000	PD
Epidermal growth factor receptor	NM_005228	EGFR	2.13	0.7161	-1.05	NA	-1.12	2.9822	PD & ICI
Tumor protein p53 binding protein 2	NM_005426	TP53BP2	2.11	0.7161	1.67	NA	1.14	4.0539	ICI
Growth factor receptor-bound protein 2	NM_002086	GRB2	1.88	1.7844	1.33	NA	1.82	0.0000	PD
Snail homolog 1 (Drosophila)	NM_005985	SNAI1	1.85	0.7161	1.96	1.1377	1.68	0.0000	
Cyclin E2	NM_057749	CCNE2	1.85	0.7161	1.41	3.1287	1.81	0.0000	PD
Vascular endothelial growth factor	NM_003376	VEGF	1.83	0.7161	1.74	2.6945	1.93	0.0000	
Cyclin-dependent kinase inhibitor 1C (p57. Kip2)	NM_000076	CDKN1C	1.82	1.7844	1.08	NA	-7.33	0.0000	PD & ICI
Activating transcription factor 3	NM_004024	ATF3	1.82	2.2677	1.24	NA	3.08	0.0000	
Serine (or cysteine) proteinase inhibitor clade A member 3	NM_001085	SERPINA3	1.80	0.7161	1.25	NA	-1.59	0.0000	ICI
Plasminogen activator. urokinase receptor	NM_002659	PLAUR	1.76	1.7844	-1.10	NA	1.46	0.0000	PD
E2F transcription factor 1	NM_005225	E2F1	1.73	1.7844	1.28	NA	1.35	0.0000	
Ribosomal protein S6 kinase. 90kDa. polypeptide 1	NM_002953	RPS6KA1	1.71	0.7161	1.90	1.1377	2.53	0.0000	
Phosphoinositide-3-kinase. catalytic. beta polypeptide	NM_006219	PIK3CB	1.70	0.7161	1.66	2.6945	2.15	0.0000	
Quiescin Q6	NM_002826	QSCN6	1.66	0.7161	1.37	3.1287	1.55	0.0000	PD
Cyclin-dependent kinase 8	NM_001260	CDK8	1.64	1.7844	1.60	1.5753	1.79	0.0000	
Signal transducer and activator of transcription 3	NM_139276	STAT3	1.64	4.3086	1.15	NA	1.56	0.0000	PD
Chromodomain helicase DNA binding protein 1 long isoform	NM_024568	CHD1L	1.63	1.2369	1.67	1.5753	1.57	0.0000	
Chromosome 1 open reading frame 71	NM_152609	FLJ32001	1.62	1.2369	1.89	1.1377	1.73	0.0000	
Parvin. alpha	NM_018222	PARVA	1.61	1.7844	1.48	1.5753	1.68	0.0000	
Heat shock 70kDa protein 9B (mortalin-2)	NM_004134	HSPA9B	1.60	3.2894	1.39	NA	-1.23	0.0000	ICI
General transcription factor IIH. polypeptide 2. 44kDa	NM_001515	GTF2H2	1.59	0.7161	1.32	NA	1.38	0.0000	
Inhibitor of growth family. member 1	NM_005537	ING1	1.59	1.2369	1.67	1.5753	1.56	0.0000	
P300/CBP-associated factor	NM_003884	PCAF	1.59	2.2677	2.16	1.5753	2.44	0.0000	
Karyopherin alpha 3 (importin alpha 4)	NM_002267	KPNA3	1.58	2.2677	2.01	1.1377	1.63	0.0000	
Denticleless homolog (Drosophila)	NM_016448	RAMP	1.56	1.7844	-1.09	NA	1.37	0.0000	PD
Insulin-like growth factor binding protein 3	NM_000598	IGFBP3	1.56	2.2677	1.35	NA	-2.54	0.0000	ICI
Thrombospondin 1	NM_003246	THBS1	1.54	2.2677	1.70	1.1377	2.17	0.0000	
MAX interactor 1	NM_005962	MXI1	1.54	3.2894	1.16	NA	1.37	0.0000	PD
Mitogen-activated protein kinase 7	NM_139033	MAPK7	1.53	2.2677	1.47	NA	1.74	0.0000	
Hydroxysteroid (17-beta) dehydrogenase 3	NM_000197	HSD17B3	1.52	3.2894	1.61	3.1287	1.24	0.7326	ICI
Myeloid cell leukemia sequence 1 (BCL2-related)	NM_021960	MCL1	1.50	3.2894	1.64	2.6945	1.32	0.0000	
Egl nine homolog 1 (C. elegans)	NM_022051	EGLN1	1.50	1.7844	1.08	NA	1.04	4.7046	PD & ICI
Growth arrest-specific 6	NM_000820	GAS6	1.48	2.2677	1.34	NA	1.64	0.0000	
H3 histone. family 3B (H3.3B)	NM_005324	H3F3B	1.48	2.2677	1.14	NA	1.15	2.8707	PD & ICI
Muskelin 1. intracellular mediator containing kelch motifs	BX648653	MKLN1	1.47	2.2677	1.12	NA	1.40	0.0000	PD
Phosphoinositide-3-kinase. class 2. beta polypeptide	NM_002646	PIK3C2B	1.46	3.2894	1.33	NA	1.98	0.0000	
Son of sevenless homolog 1 (Drosophila)	NM_005633	SOS1	1.45	2.2677	1.83	1.1377	1.13	2.8707	ICI
CDC6 cell division cycle 6 homolog (S. cerevisiae)	NM_001254	CDC6	1.45	3.2894	-1.05	NA	1.36	0.0000	PD
Signal transducer and activator of transcription 5B	NM_012448	STAT5B	1.43	2.2677	1.69	1.5753	1.55	0.0000	
Neogenin homolog 1 (chicken)	NM_002499	NEO1	1.43	2.2677	1.23	NA	1.51	0.0000	PD
Patched homolog 2 (Drosophila)	NM_003738	PTCH2	1.43	3.2894	-1.07	NA	-2.36	0.0000	PD & ICI
E2F transcription factor 3	NM_001949	E2F3	1.42	4.3086	1.62	2.6945	1.24	1.8657	ICI
Small nuclear ribonucleoprotein polypeptide N	NM_005678	SNRPN	1.42	3.2894	1.33	NA	1.52	0.0000	
Spermatid perinuclear RNA binding protein	NM_018387	STRBP	1.41	3.2894	1.51	3.1287	1.70	0.0000	
Origin recognition complex. subunit 6 homolog-like (yeast)	NM_014321	ORC6L	1.41	3.2894	1.22	NA	1.06	NA	PD & ICI
Fanconi anemia. complementation group A	NM_000135	FANCA	1.38	3.2894	-1.03	NA	1.11	2.8707	PD & ICI

Appendix A8: Inhibition by PD and/or ICI of R5020 responsive genes

17/12/2007

			T0 and R5020 6 hr	T0 and R5020 6 hr	T0+PD and R5020+PD 6h	T0+PD and R5020+PD 6h	T0 and R5020+ICI 6 hr	T0 and R5020+ICI 6 hr	
Gene name	GenBank Acc. No.	Gene symbol	Fold	*q*-value	Fold	*q*-value	Fold	*q*-value	Inhibited by
P21 (CDKN1A)-activated kinase 2	NM_002577	PAK2	-1.37	4.3086	-1.32	NA	-1.53	0.0000	
Protein phosphatase 1D magnesium-dependent. delta isoform	NM_003620	PPM1D	-1.38	4.3086	-1.22	NA	-1.93	0.0000	
Metastasis associated 1	NM_004689	MTA1	-1.43	4.3086	-1.15	NA	-1.39	0.0000	PD
Phosphoinositide-3-kinase. catalytic. alpha polypeptide	NM_006218	PIK3CA	-1.46	4.3086	-1.25	NA	-1.15	0.0000	
Chromodomain helicase DNA binding protein 4	NM_001273	CHD4	-1.46	4.3086	-1.42	NA	-1.56	0.0000	
LIM homeobox 3	NM_000107	DDB2	-1.46	3.4015	-1.21	NA	-1.27	0.0000	
SWI/SNF related, matrix associated, subfamily a. member 2	NM_003070	SMARCA2	-1.47	4.3086	-1.38	3.3663	-1.26	0.0000	
Ligase III. DNA. ATP-dependent	NM_013975	LIG3	-1.48	4.3086	-1.23	NA	-1.32	0.0000	
Nuclear receptor interacting protein 1	NM_003489	NRIP1	-1.49	4.3086	-1.38	NA	-1.58	0.0000	
Nuclear receptor coactivator 3	NM_006534	NCOA3	-1.49	3.4015	-1.11	NA	-1.30	0.0000	PD
Inhibitor of DNA binding 4. dominant negative helix-loop-helix pr	NM_001546	ID4	-1.50	4.3086	-1.32	NA	-1.76	0.0000	
Polymerase (DNA-directed). delta 4	NM_021173	POLD4	-1.50	4.3086	-1.23	NA	-1.24	0.0000	PD & ICI
Cell division cycle 25C	NM_001790	CDC25C	-1.51	3.4015	-1.62	3.3663	-1.39	0.0000	
CHK2 checkpoint homolog (S. pombe)	NM_007194	CHEK2	-1.51	4.3086	-1.29	NA	-1.45	0.0000	PD
Homeo box B7	NM_004502	HOXB7	-1.52	3.4015	-1.75	2.8979	-2.41	0.0000	
Rho GTPase activating protein 5	NM_001173	ARHGAP5	-1.53	4.3086	-1.16	NA	-1.30	0.0000	PD
Fibroblast growth factor receptor 2	NM_023028	FGFR2	-1.53	4.3086	-1.52	3.3663	-1.41	0.0000	
X-ray repair complementing defective repair in Chinese hamster	NM_022550	XRCC4	-1.54	4.3086	-1.45	NA	-1.57	0.0000	
Chromosome 14 open reading frame 109	BU739864	C14orf109	-1.60	3.4015	-1.40	3.3663	-1.78	0.0000	
Nuclear factor I	NM_005596	NFIB	-1.61	1.7844	-1.52	3.3663	-1.14	1.8657	ICI
Nibrin	NM_002485	NBS1	-1.62	3.4015	-1.30	NA	-1.72	0.0000	PD
Cyclin G2	NM_004354	CCNG2	-1.64	3.4015	-1.39	NA	-1.42	0.0000	
Transforming. acidic coiled-coil containing protein 1	NM_006283	TACC1	-1.65	4.3086	-1.40	NA	-1.77	0.0000	
Cyclin F	NM_001761	CCNF	-1.67	3.4015	-1.85	2.8979	-1.60	0.0000	
Insulin-like growth factor binding protein 4	NM_001552	IGFBP4	-1.67	3.4015	-1.13	NA	-1.55	0.0000	PD
Integrin. alpha 5 (fibronectin receptor. alpha polypeptide)	NM_002205	ITGA5	-1.71	3.4015	1.08	NA	-1.70	0.0000	PD
Transforming. acidic coiled-coil containing protein 2	BX111019	TACC2	-1.72	1.7844	-1.94	2.8979	-1.80	0.0000	
Histone 1. H2ac	NM_003512	HIST1H2AC	-1.73	4.3086	-1.36	NA	-2.03	0.0000	
Lamin A/C	NM_170707	LMNA	-1.74	3.4015	-1.39	3.3663	-1.21	0.0000	ICI
V-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avia	NM_001982	ERBB3	-1.75	1.7844	-1.98	3.3663	-1.63	0.0000	
Sp1 transcription factor	NM_138473	SP1	-1.77	1.6007	-1.70	3.3663	-2.24	0.0000	
Alkaline phosphatase. placental (Regan isozyme)	NM_001632	ALPP	-1.77	3.4015	-1.85	3.3663	-1.38	0.0000	
Notch homolog 3 (Drosophila)	NM_000435	NOTCH3	-1.88	1.6007	-1.54	3.3663	-1.20	0.0000	ICI
CDC14 cell division cycle 14 homolog A (S. cerevisiae)	NM_003672	CDC14A	-1.90	1.6007	-1.77	3.3663	-2.60	0.0000	
Integrin. alpha 2 (CD49B. alpha 2 subunit of VLA-2 receptor)	NM_002203	ITGA2	-1.94	1.6007	-1.73	2.8979	-2.02	0.0000	
H2A histone family. member J	NM_018267	H2AFJ	-1.94	1.6007	-3.06	0.0000	-2.17	0.0000	
DNA (cytosine-5-)-methyltransferase 3 alpha	NM_175629	DNMT3A	-1.95	4.3086	-1.85	NA	-6.36	0.0000	
Nuclear receptor co-repressor 2	NM_006312	NCOR2	-1.97	1.6007	-1.94	3.3663	-2.74	0.0000	
Insulin-like growth factor binding protein 5	NM_000599	IGFBP5	-1.98	1.6007	-1.05	NA	-1.25	0.0000	PD & ICI
PP12104	AF370408	PP12104	-2.01	1.7844	-1.99	2.8979	-2.51	0.0000	
Cell cycle progression 1	NM_004748	CCPG1	-2.03	1.6007	-1.40	3.3663	-1.99	0.0000	
GATA binding protein 3	NM_002051	GATA3	-2.22	1.6007	-2.68	0.0000	-1.44	0.0000	ICI
Interferon-induced protein with tetratricopeptide repeats 2	NM_001547	IFIT2	-2.26	1.6007	-5.74	2.8979	-3.88	0.0000	
Androgen receptor	NM_000044	AR	-2.29	1.7844	-1.64	NA	1.30	1.8657	ICI
Cadherin 13. H-cadherin (heart)	NM_001257	CDH13	-2.31	3.4015	-2.15	NA	-3.86	0.0000	
Histone deacetylase 9	NM_178423	HDAC9	-2.32	1.6007	-3.50	2.8979	-6.28	0.0000	
PiggyBac transposable element derived 3	NM_000124	PGBD3	-2.47	4.3086	-1.82	3.3663	-1.34	0.0732	
Calcium/calmodulin-dependent protein kinase II inhibitor 1	NM_018584	CaMKIINalph	-2.63	1.6007	-2.00	3.3663	-2.34	0.0000	
Zinc finger protein 350	NM_021632	ZNF350	-2.67	1.6007	-1.84	1.6167	-2.13	0.0000	
Uracil-DNA glycosylase 2	NM_021147	UNG2	-2.78	1.6007	-1.81	3.3663	1.07	4.9641	ICI
Cyclin-dependent kinase inhibitor 2B (p15. inhibits CDK4)	NM_078487	CDKN2B	-3.28	1.6007	-5.37	0.0000	-3.28	0.0000	

Appendix A9: Inhibition by PD and/or ICI of Estradiol responsive genes

17/12/2007

			T0 and E2 1 hr	T0 and E2 1 hr	T0+PD and E2+PD 1h	T0+PD and E2+PD 1h	T0 and E2+ICI 1 hr	T0 and E2+ICI 1 hr	
Gene name	GenBank Acc. No.	Gene symbol	Fold	*q*-value	Fold	*q*-value	Fold	*q*-value	Inhibited by
V-fos FBJ murine osteosarcoma viral oncogene homolog	NM_005252	FOS	9.05	0.0000	22.64	NA	3.19	NA	ICI
V-myc myelocytomatosis viral oncogene homolog (avian)	NM_002467	MYC	4.15	0.0000	4.30	0.0000	1.68	NA	ICI
Early growth response 1	NM_001964	EGR1	4.09	0.0000	3.45	NA	2.99	NA	
V-jun sarcoma virus 17 oncogene homolog (avian)	NM_002228	JUN	3.49	NA	3.12	NA	2.73	NA	
Activating transcription factor 3	NM_004024	ATF3	3.15	NA	1.32	NA	2.50	NA	PD
Dual specificity phosphatase 1	NM_004417	DUSP1	2.91	NA	5.58	NA	2.52	NA	
Trefoil factor 1	NM_003225	TFF1	2.62	NA	1.99	NA	-5.70	0.0000	ICI
Topoisomerase (DNA) III β	NM_003935	TOP3B	2.51	NA	-1.16	NA	1.46	NA	PD&ICI
Snail homolog 1 (Drosophila)	NM_005985	SNAI1	2.23	0.0000	1.46	NA	1.03	NA	PD&ICI
Insulin-like growth factor binding protein 3	NM_000598	IGFBP3	2.06	NA	1.57	NA	-1.43	NA	PD&ICI
Growth arrest and DNA-damage-inducible α	NM_001924	GADD45A	2.03	NA	-1.14	NA	1.28	NA	PD&ICI
Growth arrest and DNA-damage-inducible β	NM_015675	GADD45B	1.96	NA	3.76	NA	-1.09	NA	ICI
Tissue inhibitor of metalloproteinase 3	NM_000362	TIMP3	1.92	NA	1.02	NA	-1.07	NA	PD&ICI
Insulin-like growth factor binding protein 1	NM_000596	IGFBP1	1.86	NA	2.43	NA	-1.10	NA	ICI
Serine (or cysteine) proteinase inhibitor clade A member 3	NM_001085	SERPINA3	1.81	NA	-1.09	NA	-1.53	NA	PD&ICI
Cyclin-dependent kinase inhibitor 1C (p57 Kip2)	NM_000076	CDKN1C	1.74	NA	-2.05	NA	-8.38	NA	PD&ICI
Protein phosphatase 1 regulatory (inhibitor) subunit 15A	NM_014330	PPP1R15A	1.72	NA	1.08	NA	-1.12	NA	PD&ICI
DNA-damage-inducible transcript 3	NM_004083	DDIT3	1.68	NA	-1.91	NA	1.69	NA	PD
A kinase (PRKA) anchor protein 9	NM_147171	AKAP9	1.52	NA	1.02	NA	-1.50	3.4057	PD&ICI
Signal transducer and activator of transcription 5A	NM_003152	STAT5A	1.51	NA	1.24	NA	-1.15	NA	PD&ICI
Stearoyl-CoA desaturase 5	NM_024906	SCD4	1.48	NA	1.02	NA	-1.04	NA	PD&ICI
V-myb myeloblastosis viral oncogene homolog (avian)	NM_005375	MYB	1.45	NA	-1.37	NA	-1.90	3.2435	PD&ICI
Integrin. α 5 (fibronectin receptor. alpha polypeptide)	NM_002205	ITGA5	1.44	NA	1.24	NA	1.47	NA	PD
Keratin 5	NM_000424	KRT5	1.44	NA	-1.11	NA	-1.65	0.0000	PD&ICI
WNT1 inducible signaling pathway protein 2	NM_003881	WISP2	1.43	NA	1.24	NA	1.25	NA	PD&ICI
Inhibitor of DNA binding 4	NM_001546	ID4	1.35	NA	1.24	NA	1.16	NA	
Ectonucleotide pyrophosphatase phosphodiesterase 2	NM_006209	ENPP2	1.35	NA	2.07	NA	-4.39	NA	ICI
Cell division cycle 42 (GTP binding protein 25kDa)	NM_001791	CDC42	1.33	NA	1.30	NA	1.17	NA	ICI
Damage-specific DNA binding protein 1 127kDa	NM_001923	DDB1	1.32	NA	-1.10	NA	1.02	NA	PD&ICI
Hydroxysteroid (17-beta) dehydrogenase 3	NM_000197	HSD17B3	1.32	NA	1.24	NA	-1.27	NA	ICI
DNA (cytosine-5-)-methyltransferase 3 β	NM_006892	DNMT3B	1.32	NA	1.06	NA	-1.36	4.9293	PD&ICI
E2F transcription factor 1	NM_005225	E2F1	1.31	NA	-1.07	NA	1.08	NA	PD&ICI
CDC28 protein kinase regulatory subunit 2	NM_001827	CKS2	1.30	NA	1.27	NA	1.23	NA	

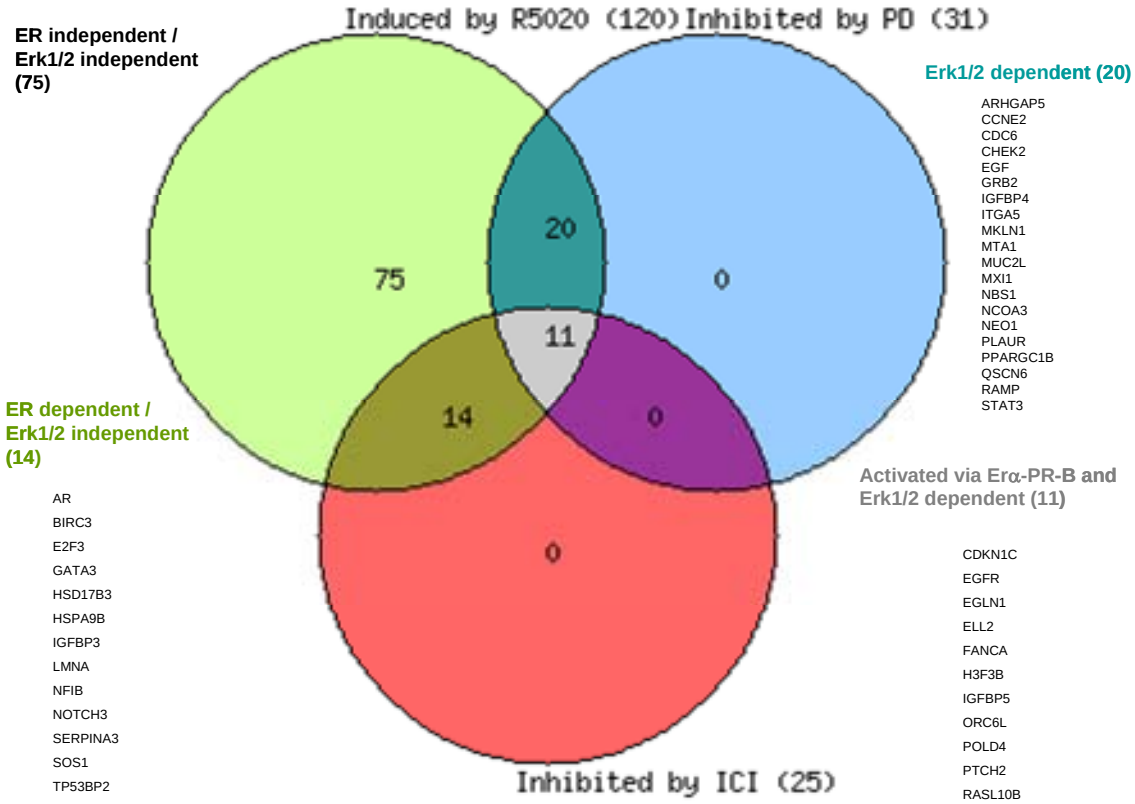
Appendix A9: Inhibition by PD and/or ICI of Estradiol responsive genes

17/12/2007

Rho GTPase activating protein 5	NM_001173	ARHGAP5	-1.31	NA	-1.14	NA	-1.20	NA	
RAD54 homolog B (S. cerevisiae)	NM_012415	RAD54B	-1.32	NA	1.10	NA	-1.10	NA	
Dynein. axonemal. heavy polypeptide 1	AB037831	XLHSRF-1	-1.33	NA	1.04	NA	-1.05	NA	
CDC14 cell division cycle 14 homolog A (S. cerevisiae)	NM_003672	CDC14A	-1.34	NA	1.70	NA	-1.60	4.1702	PD
Mucin 1. transmembrane	NM_002456	MUC1	-1.34	NA	1.10	NA	-1.13	NA	
Cyclin-dependent kinase inhibitor 2A (p16 inhibits CDK4)	NM_058197	CDKN2A	-1.34	NA	-1.18	NA	-1.37	NA	
Laminin α 3	NM_000227	LAMA3	-1.34	NA	1.19	NA	-1.35	4.9293	
Tumor protein p53 inducible protein 3	AF010309	TP53I3	-1.35	NA	-1.30	NA	-1.06	NA	
Cyclin-dependent kinase inhibitor 1A (p21 Cip1)	NM_078467	CDKN1A	-1.35	NA	-1.52	NA	1.10	NA	
Alkaline phosphatase. placental	NM_001632	ALPP	-1.35	NA	-1.15	NA	-1.35	4.1702	
Patched homolog (Drosophila)	AK124593	PTCH	-1.35	NA	1.26	NA	1.33	NA	ICI
Actin-like 6A	NM_178042	ACTL6A	-1.35	NA	1.08	NA	-1.07	NA	
Cell division cycle 25A	NM_001789	CDC25A	-1.36	NA	-1.29	NA	-1.27	NA	
Tissue inhibitor of metalloproteinase 3	NM_000362	TIMP3	-1.37	NA	-1.57	NA	-1.98	4.5409	
Zinc finger protein 350	NM_021632	ZNF350	-1.37	NA	1.37	NA	1.12	NA	PD
General transcription factor IIH. polypeptide 1 62kDa	NM_005316	GTF2H1	-1.38	NA	-1.39	NA	-1.12	NA	
Cyclin G2	NM_004354	CCNG2	-1.38	NA	-1.07	NA	-1.41	3.4057	
Pitrilysin metalloproteinase 1	NM_014889	PITRM1	-1.38	NA	-1.09	NA	-1.04	NA	
Mitogen-activated protein kinase kinase kinase 5	NM_005923	MAP3K5	-1.38	NA	1.30	NA	-1.47	4.5409	PD
Phosphoinositide-3-kinase class 2 α polypeptide	NM_002645	PIK3C2A	-1.38	NA	-1.15	NA	-1.12	NA	
Mitogen-activated protein kinase kinase kinase 6	NM_004672	MAP3K6	-1.39	NA	-1.37	NA	-1.12	NA	
Phosphoinositide-3-kinase catalytic α polypeptide	NM_006218	PIK3CA	-1.39	NA	-1.05	NA	-1.03	NA	PD
Adrenergic. alpha-1B-. receptor	NM_000679	ADRA1B	-1.39	NA	1.02	NA	-1.25	NA	PD
Forkhead box O3A	NM_001455	FOXO3A	-1.40	NA	-1.72	NA	-1.38	NA	
Breast cancer 2 early onset	NM_000059	BRCA2	-1.41	NA	1.23	NA	-1.27	NA	PD
Sin3-associated polypeptide 18kDa	NM_005870	SAP18	-1.41	NA	-1.17	NA	-1.57	3.4057	PD
RUN domain containing 1	NM_173079	RUNDC1	-1.42	NA	1.35	NA	-1.41	2.1285	PD
3-oxoacid CoA transferase 1	NM_000436	OXCT1	-1.43	NA	-1.35	NA	-2.08	4.1702	
Small EDRK-rich factor 1A (telomeric)	NM_021967	SERF1A	-1.44	NA	-1.17	NA	-1.97	2.1285	
Interleukin 6 signal transducer (gp130 oncostatin M receptor)	NM_002184	IL6ST	-1.44	NA	1.22	NA	-1.13	NA	PD
G protein-coupled receptor 126	NM_020455	GPR126	-1.45	NA	1.12	NA	-1.51	NA	PD
Catenin (cadherin-associated protein) β 1. 88 kDa	NM_001904	CTNNB1	-1.46	NA	1.15	NA	-5.53	0.0000	PD
Wilms tumor 1	NM_024426	WT1	-1.46	NA	-1.63	NA	-1.05	NA	ICI
Nuclear receptor coactivator 2	NM_006540	NCOA2	-1.47	NA	1.12	NA	-1.73	3.2435	PD
Aldehyde dehydrogenase 4 family member A1	NM_003748	ALDH4A1	-1.47	NA	-1.15	NA	-1.42	NA	
Mitogen-activated protein kinase 8	NM_139046	MAPK8	-1.47	NA	-1.11	NA	-1.66	3.4057	
Cell division cycle 27	NM_001256	CDC27	-1.48	NA	-1.17	NA	-1.31	3.2435	
Fms-related tyrosine kinase 1 (vascular endothelial growth factor)	NM_002019	FLT1	-1.51	NA	1.23	NA	-1.57	NA	PD
Estrogen-related receptor β	NM_004452	ESRRB	-1.54	NA	-1.02	NA	-2.28	4.5409	PD
Chromosome 20 open reading frame 46	NM_018354	C20orf46	-1.54	NA	1.91	NA	-2.52	4.5409	PD
Calcium/calmodulin-dependent protein kinase II inhibitor 1	NM_018584	CaMKIINalpha	-1.55	NA	-1.03	NA	-1.42	NA	PD
Uracil-DNA glycosylase 2	NM_021147	UNG2	-1.55	NA	1.13	NA	-1.53	4.1702	PD
Wingless-type MMTV integration site family. member 5B	NM_032642	WNT5B	-1.57	NA	1.34	NA	-2.52	3.2435	PD
V-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)	NM_001982	ERBB3	-1.59	NA	-1.04	NA	-1.32	NA	PD
Mitogen-activated protein kinase 13	NM_002754	MAPK13	-1.59	NA	1.14	NA	-1.55	3.2435	PD
Hydroxysteroid (17-β) dehydrogenase 7	NM_016371	HSD17B7	-1.60	NA	-1.41	NA	-1.12	NA	ICI
Ataxia telangiectasia mutated	NM_000051	ATM	-1.64	NA	1.22	NA	-1.45	NA	PD
GATA binding protein 3	NM_002051	GATA3	-1.66	NA	1.14	NA	-1.03	NA	PD&ICI
Microsomal glutathione S-transferase 1	NM_020300	MGST1	-1.70	NA	-1.16	NA	1.01	NA	PD&ICI
Glutathione S-transferase M3 (brain)	NM_000849	GSTM3	-1.72	NA	-1.10	NA	1.23	NA	PD&ICI
Ribosomal protein S6 kinase. 90kDa. polypeptide 5	NM_004755	RPS6KA5	-1.72	NA	-1.00	NA	-1.17	NA	PD&ICI
BRCA1 associated RING domain 1	NM_000465	BARD1	-1.75	NA	-1.37	NA	-1.58	3.4057	
Small nuclear ribonucleoprotein polypeptide N	NM_005678	SNRPN	-1.76	NA	-1.11	NA	-1.70	4.1702	PD
A disintegrin and metalloproteinase domain 15 (metargidin)	NM_003815	ADAM15	-1.79	NA	1.04	NA	-1.85	0.0000	PD
HIR histone cell cycle regulation defective homolog A (S. cerevisiae)	NM_003325	HIRA	-1.79	NA	1.02	NA	-1.93	NA	PD
Peroxisome proliferative activated receptor γ coactivator 1 β	NM_133263	PPARGC1B	-1.79	NA	-1.77	NA	-5.36	3.2435	
Androgen receptor	NM_000044	AR	-1.81	NA	1.40	NA	-1.62	4.1702	PD
Mdm2. transformed 3T3 cell double minute 2	NM_002392	MDM2	-1.82	NA	-1.34	NA	-1.68	NA	
Transforming growth factor β 3	NM_003239	TGFB3	-2.25	NA	8.25	0.0000	1.39	NA	PD&ICI
Crystallin. lambda 1	BX092299	CRYL1	-2.32	NA	-1.55	NA	-2.14	3.4057	
DNA (cytosine-5-)-methyltransferase 3 alpha	NM_175629	DNMT3A	-2.34	NA	-1.16	NA	-3.19	3.2435	PD
Complement component (3b/4b) receptor 1-like	XM_114735	CR1L	-2.69	NA	1.06	NA	-1.01	NA	
Insulin-like growth factor 1 (somatomedin C)	NM_000618	IGF1	-2.92	NA	1.93	NA	-1.12	NA	PD&ICI

Appendix A10: Venn diagram generated with genes induced by R5020 or estradiol and inhibited by PD, ICI, or both. (http://www.pangloss.com/seidel/Protocols/venn.cgi)

A Venn diagram of R5020 induced genes



ER independent / Erk1/2 independent (75)

Erk1/2 dependent (20)

ARHGAP5
CCNE2
CDC6
CHEK2
EGF
GRB2
IGFBP4
ITGA5
MKLN1
MTA1
MUC2L
MXI1
NBS1
NCOA3
NEO1
PLAUR
PPARGC1B
QSCN6
RAMP
STAT3

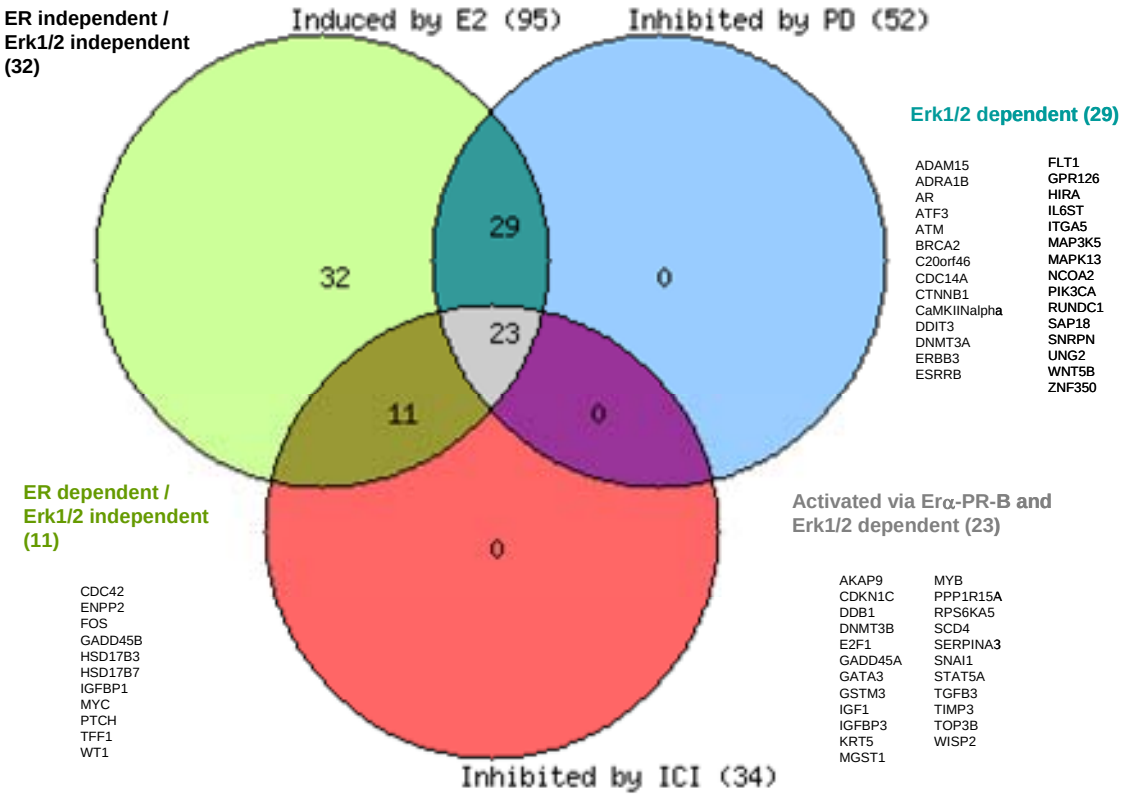
ER dependent / Erk1/2 independent (14)

AR
BIRC3
E2F3
GATA3
HSD17B3
HSPA9B
IGFBP3
LMNA
NFIB
NOTCH3
SERPINA3
SOS1
TP53BP2
UNG2

Activated via Erα-PR-B and Erk1/2 dependent (11)

CDKN1C
EGFR
EGLN1
ELL2
FANCA
H3F3B
IGFBP5
ORC6L
POLD4
PTCH2
RASL10B

B Venn diagram of E2 induced genes



ER independent / Erk1/2 independent (32)

Erk1/2 dependent (29)

ADAM15
ADRA1B
AR
ATF3
ATM
BRCA2
C20orf46
CDC14A
CTNNB1
CaMKIINalpha
DDIT3
DNMT3A
ERBB3
ESRRB
FLT1
GPR126
HIRA
IL6ST
ITGA5
MAP3K5
MAPK13
NCOA2
PIK3CA
RUNDC1
SAP18
SNRPN
UNG2
WNT5B
ZNF350

ER dependent / Erk1/2 independent (11)

CDC42
ENPP2
FOS
GADD45B
HSD17B3
HSD17B7
IGFBP1
MYC
PTCH
TFF1
WT1

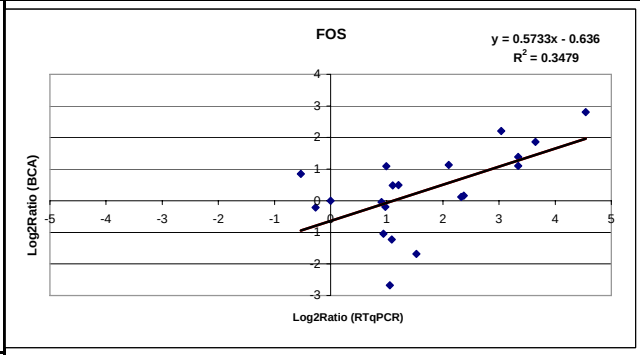
Activated via Erα-PR-B and Erk1/2 dependent (23)

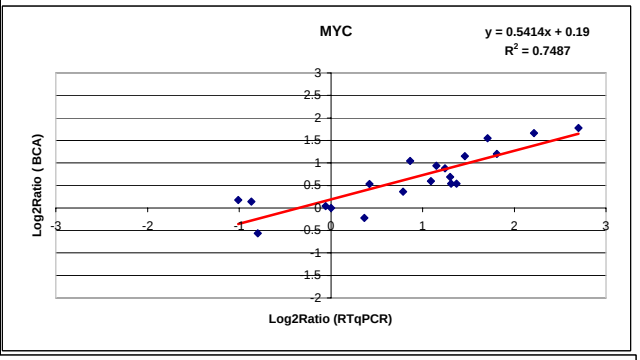
AKAP9
CDKN1C
DDB1
DNMT3B
E2F1
GADD45A
GATA3
GSTM3
IGF1
IGFBP3
KRT5
MGST1
MYB
PPP1R15A
RPS6KA5
SCD4
SERPINA3
SNAI1
STAT5A
TGFB3
TIMP3
TOP3B
WISP2

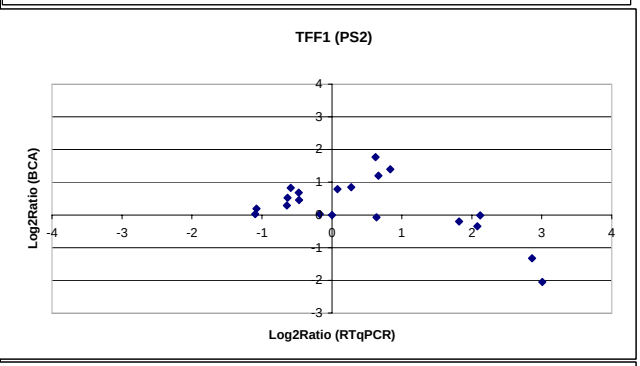
Appendix A11: Comparison Real Time qPCR and microarray values

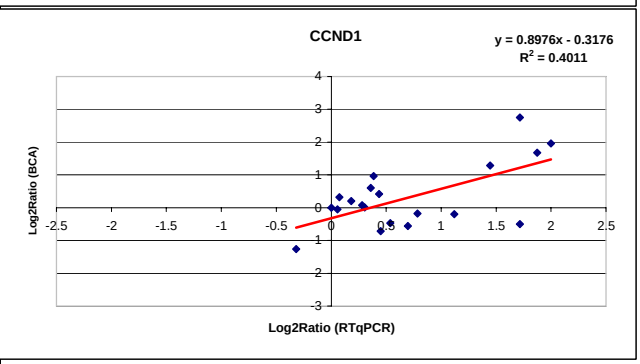
17/12/2007

Gene	Sample	Log_2Ratio (RTqPCR)	Log_2Ratio (Microarray BCA)	Graphical representation of the comparison between Real Time qPCR and Microarray obtained values
FOS	T0	0.00	0.00	
	R5020 30 min	3.34	1.10	
	R5020 60 min	3.34	1.38	
	R5020 2 hr	1.10	0.48	
	R5020 6 hr	0.99	1.09	
	R5020 12 hr	0.90	-0.04	
	R5020 24 hr	1.21	0.49	
	R5020 48 hr	2.10	1.14	
	E2 30 min	3.65	1.86	
	E2 60 min	4.54	2.80	
	E2 2 hr	2.37	0.16	
	E2 6 hr	1.09	-1.23	
	E2 12 hr	0.94	-1.05	
	E2 24 hr	1.06	-2.67	
	E2 48 hr	1.53	-1.68	
	T0+PD	-0.53	0.85	
	PD+R5020 6 hr	-0.27	-0.22	
	ICI+R5020 6 hr	0.97	-0.19	
	ICI+E2 60 min	3.04	2.21	
	PD+E2 60 min	2.33	0.12	
MYC	T0	0.00	0.00	
	R5020 30 min	1.30	0.69	
	R5020 60 min	1.81	1.20	
	R5020 2 hr	1.24	0.88	
	R5020 6 hr	-0.06	0.04	
	R5020 12 hr	-0.80	-0.56	
	R5020 24 hr	-1.01	0.18	
	R5020 48 hr	-0.87	0.14	
	E2 30 min	1.15	0.94	
	E2 60 min	2.21	1.67	
	E2 2 hr	1.71	1.55	
	E2 6 hr	1.46	1.15	
	E2 12 hr	0.87	1.05	
	E2 24 hr	1.37	0.54	
	E2 48 hr	1.09	0.59	
	T0+PD	0.37	-0.22	
	PD+R5020 6 hr	0.79	0.36	
	ICI+R5020 6 hr	0.42	0.53	
	ICI+E2 60 min	1.31	0.54	
	PD+E2 60 min	2.70	1.78	
TFF1	T0	1.00	0.00	
	R5020 30 min	-1.56	-0.64	
	R5020 60 min	-1.38	-0.47	
	R5020 2 hr	-1.39	-0.47	
	R5020 6 hr	-1.55	-0.64	
	R5020 12 hr	-2.14	-1.10	
	R5020 24 hr	-2.11	-1.08	
	R5020 48 hr	-1.50	-0.59	
	E2 30 min	-1.12	-0.17	
	E2 60 min	1.79	0.84	
	E2 2 hr	3.53	1.82	
	E2 6 hr	4.34	2.12	
	E2 12 hr	4.22	2.08	
	E2 24 hr	8.05	3.01	
	E2 48 hr	7.27	2.86	
	T0+PD	1.54	0.62	
	PD+R5020 6 hr	1.21	0.28	
	ICI+R5020 6 hr	1.06	0.08	
	ICI+E2 60 min	1.59	0.67	
	PD+E2 60 min	1.56	0.64	
CCND1	T0	0.00	0.00	
	R5020 30 min	0.06	-0.06	
	R5020 60 min	0.43	0.41	
	R5020 2 hr	1.45	1.28	
	R5020 6 hr	1.88	1.67	
	R5020 12 hr	0.07	0.32	
	R5020 24 hr	0.18	0.20	
	R5020 48 hr	0.36	0.60	
	E2 30 min	1.12	-0.19	
	E2 60 min	0.30	0.01	
	E2 2 hr	0.78	-0.18	
	E2 6 hr	0.70	-0.56	
	E2 12 hr	0.54	-0.47	
	E2 24 hr	0.45	-0.72	
	E2 48 hr	1.72	-0.51	
	T0+PD	-0.32	-1.26	
	PD+R5020 6 hr	1.72	2.75	
	ICI+R5020 6 hr	2.00	1.96	
	ICI+E2 60 min	0.28	0.07	
	PD+E2 60 min	0.39	0.96	



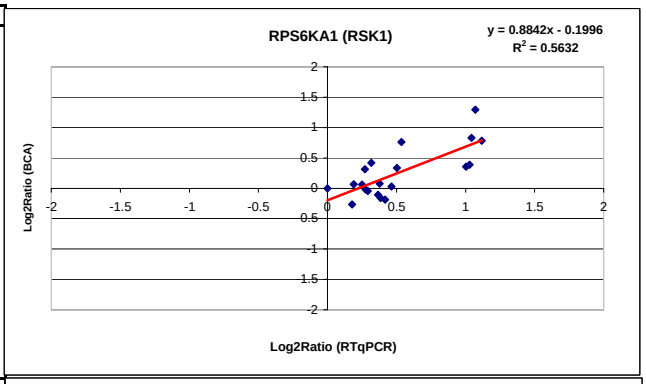


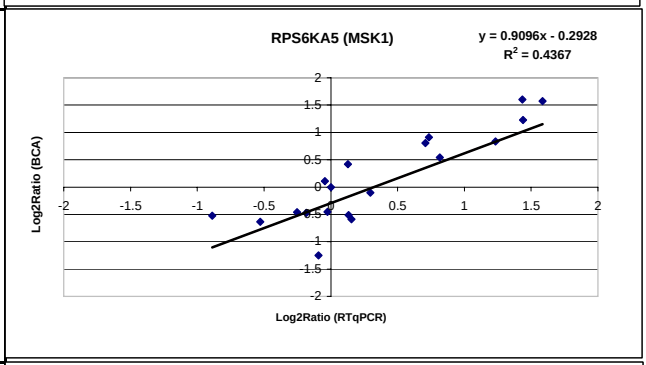


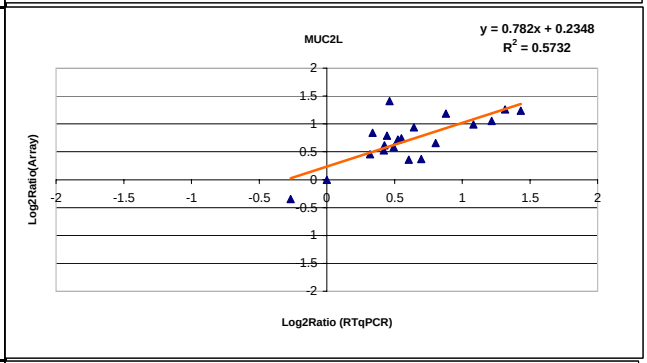


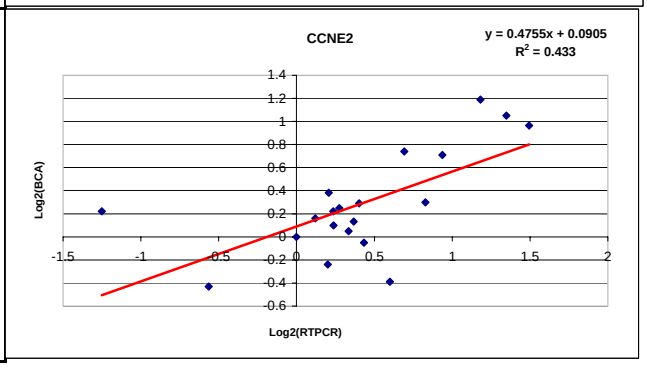
Appendix A11: Comparison Real Time qPCR and microarray values

Gene	Sample	Log_2Ratio (RTqPCR)	Log_2Ratio (Microarray BCA)
RPS6KA1 (p90/RSK1)	T0	0.00	0.00
	R5020 30 min	0.18	-0.27
	R5020 60 min	0.37	-0.11
	R5020 2 hr	0.46	0.03
	R5020 6 hr	1.04	0.83
	R5020 12 hr	0.54	0.76
	R5020 24 hr	0.29	-0.04
	R5020 48 hr	0.19	0.06
	E2 30 min	1.00	0.36
	E2 60 min	0.27	0.31
	E2 2 hr	0.50	0.34
	E2 6 hr	0.38	0.08
	E2 12 hr	0.28	-0.02
	E2 24 hr	0.32	0.42
	E2 48 hr	1.03	0.39
	T0+PD	0.25	0.06
	PD+R5020 6 hr	1.12	0.78
	ICI+R5020 6 hr	1.07	1.29
	ICI+E2 60 min	0.42	-0.19
	PD+E2 60 min	0.39	-0.16
RPS6KA5 (MSK1)	T0	0.00	0.00
	R5020 30 min	-0.25	-0.46
	R5020 60 min	-0.18	-0.47
	R5020 2 hr	0.82	0.54
	R5020 6 hr	1.44	1.22
	R5020 12 hr	0.73	0.91
	R5020 24 hr	0.71	0.80
	R5020 48 hr	1.23	0.83
	E2 30 min	0.15	-0.59
	E2 60 min	-0.09	-1.25
	E2 2 hr	0.13	0.42
	E2 6 hr	0.79	-2.12
	E2 12 hr	-0.89	-0.52
	E2 24 hr	-0.53	-0.64
	E2 48 hr	0.13	-0.51
	T0+PD	0.30	-0.10
	PD+R5020 6 hr	1.58	1.57
	ICI+R5020 6 hr	1.43	1.60
	ICI+E2 60 min	-0.03	-0.45
	PD+E2 60 min	-0.04	0.11
MUC2L	T0	0.00	0.00
	R5020 30 min	0.80	0.66
	R5020 60 min	1.08	0.99
	R5020 2 hr	1.43	1.24
	R5020 6 hr	1.32	1.26
	R5020 12 hr	0.64	0.94
	R5020 24 hr	0.55	0.74
	R5020 48 hr	0.70	0.37
	E2 30 min	0.49	0.58
	E2 60 min	0.32	0.45
	E2 2 hr	0.52	0.72
	E2 6 hr	0.60	0.36
	E2 12 hr	0.34	0.84
	E2 24 hr	0.46	1.41
	E2 48 hr	0.88	1.19
	T0+PD	-0.27	-0.35
	PD+R5020 6 hr	0.43	0.62
	ICI+R5020 6 hr	1.22	1.05
	ICI+E2 60 min	0.42	0.53
	PD+E2 60 min	0.44	0.79
CCNE2	T0	0.00	0.00
	R5020 30 min	0.37	0.13
	R5020 60 min	0.21	0.38
	R5020 2 hr	0.94	0.71
	R5020 6 hr	1.49	0.97
	R5020 12 hr	1.35	1.05
	R5020 24 hr	0.24	0.10
	R5020 48 hr	-0.56	-0.43
	E2 30 min	-1.25	0.22
	E2 60 min	0.12	0.16
	E2 2 hr	0.27	0.25
	E2 6 hr	0.40	0.29
	E2 12 hr	0.24	0.22
	E2 24 hr	1.18	1.19
	E2 48 hr	0.69	0.74
	T0+PD	0.60	-0.39
	PD+R5020 6 hr	0.33	0.05
	ICI+R5020 6 hr	0.83	0.30
	ICI+E2 60 min	0.43	-0.05
	PD+E2 60 min	0.20	-0.24









Appendix A12: Comparison between Real Time qPCR and micrioarrays

17/12/2007

PAM class	3	3	5	5	2	3	3	5	1	6	1	Unclass	3	3	3	3	4	3	3	3	5	3	6	5	2	2	1	1	1	1	1
Ratio (RTqPCR)	**T84**	**T85**	**T86**	**T87**	**T88**	**T89**	**T90**	**T94**	**T95**	**T96**	**T97**	**T98**	**T99**	**T100**	**T101**	**T102**	**T103**	**T104**	**T105**	**T106**	**T107**	**T108**	**T109**	**T110**	**T111**	**T112**	**T113**	**T114**	**T115**	**T116**	**T117**
SERPINB2	0.014			0.004		0.080	0.082	0.020	0.011	0.012	0.004	0.164	0.012	0.014	0.008	0.004	0.005	0.002	0.005	0.004	########	0.004	0.042	0.006	0.003	0.004	0.004	#######	0.002	0.028	0.014
CEBPD	0.315	0.650	0.451	0.368	0.368	0.193	0.807	0.366	0.652	0.710	0.664	0.435	0.366	1.186	1.049	2.983	0.825	0.421	0.524	2.231	0.359	1.158	2.415	0.472	0.769	0.832	0.391	0.412	0.524	0.872	1.751
TFF1	0.242	4.237		0.000	2.938	1.642	0.013		0.032	0.112	0.039	0.097	1.207	6.723	0.004	0.062	0.188	0.050	0.031	1.041		1.361	0.026	0.000	2.597	1.352	0.009	0.000	0.520	0.255	0.390
GATA3	0.367	0.764	0.091	0.049		0.495	0.519	0.006	1.248	0.089	1.068	0.093	1.017	0.983	1.644	1.326	0.345	0.330	0.638	1.204	0.047	0.552	0.174	0.021	1.592	1.247	0.636	0.342	0.829	0.871	0.368
SERPINA3	0.220	2.231	1.968	1.597	1.045	0.257	1.258	0.137	0.134	0.065	0.217	0.345	1.903	1.343	2.136	0.884	0.125	0.220	1.753	0.915	0.010	4.675	0.813	0.123	3.803	0.829	0.629	1.601	4.865	0.497	2.116
CAV1	0.197	0.107	0.035	0.021	0.034	0.078	0.115	0.101	0.149	2.350	0.025	0.079	0.133									0.351	1.044	0.044	0.065	0.062	0.234	0.169	0.164	0.153	0.266
APOD	6.78	2.45	0.36		2.47	3.02	13.82	0.89	3.97	52.10	6.42	22.50	3.67	16.92	30.13	5.58	107.74	29.27	2.89	5.40	2.41	29.48	33.78	2.07	22.41	0.40	13.07	2.42	35.26	9.84	23.10
DUSP1	2.750	2.708	4.494		2.109	2.164	4.031	1.130	1.742	7.731	0.890	1.265	1.197	6.858	7.061	7.321	1.981	2.304	2.158	7.755	0.818	4.058	12.317	1.464	2.142	5.842	1.546	6.512	1.886	4.390	19.181
TOP2A	0.014	0.005	0.009	0.007	0.013	0.013	0.012	0.022	0.014	0.002	0.043	0.023	0.026	0.052	0.004	0.145	0.027	0.021	0.019	0.136	0.033	0.007	0.002	0.043	0.013	0.017	0.025	0.009	0.012	0.018	0.004
ERBB2	0.823	0.105			0.110	0.177	0.225	0.104	0.232	0.137	1.673	0.313	1.460	0.777	0.559	2.524	0.463	3.250	0.597	1.314	0.823	0.718	0.959	0.232	0.810	0.348	0.435	0.908	0.396	0.690	0.760
SMARCE1	2.815	2.867	1.696	1.204	1.525	2.167	3.414	0.444	0.570	0.746	0.709	0.747	1.032	6.179	3.648	6.724	8.285	4.215	7.024	14.132	2.929	5.894	6.569	5.130	3.709	5.738	3.755	8.370	4.076	5.820	3.871
SMARCA4	1.514	1.710	2.361	1.470	0.940	1.098	0.662	2.112	2.033	1.481	7.876	2.312	5.081	3.777	3.320	5.973	6.166	2.792	4.970	4.745	4.722	3.423	4.091	8.379	3.218	5.638	3.576	3.335	2.470	4.607	2.261
DUSP6	1.221	0.771	0.225	0.157	0.412	0.598	0.351	0.118	0.205	2.015	0.173	0.393	0.742	1.516	1.761	1.565	0.672	1.855	0.626	0.929	0.476	1.487	4.938	0.412	0.361	3.294	1.125	0.385	2.166	0.843	1.927
KRT5	204.67	22.67	1322.97	734.46	23.45	55.07	198.40	373.01	27.72	420.97	20.94	110.26	244.91	126.95	861.27	105.21	184.56	42.19	20.29	14.84	1585.60	283.80	1449.10	67.66		35.08	226.75	115.25	209.68	81.37	281.38
S100A2	5.307	3.999	19.655	16.875	2.540	8.124	9.438	9.349	1.674	1.293	7.581	4.442	20.931	6.008	51.396	8.987	12.727	22.970	3.072	3.332	27.206	35.525	32.626	32.807	0.731	8.135	17.495	4.399	19.729	21.902	16.959
Log2Ratio (RTqPCR)	**T84**	**T85**	**T86**	**T87**	**T88**	**T89**	**T90**	**T94**	**T95**	**T96**	**T97**	**T98**	**T99**	**T100**	**T101**	**T102**	**T103**	**T104**	**T105**	**T106**	**T107**	**T108**	**T109**	**T110**	**T111**	**T112**	**T113**	**T114**	**T115**	**T116**	**T117**
SERPINB2	-6.152			-8.100		-3.650	-3.606	-5.661	-6.568	-6.349	-8.059	-2.608	-6.366	-6.164	-7.032	-8.029	-7.537	-9.377	-7.652	-7.817	21.366	-7.913	-4.568	-7.306	-8.343	-7.801	-8.117	23.690	-8.888	-5.138	-6.200
CEBPD	-1.666	-0.621	-1.148	-1.442	-1.443	-2.377	-0.310	-1.448	-0.618	-0.494	-0.590	-1.201	-1.449	0.246	0.069	1.577	-0.278	-1.249	-0.932	1.158	-1.477	0.212	1.272	-1.084	-0.380	-0.266	-1.355	-1.278	-0.933	-0.198	0.808
TFF1	-2.045	2.083		-12.155	1.555	0.715	-6.317		-4.974	-3.161	-4.683	-3.362	0.272	2.749	-7.965	-4.004	-2.410	-4.319	-5.010	0.058		0.444	-5.261	-12.447	1.377	0.435	-6.835	-11.671	-0.944	-1.974	-1.358
GATA3	-1.446	-0.389	-3.464	-4.366		-1.014	-0.946	-7.431	0.320	-3.486	0.095	-3.434	0.025	-0.025	0.718	0.407	-1.536	-1.599	-0.649	0.268	-4.412	-0.857	-2.525	-5.603	0.671	0.319	-0.653	-1.547	-0.271	-0.200	-1.443
SERPINA3	-2.185	1.158	0.977	0.676	0.064	-1.959	0.331	-2.870	-2.896	-3.934	-2.207	-1.536	0.929	0.425	1.095	-0.178	-3.005	-2.183	0.810	-0.129	-6.638	2.225	-0.299	-3.025	1.927	-0.271	-0.669	0.679	2.282	-1.008	1.082
CAV1	-2.340	-3.229	-4.851	-5.576	-4.888	-3.681	-3.121	-3.311	-2.747	1.233	-5.344	-3.664	-2.909									-1.509	0.062	-4.498	-3.935	-4.016	-2.098	-2.562	-2.612	-2.713	-1.911
APOD	2.761	1.293	-1.474		1.306	1.597	3.789	-0.160	1.988	5.703	2.683	4.492	1.878	4.081	4.913	2.480	6.751	4.871	1.531	2.432	1.268	4.882	5.078	1.052	4.486	-1.337	3.708	1.276	5.140	3.299	4.530
DUSP1	1.459	1.437	2.168		1.077	1.114	2.011	0.177	0.801	2.951	-0.169	0.339	0.260	2.778	2.820	2.872	0.986	1.204	1.110	2.955	-0.290	2.021	3.623	0.550	1.099	2.546	0.629	2.703	0.915	2.134	4.262
TOP2A	-6.153	-7.786	-6.826	-7.097	-6.255	-6.259	-6.401	-5.522	-6.157	-8.976	-4.537	-5.421	-5.246	-4.264	-7.868	-2.789	-5.192	-5.552	-5.727	-2.874	-4.922	-7.236	-8.948	-4.546	-6.249	-5.873	-5.325	-6.761	-6.356	-5.806	-7.818
ERBB2	-0.281	-3.249			-3.190	-2.502	-2.154	-3.261	-2.110	-2.873	0.742	-1.677	0.546	-0.364	-0.840	1.336	-1.110	1.700	-0.744	0.394	-0.281	-0.477	-0.060	-2.111	-0.303	-1.522	-1.200	-0.139	-1.335	-0.535	-0.397
SMARCE1	1.493	1.519	0.762	0.267	0.609	1.115	1.772	-1.171	-0.811	-0.424	-0.495	-0.421	0.046	2.627	1.867	2.749	3.051	2.076	2.812	3.821	1.550	2.559	2.716	2.359	1.891	2.521	1.909	3.065	2.027	2.541	1.953
SMARCA4	0.598	0.774	1.239	0.555	-0.089	0.135	-0.595	1.079	1.024	0.566	2.977	1.209	2.345	1.917	1.731	2.578	2.624	1.481	2.313	2.247	2.239	1.775	2.032	3.067	1.686	2.495	1.838	1.738	1.304	2.204	1.177
DUSP6	0.288	-0.376	-2.155	-2.669	-1.279	-0.742	-1.512	-3.081	-2.284	1.011	-2.535	-1.348	-0.430	0.600	0.816	0.646	-0.573	0.891	-0.677	-0.107	-1.071	0.572	2.304	-1.279	-1.469	1.720	0.171	-1.378	1.115	-0.247	0.946
KRT5	7.677	4.503	10.370	9.521	4.552	5.783	7.632	8.543	4.793	8.718	4.388	6.785	7.936	6.988	9.750	6.717	7.528	5.399	4.343	3.891	10.631	8.149	10.501	6.080		5.133	7.825	6.849	7.712	6.346	8.136
S100A2	2.408	1.999	4.297	4.077	1.345	3.022	3.239	3.225	0.743	0.371	2.922	2.151	4.388	2.587	5.684	3.168	3.670	4.522	1.619	1.737	4.766	5.151	5.028	5.036	-0.453	3.024	4.129	2.137	4.302	4.453	4.084
Log2Ratio (Microarray)	**T84**	**T85**	**T86**	**T87**	**T88**	**T89**	**T90**	**T94**	**T95**	**T96**	**T97**	**T98**	**T99**	**T100**	**T101**	**T102**	**T103**	**T104**	**T105**	**T106**	**T107**	**T108**	**T109**	**T110**	**T111**	**T112**	**T113**	**T114**	**T115**	**T116**	**T117**
SERPINB2	-0.552	-0.463	-0.395	-0.566	-0.439	-0.530	-0.610	-0.484	-0.050	-0.034	0.103	-0.772	-0.442	-0.592	-0.462	-0.488	-0.503	-0.503	-0.512	-0.515	-0.366	-0.308	-0.284	-0.322	-0.347	-0.165	-0.161	-0.256	-0.263	-0.261	-0.101
CEBPD	0.073	0.379	0.654	0.328	0.095	-0.036	0.245	0.154	-0.053	0.283	0.594	0.112	0.233	0.137	0.372	1.239	0.088	0.136	0.216	1.068	0.364	0.112	1.043	0.071	0.310	0.272	-0.216	-0.134	0.027	0.055	0.264
TFF1	-0.228	0.351	-0.198	-0.179	-0.267	-0.001	-0.178	-0.189	0.318	-0.200	-0.178	-0.036	0.547	0.167	-0.092	-0.199	-0.285	-0.391	0.132	-0.042	-0.294	0.249	0.027	-0.337	-0.060	-0.027	-0.246	-0.326	0.188	-0.054	0.250
GATA3	0.370	1.007	-0.748	-0.983	3.376	1.716	1.870	-2.697	1.885	-0.220	0.452	1.670	2.193	1.817	3.374	2.746	0.456	1.084	2.471	2.501	-0.263	1.422	-0.445	-2.255	2.109	1.830	1.762	1.199	1.937	1.620	0.686
SERPINA3	-0.848	0.838	0.712	1.658	1.605	0.069	2.561	-0.608	0.144	-0.815	-0.542	0.139	4.336	0.370	3.180	0.139	-1.969	-1.863	0.626	0.301	-2.534	2.437	-1.074	-2.747	2.071	0.630	0.251	0.449	1.833	-1.530	0.571
CAV1	-0.163	-0.046	0.032	-0.129	-0.126	-0.103	-0.146	-0.116	-0.024	0.447	-0.046	-0.072	-0.662	-0.483	-0.205	-0.223	0.001	-0.112	-0.134	-0.182	-0.179	0.005	0.516	0.066	0.275	-0.083	-0.162	-0.245	0.156	0.003	0.286
APOD	1.834	0.697	-0.048	0.286	2.774	1.194	3.580	1.213	3.886	5.490	0.306	4.120	3.243	3.607	6.600	3.180	3.872	5.882	2.370	2.937	1.769	3.843	3.112	1.053	3.736	-0.226	3.973	1.827	5.708	3.068	2.832
DUSP1	-0.237	0.434	0.325	0.263	-0.062	0.022	-0.031	-0.218	-0.056	1.665	0.210	0.062	0.201	-0.015	0.678	0.530	0.006	-0.240	-0.241	0.242	-0.404	-0.010	0.591	-0.433	0.208	0.456	-0.128	0.255	0.033	0.174	0.546
TOP2A	-3.421	-3.396	-3.272	-4.361	-5.524	-4.056	-6.369	-7.093	-7.254	-4.340	-3.548	-3.910	-5.817	-3.820	-7.110	-3.724	-3.743	-5.803	-4.380	-2.904	-2.305	-4.636	-3.920	-5.180	-2.942	-3.299	-4.001	-5.503	-5.386	-5.321	-4.779
ERBB2	0.053	-0.121	0.205	-0.265	-0.043	0.225	0.361	-0.019	-0.096	0.080	-0.033	0.130	-0.083	0.018	0.047	-0.110	-0.169	0.443	-0.146	0.187	0.067	0.058	0.083	0.147	0.041	0.109	-0.171	0.177	0.114	0.091	0.371
SMARCE1	-0.693	-0.664	-1.233	-1.318	-1.376	-0.285	-0.471	-1.885	-4.917	-0.114	-1.214	0.000	-0.291	-0.458	-1.238	-0.774	0.033	-0.851	0.205	1.286	-0.259	-0.227	-0.740	-0.199	-0.292	-0.256	-0.228	0.342	-0.247	-0.352	-0.364
SMARCA4	0.753	-0.054	0.226	0.040	1.366	0.607	-0.460	0.152	1.963	-1.199	-0.160	-0.048	-0.249	0.226	0.348	1.191	0.027	-0.233	0.518	0.333	0.203	-0.991	-0.992	0.764	0.352	0.116	-0.641	-1.565	-0.082	-0.130	-1.161
DUSP6	0.678	1.407	1.576	0.589	-0.131	-0.182	0.648	0.477	-1.775	0.390	2.458	0.529	0.102	0.049	0.003	1.880	1.578	1.025	1.668	0.658	0.816	0.003	1.254	0.127	0.123	1.040	0.105	1.422	0.500	1.328	0.515
KRT5	1.540	-0.100	3.343	3.496	0.253	0.442	2.147	5.613	0.470	2.693	0.346	2.553	3.499	0.517	2.903	0.912	2.003	0.652	-0.139	-0.019	2.575	2.146	3.713	1.331	-0.579	0.032	1.647	1.128	2.080	0.720	1.349
S100A2	0.339	0.278	0.366	0.679	0.352	0.263	0.813	0.759	0.139	0.152	0.296	0.313	1.237	0.247	1.015	0.211	0.348	0.204	-0.005	0.226	0.303	0.354	0.179	0.370	0.166	0.143	0.514	0.221	1.333	1.216	0.849
CORRELATION COEFFICIENT R	0.735	0.633	0.737	0.638	0.573	0.573	0.642	0.761	0.485	0.813	0.217	0.721	0.661	0.546	0.633	0.401	0.731	0.552	0.439	0.455	0.420	0.664	0.784	0.439	0.669	0.468	0.544	0.234	0.670	0.607	0.693
R^2	0.540	0.401	0.543	0.407	0.329	0.328	0.412	0.579	0.236	0.661	0.047	0.520	0.437	0.298	0.400	0.161	0.534	0.305	0.192	0.207	0.177	0.441	0.614	0.193	0.448	0.219	0.296	0.055	0.449	0.369	0.480

Publications

It is expected that after the conclusion of the follow-up time of the breast cancer patients, prognosis studies carried out with the help of a statistician, this work will be published.

Publications from collaborative studies with the group of Albert Jordán and Miguel Beato (CRG, Barcelona) are also expected to be published soon.

During these four years period I participated in several publications which I attach.

- Galy B, Ferring D, Minana B, Bell O, Janser HG, Muckenthaler M, Schümann K, Hentze MW. 2005 Altered body iron distribution and microcytosis in mice deficient in iron regulatory protein 2 (IRP2). 106(7):2580-9.
- Muckenthaler MU, Rodrigues P, Macedo MG, Minana B, Brennan K, Cardoso EM, Hentze MW, de Sousa M. 2004 Molecular analysis of iron overload in beta2-microglobulin-deficient mice. Blood Cells Mol Dis. 33(2):125-31.

My contribution on these two articles was in the design of the experiment, microarray platform establishment; carry out microarray experiments and in the analysis of the microarray data.

- Mengual L, Burset M, Ars E, Ribal MJ, Lozano JJ, Minana B, Sumoy L, Alcaraz A. 2006 Partially degraded RNA from bladder washing is a suitable sample for studying gene expression profiles in bladder cancer. Eur Urol. 50(6):1347-55

I contributed in the design of the experiment, collected all relevant publications, carried out RNA amplification, sample labeling, quality control, microarray hybridization and analysis.

Mengual L, Burset M, Ars E, Ribal MJ, Lozano JJ, Minana B, Sumoy L, Alcaraz A.
Partially degraded RNA from bladder washing is a suitable sample for studying gene expression profiles in bladder cancer.
Eur Urol. 2006 Dec;50(6):1347-55; discussion 1355-6. Epub 2006 Jun 15.

Galy B, Ferring D, Minana B, Bell O, Janser HG, Muckenthaler M, Schümann K, Hentze MW.
Altered body iron distribution and microcytosis in mice deficient in iron regulatory protein 2 (IRP2).
Blood. 2005 Oct 1;106(7):2580-9. Epub 2005 Jun 14.

Muckenthaler MU, Rodrigues P, Macedo MG, Minana B, Brennan K, Cardoso EM, Hentze MW, de Sousa M.
Molecular analysis of iron overload in beta2-microglobulin-deficient mice.
Blood Cells Mol Dis. 2004 Sep-Oct;33(2):125-31.