Università degli Studi di Padova

Facoltà di Scienze MM.FF.NN.
Dipartimento di Biologia

SCUOLA DI DOTTORATO DI RICERCA IN BIOCHIMICA E BIOTECNOLOGIE

INDIRIZZO IN BIOTECNOLOGIE

CICLO XXI

IDENTIFICATION OF DRUG-RESISTANCE PREDICTIVE GENES IN BREAST CANCER NEOADJUVANT CHEMOTHERAPY

Direttore della Scuola
Ch.mo Prof. Giuseppe Zanotti

Supervisore
Ch.mo Prof. Giorgio Valle

Dottorando
Lorenza Mittempergher

A.A. 2008/2009

CONTENTS

SUMMARY

Breast cancer is a heterogeneous disease where markers for therapy response remain poorly defined. Since the effectiveness of treatment differs between individual patients, during the last years much effort has being invested in the identification of new markers, to estimate patients's outcome (prognostic markers) and to indicate which treatment is most effective for an individual patient (predictive markers).

The implementation of predictive factors in clinical setting is a big challenge of the cancer research and it will provide the opportunity to guide treatment decisions. Only patients that are likely to benefit from a specific treatment will receive this specific treatment. An individualized therapy will avoid the administration of ineffective chemotherapy that increases mortality and decreases quality of life in cancer patients.

For many years research has focused on the identification of single markers predicting tumour response to chemotherapy. However it is unlikely that the chemotherapy resistance/responsiveness in breast cancer is the result of one or limited number of genes, because of the complexity of pathways involved in tumour response to chemotherapy and the heterogeneity of the individual tumours. The microarray technology made possible to study gene expression profiling of breast cancer on a global scale. It was successfully applied on the identification of breast cancer subgroups and *in* the determination of profiles predicting patient's prognosis. More recently microarray analysis of gene expression has been used as a possible approach for predicting response to chemotherapy. With the introduction of preoperative chemotherapy (neoadjuvant chemotherapy) it has become possible to directly evaluate the sensitivity of breast cancer to chemotherapy by the clinical/pathological response of the patient to the treatment. The main goal of this thesis was to identify predictive genes of response to a specific neoadjuvant chemotherapy regimen based on paclitaxel and anthracyclines (doxorubicin and epirubicin) drugs in breast cancer patients.

From 41 pre-treatment breast tumours biopsies good quality RNA was obtained and gene expression profiling was performed. Gene expression patterns of 37 patients were analyzed using Operon v2.0 70mer oligos collection at CRIBI Biotech centre and 4 patients were profiled with Operon v3.0 70mer oligos collection at Netherlands Cancer Institute. Clinical responses of 34 (out of 41) patients were recorded after administration of the neoadjuvant chemotherapy. Complete Responses (CR) to the treatment were observed in 3 patients, Partial Responses in 18 (PR) patients, No Change of the tumour mass (NC) in 11 patients and Progressive Disease (PD) in 2 patients.

First of all, a correlation analysis between the ImmunoHistoChemical data of six prognostic markers (ER, PR, Erb-B2, Bcl-2, Ki-67, p53) and the gene expression data was carried out. The results showed a significant correlation for ER, PR and Bcl-2 markers. Moreover Bcl-2 status measured by

ImmunoHistoChemistry (IHC) was significantly associated with the clinical response to neoadjuvant chemotherapy.

The molecular subtypes of 37 breast tumours analyzed with Operon v2.0 were identified using the "intrinsic gene signature" of Perou and colleagues. Most part of the patients were luminal-like subtype (28 of 37), 7 patients showed an erb-B2+ molecular subtype and 2 patients belonged to the basal-like group. Since it was reported that breast cancer molecular subtypes respond differently to neoadjuvant chemotherapy, I also checked how the clinical response to the treatment were associated to the molecular subtypes. From the analysis emerged that the luminal-like and erb-B2+ molecular subtypes were enriched of PR patients.

A hierarchical cluster analysis on the pre-treatment tumours (analyzed with Operon v2.0 and with clinical response available) was performed in order to evaluate how the patients would have been separated on the basis of their gene expression profile, using an unsupervised approach. As expected, no clear separation between Responders (PR + CR) and Non Responders (NC + PD) was found. The results did not change if we included in the responder group only the PR patients. We hypothesized that the predictive genes of resistance/sensitivity to the chemotherapy are a subtle set. The high number of differentially expressed genes would have masked the "real" predictive gene set, leading to a clustering of the patients based on biological parameters different from the clinical response. In addition the small size of the dataset was a limiting factor in the analysis.

In light of this result we opted for a supervised approach that consisted in dividing the tumours into Responders and Non Responders and searching for the genes (the drug-resistance predictive genes) that could correctly distinguish the two classes of response. I considered two datasets of patients, the dataset I including PR patients against not responders patients (NC + PD) and the dataset II with responders patients (PR and CR) against not responders patients (NC + PD).

The first approach, based on the software PAM (Prediction Analysis of Microarray), did not give a good prediction performance on both dataset of patients, misclassifying approximately 36% of patients. Therefore, a more effective analysis in terms of classification accuracy was requested. A gene selection process based on the Support Vector Machines (SVMs) was considered a good choice in light of the characteristics of the study: low number of patients (examples) and high number of genes (or features). SVMs are a supervised learning algorithm that work well at high dimensionality, overcoming the risk of overfitting due to a number of features much larger than the numbers of examples. A specific recursively feature selection procedure based on SVMs (R-SVM) was used to select the set of genes with the lowest error of classification on the dataset of patients. Because of the small sample size, it was not possible to have a training set and a test set completely separated, so a Leave-One-Out Cross Validation (LOO-CV) procedure was used to assess the performance of the feature selection process. The analysis identified a set of 54 genes able to classify the 28 patients of the dataset I with an accuracy of 85% (4 patients misclassified over 28) and a set of 14 genes able to classify the 30 patients of the dataset II with an accuracy of 76% (7 patients misclassified over 30).

The lower accuracy obtained on the dataset II was attributed to the introduction of the cCR patients in the group of Responders. The cCR patients were probably too much dissimilar in terms of clinical response in respect to the PR patients, thus reducing the homogeneity of the group of Responders. For this reason I focused the following analysis only on the dataset I.

The accuracy of 85% obtained for the dataset I was an encouraging result although the small size of the dataset.

The biological function and cellular localization of the 54 genes was examined by using GoMiner, a web tool to find associations of Gene Ontology categories within a specific group of genes. As emerged from the analysis, there were several functional categories related to the tumourigenesis processes ("cell adhesion", "insulin receptor signaling pathway", "cell proliferation", "regulation of cell proliferation"). Some categories were more closely related to cellular processes and compartments target of the chemotherapy agents used in this study ("cell cycle", "cell cycle arrest", "nucleus") and to responsiveness to the treatment ("response to hypoxia").

A literature research focused on each gene of the predictive signature showed that some of these genes (MYC, NUF2, SPC25; KFL5, CDKN1b, ITGA6, POSTN) are 'biologically plausible', since they have some connections with the drug resitance phenomenon investigated in this study. Others of the 54 genes are related to breast cancer progression and metastasis (CXCL9, CEBPD, IRS2, TCF8, ADAMTS5, PPARGC1A), but their direct involvement in drug resistance to paclitaxel/anthracycline neoadjuvant chemotherapy did not emerged.

At this point of my analysis, I tried to find out how to use the 54 genes signature as a predictive tool of responsiveness to paclitaxel/anthracyclines chemotherapy treatment. To achieve this objective, a SVM model was trained on the basis of the 54 genes to classify a new patients as putative partial responder or not responder. However, the SVM output is a value not so easily usable in statistics prediction problems. Therefore using a sigmoid function, we translated the SVM outputs into probability values that offered a more direct evaluation of the response class of the patient. In practice we transformed the SVM scores in a value, ranging from 0 to 1, that expresses the probability to belong to the positive class of response (PR patients). Using the trained SVM model on a new, not-yet classified patient, it is possible to map his SVM score on the sigmoid function and to obtain a corresponding probability value to belong to the positive class of response.

The results reported in this thesis look promising but have to be considered as preliminary, since they were obtained from a study investigating only a small number of patients and need to be validated in a completely independent test set of patients. Thus a validated gene expression signature may improve our understanding of neoadjuvant chemotherapy response mechanisms and in the future may lead to more individual, patient-tailored therapy decisions.

RIASSUNTO

Il tumore al seno è una patologia clinicamente eterogenea e *marker* biologici in grado di predirne in modo affidabile evoluzione e soprattutto sensibilità ai trattamenti farmacologici rimangono poco definiti. Negli ultimi anni la ricerca ha cercato così di identificare nuovi *marker* predittivi di risposta, per consentire trattamenti più efficace per ogni singola paziente. Riuscire ad implementare i nuovi fattori predittivi nella pratica clinica rappresenta un importante obiettivo nella ricerca sul tumore al seno. Si potranno così evitare a priori trattamenti inefficaci, che inciderebbero solo negativamente sulla qualità di vita delle pazienti.

Per molti anni si è parlato di *marker* singoli di risposta, ma, alla luce della complessità dei *pathway* cellulari coinvolti nella risposta del tumore alla chemioterapia ed all'eterogeneità tra i singoli tumori, è improbabile che la risposta o la resistenza ad un trattamento sia determinata dall'azione di un numero limitato di geni.

La tecnologia dei *microarray* ha reso così possibile un'analisi su larga scala dei profili di espressione genica dei tumori al seno ed è stata uno strumento efficace per identificarne sottogruppi molecolari e profili di espressione con valore prognostico. Più recentemente i *microarray* sono stati anche applicati alla ricerca di geni predittivi di risposta alla chemioterapia.

Con l'introduzione della chemioterapia neoadiuvante, ossia somministrata prima dell'intervento chirurgico, è divenuto possibile valutare direttamente la sensibilità del tumore al trattamento chemioterapico attraverso la risposta clinica e patologica della paziente.

L'obiettivo principale di questa tesi è stato infatti quello di identificare un *set* di geni predittivo della risposta ad un particolare trattamento chemioterapico neoadiuvante basato su taxani (paclitaxel) e antracicline (adriamicina o epirubicina).

Sono stati analizzati mediante *microarray* di oligonucleotidi 41 biopsie di tumore al seno prima della somministrazione della chemioterapia neoadiuvante. Delle 41 biopsie raccolte, 37 sono state analizzate con la piattaforma di oligonucleotidi Operon v2.0 presso il CRIBI e 4 sono state analizzate presso il *Netherlands Cancer Institute* con la piattaforma Operon v3.0. Al termine del trattamento è stato rese noto per 37 pazienti (su 41) l'esito della chemioterapia: 3 pazienti hanno mostrato una risposta clinica completa (cCR), 18 una risposta parziale al trattamento (PR), 13 pazienti non hanno risposto al trattamento, in 11 casi non si è avuto nessun cambiamento nella grandezza della massa tumorale (NC) ed in 2 casi un aumento di quest'ultima (PD).

La prima analisi condotta è stata quella volta a verificare la correlazione tra i dati di immunoistochimica (IHC) ottenuti per i 6 marker prognostici ER, PR, Erb-B2, Bcl-2, Ki-67 e p53 ed i livelli di espressione dei rispettivi geni misurati con i *microarray*. Una significativa correlazione è stata trovata per ER, PR e Bcl-2. Il livello di Bcl-2 ottenuto dall'analisi IHC si è rivelato inoltre significativamente associato con la risposta alla chemioterapia neoadiuvante.

Successivamente sono stati identificati i sottotipi molecolari dei 37 tumori analizzati con la piattaforma Operon v2.0 utilizzando l'*intrinsic gene set* individuato da Perou e colleghi. La maggior parte dei pazienti apparteneva al sottotipo luminale (28 su 37), 7 a quello erb-B2+ e 2 a quello basale. Poiché è stato riportato in letteratura che i sottotipi molecolari di tumore al seno rispondono in modo differente alla chemioterapia neoadiuvante, ho valutato come fossero distribuiti quelli da me identificati rispetto alla risposta clinica al trattamento, se disponibile. Dall'analisi è emerso che i sottogruppi luminale e erb-B2+ erano arricchiti di pazienti PR.

E' stata quindi eseguita una *cluster analysis* gerarchica dei 30 profili di espressione genica (ottenuti con Operon v2.0) delle pazienti di cui era disponibile la risposta alla chemioterapia, per valutare come si sarebbero separate sulla base dell'intero profilo di espressione con un approccio *unsupervised* (senza cioè dare a priori l'informazione sul tipo di risposta clinica). Le pazienti non si sono separati in sensibili (cCR + PR) e resistenti (NC + PD) al trattamento. Questo risultato ha confermato l'ipotesi che il *set* di geni predittivi fosse ristretto e che probabilmente venisse mascherato dal grande numero di geni differenzialmente espressi dal tumore. Inoltre il numero limitato di paziente è stato un fattore limitante all'analisi.

Sono passata quindi ad un approccio di tipo *supervised* cercando quei geni in grado di distinguere tumori sensibili e tumori resistenti al trattamento, cioè i geni predittivi della farmacoresistenza. Ho considerato due *dataset* di pazienti, il dataset I che includeva pazienti PR *vs* pazienti resistenti (NC e PD) e il dataset II che considerava anche i pazienti cCR nel gruppo di tumori sensibili al trattamento.

Il programma PAM (*Prediction Analysis of Microarray*) ha individuato *set* di geni predittivi con una bassa *performance* di classificazione dei pazienti in entrambi i *dataset* (il 36% dei pazienti veniva classificato in modo sbagliato). Si è reso quindi necessario un nuovo metodo di analisi, più efficace in termini di *accuracy* di classificazione. Una selezione dei geni significativi basata sulle *Support Vector Machines* (SVM) è stata considerata una scelta appropriata alla luce delle caratteristiche dello studio: basso numero di pazienti (o esempi) e alto numero di geni (o *features*). Le SVM infatti sono degli algoritmi di apprendimento supervisionati che lavorano bene in questi casi abbassando il rischio di *overfitting*, dovuto al numero troppo elevato di *features* rispetto agli esempi da classificare. In particolare è stato utilizzato l'algoritmo di *feature selection* R-SVM (*Recursive Support Vector Machine*) per selezionare quel *set* di geni con il più basso errore di classificazione sul *dataset* di pazienti (I e II). Per validare la *performance* di classificazione dei *set* di geni selezionati è stata usata una *Leave One Out Cross Validation* non essendo possibile, a causa del numero ridotto di pazienti, suddividere i *dataset* in un *training* and in un *test set* indipendenti. L'analisi R-SVM ha identificato un *set* di 54 geni in grado di classificare i 28 pazienti del *dataset* con un'accuratezza pari all'85% (4 pazienti sbagliati su 28) e un *set* di 14 geni in grado di classificare le 30 pazienti del *dataset* II con un'accuratezza del 76% (7 pazienti sbagliati su 30). L'abbassamento del grado di *accuracy* nel *dataset* II è stato attribuito al fatto di aver incluso nel gruppo dei pazienti sensibili al trattamento anche i pazienti cCR; in realtà essi avrebbero costituito una classe troppo diversa dai pazienti PR tale da non poter essere inclusa nello stesso gruppo di questi ultimi. Alla luce di quanto detto ho considerato solo il dataset I nelle analisi successive.

L'analisi di *Gene Ontology* sui 54 geni identificati nel *dataset* I ha rivelato che alcuni di questi geni sono annotati a livello di processi biologici caratteristici della tumorigenesi in generale ("adesione cellulare", "vie di segnalazione dell'insulina", "proliferazione cellulare", "regolazione della proliferazione cellulare"). Alcune categorie funzionali sono invece più legate a processi e compartimenti cellulari *target* dei farmaci utilizzati in questo studio ("ciclo cellulare", "arresto del ciclo cellulare", "nucleo") ed alla risposta al trattamento ("risposta all'ipossia"). Da una ricerca in letteratura mirata a ciascuno dei 54 geni della lista è emerso che alcuni di essi (MYC, NUF2, SPC25; KFL5, CDKN1b, ITGA6, POSTN) sono implicati nel fenomeno di resistenza a paclitaxel ed antracicline. Altri (CXCL9, CEBPD, IRS2, TCF8, ADAMTS5, PPARGC1A) dimostrano di avere un ruolo in processi collegati a progressione tumorale ed a metastasi ma non hanno un coinvolgimento diretto con la farmacoresistenza oggetto dello studio.

A questo punto del lavoro è stato naturale chiedersi come utilizzare il modello SVM allenato usando i 54 geni per predire la risposta alla chemioterapia (con paclitaxel ed antracicline) di un nuovo paziente, non ancora classificato come sensibile o resistente al trattamento. Dal momento che l'*output* di una SVM è una misura di distanza dall'iperpiano che separa i pazienti positivi (sensibili al trattamento) da quelli negativi (resistenti al trattamento) a cui non è associato un significato statistico, si è pensato di trasformare questo valore in una misura di probabilità di appartenenza alla classe positiva di risposta. Per fare questo è stato utilizzato un modello parametrico definito da una sigmoide che ha consentito di trasformare gli *output* SVM dei 28 pazienti in corrispondenti valori di probabilità.

I risultati ottenuti in questa tesi si sono rivelati interessanti anche se vanno considerati preliminari alla luce del numero limitato di pazienti. Si renderà necessaria pertanto una validazione su un gruppo indipendente di pazienti e, in caso di conferma dei risultati, questo lavoro potrà contribuire alla scelta di trattamenti più efficaci per il tumore al seno.

1 INTRODUCTION

1.1 STRUCTURAL ORGANIZATION OF THE MAMMARY GLAND

The mammary gland is an exocrine gland comprised of parenchymal structures that invade the mammary fat pad [1]. In simple terms the breast comprises a branching system of ducts leading down from the nipple ending in glands (acini aggregated into lobules) which have the potential to secrete milk. Approximately 12 large ducts emerge from the breast at the nipple as lactiferous ducts (fig. 1.1).

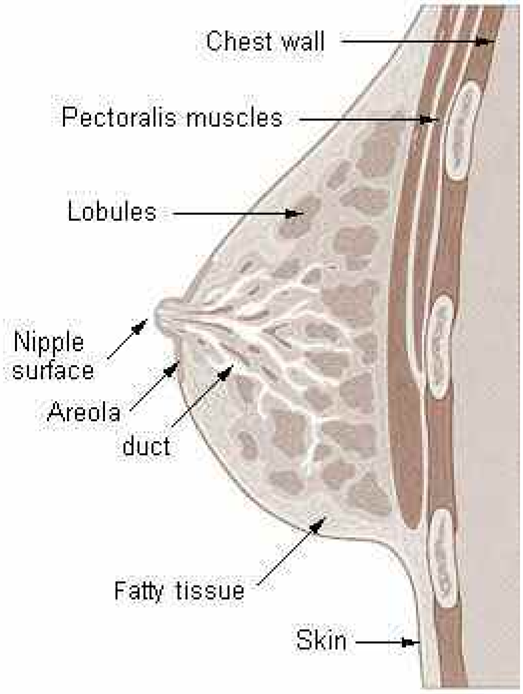


Figure 1.1: Cross section of the breast of a human female: lobules are organized in acini (~20-30) (http://training.seer.cancer.gov/ss_module01_breast/unit02_sec01_anatomy.html)

It is organized into a tree-like structure composed of hollow branches. These have an inner layer of luminal epithelial cells that face the lumen and are surrounded by an outer layer of myoepithelial cells that secretes the basal lamina, separating the mammary parenchyma from the stroma [1]. Within the mammary arbour, the ductal cells are those that line the ducts of the mammary gland. Lobular cells form secretory acinar structures (acini) at the end of each branch and, upon pregnancy and lactation, become alveolar cells that produce milk proteins [2] (fig. 1.2). Each lobule (fig. 1.3) consists of 20 – 30 acini which drain into a terminal duct called Terminal Duct Lobular Unit (TDLU). Like ducts, each acinus is double layered with the epithelial layer lining the lumen and ensuring the synthesis and secretion of milk, and the myoepithelial layer lining the basal membrane. This cell type appears to be useful for diagnostic purposes: invasive carcinoma is devoid of such cells [3]. The specialized connective tissue surrounding the acini is called palleal tissue.

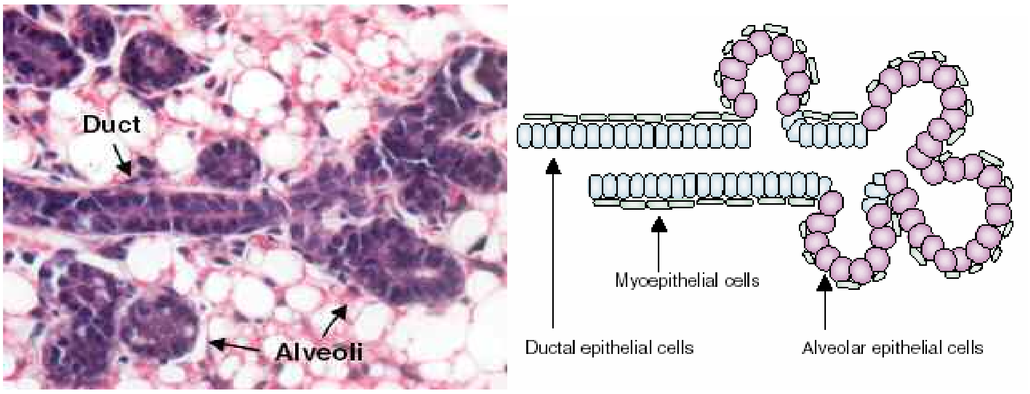


Figure 1.2: On the left section, stained with hematoxylin and eosin, of a midpregnant mammary gland from C57BL/6 mice indicating the locations of the ductal and alveolar cells. On the right schematic view of the ductal and alveolar cells during midpregnancy. The ducts are surrounded by a basal layer of overlapping myoepithelial cells, whereas the alveoli cells are surrounded by a basket-like layer of myoepithelial cells. [2].

Ducts and lobules are surrounded from connective tissue containing blood and lymphatic vessels, fat and fibrous tissue in varying proportion called stroma.

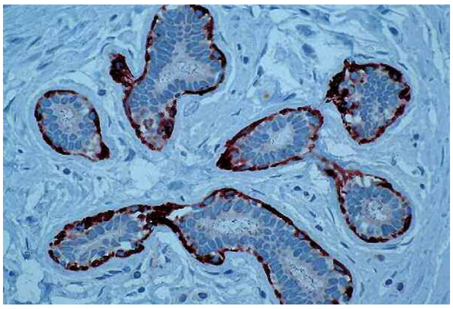


Figure 1.3: Normal lobule: high molecular weight myosin staining x200. Each acinus associates an external layer around the basal membrane of myoepithelial cells stained in red, and an internal epithelial layer in blue [3].

1.2 MAMMARY STEM CELLS AND THEIR ROLE IN BREAST TUMOURIGENESIS

The ability to replenish the mammary gland through cycles of pregnancy, lactation and involution throughout a woman's lifetime is attributed to stem cells that are proposed to reside in the mammary gland [2]. These cells are proposed to serve three functions:

- to give rise to the tissues of the adult mammary gland during development;
- to allow the enormous tissue expansion and remodelling that occurs in the mammary gland during multiple cycles of pregnancy, lactation and involution;
- to serve as a reserve for repair in the event of tissue damage (rarely).

At the onset of puberty, the immature mammary gland undergoes rapid growth and differentiation at the tip of the Terminal End Buds (TEBs) (fig. 1.4). The cap cell layer surrounding the TEB can take on a myoepithelial lineage, and therefore cap cells are thought to be multipotent stem cells. However, the TEBs are considered to be only a temporary niche, since TEBs are transient structures that disappear once the duct reaches the end of the fat pad [2].

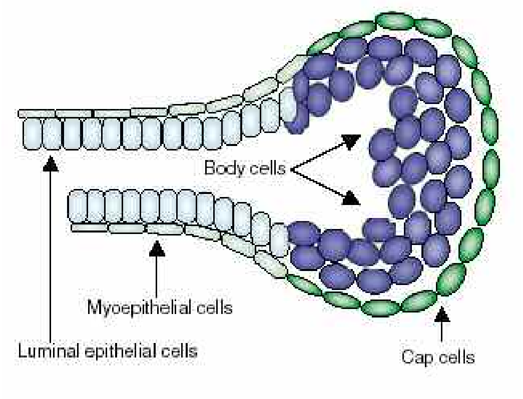


Figure 1.4: The terminal End Bud (TEB). The TEB appears at the onset of puberty, undergoing rapid growth and differentiation [2].

Recent research in breast biology has provided support for the cancer stem-cell hypothesis. Two important components of this hypothesis are that tumours originate in mammary stem or progenitor cells as a result of dysregulation of the normally tightly regulated process of self-renewal. As a result, tumours contain and are driven by a cellular subcomponent that retains key stem-cell properties including self-renewal, which drives tumourigenesis and differentiation that contributes to cellular heterogeneity [4]. In fact stem cells make an attractive candidate for the cellular origin of cancer since they possess many features of the tumour phenotype, including self-renewal and essentially unlimited replicative potential [5].

Data identifying cancer stem cells in breast cancer highlight the need for a dramatic shift in the way we design cancer therapies. Since a small population of cancer stem cells can recapitulate the entire tumour, the current cancer therapy has to be able to eradicate efficiently this small population, which probably drives cancer recurrence [2]. Conventionally cancer therapy that targets proliferating, terminally differentiated cells with limited replicative potential may initially lead to a favourable clinical response but fail to eliminate the cancer stem cells that underpin recurrence [2] (fig. 1.5). Ideally, tumour stem cell therapies would specifically target tumour stem cells. Used alone, they might lead to tumour regression, but not dissolve tumour bulk. Combining conventional therapy with treatment targeting tumour stem cells may effectively eliminate both tumour bulk and tumour stem cells (fig. 1.5). Thus, investigation of the mechanisms and signalling pathways that support stem cell renewal in normal and malignant tissue, may provide new targets for therapies designed to complement existing approaches and reduce tumour recurrence [2].

1.3 BREAST CANCER EPIDEMIOLOGY AND RISK FACTORS

Breast cancer is the most common cancer in women worldwide, comprising 23% of all cancers, with more than one million new cases per year [6] and it is the second cause of cancer death among women globally (411'000 annual world deaths in 2002 [7]) after lung cancer. According to the American Cancer Society, about 1.3 million women will be diagnosed with breast cancer annually worldwide and about 465'000 will die from this disease.

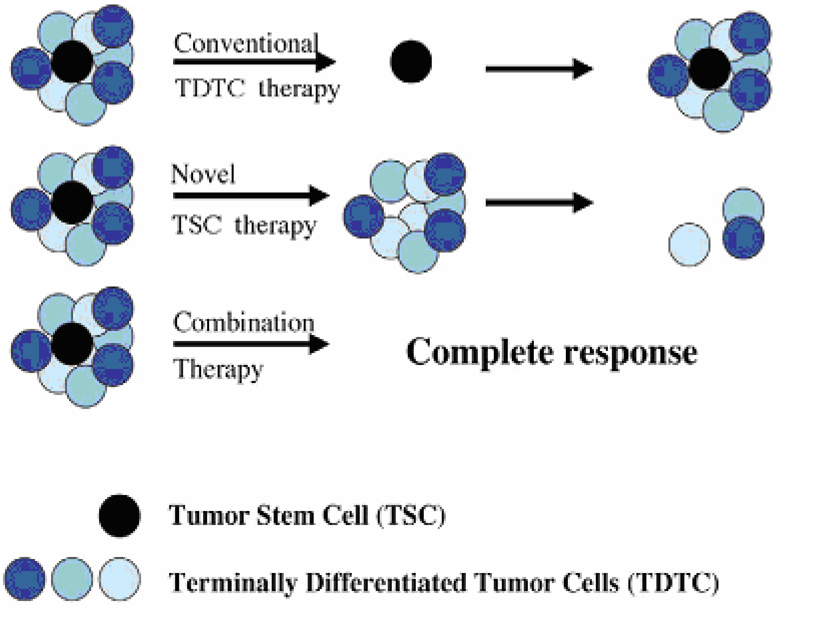


Figure 1.5: Cancer therapy approaches [2].

More than half of all cases occur in industrialized countries, about 361.000 in Europe and 230'000 in North America [7]. The high incidence in the more affluent world areas is likely due to the presence of screening programs that detect early invasive cancers, some of which would otherwise have been diagnosed later or not at all [8]. The prognosis of the breast cancer is generally rather good, so that this cancer ranks as the fifth cause of death from cancer overall (although it is still the leading cause of cancer mortality in women). The very favourable survival of breast cancer cases in western countries is also in part a consequence of the presence of screening programs and of improvements of the treatment.

Cancer results from a combination of many factors including inherited mutations or polymorphisms of cancer susceptibility genes, environmental agents that influence the acquisition of somatic genetic changes and several other systemic and local factors (fig. 1.6) [9]. A risk factor is anything that affects the chance of getting a disease and for breast cancer they are:

- gender: breast cancer is about 100 times more common among women than men;
- aging: risk of developing breast cancer increases as you get older;

- genetic risk factors: ~10% of breast cancer cases are attributable to inherited mutations in highly penetrant breast cancer susceptibility genes, two of which, BRCA1 and BRCA2 have been identified based on genetic linkage studies of affected families; in addition germline mutations of PTEN, LKB1, ATM, p53, MSH2/MLH1, CHEK2 and BACH-1 are associated with breast cancer but to a much more limited extent then the BRCA genes [9]. Polymorphism in several metabolic and detoxifying enzymes (GSTM1, CYP1A1, CYP17, NAT2, SULT1A1, COMT, SOD), components of hormonal signalling pathways (oestrogen and androgen receptor), proto-oncogenes (H-ras-VNTR), DNA repair genes (XRCC1, XRCC3) and HLA alleles have been shown to influence breast cancer susceptibility [9];
- family history of breast cancer: breast cancer risk is higher among women whose close blood relatives have this disease;
- race and ethnicity: caucasian women are slightly more likely to develop breast cancer than are African-American women;
- personal history of breast cancer: a woman with cancer in one breast has a 3- to 4-fold increased risk of developing a new cancer in the other breast or in another part of the same breast;
- dense breast tissue: women with denser breast tissue (as seen on a mammogram) have more glandular tissue and less fatty tissue, and have a higher risk of breast cancer;
- certain benign breast conditions: women diagnosed with certain benign breast conditions may have an increased risk of breast cancer;
- menstrual periods: women who have had more menstrual cycles because they started menstruating at an early age (before age 12) and/or went through menopause at a later age (after age 55) have a slightly higher risk of breast cancer;
- previous chest radiation: women who, as child or young adult, had radiation therapy to the chest area as treatment for another cancer (such as Hodgkin disease or non-Hodgkin lymphoma) are at significantly increased risk for breast cancer;
- not having children, or having them later in life: women who have had no children or who had their first child after age 30 have a slightly higher breast cancer risk;
- using post-menopausal hormone therapy: Long-term use (several years or more) of combined post-menopausal hormone therapy increases the risk of breast cancer and may also increase the chances of dying of breast cancer;
- alcohol: use of alcohol is clearly linked to an increased risk of developing breast cancer;
- being overweight or obese: this condition has been found to increase breast cancer risk, especially for women after menopause.

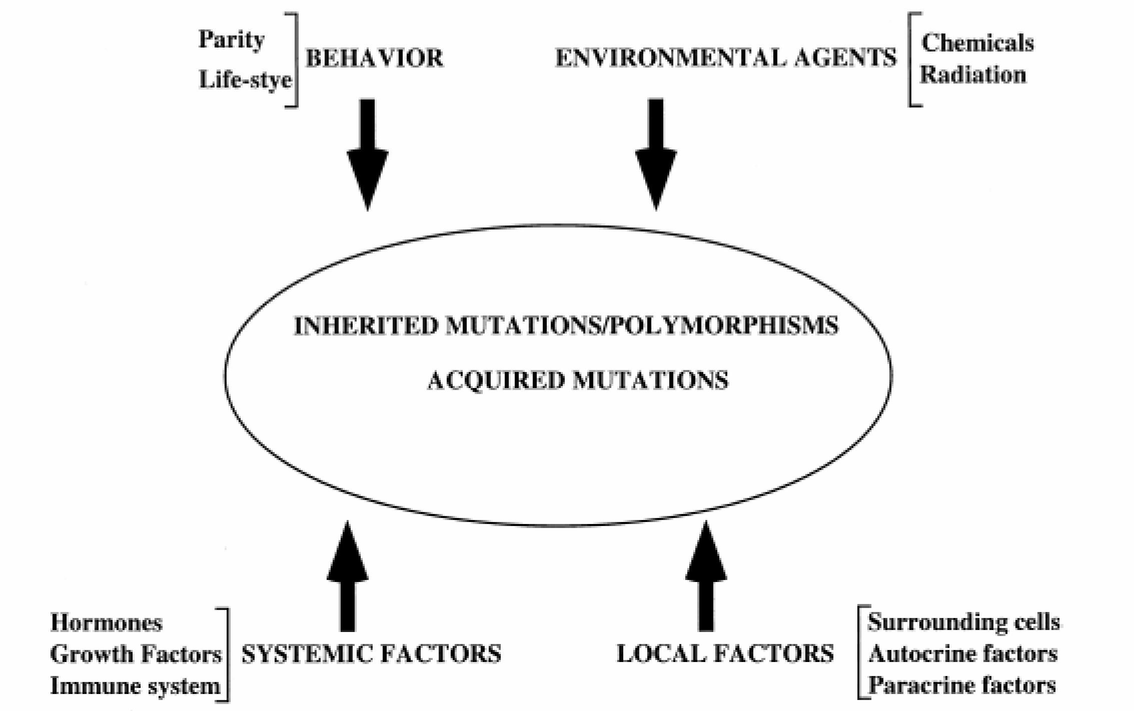


Figure 1.6: Summary of factors influencing breast carcinogenesis [9].

1.4 BREAST CANCER CLASSIFICATION

Breast cancer classification divides all forms of breast cancer according to four different schemes, each based on different criteria. The four approaches consider the **histology** (1.4.1), the **grade** (1.4.2), the **stage** (1.4.3) and the **gene expression profile** of the tumour (1.4.4).

The morphological attributes of a breast cancer, as assessed by histological examination, supply the breast cancer care team with invaluable prognostic and predictive information. Together with the great advances in molecular techniques availability, morphology remains indispensable [10]. Studies of most large cohorts of unselected breast cancers continue to show that grade (see 1.4.2), nodal status and tumour size (see 1.4.3) remain powerful prognostic factors in a multivariable analysis [10]. These are the parameters combined in the Nottingham Prognostic Index (NPI), the most used prognostic tool in use in the UK [11]. Tumour size and nodal status are very much temporal factors, whereas grade is a morphological attribute and is qualitative. It is a reflection of the intrinsic qualities of a tumour and it will give an indication of features such as the rapidity of growth and probability of metastasis [10].

1.4.1 BREAST CANCER HISTOLOGICAL CLASSIFICATION

The most significant effort in the classification of tumours of the breast was that produced by the World Health Organization (WHO) in 2003. All carcinomas of the breast, both invasive and non-invasive (*in situ*), are classified on the basis of the histological and/or cytological appearance [12, 13].

Carcinoma *in situ*
- Ductal carcinoma *in situ*
- Lobular carcinoma *in situ*

Invasive Carcinoma
- Invasive ductal carcinoma, Not Otherwise Specified (NOS)
    - Mixed type carcinoma
    - Pleomorphic carcinoma
    - Carcinoma with osteoclastic giant cells
- Invasive lobular carcinoma
- Tubular carcinoma
- Invasive cribriform carcinoma
- Medullary carcinoma
- Mucinous carcinoma and other tumours with abundant mucin
    - Mucinous carcinoma
    - Cystadenocarcinoma and columnar cell mucinous carcinoma
    - Signet ring cell carcinoma
- Invasive papillary carcinoma
- Invasive micropapillary carcinoma
- Apocrine carcinoma
- Metaplastic carcinomas
- Pure epithelial metaplastic carcinomas
- Mixed epithelial/mesenchymal metaplastic carcinomas
- Lipid-rich carcinoma
- Adenoid cystic carcinoma
- Acinic cell carcinoma
- Glycogen-rich clear cell carcinoma
- Inflammatory carcinoma
- Microinvasive carcinoma

Table 1.1: Histological classification of breast carcinoma (adapted from [12])

As reported in the table 1.1, there are two major groups of breast tumours: carcinoma *in situ* and invasive carcinoma.

1.4.1.1 Carcinoma *in situ*

Carcinoma *in situ* is a proliferation of malignant epithelial cells within the ductulo-lobular system of the breast that on light microscopy shows no evidence of breaching the basement membrane to invade the adjacent stroma. There are two forms: ductal and lobular. Lobular Intraepithelial Neoplasia (LIN) (fig. 1.7) is located within the terminal duct-lobular unit, often accompanied by pagetoid involvement of the adjacent terminal ducts (fig. 1.7). These are markedly distended by a proliferation of monomorphous cells that have effaced the lumen. LIN is associated with an increase in the risk of developing invasive breast cancer [13]. Ductal Carcinoma *In Situ* (DCIS) (fig. 1.7) is a heterogeneous group of pre-malignant lesions that may be identifiable on mammography as foci of microcalcification [14]. For DCIS are defined three categories on the basis of cytonuclear differentiation: poorly differentiated, intermediately differentiated and well-differentiated. Lesions in the poorly differentiated group are usually Erb-B2 positive and are less frequently oestrogen and progesterone receptor positive, conversely to those in the well-differentiated group. The treatment of DCIS depends on the size and distribution of the lesion. The status of excision margins around the tumour remains the most important factor in terms of risk of local recurrence. Microinvasive carcinoma (size limit of 1 mm) is rare and occurs mostly in association with *in situ* carcinoma, usually of the poorly differentiated type [13].

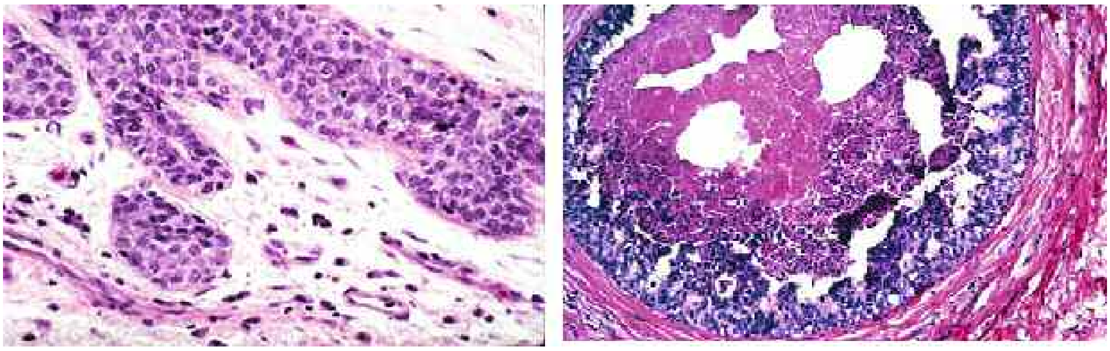


Figure 1.7: Left: lobular intraepithelial neoplasia. Right: poorly differentiated ductal carcinoma in situ, adapted from [13]

1.4.1.2 Invasive carcinoma

Invasive breast cancer is a group of malignant epithelial tumours characterized by invasion of adjacent tissue and a marked tendency to metastasize to distant sites. Breast cancer arises from the mammary epithelium, most frequently from the cells of the terminal duct lobular unit [13]. Numerous histological types of breast carcinoma have been identified (see tab. 1.1) but one in particular, invasive ductal carcinoma (Not Otherwise Specified, NOS) (fig. 1.8) largely predominates and represents more than 75% of all cases [3].

Generally, three different components are associated, the stroma, the invasive component and the *in situ* component. Stroma is essential for tumour growth. It is composed of vessels which ensure the nutrional supply, inflammatory cells and a hyaline and elastosic tissue. Ductal NOS tumours are less common below the age of 40. These tumours do not have specific macroscopic features. Invasive carcinoma is often associated with high grade ductal carcinoma *in situ*. If a ductal carcinoma NOS is accompanied by a second distinct morphologic pattern (lobular), the cancer is defined as mixed. Approximately 70-80% of ductal NOS breast cancers are Estrogen Receptor (ER) positive, and between 15-30% of cases are Erb-B2 positive.

Many other types of lesions have been identified according to the cytological and the architectural presentation. Examples are medullary, mucinous and papillary carcinomas. They all contain a particular stroma which is either inflammatory (medullary carcinoma), colloid (mucinous carcinoma) or vascular (papillary carcinoma) with little or no fibrous tissue. These tumours do not attract the surrounding tissue but push it as they grow, giving rise to a nodular shape [3]. Medullary, mucinous and papillary carcinomas have relatively good prognosis and represent between 1 and 7% of all breast cancers [13].

The second most frequent type is lobular carcinoma (fig. 1.8), accounting for almost 10% of invasive carcinomas. It is characterized by indistinct tumours margins. About 70-95% of lobular carcinomas are ER positive and 60%-70% are Progesteron Receptor (PR) positive. Overexpression of Erb-B2 is lower than in invasive ductal carcinoma [13].

Tubular carcinoma shows a favourable prognosis and accounts for under 2% of invasive breast cancer in most series. Ductal carcinoma *in situ* is found in association; occasionally the *in situ* component is of lobular type. Oestrogen and Progesteron receptors are always positive and Erb-B2 is negative [13].

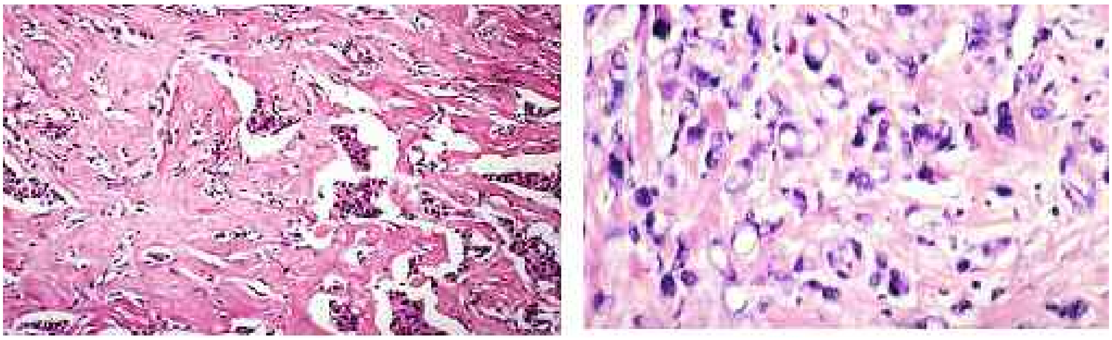


Figure 1.8: Left: invasive ductal carcinoma (NOS). Right: invasive lobular carcinoma, adapted from [13].

1.4.2 BREAST CANCER GRADING

In situ ductal carcinoma and all invasive tumours are routinely graded. Among the various grading systems that have been proposed, the combined grading method of Elston and colleagues from Nottingham is currently the most widely used in Europe [15]. In this system three parameters are evaluated: tubule formation, nuclear polymorphism and mitotic rate. A numerical scoring system of 1-3 is used to ensure that each factor is assessed individually [13].

The three values are combined together and produce scores of 3 to 9, to which the grade is assigned:

- total score 5: grade 1, well differentiated;
- total score 6-7: grade 2, moderately differentiated;
- total score 8-9: grade 3, poorly differentiated.

1.4.3 CLINICAL CLASSIFICATION OR TNM STAGING SYSTEM FOR THE BREAST CANCER

In 2002 the American Joint Committee on Cancer (AJCC) published the sixth edition of the Cancer Staging Manual, that reports additions made to the staging system, designed to facilitate the uniform collection of clinically relevant information about new techniques for the detection of metastatic cells [16, 17]. These additions include quantitative criteria to distinguish micrometastases from isolated tumour cells, and specific identifiers to record the use of sentinel lymph node biopsy, immunohistochemical (IHC) staining, and molecular biology techniques. Revisions of the previous staging system are related to the number of affected axillary lymph nodes and to their classification [17].

Primary tumour (T)

TX		Primary tumour cannot be assessed
T0		No evidence of primary tumour
Tis		Carcinoma *in situ*
	Tis (DCIS)	Ductal Carcinoma *In Situ*
	Tis (LCIS)	Lobular carcinoma *In Situ*
	Tis (Paget)	Paget's disease of the nipple with no tumour Note: Paget's disease associated with a tumour is classified according to the size of the tumour.
T1		Tumour ≤ 2 cm in greatest dimension
	T1mic	Microinvasion ≤ 0.1 cm in greatest dimension
	T1a	Tumour > 0.1 cm but not > 0.5 cm in greatest dimension
	T1b	Tumour > 0.5 cm but not > 1 cm in greatest dimension
	T1c	Tumour > 1 cm but not > 2 cm in greatest dimension
T2		Tumour > 2 cm but not > 5 cm in greatest dimension
T3		Tumour > 5 cm in greatest dimension
T4		Tumour of any size with direct extension to (a) chest wall or (b) skin, only as described below
	T4a	Extension to chest wall, not including pectoralis muscle
	T4b	Edema (including peau d'orange) or ulceration of the skin of the breast, or satellite skin nodules confined to the same breast.
	T4c	Both T4a and T4b
	T4d	Inflammatory carcinoma

Regional lymph nodes (N)

NX		Regional lymph nodes cannot be assessed (eg, previously removed)
N0		No regional lymph node metastasis
N1		Metastasis in movable ipsilateral axillary lymph node(s)
N2		Metastases in ipsilateral axillary lymph nodes fixed or matted, or in clinically apparent* ipsilateral internal mammary nodes in the absence of clinically evident axillary lymph node metastasis.
N2a		Metastasis in ipsilateral axillary lymph nodes fixed to one another (matted) or to other structures.
N2b		Metastasis only in clinically apparent* ipsilateral internal mammary nodes and in the absence of clinically evident axillary lymph node metastasis.
N3		Metastasis in ipsilateral infraclavicular lymph node(s), or in clinically apparent* ipsilateral internal mammary lymph node(s) and in the presence of clinically evident axillary lymph node metastasis; or metastasis in ipsilateral supraclavicular lymph node(s) with or without axillary or internal mammary lymph nod involvement.
	N3a	Metastasis in ipsilateral infraclavicular lymph node(s) and axillary lymph node(s).
	N3b	Metastasis in ipsilateral internal mammary lymph node(s) and axillary lymph node(s).
	N3c	Metastasis in ipsilateral supraclavicular lymph node(s)

Regional lymph nodes (pN)†

pNX		Regional lymph nodes cannot be assessed (eg, previously removed or not removed for pathologic study).
pN0		No regional lymph node metastasis histologically, no additional examination for isolated tumour cells‡.
	pN0(i_)	No regional lymph node metastasis histologically, negative IHC.
	pN0(i+)	No regional lymph node metastasis histologically, positive IHC, no IHC cluster > 0.2 mm.
	pN0(mol_)	No regional lymph node metastasis histologically, negative molecular findings (RT-PCR)
	pN0(mol+)	No regional lymph node metastasis histologically, positive molecular findings (RT-PCR)
	pN1mi	Micrometastasis (_ 0.2 mm, none _ 2.0 mm)
	pN1	Metastasis in one to three axillary lymph nodes and/or in internal mammary nodes with microscopic disease detected by sentinel lymph node dissection but not clinically apparent§
	pN1a	Metastasis in one to three axillary lymph nodes
	pN1b	Metastasis in internal mammary nodes with microscopic disease detected by sentinel lymph node dissection but not clinically apparent§.
pN1c		Metastasis in one to three axillary lymph nodes and in internal mammary lymph nodes with microscopic disease detected by sentinel lymph node dissection but not clinically apparent§.

pN2	Metastasis in four to nine axillary lymph nodes, or in clinically apparent* internal mammary lymph nodes in the absence of axillary lymph node metastasis.
pN2a	Metastasis in four to nine axillary lymph nodes (at least one tumour deposit _ 2.0 mm).
pN2b	Metastasis in clinically apparent* internal mammary lymph nodes in the absence of axillary lymph node metastasis.
pN3	Metastasis in 10 or more axillary lymph nodes, or in infraclavicular lymph nodes, or in clinically apparent* ipsilateral internal mammary lymph nodes in the presence of one or more positive axillary lymph nodes; or in more than three axillary lymph nodes with clinically negative microscopic metastasis in internal mammary lymph nodes; or in ipsilateral supraclavicular lymph nodes
pN3a	Metastasis in 10 or more axillary lymph nodes (at least one tumour deposit > 2.0 mm), or metastasis to the infraclavicular lymph nodes
pN3b	Metastasis in clinically apparent* ipsilateral internal mammary lymph nodes in the presence of one or more positive axillary lymph nodes; or in more than three axillary lymph nodes and in internal mammary lymph nodes with microscopic disease detected by sentinel lymph node dissection but not clinically apparent§.
pN3c	Metastasis in ipsilateral supraclavicular lymph nodes

Distant metastasis (M)

MX	Distant metastasis cannot be assessed
M0	No distant metastasis
M1	Distant metastasis

Table 1.2 TNM Staging System for Breast Cancer. Adapted from [17].
IHC, ImmunoHistoChemistry; RT-PCR, Reverse Transcriptase Polymerase Chain Reaction. *****"Clinically apparent" is defined as detected by imaging studies (excluding lymphoscintigraphy) or by clinical examination. **†**Classification is based on axillary lymph node dissection with or without sentinel lymph node dissection. Classification based solely on sentinel lymph node dissection without subsequent axillary lymph node dissection is designated (sn) for "sentinel node" (eg, pN0(i+)(sn)). **‡**Isolated tumour cells are defined as single tumour cells or small cell clusters not greater than 0.2 mm, usually detected only by immunohistochemical or molecular methods but which may be verified on hematoxylin and eosin stains. Isolated tumour cells do not usually show evidence of metastatic activity (eg, proliferation or stromal reaction). **§**"Not clinically apparent" is defined as not detected by imaging studies (excluding lymphoscintigraphy) or by clinical examination.

The staging system is called **TNM**: **T** describes the size of the tumour and whether it has invaded nearby tissue, **N** describes regional lymph nodes that are involved, and **M** describes distant metastasis (spread of cancer from one body part to another) (tab. 1.2). In the table 1.3 I reported a summary of the TNM staging system. The stage of the tumour is defined as a combination of T, N and M parameters. The adoption and routine utilization of the staging system is critically important in laying the groundwork for future decisions regarding breast cancer prognosis and treatment [17].

Stage	Tumour size	Lymph nodes	Metastasis
0	Tis	N0	M0
I	T1	N0	M0
IIA	T0	N1	M0
	T1	N1	M0
	T2	N0	M0
IIB	T2	N1	M0
	T3	N0	M0
IIIA	T0	N2	M0
	T1	N2	M0
	T2	N2	M0
	T3	N1	M0
	T3	N2	M0

IIIB	T4	N0	M0
	T4	N1	M0
	T4	N2	M0
IIIC	Any T	N3	M0
IV	Any T	Any N	M1

Table 1.3: TNM Stage Grouping for breast cancer (adapted from [18]).

The survival rate (see table 1.4) depends on the stage of the tumour and also on others factors, like the medical situation of the patient, etc. A survival rate of five years is the average number of patients still alive after five years from the first diagnosis. After seven years the survival rate decreases in all classes (stage I 92%, stage II 71%, stage III 39%, stage IV 11%).

Stage	5-years survival rate
0	100%
I	98%
IIA	88%
IIB	76%
IIIA	56%
IIIB	49%
IV	16%

Table 1.4: Breast cancer survival rate after five years from the first diagnosis (adapted from [18]).

1.4.4 MOLECULAR CLASSIFICATION OF BREAST CANCER BASED ON GENE EXPRESSION PROFILE

Breast cancer is a clinically heterogeneous disease. Histologically similar tumours may have different prognoses and may respond to therapy differently [19]. It is believed that these differences in clinical behaviour are due to molecular differences between histologically similar tumours. DNA microarray technology is ideally suited to reveal such molecular differences because it is a powerful tool to look at genome-wide gene expression [20].

It is important to remind that the transcriptional profile of the tumour reflects the contribution of a large number of cellular components including normal breast epithelium, cancer cells, fibroblasts, adipocytes, infiltrating leukocytes and vascular components [21]. This tissue heterogeneity could be a confounding factor and makes data analysis difficult because of the high background. Also the profiles are distorted by transcriptional response to surgical stress, tissue handling, and general anesthetics or other drugs administered before and during surgery [21]. The expression of genes involved in signal transduction and response to stress or hypoxia may be particularly sensitive to rapid changes in the tissue microenvironment that occur during surgery. For these reasons, it is essential using suitable sampling procedure. Pustztai and colleagues reported [21] that the Fine-Needle Aspiration (FNA) is a reliable technique to collect the tumour sample. FNA is minimally invasive and the cells removed with this method frequently represent relatively pure tumour cells. In recent years has been also developed another technique, the Laser Capture microdissection (LCM) for isolating individual cells or subcellular structures from a heterogeneous cell population. Ma and colleagues [22] applied successfully LCM and DNA microarray technology to identify different pathological stages of breast tumours.

Perou and colleagues have proposed for the fist time, that the phenotypic diversity of breast tumours might be accompanied by a corresponding diversity in gene expression pattern [23]. They used microarrays to investigate gene expression pattern in 42 breast cancer patients by unsupervised hierarchical cluster analysis. A set of 496 genes ("intrinsic gene set") was selected based on large variation in expression between 2 biopsies from one patient (pre and post-neoadjuvant chemotherapy treatment). By clustering tumours using this "intrinsic gene list" (supervised approach) were identified 4 subgroups of cancers with separate gene expression profiles: the luminal/epithelial ER+, the basal type, the normal-like and the Erb-B2+ groups (fig. 1.9). In a subsequent study, by Sorlie and colleagues [24], the luminal group was subdivided into the luminal A and luminal B subgroups. These studies highlighted that the molecular subgroups were strongly associated with ER status: luminal types are ER positive, basal-like and Erb-B2+ mostly ER negative.

Subsequently Sotiriou and colleagues [25] correlated gene expression patterns generated from cDNA microarrays with clinico-pathological characteristics and clinical outcome in an unselected group of 99 node-negative and node-positive breast cancer patients. Gene expression pattern were found highly related with ER status, as reported from [23, 24] and moderately associated with grade, but not associated with menopausal status, nodal status or tumour size. They showed that ER status of the tumour was, indeed, the most important discriminator of expression subtypes and that tumour grade was a distant second. This finding confirms that ER biology plays a central role in breast carcinogenesis defining the configuration of the final tumour. Moreover, a hierarchical cluster analysis segregated their population into two distinct subgroups with different relapse-free survival: the basal-like and Erb-B2+ subgroups had the shortest relapse-free and overall survival, whereas the luminal-type tumours had a more favourable clinical outcome [19].

1.4.4.1 Luminal type breast tumours

The luminal-type tumours, also called ER+ cluster, are characterized by the relatively high expression of many genes typical of breast luminal cells [23]. They show a high expression of luminal cell keratins 8/18, ESR1 (Oestrogen Receptor 1) and other genes involved in the ESR1 activation, like SLC39A6 (Solute Carrier Family 39 zinc transporter, member 6) and CCND1 (cyclin D1) [23, 25]. Less than 20% of luminal-type tumours contain TP53 mutations [24, 25]. Sørlie and colleagues [23], increasing the sample size of their previous study [23], identified within the luminal group three distinct subgroups: luminal A, luminal B and luminal C. The subtype luminal A shows the highest expression of ESR1, GATA binding protein 3, X-box binding protein 1, trefoil factor 3, hepatocyte nuclear facto 3 alpha and oestrogen-regulated LIV-1. The others two subtypes, B and C, show low to moderate expression of the luminal – specific genes including the ER cluster. Luminal subtype C was further distinguished from luminal A and B by the high expression of a novel set of genes which are a feature they share with the basal-like and Erb-B2+ subtypes [24]. The luminal subtypes show differences in clinical outcome [25]. The luminal A has a better outcome than luminal subtypes B and C, in fact they both might represent a clinically distinct group with a different and worse disease course, in particular with respect to relapse [24]. Luminal subtype C expresses some of the genes characteristic of the ER-negative tumours in the basal-like and Erb-B2+ subtypes, suggesting the poor disease outcome. The different prognosis of three luminal subgroups could be related to chemotherapy response. The luminal tumours are treated with hormonal therapy, because several studies showed that the traditional chemotherapy fails in ER+ tumours [26].

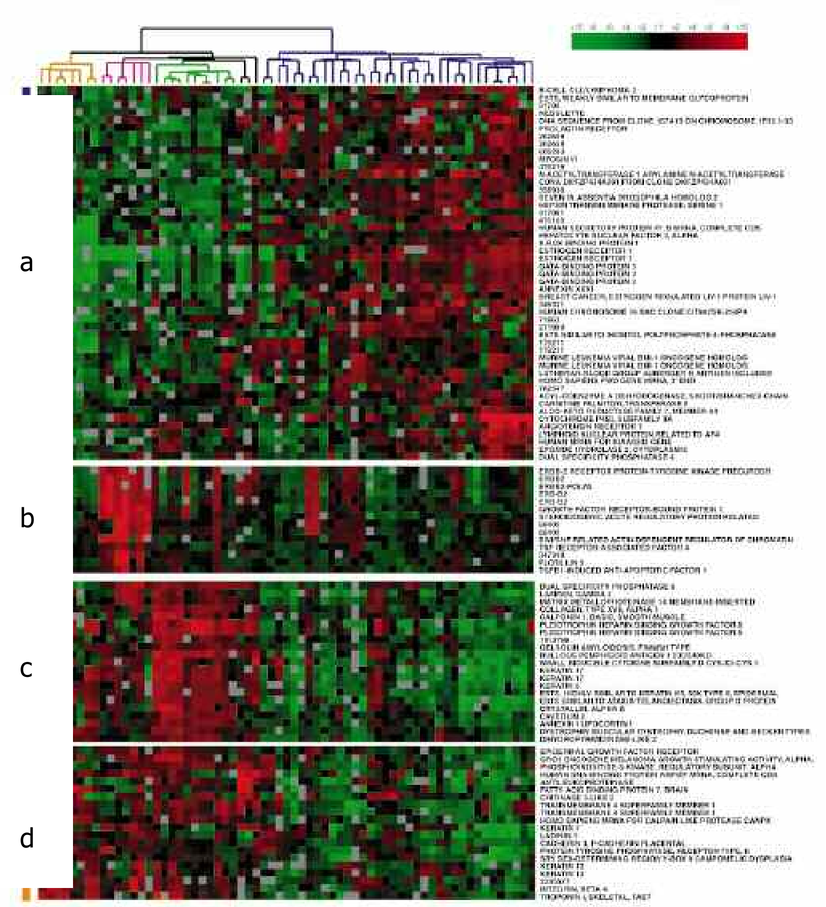


Figure 1.9: Cluster analysis using the "intrinsic"gene subset. Two large branches were apparent in the dendogram, and within these large branches were smaller branches for which common biological themes could be inferred. Branches are coloured accordingly: basal-like, yellow; Erb-B2+, pink; normal-breast like, light green; and luminal epithelial/ER+, dark blue. There are 4 clusters of interesting genes: luminal epithelial/ER+ (a), Erb-B2+ (b), basal epithelial cell associated cluster containing keratin 5 and 17 (c), a second basal epithelial-cell-enriched gene cluster. Adapted from [23].

Recently it was reported [27] a better survival in metastatic luminal breast cancer treated with Bevacizumab (Avastin, Genentech), a humanized monoclonal antibody directed against all isoforms of VEGF-A. The luminal-type breast tumours are the most common tumours type, ~60-70% in the caucasian woman population, as the Caroline Breast Cancer Study has reported [28]. Recently Rouzier and colleagues [19] showed that breast cancer molecular subtypes respond differently to preoperative chemotherapy. The luminal tumours, together with the normal-like tumours, have lower pathologic Complete Response (pCR) rates than the basal-like and Erb-B2 tumours. Instead the basal-like (par. 1.4.4.2) and Erb-B2 tumours (par. 1.4.4.4) are predominantly high nuclear grade and the basal-like tumours are often ER negative: both of these characteristics are known to be associated with higher likelihood of pathologic CR to preoperative chemotherapy [29].

1.4.4.2 Basal-like breast tumours

Basal-like tumours typically show low expression of Erb-B2 and ER and exhibit high expression of genes characteristic of the basal epithelial cell layer, including expression of cytokeratins 5, 6 and 17 [27]. Basal-like breast carcinomas account for up to 15% of all breast cancers and often affect younger patients [23, 24, 30]. These tumours show either p53 immunohistochemical expression or TP53 mutations in up to 85% of cases, display exceedingly high levels of proliferation-related genes and express Epidermal Growth Factor Receptor (EGFR) in >60% of cases [22, 23, 30, 31]. Morphologically, basal-like breast carcinomas are characterized by high histological grade, high mitotic index and the presence of metaplastic elements [32]. In fact, recent studies have demonstrated that >90% of metaplastic breast carcinomas show a basal-like phenotype [33]. Basal-like cancers have a more aggressive clinical behaviour, some studies have demonstrated that the expression of basal keratins is a prognostic factor independent of tumour size, grade and lymph node status [34]. Given that by microarray-based expression analysis, basal-like cancers are preferentially negative for ER and PR and lack Erb-B2 expression, they are often used as synonym for the triple-negative breast cancers, although not all basal-like tumours are negative for ER, PR and Erb-B2; in fact 15%-54% of them express at least one of these markers [20, 25, 26].

Although the basal-like breast cancers show high rates of objective response to neoadjuvant chemotherapy, patients with this tumour-type that have not evolved to pathological complete response still show a significantly poorer prognosis than those with tumours pertaining to other molecular subgroups [35] This could be explained by the limited therapeutical options for the basal-like subtype due to the triple negativity for the reported markers and not by a drug-resistance of the primary tumour.

There is increasingly more coherent evidence to suggest a link between the BRCA1 (BReast CAncer 1) pathway and basal-like breast cancers. In fact, the vast majority of tumours arising in BRCA1 germ-line mutation carriers, in particular those diagnosed before 50 years of age, have morphological features similar to those described in basal-like cancers and they display a basal-like phenotype [36]. It has recently demonstrated that the BRCA1 pathway may be dysfunctional in sporadic basal-like tumours [37]. BRCA1 protein expression levels have been shown to be significantly lower in tumours of high histological grade, lacking ER and PR expression and of basal-like phenotype [38]. Since BRCA1 is involved in the DNA repair system, mutations in this gene could cause alterations in DNA-damage-repair pathway and increase the tumour sensitivity to chemotherapic agents that damage DNA.

Taken together, there is a evidence that BRCA1 pathway dysfunction is integral to the biology of basal-like breast carcinomas. They show, in fact, a sensitivity to cross-linking agents (e.g. platinum salts) and to inhibitors of the poly ADP-ribose polymerase (PARP) enzyme [39, 40]. These findings suggest new therapeutic strategies for the management of patients with basal-like breast cancers for testing in clinical trials.

1.4.4.3 Normal-like breast tumours

The normal breast-like group shows the highest expression of many genes known to be expressed by adipose tissue and other nonepitelial cell types [24]. These tumours also are characterized from strong expression of basal epithelial genes and low expression of luminal epithelial genes [24]. The normal-like tumours cluster close to Erb-B2 tumours and share with them the short relapse-free and overall survival respect the luminal-type tumours. Rouzier and colleagues [19] showed that the normal-like tumours, together the luminal-type ones, are less sensitive to paclitaxel- and doxorubicin-containing preoperative chemotherapy than the basal-like and Erb-B2 tumours.

1.4.4.4 Erb-B2 type breast tumours

The Erb-B2 subtype is characterized by high expression of several genes in the Erb-B2 amplicon at 17q22.24 including Erb-B2 and GRB7 (Growth factor Receptor-Bound protein 7) [24], index of a possible genic amplification. These tumours also show low levels of expression of ER and of almost all of the others genes associated with ER expression, a trait they share with the basal-like tumours [23].

The Erb-B2 tumours have often TP53 mutations (71% of cases). Risk factors associated to this type of tumours are not yet characterized, but some studies showed that often they are high grade (III) tumours, poorly differentiated, with a double probability to be sentinel lymph node positive than the luminal type ones [27].

Sørlie and colleagues [24] reported that overexpression of Erb-B2 gene is a well-known prognostic factor associated with poor survival in breast cancer, which also was found for the Erb-B2 molecular subtype.

Although the poor prognosis, the Erb-B2 subtype is, similarly to basal-like subtype, more sensitive to paclitaxel- and doxorubicin-containing preoperative chemotherapy than the luminal and normal like-cancers.

The Erb-B2 tumours are the only molecular subtype target of Trastuzumab commercial name Herceptin®), a humanized receptor antibody directed against Erb-B2. Patients who received concurrent preoperative chemotherapy and trastuzumab had a significantly higher pCR than those who received chemotherapy alone [41]. Resistance to trastuzumab is an active research field. Several known mechanisms of resistance have been identified , like increased production of insulin-like growth factor, dysregulation of p27, overexpression of epidermal growth factor receptor with activation of the Akt pathway, and decreased PTEN (Phosphatase and TENsin homolog) [42].

1.5 MOLECULAR FORECASTING IN BREAST CANCER: PREDICTION OF THE TUMOUR COURSE BASED ON THE GENE EXPRESSION PROFILING

Like all malignancies, breast cancer arises as a result of the accumulation of genetic alterations, most importantly deregulation of the expression of oncogenes and tumour suppressor genes. As a consequence this will lead to highly proliferating cells that lose their differentiation and have the ability to become invasive and metastatic. There are various genetic pathways that have been identified and it has become clear that breast cancer represents a heterogeneous disease [43]. This heterogeneity of breast cancer is also reflected in the variable clinical courses of the disease. Some patients will develop metastases at an early stage, other tumours will never metastasize. Treatment that is effective in one patient may not show the same efficacy in other patients with a similar tumour type. Since the effectiveness of treatment differs between individual patients, much effort is being invested in the identification of new prognostic and predictive markers. Prognostic factors are important to estimate patient's outcome and can be used to decide which patients will need additional adjuvant systemic treatment (systemic therapy to eliminate any remaining tumour cells after surgical removal of the primary tumour). Instead predictive factors indicate which treatment is most effective for an individual patient. More specifically, prognostic biomarkers predict the clinical outcome for a patient if no anticancer drugs are administered, whereas predictive biomarkers predict the outcome of a specific therapy for a patient [44]. The implementation of predictive factors in clinical decision making will help to ensure that only patients that are likely to benefit from a specific treatment will receive this specific therapy [43]. So far, clinical and pathological factors guide important decision in the treatment of breast sample patients. However the clinical and pathological factors now available, are not accurately reflecting the heterogeneity in the prognosis and responsiveness to various therapies.

About 10 years ago, microarray technology has enabled scientist to analyze the expression of thousands of mRNAs simultaneously. The mayor improvement came in the 1990s, when the first papers were published [45, 46] using a two-colors microarray, that allowed the analysis of the relative abundance of thousands of mRNAs in one experiment. The possibility to analyze the expression of thousands of genes in one experiment, instead of performing single markers studies, has provided a powerful tool to gain new insights in tumour biology that will help to develop diagnostic tool for clinical routine [43].

After large-scale gene-expression data sets have been collected, there are two different ways to analyze them [43]. The first approach, called **unsupervised classification** (or hierarchical clustering), is to ask whether in a group of samples, there are subgroups or clusters of samples with similar gene-expression patterns. These similarities in gene expression can be used to classify a cancer into subtypes that could have similarities in biological behaviour. This type of data analysis has the advantage that additional clinical data are not required [44]. For example unsupervised analysis identified five molecular subtypes of breast cancer that differ markedly in their aggressiveness and prognosis (see 1.4.4).

The second approach to data analysis is known as **supervised classification**. Samples are divided into groups that are known to have different clinical end points (for example, recurrence versus no recurrence, drug response versus no drug response), and genes that can correctly identify the distinct groups are searched for. One set of tumours (called training set) is used to identify the genes that discriminate between the groups - the gene expression signature- and then a second, independent, set of tumours (called the validation set) is used to test how well these genes can classify samples that have not been grouped [44].

Although unsupervised approaches seem to be less biased, the possibility to identify informative molecular details of clinical subgroups is enhanced when all additional available clinical information is included in the analysis [43].

We could describe three approaches how to find connections between the patterns of gene expression by tumour cells and the behaviour of these cells: the data-driven approach, the knowledge-driven approach and the model-driven approach [44]. The most straightforward is the **data-driven approach**, in which a genome-wide analysis of gene expression is carried out, and then correlates between patterns of gene expression and certain tumour traits are searched for. The strength of this approach is that it is unbiased: there are no assumptions about which genes are likely to be involved in the process of interest. A drawback of this approach is that the outcome relies solely on the quality of the data (and the samples).

By contrast, using the **knowledge-driven approach**, genes that are thought to be relevant to a particular cancer trait, are selected on the basis of the scientific literature. This approach is often used when only formalin-fixed paraffin-embedded tumour tissue is available. The RNA isolated from such tissue is fragmented, and such poor-quality RNA is far from ideal for genome-wide quantitative analysis using DNA microarrays. It can, however, be analysed by PCR with reverse transcription, although this approach precludes genome-scale analysis of gene expression. Thus, in studies involving formalin-fixed paraffin-embedded material, sets of 'likely suspect' genes are tested. A drawback of this approach is that the genes that are not known to be involved in a process cannot be considered.

In the **model-driven approach**, the transcriptional responses of cells after exposure to specific stimuli are used to predict tumour traits. For example, a gene-expression signature for wound healing has been used to predict the survival of individuals with breast cancer (see 1.5.1). Similarly, gene expression signatures that reflect the activation of specific oncogenic pathways have been used both to determine prognosis and to predict responses to anticancer drugs [44]. This approach has the drawback that the experimental model used might not accurately reflect the processes that occur in tumours.

If the presence of a certain transcription factor is known to affect the prognosis of individuals with a particular cancer, then in the knowledge-driven approach, the gene encoding this transcription factor would be incorporated into a prognostic signature. In some cancers, however, this gene might be expressed, but its product might be not-functional (for example, as a result of a missense mutation). For this reason, in a data-driven approach, targets downstream of a transcription factor of interest are often found to be distinguishing features, rather than the gene encoding the transcription factor itself, because the expression of these targets provides more relevant information on the activity of the transcription factor [44]. For example the Paik's gene signature (see 1.5.1) for the prognosis of breast cancer that was derived from 250 'candidate' genes selected on the basis of published studies, includes ESR1 which encodes oestrogen receptor-α (ER-α; a transcription factor that is expressed by most breast cancers). By contrast, the van 't Veer's signature (see 1.5.1) for assessing breast-cancer prognosis that was identified by a data-driven approach, does not include ESR1 itself but includes several genes that are targets of ER-α4 [44].

1.5.1 PROGNOSTIC PROFILES

Prognostic indicators based on currently available clinical and histopathologic variables already exist and are used in clinical practice. Examples of such indicators include the Nottingham Prognostic Indicator (NPI), the St Gallen criteria, the NIH consensus guidelines and Adjuvant!Online (Adjuvant! Inc, San Antonio) which use criteria like tumour size, tumour grade, lymph node status and hormone receptor status to predict a patient's outcome [47, 48, 49, 50]. However, these indicators are still inadequate in that within a given patient population with a specific predicted risk of recurrence, there are always patients whose actual clinical outcome does not match that predicted by the indicator [51]. Even well-validated tools like Adjuvant Online, which are used to predict recurrence, mortality risks and the benefit of adjuvant systemic therapy, can still lead to patients being unnecessarily treated with toxic therapies or not treated when they have a poor outcome [51]. More than 80% of patients with breast cancer receive adjuvant chemotherapy, although only approximately 40% of them will relapse and ultimately die of metastatic breast cancer [52]. Therefore new prognostic are needed to identify patients who are at the highest risk for developing metastases, which might enable oncologist to begin tailoring treatment strategies to individual patients [52]. For example, it is already known that prognosis for breast cancer patients with lymph node positive disease is poorer and that adjuvant systemic therapy decreases their risk of recurrence, but for patients with lymph node negative disease (LNN), the benefit of adjuvant systemic therapy is not so clear. Thus the ability to risk stratify LNN patients according to prognosis could provide important information for the patient and the treating oncologist [51].

Although many different prognostic indicators are in development, there are seven that are relatively well characterized, four of which have been specifically developed to address this question of prognosis in LNN patients (1, 3, 4, 5) [51]:

1. Amsterdam 70-gene profile
2. Genomic Grade index (GGI)

3. Recurrence Score (RS)
4. Rotterdam 76-gene signature
5. Wound response signature
6. Invasiveness gene set (IGS)
7. Intrinsic gene subtypes

The Amsterdam 70-gene profile (**1**) was first developed from supervised gene expression profiling analysis of frozen tumour samples from two distinct patient populations using the Agilent microarray platform (see fig. 1.10). All patients were <55 years of age and had lymph node negative disease but 34 of the 78 (44%) patients had distant metastasis within 5 years of completing treatment and 44 of the 78 (56%) patients did not develop distant metastasis within 5 year [53]. By comparing the gene expression profile of these two groups, a signature 70-gene set was identified that correlated with clinical outcome [51]. The hypothesis of van 't Veer and colleagues was that a tumour would have an intrinsic capacity to metastasize, regardless of size, and that this features could be captured by gene expression profiling [20]. The expression profile of these 70 genes classified the primary breast tumours as having either a **poor-prognosis** signature, which means they were likely to metastasize, or a **good-prognosis**, meaning that the development of metastases was unlikely [52].
Internal validation of the set indicated that it could accurately predict disease outcome for 65 of the 78 (83%) patients using the 70-gene signature [51]. The poor-prognosis signature included genes involved in the cell cycle, invasion and metastasis, angiogenesis and signal transduction. Interestingly, it also comprised genes that are almost exclusively expressed by the stromal cells that surround the epithelial cells in a tumour. For example, these include MMP1 and MMP9, which are required for ECM (ExtraCellular Matrix) degradation and tumour invasion [52]. The upregulation of genes that are highly expressed by stromal cells in a prognosis signature for breast cancer metastasis, and their defined role in invasion, again underlines the influence of the tumour microenvironment on tumour progression. Focusing on epithelial cells using microdissection to understand breast cancer progression and detection of prognostic markers could be not sufficient to provide a successfully prognostic gene signature [52]. External validation of the Amsterdam 70-gene prognostic indicator came from a retrospective analysis of 295 young patients (age <53 years) with both lymph node negative and lymph node positive disease, some of whom were included in the earlier trial [54]. The mean 5 year overall survival for the poor prognosis group of patients was 74% as compared to 97% for the good prognosis patients. In a second validation series by the TRANSBIG (TRANslating molecular knowledge into early breast cancer management building on the Breast International Group) consortium of 307 early-stage LNN breast cancer patients who did not receive adjuvant systemic treatment from 5 different European centers, the 70-gene signature holds up as an independent predictor outcome [20]. In these patients, the 70-gene prognostic indicator was better at predicting time to distant metastasis and overall survival compared to the clinical variables used by Adjuvant! Online. The TRANSBIG validation showed that 70-genes diagnostic test called Mammaprint® (Agendia, Amsterdam, The

Netherlands), exhibits a significant prognostic value (Hazard Ratio [HR] for distant metastasis: 2.32, 95% Confidence Interval [CI]: 1.35-4). The sensitivity to predict 10-years breast cancer death was 0.84 (0.73-0.92) and the specificity was 0.42 (0.36-0.48). While Adjuvant! Online exhibited a similar sensitivity 0.82 (0.71-0.90), its specificity was lower 0.29 (0.23-0.35) [55].

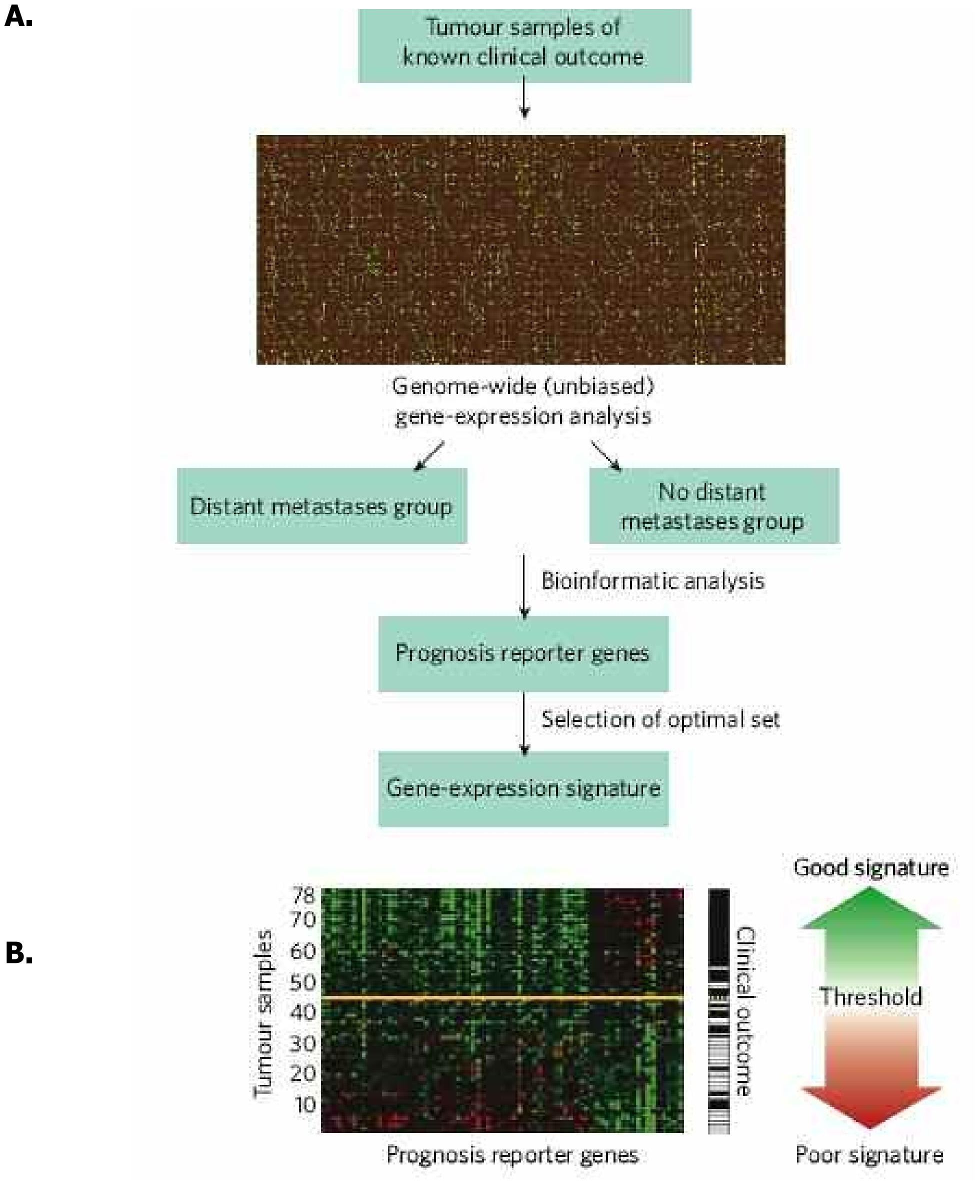


Figure 1.10: Predicting disease outcome by using gene expression test. **A.** Generating a prognostic gene-expression signature by using supervised classification. The gene expression of cells in a set of tumours of known clinical outcome is analysed by using whole-genome microarrays. The results for each tumour sample are then classified into two categories: tumours with a good outcome (no distant metastases developed) and tumours with a poor outcome (distant metastases developed). Using bioinformatic analysis, genes whose expression is significantly correlated with disease outcome are identified, and these are known as prognosis reporter genes. An optimal set of genes is then selected from the prognosis reporter genes by using bioinformatic algorithms and the pattern of expression of this multigene set is known as a gene-expression signature (or classifier). **B.** The gene-expression signature generated in a is shown as a 'heat map'. The expression of the 70 prognosis reporter genes selected as the optimal set (vertical columns) is shown for 78

tumours (horizontal lines). The outcome of the disease is shown on the right: white indicates metastasis; black indicates no metastasis; and yellow indicates the threshold for metastasis. (adapted from [44]).

These findings suggested that Mammaprint® could potentially increase the detection rate of patients with good prognosis, and thereby allow a decrease in the use of adjuvant chemotherapy. The clinical merit of Mammaprint® is the main research question of a large phase III randomized trial called MINDACT (Microarray In Node negative Disease may Avoid ChemoTherapy) [20]. Risk assessment for LNN breast cancer patients will be determined by the both 70-gene and clinical criteria (assessed by using Adjuvant! Online) (see fig. 1.11).

For patients with hormone receptor-positive breast cancer, who are designated as having a good prognosis both by Adjuvant!Online and genomic test, only adjuvant hormonal treatment is advised. In the case of a concordant high-risk assessment, patients will be advised to undergo chemotherapy and hormonal therapy for endocrine-responsive disease. The merit of the genetic predictive assay is tested in the third group, in which the clinical and genomic criteria are discordant. These patients will be randomized to either undergo adjuvant chemotherapy treatment based on genomic or based on clinical criteria [20]. If the main hypothesis of this trial (MINDACT) is validated, this study will be the first one to provide a level I evidence for a decrease in the indications of adjuvant chemotherapy [55].

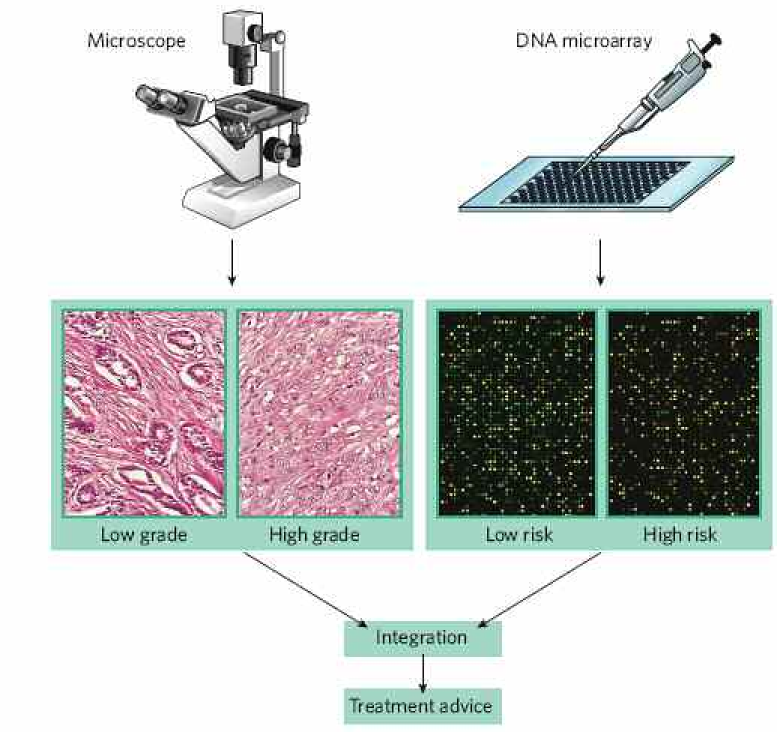


Figure 1.11: Conventional and molecular diagnostic testing for cancer. Conventional diagnostic tests rely heavily on morphological criteria to judge the aggressiveness of cancer, a process known as grading. More recently, multigene-expression tests (e.g. Mammaprint®) have been shown to be powerful tools for predicting disease outcome. One current challenge is how to integrate the knowledge obtained from these conventional tests and molecular diagnostic tests into a single recommendation for the oncologist treating the patient (adapted from [44]).

Two key biological processes captured by the 70-gene signature, proliferation and cell cycle, are also the main biological determinants of the Genomic Grade Index (GGI) (**2**). Sotiriou and colleagues [56] used microarray analysis as a tool to further subclassify the intermediate-risk group of histological grade II tumours [20]. These tumours, which represent 30 - 60% of cases, are the major source of inter-observer discrepancy and may display intermediate phenotype and survival, making treatment decisions for these patients a great challenge, with subsequent under- or over-treatment [55]. They developed a GGI score based on 97 genes, that were consistently differentially expressed between low and high grade breast carcinomas. The GGI, which essentially quantifies the degree of similarity between the tumour expression pattern of these 97 genes and tumour grade, was able to reclassify patients with histological grade II tumours into two groups with distinct clinical outcomes similar to those with histological grades I and III, respectively [55].

The 21 gene Recurrence Score prognostic indicator (Oncotype Dx™) (**3**) was developed using slightly different methods than those described above [57]. In this series of experiments, 250 candidate genes were selected from the published literature, genomic databases, and gene expression profiling experiments (using RT-PCR technique) and correlated with breast cancer recurrence in 447 patients [51]. From these 250 genes, 16 cancer-related genes and five reference genes were selected and their expression levels used to develop the Recurrence Score assay, which is unique in that it can be performed on Formalin-Fixed, Paraffin-Embedded (FFPE) tumour samples and does not require frozen sample [51]. This signature is a combined score of these 16 genes of interest, including ER, PR, Erb-B2 and Ki67 [55]. This predictor has allowed to identify a population (recurrence score < 18) that presents a very good prognosis (6.8% of 10-year distant metastasis), and could be spared from adjuvant chemotherapy. External validation of the 21 gene Recurrence Score came from the application of this prognostic indicator to patient samples collected in the large multicenter NSABP (National Surgical Adjuvant Breast and Bowel Project) B-14 trial, that examined the benefit of adjuvant tamoxifen in patients with hormone receptor-positive, LNN breast cancer [51]. As for the 70-gene signature, the process of clinical validation to achieve level I evidence is ongoing for this signature in a clinical trial named TAILORx (Trial Assigning IndividuaLized Options for Treatment (Rx)) [55].

Also another study [58] reported novel markers to predict distant metastasis risk and clinical outcome in patients with oestrogen-receptor-positive breast tumours treated with adjuvant tamoxifen [52]. Ma and colleagues identified three genes, HOXB13, IL17BR and CHDH, and the HOXB13:IL17BR ratio index in particular, that strongly predicted clinical outcome in breast cancer patients receiving tamoxifen monotherapy. In a subsequent larger independent patient cohort they showed that HOXB13:IL17BR index is a strong independent prognostic factor for ER+ node-negative patients irrespective of tamoxifen therapy [59].

The Rotterdam 76-signature (**4**) was specifically developed to address the clinical question of how to identify those patients with LNN breast cancer that would benefit from adjuvant systemic therapy, regardless of hormone receptor status, since these patients are cured with locoregional treatment (e.g. application of radiotherapy after mastectomy) [60]. 286 with LNN breast cancer that had not received adjuvant therapy were divided into ER− and ER+ groups and subjected to gene expression profiling. 115 patients served as the source of the training set data, from which a prognostic model was created by combining the 76 genes selected from the profiling experiments with ER status data. The remaining 171 mixed ER+ (75%) and ER− (25%) tumours served as the validation set. The sensitivity of the 76-gene test in predicting distant metastasis was 93%, and the specificity was 48%. In multivariate analysis of distant metastasis-free survival, the 76-gene prognostic indicator outperformed clinical variables and was the only significant variable to contribute to prognosis prediction. In a subsequent study, the Rotterdam 76-gene signature was also externally validated using a retrospective analysis of an independent data set of 180 LNN patients who did not receive adjuvant systemic therapy. The 76-gene signature was able to accurately identify poor prognosis patients (increased risk of distant metastasis within 5 years) versus good prognosis patients with a hazard ratio of 7.41 (95% CI 2.63–20.9). Only 16 patients had ER-negative disease, making generalizations to this subset difficult. In multivariate analysis of distant metastasis free survival, the Rotterdam 76-gene signature was the only factor significantly affecting prognosis [51]. Biological processes underlying this 76-gene signature are cell cycle, proliferation, cell death, DNA replication, and repair [20].

Huang and colleagues [61], with a different approach, identified aggregate pattern of gene expression, called metagenes, that were associated with lymph-node status and a 3-years-recurrence risk in breast cancer patient of all ages. Because of the small sample numbers, cross-validation is used to determine the accuracy of the 3-year-recurrence predictor instead of a second independent set of tumour samples, as previous studies. These metagenes, associated with lymph node status and recurrence, were capable of predicting outcomes in individual patients with about 90% accuracy. The metagenes defined distinct groups of genes, suggesting different biological processes underlying these two characteristics of breast cancer.

There are several prognostic signatures that are less clinically developed but are of interest. The Wound response signature (**5**) arose from the identification of core serum response (CSR) genes that changed expression levels when cultured fibroblasts were activated with serum [51, 62]. Evaluation of the CSR genes suggested that they represent important processes in wound healing like matrix remodeling, cell motility and angiogenesis, all of which are predicted to play a role in cancer invasion and metastasis [51]. Subsequent evaluation of the expression of these CSR genes in an external gene expression profiling data set generated from 295 patient samples used to validate the Amsterdam 70-gene profile, indicated that patients with tumours that expressed an activated wound response signature had a significantly decreased survival and increased probability of distant metastasis as compared to patients whose tumours expressed a quiescent wound response signature [62]. In addition, multivariate analysis of metastasis and death in this patient population indicated that the wound response signature was an independent predictor of prognosis. Genes associated with proliferation alone may also provide prognostic information within a subset of patients. The proliferation gene profile was derived from the Amsterdam 70-gene dataset [53, 54], in which investigators noted that outcome heterogeneity still existed within patient populations classified as having good and poor outcome signatures. They found that after stratification by ER expression and age, the expression level of a group of 50 cell cycle-related genes predicted outcome among those patients identified as having higher than expected ER expression levels for their age [63]. The proliferation signature is an example of a prognostic indicator that may play a role in a specific patient population. Expression levels of hypoxia-induced genes are also prognostic in early stage breast cancer [64]. While the independent contribution of this signature is not yet clear, it may be therapeutically relevant since we currently have no strategies for selecting appropriate patients for antiangiogenic strategies [51].

Recent reports have focused upon the genes associated with the putative cancer stem cell, which comprise less than 10% of the cells in breast cancer and are highly tumourigenic [51]. These cells are characterized by high expression of the cell surface marker CD44, which is implicated in cell adhesion, migration, and proliferation, and low expression of the less well-characterized CD24. Comparison of CD44+/CD24− cells with normal epithelial cells identified 186 genes associated with the tumourigenic cells, called the Invasiveness Gene Set (IGS) (**6**), which showed a prognostic value in both breast and other tumour types. Examination of the 295-patient Amsterdam dataset revealed that the IGS is prognostic independent of clinical characteristics, and appears to be particularly so among ER-positive or intermediate grade tumours. The IGS gene set overlapped little with other prognostic gene sets, and its impact was independent of the wound response signature [65].

Although the Intrinsic gene subtypes (**7**) described by Perou and colleagues (see 1.4.4) were not originally intended to function as prognostic indicators, they correlated with prognosis in the original population of 49 patients with relatively locally advanced tumours who had been treated with neoadjuvant doxorubicin on a clinical trial [24, 51]. Patients with the Luminal A subtype had the best prognosis as evaluated by overall survival (OS) and relapse-free survival (RFS) followed by Luminal B. Both the basal-like and Erb-B2+/ER− subtypes had the worst OS and RFS rates. Correlation of outcome with subtype in the independent Amsterdam dataset revealed a significantly longer time to development of distant metastasis among patients with Luminal A tumours compared to patients with basal-like or Erb-B2+/ER− tumours [30, 51].

In the table 1.5 are summarized the principals characteristics of the prognostic profiles described above.

Profile	Developed from	Technology	Validation	Rationale	Clinical use
Amsterdam 70 gene profile(Mammaprint®) [53, 54]	78 LNN pts, age<55 years, followed for >5 years	oligonucleotide microarray	independent training set	good signature is related to low metastasis risk; poor signature is associated with a high metastasis risk	predictor of distant metastasis in stage I-II; requires frozen tissue
Genomic Grade index 97-gene signature [56]	intermediate-risk group of histological grade II tumours	oligonucleotide microarray	independent training set	GGI reclassifies patients with histological grade II tumours into two groups	none at this time
Recurrence Score (Oncotype Dx™) [57]	candidate list of 250 genes applied to 447 pts with LNN and LNP disease, ER+ and ER-	RT-PCR	independent training set	likelihood of distant recurrence in tamoxifen-treated pts	predictor of distant relapse in pts ER+, LNN disease. Can be performed in fixed archival tissue
Rotterdam 76-gene signature [60]	115 pts LNN disease, no systemic neoadjuvant or adjuvant Rx, followed for > 5 years	oligonucleotide microarray	independent training set	good 76-gene signature versus poor 76-gene-signature for distant metastasis-free survival	predictor of distant metastasis-free survival in pts with LNN disease not treated with systemic therapy. Validated primarily in ER+. Requires frozen tissue.
Wound response signature [62]	identification of core serum response genes (446) expressed in serum-stimulated fibroblast	cDNA microarray	independent training set	expression of serum activated signature versus no expression to predict survival and distant metastasis	none at this time
Invasiveness gene set (IGS) [65]	identification of 186 genes that differentiate tumourigenic CD44+/CD24- cells from normal breast epithelium	oligonucleotide microarray	independent training set	expression of Invasiveness gene set related to metastasis-free survival and overall survival	none at this time
Intrinsic gene subtypes [23, 24]	gene list from unsupervised analysis, 49 pts with locally advanced disease, Rx neoadjuvant doxorubicin	cDNA	cross-validation	luminal A tumours have a better outcome than luminal B tumours. Worst outcome is for basal-like and Erb-B2 tumours	none at this time

Table 1.5: Prognostic profiles (for more details see text). pts:patients, LNN: lymph node negative, LNP: lymph node positive, Rx: treatment. Adapted from [51, 52].

1.5.2 COMPARISON OF PROGNOSTIC PROFILES

In order for a new prognostic or predictive assay to be clinically accepted it must be accurate, reproducible, feasible using clinical samples and it has to provide better information for clinical decision-making [51]. As described in the previous paragraph, there are currently several tools, each incorporating slightly different clinical and histopathologic variables into a prognostic model, available to the practicing oncologist to guide breast cancer treatment decisions. These conventional clinicalpathologic tools are useful but sufficiently inaccurate in predicting either good or bad outcomes, such that many patients are either undertreated or overtreated with adjuvant therapy.

Comparison of the Amsterdam 70-gene signature with the St. Gallen or NIH criteria reveals that the 70-gene signature assigns more LNN patients to the low risk prognosis group than either of the other two clinical indicators: 40% versus 15% versus 7% respectively [54]. Those patients identified as low risk by the 70-gene profile had a higher likelihood of metastasis-free survival than those identified as low risk by the other two methods, thereby indicating that use of the Amsterdam signature could still identify those patients with high risk disease while resulting in fewer patients being inappropriately treated. Comparison of the Amsterdam 70-gene signature to the Adjuvant! Online risk assessment also confirmed the added benefit of the 70-gene profile to clinical risk assessment. The additional benefit of this and similar genomic tools over conventional clinical-pathologic criteria is still controversial [66]. The Rotterdam 76-gene signature also appears to be superior to both the St. Gallen and NIH consensus criteria, with respect to being able to identify those patients with high risk disease, while reducing the numbers of patients with LNN disease unnecessarily exposed to the toxicity of adjuvant systemic therapy [60]. More specifically, 40% of patients classified as average or high risk patients by St. Gallen and 41% of patients classified as average or high risk by NIH would have been reclassified accurately as low risk using the 76-gene signature [67]. As this analysis suggests, these molecular profiling prognosticators will likely provide the most impact when applied in conjunction with clinical prognostic variables rather than instead of clinical variables.

Fan and colleagues [66] obtained a single dataset of 295 samples and applied five gene-expression-based models (intrinsic subtypes, 70 gene profile, wound response, recurrence score and the two-gene ratio) to compare the predictions derived from these gene sets for individual samples. They found that most models had high rates of concordance in their outcome predictions for the individual samples. In particular, almost all tumours identified as having an intrinsic subtype of basal-like, Erb-B2-positive and ER, or luminal B (associated with a poor prognosis), were also classified as having a poor 70-gene profile, activated wound response, and high recurrence score. The 70-gene and recurrence-score models, which are beginning to be used in the clinical setting, showed 77% to 81% agreement in outcome classification [66].

The most interesting observation to be made from the concordance of the different gene expression profile prognostic indicators with respect to predicting clinical outcome, is that there is little gene overlap between different prognostic signatures, i.e. they have few genes in common among all of them [66, 51]. For example the 70-gene and 76-gene signatures have only three genes in common, similarly the 70-gene and recurrence-score profiles overlapped by only 1 gene. This finding was interpreted by some to indicate that such gene-expression signatures are highly unstable [68, 44] but, because these gene-expression signatures could be independently validated in large groups of patients, it is more likely that different signatures use different genes to monitor the same biological processes. Although different gene sets are being used as predictors, they each track a common set of biologic characteristics that are present in different groups of patients with breast cancer, resulting in similar predictions outcome [66]. The ability to predict clinical outcome is not related to the expression of a specific and unique set of breast cancer-promoting genes, but there are a multiple gene sets within important pathways that can serve as correlates for the biological processes driving these tumours [68]. The overlap in gene identity among gene-expression profiles is not a good measure of reproducibility and the classification of individual samples is the relevant measure of concordance [66]. Some criticisms have been raised regarding the prognostic gene signatures reported so far. First, some argue that most of the signatures only add little information compared to a clinico-pathological score that would include ER, Erb-B2 and Ki67 in addition to the conventional clinical parameters. It must also be remarked that most of the genes included in the various published prognostic gene signatures are related to cell proliferation, and the question then arises as whether a simpler biomarker for such parameter like Ki67, which has been measured routinely for many years, could have provided similar results [55, 69]. However, gene expression profiling studies suggest that measuring proliferation with a more objective, automated and quantitative assay may be more robust than less quantitative assays such as immunohistochemistry [51].

Another criticism relates to the fact that most of the predictors were generated using a mix of molecularly heterogeneous tumours. Since breast cancer population is a mix of at least four different molecular classes [see 1.4.4] and oncogenic events are different across these subtypes, some have suggested that optimal predictors should be set up in each molecular class [70]. This was applied, for example, by Wang and colleagues who developed a 76-gene signature to identify patients at a high [60] risk of distant recurrence based on the prognostic genes separately identified in ER– and ER+ tumours. Desmedt and colleagues, in a recent meta-analysis of publicly available gene expression breast cancer data, showed also that proliferation is the strongest parameter predicting clinical outcome in the ER+/Erb-B2– subgroup of patients only, whereas immune response and tumour invasion appear to be the main biological processes associated with prognosis in the ER–/Erb-B2– and ERB-B2+ subgroups, respectively [55, 71]. This implies that the molecular background of the tumour should be taken into consideration to make prediction regarding prognosis [55].

While sensitivity of gene signatures for prognostic purpose looks good, there is still around 5–10% of patients who will present a metastatic relapse in the group predicted as low risk of relapse. Since the newer generation chemotherapy decreases by 30% the risk of breast cancer death in the whole population of breast cancer, it can be that an optimal chemotherapy could provide benefit even in this good prognostic population [55]. Nevertheless, it must be emphasised that most of the tumours classified as good prognosis are actually predicted to be resistant to chemotherapy [72]. Recently, it has been reported that some gene signatures present a strong time-dependency [73, 74]. This finding makes sense, since these signatures were built to predict the occurrence of metastases within the first 5 years, and are enriched in genes involved in cell proliferation. Thus predictors for metastatic relapse should be designed to predict both early and late relapses [55]. Several gene signatures for prognostic purpose have been generated at this time. At least two of them are being validated in prospective trials [53, 54, 57]. Although they allow an increase in the rate of patients who could be spared adjuvant chemotherapy while still correctly identifying the high-risk patients, they present some limitations that will have to be taken into account to generate more accurate 'second generation' gene signatures [55].

1.5.3 PREDICTIVE MARKERS OF RESPONSE TO NEOADJUVANT CHEMOTHERAPY IN BREAST CANCER

1.5.3.1 Neoadjuvant chemotherapy

NeoAdjuvant ChemoTherapy (NACT), also known as primary or induction chemotherapy, refers to administration of chemotherapy before locoregional treatment, with surgery and/or irradiation [75]. Since its initial use in the early 1970s [76], NACT has become the standard of care in the management of Locally Advanced Breast Cancer (LABC) primarily due to its ability to downsize large tumours. Lately NACT is increasingly being used for treatment of early-stage breast cancer.

However despite high response rates to NACT, a small proportion of patients fail to respond or even progress during therapy. The early identification of these non-responders assumes importance to plan alternative treatment options for such patients. Thus, biological markers that can reliably predict clinical or pathological response early during the course of treatment, have considerable clinical potential [75]. Clinically, neoadjuvant chemotherapy has several advantages:

- it can downsize large tumours, thus allowing breast-conserving surgery;
- it provides information on tumour response to a specific chemotherapeutic agent, allowing to investigate molecular determinants of chemotherapy response;
- it helps in achieving longer disease-free survival (DFS) and overall survival (OS), presumably through early treatment of systemic micrometastatic disease [77].

From biological point of view, justification for the evaluation of NACT derives from several hypothetical bases [75]:

- Almost all patients with LABC harbor micrometastases that can eventually lead to their death. Even about 10% patients with early breast cancer have circulating cancer cells in the peripheral blood [78]. It is now known that primary tumours contain cell variants destined to form metastases [75]. If these cells are forced into the circulation by local perturbation of the primary tumour, as would occur during surgery, they may then establish micrometastases elsewhere in the body. Surgery performed after effective chemotherapy would greatly reduce the chances of these clonogenic cells being disseminated in significant quantities and also take care of those already present in the circulation [75].

- There is evidence from animal models that surgical removal of the primary tumour can lead to enhanced metabolic activity in metastatic deposits with a risk of further dissemination [75]. Retsky et colleagues discussed the role of anti-angiogenic factors secreted by cancers; removal of the tumour may remove these inhibitory factors, inducing angiogenesis in the micrometastatic tumour bed leading to rapid growth and further dissemination of tumour cells [79]. It is suspected that such stimulated angiogenesis may occur in up to 20% of premenopausal node-positive breast cancer patients. Chemotherapy given preoperatively may help to counteract the stimulation of the growth of metastases by these tumour substances released into the circulation as a result of surgery [75].
- Successful early treatment with systemic therapy is consistent with the Goldie-Cold man hypothesis, whereby metastases are treated prior to the emergence of chemo-resistant clones [80]. They hypothesized that when a tumour cell population increases, the absolute number and also the percentage of drug resistant cells in the tumour increases, possibly because of spontaneous somatic mutations. With the enhanced proliferation of cells following primary tumour removal, it is likely that the number of resistant phenotypes in the metastatic population will increase. Hence, NACT should not only destroy cells made more sensitive by their kinetic alteration but also prevent cell proliferation and a consequent increase in the resistant cells [75].
- Also earlier systemic treatment of the primary tumour when it is still small, allows chemotherapy to make maximal use of tumour kinetics by attacking tumour cells when their proliferative activity is at the highest level [81].
- Finally NACT ensures better delivery of anticancer agents to the tumour because of an intact tumour vasculature [75].

The complete response to NACT is the ultimate goal of successful chemotherapy treatment. The response can be assessed clinically and/or pathologically. Clinical response is mainly evaluated by the conventional techniques of clinical measurement or imaging [75]. Reduction in the tumour size is a good indicator of response to NACT. Several guidelines to define tumour response have been proposed [82, 83]. According to these guidelines, responses can be classified as a complete response, partial response (the reduction of the tumour mass by at least 50%), stable disease (minor response or the increase of tumour volume of not more than 25%) and progressive disease (increase of more than 25% tumour volume or the appearance of new lesions). In general 60-90% of patients achieve clinical response to NACT. Complete pathologic remissions are, however, noted in only 3-30% of patients in most breast cancer trials [75]. Although few patients show mixed responses (response in the primary tumour and no response in the lymph nodes and vice versa) for most patients, the responses are similar in all sites of the tumour involvement [75]. Pathological complete response (pCR) is considered to be the most powerful predictor of outcome in terms of survival [84]. There are different systems for assessing pathological responses and this represents a limitation in comparing results across studies. Some studies defined pathologic complete response as no residual invasive cancer in the breast after neoadjuvant therapy and at the time of surgery [85], whereas other groups also take node status and noninvasive cancer into account [86]. Thus, the inconsistency in pCR definition should be considered when evaluating results from published neoadjuvant clinical trials.

The most important drugs used in the treatment of breast cancer are anthracyclines (e.g doxorubicin), methotrexate, cyclophosphamide and taxanes (docetaxel, paclitaxel) [87]. In advanced disease single agent therapy with these drugs leads to response rates between 20% and 60% with a relatively short duration of response of three to nine months. Combination chemotherapy, when compared to monotherapy, significantly increases both the response rates and duration of response rates for advanced stage breast cancer [87]. Anthracycline-based combination therapies have been shown to be more effective than methotrexate and taxanes add to the efficacy of anthracyclines [87, 88].

Several single biomarkers have been identified that predict response to NACT to a variable extent and some of them can be considered also with a prognostic power.

Tumour size: complete response was 50% for patients with tumours < 2 cm, 38% in T 2 - 4 cm and only 18% if T > 5 cm in size. It is also known that patients with large tumours and N2 nodes (higher stage breast cancer) are less likely to have a complete pathological response compared to tumours in lesser stage [75].

Hormone receptor status: oestrogen and progesterone receptors (ER, PR respectively) are hormone activated nuclear transcription factors that influence directly the mammary epithelial growth, differentiation, and survival [75, 87]. Hormone receptor status of a tumour is identified as an independent variable that is significantly associated with the likelihood of achieving pathologic complete response. pCR rates were significantly higher in patients with hormone receptor negative tumours. In particular there are several evidences that ER-negative tumours tend to respond better to chemotherapy than ER-positive tumours [75].

Tumour type and differentiation: a retrospective analysis of patients who received anthracycline-based NACT revealed that patients with invasive lobular carcinoma were less likely to achieve a pCR compared with patients with invasive ductal carcinoma [75, 89]. A possible explanation could be that histological and biological factors predicting a poor response to NACT (histological grade, ER, Ki-67 and p53 status) were more frequent in ILC than in IDC patients. Few other pathologic characteristics, such as poor differentiation and high nuclear grade also make a tumour more sensitive to NACT compared with tumours that are well differentiated [75].

Human epidermal growth factor receptor 2 (Erb-B2 o HER2) status: Erb-B2 (also referred as neu) belongs to the epidermal growth factor receptor (EGFR) family of tyrosine kinase receptors. Erb-B2 does not have its own ligand and forms heterodimers with other members of EGFR family [87, 90].Erb-B2 overexpression (15-25% patients with breast cancer) has been associated with poor outcome, especially in LNP patients. The association between Erb-B2 expression and the response to NACT is not so clear and published data are conflicting. An improved response to anthracycline-based NACT in Erb-B2 positive patients has only been demonstrated in some of the studies [87].

Topoisomerase IIα expression: there is a growing evidence that topoisomerase IIa (Top II) is a marker for anthracycline, and microtubule-associated protein tau (MAPT) for taxane sensitivity. The most commonly used drug in breast cancer, the anthracyclines, interact with the nuclear enzyme Top II. Top II reduces DNA twisting and super-coiling, allowing selected regions of DNA to untangle and thus engage in transcription, replication, or repair processes [75]. Due to the close location of Top II and Erb-B2 genes on chromosome 17, Top II gene aberrations (either amplification or deletion) are mainly associated with Erb-B2 gene amplification [75]. These observations have led to the hypothesis that Erb-B2 amplification is only a surrogate marker and the Top II amplification/overexpression could be the real predictive marker of response to anthracycline-based chemotherapy. Top II amplification is present in 5% of the total population, one third of Erb-B2 amplified tumours. It has been noted that, contrary to Erb-B2, where gene amplification is almost always correlated with protein expression, there was no correlation between Top II gene amplification and protein overexpression, as protein expression of Top II is highly cell-cycle dependent and associated with high proliferation [87, 91]. In literature there are controversial results respect to the link between Top II and the clinical response to neoadjuvant anthracycline-based chemotherapy, thus Top II deserves further testing in a prospective setting as a predictive marker [87].

Tumour proliferation Ki-67: uncontrolled proliferation is the key element of malignant transformation [75]. The MIB-1 (Ki-67) nuclear antigen is expressed in the G1 (gap 1), S (synthesis) G2 (gap 2), and M (mitosis) phases of continuously cycling cells, but is absent in G0 (quiescent phase) cells. Therefore, immunostaining with monoclonal antibody MIB-1 serves as a measure of cell proliferation. This index is the most practical method of monitoring cell proliferation [75]. Tumours with high cell proliferation should respond well to chemotherapy. Breast carcinomas with a high Ki-67 positive count, show improved response to chemotherapy in several studies and the Ki-67 expression is found to be decreased after NACT [87]. However, because other studies reported different findings, there is still not a consensus view on this marker [75].

Apoptosis related genes (p53, BAX and Bcl-2): it has been shown that many chemotherapeutic agents kill cancer cells by inducing apoptosis. Therefore, the proteins (p53 and Bcl-2 homologous family proteins) involved in the apoptotic pathway have been studied with a view to predicting chemoresponsiveness [75].

The p53 gene, the “guardian of the genome”, is a tumour suppressor gene. In contrast to the normal p53 gene product (protein), the mutated p53 gene product (protein) tends to accumulate in cell nuclei which can be detected by immunohistochemistry. It is mutated in at least 50% of human cancers ("the most mutated gene" in human cancer). The p53 encoded protein is involved in the apoptotic pathway, by inducing cell cycle arrest and initiating apoptosis. The use of p53 as a biological marker to predict response to chemotherapy, however, is still a controversial field of research [75].

The Bcl-2 gene encodes a 26-kDa protein involved in mainly inhibiting apoptosis. BAX, another protein of the same family of Bcl-2, shows functions closely with Bcl-2, but in the opposite manner, as it is a pro-apoptotic agent. It is reported that, either elevation of Bcl-2 or a reduction of BAX, may predispose to enhanced resistance to chemotherapeutic drugs. However also for these two markers, in the clinical context, the relationship between them and the chemoresponsiveness is still unclear [75, 92].

1.5.3.2 Gene expression profiling to identify predictive signatures

For many years, the research has focused on the identification of single markers predicting tumour response to chemotherapy. It is unlikely that the action of one or only a limited number of genes will confer chemotherapy resistance/responsiveness in breast cancer, since the pathways involved in tumour response to chemotherapy are complex and different between individual tumours [87]. Therefore, microarray technology, giving the possibility to analyze gene expression on a global scale, looks promising in the study of chemotherapy responsiveness of breast cancer. Since response to chemotherapy can be monitored in the neoadjuvant setting, these studies have been performed by giving chemotherapy preoperatively (fig. 1.12) [87].

The first indication that molecular profiling could predict chemosensitivity came from gene expression profiling experiments in cell culture lines where cell lines were classified as sensitive or resistant to a specific compound. Evaluation of 60 cell lines and 232 compounds revealed that 88 of 232 (38%) of profiles could accurately predict sensitivity or resistance to a given compound while only 12 of 232 (5%) of such profiles would be predicted to do so if the profiles were created by chance [61]. These data suggested that gene expression profiles differed between cells that were sensitive or resistant to a given drug, and that evaluation of these differences might be used in a predictive way [51, 93]. On the basis of this study, a number of efforts in identifying gene predictive of chemotherapy response in breast cancer tumours, have been published so far.

Ayers and colleagues [94] determined a 74 gene profile predicting response to neoadjuvant paclitaxel/fluorouracil plus doxorubicin plus cyclophosphamide in 24 patients. This profile could be validated in 18 additional patients with a predictive accuracy of 78% and a sensitivity of 43% [87]. They used frozen material and analyzed the gene expression profile with Affymetrix platform.

Chang and colleagues [95] correlated the expression of 92 genes and response to neoadjuvant docetaxel monotherapy in 24 patients and the prediction accuracy was 88%. In a subsequent study by the same group, comparison of the gene expression profiles of all tumours after docetaxel treatment (regardless of whether they were classified as sensitive or resistant), revealed that the profiles of tumours subjected to the selection pressure of docetaxel were relatively homogeneous. This observation suggested that those tumours originally sensitive to docetaxel may have developed resistance to the drug or led to selection of a resistant clone as evidenced by convergence of the gene expression profiles of sensitive and resistant tumours [51, 96]. Chang and group used frozen material to perform the microarray experiment with a cDNA based-platform.

Other studies did not see statistical significant differences in gene expression between patients with pCR compared with patients without pCR. Hannemann and colleagues [97] compared the gene expression profiles of 48 patients either treated with six courses of doxorubicin/cyclophosphamide (AC) or six courses of doxorubicin/docetaxel (AD). They reported that tumours that did not respond to chemotherapy showed minimal changes in overall gene expression profile before and after therapy, whereas tumours showing a partial remission show major changes in gene expression after treatment [87, 97]. Hannemann and colleagues used frozen material to perform the microarray experiment with a cDNA based-platform.

Another study by Gianni and colleagues correlated the expression of 384 genes with pCR following paclitaxel and doxorubicin based NACT in patients with LABC [98]. Differently from previous studies, RNA was extracted from the pretreatment formalin-fixed paraffin-embedded core biopsies and the experiments were performed using both RT-PCR and microarray. They found 86 genes correlated with pCR, that was more likely with higher expression of proliferation-related genes and immune-related genes, and with lower expression of ER-related genes [75].

Recently, several other signatures have been published.

Thuerigen and colleagues identified a signature of 512 genes predicting response to neoadjuvant gemcitabine/epirubicin/docetaxel with a overall accuracy of 88% in a validation set of 48 patients [99]. They started from frozen material and analyzed the gene expression profiling with 21K oligo microarray platform.

Dressman and colleagues described a set of 38 genes that predicted response to chemotherapy containing doxorubicin/paclitaxel in 36 patients [100]. Some of the genes identified have been previously linked to breast cancer outcome and metastasis. They also started from frozen material and analyzed the gene expression profiling with Affymetrix microarray platform.

The largest study so far comes from Hess and colleagues which included 133 patients with stages I-III breast cancer who all received preoperative weekly paclitaxel and 5-fluorouracil, doxorubicin, and cyclophosphamide (T/FAC) chemotherapy [101]. They used 82 patients to develop a 30-gene molecular predictor of pathologic complete response and the remaining 52 patients were used to assess the accuracy of the predictor.

The test misclassified one of the patients who achieved pCR (12 of 13) and one of those who had residual cancer (27 of 28) in the validation set. It showed significantly higher sensitivity than a clinical variable-based predictor including age, grade, and ER status (92% vs. 61%). The high sensitivity indicates that the test correctly identified almost all of the patients (92%) who actually achieved pathologic complete response. However, to what extent this genomic predictor of sensitivity is specific to T/FAC therapy, rather than being a generic marker of chemotherapy sensitivity, is yet to be determined. In this study, the combination of genomic and clinical information was not significantly better than using the 30-gene predictor alone. Hess and colleagues used FNA (Fine Needle Aspiration) specimens for the microarray analysis with Affymetrix platform.

Another approach is to identify gene expression patterns of response or resistance *in vitro* and then transfer these results to the clinical settings [87]. One recent study describes two gene sets correlating with *in vitro* resistance to doxorubicin or mitoxantrone, respectively [87, 102]. The EORTC [103] (European Organization for Research and Treatment of Cancer) clinical phase III trial published a substudy recently involving 212 patients with ER-negative breast cancer treated with either a non-taxane regimen (fluorouracil, epirubicin, and cyclophosphamide [FEC]) or a taxane regimen (docetaxel followed from epirubicin plus docetaxel [TET]). The RNA, extracted from sections of frozen biopsies was hybridized to Affymetrix microarrays. In vitro single agent drug sensitivity signatures were combined to obtain FEC and TET regimen-specific signatures. The regimen-specific signatures significantly predicted pathological complete response in patients treated with the appropriate regimen. The FEC predictor had a sensitivity of 96% and specificity of 66%, the TET predictor had a sensitivity of 93% and specificity 69%.

Gene expression profiling is promising not only for predicting sensitivity to chemotherapy, it has also been used to predict sensitivity to endocrine therapies like tamoxifen. Although current histopathologic evaluation of breast cancer tumours involves determination of ER status which, in general, correlates with response to endocrine therapy, a large percentage of patients with ER+ disease will display *de novo* resistance to endocrine therapy or will develop resistance over time. As reported above (see 1.5.1) the Recurrence Score identifies those most likely to develop distant metastases despite adjuvant tamoxifen [51, 57].

Jansen and colleagues, performing an unsupervised gene expression profiling analysis of microarrays created from ER-positive tumours, also revealed a 44-gene signature that correlated with tamoxifen resistance in 77% of patients [105]. Clinical ER status correctly predicts response to tamoxifen in only 50–60% of patients. This tamoxifen resistance profile is undergoing independent validation. As expected, functional analysis of the gene signature revealed a large number of genes known to be regulated by oestrogen, although genes involved in apoptosis and extracellular matrix remodeling were also detected [51]. Between all these predictive signatures there is hardly any overlap, indicating that there may be not only one profile, but that several combinations of probes may predict response to chemotherapy with the same accuracy [87]. All studies described above have aimed to predict sensitivity to taxanes, anthracycline based regimens or a combination of both. The predictive profiles identified so far seem very promising , although they are based on studies with a relatively small sample size and none of them have been implicated into clinical routine.

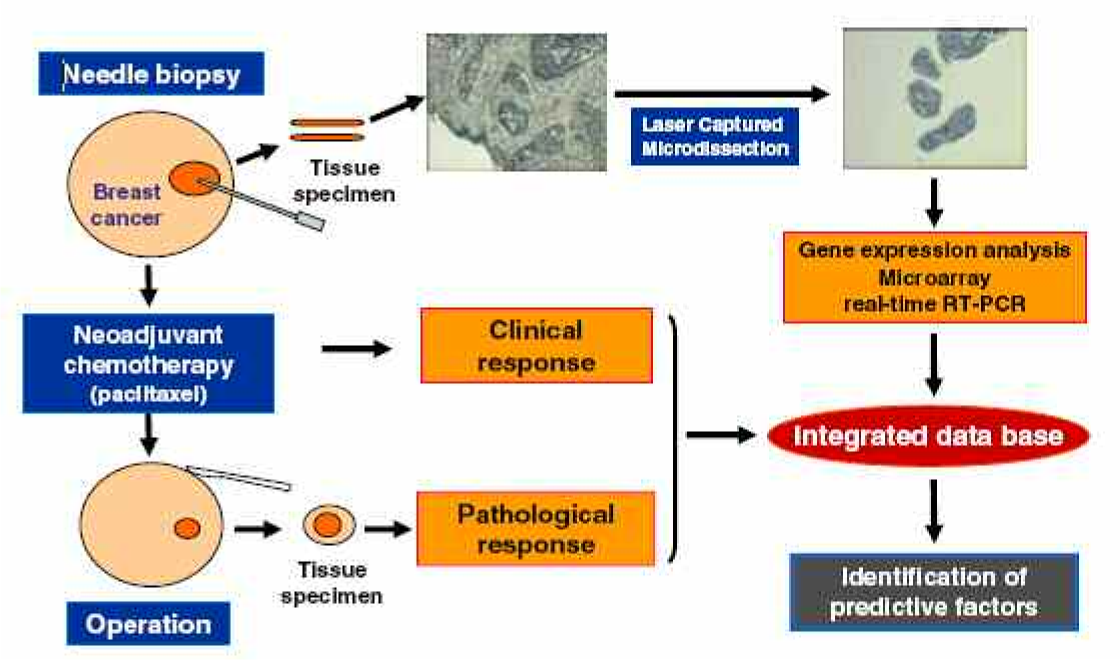


Figure 1.12: Study of identifying genetic markers for NeoAdjuvant ChemoTherapy (NACT). Some studies can combine the Laser Captured Microdissection (LCM) (see for details 1.4.4) method with the microarray technologies. Tumour cells were selectively collected by LCM to exclude most of the stromal tissues to analyze cancer cells and assessed gene expression analyses. Clinical and pathological responses are evaluated at completion of treatment. Differentially expressed genes were selected for discriminating between not-responder and responder (adapted from [104]).

It is emerging that responses to anticancer drugs are more difficult to predict by using molecular tests than prognosis is [44]. One of the main reasons for this difficulty is that resistance to anticancer agents can result from a variety of mechanisms and it might result from subtle mutations that do not cause evident changes in gene expression [44]. Therefore, although these molecular profiles provide interesting information about sensitivity to commonly used breast cancer drugs, all of these assays require further validation. It is also important to consider the differences in term of tumour sizes, patient populations and methodologies, when a comparison between each others is done.

1.5.4 LONG TERM PERSPECTIVE: WILL MICROARRAYS BE STILL USEFUL IN THE FUTURE?

Although most of the predictors previously discussed will be determined either in a central laboratory (Oncotype DX) and/or using dedicated array (MammaPrint®), it must be emphasised that DNA arrays offer the major advantage of being able to provide genome-wide gene expression measurements [55].

It has also to be remarked that DNA arrays can be used for many other purposes than to predict prognosis and treatment efficacy. Indeed, this technology has been used to generate molecular predictors for the diagnosis of malignancy, organ-specific metastases, lymph node involvement, as well as the identification of activated pathways and the expression of therapeutic targets [55]. Moreover recent technological advances have made it possible to measure gene expression in single cells [106, 107] allowing to evaluate gene expression profile of circulating tumour cells, cancer stem cells or some other cells of interest, including mesenchymal stem cells and endothelial progenitor cells [55].

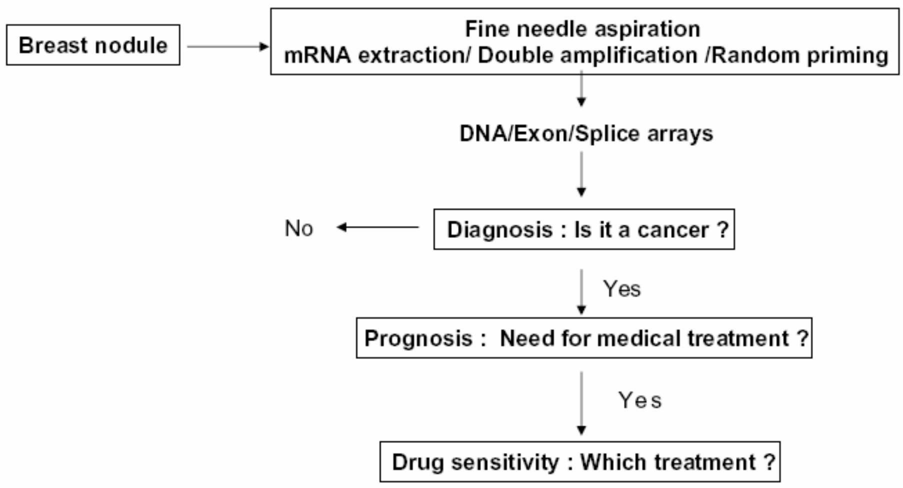

Breast nodule
Fine needle aspiration
mRNA extraction/ Double amplification /Random priming
DNA/Exon/Splice arrays
No
Diagnosis : Is it a cancer ?
Yes
Prognosis : Need for medical treatment ?
Yes
Drug sensitivity : Which treatment ?

Figure 1.13: Potential of molecular predictors for breast cancer in the next decades. This figure reports how gene signature will be integrated in the clinical practice in the future (adapted from [55]).

This suggests that besides prognostic and predictive information related to chemotherapy and endocrine therapy, DNA arrays could allow in the future to determine all the information needed for optimal patient's care. Since the current trend is to increase the number of independent bioassays or analyses to be done in a single sample (conventional pathology, immunohistochemistry, FISH, RT-PCR), the use of DNA arrays could change this tendency by doing all analyses in a single experiment in certified laboratories. In addition to facilitate logistics, decrease time before treatment decision, this approach would probably save costs by decreasing the rate of bioassays to be done [55]. Several new arrays [108], including Splice Arrays, are detecting both gene expression and splicing events, thereby providing a more functional picture of genomic program in every single patient.

The figure 1.13 shows how genome-wide DNA arrays could affect patient's management in the next decades.

1.6 ROADMAP FOR DEVELOPING AND VALIDATING THERAPEUTICALLY RELEVANT GENOMIC CLASSIFIERS

As reported so far, microarray expression profiling provides an exciting new technology for relating tumour gene expression to patient outcome, but it also provides increased challenges for translating initial research findings into robust diagnostics that benefit patients and physicians in therapeutic decision making. Development of biomarker classifiers useful for improving treatment decisions and sufficiently validated for broad clinical application is difficult, and more difficult for expression signature classifiers. In a recent paper of Simon *et al.* they tried to define some of the key steps in obtaining a classifier reliable and potentially useful in clinical setting. These characteristics are summarized in the table 1.6 [109].

Key Steps in Development and Validation of Therapeutically Relevant Genomic Classifiers
Develop classifier for addressing a specific important therapeutic decision Patients are sufficiently homogeneous and receiving uniform treatment so that results are therapeutically relevant Treatment options and costs of mis-classification are such that a classifier is likely to be used
Perform internal validation of classifier to assess whether it appears sufficiently accurate relative to standard prognostic factors that it is worth further development
Translate classifier to platform that would be used for broad clinical application
Demonstrate that the classifier is reproducible
Independent validation of the completely specified classifier on a prospectively planned study

Table 1.6: Key Steps in Development and Validation of Therapeutically Relevant Genomic Classifiers (from [109]).

A multigene expression signature classifier is a function that provides a classification of a tumour based on the expression levels of the component genes. The gene sets identified as associated with outcome tend to be unstable because gene groups are correlated by co-regulation and the stringent criteria used for identifying differentially expressed genes results in reduced statistical power for gene selection. It is often much easier to develop a classifier that performs accurately than it is to identify exactly the optimal gene set [109]. Although it would be desiderable to understand the mechanistic relationship of the components of an expression signatures (the genes), the classifier can be validated without such understanding, also because a clear biologic interpretation may be more difficult to achieve than an accurate classification. As Simon reports, the concept of validation of these "new" gene classifiers is different compared to the concept of validation of "traditional" disease biomarkers. For example, an expression signature should be developed to predict outcome for a well-defined set of patients (training set) who receive a well-defined therapy. The signature classifier would be developed using data from such patients and would be validated for an independent set of such patients. The "developmental study" would identify the genes into a completely specified classifier that can be used and potentially validated in a subsequent study. The validation does not consist of seeing whether the same genes are prognostic in the subsequent study. The validation should be focused on addressing whether the application of the previously defined classifier to a new set of patients results in clinical benefit [109].

Many algorithms have been used effectively with DNA microarray data for class prediction. A linear discriminant is a function

$$l(x) = \Sigma_{i \in F} w_i x_i$$

where xi is the expression measurement for the gene, wi is the weight given to that gene, and the summation is over the set F of features (genes) selected for inclusion in the classifier. For a two-class problem, there is a threshold value c that must be defined; a sample with expression profile defined by a vector x of values is predicted to be in class 1 or class 2 depending on whether l(x) as computed from the equation is less than or greater than c [109]. Many types of classifiers used in the literature have the form shown in the preceding equation. They differ with regard to how the weights are determined. These classifiers include for example Fisher's linear discriminant analysis and diagonal discriminant analysis or support vector machines. When the number of genes (p) is greater than the number of cases (n), perfect separation of a training set is always possible with a linear classifier. In fact, there are an infinite number of linear classifiers that achieve perfect separation. That suggests that there may not be sufficient information in most datasets to effectively utilize nonlinear classifiers. Although complex nonlinear classifiers are popular, there is little evidence that they perform any better than simpler methods [109].

Since most classifiers do not use all of the genes whose expression is measured, one step is determining which genes to include in the classifier; this process is called feature selection. The number of "informative genes" is usually small compared to the number of "noise genes". Including too many noise genes can mask the informative genes and reduce the accuracy of prediction. It is sometimes possible to distinguish different cell types based on expression levels of a small number of genes, very differentially expressed in the two cell types. However, this is often not the case for more difficult classification problems. In these situations, there may be a dozen or more differentially expressed genes, but the fold differences in expression may not be large and it is difficult to identify these genes from among the thousands of noise genes. It also important to remark that omitting informative genes from a classifier has a greater deleterious effect on classification accuracy than does inclusion of noise genes, if the number of noise genes included is not too great [109]. We can divide genomic classifier studies into **developmental studies** and **validation studies**.

Developmental studies are the ones that first develop the classifiers. It is delicate to evaluate whether a genomic classifier is promising based on a developmental study. The difficulty derives from the fact that the number of candidate genes available for use in the classifier is much larger than the number of cases available for analysis. In such situations, it would be always possible to find classifiers that accurately classify the data on which they were developed. Consequently, even in developmental studies, the validation on data not used for developing the model is necessary. This internal validation is usually accomplished either by splitting the data into two portions, one used for training the model and the other for testing the model, or some form of Cross Validation (CV) based on repeated model development and testing on random data partitions [109]. The most straightforward method of estimating the accuracy of future prediction is the split-sample validation, nevertheless the cross-validation approach is a valid alternative when the number of cases is small (and would be difficult to split the dataset in training and test sets). A type of cross-validation procedure is the Leave One Out CV (LOO-CV). LOO-CV starts like split-sample cross validation in forming a training set of samples and a test set. The test set consists of only a single sample; the rest of the samples are placed in the training set. The sample in the test set is placed aside and not utilized at all in the development of the class prediction model. Using only the training set, the informative genes are selected and the parameters of the model are fit to the data. The process is repeated leaving each of the n biologically independent samples out of the training set, one at a time. During the steps, n different models are created and each one is used to predict the class of the omitted sample. The number of prediction errors is totalled and reported as the leave-one-out crossvalidated estimate of the prediction error [109].

Often the initial developmental study is not large enough to estimate the positive and negative predictive values of the test with sufficient precision to determine whether the test has real clinical utility. It is important that the clinical use of the classifier be carefully considered in planning the external **validation study** so that these performance characteristics can be adequately estimated. The objective of external validation is to determine whether use of a completely specified diagnostic classifier for therapeutic decision making in a defined clinical context results in patient benefit. An independent validation study could be a prospective clinical trial in which patients are randomly assigned to treatment assignment without use of the classifier versus treatment assignment with the aid of the classifier [109].

From this overview it is emerging that the steps needed to translate research findings of correlations between gene expression and prognosis into clinical diagnostic tests, are neither easy nor immediate. Nevertheless they are necessary to move genomic signatures into clinical application as therapeutically relevant and robust diagnostics.

1.7 DRUG RESISTANCE MECHANISMS

In general, systemic drugs are active at the beginning of therapy in 90% of primary breast cancers and 50% of metastases. This is demonstrated by reduced tumour volume, improved symptoms and decreased serological tumour markers [110]. However, after a variable period, the tumours can become resistant to the chemotherapic drugs, with a consequent failure of the treatment. Chemotherapy resistance is a major problem in the management of breast cancer, therefore detecting drug resistance before first line chemotherapy may increase the patient's survival.

Resistance to therapy is caused by a genetic amplification or by point mutations in specific genes. In the genic amplification the treatment regimen with a single systemic agent selects a group of cancer cells that is increasingly resistant to therapy, decreasing the rate of response to further therapy [110]. The genic amplification could also due to the action mechanism of the drug itself. Some of them showed mutagen properties that accelerate the resistance development. Also drugs that reduce the progression of cellular cycle during the DNA synthesis (e.g. Topoisomerase inhibitor) can increase the level of genic amplification. The resistance due to point mutations is not established gradually but in a single phase and it is stable.

There are two general classes of resistance to anticancer drugs: those that impair delivery of anticancer drugs to tumour cells (intrinsic resistance), and those that arise in the cancer cell itself due to genetic and epigenetic alterations that affect drug sensitivity (acquired resistance). In the intrinsic resistance, impaired drug delivery can result from poor absorption of orally administered drugs, increased drug metabolism or increased excretion, resulting in lower levels of drug in the blood and reduced diffusion of drugs from the blood into the tumour mass. Recent studies have remarked the importance of the tumour vasculature and an appropriate pressure gradient for adequate drug delivery to the tumour [111]. Since some cancer cells, that are sensitive to chemotherapy as monolayer cells in culture, become resistant when transplanted into animal models, it was supposed that environmental factors, such as the extracellular matrix or tumour geometry, might be involved in drug resistance. Much remains to be learned about this type of drug resistance and its role in clinical oncology.

In the acquired resistance, cancer cells in culture become resistant to a single drug, or a class of drugs with a similar mechanism of action, by altering the drug cellular target or by increasing repair of drug-induced damage, frequently to DNA.

After selection for resistance to a single drug, cells might also show cross-resistance to other structurally and mechanistically unrelated drugs — a phenomenon that is known as MultiDrugResistance (MDR).This might explain why treatment regimens that combine multiple agents with different targets are not always more effective [110].

Interestingly, a study showed that cells at early stages of tumourigenesis process were able to develop doxorubicin-resistant derivates [112]. This demonstrates that, at least in this cell model system, the ability to acquire drug resistance is not a consequence of the accumulation of mutations that occur during the proliferation of a transformed cell, but it is an intrinsic characteristic that appears before the complete set of genetic transforming alterations [113].

However MDR is highly relevant to protection against mutation and cancer, since these same mechanisms will not only exclude anticancer drugs (many of which are mutagenic and carcinogenic), but also a wider range of endogenous mutagens and carcinogens froma range of cells [114].

As illustrated in fig. 1.14, different types of cellular multidrug resistance have been described.

Resistance to natural-product hydrophobic drugs — sometimes known as classical multidrug resistance — generally results from altered expression and/or activity of ATP-dependent efflux pumps with broad drug specificity. These pumps belong to a 48 family of ATP-binding cassette (ABC) transporters that share sequence and structural homology. So far, 48 human ABC genes have been identified and divided into seven distinct subfamilies (ABCA–ABCG) on the basis of their sequence homology and domain organization [111]. The increased drug efflux lowers intracellular drug concentrations with consequent resistance. Drugs that are affected by classical multidrug resistance include the vinca alkaloid (vinblastine and vincristine), the anthracyclines (doxorubicin and daunorubicin), the RNA transcription inhibitor actinomycin-D and the microtubule-stabilizing drug paclitaxel [111].

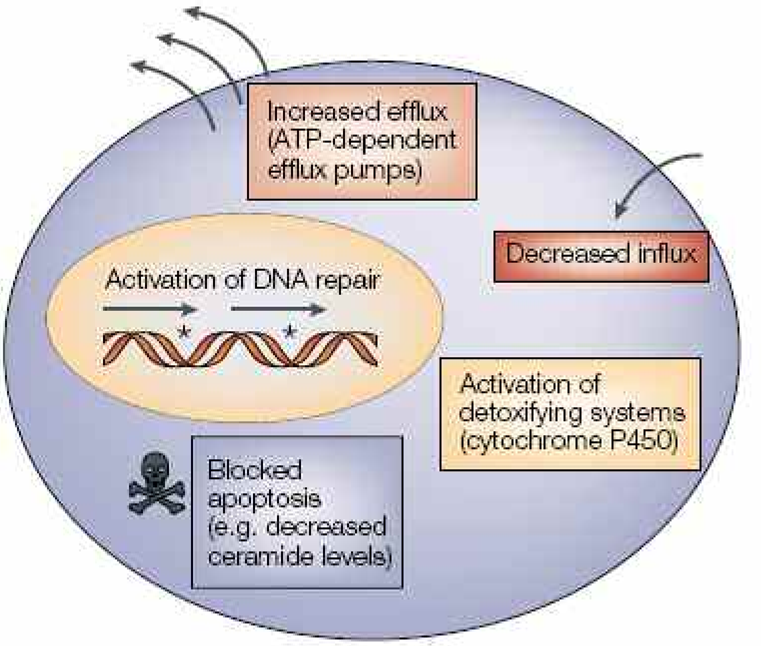

Increased efflux (ATP-dependent efflux pumps)
Decreased influx
Activation of DNA repair
Activation of detoxifying systems (cytochrome P450)
Blocked apoptosis (e.g. decreased ceramide levels)

Figure 1.14: Cellular factors that cause drug resistance. Cancer cells become resistant to anticancer drugs by several mechanisms. One way is to pump drugs out of cells by increasing the activity of efflux pumps, such as ATP-dependent transporters. Alternatively, resistance can occur as a result of reduced drug influx — a mechanism reported for agents that 'piggyback' on intracellular carriers or enter the cell by means of endocytosis. In cases in which drug accumulation is unchanged, activation of detoxifying proteins, such as cytochrome P450 mixed-function oxidases, can promote drug resistance. Cells can also activate mechanisms that repair drug-induced DNA damage. Finally, disruptions in apoptotic signaling pathways (e.g. p53 or ceramide) allow cells to become resistant to drug-induced cell death [110].

It is known that cancer cells can survive in a toxic environment by pumping drugs out of their cytoplasm. *In vitro* assays show increased expression of ATP transporter proteins as **P-gp** (encoded by the **MDR1 gene**), **MultiDrug Resistance associated Proteins** (**MRPs**) and **Breast Cancer Resistance Protein** (**BCRP**) in the membranes of cells grown under cytotoxic conditions [87].

MDR1 gene transfection into drug sensitive cells results in P-gp overexpression, decreased drug accumulation and in the MDR phenotype [87]. P-gp expression has been detected in many human cancers and several of the clinically applied drugs in cancer treatment are substrates for P-gp mediated transport, e.g. docetaxel, doxorubicin and epirubicin [87].

Two inhibitors that are used in the laboratory and in clinical trials that attempted to reverse drug resistance are the calcium channel blocker verapamil and the immunosuppressant cyclosporin A [111]. It is still unclear whether P-gp expression plays a role in drug resistance in breast cancer in clinical setting. Studies that compared mRNA levels in untreated and treated samples indicated a subtle though significant increase of P-gp expression after anthracycline-based chemotherapy. It could be supposed that higher levels of P-gp expression should result in an increased cellular tolerance for anthracyclines and a decreased tumour response. P-gp could best be used as a predictor of response in case of an association between P-gp levels in samples taken before anthracycline treatment [87]. However, at present there are not enough evidences to suggest a role for MDR1/P-gp in the responsiveness on breast cancer to chemotherapy.

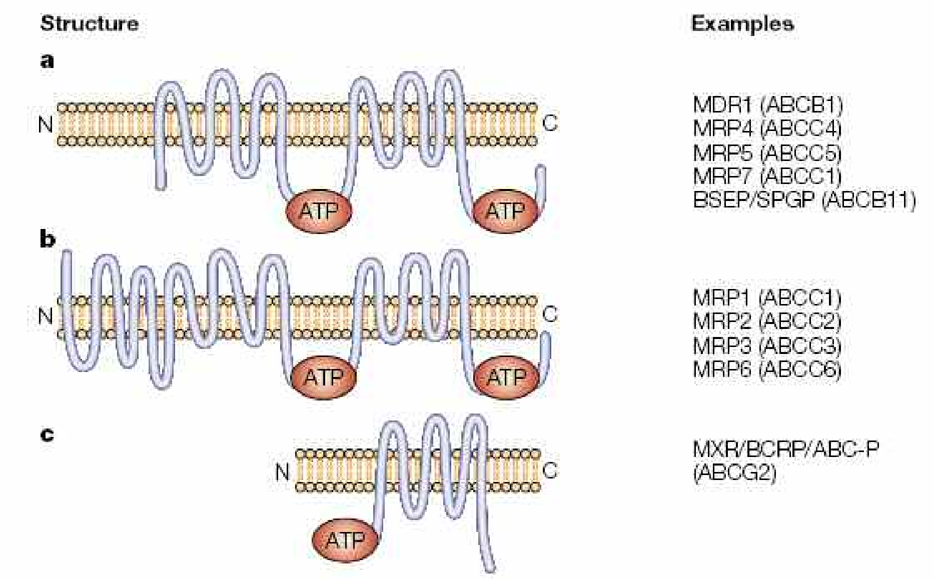


Figure 1.15: Structures of ABC transporters known to confer drug resistance. The structures of three categories of ABC transporter. **a.** ABC transporters such as multidrug resistance MDR1 and multidrug-resistance-associated protein 4 MRP4 have 12 transmembrane domains and two ATPbinding sites. **b.** The structures of MRP1, 2, 3 and 6 are similar in that they possess two ATPbinding regions. They also contain an additional domain that is composed of five transmembrane segments at the amino-terminal end, giving them a total of 17 transmembrane domains. **c.** ABCG2 contains six transmembrane domains and one ATP-binding region ("half-transporter") — in this case, on the amino-terminal side (N) of the transmembrane domain. In other half-transporters, such as the transporter associated with antigen processing (TAP), the ATP-binding cassette is found on the carboxy-terminal (C) side of the transmembrane domain. Half-transporters are thought to homodimerize or heterodimerize to function (from [111]).

Since not all multidrug-resistant cells express P-gp, the research led to the discovery of the Multidrug-Resistance-associated Protein 1 (MRP1, or ABCC1) [111]. MRP1 is similar to P-gp in structure, with the exception of an aminoterminal extension that contains five-membranespanning domains attached to a P-gp-like core (fig. 1.16). As for P-gp, the literature is not conclusive about the role of MRP1 in drug resistance in the clinical setting for breast cancer. The discovery of MRP1 stimulated a genomic search for homologues, leading to the identification of eight additional members of the ABCC subfamily of transporters (1.15 b).

The Breast Cancer Resistance Protein (BCRP) -also known as MXR (mitoxantrone-resistance gene) or ABC-P (ABC transporter in placenta)- was identified using breast cancer cell lines without overexpression of P-gp and MDR1. This transporter is a homodimer of two half-transporters, each containing an ATP-binding domain at the amino-terminal end of the molecule and six transmembrane segments (fig. 1.16). The protein is located in the cytoplasmic membrane of cells and in *in vitro* experiments it confers resistance to a variety of drugs used in cancer treatment, including anthracyclines [87]. So far, a role for BRCP as a predictor of response to chemotherapy in the clinical has not been demonstrated.

The **Lung Resistance Protein** (**LRP**), not belonging to the ABC transporters family, it is expressed at high levels in drug-resistant cell lines and some tumours [115]. LRP is a major vault protein (vaults proteins are large ribonucleoprotein particles present in all eukaryotic cells) found in the cytoplasm and on the nuclear membrane. Although their role in normal physiology is not yet established, vaults might confer drug resistance by redistributing drugs away from intracellular targets [111].

Resistance can also be mediated by reduced drug uptake. Water-soluble drugs that are carried on transporters and carriers that are used to bring nutrients into the cell, or agents that enter by means of endocytosis, might fail to accumulate without evidence of increased efflux. Examples include the antifolate methotrexate, nucleotide analogues, such as 5-fluorouracil and 8-azaguanine, and cisplatin [111].

Multidrug resistance can also result from activation of coordinately regulated detoxifying systems, such as DNA repair and the cytochrome P450 mixed function oxidases. Indeed, it has been showed a coordinate induction of the multidrug transporter P-glycoprotein (P-gp) and cytochrome P450 3A [116]. This type of multidrug resistance can be induced after exposure to any drug.

Resistance can also result from defective apoptotic pathways. This might occur as a result of malignant transformation; for example, in cancers with mutant or
non-functional p53 . Alternatively, cells might acquire changes in apoptotic pathways during exposure to chemotherapy, such as alteration of ceramide levels or changes in cell-cycle machinery, which activate checkpoints and prevent initiation of apoptosis [111].

From a molecular point of view, there are several cellular pathways that influence the drug resistance. In certain cancer types, expression of the Raf/MEK/ERK pathway can modulate the expression of drug pumps and anti-apoptotic molecules such as Bcl-2 [117]. Studies reported that ectopic expression of Raf will increases the levels of both the MDR-1 drug pump and the anti-apoptotic Bcl-2 protein in breast cancer cells. The increased expression of MDR-1 and Bcl-2 most likely occurs by a transcriptional mechanism by downstream target kinases of the Raf/MEK/ERK pathway inducing the phosphorylation of transcription factors, which bind the promoter regions of MDR-1 and Bcl-2 and stimulate transcription [117]. Also the PI3K/PTEN/Akt pathway shows effects on the drug resistance of breast cancer cells. Expression of Akt conferred resistance of the MCF-7 breast cancer cell line to 4HT, a drug commonly used to treat ER+ breast cancer patients. PTEN activity or lack thereof can also regulate drug resistance in breast cancer [117].

An important principle in multidrug resistance is that cancer cells are genetically heterogeneous. Although the process that results in uncontrolled cell growth in cancer favours clonal expansion, tumour cells that are exposed to chemotherapeutic agents will be selected for their ability to survive and grow in the presence of cytotoxic drugs. These cancer cells are likely to be genetically heterogeneous because of the mutator phenotype. So, in any population of cancer cells that is exposed to chemotherapy, more than one mechanism of multidrug resistance can be present. This phenomenon has been called Multifactorial Multidrug Resistance.

Detailed knowledge about the causes and mechanisms of drug resistance might make it possible, in the future, to predict effectively the response of a human cancer to chemotherapy. Once all the main causes of drug resistance have been catalogued and molecular probes have been defined, it should be possible to determine their expression in individual cancer cells, even specific mechanisms of resistance expressed in a subpopulation of cells [111]. In the last years genomic-wide analysis methodologies (Comparative Genomic Hybridization, SNP-arrays, expression microarrays) are improving our ability to determine which drug-resistance and drug-metabolizing genes are upregulated in different tumours, and these results can then be correlated with clinical responses to specific types of chemotherapy (see 1.5.3.2).

As emerged from the general overview above reported, drug resistance is a complex and dynamic phenotype. Unraveling the basic mechanisms giving rise to this multifactorial phenomenon and translating these finding in the design of novel therapeutic strategies in the clinic, is the next challenge that both scientists and clinicians have for the near future [113].

1.8 CHEMOTHERAPIC DRUGS USED IN THIS STUDY

1.8.1 PACLITAXEL

Paclitaxel (commercial name Taxol$^{®}$) is an antimicrotubule agent belonging to the taxanes class of antineoplastic compounds, with established antitumour activity in a variety of cancers including breast cancer, ovarian cancer, lung cancer and Kaposi sarcoma [118]. Interest in the taxanes began in 1963 when a crude extract of bark from the Pacific yew, *Taxus brevifolia*, was shown to have broad antitumour activity in preclinical tumour models. In 1971 was identified paclitaxel as the active constituent of the bark extract. The search for taxanes led to the development of docetaxel, that is a synthetic derivate of an inactive taxane precursor [119].

Microtubules are composed of polymers of tubulin in dynamic equilibrium with tubulin heterodimers composed of alpha and beta protein subunits. Although their principal function is the formation of the mitotic spindle during cell division, microtubules are also involved in many vital interphase functions, including the maintenance of shape, motility, signal transmission, and intracellular transport. Unlike other antimicrotubule drugs, such as vinca alkaloids, which induce the disassembly of microtubules, paclitaxel promotes the polymerization of tubulin. At subnanomolar concentrations, paclitaxel inhibits the disassembly of microtubules, whereas it increases their mass and numbers at higher, albeit clinically achievable, concentrations. The microtubules formed in the presence of paclitaxel are extraordinarily stable and dysfunctional, thereby causing the death of the cell by disrupting the normal microtubule dynamics required for cell division and vital interphase processes. Paclitaxel binds to the N-terminal 31 amino acids of the beta-tubulin subunit in the microtubule, rather than to tubulin dimers. In intact cells, paclitaxel induces the bundling of microtubules, which may be an useful clinical correlate of a lethal drug effect, and the formation of large numbers of asters of mitotic spindles. It also enhances the cytotoxic effects of ionizing radiation in vitro, possibly by inducing arrest in the premitotic G2 and mitotic phases of the cell cycle, which are the most radiosensitive phases [120]. The efficacy of intravenous paclitaxel as adjuvant therapy for early breast cancer has been investigate in two large randomized trials; it was administered sequentially to standard doxorubicin-cyclophosphamide (AC) combination therapy and compared with cycles of AC alone [118]. In both trials, the addition of sequentially administered paclitaxel to the AC regimen, significantly improved disease-free survival at 5 years compared with AC alone. In one of the trials, women who received paclitaxel also had a significant improvement in overall 5-year survival time. In randomized trials of neoadjuvant therapy for women with early breast cancer, paclitaxel or paclitaxel-containing regimens showed efficacy in terms of response/remission rates, local breast tumour recurrence and proportion of patients eligible for breast-conserving surgery [118].

Two principal mechanisms of acquired resistance to the taxane have been characterized. First, some tumours contain alpha- and beta-tubulin with an impaired ability to polymerize into microtubules and have an inherently slow rate of microtubule assembly that is normalized by the taxanes [120]. Therefore a possible mechanism for paclitaxel resistance involved alterations in microtubule dynamicity [121]. Such alterations include mutations in alpha- or beta-tubulin that affect lateral/longitudinal interactions in the microtubule lattice and/or the binding of regulatory proteins that could result in more dynamic microtubules (stathmin and MAP4). In paclitaxel-resistant cells, stathmin, a microtubule-destabilizing protein, can not bind microtubules because of a mutated alpha-tubulin. It has been noted that stathmin is up-regulated in breast carcinoma cell from patients with more aggressive disease, similar to the increase observed in paclitaxel-resistant cells. One hypothesis is that the alpha-tubulin mutation may alter the binding of stathmin to alpha-tubulin, and combined with the increased protein expression of stathmin, lead to hypostable microtubules (stathmin sequesters tubulin dimers). The paclitaxel-resistant cells would be less affected by the hyperstabilizing of paclitaxel because of the combined effects of the alpha-tubulin mutation and the stathmin changes [122]. The decrease of expression of MAP4, a microtubule-stabilizer protein, in addition with the increased levels of stathmin, could also cause an additional destabilization of the microtubule network in paclitaxel-resistant cell lines [122].

A second mechanism involves the amplification of membrane phosphoglycoproteins that function as drug-efflux pumps (ABC transporters, see 1.6) [120]. The MDR phenotype of tumour cells confers cross-resistance to various structurally bulky natural products, including anthracyclines, etoposide, vinca alkaloids, colchicine and taxanes.

The upregulation of caveolin-1, a membrane component involved in small molecule transport and intracellular signaling, has also been found to be related to taxane resistance [110].

Recently, Rouzier and colleagues, using microarrays to identify genes associated with pCR to preoperative paclitaxel-containing therapy in breast cancer patients, found that the microtubule-associated protein Tau was the most significantly differentially expressed gene [123]. Tau mRNA expression was low in cases with pCR. Tau protein promotes tubulin polymerization and stabilizes microtubules. They reported that down-regulation of tau increased sensitivity of breast cancer cells to paclitaxel. Their data suggested that low tau expression increases the "vulnerability" of microtubules to paclitaxel and makes breast cancer cells hypersensitive to this drug. Low tau expression may be used as a marker to select patients for paclitaxel therapy. Inhibition of tau function might be exploited as a therapeutic strategy to increase sensitivity to paclitaxel [123].

1.8.2 DOXORUBICIN AND EPIRUBICIN

Doxorubicin (commercial name Adriamycin®) is an anticancer agent belonging to the anthracyclines antibiotic class, one of the most commonly used classes of anticancer drugs, used in clinical practice since the 1960s. Doxorubicin, together with daunorubicin, was the first anthracyclin in clinical use, poduced by the *Streptomyces* species [124]. Although the development of the second-generation synthetic anthracyclines (e.g. idarubicin or epirubicin), doxorubicin still remains the most widely used in lymphoma, leukemia, sarcoma and breast cancer treatment.

The molecular target of doxorubicin is type II DNA topoisomerase enzymes (Top II) that control and modify the topological states of DNA. The mechanisms of these enzymes involve DNA cleavage and strand passage through the break, followed by religation of the cleaved DNA; the precise manner by which these events occur in a single cell is the source of intense research [124]. In mammalian cells, these enzymes have been differentiated, based on their mechanistic and physical properties, into two types, type I and type II. In contrast to that of Top I, the function of Top II is ATP dependent. Once Top II binds to duplex DNA, nucleophilic reactions sequentially cleave the two complementary strands of DNA four base pairs apart, and the resulting 5'-phosphoryl groups become covalently linked to a pair of tyrosine groups, one in each half of the dimeric Top II enzyme [124]. Once the double-strand break has been made, the cleaved ends must be moved apart and a second double-strand segment of DNA passed through the break. Once strand passage is complete, the cleaved DNA is religated. Two Top II isoforms have been identified in humans: the alpha-form (α) and the beta-form (β). The α form is encoded by a single-copy gene located on chromosome 17 and the β form has been mapped to chromosome 3. The drugs that have Top II as molecular target inhibit religation of DNA cleaved by Top II and induce protein-linked breaks in the DNA. When drug is removed, these breaks are reversible. All mammalian Top II inhibitors are DNA intercalators that insert a planar moiety between two adjacent base pairs in duplex DNA. The anthracyclines induce formation of covalent topoisomerase-DNA complexes, and prevent the enzyme from completing the religation portion of the ligation-religation reaction [124]. The anthracyclines interact with DNA TopII complex in a sequence-specific manner, since they can stimulate the DNA cut only on specific sites and not on all sites recognized from Top II enzyme [125]. The anthracyclines are also

DNA intercalators that insert part of their planar structures between two adjacent base pairs in DNA, causing single-stranded and double-stranded breaks. These agent can undergo chemical reduction through enzymatically catalyzed or iron-catalyzed pathways to yield reactive free-radical intermediates, that can cause oxidative damage to cellular proteins [125]. Although the anthracyclines are associated with all these reactions, it is their interaction with Top II the most important mechanism of cytotoxicity. The DNA damage induced from anthracyclines leads to transcriptional activation of protooncogenes (c-fos, c-jun) and oncosuppressors (p53) with consequent apoptosis (if the damage is too heavy).

The mechanisms of resistance in anthracyclines are two:

- over-expression of the MDR1 gene (see 1.6);
- under-expression or mutation of the gene that encodes TopII α enzyme, that shows a reduction of its activity and sensitivity. For example, is known that the overexpression of heat shock protein 27 (HSP27) in a variety of cancer (e.g. breast, ovarian) is related to doxorubicin resistance. The HSP27 overexpression inhibits doxorubicin-induced apoptosis by decreasing the expression of Top II. Since paclitaxel it was reported to suppress HSP27 expression, a combination with doxorubicin can sensitize breast cancer cells with HSP27 overexpression to doxorubicin [126].

Kubo and colleagues reported that the point mutations of the topoisomerase II alpha gene do not have an essential role in drug resistance, instead the alteration of the ABC efflux-pumps activity play a key role in this phenomenon [127].

Most of the published data support the preferential use of an anthracycline-containing adjuvant regimen for individuals with Erb-B2 positive tumours [128]. As previously reported (1.5.3.1) due to the close location of Top II and Erb-B2 genes on chromosome 17, Top II gene aberrations are mainly associated with Erb-B2 gene amplification [75]. Therefore it was suggested that Erb-B2 amplification could be only a surrogate marker and that the Top II amplification/overexpression the real predictive marker of response to anthracycline-based chemotherapy. It was also noted that there is a positive response to preoperative doxorubicin treatment in patients Erb-B2 positive, without Top II amplification, then Top II amplification can not be the only explanation of this chemosensitivity [129].

The anthracyclines are associated with both acute and chronic cardiac toxicity and strategies to minimize this toxicity have been proposed, including patient selection on the basis of preexisting cardiac risk, monitoring of cardiac function during treatment and early management of cardiac dysfunction [131]. The mechanisms of this cardiotoxicity include enzymatic-mediated formation of oxygen free radicals that initiate lipid peroxidation and a nonenzymatic pathway for free radicals formation [127]. The use of less cardiotoxic anthracyclines may be a strategy to reduce the risk of cardiotoxicity. Liposomal doxorubicin products offer similar efficacy compared with conventional doxorubicin and have been successfully used in combination with trastuzumab in the metastatic and neoadjuvant setting in Erb-B2-positive breast cancers [128].

Epirubicin (commercial name Ellence™ or Pharmorubicin™), a semisynthetic derivative of doxorubicin, is less cardiotoxic than doxorubicin and it has been combined with paclitaxel in the treatment of metastatic breast cancer with effective results [130]. Similarly to doxorubicin, epirubicin acts by intercalating DNA strands. Intercalation results in complex formation which inhibits DNA and RNA synthesis. It triggers DNA cleavage by topoisomerase II, resulting in mechanisms that lead to cell death. Epirubicin is favoured over doxorubicin, the most popular anthracycline, in some chemotherapy regimens as it appears to cause fewer side-effects. Epirubicin has a different spatial orientation of the hydroxyl group at the 4' carbon of the sugar, which may account for its faster elimination and reduced toxicity.

2 AIM

Microarray profiling technology provided biological evidence for the heterogeneity of breast cancer, since different expression patterns can be identified within distinct tumour groups. Expression array studies of breast cancer described genes associated with histology, grade and hormonal receptor status but now, the most exciting challenge of microarrays to breast cancer research is the development of prognostic profiles and predictive signatures of responsiveness to chemotherapic treatment.

All eligible women are often treated in the same manner even though *de novo* drug resistance results in treatment failures in many breast cancer patients. The administration of ineffective chemotherapy increases mortality and decreases quality of life in cancer patients. This emphasizes the need to evaluate every patient's probability of responding to the chemotherapic treatment, limiting the drugs used to those most likely to be effective.

Expression profiling to identify new predictive signatures and markers of response has been applied to tumours treated with a number of different standard neoadjuvant systemic therapy regimens (treatment given before surgery) and the response to NeoAdjuvant ChemoTherapy (NACT) was used to test the efficacy of the treatment.

The study reported here was conducted to evaluate the use of gene expression profiling to predict the response to a specific neoadjuvant chemotherapy regimen based on paclitaxel and anthracyclines (doxorubicin and epirubicin) drugs. Gene expression profiles of pre-treatment breast tumour biopsies were correlated with the clinical response to the treatment in order to develop a gene predictive signature of response to chemotherapy.

3 METHODS

Abbreviations

AF = autoclaved and filtered
BSA = bovine serum albumin
Cy3 = 5-NN'-diethyl-tethrametylindocarbocyanine
Cy5 = 5-NN'-diethyl-tethrametylindo di carbocyanine
CH1, CH2 = two microarray channels
cDNA = complementary DNA
ds DNA = double strand DNA
ss DNA = single strand DNA
DEPC = diethylpyrocarbonate
DMSO = dimethylsulfoxide
EDTA = ethylenediaminetetraacetic acid
mQ H_2O= purified water with the Milli RO 15 (Millipore) system
DEPC H_2O= mQ water treated with DEPC
min. = minutes
nt = nucleotides
dNTPs = 3'- deoxynucleotide triphosphates (dATP, dGTP, dCTP, dTTP)
o.n. = over night
pb = base pair
PBS = phosphate buffered saline
PVP = polyvinylpyrrolidone
mRNA = messenger RNA
SDS = sodium dodecyl sulfate
SSC = sodium chloride-sodium citrate
sec. = seconds
Tris = tris(hydroxymethyl)aminomethane

3.1 BUFFERS AND SOLUTIONS

SSC 20 X

175,3 gr	NaCl
88,2 gr	sodium citrate
800 ml	H_2O

Bring to pH 7 with NaOH 10 N.
Add water to 1 litre

Pre-hybridization buffer

500 µl	SSC 20 X
20 µl	SDS 10%
200 µl	ss DNA 2 µg/µl
200 µl	Denhardt's solution 50 X
1180 µl	H_2O mQ AF

Hybridization buffer

500 µl	SSC 20 X
20 µl	SDS 10%
50 µl	ss DNA 2 µg/µl
500 µl	formamide
930 µl	H_2O mQ AF

aRNA fragmentation

10 X Fragmentation Reagents: 200 ml Zn salt buffered solution
Stop Solution: 200 mM EDTA pH 8.0

Post-hybridization wash

1st wash: 2 ml SSC 20X, 800 µl SDS 10% in 40 ml H_2O mQ AF
2th wash: 200 µl SSC 20X, 800 µl SDS 10% in 40 ml H_2O mQ AF
3th wash: 400 µl SSC 20X in 40 ml H_2O mQ AF
4th wash: 200 µl SSC 20X in 40 ml H_2O mQ AF

Denhardt's solution 50X

1 gr BSA;
1 gr Ficoll;
1 gr PVP;
100 ml H_2O mQ A.

3.2 DESIGN OF THE STUDY

The project of this thesis was performed at CRIBI (*Centro di Ricerca Interdipartimentale per le Biotecnologie Innovative*), in the Biology Department of the University of Padova in collaboration with AB ANALITICA srl, Oncology and Surgery sections of Dolo, Mirano and Noale Hospitals.
The project was approved and granted from MIUR (Ministry of University and Scientific and Technological Research). Written informed consent was obtained from all patients. The patients are free to leave the study at any time, without any consequences for theirself.

3.2.1 SELECTION OF PATIENTS

Patients were eligible for the study with the following criterions:

- histological diagnosis of mammary carcinoma,
- stage II breast cancer ≥ 2cm in diameter, clinically defined and confirmed with mammography and ecography,
- no previous chemotherapic or radiotherapic treatment,
- ≤ 70 years-old (but there are some patients ≥ 70 years-old),
- standard medullary (neutophils > 2000; platelets >150.000), epatic and renal activity,
- standard PAO (patients with hypertension could be included but with PAO under control),
- negative pathological medical history for heart disease,
- no other diseases under way (except the breast neoplasy),
- no abnormality detected in chest X-ray, liver X-ray and electrocardiogram

3.2.2 SAMPLING AND NEOADJUVANT CHEMOTHERAPY

The patients underwent a diagnostic biopsy or a fine needle aspiration biopsy to obtain a histological diagnosis of breast cancer. Estrogen receptor (ER), Progesteron receptor (PR), c-erbB-2 (HER2), p53, Ki67 and Bcl-2 were assessed using ImmunoHistoChemistry (IHC). Another biopsy was taken, fresh-frozen in liquid nitrogen (- 196 °C) and used for RNA isolation and microarray expression analysis (see 3.5).
After the biopsy, the patients received four courses of Doxorubicin 60 mg/mq and Taxol 175 mg/mq (AT) or Epirubicin 60 mg/mq and Taxol 175 mg/mq (ET), every three weeks. At each course the response to chemotherapy was evaluated, if there was not a reduction of the tumour diameter after two courses, the patient went to the surgery, as well if the diameter of the tumour increased after the first course. After the first course was done a clinical exam and, if it is necessary, an instrumental check. After the second and the fourth course were done mammography and ecography.

3.2.3 HISTOLOGICAL CLASSIFICATION AND IMMUNOHISTOCHEMICAL ASSESSMENT

Microscopic slides containing 3-4 µm sections were stained with commercially available antibodies: ER (cone 1D5), PR (clone PgR 636), c-erbB-2 (clone CB11), Ki67 (clone MIB1), Bcl-2 (clone 124), p53 (clone BP53-12).
All tissue sections were reviewed by one pathologist for histological classification and immunohistochemical assessment. Samples were scored as ER, PR, c-erbB-2, Ki67, p53 positive by IHC when at least 10% of the tumour cells showed staining of the receptors or proteins. Specimens with ≥ 25% of the cells stained for Bcl-2 were considered positive. A sample was scored as being HER2 positive (Erb-B2 positive) when a membrane staining (indicated as 3+ or 2+) could be observed by IHC, using HercepTest^{TM} (DAKO); the sample was scored as being HER2 negative when a membrane staining could be not detected (indicated as 1+).

3.2.4 RESPONSE EVALUATION AFTER COMPLETION OF NEOADJUVANT CHEMOTHERAPY

The response of the primary tumour to chemotherapy was evaluated after the first and the fourth course of chemotherapy with mammography and ecography, as reported above and by pathological examination following surgery after the completion of chemotherapy.
The clinical responses to chemotherapy could be:
cCR clinical Complete Response (residual tumour mass < 25%, used instead of pCR when the pathological examination is not available);
PR Partial Response (residual tumour mass < 50% and ≥ 25%);
NC No Change (residual tumour mass > 50%);
PD Progressive Disease (increase of tumour mass).
After the pathological examination there were two possible responses:
pPR pathological Partial Response (residual tumour mass < 50%)
pCR pathological Complete Remission (absence of residual vital tumour cells at microscopy).

3.3 MICROARRAY SYNTHESIS

Microarray expression was analyzed using the Operon 70 mer oligos collection (Human Version 2.0) containing 21.521 oligonucleotides spotted in duplicate on MICROMAX glass slides- SuperChip I provided by PerkinElmer Life Sciences Inc. (Boston, USA). This set consists of oligonucleotides designed on Human Unigene clusters, mainly in the 3'-terminal region.

3.3.1 PREPARATION OF THE OLIGONUCLEOTIDES

The synthetic oligonucleotides were delivered lyophilized in 384-well formats in polypropylene Microarray plates (Genetix) and were re-suspended in 15 µl Micro Spotting Solution (Teleken); this buffer ensures a good quality of the spots and helps the binding of the oligos to the microarray slide. Before the resuspension, the plates are centrifuged at 3500 rpm for 1 minute in order to to gather the oligos on the bottom of the tube, avoiding possible contaminations during the aluminium cover removal. The microarray plates are agitated for 12 h at 4°C at 900-1000 rpm, until a complete re-suspension of the oligos.

3.3.2 PRINTING OF THE OLIGONUCLEOTIDES

Oligos (probes) were printed on the microarray slide using Biorobotics Microgrid II spotter (fig. 3.1), it is designed for high throughput sample handling and it could work with 96, 384 or 1536 well microplates. Up to 24 microplates at a time can be accommodated in the BioBank loading cartridge and a maximum of 10 BioBanks (240 microplates) may be programmed into a single run. The spotter can keep up to 120 slides at a time.

Figure 3.1: Microgrid II biorobotics

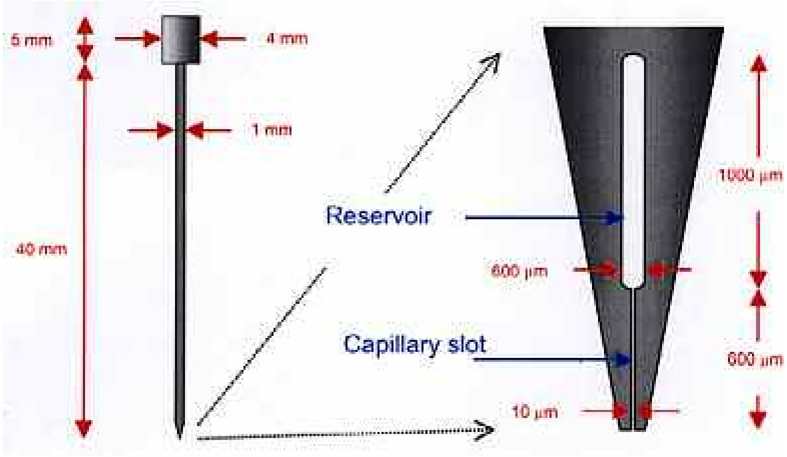


Figure 3.2: MicroSpot pin tip (http://www.arrayit.com/)

The metal pins (fig. 3.2) of the printhead load the oligos from the plate (fig. 3.3), then transfer them on the slide making the spots, with a diameter between 40 μm and 80 μm and spaced each other of 100 μm. The Operon collection was spotted using a printhead loaded with 48 pins.

The slides used are the MICROMAX Glass Slides, SuperChip I (Cat No MPS696) (Perkin Elmere), 25mm x 75mm x 1mm, uniformly coated with aminopropylsilane. During array spotting, positively charged primary amine groups bonded to the glass surface react with the negatively charged sugar phosphate backbone of acid nucleic molecules (fig. 3.4); coupling takes place at or near neutral pH.

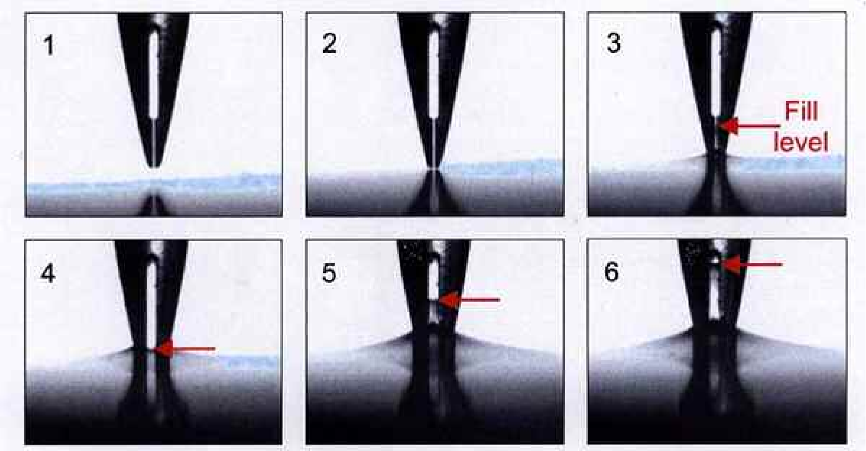


Figure 3.3: Pins refilling by capillarity (http://www.arrayit.com/)

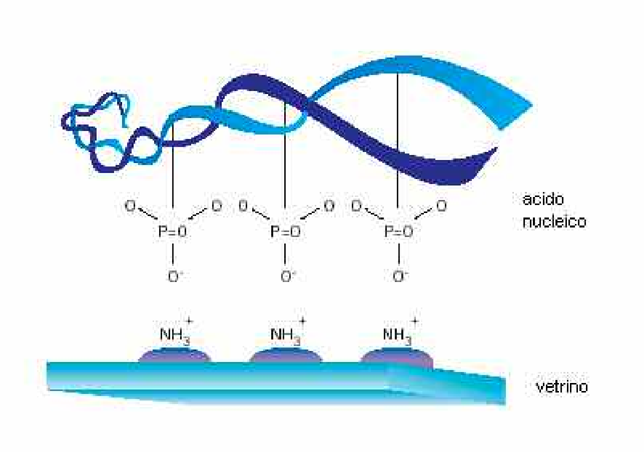


Figure 3.4: Aminopropylsilane surface of the microarray slide (adapted from Amersham, 2002)

A relative humidity near 50% has to be kept inside the spotter, in order to avoid that the solution evaporates from the pins and the plates, before the completion of the oligos transfer on the glass slide. If the humidity goes over 70% there could be the spot swelling.

The air coming into the spotter is forced through the HEPA (High Efficiency Particulate Air) filters that are able to retain the dust; indeed it could block the pins and, because it is fluorescent, alters the microarray image resolution. It is very important to keep the slides in airtight boxes. After use, the microplates are stored at -20°C and re-used for others transfer cycles; it allows to have already the re-suspended oligos solutions for next times.

After every transfer the pins are washed with H_2O mQ and dried with a vacuum pump.

After the completion of probe transfer, the slides were left into the spotter to dry out in a controlled humidity, for 30 min.; then the slides were rehydrated at room temperature for 5 min. This treatment causes the swelling of the spots, that become translucent and show a more homogeneous distribution of the probe.

After the drying of the spots, they show salt crystals left from the Micro Spotting Solution, used to check, by scanning with a confocal laser scanner, if the microarrays are correctly printed.

Then the oligos were linked to the slide surface by UV irradiation (300 mJ) with the Stratalinker 1.800 (Stratagene) (fig. 3.5).

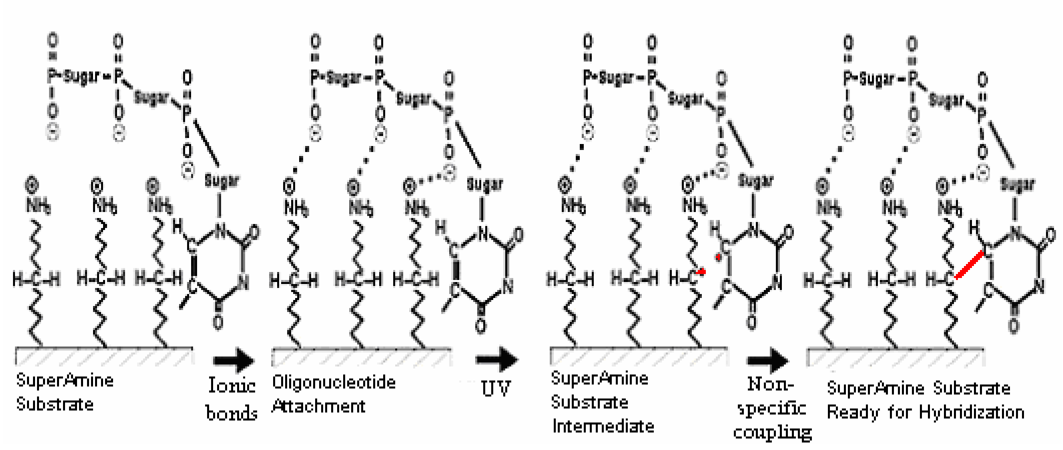

SuperAmine Substrate
Ionic bonds
Oligonucleotide Attachment
UV
SuperAmine Substrate Intermediate
Non-specific coupling
SuperAmine Substrate Ready for Hybridization

Figure 3.5: UV binding of the probe to the microarray slide (http://www.arrayit.com/)

Subsequently the slides are washed following this procedure:

- wash with SDS 1% and SSC 3X in mQ H_2O: shake vigorously to remove the probe weakly bound that could interfere with the target hybridization;
- wash twice with mQ H_2O for 5 minutes at room temperature.

Finally the slides were dried by centrifugation for 1 min. at 500 x g in ALC 4237R centrifuge. Printed slides are stored in the dark, at room temperature and are stable for 6-12 months.

3.4 RNA EXTRACTION

Total RNA was extracted using TRIzol^{TM} (Gibco BRL), a mono-phasic solution of phenol and guanidine isothiocyanate; the reagent is an improvement to the single-step RNA isolation method developed by Chomczynski and Sacchi [131]. During sample homogenization, TRIzol^{TM} reagent maintains the integrity of the RNA, while disrupting cells and dissolving cell components.
Before to start with the RNA extraction the frozen samples were weighed and cut with a pestle in a mortar. It is very important to keep the tissue frozen using liquid nitrogen to avoid RNA degradation.
The RNA extraction consists of five steps:
1. Homogenization: homogenize tissue samples in 1 ml of TRIzol^{TM} per 50-100 mg of tissue using homogenizer ultra-turrax-t8 (IKA$^{®}$ - WERKE).
The homogenizer tips were washed prior to use it:

o 10 min. NaOH 0.5 M
o min. H_2O DEPC
o 3 min. H_2O DEPC
o 2 min. H_2O DEPC

and between the homogenization of different samples inserted in the homogenize and immersed in the solutions:

o 10 sec. NaOH 0.5 M

- o 10 sec. Tris-HCl 0.1 M pH 7.5
- o 10 sec. H_2O DEPC
- o 20 sec. H_2O DEPC
- o 30 sec. H_2O DEPC

2. Phase separation: incubate the homogenized samples for 5 minutes at 15 to 30°C to permit the complete dissociation of nucleoprotein complexes. Add 0.2 ml of chloroform per 1 ml of TRIzol^{TM}. Shake tubes vigorously for 15 seconds and incubate them at 15 to 30°C for 2 to 3 minutes. Centrifuge the samples at no more than 12.000 × g for 15 minutes at 2 to 8°C. Following centrifugation, the mixture separates into a lower red, phenol-chloroform phase, an interphase, and a colorless upper aqueous phase. RNA remains exclusively in the aqueous phase. The volume of the aqueous phase is about 60% of the volume of TRIzol^{TM} used for homogenization.

3. RNA precipitation: transfer the aqueous phase to a fresh tube, and save the organic phase if isolation of DNA or proteins is desired. Precipitate the RNA from the aqueous phase by mixing with isopropyl alcohol. Use 0.5 ml of isopropyl alcohol per 1 ml of TRIzol^{TM} used for the initial homogenization. Incubate samples at 15 to 30°C for 10 minutes and centrifuge at no more than 12.000 × g for 10 minutes at 2 to 8°C. The RNA precipitate, often invisible before centrifugation, forms a gel-like pellet on the side and bottom of the tube.

4. RNA wash: remove the supernatant. Wash the RNA pellet once with 75% ethanol, adding at least 1 ml of 75% ethanol per 1 ml of TRIzol^{TM} used for the initial homogenization. Mix the sample by vortexing and centrifuge at no more than 7,500 × g for 5 minutes at 2 to 8°C.

5. Redissolving the RNA: at the end of the procedure, briefly dry the RNA pellet (air-dry or vacuum-dry for 5-10 minutes). It is important not to let the RNA pellet dry completely as this will greatly decrease its solubility. Partially dissolved RNA samples have an A260/280 ratio < 1.6. Dissolve RNA in RNase-free water by passing the solution a few times through a pipette tip, and incubating for 10 minutes at 55 to 60°C. and stored at -70°C. It is possible to store the RNA pellet at -80°C, if it is not immediately used.

The RNA control (reference) used in my experiments is "Adult Total Breast RNA" (Stratagene^{TM}) of pooled breast normal tissues of females, 56 years old, Caucasian race.

3.5 ASSESSMENT RNA YIELD AND QUALITY

3.5.1 RNA QUANTITATION

The concentration of the RNA solution was determined by measuring its absorbance at 260 nm:

µg RNA/ml = A_{260} x dilution factor x 40

where 40 is the extinction coefficient: 1 A_{260} = 40 µg RNA/ml

To quantitate the extracted RNA I used NanoDrop 1000A Spectrophotometer (http://www.nanodrop.com/) (fig. 3.6).

The Thermo Scientific NanoDrop™ 1000 Spectrophotometer measures 1 ul samples with high accuracy and reproducibility. The full spectrum (220nm-750nm) spectrophotometer utilizes a patented sample retention technology that employs surface tension alone to hold the sample in place. This eliminates the need for cumbersome cuvettes and other sample containment devices and allows for clean up in seconds. In addition, the NanoDrop 1000 Spectrophotometer has the capability to measure highly concentrated samples without dilution (50X higher concentration than the samples measured by a standard cuvette spectrophotometer).

Figure 3.6: NanoDrop® ND-1000 (http://www.nanodrop.com/).

A 1.4 µl RNA sample is pipetted onto the end of a fiber optic cable (the receiving fiber). A second fiber optic cable (the source fiber) is then brought into contact with the liquid sample causing the liquid to bridge the gap between the fiber optic ends. The gap is controlled to both 1mm and 0.2 mm paths. A pulsed xenon flash lamp provides the light source and a spectrometer utilizing a linear CCD array is used to analyze the light after passing through the sample. The instrument is controlled by PC based software, and the data is logged in an archive file on the PC.

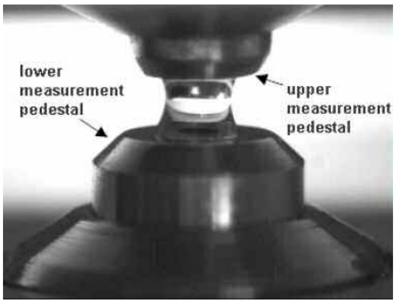


Figure 3.7: Fiber optic cables of NanoDrop® ND-1000 (http://www.nanodrop.com/).

To measure nucleic acid samples select the 'Nucleic Acid' application module and the program show the following parameters:

<u>sample type</u>: used to select the type of nucleic acid being measured. The user can select 'DNA-50' for dsDNA, 'RNA-40' for RNA, 'ssDNA-33' for single-stranded DNA, or 'Other' for other nucleic acids.

<u>λ and abs</u>: the user selected wavelength and corresponding absorbance. The wavelength can be selected by moving the cursor or using the up/down arrows to the right of the wavelength box.

<u>A260 10 mm path</u>: absorbance of the sample at 260 nm represented as if measured with a conventional 10 mm path. Note: This is 10X the

absorbance actually measured using the 1 mm path length and 50X the absorbance actually measured using the 0.2 mm path length.

A280 10 mm path: sample absorbance at 280 nm represented as if measured with a conventional 10 mm path. Note: This is 10X the absorbance actually measured using the 1 mm path length and 50X the absorbance actually measured using the 0.2 mm path length.

260/280: ratio of sample absorbance at 260 and 280 nm. The ratio of absorbance at 260 and 280 nm is used to assess the purity of DNA and RNA. A ratio of ~1.8 is generally accepted as "pure" for DNA; a ratio of ~2.0 is generally accepted as "pure" for RNA. If the ratio is appreciably lower in either case, it may indicate the presence of protein, phenol or other contaminants that absorb strongly at or near 280 nm. See "260/280 Ratio" in the Troubleshooting section for more details on factors that can affect this ratio.

260/230: ratio of sample absorbance at 260 and 230 nm. This is a secondary measure of nucleic acid purity. The 260/230 values for "pure" nucleic acid are often higher than the respective 260/280 values. They are commonly in the range of 1.8-2.2. If the ratio is appreciably lower, this may indicate the presence of co-purified contaminants.

ng/μl: sample concentration in ng/ul based on absorbance at 260 nm and the selected analysis constant.

3.5.2 RNA QUALITY ANALYSIS

Due to the presence of RNases, integrity check is an essential steps before any RNA dependent application. I used the 2100 Agilent Bioanalyzer (http://www.chem.agilent.com/) to analyze the quality of the extracted total RNA. The instrument gives also a RNA Integrity Number (RIN), an estimate of the quantity, and calculates ribosomal ratios (rRNA 18S and rRNA 28S) of the total RNA sample.

Figure 3.8: 2100 Agilent Bioanalyzer (http://www.chem.agilent.com/)

The 2100 Bioanalyzer is a microfluidics-based platform based on the "Lab-on-a-Chip technology". This technology utilizes a network of channels and wells that are etched onto glass or polymer chips to build mini-labs. Pressure or electrokinetic forces move picoliter volumes in finely controlled manner through the channels. Lab-on-a-Chip enables sample handling, mixing, dilution, electrophoresis and detection on single integrated systems. The main advantages of Lab-on-a-Chip are ease-of-use, speed of analysis, low

sample (detects until 200 pg/µl of total RNA) and reagent consumption and high reproducibility due to standardization and automation.
In my analysis I used the RNA 6000 Nano LabChip, which is able to detect 5 ng/µl of total RNA and 25 ng/µl of mRNA.

Figure 3.9: RNA 6000 Nano LabChip (http://www.chem.agilent.com/)

This chip contains a network of interconnected channels where the samples are loaded. A gel matrix with a specific RNA dye fills the chip. Each sample migrates into the matrix as in a capillary electrophoresis (fig. 3.10).

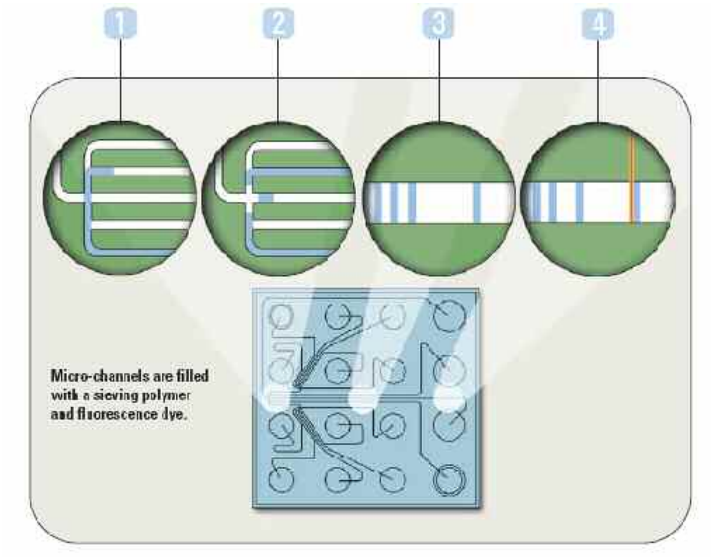

1
2
3
4
Micro-channels are filled with a sieving polymer and fluorescence dye.

Figure 3.10: The sample migrates from the wells through the microchannel of the chip (1), until the microchannel of separation (2). The different components of the sample are separated by electrophoresis (3), the fluorescence was detected and it was converted in the classical electropherogram image (4). (http://www.home.agilent.com).

A degradated RNA molecule is easily detected from the graph (fig. 3.11) because of:

- lower ratio rRNA18S/rRNA28S,
- other peaks between rRNA18S peak and rRNA28S peak,
- decrease of the global signal detected from the instrument,
- shift of the peaks to lower molecular weight.

The software calculates on the basis of the whole electrophoresis migration and the presence/absence of degradation products, the RIN value (RNA Integrity Number) with a range of numbers from 1 (completely degraded RNA) to 10 (totally intact RNA).

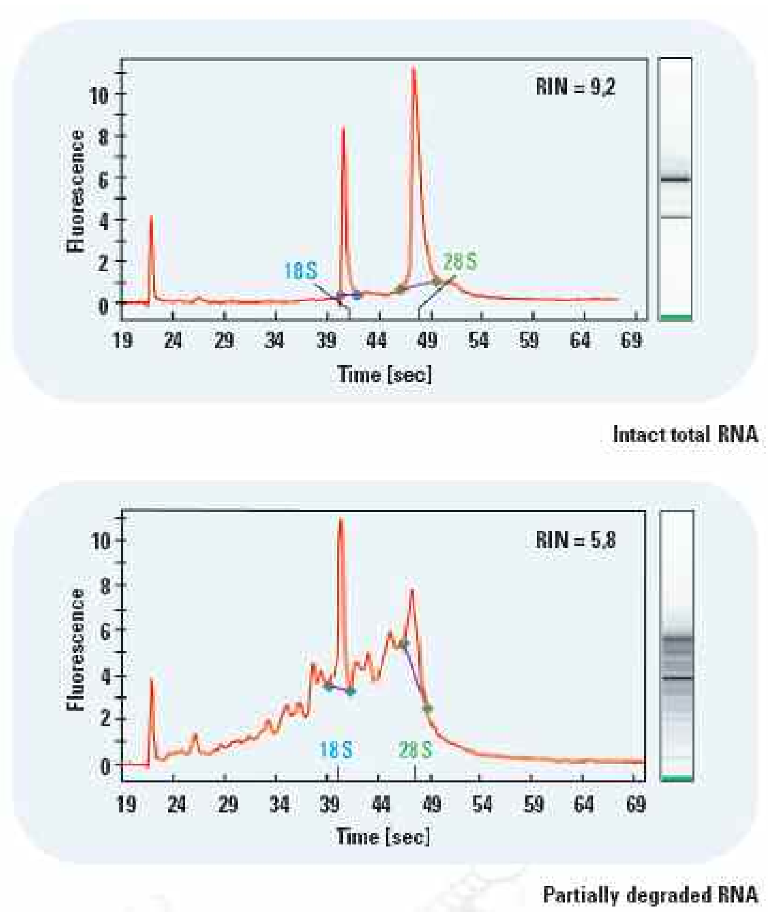


Figure 3.11: Electropherograms of two total RNA samples: above a high quality sample, below a partially degraded sample(http://www.home.agilent.com). RIN= RNA Integrity Number

3.6 RNA AMPLIFICATION

3.6.1 Amino Allyl MessageAmp^{TM} aRNA

RNA amplification was originally developed as a method to amplify RNA samples to produce enough material for array hybridization [132].

The Amino Allyl MessageAmp^{TM} aRNA Amplification Kit is based on the RNA amplification protocol developed in the Eberwine laboratory [133]. The procedure consists of reverse transcription with an oligo(dT) primer bearing a T7 promoter and using ArrayScript^{TM}, a reverse transcriptase (RT) engineered to produce higher yields of first strand cDNA than wild type enzymes. ArrayScript catalyzes the synthesis of full-length cDNA that then undergoes a second strand synthesis and clean-up to become a template for in vitro transcription (IVT) with T7 RNA polymerase. This enzyme generates from hundreds to thousands of antisense RNA copies (aRNA) of each mRNA in a sample. The quantity of RNA for the amplification is 100-1000 ng total RNA or 10-100 ng mRNA.

Several groups have tried to determine whether amplification of RNA introduces bias and they reported that any bias is minimal [134, 135].

In summary the Amino Allyl MessageAmp aRNA amplification procedure consists in five steps:

- reverse transcription to synthesize first strand cDNA (3.6.2),
- second strand cDNA synthesis (3.6.3),
- cDNA purification (3.6.4),
- in vitro transcription to synthesize amino allyl-modified aRNA (3.6.5),
- aRNA purification (3.6.6).

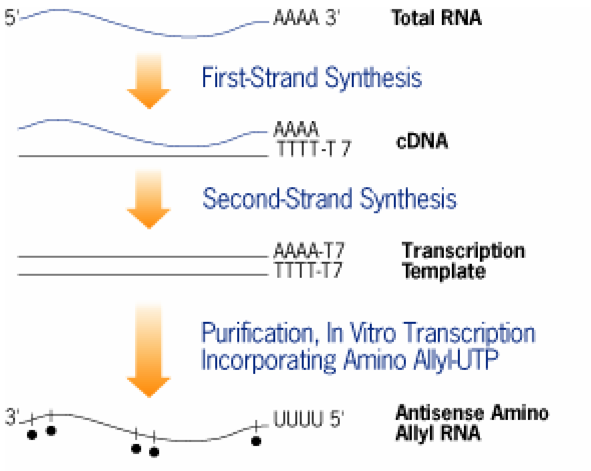


Figure 3.12: Amino Allyl MessageAmp II aRNA amplification procedure (http://www.ambion.com)

3.6.2 REVERSE TRASCRIPTION TO SYNTHESIZE FIRST STRAND cDNA

This reaction is primed with the T7 Oligo(dT) primer that binds to the 3' poli A tail of mRNA to synthesie cDNA containing a T7 promoter.
The protocol consists of eight steps:

- place 1 µg of total RNA into a sterile RNase-free tube;
- add 1 µl of T7 Oligo(dT) primer;
- add nuclease-free water to a final volume of 12 µl, vortex briefly to mix, then centrifuge to collect the mixture at the bottom of the tube;
- incubate 10 min. at 70°C in a thermal cycler to denaturate the RNA secondary structures;
- centrifuge samples briefly to collect them at the bottom of the tube, place the mixture on ice;
- at room temperature prepare reverse transcription master mix in a nuclease-free tube assembling the reagents in the order shown below:

COMPONENT	**QUANTITY**
10X First Strand Buffer	2 µl
dNTP Mix	4 µl
RNAse Inhibitor	1 µl
ArrayScript	1 µl

- transfer 8 µl of mix to each RNA sample mixing well and incubate 2 hours at 42°C;

- after the incubation place the tubes on ice and proceed to the second strand cDNA synthesis.

3.6.3 SECOND STRAND cDNA SYNTHESIS

On ice, prepare a Second strand master mix in a nuclease-free tube in the order listed below:

COMPONENT	QUANTITY
H_2O Nuclease-free	63 µl
10X Second Strand Buffer	10 µl
dNTP Mix	4 µl
DNA Polymerase	2 µl
RNase H	1 µl

- mix well by gently vortexing;
- transfer 80 µl of mix to each sample mixing well;
- incubate 2 hours in a 16°C thermal cycler;
- after the incubation place the reaction on ice and proceed to the cDNA purification step.

3.6.4 cDNA PURIFICATION

cDNA purification removes RNA, primers, enzymes and salts that would inhibit in vitro transcription. This protocol uses specific cDNA filter cartridges containing silicon membranes to bind the cDNA. Before the purification procedure the cDNA filter is equilibrated with 50 µl of cDNA Binding Buffer and is incubated 5 min at room temperature.
The protocol consists of:

- add 250 µl of cDNA Binding Buffer and mix thoroughly;
- pipet the solution onto the center of the cDNA filter cartridge (cDNA mixture binds the silica filter at pH < 7.5);
- centrifuge for ~ 1 min. at 10.000 x g or until the mixture goes through the filter; (only the cDNA > 100 bp binds the filter);
- discard the flow-through and replace the cDNA filter cartridge in the wash tube;
- apply 500 µl of Wash Buffer and centrifuge for ~ 1 min. at 10.000 x g; discard the flow-through and spin the cDNA cartridge for an additional minute;
- transfer cDNA filter cartridge to a new cDNA elution tube;
- apply 9 µl of nuclease-free water, preheated to 50°C-55°C, to the center of the filter in the cDNA filter cartridge;
- leave at room temperature for 2 min. and then centrifuge for 1.5 min. at 10.000 x g, or until all the nuclease-free water is through the filter;
- eluate with a second 9 µl of preheated nuclease-free water.

3.6.5 AMINO ALLYL-MODIFIED aRNA IN VITRO TRANSCRIPTION (IVT)

IVT generates multiple copies of amino allyl-modified aRNA from the double-stranded cDNA templates. This protocol uses a modified nucleotide, 5-(3-aminoallyl)-UTP (aaUTP) that contains a primary aminic group on C5 of uracil. This group reacts with the carbossilic group of the fluorophore molecule during the dye coupling reaction (see par 3.8.1) (fig. 3.13).

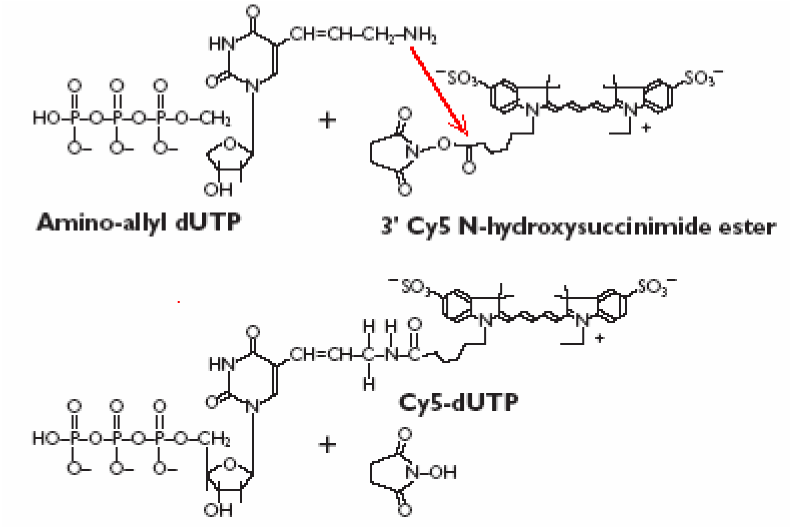


Figure 3.13: Amino Allyl Labeling Reaction (adapted from MessageAmp^{TM} II aRNA kit)

The IVT master mix was prepared at room temperature by adding the following reagents:

COMPONENT	QUANTITY
aaUTP Solution (50mM)	3 µl
ATP, CTP, GTP Mix (25mM)	12 µl
UTP Solution (50mM)	3 µl
T7 10X Reaction Buffer	4 µl
T7 Enzyme Mix	4 µl

- mix well the mix by gently vortexing and transfer 26 µl to each sample;
- incubate for 4-14 hours at 37°C;
- after the incubation add 2 µl of DNase I to completely remove the cDNA, incubate 30 min. at 37°C.

3.6.6 aRNA PURIFICATION

The purification removes unincorporated aaUTP and Tris from IVT reactions that would otherwise compete with the aRNA for dye coupling; it also removes enzymes, salts and other unincorporated nucleotides. Similarly to cDNA purification, aRNA purification uses specific aRNA filter cartridges in silicon material.

The protocol is described below:

- add 58 µl nuclease-free H_2O, 350 µl aRNA Binding Buffer, 250 µl 100% ethanol and mix by pipetting the mixture;
- pipet the mixture onto the center of the filter in the aRNA filter cartridge and centrifuge for ~ 1 min. at 10.000 x g, continue until the mixture has passed through the filter, now the aRNA is bound to the filter;
- wash the filter with 650 µl of Wash buffer and centrifuge for ~ 1 min. at 10.000 x g, repeat the centrifugation to remove trace amounts of ethanol contained in Wash Buffer;
- transfer filter cartridge to a fresh aRNA collection tube and add 100 µl of nuclease-free H_2O preheated to 50 – 60°C;
- leave at room temperature for 2 min. and then centrifuge for ~1.5 min. at 10.000 x g or until the nuclease-free water is through the filter; the aRNA will now be in the aRNA collection tube
- store purified aRNA at -20°C overnight or at -80°C for longer times if desired.

The concentration of the aRNA solution was determined using the NanoDrop 1000A as previously described (see par. 3.6.1). Usually a good yield of aRNA is 20-30 µg starting from 1 µg of total RNA.

3.7 DYE COUPLING AND LABELED aRNA CLEANUP

3.7.1 aRNA DYE COUPLING REACTION

To prepare the labelled target to hybridize with the probes on the array I used the indirect labelling method. This procedure is more convenient than the direct labelling. The aaUTP incorporated during the IVT has only a minor effect on the reaction efficiency and yield [136] because the T7 RNA polymerase incorporates with high efficiency both aaUTP and UTP. Moreover dye coupling reaction using Cy3 and Cy5 has similar efficiency and labelled samples will not have the biases that can result from direct incorporation of modified nucleotides by in vitro transcription.

To label the target aRNA I used the mono-reactive NHS esters of Cy3 and Cy5 (Amersham Biosciences) (fig. 3.14).

Cy3 and Cy5 fluorescent dyes are used very often in the microarray technology because they are photostable and they have a high signal of fluorescence emission. Since the absorption spectrums of Cy3 and Cy5 have a small overlap, they can be excited separately to detect the single fluorescence (3.15).

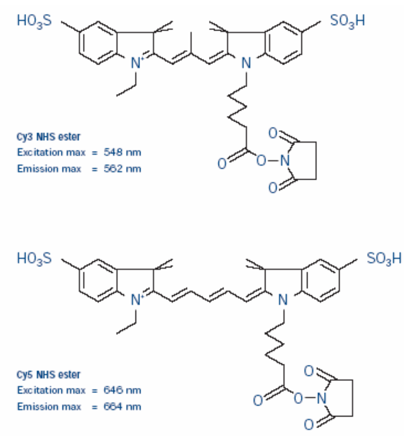


Figure 3.14: Chemical structures of Cy3 (5-NN'-diethyl-tethrametylindocarbocyanine) and Cy5 (5-NN'-diethyl-tethrametylindo di carbocyanine) molecules (MicroArray Handbook Amersham, 2002)

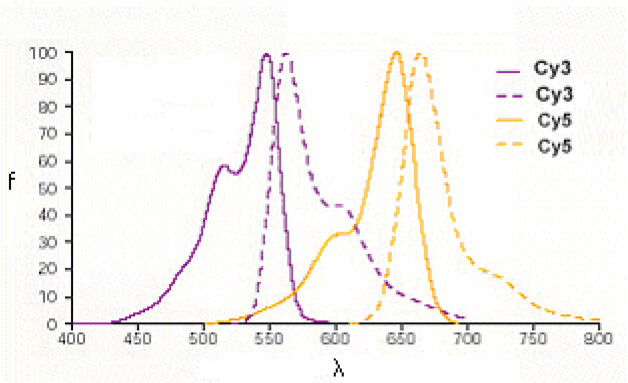


Figure 3.15: Absorption and emission spectrum of Cy3 and Cy5. f=fluorescence, λ wavelength (MicroArray Handbook Amersham, 2002)

The labelling protocol is described below:

- prepare dye before starting the dye coupling procedure adding 11 µl of DMSO to Cy3 and Cy5 reactive dye and mixing thoroughly, keep the resuspended dye in the dark at room temperature for up to 1 hour. It is important that the dye compounds remain dry before and after dissolving in DMSO because any water that is introduced will cause hydrolysis of NHS esters, lowering the efficiency of coupling;
- add to 20-25 µg of aRNA previously lyophilized 9 µl of Coupling Buffer and resuspend thoroughly by gentle vortexing;
- add the DMSO dyes to the aRNA coupling buffer mixture and mix well by vortexing gently;
- incubate the solution 30 min. at room temperature in the dark: this incubation allows the dye coupling reaction to occur;
- add 4.5 µl of 4M Hydroxylamine and incubate the reaction in the dark for 15 min.; the large molar excess of hydroxylamine quenches the amine-reactive groups on the unreacted dye molecules.

3.7.2 DYE LABELED aRNA PURIFICATION

This purification removes excess dye from labeled aRNA. The procedure follows the same protocol described above (par. 3.7.6) for the aRNA purification. The labeled aRNA purified can be stored at -20°C for some months without damage for incorporated fluorophores.

3.7.3 SPECTROPHOTOMETRIC ANALYSIS OF DYE INCORPORATION

To calculate how many picomoles (pmol) of fluorophores were incorporated after aRNA labeling, I used the NanoDrop 1000A Spectrophotometer (http://www.nanodrop.com/), as described in paragraph 3.6.1. In this case I selected the application module "Microarray" that measures the dye incorporation concentration. The instrument has a very low detection limit, 0.20 pmol/µl for Cy3 and 0.12 pmol/µl for Cy5 and accurately measures concentrations up to 100 pmol/µl for Cy3 and 60 pmol/µl for Cy5.
The output of the program reports:

Cy3, Cy5 abs.norm: normalized absorbance of selected Dye at the 1 mm pathlength

pmol/ul: concentration based upon selected Dye's extinction coefficient (0,15 µM/cm Cy3 and 0,25 µM/cm Cy5)

ng/ul: concentration of nucleic acids in the sample calculated using the absorbance at 260 nm minus the absorbance at 340 nm (i.e. normalized at 340 nm) and the nucleic acid analysis constant

260/280: ratio of sample absorbance at 260 and 280 nm. The ratio of absorbance at 260 and 280 nm is used to assess the purity of RNA, a ratio of ~2.0 is generally accepted as "pure" for RNA. If the ratio is appreciably lower, it may indicate the presence of protein, phenol or other contaminants that absorb strongly at or near 280 nm.
To calculate the number of picomol the software uses the following calculation:

pmolCy3 = [(A_{550} – A_{700}) * vol* dilution factor]/ 0.15

pmolCy3 = [(A_{650} – A_{700}) * vol* dilution factor]/ 0.25

where:

A_{550} = Cy3 absorbance measured at the wavelength of the maximum absorption of Cy3 (550 nm)

A_{650} = Cy5 absorbance measured at the wavelength of the maximum absorption of Cy5 (650 nm)

A_{700} = absorbance at wavelength = 700 nm to determine background

0.15 = Cy3 extinction coefficient (ε)

0.25 = Cy5 extinction coefficient (ε)

It is important to estimate also the number of dye molecules incorporated per 1000 nucleotides (nt):

nr dye molecules/1000 nt = A_{dye}/A_{260} * 9010 cm-1M-1/dye extinction coefficient * 1000

The expected incorporation rate is 30-60 dye molecules per 1000 nucleotides.

3.8 PRE-HYBRIDIZATION AND HYBRIDIZATION REACTIONS

3.8.1 LABELED aRNA PRECIPITATION

The Cy3 labeled aRNA and the Cy5 labeled aRNA are pooled together in the same solution. It is necessary that the two samples have a similar number of picomoles of fluorophore and also that the quantity of aRNA is comparable in two labeled aRNA (Cy3 and Cy5). I struck a balance between these two parameters and also I had to consider the different quantum yield of Cy3 and Cy5. Indeed Cy3 binds the aRNA more efficiently than Cy5 but has a lower quantum yield when it is detected from the scanner. After the optimization of these parameters, the probes are precipitated as reported below:

- add 4/5 of total volume of CH_3COONH_4 and 2,5 volumes of EtOH absolute mixing well;
- leave the solution 30 min. at 4°C (or at -20°C for longer time) and centrifuge at room temperature for 15 min. at 14.000 x g: it is now visible on the bottom of the tube a pellet;
- remove the supernatant and wash the pellet with EtOH 75%;
- centrifuge for 5 min. at 14.000 x g: the pellet is firmly attached to the bottom of the tube:
- repeat the wash with EtOH 75% and dry the pellet under a laminar flux hood, in order to remove any traces of ethanol.

3.8.2 LABELED aRNA FRAGMENTATION

Many protocols for using amplified RNA in microarray analysis recommend the aRNA fragmentation prior to hybridization to an oligonucleotide microarray. This fragmentation step improves hybridization kinetics with the arrayed oligonucleotides and can lead to enhanced signal.

The protocol for 2–20 µg of RNA, that uses the Ambion's RNA Fragmentation Reagents^{TM}, is reported below:

- bring the RNA sample volume to 9 µl with Nuclease-free Water. Add 1 µl of the 10X Fragmentation Buffer to the RNA sample;
- mix, spin briefly, and incubate at 70°C for 15 min in a heating block, thermocycler, or water bath;
- add 1 µl of Stop Solution, place it on ice until use, or store it at –80°C.

The average size of the resulting fragments will be 60-200 nucleotides.

3.8.3 PREHYBRIDIZATION REACTION

The pre-hybridization reaction saturates not-specific sites on the surface of the microarray slide to avoid an high background signal due to the not-

specific binding between the probes and the target.
The protocol for the pre-hybridization is reported below:

- add 70 µl of pre-hybridization buffer (see par. 3.1) preheated at 48°C on the surface of microarray slide where the oligos are spotted, cover with a coverslip;
- put the microarray slide in a hybridization chamber (HybChamber^{TM} Gene Machines, #HYB-03) (fig. 3.16) that contains in the reservoir groove ~ 100 µl of H_2O mQ to maintain enough humidity during the incubation;
- leave at least 1 hour the chamber in a water bath at 48°C; if the incubation is prolonged until 18-24 hours, the slide show a lower background;
- remove the coverslip gently (wash the slide with H_2O mQ and dry with compressed air.

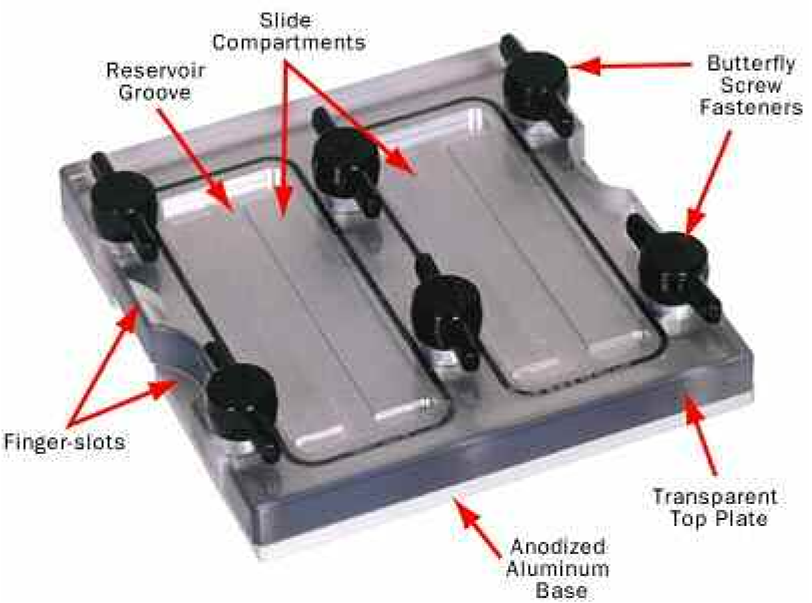


Figure 3.16: Hybridization chamber (HybChamber^{TM} Gene Machines, #HYB-03) used for the pre-hybridization reaction.

3.8.4 HYBRIDIZATION REACTION

The hybridization between the labeled aRNA and the probe (oligonucleotide) on the microarray slide requires a proper stringency. There are many factors that influence the reaction:

- temperature,
- salt concentration,
- formamide concentration in the hybridization buffer.

High temperature and high salt concentration increase the stringency, instead low temperature and low salt concentration decrease it. It is necessary to find a balance between these extreme conditions, in fact too high stringency could cause too low signal but also there is a risk of low specificity if the stringency is low.
For the hybridization I used a particular microarray hybridization station that performs a incubation by micro agitation, the ArrayBooster^{TM} (Advalitix Instruments) (fig. 3.17).

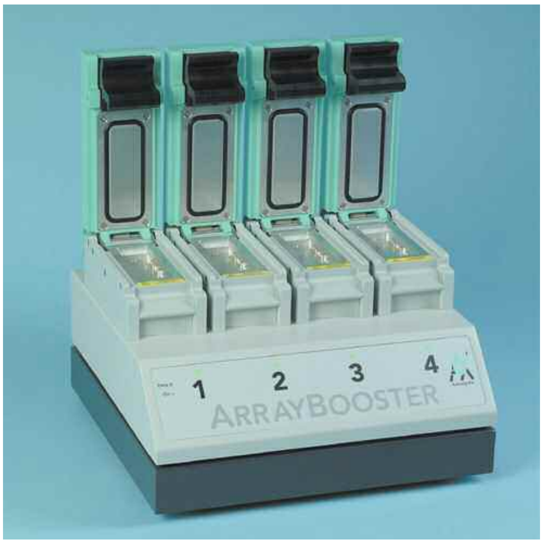
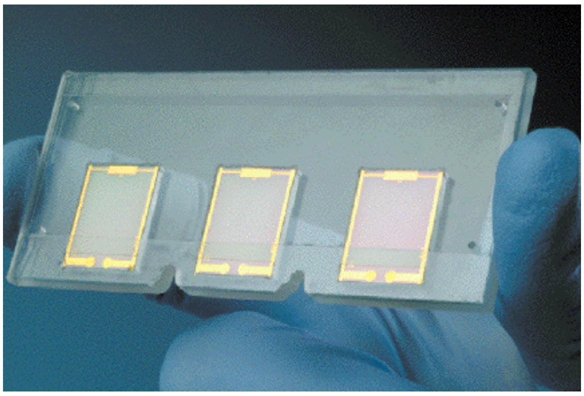


Figure 3.17: ArrayBooster^{TM} (Advalitix Instruments) (http://www.advalytix.de/)

The hybridization of oligos microarrays is diffusion limited, i.e. the signal intensity, especially for low expression genes, is determined by the number of target molecules reaching the probe. However, it is known that diffusion is a slow process for large molecules so that, even during overnight hybridization, the system does not reach equilibrium. Agitation is the obvious solution to overcome the diffusion limitation of acid nucleic hybridization.
The ArrayBooster™ uses Surface Acoustic Waves (SAW) to effectively agitate the sample solution during the incubation. The ArrayBooster™ has four independently controlled chambers that accept all standard slide formats. As the incubation chambers have no valves or tubes the system works without any dead volume. Sample volumes as low as 10 microliters can be incubated and agitated.



3.18: AdvaCard™ with three chips (http://www.advalytix.de/)

The core of the ArrayBooster™ is the AdvaCard™ (see fig. 3.18). The sample solution is sandwiched between the AdvaCard™ and the microarray. Special fluidics prevent bubble formation in the sample loading steps. The AdvaCard™ is a micro-agitation chip card available in three different sizes adapted to different spotting areas. To ensure optimal agitation efficiency

Advacards™ contain one, two or three agitation chips. In my case I used Advacards™ with three chips. A radio frequency voltage feeding the nano pumps on the chips induces nearly-chaotic streaming patterns in the sample solution. The software that controls the instrument, allows the user to program different parameters, temperature, time and grade of agitation, for each chamber independently (see fig. 3.19).

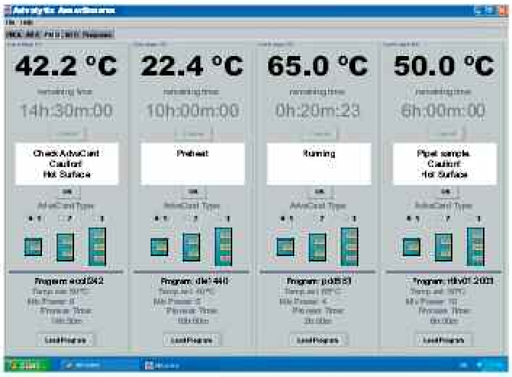


Figure 3.19: ArrayBooster™ control software (http://www.advalytix.de/)

The AdvaCard^{TM} is placed in the incubation chamber over the microarray slide. Before the incubation, I put on both sides of the hybridization chamber the diluted hybridization buffer (250 µl of buffer and 250 µl of mQ H_2O) to ensure a constant humidity during the reaction. The target solution (labeled aRNA) was loaded between the advacard and the microarray slide (see fig. 3.20). Hybridization was carried out at 48°C for 12-16 hours.

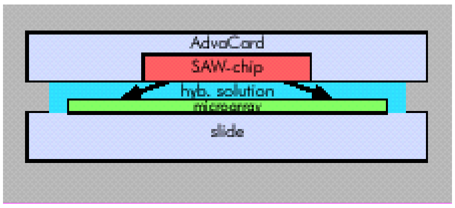


Figure 3.20: Placement of the AdvaCard^{TM}, target solution (hyb.solution) and microarray slide in the ArrayBooster chamber (http://www.advalytix.de/).

For each experiment three replicates (three hybridizations) were performed including a dye-swap procedure (see 3.12.3).

3.8.5 POST-HYBRIDIZATION WASH

After the hybridization process, the microarray slide was washed to remove the target not-hybridizated. The slide was placed in a falcon with 40 ml of wash solution and washed as reported below:

- 1X SSC and 0,2 % SDS for 4 min. at room temperature;
- 0,1 X SSC and 0,2 % SDS for 4 min at room temperature;
- 0,2 X SSC for 4 min. at room temperature (twice);
- 0,1 X SSC for 3 min. at room temperature (twice).

It is important to change the 50 ml tube between the second and the third wash since the SDS is fluorescent could increase the background signal of the microarray. Then the slide was dried with compressed air and stored in the dark until the use.

3.9 MICROARRAY SCANNING USING SCANARRAY LITE^{TM}

Microarrays were scanned using ScanArray Lite (PerkinElmer^{TM}), a confocal laser scanner that it is able to excite the fluorophores Cy3 and Cy5 and detect their fluorescence emission (fig. 3.21). The working principle of this scanner is similar to a confocal microscopy: a confocal microscope uses point illumination and a pinhole in an optically conjugate plane in front of the detector to eliminate out-of-focus information. Only the light within the focal plane can be detected, so the image quality is much better than that of wide-field images (http://micro.magnet.fsu.edu/) [137]. ScanArray Lite utilizes confocal technology to collect more signal of interest and to automatically reduce background noise in order to achieve higher signal-to-noise ratios.

Figure 3.21: ScanArray Lite (Packard)

In brief ScanArray Lite works as described below:

- it has two fixed internal lasers: the green laser He-Ne excites Cy3 with a excitation wavelength of 543 nm and the red laser He-Ne excites Cy5 with a excitation wavelength of 633 nm. Since there is only one photomultiplier tube (PMT) that detects the fluorescence signal, the lasers do not work at the same time but excite one dye at a time;
- the lens collects the light generated from the excited fluorophores;
- the light passes through two emission filters (for 570 nm and 670 nm wavelengths) that adsorb the light reflected from the microarray slide;
- the detector focuses the light into a pinhole in front of the PMT tube, in order to detect the light within the focal plane of the slide;
- the PMT converts the light signal into electric signal and then into digital image.
- the software that controls the instrument assigns ″false″ colours to each pixel of the image: the spots with a low fluorescence signal look blue, the spots with a higher signal in one of two channels (Cy3 or Cy5) green or red, the saturated spots white.

The software ScanArray Express^{TM} controls the acquisition and analysis of the image. If the laser power and the PMT gain are increased, the fluorescence signal is intensified and, also, the background signal.
The "tiff" image format is 16 bit and it means that pixels could have 2^{16}-1

different values: each pixel could have a value between 0 and 65535.

For each slide are acquired two separate images, one for the Cy3 labeled sample and the other one for the Cy5 labeled sample: the result is a "false" colours image with Cy3 = green and Cy5 = red. Then the two images are combined to obtain a composite image: green spots are the over-expressed genes in the Cy3 labeled aRNA, instead the red spots are the over-expressed genes in the Cy5 labeled, the yellow spots represent the genes not differentially expressed between Cy3 and Cy5
labeled aRNA.

Signal intensity is balanced between two fluorophores in an array by automatically adjusting the laser power or PMT gain to a user defined target of pixel intensity for each fluorophore.

ScanArray Lite is able to detect until 0.05 molecules of fluorophore/μm^2 and produces highly reproducible results: repeatability of results with multiple scans and uniformity of results across the slide are both less than 5% CV.

3.10 IMAGE ANALYSIS WITH SCANARRAY EXPRESS

After image acquisition by scanning, I selected a quantitation protocol in ScanArray Express ^{TM} to analyze the images. The software requires a ".gal" file created during the spotting of the probes on the slide. This file contains all specifications that define the geometry of the array: gene names and positions on the slide, subarray number, spot diameter, spacing between array columns and rows.

The user can select specific parameters for the image quantitation, in my protocol I set the options reported below:

- Spots quality measurement method: Footprint. For each subarray, the software calculated the difference between the center of the nominal spots and the center of the found spot. Let the shifted nominal position be (X,Y), the found position to be (x,y), the footprint is:
  $(X - x)^2 + (Y - y)^2$
  Spots with a calculated footprint lower than the maximum specified in the application settings (100 μm in my experiment) are considered for the subsequent analysis.
- Signal quantitation method: Fixed circle. ScanArray Express determines the center for each spot and the corresponding patch. The patch is a rectangle that is constructed around the center of the spot with the dimensions indicated in the template. Both the spot and the background must be defined with the patch. The quantitation method selected then constructs masks for the spot and the background. A mask (fig. 3.22) is a pixel by pixel map that indicates the property of each pixel. The Fixed circle method fits all spots in the image with circles of fixed diameter. Using this option the spot mask and the background mask are constructed using the parameters of the spot diameter, and the background inner and outer dimensions. This method works well if all spots have the same size and shape.

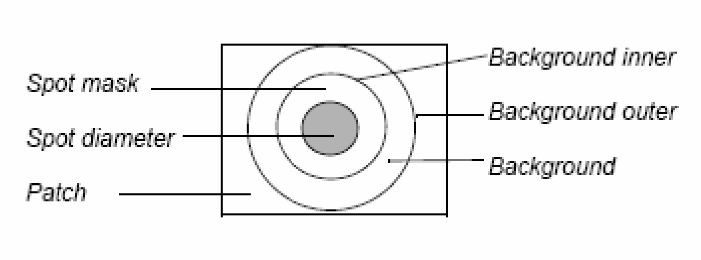


Fig. 3.22: Example of mask constructed around a spot from ScanArray Express (ScanArrayExpress User Manual).

The software, identified the pixels belonging to spots and background, calculates the difference and the mean of the fluorescence intensity for both channels (Cy3 and Cy5) and produces a single value. I used the median value instead of the mean value, because this measure is less influenced from groups of pixels with values too different from the mean value.

- Normalization method: LOWESS (Locally Weighted Scatter Plot Smoothing). The LOWESS method carries out robust locally-weighted scatter plot smoothing for both equally spaced and non-equally spaced data [138].

The quantitation results were displayed in the main window as a Spreadsheet, a Scatter Plot (fig. 3.23), and a Distribution Plot (fig. 3.24).

- Spreadsheet: each row in the spreadsheet is the data from one spot, including the gene names and ID numbers that were imported from the .GAL file. You can scroll vertically to see the data for each spot and horizontally to view the 55 columns of data for any spot. It is useful to evaluate if the spot fluorescence level is high enough to perform a reliable statistical analysis.
- Scatter Plot: this plot allows you to see any column of data plotted against any other column. When you select a data point on the scatter plot, the corresponding data point are displayed in the spot viewer and is highlighted in both the Spreadsheet and Image viewer when you switch back to those tabs.

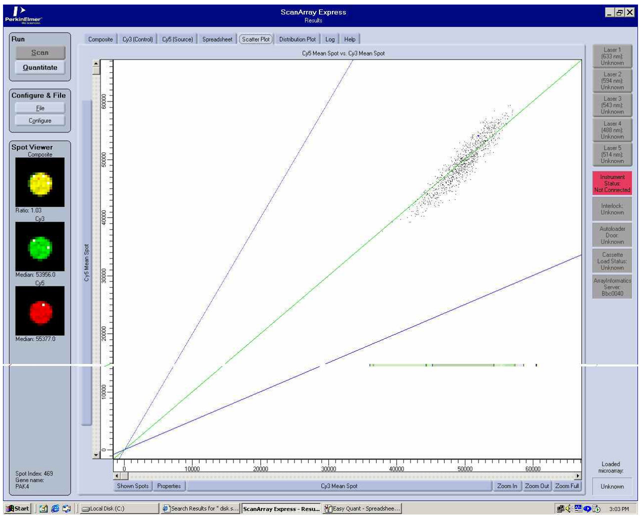
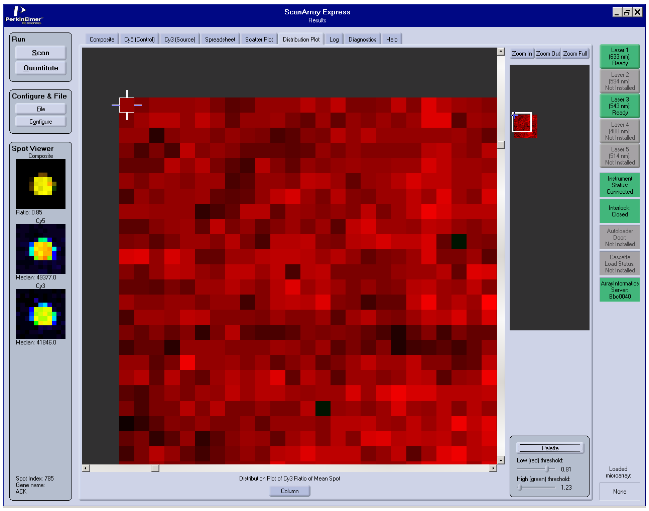


Figure 3.23: Scatter Plot (ScanArrayExpress User Manual)

- Distribution Plot: this plot allows you to see trends that are area-sensitive on the slide; for example, if one side of the slide is over- or under-washed, or if a pin is partially clogged. A good (high quality) microarray should not have any spatial correlations for any parameters.



Ratio: 0.85

Median: 49377.0

Median: 41846.0

Spot Index: 785
Gene name:
ACK

Distribution Plot of Cy3 Ratio of Mean Spot

Low (red) threshold: 0.81

High (green) threshold: 1.23

Laser 1 (633 nm): Ready

Laser 2 (594 nm): Not Installed

Laser 3 (543 nm): Ready

Laser 4 (488 nm): Not Installed

Laser 5 (514 nm): Not Installed

Instrument Status: Connected

Interlock: Closed

Autoloader Door: Not Installed

Cassette Load Status: Not Installed

ArrayInformatics Server: Bbc0040

Loaded microarray: None

Figure 3.24: Distribution Plot (ScanArrayExpress User Manual)

3.11 STATISTICAL ANALYSIS

3.11.1 FILTERING ANALYSIS OF FLUORESCENCE VALUES

The values within a slide were filtered considering their fluorescence intensity and the standard deviation (SD) value; this step was performed before the normalization step (see 3.12.2). The procedure is reported below:

- spots with a median fluorescence pixel intensity below 300 (calculated considering negative control intensity) on both Cy3 and Cy5 channels were filtered out. The value 300 represents the threshold under that the software ScanArray Express can not detect consistently the fluorescence intensity;
- spots with a median fluorescence pixel intensity of zero or less in only one channel were set to 100 to prevent their elimination during normalization (see 3.12.2);
- based on the method suggested by Yang et al. [139] R1 = (CH1 intensity/ CH2 intensity) and R2 = (CH1 intensity/ CH2 intensity) values for two replicates of the same gene on the microarray and $\log_2$(R1/R2) were calculated. We indicated the two replicates of the spot as R1 and R2. Then we calculated the mean and SD for the $\log_2$(R1/R2) values of all microarray spots. Those with a $\log_2$ ratio higher than $|3\ \mathrm{SD}|$ were rejected due to replicate inconsistency. The microarrays used have two replicates for each gene, so it is important to look at the consistency of the two values for the Cy3- and Cy5-channel. The majority of the replicates have similar values and the $\log_2$(R1/R2) should be around zero. If the values are too different, and so the $\log_2$(R1/R2) is not around zero, it is not possible to establish which is the "real" value. I used this method

based on the SD calculation for each $\log_2$(R1/R2) to remove the replicates with a $\log_2$(R1/R2) that highly differ from zero

- geometric mean for the two intra-array replicates of the remaining genes was calculated based on the method suggested by Quackenbush [140].

Output files were saved in "tav" format to make them suitable for the MIDAS software (see 3.12.2).

3.11.2 DATA NORMALIZATION WITH TIGR MICROARRAY DATA ANALYSIS SYSTEM (MIDAS) (http://www.tm4.org/midas.html)

The "tav" file for each microarray experiment was normalized with MIDAS software using the LOWESS (Localised weighted smother estimator) method.
There are many sources of systematic variation in microarray experiments which can affect the gene expression levels: differences in labelling efficiency between the two fluorescent dyes, experimental variability in hybridization and processing procedures or scanner settings at the data collection step. For example, since the red (Cy5) and green (Cy3) dyes differ in physical properties such as heat and light sensitivity, usually the mean fluorescence intensity of Cy3 is lower than the mean fluorescence intensity of Cy5. The purpose of the normalization is to minimize systematic variations in the measured expression levels of two co-hybridized mRNA samples, so that biological differences can be more easily distinguished, as well as to allow the comparison of expression levels across slides [141].
The relationship between dye-bias and intensity could be better evaluated in an MA-plot (fig. 3.25) which is a scatterplot of the log-ratios called M-values (minus) against the log-intensities called A-values (add) for an array. MA-plots may also be called RI-plots (ratio-intensity). In detail:

log-ratios for each spot: M = log2Ai-log2Bi = log2(Ai / Bi)
log-intensity of each spot: A = (log2Ai + log2Bi)/2 = log2√(Ai x Bi)

This plot can reveal intensity specific artefacts in the log2(R/G) measurements. In a perfect situation, the log-ratios M in an MA-plot should be evenly distributed around zero across all intensity-values A. However, this is rarely the case. Imbalance of the hybridization intensities of the different dyes can be seen as a curve in the plot (graph A in the figure). This systematic error can be removed with normalisation, which is normally applied to the log-ratios (M-values). In my case, normalisation is performed by applying a statistical regression method to the MA plot called locally weighted linear regression analysis (Lowess). Lowess removes the intensity dependent curvature of the data in the MA-plot. The Lowess curve is constructed performing a series of local regressions, one for each point in the scatterplot. The local regressions are based on a (user defined) percentage of spots (f parameter) that are closest in terms of intensity-values (A-values) to the spot, for which the local regression is being predicted. From these neighbouring spots a weighted average of the log-ratio, log2(Ai / Bi) is calculated.

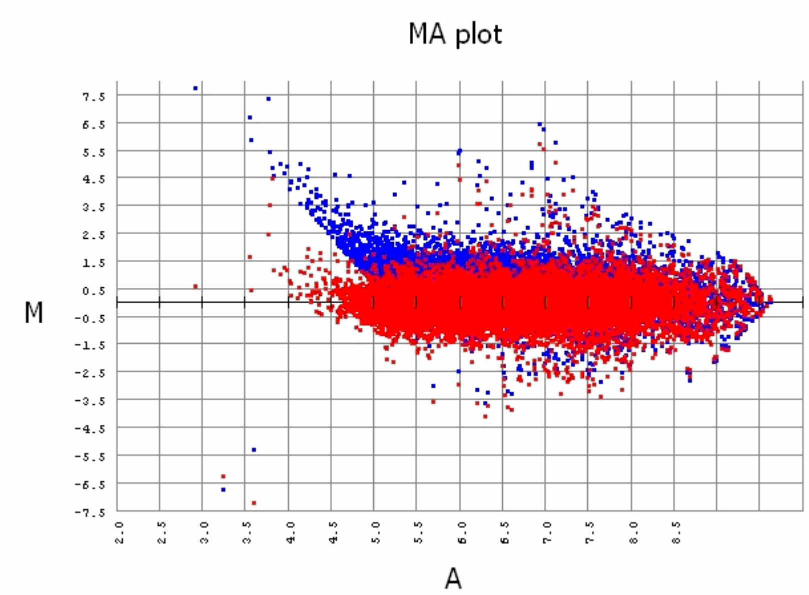


Figure 3.25: MA-plots generated from MIDAS before (blue graph) and after (red graph) Lowess normalisation (TIGR MIDAS User Manual). The figure highlights that for low signal there is a systematic increase in one of two channels (left part blue graph) that it is corrected by the Lowess normalization (red graph) (TIGR MIDAS User Manual).

This means that the neighbouring spots are weighted differently depending on how far they are from the target A-value. In summary, observations further from the target A-value are down-weighted compared with values close to the target A-value. The weighted average value (the Lowess value) is then subtracted from the experimentally observed log-ratio of the spot (M-value): M' = M - Lowess. Thus, each gene is normalised with a different normalization value dependent on its hybridisation intensity value (A-value) (http://www.systemsbiology.nl/).

The degree of curve-smoothing is determined by the window width parameter. A larger window width results in a smoother curve, a smaller window results in more local variation. The Lowess scatter plot smoother is not affected by a small percentage of differentially expressed genes, which appear as outliers in the MA-plot [141]. The larger the f value (the fraction of the data used for smoothing at each point) is, the smoother the fit [141].

TIGR Microarray Data Analysis System (MIDAS) is one member of a suite of microarray data management and analysis applications developed at The Institute for Genomic Research (TIGR). This program is open-source and is freely available through the TIGR website, www.tigr.org/software/tm4.

MIDAS provides two methods to normalize the data: total intensity normalization and Locfit (LOWESS) normalization. I used the LOWESS normalization since the global normalization approach would be not adequate in situation where dye biases can depend on spot overall intensity and/or spatial location within the array [141].

The MIDAS user can select some parameters to perform the normalization:

- Mode: MIDAS provides two "modes" to apply LOWESS algorithm on an input data file: either computing the LOWESS factor for each spot by assuming that all spots within a slide contribute to the bias of this spots intensity (Global mode), or computing the LOWESS factor for each spot by assuming that only those spots within the same block as this spot contribute to the bias of this spot intensity (Block mode). I chose the Blok mode option.

- Smooth parameter: percentage number used by MIDAS to compute LOWESS factor for each spot. The higher the smooth parameter is set, the more severe the channel A or channel B intensity of raw input data will be adjusted. I selected the default value that is set to 33%.
- Reference: either Cy3, I(A) or Cy5 I(B) can be selected as a reference. If Cy3 is selected as the reference, then I(A) of each spot will not be changed in the output file; however, I(B) of each spot will be adjusted by the calculated LOWESS factor for the spot. I set this parameter to Cy3 as by default.

This approach assumes that the majority of the genes on the array are non-differentially expressed between the two channels and the number of over-expressed genes is similar to the number of under-expressed genes. There is not a correlation between differential gene expression and localization of the spots (spatial bias).

3.11.3 LOGARITMIC TRANSFORMATION OF THE EXPRESSION RATIOS VALUES AND ANALYSIS OF THE MICROARRAY REPLICATES

In the microarray technology the simplest approach to identify the differentially expressed genes calculates the ratio between CH1 expression values and CH2 expression values for each spot, where CH1 and CH2 are the two channels in a microarray experiment. In my case CH1 is the query sample (patient) and CH2 the reference sample (control). Although ratios provide an intuitive measure of expression changes, they have the disadvantage of treating up- and down-regulated genes differently. Genes upregulated by a factor of 2 have an expression ratio of 2, whereas those downregulated by the same factor have an expression ratio lower than 0.5. The most widely used alternative transformation of the ratio is the logarithm base 2, which has the advantage of producing a continuous spectrum of values and considering up- and downregulated genes in a similar fashion. The logarithms of the expression ratios are treated symmetrically, so that a gene upregulated by a factor of 2 has a $\log_2$ratio of 1, a gene down-regulated by a factor of 2 has a $\log_2$ratio of -1, and a gene expressed at a constant level (with a ratio of 1) has a $\log_2$ratio equal to zero [140].

Before the logarithmic transformation of the expression ratios I obtained a single expression value for each gene from the replicates performed for that gene. Replication is essential for identifying and reducing the variation in microarray experiments [140]. Technical replicates provide information on the natural and systemic variability that occurs in the assay. Technical replicates include multiple independent elements for a particular gene within an array (such as independent oligos for a particular gene) and replicates hybridizations for a particular sample [140].

In my experiment there were two replicates within the microarray (see 3.12.1) and each experiment was replicated three times including a dye-swap procedure to avoid a bias due to the labeling. In a dye swap experiment the reference RNA and the query RNA are labeled with Cy3 and Cy5 respectively, on the first array. These dyes are then reversed for the second array. I calculated the geometric mean of two expression values of CH1 and CH2 for each spot within the array (see 3.12.1) and then the arithmetic mean between the three values of CH1 and CH2 for the replicates. Finally I performed the $\log_2$ratio transformation on the averaged value of CH1 and for CH2.

3.11.4 K-NEAREST NEIGHBOR (KNN) ALGORITHM

After the statistical filtering procedure (3.12.1), some genes did not have an expression value for each patient (sample). So the genes that had less than half plus one missing values were eliminated while, for the other genes, a KNN procedure [141] was used to determine the missing values. This algorithm works as follows.

For each gene i having at least one missing value:

1. let Si be the samples for which gene i has no missing values;
2. find the K nearest neighbours to gene i using only samples Si to compute the Euclidean distance. When computing E distances, other genes could be have missing values for some of the samples Si; the distance is averaged over the non-missing entries in each comparison;
3. impute the missing sample values in gene i, using the averages of the non-missing entries for the corresponding sample for that gene.

If a gene still has missing values after the above steps, impute the missing values using the average (non-missing) expression for that gene.

I tried three different values of K (3, 4, 5) and I chose K=5 because it corresponded to the best performance in the classification procedure.

The data from microarray experiments is usually in the form of large matrices of expression levels of genes (rows) under different experimental conditions (columns) and frequently with some missing values. Missing values occur for diverse reasons and are usually manually flagged and excluded from subsequent analysis. Many analysis methods, such as Support Vector Machines (see par. 3.15), require complete matrices; one solution is to repeat the experiment or eliminate the genes with missing values. The first strategy can be expensive and the second one will occur into loss of data. Missing data are often replaced by zeros or, less often, by the average expression over the row, or "row average". These approaches are not optimal, since they do not take into consideration the correlation structure of the data. Instead KNN impute seems to provide a more robust and sensitive method for missing value estimation since it takes advantage from the correlation structure of the data [141].

I used the KNN algorithm implemented in PAM (Prediction Analysis of Microarray) [143] to determine the missing values.

3.12 HIERARCHICHAL CLUSTERING WITH TMEV (*TIGR MultiExperimentViewer*) SOFTWARE

TIGR MultiExperiment Viewer (TMEV), one member of the suite of microarray data analysis programs is an application that allows the visualization of processed microarray slide representations and the identification of genes and expression patterns of interest.

TMEV is composed by several modules (fig. 3.26), to perform different types of analysis in the same work session. Each program implemented in TMEV has a dialog window where the user can insert the parameters of interest. MEV can interpret different file formats, including the MultiExperiment Viewer format (.mev), the TIGR ArrayViewer format (.tav), the TDMS file format (Tab Delimited, Multiple Sample format), the Affymetrix file format, and GenePix fileformat (.gpr).

In my analysis the input file, a TDMS file, contains a matrix of log2ratio expression values for each gene (rows) in each patient (columns).

To perform an unsupervised cluster analysis I used the HCL (Hierarchical Clustering) module of TMEV, an agglomerative algorithm that arranges genes and patients according to similarity in the gene expression pattern.

The object of a hierarchical clustering is to compute a dendrogram that assembles all elements into a single tree [144]. For any set of n genes, an upper-diagonal similarity matrix is computed, which contains similarity scores for all pairs of genes. The matrix is scanned to identify the highest value (representing the most similar pair of genes). A node is created joining these two genes, and a gene expression profile is computed for the node by averaging observation for the joined elements. The similarity matrix is updated with this new node replacing the two joined elements, and the process is repeated n-1 times until only a single element remains [144]. Agglomerative algorithms begin with each element as a separate cluster and merge them into larger clusters. An important step in any clustering process is to select a distance measure, which will determine how the similarity of two elements is calculated. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and further away according to another. TMEV allows to calculate the distance with different approaches, in this study I chose the Euclidean distance method. Another parameter to set is the "Linkage Method" that indicates the approach used for determining cluster-to-cluster distances, when constructing the hierarchical tree. I used the "Average Linkage" method that uses the average distance of each member of one cluster to each member of the other cluster as a measure of cluster-to-cluster distance. This option in MeV is determined by a weighted average of distances of cluster members.

The cluster analysis visualization of TMEV consists of colored rectangles, representing genes (fig. 3.27). Each column represents all the genes from a single experiment, and each row represents the expression of a gene across all experiments. The default color scheme used to represent expression level is red/green (red for overexpression, green for underexpression); black rectangles are not-differentially expressed genes. In the upper and left part of the graph is reported the dendogram structure that represents the correlation between genes (or patients): more nodes separate the genes.

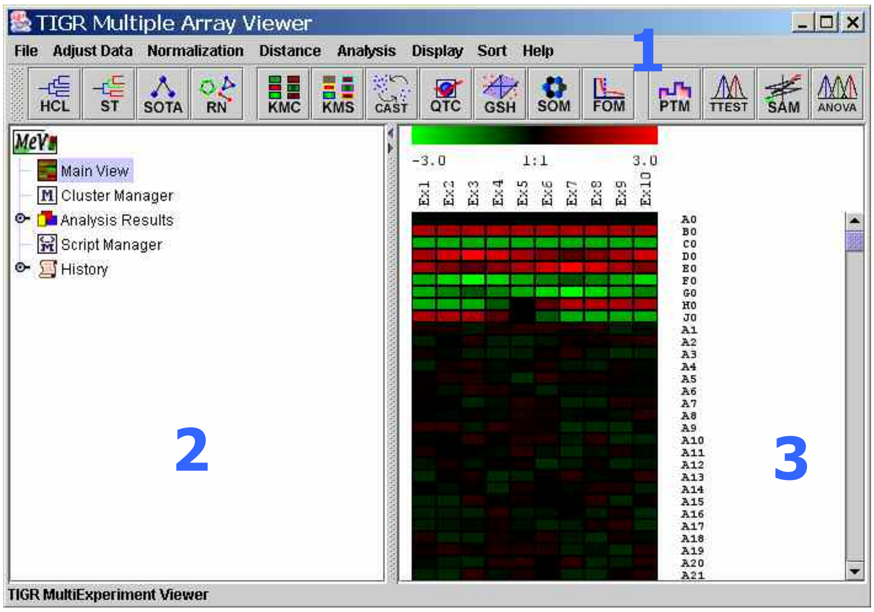

TIGR Multiple Array Viewer
File Adjust Data Normalization Distance Analysis Display Sort Help
1
HCL ST SOTA RN KMC KMS CAST QTC GSH SOM FOM PTM TTEST SAM ANOVA
MeV
Main View
Cluster Manager
Analysis Results
Script Manager
History
2
-3.0 1:1 3.0
Ex1 Ex2 Ex3 Ex4 Ex5 Ex6 Ex7 Ex8 Ex9 Ex10
A0 B0 C0 D0 E0 F0 G0 H0 J0 A1 A2 A3 A4 A5 A6 A7 A8 A9 A10 A11 A12 A13 A14 A15 A16 A17 A18 A19 A20 A21
3
TIGR MultiExperiment Viewer

Figure 3.26: Main view in TMEV: 1= TMEV modules 2= windows containing the Analysis Results 3= graphical output of the analysis (TMEV User Manual)

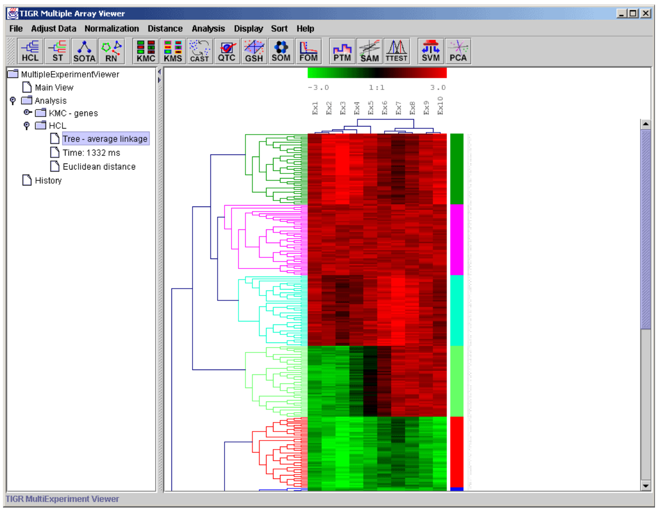

TIGR Multiple Array Viewer
File Adjust Data Normalization Distance Analysis Display Sort Help
HCL ST SOTA RN KMC KMS CAST QTC GSH SOM FOM PTM SAM TTEST SVM PCA
MultipleExperimentViewer
Main View
Analysis
KMC - genes
HCL
Tree - average linkage
Time: 1332 ms
Euclidean distance
History
-3.0 1:1 3.0
Ex1 Ex2 Ex3 Ex4 Ex5 Ex6 Ex7 Ex8 Ex9 Ex10
TIGR MultiExperiment Viewer

Figure 3.27: Hierarchical tree with clusters selected (TMEV User Manual)

3.13 PAM (PREDICTION ANALYSIS OF MICROARRAY)

PAM (Prediction Analysis of Microarray) (http://www-stat.stanford.edu/~tibs/PAM/) is a statistical software for class prediction from gene expression data using nearest shrunken centroids; it is described in Tibshirani and colleagues [145]. The method of nearest shrunken centroids identifies subsets of genes that best characterize each class. It 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 that it is closest to, in squared distance, is the predicted class for that new sample. Nearest shrunken centroid classification "shrinks" each of the class centroids toward the overall centroid for all classes by an amount they call the threshold. This shrinkage consists of moving the centroid towards zero by threshold, setting it equal to zero if it hits zero. For example if threshold was 2.0, a centroid of 3.2 would be shrunk to 1.2, a centroid of -3.4 would be shrunk to -1.4, and a centroid of 1.2 would be shrunk to 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 it shows that high or low expression for that gene characterizes that class. The user selects the threshold value on the basis of the K-fold cross-validation procedure that PAM performs for a range of threshold values. The samples are divided up at random into K roughly equally sized parts (K is set to 10). For each part in turn, the classifier is built on the other K-1 parts then tested on the remaining part. This is done for a range of threshold values, and the cross-validated misclassification error rate is reported for each threshold value. Typically, the user would choose the threshold value giving the minimum cross-validated misclassification error rate.

PAM handles three different problems: a standard classification problem (my choice), survival analysis and regression. The input file is an excel spreadsheet, .xls format. In a standard classification problem PAM requires the training data set containing the "Class Labels".

After setting the threshold, the program supplies three more worksheets to the workbook: 1- "PAM plots" contains plots that can be produced 2- "PAM Output" contains the list of significant genes (when it is asked for) 3- "PAM Worksheet" used for writing intermediate calculations and data used for plotting.

3.14 SUPPORT VECTOR MACHINES (SVM$_S$) AND FEATURE SELECTION

3.14.1 SUPPORT VECTOR MACHINES (SVMs)

Molecular classification approaches based on SVMs applied to microarray data have shown to have statistical and clinical relevance [146]. Support Vector Machines (SVMs) [147] are a particular machine learning algorithm used for classification and regression problems. In this study we considered a binary classification problem with linearly separable patterns (or vectors). Assuming that input data are two sets of vectors, positive and negative, in an *n*-dimensional space, a SVM will construct a separating hyperplane in that space, that maximizes the margin of separation between the two data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are "pushed up against" the two data sets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance from the neighbouring data points of both classes, since in general the larger the margin the better the generalization error of the classifier. The generalization error is a function that indicates the capacity of the SVM to classify also the examples not included in the set of examples (training set) used to generate the classification function.

In my study a vector, that represented a patient, was constituted by a number of components or "features" (n), which were gene expression coefficients. As I previously reported, the SVM analysis was applied to a two-class classification problems, positive examples (responders patients) and negative examples (not responders patients), that constituted the training set. When the training set was linearly separable, a linear SVM was a maximum margin classifier [148]. The decision boundary was positioned in order to have the largest possible margin on either side.

A particularity of SVMs is that the weights w_i of the decision function are a function only of a small subset of the training examples, called "support vectors". Those are the examples that are closest to the decision boundary and lie on the margin [148]. A geometric interpretation of the SVM illustrates how this idea of smoothness or stability gives rise to a geometric quantity called margin which is a measure of how well separated the two classes can be [146]. We start by assuming that the classification function is linear:

$$\boxed{f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} = \sum_{i=l} w_i x_i}$$

where x_i and w_i are the i_{th} elements of the vector **x** and **w**, respectively. The operation $\mathbf{w} \cdot \mathbf{x}$ is called a "dot product". The label of a new point $\mathbf{x_{new}}$ is the sign of the above function, y_{new} = sign [$f(\mathbf{x_{new}})$]. The classification boundary, all values of **x** for which $f(\mathbf{x}) = 0$, is a hyperplane defined by its normal vector **w** (fig. 3.28).

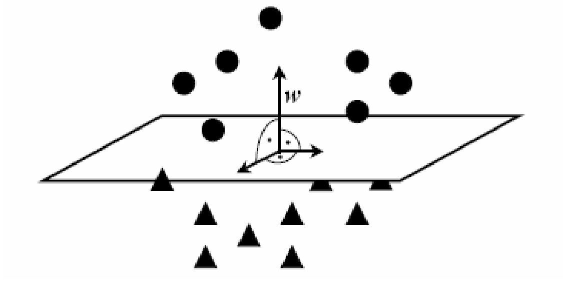


Figure 3.28: The hyperplane separating two classes. The circles and the triangles designate the members of the two classes. The normal vector of the hyperplane is the vector **w** [146].

The linearly separable SVM problem (i.e. the hyperplane that separates the two classes of vectors) is written as:

$$\min_{(w,b)} 1/2\ ||\mathbf{w}||^2 \qquad \text{subject to} \qquad yi(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1 \qquad [*]$$

where **w** is the normal vector of the hyperplane and b is a free threshold parameter that translates the optimal hyperplane relative to the origin. The distance from the hyperplane to the closest points of the two classes is the margin that is defined by $1/||\mathbf{w}||^2$. SVMs find the hyperplane that maximize the margin. The figure 3.29 illustrates the advantage of a large margin.

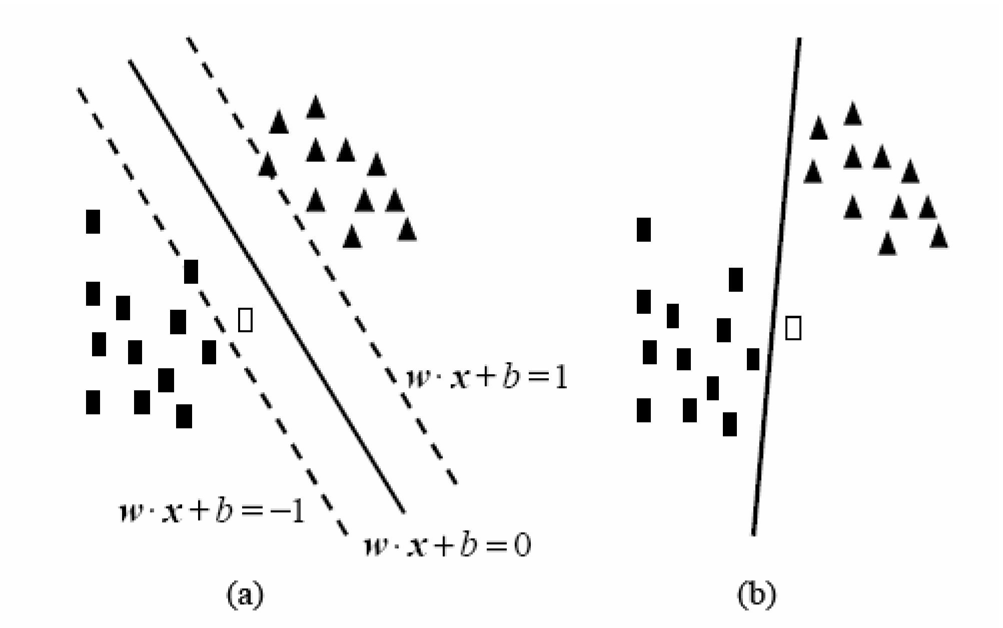


Figure 3.29: (a) The maximum margin hyperplane separating two classes. The solid black line is the hyperplane (**w** * **x** + b = 0). The two dashed lines are those for the points in the two classes closest to the hyperplane (**w** * **x** + b = ±1). A new point, the black retangle, is classified correctly in (a). Note, the larger the margin the greater the deviation allowed or margin for error. (b) A non-maximum margin hyperplane separating the two classes. Note that the same new point is now classified incorrectly. There is less margin for error [146].

Data sets are often not linearly separable. To deal with this situation, slack variables (ξ_i) are added that allow to violate the original distance constraints. The equation [*] becomes now:

$$\min_{(\mathbf{w}, b, \xi)} 1/2 \, ||\mathbf{w}||^2 + C \, \Sigma \, \xi_i \quad \text{subject to} \quad yi(\mathbf{w} \, \mathbf{x_i} + b) \geq 1 - \xi_i \qquad [**]$$

where $\xi_\mathbf{i} \geq 0$ for all i. The new formulation [**] trades off the two goals of finding a hyperplane with large margin (minimizing ||**w**||) and finding a hyperplane that separates the data well (minimizing the ξ_i). The parameter C controls this trade-off determining the "soft margin SVM" [146]. The figure 3.30 illustrates the new approach.

SVMs can also be used to construct nonlinear separating surface. The basic idea here is to nonlinearly map the data to a feature space of high or possible infinite dimensions, **x** → Φ (**x**). Then the linear SVM is applied in this feature space. A linear separating hyperplane in the feature space corresponds to a nonlinear surface in the original space. In this situation, the dot product can be computed without explicitly mapping the points into feature space by a "kernel function", which is defined as the dot product for two points in the feature space:

$$K \, (\mathbf{x_i}, \mathbf{x_j}) \equiv \Phi \, (\mathbf{x_i}) \; * \; \Phi \, (\mathbf{x_j})$$

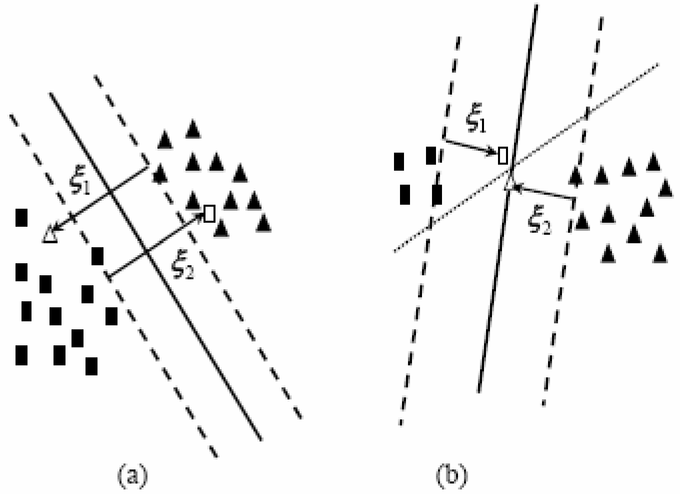


Figure 3.30: a) The data points are not linearly separable. The solid black line is the SVM solution. The white triangle and the white rectangle are misclassified. The slack variables designate the distance of these points from the dashed lines for the corresponding classes. b) The classes are separable. The dotted line is the solution when the trade-off parameter C is very large (e.g., infinite), and this gives us the maximum margin classifier for the separable case. If the trade-off parameter is small, then one allows errors (given by the two slack variables), but one gets a much larger margin [146].

3.14.2 FEATURE SELECTION USING SUPPORT VECTOR MACHINES

Since my goal was to select a subset of features with the maximum discriminatory power between the two classes of patients (see chapter 2), it was important to know which genes were most relevant to the binary classification task. The gene selection problem is an example of what is called "feature selection" in machine learning.
A known problem in classification is to find ways to reduce the dimensionality *n* of the feature space F to overcome the risk of "overfitting" [148]. Data overfitting arises when the number *n* of features is large (in this case thousand of genes) and the number l of training patterns is comparatively small (in this case a few dozen patients). In such a situation, one can easily find a decision function that separates the training data but will perform poorly on test data. This makes many standard pattern classification algorithm fail [149]. For machine learning methods such as SVMs that can work at high-dimensionality, dimension reduction can improve the performance. However, when validating the performance of a classification algorithm with feature-selection steps, the feature selection procedure should also be validated simultaneously to avoid bias in the assessment [150]. Also, due to the small sample size, the cross-validation prediction of the algorithm's performance tends to have a high variance. Thus it is necessary to pay more attention to properties related to generalization ability rather than prediction performance *per se*.
Methods for automated feature selection can be divided into two categories (fig. 3.31): filtering approaches, meaning that feature selection is carried out in a pre-processing step of classification, independent from the choice of the classification method, and wrapper approaches, meaning that a classifier is used to generate scores for features in the selection process and feature selection depends on the choice of the classifier. Filtering methods are not

the best choices because they score the importance of features independently, ignoring the correlations among them [151].
The SVM feature selection is a wrapped method.

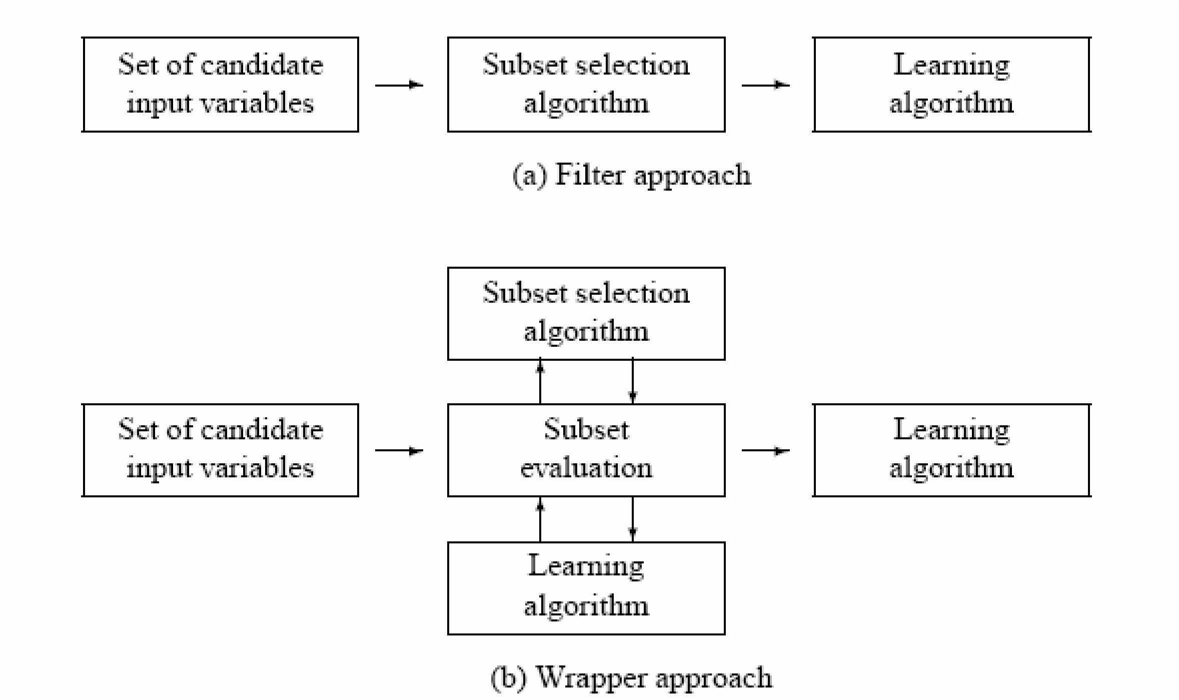

Set of candidate input variables
Subset selection algorithm
Learning algorithm
(a) Filter approach
Subset selection algorithm
Set of candidate input variables
Subset evaluation
Learning algorithm
Learning algorithm
(b) Wrapper approach

Figure 3.31: Two approaches of feature selection: (a) Filter procedure and (b) Wrapper procedure [152].

In my analysis I evaluated two different approaches to perform the feature selection: Recursive Feature Elimination SVM (RFE-SVM) [148] and Recursive SVM feature selection (R-SVM) [150]. Finally I chose the R-SVM approach because was the most correct in terms of cross-validation scheme (see par. 3.14.2.3 for more details).

3.14.2.1 Recursive Feature Elimination SVM (RFE-SVM)

The method recursively removes features based upon the absolute magnitude of the hyperplane elements. Given microarray data with n genes per sample, the SVM outputs the normal to the hyperplane, w, which is a vector with n components, each corresponding to the expression of a particular gene (see 3.15.1) [146]. Assuming that the expression values of each gene have similar ranges, the absolute magnitude of each element in w determines its importance in classifying a sample, since the following equation holds:

$$f(\mathbf{x}) = \mathbf{w} * \mathbf{x} + b = \sum_{i=l} w_i x_i + b$$

The idea behind RFE is to eliminate elements of w that have small magnitude, since they do not contribute much in the classification function. The SVM is trained with all genes; then is compute the following statistic for each gene:

$$S(j) = |w_j|$$

where w_j is the value of the j^{th} element of **w**. Then are sorted S from largest to smallest value and are removed the genes corresponding to the indices that fall at the bottom of the sorted list S. The SVM is retrained on this smaller gene expression set, and the procedure is repeated until a desired number of genes, m, is obtained. The percentage of features with smallest ranking criterion that have to be removed at each iteration, is selected from the user (10%, 50%, 75% etc.). For my analysis I used the RFE-SVM method implemented in the software package Gist 2.3 (http://bioinformatics.ubc.ca/gist/). In particular Gist 2.3 uses a ranking criterion based on the square of weight w_j^2 (instead of $|w_j|$) and removes the features that fall in the bottom 50% of the sorted list S.

3.14.2.2 Recursive SVM feature selection (R-SVM)

This method uses a similar recursively procedure of RFE-SVM but ranks the features according to a different ranking criterion, s_j called "contribution factor of feature j", computed as:

$$s_j = w_j (m_j^+ - m_j^-)$$

where m_j^+ and m_j^- are the means of feature j in the two classes. Thus the factor s_j is not only decided by the weight w_j in the classifier function, but also by the data (class-means) [153]. To perform the R-SVM feature selection I used the algorithm freely available in the software package of Zhang and colleagues [150] (http://www.hsph.harvard.edu/bioinfocore/RSVMhome/R-SVM.html).
Also in R-SVM method was removed the 50% of features low-ranked each time of the iterative procedure. The feature selection procedure proceeds until the number of features selected is bigger than the threshold set from the user (in this case we used the default value equal to 5).

3.14.2.3 Assessing the performance of feature selection

Since an independent test set is not available in many investigations, cross-validation (e.g. leave one out cross-validation or LOO-CV) is often used to assess the accuracy of classifiers. LOO-CV uses a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. This is the same as a K-fold cross-validation with K being equal to the number of observations in the original sample. It should be noted that feature selection results may vary with even a single-case difference in the training set when the sample size is small (as the case of my study) [150]. There are two approaches of assessing the performance of feature selection:

1) CV1: feature selection steps are external to the cross-validation procedure, i.e. the feature selection is done with all the samples and the cross-validation is only done for the classification procedure. CV1

may severely bias the evaluation in favour of the studied method due to “information leak” in the feature selection step [150].
2) CV2: feature selection steps are included in the cross validation procedure, i.e., to leave the test sample(s) out from the training set before undergoing any feature selection. In this way, not only the classification algorithm, but also the feature selection method is validated [150].

The RFE-SVM approach uses the CV1 scheme. In fact ranks the genes only once using all samples, and uses the top ranked genes in the succeeding cross-validation for the classifier. This scheme generates a biased estimation of errors
The R-SVM approach follows the CV2 scheme to estimate the error rate at each level (see fig. 3.32). In cross-validation experiments, different training subsets generate different lists of features. In the R-SVM method after the recursive feature selection steps on each subset, are counted at each of the d_i levels the frequency of the features being selected among all rounds of cross-validation experiments. The top d_i most frequently selected features are reported as the final d_i features, called “the top features” [150].

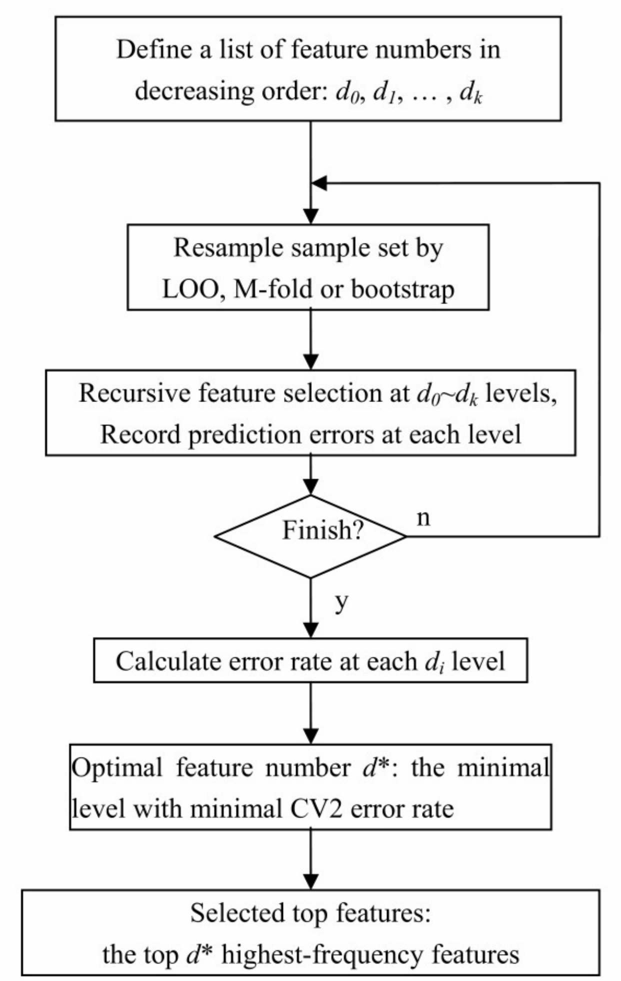
3.14.3 PROBABILISTIC OUTPUTS FOR SUPPORT VECTOR MACHINES

Since SVMs produce an uncalibrated value that is not a probability; constructing a classifier to produce a posterior probability (P) is very useful in practical recognition situations [153]. We used the Platt’s algorithm [153] to map the SVM outputs into probabilities.
Platt uses a parametric model to fit the posterior probability P directly. The parameters of the model are adapted to give the best probability outputs. The form of the parametric model is a sigmoid (fig. 3.33):

$$P(y=1 \mid f) = 1/\ (1 + \exp\ (A\ f + B)$$

This sigmoid model is equivalent to assuming that the output of the SVM is proportional to the log odds of a positive example. The sigmoid function has two parameters, A and B, trained discriminatively.
Simply speaking the procedure consisted of training the SVM using the features selected during the feature selection procedure (3.15.2), then training the parameters of the sigmoid function to map the SVM outputs into probabilities. In practice the SVM outputs were a measure of distance of the patients (vectors) from the optimal hyperplane (see 3.15.1) and the sigmoid function translated this distance in measure of probability, more useful in statistics. It is reported that the sigmoid fit works well even beyond the margins and seems to be close to the true model [153]. In order to avoid a biased training set, the sigmoid function was trained using a LOO-CV for a number of times depending from the number of the examples (vectors). It means that the sigmoid was trained on N-1 examples (where N= total number of examples or

patients) for N times (i.e. obtaining N sigmoids) and tested on the left example. The final parameters of the sigmoid function were derived from the N pairs of A and B parameters obtained with the sigmoid training.



Define a list of feature numbers in decreasing order: $d_0, d_1, \dots, d_k$
Resample sample set by LOO, M-fold or bootstrap
Recursive feature selection at d_0~d_k levels, Record prediction errors at each level
Finish?
n
y
Calculate error rate at each d_i level
Optimal feature number d^*: the minimal level with minimal CV2 error rate
Selected top features: the top d^* highest-frequency features

Figure 3.32: Workflow of the R-SVM algorithm [150]

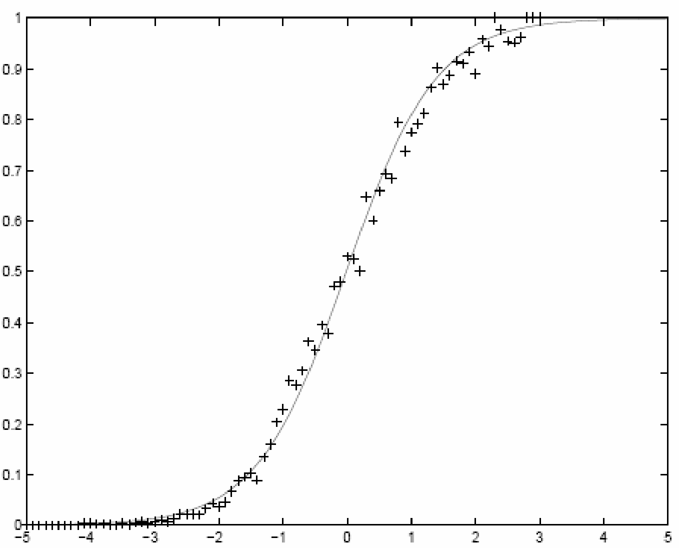


Figure 3.33: The fit of the sigmoid to the data for a linear SVM on a dataset taken as

example. Each plus mark is the posterior probability computed for all examples falling into a bin of width 0.1. The solid line is the best-fit sigmoid to the posterior, using the Platt's algorithm [153].

3.15 STATISTICAL TESTS USED IN THE STUDY

3.15.1 SPEARMAN'S RANK CORRELATION COEFFICIENT

In statistics, Spearman's rank correlation coefficient or Spearman's rho, ρ, is a non-parametric measure of correlation: it assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables [154]. Spearman's correlation does not require the assumption that the relationship between the variables is linear, nor does it require the variables to be measured on interval scales; it can be used for variables measured at the ordinal level. In principle, ρ is simply a special case of the Pearson product-moment coefficient in which the data are converted to rankings before calculating the coefficient. The raw scores are converted to ranks, and the differences d between the ranks of each observation on the two variables are calculated. ρ is given by:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where:
d_i = the difference between each rank of corresponding variables
n = the number of pairs of values

The size of this correlation is evaluated as follows:

- rs <0,33: small correlation between two variables
- 0,33<rs<0,67: medium correlation between two variables
- rs>0,67: large correlation between two variables

I used this test to asses the correlation between immunohistochemical prognostic markers (see 3.2.3) and microarray expression data. I correlated log2 ratio of microarray fluorescence intensity and positive ("1") or negative ("0") immunostaining of the markers.

The calculation of Spearman's correlation coefficient was done using a function implemented in the Winstat package (http://www.winstat.com).

3.15.2 FISHER'S EXACT TEST

Fisher's exact test is a non-parametric statistical significance test used in the analysis of categorical data where sample sizes are small [155].

The test is usually used to examine the significance of the association between two variables in a 2 x 2 contingency table. The null hypothesis is that the relative proportions of one variable are independent of the second variable. With large samples, a chi-square test can be used in this situation. However, this test is not suitable when the expected values in any of the cells of the table are below 10: the sampling distribution of the test statistic that is calculated is only approximately equal to the theoretical chi-squared distribution, and the approximation is inadequate when sample sizes are small. The Fisher's exact test is, as its name states, exact, and it can therefore be used regardless of the sample characteristics. The test does not make any assumptions about the frequency distribution of the variables.

I used this test to assess the correlation between immunohistochemical prognostic markers (see 3.2.3) and clinical response to neoadjuvantchemotherapy. I performed the Fisher's exact test using a function implemented in the Winstat package (http://www.winstat.com).

3.16 BIOINFORMATIC TOOLS AND DATABASES

3.16.1 GENE ONTOLOGY (http://www.geneontology.org)

The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The project began as collaboration between three model organism databases, FlyBase external link (Drosophila), the Saccharomyces Genome Database external link (SGD) and the Mouse Genome Database external link (MGD), in 1998. The GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. In particular:

- Biological process: is series of events accomplished by one or more ordered assemblies of molecular functions. It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps.
- Cellular component: is a component of a cell, but with the proviso that it is part of some larger object; this may be an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).
- Molecular function: describes activities, such as catalytic or binding activities, that occur at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products. It is easy to confuse a gene product name with its molecular function, and for that reason many GO molecular functions are appended with the word "activity".

A gene product might be associated with or located in one or more cellular components; it is active in one or more biological processes, during which it performs one or more molecular functions.

The building blocks of the Gene Ontology are the terms identified a a entry in GO database with a unique numerical identifier of the form GO:nnnnnnn, and a term name, e.g. cell, fibroblast growth factor receptor binding or signal transduction. Each term is also assigned to one of the three ontologies, molecular function, cellular component or biological process. The ontologies are structured as directed acyclic graphs (DAGs), which are similar to hierarchies but differ in that a more specialized term (child) can be related to more than one less specialized term (parent)

The GO consortium takes care of three aspects:

- development and maintenance of the ontologies themselves;
- annotation of gene products, which entails making associations between the ontologies and the genes and gene products in the collaborating databases;
- development of tools that facilitate the creation, maintenance and use of ontologies

Collaborating databases annotate their genes or gene products with GO terms, providing references and indicating what kind of evidence is available to support the annotations.

3.16.2 GOMINER (http://discover.nci.nih.gov/gominer/index.jsp)

GoMiner is a program package that organizes lists of "interesting" genes (for example, under- and overexpressed genes from a microarray experiment) for biological interpretation in the context of the Gene Ontology (see 3.17.1) [157]. GoMiner is a freely available computer resource that fully incorporates the hierarchical structure of the Gene Ontology to automate the functional categorization of gene lists of any length [157]. GoMiner was developed particularly for biological interpretation of microarray data; one can input a list of under- and overexpressed genes and a list of all genes on the array, and then calculate enrichment or depletion of categories with genes that have changed expression [157]. The user flag the genes overexpressed with "1"and the genes underexpressed with "-1" and are accepted all types of identifiers (EntrezGene ID, Genbank ID, Unigene ID, etc.) used from the GO consortium. GoMiner displays the genes within the framework of the Gene Ontology hierarchy, both as a directed acyclic graph (DAG) and as the equivalent tree structure. Each category is annotated to reflect the number of genes from the user's experiment assigned to that category plus the number assigned to its progeny categories.

The most important parameter for purposes of interpretation is the enrichment (or depletion) of a category with respect to flagged genes. The two-sided Fisher's exact test p-value for a category reflects a test of the null hypothesis that the category is neither enriched in, nor depleted of, flagged genes with respect to what relative to what would have been expected by chance alone. It reflects the null hypothesis (1) that, for each category, there is no difference between the proportion of flagged genes that fall into the category (p_1) and the proportion of flagged genes that do not fall into the category (p_2):

$$H_o: p_1 - p_2 = 0 \quad (1)$$

where $p_1 = n_f/n$ and
$p_2 = (N_f - n_f)/(N - n)$
n_f number of flagged genes in category
n total number of genes in the category

N_f number of flagged genes on the microarray
N total number of genes on the microarray

Another useful measure that the program calculates is the relative enrichment factor R_e defined as:

$$R_e = (nf/n)/(Nf/N) \quad (2)$$

4. RESULTS AND DISCUSSION

4.1 PATIENT CHARACTERISTICS

At the time of analysis 54 patients has been collected in this study. From 41 pre-treatment biopsies of these patients, good quality RNA was obtained and gene expression profiling was performed (see the fig. 4.1 to have a complete overview). Since 9 of these 41 patients left the study before completing the chemotherapy treatment, I could not include them in the subsequent statistical analysis. Therefore I considered 34 patients (out of 41), of which were available informations of the NeoAdjuvant ChemoTherapy (NACT) treatment: 28 have received Adriamycin and Taxol (AT), 5 have received Epirubicin and Taxol (ET) and 1 has received Adriamycin alone (A). Four courses of NACT were administered every three weeks to the patients.

From these 34 patients we collected the clinical responses after the treatment: 3 patients achieved a clinical Complete Response (cCR), 18 patients had a Partial Response (PR), 11 patients showed No Change (NC) in the tumour mass size and 2 patients showed Progressive Disease (PD). Unfortunately it was not possible to obtain the pathological responses for all the patients (only 13/34 pathological responses available), so I decided to consider for the study the clinical responses.

For 34 patients out of 41 were available both the immunohystochemical data for ER, PR, c-erbB-2, Ki67, p53 and Bcl-2 markers (see par. 3.2.3) and the array data.

The patient characteristics are summarized in the table 4.1.

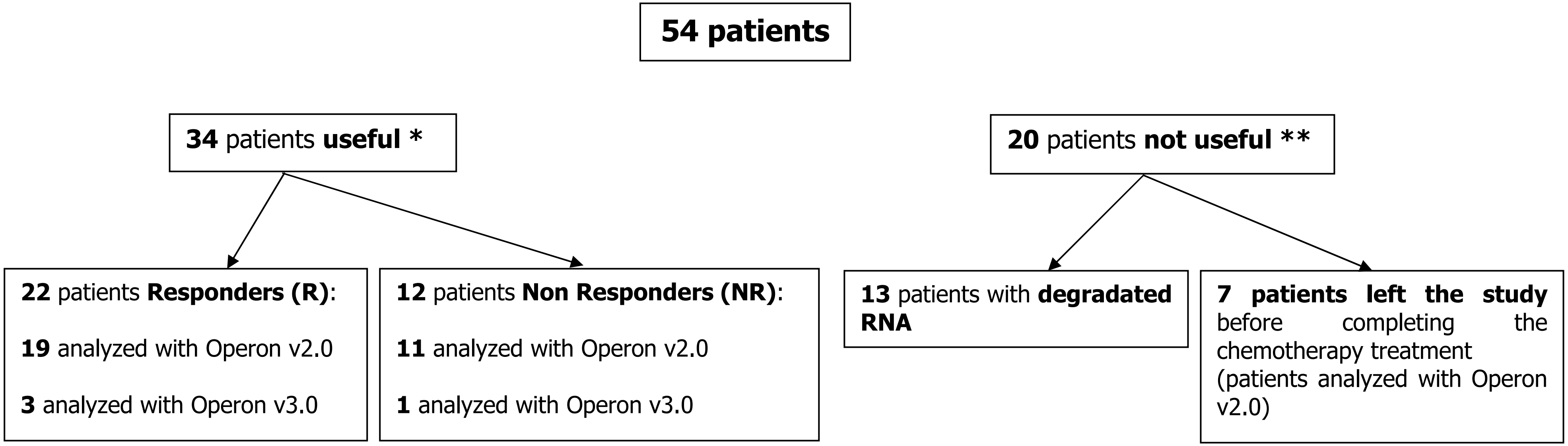
Patient characteristics		
		No. of Patients (41 patients considered)
Age		
Median (years)	56	
Range (years)	36-82	
Histology		
IDC		24
ILC		3
DCIS		1
not assessable		13
Tumour diameter	≥ 2 cm	
Immunostaining		
ER +/-		26/8
PR +/-		23/11
c-erbB-2 +/-		19/14 (1 na**)

Herceptest +/-		3/3 (28 na**)			
p53 +/-		15/19			
Bcl-2 +/-		22/12			
Ki67 +/-		25/9			
na *		7			
Neoadjuvant Chemotherapy					
4 X AT		28			
4 X ET		5			
4 X A		1			
not assessable*		7			
Clinical responses					
		cCR	**PR**	**NC**	**PD**
4 X AT		3	15	8	2
4 X ET		0	3	2	0
4 X A		0	0	1	0
not assessable*	7				

Table 4.1: Patient characteristics. Samples were considered to be positive (+) for ER/PR/c-erbB-2/p53/Ki67 when at least 10% of the tumour cells were stained positive, negative if < 10% (-) of the tumour cells were stained positive. Samples were scored as Bcl-2 positive (+) with at least 25% of the tumour cells positive, Bcl-2 negative (-) with < 25% of the tumour cells positive. IDC Invasive Ductal Carcinoma, ILC Invasive Lobular Carcinoma, DCIS Ductal Carcinoma *In Situ*; ER Estrogen Receptor, PR Progesteron Receptor, +/- positive/negative; AT Adriamycin/Taxol, A Adriamycin, ET Epirubicin/Taxol; cCR clinical Complete Response, PR Partial Response, NC No Change, PD Progressive Disease.*na (not assessable) on the total number of patients (41), **na on the total number of patients with available IHC data (34).

Because of the low number of the cCR and PD patients, 3 and 2 cases respectively, I chose to divide the patients in two main groups in term of NACT clinical response, the Responders (R) and the Non Responders (NR). The Responders included the patients with cCR and PR clinical responses, the Non Responders the patients with NC and PD clinical responses.

The common trait of the R (cCR + PR) patients was that they showed a positive response to the NACT treatment, although on different levels. In fact it is important to remark that a cCR patient shows a complete response, meaning that the treatment, at the clinical exam, is successful. Instead a PR patient shows a partial remission of the tumour, but not a complete remission, meaning that the treatment is not fully effective at the check point. Nevertheless, when I performed an unsupervised clustering analysis considering all responder patients (cCR + PR), the 2 cCR patients considered (only the cCR analyzed with Operon v2.0) did not cluster separately from the PR patients (fig. 4.2). This result could indicate that in terms of whole gene expression profile the cCR patients did not show so evident differences in respect to the PR patients. Thus I decided to include in the same group of Responders, cCR and PR patients.


54 patients
34 patients useful *
20 patients not useful **
22 patients Responders (R):
19 analyzed with Operon v2.0
3 analyzed with Operon v3.0
12 patients Non Responders (NR):
11 analyzed with Operon v2.0
1 analyzed with Operon v3.0
13 patients with degradated RNA
7 patients left the study before completing the chemotherapy treatment (patients analyzed with Operon v2.0)

*** 34 can be included** in the study, but so far were considered only the **19** + **11** patients analyzed with Operon v2.0.

**** 20** patients **can not be included** in the study because was not possible to perform the microarray experiments (poor quality RNA) or recover the clinical responses (the patients left the study). But: 7 of these patients are used for some statistical analyses (see par. 4.2 and par. 4.4 because were available the microarray data.

Figure 4.1: Overview of the patients collected in the study. The patients were called as **useful** if the clinical responses to the NACT treatment were available, as **not useful** on the contrary. The 34 patients useful were divided in Responders (Partial Response and Complete Response patients) and Non Responders (No Change and Progressive disease patients). Of the 34 patients useful, 4 patients were analyzed at the Netherlands Cancer Institute (NKI) with the Operon platform v3.0 (see par. 4.5). The 7 patients which left the study before completing the chemotherapy treatment were anyway analyzed with Operon v2.0 platform and included for some statistical analysis (see par 4.2 and 4.4).

Moreover, it has to be reminded that the patients defined PR can show a reduction of tumour mass from 50% until 75%; thus, it could be possible that a patient with a reduction equal to 75% is closer to a cCR patient than a patient with a reduction of 50%. Unfortunately, I could not stratify the PR samples on the basis of the percentage of tumour mass reduction because the data were not available. Hannemann and colleagues in a similar study [97], did not include the PR patients in the responder group, considering them a group not enough homogenous in terms of clinical response. However, since in this study the PR group was the most numerous group, including it in the analysis was the only possible choice in terms of a statistical evaluation. The Non Responders are a more homogenous group than the Responders because in both cases (NC and PD patients) the treatment is ineffective.

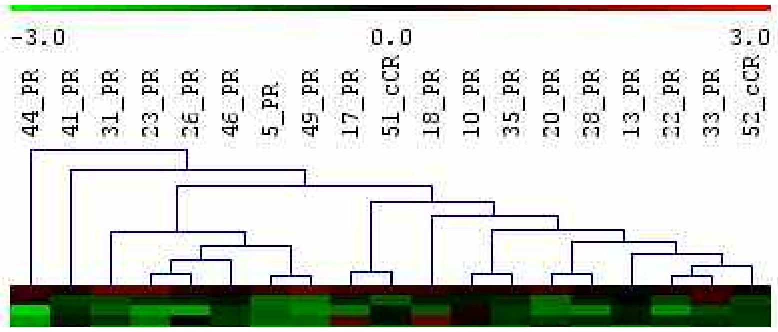


Figure 4.2: Unsupervised clustering analysis of all responder patients analyzed with Operon v2.0 platform. The cCR patients did not cluster separately from the PR patients. PR (Partial Response) and cCR (complete Clinical Response). TMEV (see par. 3.13) was used to perform the hierarchical clustering with the average linkage method.

4.2 CORRELATION ANALYSIS BETWEEN IMMUNOHISTOCHEMICAL DATA AND MICROARRAY DATA

Although the main goal of this study was to identify a predictive signature of responsiveness to NACT, I decided to evaluate, first of all, if there was a correlation between the ImmunoHistoChemical (IHC) data and the array data. For this analysis I considered only the patients analyzed with the Operon v2.0 platform (see fig. 4.1) with IHC and array data available (see tab. 4.2 for details), 34 patients in all. The markers that I considered, ER (Estrogen Receptor), PR (Progesteron Receptor), Erb-B2 (v-Erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog [*avian*]), Bcl-2 (B-cell CLL/lymphoma 2), Ki67 (antigen identified by monoclonal antibody Ki-67) and p53 (tumour protein 53), are prognostic markers routinely used in breast cancer clinical diagnosis. However, so far, their association with the response to NACT is not yet fully demonstrated and published data are controversial (see par. 4.3).

In order to perform the comparison between the percentage of positive tumour cells for a specific marker, analyzed by immunohistochemistry, and its mRNA abundance, measured with microarrays, I decided to consider the averaged logarithmic-transformed ratio (patient/control) of the gene expression value of the three performed replicates (see Methods). Indeed it is more correct to take a ratio, instead the fluorescence absolute intensity value, to reduce the experimental bias. I substituted the real IHC values (percentage of positive cells in the immunostaining) with "1" if the marker was positive and with "0" if it was negative. As I previously reported (see par. 3.2.3), samples were scored as ER, PR, c-erbB-2, Ki67, p53 positive by IHC when at least 10% of the tumour cells showed staining of these markers and Bcl-2 positive with ≥ 25% of the cells stained for Bcl-2. This transformation makes easier the statistical tests without altering the reliability of the results. It is a method commonly used to compare two series of data with a different range of values, as the case of microarray data and immunohistochemical data. For the microarray values I could hypothesize a normal distribution, but not for the IHC values ("1" or "0"); so it was not possible to apply a classical t-test to perform the correlation analysis. For this reason I used the Spearman's rank correlation coefficient (r_s) (see par. 3.15.1), a statistical test that measures the correlation between two variables, without making any assumptions about their frequency distribution.

In the table 4.3 are shown the results of the Spearman's rank correlation test. There was a good correlation between the IHC data and microarray data for ER (rs=0.678, p-value=5.169E-06) and PR (rs=0.678, p-value=9.047E-05), two of the most used prognostic markers in the breast cancer research. This result is in agreement with the study of Pusztai and colleagues but they also reported a significant correlation with Erb-B2 that I did not find in my analysis [21].

Patient code	ER %	PR %	Erb-B2		Bcl-2 %	Ki67 %	p53 %
			%	Herceptest			
4	70 (1)	60 (1)	80 (1)	na	40 (1)	55 (1)	0 (0)
5	80 (1)	80 (1)	40 (1)	na	80 (1)	10 (1)	5 (0)
6	95 (1)	95 (1)	0 (0)	na	0 (0)	15 (1)	2 (0)
7	100 (1)	30 (1)	0 (0)	na	90 (1)	20 (1)	5 (0)
8	90 (1)	90 (1)	0 (0)	na	10 (0)	7 (0)	70 (1)
10	75 (1)	75 (1)	5 (0)	na	15 (0)	45 (1)	70 (1)
12	100 (1)	100 (1)	0 (0)	na	100 (1)	20 (1)	5 (0)
13	20 (1)	0 (0)	100 (1)	na	20 (0)	10 (1)	50 (1)
14	100 (1)	100 (1)	10 (1)	na	100 (1)	20 (1)	10 (1)
17	100 (1)	60 (1)	30 (1)	na	0 (0)	5 (0)	0 (0)
18	0 (0)	0 (0)	40 (1)	na	20 (0)	40 (1)	0 (0)
19	100 (1)	70 (1)	0 (0)	na	100 (1)	10 (1)	0 (0)
20	100 (1)	90 (1)	20 (1)	na	100 (1)	20 (1)	2 (0)
21	90 (1)	90 (1)	10 (1)	na	100 (1)	3 (0)	0 (0)
22	90 (1)	20 (1)	30 (1)	na	100 (1)	15 (1)	1 (0)
23	0 (0)	0 (0)	80 (1)	na	0 (0)	30 (1)	0 (0)
26	5 (0)	3 (0)	2 (0)	na	10 (0)	15 (1)	80 (1)
27	80 (1)	0 (0)	0 (0)	na	90 (1)	20 (1)	2 (0)
31	90 (1)	20 (1)	0 (0)	na	90 (1)	2 (0)	0 (0)
32	3(0)	3 (0)	90 (1)	na	40 (1)	5 (0)	50 (1)
33	95 (1)	10 (1)	10 (1)	na	95 (1)	10 (1)	0 (0)
35	1 (0)	0 (0)	100 (1)	na	0 (0)	60 (1)	90 (1)
38	100 (1)	100 (1)	0 (0)	na	100 (1)	5 (0)	20 (1)

41	100 (1)	100 (1)	60 (1)	3+	100 (1)	10 (0)	0 (0)
44	53 (1)	0 (0)	na	3+	0 (0)	3 (0)	40 (1)
45	90 (1)	90 (1)	40 (1)	3+	100 (1)	10 (0)	20 (1)
46	70 (1)	90 (1)	5 (0)	1+	100 (1)	20 (1)	20 (1)
47	100 (1)	40 (1)	40 (1)	1+	100 (1)	20 (1)	60 (1)
49	95 (1)	90 (1)	40 (1)	na	80 (1)	30 (1)	5 (0)
51	0 (0)	0 (0)	80 (1)	na	0 (0)	10 (1)	20 (1)
52	5 (0)	0 (0)	0 (0)	na	0 (0)	2 (0)	0 (0)
53	100 (1)	1 (0)	0 (0)	na	90 (1)	1 (0)	0 (0)
55	90 (1)	40 (1)	9 (0)	1+	100 (1)	25 (1)	15 (1)

Table 4.2: For each patient (patient code) is reported the percentage of positive stained tumour cells for the 6 immunohistochemical markers: ER (Estrogen Receptor), PR (Progesteron Receptor), Erb-B2 (v-Erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog [*avian*]), Bcl-2 (B-cell CLL/lymphoma 2), Ki67 (antigen identified by monoclonal antibody Ki-67) and p53 (tumour protein 53). Samples were scored as ER, PR, c-erbB-2, Ki67, p53 positive with value ≥ 10%, as Bcl-2 positive with value ≥ 25%. For Erb-B2 marker is also reported the result of the Herceptest (if available): 1+, 2+, 3+ (see par. 3.2.3 for details) that measures the overexpression of the protein. Enclosed in parenthesis the transformed values are reported: 1 as a positive case, 0 as a negative case. na not available

The discordant result for Erb-B2, could be explained because the antibody against Erb-B2 (CB11 clone) is able to detect only the over-expression of Erb-B2 often due to a genic amplification, whereas it is not effective for detecting genic-overexpression without genic amplification.

In my dataset there are 11 patients Erb-B2 negative by IHC analysis and if we observe the ratio of the fluorescence intensity patient/control, we notice that in 5 cases the ratio is above 2. This value could indicate that there was a genic over-expression without genic amplification.

	ER	**PR**	**Erb-B2**	**Bcl-2**	**Ki67**	**p53**
Correlation coefficient (r_s)	**0.678**	**0.59**	0.019	**0.464**	0.125	0.078
valid cases	34	34	33	34	34	34
p-value	**5.169E-06**	**9.047E-05**	0.457	**0.002**	0.239	0.329

Table 4.3: For each of 6 markers the Spearman's correlation value and the p-value associated are reported. On the grey background are indicated the significant correlations (ER, PR, Bcl-2).

From the correlation analysis emerged that Bcl-2 showed a significant correlation IHC/microarray (rs=0.464, p-value=0.002). This result agrees with other studies, that reported a good correlation between cDNA array and IHC for Bcl-2 [158, 159].

The table 4.3 shows that there was not correlation for Ki-67 and p53, a not encouraging result at a first analysis.

The discrepancy for p53 could be expected because p53 protein detection is not dependent on mRNA overexpression, but on the increased half-life of a mutated protein. In normal cells, p53 protein half-life is short and expression levels are low and undetectable by IHC. In cancer cells, most p53 mutations lead to products that are not ubiquitinated and accumulate in the nuclei where they can be detected [160]. The technical problem in assessing p53 status for IHC is that this technique measures a stabilized p53 protein, due to point mutations which lead to a stabilization of the protein structure. Approximately 20% of the p53 mutations causes the truncation of the protein and these mutations are not picked up if IHC is used [87].

The nuclear antigen Ki-67 is a marker of cell proliferation and is expressed in S, G2 and M phase of the cell cycle but not in G0 phase (resting cells). A recent work of Urruticoechea and colleagues [161], which reviewed in detail all studies performed so far on Ki-67, reported that the correlation between Ki-67 mRNA levels and the presence of the protein identified by immunohistochemistry, has not been yet fully proved.

In light of what stated above, the informations obtained from IHC and microarray tecnologies could be both used in the analysis of prognostic/predictive markers of responsiveness to chemotherapy. However, since IHC and microarrays have a different sensitivity in terms of signal detection (protein level/mRNA level respectively), the informations provided from these techniques should be evaluated separately if in the presence of discordant results.

4.3 CORRELATION ANALYSIS BETWEEN IHC PROGNOSTIC MARKERS AND CLINICAL RESPONSE TO NEOADJUVANT CHEMOTHERAPY

After the correlation analysis IHC/array, I thought it was also interesting to see if there was an association between the positivity/negativity of the 6 prognostic markers by IHC (ER, PR, Erb-B2, Bcl-2, Ki-67, p53) and the clinical response to the treatment. I included all the patients with available IHC and clinical response data, in all 31 (30 for Erb-B2).

This analysis showed two main limitations: the small number of patients and the absence of two clearly distinct classes of clinical response, as would have been the case of complete response and progressive disease. In fact I considered the groups of Responders (cCR + PR) and Non Responders (PD + NC), because I could not consider only cCR and PD patients that were 5 cases in all (3 cCR and 2 PD). Moreover, when I did not include cCR and PD patients, the results did not change significantly from taking the whole dataset of patients.

To measure the association between the two variables, IHC staining and clinical response, I used the Fisher's exact test (see par. 3.15.2), that is suitable for a small number of samples.

The results are reported in the table 4.4 with the statistical significance values (one-tail p-value and two-tail p-value, left and right). In my analysis I referred to a two-tail p-value because I did not assume *a priori* the direction of the association (negative or positive). If a significant association is assessed with a two-tail test, then it is possible to perform a directional test (one-tail p-value) and establish the type of association.

MARKER	FISHER'S EXACT TEST				p-value
ER		R	NR	tot	one-tail left: 0.173 one-tail right: 0.974 two-tail 0.221
	ER+	**14**	**7**	21	
	ER-	**9**	**1**	9	
	tot	23	8	**31**	
PR		R	NR	tot	one-tail left: 0.280 one-tail right: 0.925 p two-tail 0.428
	PR+	**13**	**8**	21	
	PR-	**8**	**2**	10	
	tot	21	10	**31**	
Erb-B2		R	NR	tot	one-tail left: 0.938 one-tail right: 0.231 two-tail 0.418
	Erb-B2+	**14**	**7**	21	
	Erb-B2-	**4**	**5**	9	
	tot	18	12	**30**	
Bcl-2		R	NR	tot	one-tail left: 0.036 one-tail right: 0.997 two-tail 0.044
	Bcl-2+	**10**	**12**	22	
	Bcl-2-	**8**	**1**	9	
	tot	18	13	**31**	
Ki-67		R	NR	tot	one-tail left: 0.852 one-tail right: 0.404 two-tail 0.685
	Ki-67+	**15**	**6**	21	
	Ki-67-	**6**	**4**	10	
	tot	21	10	**31**	
p53		R	NR	tot	one-tail left: 0.546 one-tail right: 0.749 two-tail 1
	p53+	**11**	**11**	22	
	p53-	**5**	**4**	9	
	tot	16	15	**31**	

Table 4.4: The table shows the results of the Fisher's exact test for the prognostic markers: ER, PR, Erb-B2, Bcl-2, Ki-67 and p53. The response classes (Responders, R and Non Responders, NR) are reported in the columns and the positivity or negativity for the marker by IHC in the rows. The last column contains the p-values, one-tail left/right and two-tail. In this analysis we have considered the p value two-tail (in yellow). For Bcl-2 I considered also the one-tail left p-value (in light yellow) (see text for more details).

The results showed that there was not a preferential association between the positive/negative IHC staining of the markers and the clinical response to the treatment, except one case (Bcl-2).

From several studies it is known that ER expression and to a smaller extent PR expression, are established prognostic factors with a favourable clinical outcome for hormone receptor positive patients [87]. However, we did not consider the clinical outcome but the response to neoadjuvant chemotherapy. In this case, the literature showed conflicting results and not yet definitive. For example, some clinical series suggested that ER negative tumours are more sensitive to chemotherapy than receptor positive ones. It has been demonstrated that ER negative cancers have an increased proliferation rate compared to ER positive tumours [87]. High proliferation rate of tumour cells is a characteristic associated to positive chemotherapy response. Therefore, hormone receptor associated sensitivity to chemotherapy could be an effect of the proliferative activity of ER negative breast cancer and not dependent on the ER status *in se*.

Also the association between Erb-B2 status and the response to NACT is not yet fully investigated and the published data are conflicting. For p53 there is some evidence that it can act as a predictor of chemosensitivity (e.g.

tumours with p53 mutations seem to respond better to paclitaxel), but these results have still to be validated for a clinical use.

Breast carcinomas with a high Ki-67 positive count show improved response to chemotherapy in several studies [87] and the Ki-67 expression is found to be decreased after NACT. This could indicate that Ki-67 negative cells are not proliferating and thus not sensitive to the chemotherapic treatment. Different results are showed from an other group [162], which have reported that a reduction of the Ki-67 fraction is not a useful predictor for chemotherapy response.

The only prognostic marker that showed a significant association with the clinical response was Bcl-2; in particular I found a p-value one tail left significant. This result means that there was a negative association between the two variables: the negativity of Bcl-2 by IHC was correlated with a positive clinical response. Pusztai and colleagues found that tumours lacking Bcl-2 show more often a pathological complete response than tumours expressing Bcl-2 after doxorubicin-based chemotherapy [163]. Therefore the significant negative correlation that I found for Bcl-2, seems consistent with the observation of Pusztai. However, is also important to remark that Bcl-2 is part of the apoptotic network and is negatively regulated by p53, so Bcl-2 should be studied not as a single predictive factor but only in the context of its signalling pathway.

In summary this analysis showed that for 5 of 6 prognostic markers there was not significant association between the IHC data and clinical response to NACT. Although this result is probably affected from the small number of patients (especially the NR class), it is important to keep in mind that also other studies with larger cohorts of patients did not still clearly answer to the question if the prognostic markers have to be consider also as predictive markers.

4.4 ANALYSIS OF THE MOLECULAR BREAST TUMOURS SUBTYPES BASED ON THE INTRINSIC GENE SIGNATURE OF PEROU *ET AL*

4.4.1 IDENTIFICATION OF THE MOLECULAR SUBTYPES

In order to identify the molecular subtypes of the breast tumours included in this study, I used the intrinsic gene signature of Perou and colleagues [23, 30]. As I previously reported (par. 1.4.4), by clustering breast tumours using this "intrinsic gene list", were identified 4 subgroups of cancers with separate gene expression profiles: the luminal A, the luminal B, the basal type, the normal-like and the Erb-B2+ groups. It has been demonstrated that these breast cancer molecular subtypes have different prognoses and they also respond differently to preoperative chemotherapy [19].

Consequently, using this approach in my study could be useful for investigating the correlation between clinical response to NACT and gene expression profile.

I considered 37 patients analyzed with the Operon v2.0 platform, 30 with clinical response available and 7 without. Before starting with the analysis, I found how many genes of the intrinsic signature were present in our microarray platform, the Operon v2.0. The most recently updated version of the intrinsic signature [30] contains 306 genes that correspond to 431 probes of the Operon v3.0 platform[1] (Human genome oligo set version 3.0 arrays, see par. 4.5.1 for details). Out of 431 probes, 289 probes (67%) were present in the Operon v2.0 platform, 113 (26%) were multiple copies and 29 (7%) were not present. I recovered the log2ratio expression values of these 289 probes in 37 patients included in the analysis. Out of 289 probes, only those probes that had a value at least in 23/37 patients were taken, in all 236 probes (fig. 4.3).

To identify the molecular subtypes (Luminal A; Luminal B, Erb-B2, Normal-like and Basal-like), I considered five subtype mean expression profiles (centroids) based upon the expression of the 236 "intrinsic genes", similarly to the procedure of Hu and colleagues [30]. The sample was then assigned to the nearest subtype/centroid as determined by Pearson correlation.

The table 4.5 shows the distribution of the 37 patients in respect with the molecular subtype. No patients with the normal-like molecular subtype were found, so I reported only the luminal A, luminal B, Erb-B2+ and basal-like groups.

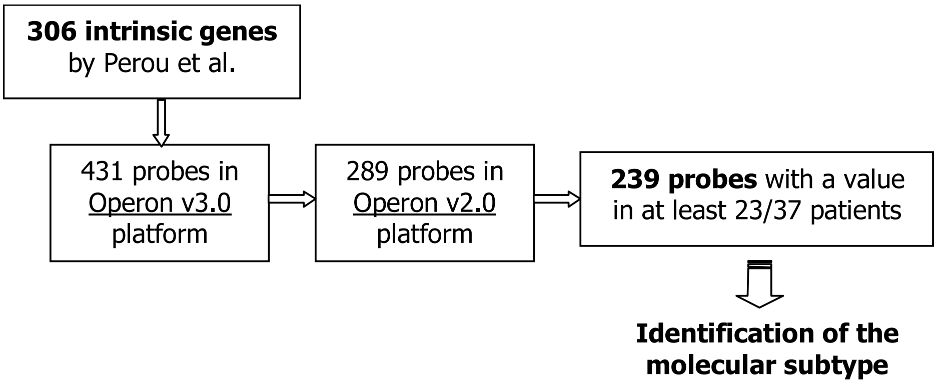


Figure 4.3: Summary of the procedure to find the "intrinsic gene list" in the microarray platform (Operon v2.0) used in this study.

As can be seen, 28 of 37 (76%) patients belonged to the luminal-type molecular subtype, and almost all of them (26/28) were luminal B type. This result is in agreement with the data reported in literature that the luminal-type breast tumours are the most common tumours type, ~60-70% in the white woman population, as the Caroline Breast Cancer Study has reported [27]. At this point I used the 239 probes of the intrinsic genes to perform an unsupervised hierarchical cluster analysis to evaluate if the patients were grouped on the basis of these 239 probes. The results are shown in the figure 4.4. As can be seen, the samples did not cluster on the basis of the molecular subtype, except the two basal-like tumours, that grouped together. The samples could not be clustered on the basis of their molecular subtypes because I did not use all intrinsic gene list of the "original" signature of Perou *et al.*, but only those genes contained in the Operon v2.0

[1] The list of the 431 probes Operon v3.0, corresponding to the 306 intrinsic genes, was supplied from Juliane Hannemann at The Netherlands Cancer Institute.

platform. In this "reduced signature" some informative genes could lack and these would be useful to separate correctly the patient. Also the small sample size might influence the result.

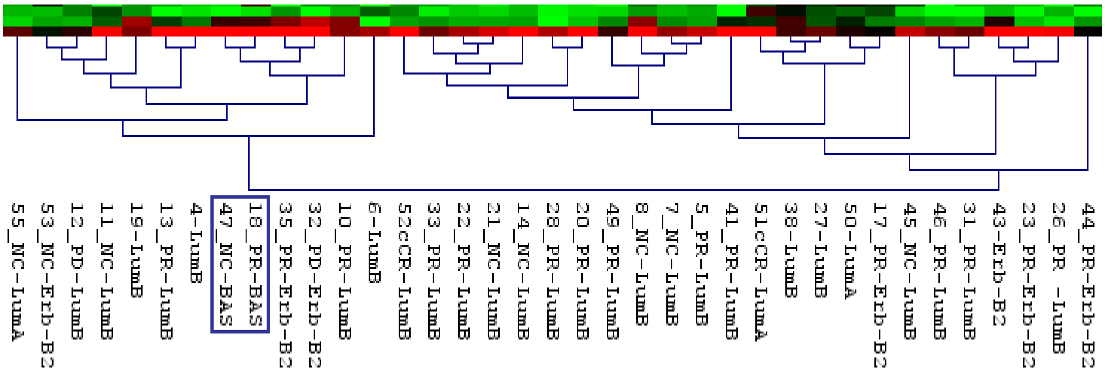


Figure 4.4: Hierarchical clustering of 37 patients (with and without clinical response) based on the 236 probes of the intrinsic gene set [30]. Where available, the clinical response of the patient has been reported (cCR complete Clinical Response, PR Partial Response, NC No Change, PD Progressive disease). LumB: Luminal B, LumA: Luminal A, BAS: BASAL-like. The basal-like patients are highlighted in the graph. TMEV (see par. 3.13) was used to perform the hierarchical clustering with the average linkage method.

Since Rouzier and colleagues showed that breast cancer molecular subtypes respond differently to preoperative chemotherapy [19], it was interesting to see the molecular subtypes of the patients with respect to the clinical response (tab. 4.5). Rouzier reported in his study that the luminal tumours tend to show lower pathologic Complete Response (pCR) rates to paclitaxel- and doxorubicin-containing preoperative chemotherapy than the basal-like and erb-B2 tumours, which have on the contrary, a higher likelihood of pCR. In my study I did not consider as response a pathologic response but a clinical response (see 4.1), so the comparison with the results obtained from Rouzier *et al.* could present some limitations. However, I considered clinical Complete Response (cCR) the closest, in term of success of the NACT treatment, to the pCR response. Two patients, out of 30, were cCR, one belonged to erb-B2+ subtype and the other one to lumB group. I would have expected that both were erb-B2+ or basal-like, but it is also important to remark that I was considering clinical responses and, most importantly, I had only two cases of cCR. The Partial Response (PR) patients belonged mainly (70%) to the luminal B type. If we consider a clinical partial response closer to the pCR, this result would disagree with what reported from Rouzier that the luminal subtype shows lower pCR rates to NACT.

MOLECULAR SUBTYPE	N° OF PATIENTS
luminal A	2
luminal B	26
erb-B2+	7
basal-like	2

	lum A	**lum B**	**erb-B2+**	**basal**
cCR	0	1	1	0
PR	0	12	4	1
NC	1	7	0	1
PD	0	1	1	0

Table 4.5: On the left: distribution of the 37 patients included in the analysis with respect to the molecular subtypes: luminal A (lum A), luminal B (lum B), erb-B2+ and basal-like

(basal). On the right: distribution of the 4 molecular subtypes with respect to the clinical response; only the patients with available clinical response are included. cCR clinical Complete Response, PR Partial Response, NC No Change, PD Progressive Disease.

To explain this result, it is important to underline that Rouzier and colleagues have considered in their study patients that showed a pCR compared to those with residual disease. Patients with residual disease included, in fact, both Partial Response and No Change/Progressive disease response. It means that I should consider, in this case, a PR patient not in the same group of cCR patients, as Responders, but a distinct group with NC and PD patients. In light of this observation, is relevant the high percentage of PR patients with a luminal B subtype.

Almost all the not responder patients (PD and NC patients) belonged to the luminal subtype (9/11, 82%). The result has to be corrected in light of the low numerosity of the not responder group and the observation that the luminal-type tumours are the most common in the population; nevertheless the result is in agreement with that reported from Rouzier and coworkers.

4.4.2 COMPARISON BETWEEN THE MOLECULAR SUBTYPES AND THE ER, ERB-B2 STATUS BASED ON IHC AND MICROARRAY

The molecular subtypes are characterized by the expression of a specific cluster of genes (see par. 1.4.4) belonging to the intrinsic gene list of Perou and colleagues [23]. In particular the luminal-type, the basal-like and the erb-B2+ breast tumours differ from each other on the basis of the expression of Estrogen Receptor 1 (ESR1), and Erb-B2. In fact, the luminal-type tumours are also called ER+ tumours, because generally show a high expression of ESR1 and other genes involved in the ESR1 activation (see par. 1.4.4.1 for details); instead the erb-B2+ subtype show low levels of expression of ESR1, as the basal-like tumours. The characteristic trait of erb-B2+ tumours is the high expression of several genes in the Erb-B2 amplicon including Erb-B2 itself. The basal-like tumours present low expression of Erb-B2 and ESR1.

In light of what stated above, I checked the microarray expression values and the positivity/negativity in ImmunoHistoChemistry (IHC) of ER and Erb-B2 in the 37 patients included in the analysis. The results are reported in the figure 4.5.

As can be seen there was a fairly good correspondence between erb-B2+ molecular subtype, the Erb-B2 gene expression values measured by microarray and the Erb-B2 IHC status. Out of 7 erb-B2+ tumours, 5 showed an Erb-B2 overexpression by microarray and a positivity staining by IHC for the protein. In the basal-like subtype we observe that the microarray results are in agreement with the characteristic of this group, that is an underexpression of both ESR1 and Erb-B2; instead the IHC data do not show a good correspondence. In particular, in the patient 47 both markers (ER and Erb-B2) were identified as positive using the IHC assessment. This conflicting result could be due to the differences in the detection of the protein level in these two techniques (IHC and microarrays), whereas the molecular subtypes were identified based on the gene expression level. It is also important to remark that for the patient 47 we found a negative score in the Herceptest (1+, see the table 4.2) for Erb-B2 although the IHC immunostaining gave a high percentage of positivity. Therefore, if we consider the Herceptest score, the result agrees with the microarray data. The luminal-type tumours in 22 of 28 cases showed an over-expression of ESR1, in agreement with the result reported by Perou and colleagues. A similar result was obtained with IHC assessment, where 24 of 26 luminal-type tumours with IHC data available, were positive for ER marker.

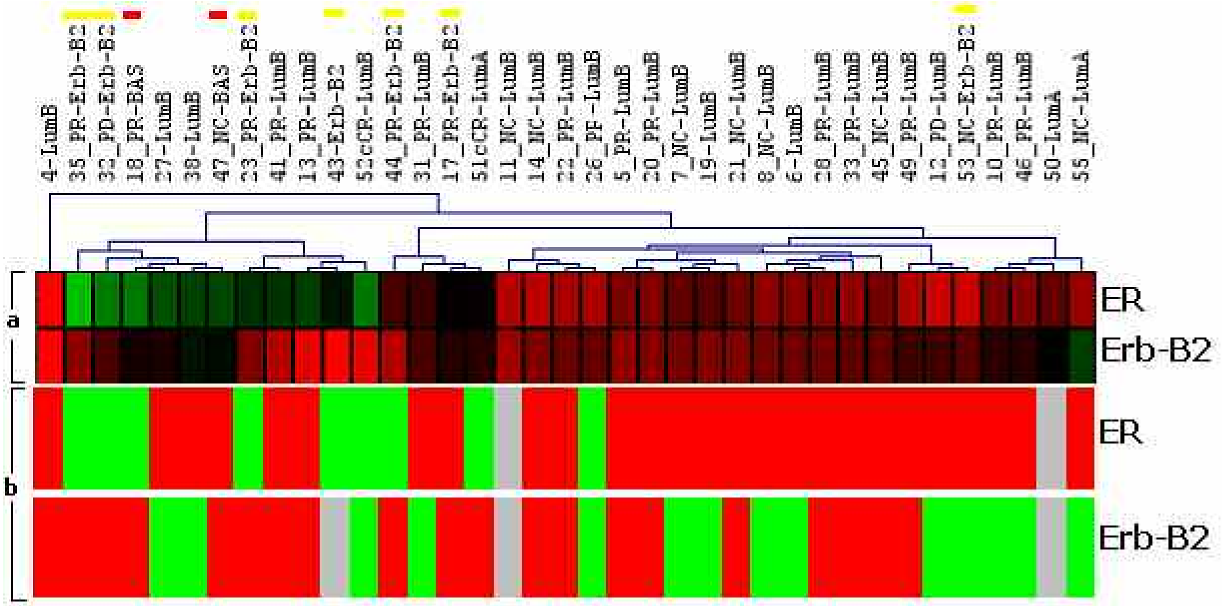


Figure 4.5: The log2ratio gene expression values (**a**) and the IHC status (**b**) of ER (Estrogen Receptor) and Erb-B2 are shown. The grey cells represent missing data in the IHC assessment. The samples were grouped on the basis of a hierarchical clustering using ER and Erb-B2. The yellow horizontal bars indicate the Erb-B2 tumours, the red horizontal bars the basal-like tumours. The labels of the patients report the molecular subtype and the clinical response, if available. Color coding: (**a**) red scale (over-expression): 0<log2ratio<+3, green scale (under-expression) -3<log2ratio<0 (**b**) red: IHC status positive (% stained cells ≥ 10), green: IHC status positive (% stained cells < 10). TMEV (see par. 3.13) was used to perform the hierarchical clustering (average linkage method). Note: for the patient 44 I reported the Erb-B2 status (3+) measured by the Herceptest (DAKO, see par. 3.2.3) because the result of immunohistochemical assessment was not available.

If we look at the clustering of the patients, it can be noticed that all patients with ESR1 gene downregulated are grouped in a cluster enriched of erb-B2+ tumours (4 out of 7 erb-B2+ tumours) and also containing the two basal-like tumours.

Taken together, the results show that the breast tumours had, in most of the cases, a gene expression of the two genes under study (ER and Erb-B2) in agreement with the key characteristics of the luminal and erb-B2+ molecular subtypes: the former overexpression of ESR1, the latter overexpression of Erb-B2. A limitation of this type of analysis is the small size of the dataset under exam and especially the limited number of tumours belonging to the basal-like (2) and erb-B2+ subtypes (7). Since they were only 9 cases in all (out of 37 tumours *in toto*), these groups were too small to draw any significant conclusion from these results.

4.5 OPERON v3.0 PLATFORM *VS* OPERON v2.0 PLATFORM

As I reported in the previous paragraphs, a limitation of my study was the small size of the dataset of patients with clinical response available (see fig. 4.1), therefore increasing the number of patients would have been useful from a statistical point of view. The figure 4.1 gives an overview of the patients collected, and, as can be seen, there were 4 patients that were analyzed with the platform Operon v3.0 instead of with the Operon v2.0. Operon v3.0 platform represents the updated version of Operon v2.0.
My goal was to evaluate if I could introduce these 4 patients in the dataset of patients analyzed with the Operon v 2.0, without introducing too much variability due to the different microarray platform. In order to do this, I chose to re-analyze four patients, already characterized with Operon v2.0 microarrays, with the Operon v3.0 platform and to check the level of correlation between the gene expression profiles obtained with the two systems. The correlation analysis was difficult for two reasons:

- Operon v3.0 platform differed from the Operon v2.0 in term of number of probes, so I should firstly find an overlap between the two platforms and then perform the comparison (see par. 4.5.1 for details);
- the microarray system implemented in the hybridization step and fluorescence signal detection is different from that one used to analyze the 30 patients (see par. 4.5.1 for details).

4.5.1 CHARACTERISTICS OF THE OPERON v3.0 PLATFORM

The Human Genome Oligo Set Version 3.0 array contains 34.580 probes representing 24.650 genes and 37.123 gene transcripts. All oligos are 70mers. The arrays consist in all of 37.632 features (oligos and controls) and were printed in a 28x28 subarray layout using 48 Biorobotic 10K-micro spot pins. These arrays were obtained from the Central Microarray Facility (CMF) at the Netherlands Cancer Institute (NKI). CMF performed the microarray hybridization and the scanning; detailed information about the protocol can be found at http://microarrays.nki.nl/research/methods.html. All hybridizations were performed in the hybridization station Tecan HS4800 (http://www.tecan.com/). The sample solution was mixed during the incubation. The Agilent DNA Microarray scanner was used to scan the slides. RNA isolation, amplification and labeling were carried out following the same procedure used for the Operon v2.0 platform (see Methods). Each experiment was replicated using a dye swap procedure. The control RNA was the same used for the Operon v2.0 platform.

4.5.2 CORRELATION ANALYSIS BETWEEN PATIENTS ANALYZED WITH OPERON v2.0 AND OPERON v3.0 PLATFORMS

A common dataset of probes between the two platforms was identified to be used to perform the correlation analysis. Based on EntrezGene ID and EnsemblGene ID of the oligonucleotides I found that 15.554 (~72%) Operon v2.0 probes represented the same genes of the Operon v3.0 probes. Although this fairly good overlap, I chose to keep only those overlapped probes with the same oligonucleotide sequence in both platforms. Thus the number of overlapping probes was reduced to 12083. The choice was done to be sure to evaluate expression values for the genes recognized by the same oligonucleotide, avoiding the risk to consider oligos that mapped on the same gene but specific for different transcripts (isoforms of the gene with different level of expression). This procedure had the drawback that a specif isoform of a gene recognized from different oligos would have been lost. However my goal was to evaluate the level of correlation between two gene expression profiles of the same patient obtained with two different platforms, thus a dataset of 12083 common probes would have been large enough to perform a statistically reliable correlation analysis.

The further step was to compare, using the Pearson correlation, the normalized gene expression value of the single channels data for the same patient analyzed with Operon v2.0 and Operon v3.0 platforms. Since were performed three replicates for each experiment with the Operon v2.0 platform and two replicates with the Operon v3.0, I considered the replicates separately in the calculation of the Pearson correlation; finally I averaged the correlation coefficient (r) for each channel (channel 1 = patient, channel 2 = control). The patients re-analyzed with Operon 3.0 platform were 7, 13, 53 and 55. The results are shown in the table 4.6.

patient code	**r (CH1_Op2.0 *vs* CH1_Op3.0) ± SD**	**r (CH2_Op2.0 *vs* CH2_Op3.0) ± SD**
7	0.79 ± 0.02	0.78 ± 0.03
13	0.80 ± 0.09	0.77 ± 0.04
53	0.34 ± 0.02	0.48 ± 0.02
55	0.21 ± 0.02	0.29 ± 0.06

Table 4.6: The Pearson correlation coefficients (r) calculated for each comparison is reported. CH1_Op2.0 (CH2_Op2.0): channel 1(2) of the Operon v2.0 platform; CH1_Op3.0 (CH2_Op3.0): channel 1(2) of the Operon v3.0 platform. SD Standard Deviation

The correlation could be considered good for the patients 7 and 13, medium for the patient 53 and low for the patient 55. The low correlation could be due to:

- different platform;
- different range of sensitivity of the detection of the fluorescence signal from the two scanners, Agilent and Packard (see par. 3.9).

In light of these results I did not consider the four new patients in the dataset of patients analyzed with Operon v2.0 platform ("old dataset"), because I did not find a good correlation for all the four patients analyzed with both platforms (Operon v2.0 and v3.0). The risk would have been to introduce some variability in the old dataset, due to a different system of microarray analysis, that could influence the results.

4.6 IDENTIFICATION OF PREDICTIVE GENES OF RESPONSIVENESS TO ANTHRACYCLINE-TAXANE BASED NEOADJUVANT CHEMOTHERAPY

As mentioned before (see chap. 2), the primary goal of this study was to identify a gene expression signature predicting the response to a NeoAdjuvant ChemoTherapy (NACT) based on Paclitaxel/Doxorubicin or Paclitaxel/Epirubicin combination of drugs. Next steps of my study were focused on the identification of this predictive set of genes for the pre-treatment tumours (patients) with available clinical responses, in all 30 patients as reported above (par. 4.1). The gene expression values were filtered on the basis of the statistical procedure previously described (see Methods, par. 3.11) to create the dataset for the subsequent analysis.

4.6.1 UNSUPERVISED HIERARCHICAL CLUSTERING ANALYSIS

First of all, I performed a hierarchical clustering on the pre-treatment tumours with clinical responses available, in order to evaluate how the patients would have been separated on the basis of their gene expression profile without giving any information about the class of response to the treatment (unsupervised approach).
I considered two datasets with a different number of patients, the first composed by the Partial Response (PR) patients in the Responder group, the second including also the clinical Complete Response (cCR) patients:

- **dataset I**: 17 PR + 11 NR (9 NC + 2 PD); it contains 13973 gene expression values (fig. 4.6A);
- **dataset II**: 19 R (17 PR + 2 cCR) + 11 NR (9 NC + 2 PD); it contains 13870 gene expression values (fig. 4.6B).

I took into account two types of dataset to evaluate if the results would have been different considering only PR patients against non responder patients or responders patients (PR and cCR). In the first case the class of response was more homogeneous (only partial response patients) in terms of response to the treatment.
The figure 4.6 displays the dendograms of the 28 samples (dataset I) and 30 samples (dataset II) based on 13973 genes and 13870 genes, respectively.
As can be seen from both dendograms (fig. 4.6A and fig. 4.6B) there was not a clear separation between the Responders (PR and cCR) and Non Responders (NC and PD). This result suggest that other set of genes separate the samples on the basis of other biological parameters not dependent on the response to the drugs. The high number of differentially expressed genes could mask the "real" predictive genes and bring to a clustering of samples independent from the class of response. Probably the patients could be better clustered in responders (sensitive) and non responders (resistant) if their number was higher: the biological differences would have been more randomly distributed between the classes of response, and the predictive genes could be the principal discriminant factor. However, the result obtained agrees with the hypothesis, already proposed from van 't Veer *et al* [44], that the predictive genes of resistance/sensitivity

to the drugs are a subtle set and it could not cause great changes in gene expression, hardly detectable with an unsupervised approach.

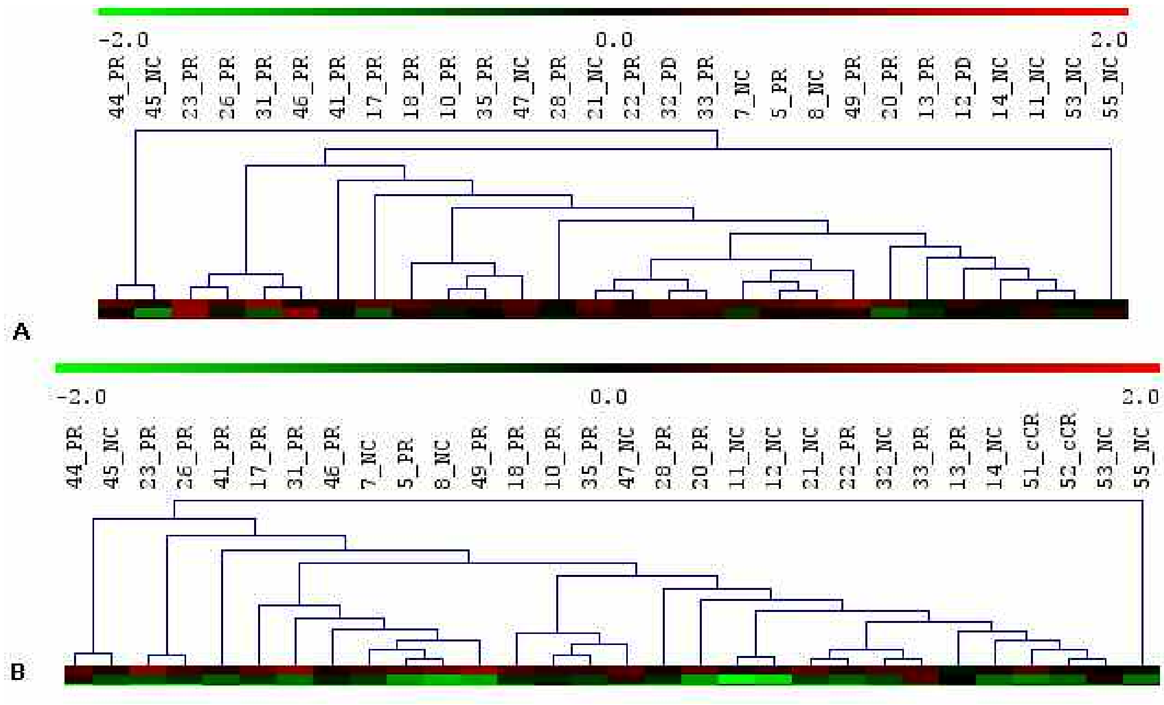


Figure 4.6: Unsupervised hierarchical clustering (average linkage) performed on 28 patients and using 13973 genes (**A**) and on 30 patients using 13870 genes (**B**). cCR clinical Complete Response, PR Partial Response, NC No Change, PD Progressive Disease. The clustering analysis was performed using TMev software (par. 3.12).

In light of what stated above, I preferred a supervised approach, which consists of dividing the tumours into groups that have different clinical responses and searching for the genes that can correctly identify the distinct groups of response, in other words the predictive genes of resistance/sensitivity to the neoadjuvant chemotherapy. This approach has been previously used in various studies similar to my study (see par. 1.5); although unsupervised approaches seems to be less biased, it emerged that the possibility to identify informative genes of clinical subgroups is enhanced when additional available clinical information (e.g. clinical response) were included in the analysis.

4.6.2 IDENTIFICATION OF DRUG-RESISTANCE PREDICTIVE GENES USING PAM (PREDICTIVE ANALYSIS OF MICROARRAY)

The first supervised approach that I used to find the predictive genes was based on the algorithm implemented in PAM (Prediction Analysis of Microarray), that I described in detail in Methods, par. 3.13. Also for PAM analysis I considered the dataset I (28 patients, 13973 genes) and the dataset II (30 patients, 13870 genes).

The results are reported in the table 4.7. As can be seen from the table 4.7, in both cases the misclassification error is high and it increased if we included in the Responder class also the cCR patients.

	nr predictive genes	true/predicted	1	2	Class error rate
dataset I	110	**1**	11	6	0.353
		2	4	8	0.363
dataset II	86	**1**	12	7	0.368
		2	4	7	0.363

Table 4.7: For the dataset I (28 patients) and the dataset II (30 patients) are reported the number of predictive genes identified from PAM, the true number of patients belonging to the class 1/class 2 and the predicted number of patients belonging to the class 1 and 2 (based on the predictive genes identified), the class error rate associated to each class of response. For the dataset I the class 1 corresponds to the PR patients and the class 2 to the NR patients (NC + PD); for the dataset II the class 1 corresponds to the R patients (PR + cCR) and the class 2 to the NR patients (NC + PD). The number of patients correctly assigned to the class of response is on grey background.

This result could be due to the fact that the cCR patients are too different respect to the PR patients in terms of biological mechanisms of response to the treatment. The tumours with a partial regression in the mass size are partially sensitive to the drug and could show a certain grade of resistance. It would make not advisable consider cCR patients in the same group of PR patients for searching of predictive genes of response to chemotherapy.

The general low accuracy in the prediction performance obtained with PAM did not exclude the existence of a better predictive profile in terms of prediction accuracy, rather than PAM could not be the proper method to find it. Therefore a new type of method was required to:

- find a predictive gene set more powerful in distinguishing the class of response to the neoadjuvant chemotherapy;
- thoroughly evaluate the performance of the gene predictive signature.

4.6.3 IDENTIFICATION OF DRUG-RESISTANCE PREDICTIVE GENES USING FEATURE SELECTION BASED ON SUPPORT VECTOR MACHINES

A key difficulty in microarray studies arises from the high dimensionality of the data: typically a microarray analysis involves tens of samples while measured genes (features) are thousands. Since the data dimensionality is much larger than the sample size, this makes many standard pattern classification algorithms fail and also increase the risk of overfitting (see par. 3.14.2). Machine learning methods such as Support Vector Machines (SVMs), can work at high-dimensionality, as observed in other studies [148], thus I chose this approach to identify the gene predictive signature. I used R-SVM, a recursively method of genes (called features) selection based on SVMs (see par. 3.14.2.2 for more details). At each iteration of feature selection process (in all 12 iterations) was selected a subset of features that had the higher contribution in the classification of the patients, ranking all the genes according to their SVM score and were eliminated the 50% of the low ranked features. I chose a Leave-One-Out Cross Validation (LOO-CV) procedure to assess the performance of the feature selection process, because of the small size of the dataset of patients (28 or 30 patients). The LOO-CV is the statistically suggested cross validation method if the number of the samples is low, as in my study. A key point is that the feature selection steps were included in the cross validation procedure in order to validate both the classification algorithm and the feature selection process. Others feature selection procedures did not use a correct validation scheme, as the case of RFE-SVM (see par. 3.14.2.2 for more details), and for this reason I used the R-SVM method.

At each iteration of the feature selection process, the subset of genes with the higher contribution in the classification was associated to an error of classification, or, in other words, to a number of patients that, based on those genes, were not classified in the true class of response. Therefore I chose the subset of genes with the lowest error of misclassification. The genes selected represented a gene-expression signature able to distinguish the responder and the non responder patients.

For the feature-selection analysis with R-SVM I considered the dataset I and the dataset II, similarly to the analysis with PAM. In the table 4.8 I reported for each dataset the number of selected genes with the highest accuracy in the classification (respect to the patients considered) determined with the LOO-CV procedure and the number of misclassified patients.

	nr of selected genes	**nr of misclassified patients**	**misclassified patients**	
dataset I **28 patients** (17 PR, 11 NR) **13973 genes**	54	4	**PR**	17
			NR	14, 21, 32
dataset II **30 patients** (19 R, 11 NR) **13870 genes**	14	7	**R**	13, 17
			NR	14, 21, 32, 47, 55

Table 4.8: For the dataset I and the dataset II the number of selected genes from R-SVM method (nr of selected genes), the number of misclassified patients using the selected genes by LOO-CV procedure (nr of misclassified patients) and the misclassified patients are reported. PR Partial Response, NR Non Responders, R Responders

The table 4.9 shows the performance of the 54-genes signature (dataset I) and the 14-genes signature (dataset II) in terms of sensitivity, specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV) and accuracy.

As emerged from the tables 4.8 and 4.9, the best performance in the classification considering the statistical parameters evaluated, was obtained using the dataset I using the 54-genes signature. The accuracy of the 54 genes set is equal to 85%, instead with the 14 genes set of the dataset II the accuracy decreased to 76%. If we look at the specificity, that is the proportion of NR patient (also called negative examples) which was correctly identified, we obtained lower values than for the sensitivity, the proportion of R patient (also called positive examples) which was correctly identified, in both datasets. This result could be expected because the NR patients were in lower number respect to the R patients, 11 Non Responders *vs* 17 (or 19 if included the cCR patients) Responders.

	54 genes (dataset I)		**14 genes** (dataset II)	
	cases	**percentage**	**cases**	**percentage**
sensitivity	16/17	**94%**	17/19	**89%**
specificity	8/11	**72%**	6/11	**52%**
PPV	16/19	84%	17/22	77%
NPV	8/9	88%	6/8	75%
accuracy	24/28	**85%**	23/30	**76%**

Table 4.9: For the dataset I and the dataset II are shown sensitivity, specificity, PPV (Positive Predictive Value), NPV (Negative Predictive Value) and Accuracy of the set of 54 genes and 14 genes. **54 genes**: TP (True Positive)=16, TN (True Negative)=8, FP (False Positive)=3, FN (False Negative)=1; **14 genes**: TP=17, TN=6, FP=5, FN=2. Sensitivity=TP/P, specificity=TN/(FP+TN), PPV=TP/(TP+FP), NPV=TN/(TN+FN), accuracy=(TP+TN)/P+N

It can be observed that both signatures, 54-genes and 14-genes, misclassified the same patients (17, 14, 21, 32) and the 14-genes signature misclassified also the patients 13, 47, 55. This result could indicate that the "system" is unstable and every time that new patients are added the error changes. However, each new patient can introduce some variability in terms of biological heterogeneity and, because of the limited number of patients, this factor plays an important effect. The difference in the performance of the classification between the 54-genes signature and the 14-genes signature is consistent with what already observed (par. 4.6.2), that the cCR patients would represent a too different class respect to the PR patients in terms of biological mechanisms of response to the treatment. The cCR patients should be treated as a different class of response, in order to have groups of response as homogeneous as possible. Therefore I chose to focus the subsequent analysis on the dataset I that included only the patients PR (partially responsive to the treatment) and the patients NR (not responsive to the treatment).

Although the number of patients *in toto* (and especially the NR class) is too small for general conclusions, the result of this exploratory supervised classification for the dataset I seems encouraging. It is also important to remark that another limitation of this study is the absence of two clearly distinct classes of clinical response, as would have been the case of only NC and cCR patients. The PR class, as reported in literature [97], may be very heterogeneous group, making it difficult to predict the treatment response for these patients correctly.

From a biological point of view, it is known that there are several mechanisms that lead to the drug-resistance phenotype (see 1.7) of a tumour cell, but they are not all active at the same time. Therefore the drug-resistance markers could be present in some patients and absent in others. As a consequence, the 54-genes signature identified could be effectively predictive of response to the neoadjuvant chemotherapy (anthracyclines plus paclitaxel) only for those patients with common markers of drug-resistance.

4.6.4 ANALYSIS OF THE 54-GENES PREDICTIVE SIGNATURE

At this point of the study I decided to analyze more in detail the 54 predictive genes listed in the table 4.10.

OligoID	Entrez Gene ID	GeneBank ID	GeneSymbol
H200002171	84159	NM_020403	**ARID5B**
H200002810	51232	NM_016441	**CRIM1**
H200002898	3434	NM_001548	**IFIT1**
H200003547	1365	NM_001306	**CLDN3**
H200003548	2353	NM_005252	**FOS**
H200004431	79682	NM_024629	**MLF1IP**
H200004583	57678	AB046780	**GPAM**
H200005130	84627	AB058761	**ZNF469**
H200005447	26872	NM_012449	**STEAP1**
H200006203	427	NM_004315	**ASAH1**
H200006389	4283	NM_002416	**CXCL9**
H200006397	5577	NM_002736	**PRKAR2B**
H200006318	1052	NM_005195	**CEBPD**
H200006446	9088	NM_004203	**PKMYT1**
H200006618	4609	NM_002467	**MYC**
H200007119	688	NM_001730	**KLF5**
H200008477	4023	NM_000237	**LPL**
H200009886	5166	NM_002612	**PDK4**
H200010330	185	NM_031850	**AGTR1**
H200011610	10580	NM_015385	**SORBS1**
H200013414	8660	AF073310	**IRS2**
H200013620	1654	NM_001356	**DDX3X**
H200013568	1289	NM_000093	**COL5A1**
H200013741	57496	AB033069	**MKL2**
H200014307	9415	NM_004265	**FADS2**
H200015305	10850	NM_006664	**CCL27**
H200015445	6935	NM_030751	**TCF8**
H200015384	10930	NM_006789	**APOBEC2**
H200017585	114783	AB067470	**LMTK3**
H200017794	57092	NM_020357	**PCNP**
H200020738	26074	AK056971	**C20orf26**
H200007115	9499	NM_006790	**MYOT**
H200015637	1027	BC001971	**CDKN1B**
H200015854	171024	AF177291	**SYNPO2**
H200018157	5507	BC012625	**PPP1R3C**
H200000922	9985	NM_005132	**REC8L1**
H200005314	11096	NM_007038	**ADAMTS5**
H200010022	57223	AB037808	**SMEK2**
H200019109	6711	AK023762	**SPTBN1**
H200006219	4089	NM_005359	**SMAD4**
H200007199	23676	NM_014332	**SMPX**
H200015396	3655	NM_000210	**ITGA6**
H200002977	57405	NM_020675	**SPC25**
H200011805	84419	NM_032413	**C15orf48**

H200000153	667	NM_001723	**DST**
H200014817	10891	NM_013261	**PPARGC1A**
H200008461	#	AK027252	**ATP8A1**
H200006689	4753	NM_006159	**NELL2**
H200013115	643008	AK055768	**LOC643008**
H200001902	7364	NM_001074	**UGT2B7**
H200015506	83540	NM_031423	**CDCA1**
H200008367	55008	NM_017912	**HERC6**
H200014220	4477	NM_002443	**MSMB**
H200007560	10631	AK023481	**POSTN**

Table 4.10: 54-genes signature identified with R-SVM for the dataset I (17 PR + 11 NR). For each gene I reported: Oligo ID of the Operon v2.0 platform, Entrez Gene ID, GenBank ID and Gene Symbol.

When we looked at the microarray expression value of each gene singularly, it emerged that they did not show a gene expression value markedly different between the two class of response (e.g. always underexpressed in the PR patients and overexpressed in the NR patients), so one could ask why they should be predictive, that is able to separate the two class of response. It should be pointed out that a single gene is not discriminating *per se*, but the genes of the predictor are optimal for the classification only if taken together. In fact, the feature-selection based on SVMs is a wrapped method because the feature selection process scored the importance of the genes in the classification considering the correlation between them.

The neoadjuvant chemotherapy treatment of this study is based on a combination of paclitaxel and anthracyclines (doxorubicin and epirubicin), two compounds whose mechanism of action involve the microtubule cellular dynamics and the DNA binding and replication, respectively (see 1.8 for more details).Therefore it could be speculated that the 54 gene classifier predicting response to paclitaxel/anthracyclines regimen, would contain a number of genes involved in these process.

In order to have a general overview about the functional categories more represented from the 54-genes signature I used GoMiner software (see par. 3.16.2). This program uses the Gene Ontology (GO) annotation to identify enriched GO categories in the subset of selected genes with respect to the whole dataset of genes. I considered the whole dataset of 13973 genes (dataset I) and, as a subset of genes, the 54 genes identified with R-SVM. With respect to the total number of genes (13973), 7579 had a GO Biological Process (GO BP) annotation, 8057 out of 13973 a GO Cellular Component (GO CC) annotation and 8056 a GO Molecular Function (GO MF) annotation. Out of the 54 predictive genes, 47 had a GO BP, 46 a GO CC and 48 a GO MF annotation.

The table 4.11 shows the selected GO categories, with a p-value < 0.05, considering the three ontologies, Biological Process, Cellular Component and Molecular Function, separately. Only the categories comprising at least 2 genes are reported.

	nr total genes	nr predictive genes	p-value
GO Biological Process			
transcription_from_RNA_polymerase_II_promoter (KLF5, PPARGC1, CEBPD, TCF8, FOS, SMAD4, MKL2, MYC)	428	8	0.002
fatty_acid_metabolic_process (PPARGC1, FADS2, LPL, ASAH1)	113	4	0.004
cell_cycle_arrest (CDKN1B, MYC, DST)	58	3	0.004
cell_adhesion (COL5A1, CLDN3, NELL2, SORBS1, ITGA6, DST, POSTN)	419	7	0.008
insulin_receptor_signaling_pathway (IRS2, SORBS1)	25	2	0.009
cell_cycle (CDCA1, CDKN1B, PCNP, PKMYT1, REC8L1, MYC, DST)	470	7	0.015
response_to_hypoxia (CLDN3, SMAD4)	35	2	0.016
cellular_carbohydrate_metabolic_process (PPARGC1, PDK4, IRS2, PPP1RC3)	180	4	0.018
regulation_of_cell_proliferation (KLF5, CDKN1B, IRS2, SMAD4, MYC)	295	5	0.024
cell_growth (CRIM1, CDKN1B, SMAD4)	121	3	0.030
actin_cytoskeleton_organization_and_biogenesis (SPTBN1, SORBS1, DST)	127	3	0.034
cell-matrix_adhesion (SORBS1, ITGA6)	56	2	0.039
lipid_metabolic_process (PPARGC1, FADS2, UGT2B7, LPL, GPAM, ASAH1)	456	6	0.041
cell_proliferation (KLF5, CDKN1B, IRS2, TCF8, SMAD4, MYC)	480	6	0.051
GO Cellular Component			
extracellular_region (COL5A1, CRIM1, NELL2, CXCL9, MSMB, LPL, CCL27, ADAMTS5, POSTN, DST)	861	10	0.019
extracellular_matrix (COL5A1, ADAMTS5, POSTN, DST)	187	4	0.020
apical_junction_complex (CLDN3, SORBS1)	49	2	0.031
apicolateral_plasma_membrane (CLDN3, SORBS1)	51	2	0.033
cell_junction (CLDN3, SORBS1, STEAP1, DST)	219	4	0.034
nucleus (CDKN1B, MLF1IP, SYNPO2, ARID5B, PCNP, ZNF469, TCF8, SORBS1, MSMB, SMAD4, REC8L1, CDCA1, PPARGC1, KLF5, DDX3X, CEBPD, FOS, SPTBN1, MKL2, SMPX, MYC)	2646	21	0.043
GO Molecular Function			
insulin_receptor_binding (IRS2, SORBS19)	11	2	0.002
heparin_binding (COL5A1, LPL, POSTN)	47	3	0.002
receptor_binding (PPARGC1, IRS2, CXCL9, SORBS1, CCL27, ADAMTS5, DST)	407	7	0.007
integrin_binding (ADAMTS5, DST)	27	2	0.010
cytoskeletal_protein_binding (SYNPO2, SORBS1, SPTBN1, MYOT, DST)	286	5	0.021
transcription_regulator_activity (PPARGC1, KFL5, CEBPD, ARID5B, TCF8, FOS, SMAD4, MKL2, MYC)	816	9	0.036
transcription_coactivator_activity (PPARGC1A, TCF8, MKL2)	133	3	0.038

Table 4.11: The table reports the Biological Process, the Cellular Component and the Molecular Function GO categories analyzing with GoMiner software the 54-predictive genes. In parenthesis are reported the predictive genes associated to each term. For each GO term are shown the total number of genes of the dataset (nr total genes), and the number of predictive genes (nr predictive genes) annotated with the term. Only the GO categories with p-value (last column) < 0.05 are considered. On grey background are evidenced the gene discussed more in detail in the text.

As emerged from the analysis with GoMiner, there were several functional categories related to the tumourigenesis processes in general ("cell adhesion", "insulin receptor signaling pathway", "cell proliferation", "regulation of cell proliferation") and this result could be expected. However it is also interesting to observe that some of these categories are more closely related to the cellular processes target of the chemotherapy agents used in this study, as the case of cell cycle or cell cycle arrest. In fact both paclitaxel and anthracyclines interfer with the normal cellular cycle, leading to cell apoptosis (see par. 1.8). Several studies reported that cell proliferation is related to the response to chemotherapy. Gianni and colleagues found that a high expression of genes related to cell-proliferation was correlated to higher sensitivity to paclitaxel/doxorubicin neoadjuvant chemotherapy [98]. It can be seen from the table 4.11 that two enriched categories are "cell proliferation" and "regulation of cell proliferation". Another interesting category that emerged from this analysis is the "response to hypoxia". It is known that hypoxia results in cellular responses that plays roles not only in tumour development and progression, but also in therapy responsiveness [100]. Hypoxia in solid tumours is associated with the development of chemoresistance. Recently it was reported that hypoxia leads to resistance to various classes of chemotherapeutic agents, including anthracyclines (daunorubicin and doxorubicin) [164].

In the GO Cellular Component an enriched category is the "nucleus". Doxorubicin and Epirubicin interact with DNA TopII complex at a nuclear level, therefore the presence of this GO term would make a biological sense.

It is clear that GoMiner software allowed us to have a general overview about the biological process and cellular localization significantly represented from the genes of the predictive signature. However, deducing from the 54 predictive genes which cellular pathway could be involved in the resistance or sensitivity of tumour cells to anthracyclines/paclitaxel based chemotherapy, can not be that easy. The GO categories identified as significantly represented in this gene set, would confirm, as reported from other studies, that the drug resistance phenotype is dependent not only on apoptotic pathways and cell cycle process, but also on other biological processes, which would have a role in establishing this condition.

In light of what stated above, I searched more in detail in the literature if the 54 genes identified were already known as involved in the drug resistance or reported in other predictive signature of response to the neoadjuvant regimen used in this study, that is paclitaxel plus antharcyclines.

Two genes were already reported from other groups in their predictive signature, MYC (v-Myc MYeloCytomatosis viral oncogene homolog (avian)) and NUF2 (NDC80 kinetochore complex component, homolog (*S. cerevisiae*)) and they look quite interesting.

C-Myc is an oncogene that functions in the stimulation of cell proliferation and apoptosis, in particular c-Myc is a down effector of the erb-B2 signaling pathway. As I previously reported (see par. 1.5.3.1), an improved response to anthracyclines-based NACT in erb-B2 positive patients was demonstrated in some studies, although its role in the sensitivity of cancer cells to chemotherapy is still unclear [165]. Recently Salter and colleagues showed in their study [166] that co-activation of MYC and E2F (E2F transcription factor 2) in tumours treated with TFAC (paclitaxel, 5-fluorouracil, adryamicin and cyclophosphamide) had the lowest percentage of responders. Taken together the presence in the predictive list of this gene would seem in agreement with what already reported from other groups. As a confirmation I found that this gene was more under-expressed in the partial responder patients than in the non responders.

NUF2 encodes a protein that is a component of a conserved protein complex associated with the centromere. Recent studies [167] demonstrated that nuf2 protein, together with hec1, are part of the stable core region of the kinetochore complex (Ndc80), an important attachment site for the microtubules during the mitotic spindle formation. Target of paclitaxel are the microtubule, therefore proteins involved in the microtubule dynamics, like Nuf2, could influence the action of this compound. Rouzier and colleagues reported NUF2 as a gene associated with pathologic complete response in basal-like breast tumours [19]. In my experiments, NUF2 was over-expressed in the PR patients, so this seems to agree with the results obtained from Rouzier *et al.* though they consider pCR as response and a specific subgroup of breast cancer (basal-like molecular subtype). Interestingly in the 54-genes signature there was another gene, SPC25 (SPC25, NDC80 kinetochore complex component, homolog (*S. cerevisiae*)) that encodes a protein that is part of the Ndc80 complex. SPC25 is an essential kinetochore component that plays a significant role in proper execution of mitotic events and is involved in kinetochore-microtubule interaction and spindle checkpoint activity [168].

As can be seen, only two genes are present in other predictive signatures but this result is in agreement with what is emerging from this type of studies. Between all the predictive signatures there is hardly any overlap, indicating that there may be not only one profile, but that several combinations of probes may predict response to chemotherapy.

After a detailed bibliographic research, also other genes of the signature showed to have some connection with the drug resistance phenomenon investigated in our study.

KFL5 (Kruppel-like Factor 5) is a transcription factor that regulates cellular signaling involved in cell proliferation and oncogenesis. Zhu and colleagues reported that KLF5 interacts with tumour suppressor p53 in regulating the expression of the inhibitor-of-apoptosis protein survivin, which plays a role in pathological process of cancer [169]. In particular KLF5 binds to the core survivin promoter and strongly induces its activity. Very recently it has been demonstrated that activation of survivin expression can induce the drug-resistance to paclitaxel in ovarian cancer [170]. Another study reported that breast cancers patients with higher KLF5 expression had shorter disease-free survival and overall survival than patients with lower KLF5 expression [171]. In my experiments, KFL5 was more markedly under-expressed in the PR patients, accordingly with the data reported above.

CDKN1b (Cyclin-dependent kinase inhibitor 1B) encodes the cycle regulatory protein p27, an inhibitor of cyclin-dependent kinase (CDK). It has been reported that this protein has a role in resistance to cancer chemotherapy, although the predictive value of p27 for chemosensitivity is not yet fully proved. However it was demonstrated *in vitro* that the sensitivity to doxorubicin was significantly higher in breast cancer cells with high CDKN1b expression [172]. Interestingly a very recent study [173] reported that c-myc (one of the genes identified with SVM) mediates the inhibitory effect of PDGF (Platelet-Derived Growth Factor) on the p27 promoter. Taken together, we could hypothesize that inhibition of the p27 expression *via* c-myc would have a role in establishing a drug resistance phenotype.

The ITGA6 gene product is the integrin alpha chain alpha 6. Integrins are integral cell-surface proteins composed of an alpha chain and a beta chain known to participate in cell adhesion as well as cell-surface mediated signaling. Liang and colleagues reported that the over-expression of ITGA6, together with others integrins subunits (ITGA2, ITGA5, ITGB4) increased the invasiveness of the tumours cells *in vitro*. It has also been demonstrated that the invasiveness phenotype is closely related to MDR (Multi Drug Resistance) phenotype [174]. A previous study reported that ITGA6 is more expressed in doxorubicin-resistant cells [175]. Gianni and colleagues [98] included ITGB2, another gene of the integrins family, in their predictive signature of response to paclitaxel/doxorubicin neoadjuvant chemotherapy. Taken together, it could be speculated that ITGA6 and the integrins in general are also involved in establishing a drug-resistance condition. In my experiment this gene showed an under-expression in the PR patients and it could be consistent with what explained above.

Also POSTN (PeriOSTiN) gene deserves more attention, although its connection with the drug resistance phenotype is less direct. This gene encodes the periostin, an osteoblast specific factor and, what is more interesting in light of our discussion, a potential marker for breast carcinoma BRCA1 mutations carriers. In fact it has been reported that *in vitro* cells with mutated BRCA1 show an up-regulation of POSTN [176]. Byrski and colleagues observed that women with a BRCA1 mutation who received docetaxel (a taxane as paclitaxel) in combination with doxorubicin (the chemotherapic used in this study) as neoadjuvant chemotherapy were less likely to respond to the treatment than women with no mutation. In contrast, BRCA1 carriers who were treated only with DNA-damaging chemotherapies, as anthracyclines, responded as frequently as non-carriers [177]. It is also reported from Rouzier *et al.* [19] that basal-like tumours are more responsive to paclitaxel/doxorubicin based chemotherapy. Basal-like tumours are typically found in BRCA1 mutations carriers, thus it would seem in contrast with the previous study. In light of what described we can not speculate too much about the connection between POSTN and drug sensitivity *via* BRCA1, only hypothesize that POSTN could play a role in drug resistance phenotype.

Other of the 54 genes are related to breast cancer progression and metastasis (CXCL9, CEBPD, IRS2, TCF8, ADAMTS5, PPARGC1A), but their direct involvement in drug resistance to paclitaxel/anthracyclines neoadjuvant chemotherapy is not reported in literature. Therefore other genes determined in this classifier, probably have additional, as yet unknown, functions in regulating drug response, which contribute to sensitivity or resistance of tumours to neoadjuvant chemotherapy.

4.7 USING THE SVM MODEL AS PREDICTIVE TOOL OF RESPONSIVENESS TO NEOADJUVANT CHEMOTHERAPY

At this point of my analysis, it was obvious to wonder how to use the SVM model as a predictive tool for the evaluation of the responsiveness to the treatment reported in this study (paclitaxel/anthracyclines based neoadjuvant chemotherapy). Predictive genes identification was the first step of the project, then it was important to understand how to use this gene signature for predicting the response to chemotherapy of a new patient, not yet classified as putative partial responder or not responder. Considering the dataset of 28 patients (17 PR and 11 NR) we identified with R-SVM feature selection process 54-genes, able to classify the patients in the two classes of response (PR and NR) with a LOO-CV accuracy of 85%. When we will have a gene expression profile of a new patient, we could use the SVM model, built with the 28 patients using the 54 genes, for classifying the new patient as partial responder or not responder based on the gene expression values of these 54 genes. The SVM output is a measure of distance of the patient, represented from a vector of 54 components (the gene expression values) from the optimal hyperplane that separates the patients in the two distinct classes of response. If we consider the new patient as a vector of 54 components (the 54 predictive genes) her class of response (positive or negative) will be established on the basis of the position of this vector respect to the separating hyperplane. For example, if the sign of the function that defines the hyperplane (see 3.14.1 for details) will be positive, then the new patient will be classified as partial responder (PR).

However, the SVM output is a value difficult to manage in statistics prediction problems. Therefore we used the sigmoid function, obtained applying the Platt's algorithm (see par. 3.14.3) to translate the SVM outputs into probability values that offered a more direct evaluation of the response class of the patient. In practice we transformed the SVM scores in a value ranging from 0 to 1, that expresses the probability to belong to the positive class of response (PR patients). A linear SVM was trained on the 28 patients using the 54 features and then was trained the sigmoid function (see par 3.14.3 for details) to map the SVM outputs into probabilities. In the figure 4.7 is shown the sigmoid function that fit the linear SVM output on the dataset of 28 patients (dataset I).

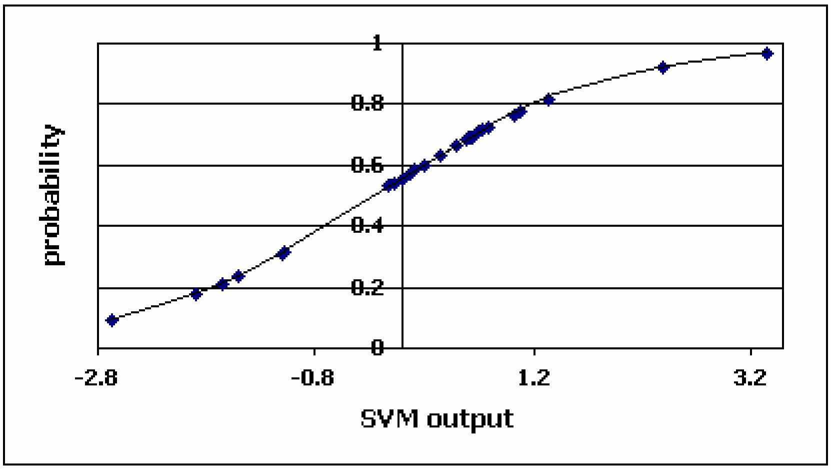


Figure 4.7: The fit of the sigmoid to the data for the linear SVM on the 28 patients dataset. Each point is the probability value computed for the 28 patients falling into a bin of width 0-1 (x-axis), corresponding to the SVM score calculated in the SVM training with the model of 54-genes signature (y-axis). The solid line is the fitted sigmoid to the probability values.

As can be seen from the figure, the sigmoid function is defined from 28 points, each corresponding to a patient.

In order to evaluate if the probability value computated with the Platt's algorithm for each SVM output score was consistent with the class of response of the patients, I correlated graphically the probability output of each patient with the class of response (PR or NR) (fig. 4.8). Since the probability value obtained with this approach was the probability of belonging to the positive class of response (PR in this case), I expected a probability value under 0.5 for the not responder patients and a probability value over 0.5 for the partial responder patients.

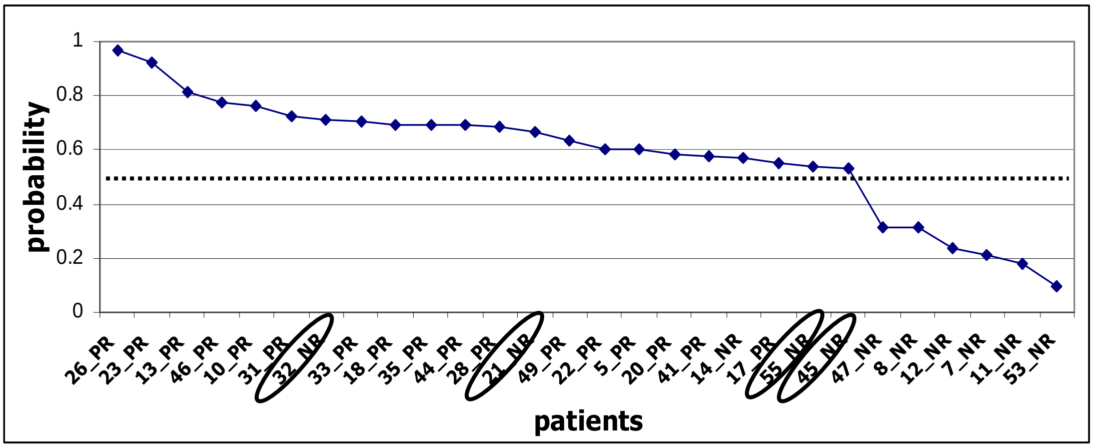


Figure 4.8: For each patient is shown the corresponding value of probability to belong to the positive (PR) class. The dashed line represents the probability threshold of 0.5 (see text for details). With a black ring are evidenced the patients with a probability value not consistent with their true class of response.

As emerged from the graph (fig. 4.8) the patients that had a probability value to belong to the PR class not concordant with their true class of response were five, so that means one plus respect to the estimation of the SVM model. These five patients were all belonging to the NR class (14 NR, 21 NR, 32 NR, 55 NR and 45 NR). Instead, another patient (17 PR), that was incorrectly classified with the 54-genes signature shows a probability value to belong to the positive class of response above 0.5, in agreement with her true class of response. Therefore, if we compare SVM outputs and probabilities, three patients were misclassified in both approaches (14 NR, 21 NR and 32 NR), one patient was misclassified only with the SVM model (17 PR) and two patients showed a low probability to belong to their true class of response using the sigmoid function (45 NR and 55 NR). The translation of SVM scores in probability values would under-estimates the accuracy of the classification computed with the LOO-CV. However, if we look at the probability values of the 2 patients misclassified with Platt's algorithm (and not with the SVM model) they show a "border line" probability value equal to 0.53. Consistently to this probability, the SVM scores of these two patients were also "border line" values, -0.1 (45 NR) and -0.09 (55 NR), compared to a point on the optimal hyperplane (that separates the positive responders and the negative responders) whose distance is 0. Also the patient 17 PR, misclassified with SVM but not with Platt's method, has to be considered a border line patient, having a SVM score of -0.01 and a probability value of 0.55. Taken together, these data indicate that these three patients (17 PR, 55 NR and 45 NR) representing "border cases", are differently treated from the two systems, that in fact have different range of values for establishing the class of response. These results, although discordant at the first analysis, could be read in a positive perspective. The translation of SVM outputs to probabilities can add informations for predicting the class of response of a new unclassified patient. In other words, we could give a statistical weight to the SVM score and make a better estimation of the prediction. For instance, the predictions of two patients, both classified as possible positive responders but with a probability of 0.7 and 0.54, could be considered differently at the moment of deciding about the most appropriate chemotherapic treatment to administer.

CONCLUDING REMARKS

During the last years breast cancer research made several efforts to identify markers predicting whether a patient will benefit from a specific chemotherapy treatment. The introduction of preoperative chemotherapy, also called NeoAdjuvant ChemoTherapy (NACT), made possible to have short-term responses to anticancer drugs, directly measuring the sensitivity/resistance of breast tumour to them [44]. It is unlikely that the resistance phenotype is the result of the action of a limited number of genes since the signaling pathways involved in tumour response to chemotherapy are complex and also dependent on the individual characteristics of the tumour.

The gene expression profling using microarray technology allowed to perform a systematic analysis of the gene-expression pattern of the tumour samples enabling researchers to better understand the tumour heterogeneity. It has been showed that gene expression profiling is a successful tool for the classification of breast cancer [23], for distinguishing prognostic subgroups [21, 53, 54], and it can also help to predict response to chemotherapy [94, 95]. Perou and colleagues [23] identified different molecular subtypes of breast cancer based on an intrinsic gene list: luminal-like, basal-like, erb-B2+ and normal-like subtypes. Subsequently it was reported that these subtypes could respond differently to neoadjuvant chemotherapy. In light of this I identified the molecular subtypes of the tumours included in my study and I evaluated how the clinical response to the treatment was associated to the molecular subtypes. It emerged that the luminal-like and erb-B2+ molecular subtypes were enriched of patients with a clinical Partial Response (PR) to the treatment.

Knowing that gene expression profiles of pre-treatment breast tumour biopsies could be correlated with the clinical response to NACT treatment, I focused my research on the identification of predictive genes of responsiveness to NACT regimen based on paclitaxel/anthracyclines (doxorubicin or epirubicin) drugs.

Two datasets of patients were considered to build the multigene predictive classifier: dataset I, containing the Partial Responders (PR) and the Non Responders (NR); dataset II, including in the group of Responders also the clinical Complete Responder (cCR) patients.

An unsupervised hierarchical cluster analysis on the dataset I and II showed that the patients grouped on the basis of biological parameters different from the response to chemotherapy. As observed also in other studies, the predictive genes of sensitivity/resistance to chemotherapy are probably a subtle set of genes, that could be masked from the high number of differentially expressed genes and hardly to be identified with an unsupervised approach.

Therefore a proper supervised approach was required to identify only those genes able to distinguish the tumours sensitive or resistant to the treatment. In order to have a more homogeneous class of positive response, I considered as Responders only the PR patients, excluding the cCR patients from the analysis.

The main obstacle of this study was the limited number of patients (28) compared to the high number of genes that defined their expression profile (more than 13000). In such a situation it would be easy to find set of genes discriminating the two classes of patients but these genes could have a low performance in classifying an independent set of patients (risk of overfitting). In light of this, a gene selection method based on Support Vector Machines (R-SVM) was considered a good approach for selecting a set of genes able to separate the PR and NR patients (gene classifier) with a good performance in classifying an independent test set. Since the small size of the dataset, I chose a Leave-One-Out Cross Validation procedure to assess the performance of the predictive genes selected with R-SVM.

Using this procedure we identified a set of 54 genes that could separate PR and NR patients with a LOO-CV accuracy of 85%. This set of genes represents a multigene predictor of sensitivity/resistance to the paclitaxel/anthracyclines NACT. Based on the 28 patients using the 54 predictive genes was trained a SVM model able to classify as responder or not responder a patient considering the gene expression values only of these selected genes.

A literature research focused on each gene of the predictive signature showed that some of these genes (MYC, NUF2, SPC25; KFL5, CDKN1b, ITGA6, POSTN) are 'biologically plausible' since they have connections with the mechanisms of resistance to paclitaxel, doxorubicin or epirubicin. Others genes are related to breast cancer progression and metastasis (CXCL9, CEBPD, IRS2, TCF8, ADAMTS5, PPARGC1A), but their direct involvement in drug resistance to paclitaxel/anthracyclines chemotherapy did not emerge in this literature research. It may be that the genes not found in the literature research as related to the drug resistance phenotype, could have additional, as yet unknown functions which contribute to sensitivity or resistance of breast tumours to NACT.

It should be pointed out that a single gene of the predictive list is not dicriminating *per se*, but the 54 genes are optimal for the classification only if taken together.

Next step of the analysis was the use of the trained SVM model as a predictive tool of responsiveness to NACT. Since the SVM outputs are uncalibrated values, not easily usable in statistics prediction problems, our idea was to translate this output in probabilities using a sigmoid function. In practice, we transformed the SVM scores, that are a measure of the distance of the patient from the optimal hyperplane that separates the responders from the non responders, in a measure of probability belonging to the positive class of response (PR patients). With a probability value we could better appreciate how the patient is classified by the SVM.

This approach, never used before in this type of studies, looks quite promising, although, so far, was applied to a small dataset of patients.

In fact we can consider this study as an "exploratory" analysis because of the small number of patients and the two classes of response, partial responders and non responders, not completely distinct in terms of regression of tumour mass. Although these limiting factors, the method adopted to face the object of the study, i.e. the identification of predictive genes of response to a particular regimen of neoadjuvant chemotherapy, has demonstrated to be effective and to produce reliable preliminary results. Applying the same methodological procedure to a larger cohoort of patients, in combination with a independent validation, could provide in the future the opportunity to better guide treatment decisions in breast cancer NACT.

Another aspect evaluated in this thesis was the analysis of the prognostic markers commonly used in the clinical setting to predict the tumour course: ER, PR, Bcl-2, p53, Erb-B2 and Ki-67. The level of the protein expression was measured with ImmunoHistoChemistry (IHC) and the mRNA abundance with microarrays. The correlation analysis performed between IHC and microarray data showed a significant correlation for ER, PR and Bcl-2 markers but not for p53, Erb-B2 and Ki-67. These results could indicate that post-transcriptional or translational mechanisms can modulate the level of some of these prognostic markers, leading IHC and microarrays to assess differently their level.

6. REFERENCES

1 Richert MM, Schwertfeger KL, Ryder JW, Anderson SM. An atlas of mouse mammary gland development. J Mammary Gland Biol Neoplasia. 2000;5(2):227-41.

2 Woodward WA, Chen MS, Behbod F, Rosen JM. On mammary stem cells. J Cell Sci. 2005; 118(Pt 16):3585-94.

3 Guinebretière JM, Menet E, Tardivon A, Cherel P, Vanel D. Normal and pathological breast, the histological basis. Eur J Radiol. 2005; 54(1):6-14.

4 Kakarala M, Wicha MS. Implications of the cancer stem-cell hypothesis for breast cancer prevention and therapy. J Clin Oncol. 2008; 26(17):2813-20.

5 Reya T, Morrison SJ, Clarke MF, Weissman IL. Stem cells, cancer, and cancer stem cells. Nature; 414(6859):105-11.

6 Kumle M. Declining breast cancer incidence and decreased HRT use. Lancet. 2008; 372(9639):608-10.

7 Parkin DM, Fernández LM. Use of statistics to assess the global burden of breast cancer. Breast J. 2006; 12 Suppl 1:S70-80.

8 Botha JL, Bray F, Sankila R, Parkin DM. Breast cancer incidence and mortality trends in 16 European countries. Eur J Cancer. 2003 (12):1718-29.

9 Polyak K. On the birth of breast cancer. Biochim Biophys Acta. 2001; 1552(1):1-13.

10 Hanby AM. Aspects of molecular phenotype and its correlations with breast cancer behaviour and taxonomy. Br J Cancer. 2005; 92(4):613-7.

11 Balslev I, Axelsson CK, Zedeler K, Rasmussen BB, Carstensen B, Mouridsen HT. The Nottingham Prognostic Index applied to 9,149 patients from the studies of the Danish Breast Cancer Cooperative Group (DBCG). Breast Cancer Res Treat. 1994; 32(3):281-90.

12 Tavassoli, FA; Devilee, P. World Health Organization Classification of Tumours. Pathology and Genetics. Tumours of the Breast and Female Genital Organs. Lyon: IARC Press; 2003. p. 98.

13 Fabbri A, Carcangiu ML, Carbone A. Histological Classification of Breast cancer. Breast Cancer Nuclear Medicine in Diagnosis and Therapeutic Options. 2007; 3-14.

14 Holland R, Peterse JL, Millis RR, Eusebi V, Faverly D, van de Vijver MJ, Zafrani B. Ductal carcinoma in situ: a proposal for a new classification. Semin Diagn Pathol. 1994 ;11(3):167-80.

15 Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. The value of histological grade in breast cancer: Experience from a large study with long-term follow-up. Histopathology 19: 403–410.

16 Singletary SE, Allred C, Ashley P, Bassett LW, Berry D, Bland KI, Borgen PI, Clark GM, Edge SB, Hayes DF, Hughes LL, Hutter RV, Morrow M, Page DL, Recht A, Theriault RL, Thor A, Weaver DL, Wieand HS, Greene FL. Staging system for breast cancer: revisions for the 6th edition of the AJCC Cancer Staging Manual. Surg Clin North Am. 2003; 83(4):803-19.

17 Singletary SE, Greene FL; Breast Task Force. Revision of breast cancer staging: the 6th edition of the TNM Classification. Semin Surg Oncol. 2003; 21(1):53-9.

18 Greene, FL, Page, DL, Fleming. AJCC (American Joint Committee on Cancer) Cancer Staging Manual, 6th ed ID Springer-Verlag, New York, 2002. Pp. 223-40.

19 Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Anderson K, Hess KR, Stec J, Ayers M, Wagner P, Morandi P, Fan C, Rabiul I, Ross JS, Hortobagyi GN, Pusztai L. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clin Cancer Res. 2005; 11(16):5678-85.

20 Nuyten DS, van de Vijver MJ. Using microarray analysis as a prognostic and predictive tool in oncology: focus on breast cancer and normal tissue toxicity. Semin Radiat Oncol. 2008; 18(2):105-14.

21 Pusztai L, Ayers M, Stec J, Clark E, Hess K, Stivers D, Damokosh A, Sneige N, Buchholz TA, Esteva FJ, Arun B, Cristofanilli M, Booser D, Rosales M, Valero V, Adams C, Hortobagyi GN, Symmans WF. Gene expression profiles obtained from fine-needle aspirations of breast cancer reliably identify routine prognostic markers and reveal large-scale molecular differences between estrogen-negative and estrogen-positive tumors. Clin Cancer Res. 2003; 9(7):2406-15.

22 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, Sgroi DC. Gene expression profiles of human breast cancer progression. Proc Natl Acad Sci U S A. 2003; 100(10):5974-9.

23 Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature. 2000; 406(6797):747-52.

24 Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lønning P, Børresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001; 98(19):10869-74.

25 Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc Natl Acad Sci U S A. 2003; 100(18):10393-8.

26 Rouzier R, Mathieu MC, Sideris L, Youmsi E, Rajan R, Garbay JR, André F, Marsiglia H, Spielmann M, Delaloge S. Breast-conserving surgery after neoadjuvant anthracycline-based chemotherapy for large breast tumors. Cancer. 2004; 101(5):918-25.

27 Millikan RC, Newman B, Tse CK, Moorman PG, Conway K, Dressler LG, Smith LV, Labbok MH, Geradts J, Bensen JT, Jackson S, Nyante S, Livasy C, Carey L, Earp HS, Perou CM. Epidemiology of basal-like breast cancer. Breast Cancer Res Treat. 2008; 109(1):123-39.

28 Carey LA, Perou CM, Livasy CA, Dressler LG, Cowan D, Conway K, Karaca G, Troester MA, Tse CK, Edmiston S, Deming SL, Geradts J, Cheang MC, Nielsen TO, Moorman PG, Earp HS, Millikan RC. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. JAMA. 2008; 295(21):2492-502.

29 Mathieu MC, Rouzier R, Llombart-Cussac A, Sideris L, Koscielny S, Travagli JP, Contesso G, Delaloge S, Spielmann M. The poor responsiveness of infiltrating lobular breast carcinomas to neoadjuvant chemotherapy can be explained by their biological profile. Eur J Cancer. 2004; 40(3):342-51.

30 Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, Livasy C, Carey LA, Reynolds E, Dressler L, Nobel A, Parker J, Ewend MG, Sawyer LR, Wu J, Liu Y, Nanda R, Tretiakova M, Ruiz Orrico A, Dreher D, Palazzo JP, Perreard L, Nelson E, Mone M, Hansen H, Mullins M, Quackenbush JF, Ellis MJ, Olopade OI, Bernard PS, Perou CM. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics. 2006; 7:96.

31 Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lønning PE, Brown PO, Børresen-Dale AL, Botstein D. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A. 2003; 100(14):8418-23.

32 Reis-Filho JS, Tutt AN. Triple negative tumours: a critical review. Histopathology. 2008; 52(1):108-18.

33 Reis-Filho JS, Milanezi F, Steele D, Savage K, Simpson PT, Nesland JM, Pereira EM, Lakhani SR, Schmitt FC. Metaplastic breast carcinomas are basal-like tumours. Histopathology. 2006; 49(1):10-21.

34 Abd El-Rehim DM, Pinder SE, Paish CE, Bell J, Blamey RW, Robertson JF, Nicholson RI, Ellis IO. Expression of luminal and basal cytokeratins in human breast carcinoma. J Pathol. 2004; 203(2):661-71.

35 Carey LA, Dees EC, Sawyer L, Gatti L, Moore DT, Collichio F, Ollila DW, Sartor CI, Graham ML, Perou CM. The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. Clin Cancer Res. 2007; 13(8):2329-34.

36 Foulkes WD, Brunet JS, Stefansson IM, Straume O, Chappuis PO, Bégin LR, Hamel N, Goffin JR, Wong N, Trudel M, Kapusta L, Porter P, Akslen LA. The prognostic implication of the basal-like (cyclin E high/p27 low/p53+/glomeruloid-microvascular-proliferation+) phenotype of BRCA1-related breast cancer. Cancer Res. 2004; 64(3):830-5.

37 Turner NC, Reis-Filho JS, Russell AM, Springall RJ, Ryder K, Steele D, Savage K, Gillett CE, Schmitt FC, Ashworth A, Tutt AN. BRCA1 dysfunction in sporadic basal-like breast cancer. Oncogene. 2007; 26(14):2126-32.

38 Abd El-Rehim DM, Ball G, Pinder SE, Rakha E, Paish C, Robertson JF, Macmillan D, Blamey RW, Ellis IO. High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. Int J Cancer. 2005 Sep 1;116(3):340-50.

39 Rottenberg S, Nygren AO, Pajic M, van Leeuwen FW, van der Heijden I, van de Wetering K, Liu X, de Visser KE, Gilhuijs KG, van Tellingen O, Schouten JP, Jonkers J, Borst P. Selective induction of chemotherapy resistance of mammary tumors in a conditional mouse model for hereditary breast cancer. Proc Natl Acad Sci U S A. 2007; 104(29):12117-22.

40 Yap TA, Boss DS, Fong PC. First in human phase I pharmacokinetic (PK) and pharmacodynamic (PD) study of KU-0059436 (Ku), a small molecule inhibitor of poly ADP-ribose polymerase (PARP) in cancer patients (p), including BRCA1/2 mutation carriers. J Clin Oncol. 2007; 25: 3529 (Abstract).

41 Sachelarie I, Grossbard ML, Chadha M, Feldman S, Ghesani M, Blum RH. Primary systemic therapy of breast cancer. Oncologist. 2006; 11(6):574-89.

42 Gonzales-Angulo AM, Morales-Vasquez F and Hortobagyi GN. Overview of Resistance to systemic therapy in patients with breast cancer. Breast Cancer Chemosensitivity. 2007; chapter 1: 1-22.

43 Hannemann J. Gene expression profiling in breast cancer: a link between biology and clinical decision making. 2008. Academisch Proefscrift (http://dare.uva.nl/document/104758) Chapter 1: 1-3.

44 van 't Veer LJ and Bernards R. Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature. 2008; 452: 564-569.

45 DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet. 1996 Dec;14(4):457-60.

46 Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci U S A. 1996; 93(20):10614-9.

47 Olivotto IA, Bajdik CD, Ravdin PM, Speers CH, Coldman AJ, Norris BD, Davis GJ, Chia SK, Gelmon KA. Population-based validation of the prognostic model ADJUVANT! for early breast cancer. J Clin Oncol. 2005; 23(12):2716-25.

48 Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. Breast Cancer Res Treat. 1992; 22(3):207-19.

49 Goldhirsch A, Wood WC, Gelber RD, Coates AS, Thürlimann B, Senn HJ. Meeting highlights: updated international expert consensus on the primary therapy of early breast cancer. J Clin Oncol. 2003; 21(17):3357-65.

50 Eifel P, Axelson JA, Costa J, Crowley J, Curran WJ Jr, Deshler A, Fulton S, Hendricks CB, Kemeny M, Kornblith AB, Louis TA, Markman M, Mayer R, Roter D. J Natl Cancer Inst. National Institutes of Health Consensus Development Conference Statement: adjuvant therapy for breast cancer, November 1-3, 2000. 2001; 93(13):979-89.

51 Morris SR, Carey LA. Curr Opin Oncol. Gene expression profiling in breast cancer. 2007; 19(6):547-51.

52 Weigelt B, Peterse JL, van 't Veer LJ. Breast cancer metastasis: markers and models. Nat Rev Cancer. 2005; 5(8):591-602.

53 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, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415(6871):530-6.

54 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, Bernards R. A gene-expression signature as a predictor of survival in breast cancer.N Engl J Med. 2002; 347(25):1999-2009.

55 Desmedt C, Ruíz-García E, André F. Gene expression predictors in breast cancer: current status, limitations and perspectives. Eur J Cancer. 2008; 44(18):2714-20.

56 Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, Desmedt C, Larsimont D, Cardoso F, Peterse H, Nuyten D, Buyse M, Van de Vijver MJ, Bergh J, Piccart M, Delorenzi M. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst. 2006; 98(4):262-72.

57 Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004; 351(27):2817-26.

58 Ma XJ, Hilsenbeck SG, Wang W, Ding L, Sgroi DC, Bender RA, Osborne CK, Allred DC, Erlander MG. The HOXB13:IL17BR expression index is a prognostic factor in early-stage breast cancer. J Clin Oncol. 2006; 24(28):4611-9.

59 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 W, Wang W, Enright E, Stecker K, Estepa-Sabal E, Smith B, Younger J, Balis U, Michaelson J, Bhan A, Habin K, Baer TM, Brugge J, Haber DA, Erlander MG, Sgroi DC. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer Cell. 2004; 5(6):607-16.

60 Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005; 365(9460):671-9.

61 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. Gene expression predictors of breast cancer outcomes. Lancet. 2003; 361(9369):1590-6.

62 Chang HY, Sneddon JB, Alizadeh AA, Sood R, West RB, Montgomery K, Chi JT, van de Rijn M, Botstein D, Brown PO. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. PLoS Biol. 2004; 2(2):E7.

63 Dai H, van't Veer L, Lamb J, He YD, Mao M, Fine BM, Bernards R, van de Vijver M, Deutsch P, Sachs A, Stoughton R, Friend S. A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. Cancer Res. 2005; 65(10):4059-66.

64 Chi JT, Wang Z, Nuyten DS, Rodriguez EH, Schaner ME, Salim A, Wang Y, Kristensen GB, Helland A, Børresen-Dale AL, Giaccia A, Longaker MT, Hastie T, Yang GP, van de Vijver MJ, Brown PO. Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. PLoS Med. 2006; 3(3):e47.

65 Liu R, Wang X, Chen GY, Dalerba P, Gurney A, Hoey T, Sherlock G, Lewicki J, Shedden K, Clarke MF. The prognostic role of a gene signature from tumorigenic breast-cancer cells. N Engl J Med. 2007; 356(3):217-26.

66 Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, van't Veer LJ, Perou CM. Concordance among gene-expression-based predictors for breast cancer. N Engl J Med. 2006; 355(6):560-9.

67 Foekens JA, Atkins D, Zhang Y, Sweep FC, Harbeck N, Paradiso A, Cufer T, Sieuwerts AM, Talantov D, Span PN, Tjan-Heijnen VC, Zito AF, Specht K, Hoefler H, Golouh R, Schittulli F, Schmitt M, Beex LV, Klijn JG, Wang Y. Multicenter validation of a gene expression-based prognostic signature in lymph node-negative primary breast cancer. J Clin Oncol. 2006; 24(11):1665-71.

68 Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics. 2005; 21(2):171-8.

69 de Azambuja E, Cardoso F, de Castro G Jr, Colozza M, Mano MS, Durbecq V, Sotiriou C, Larsimont D, Piccart-Gebhart MJ, Paesmans M. Ki-67 as prognostic marker in early breast cancer: a meta-analysis of published studies involving 12155 patients. Br J Cancer. 2007; 96(10):1504-13.

70 Molecular classification of breast cancer: implications for selection of adjuvant chemotherapy. Andre F, Pusztai L. Nat Clin Pract Oncol. 2006; 3(11):621-32.

71 Desmedt C, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, Bontempi G, Delorenzi M, Piccart M, Sotiriou C. Biological processes associated with breast cancer clinical outcome depend on the molecular subtype. Clin Cancer Res. 2008; 14(16): 5158-65.

72 Paik S, Tang G, Shak S, Kim C, Baker J, Kim W, Cronin M, Baehner FL, Watson D, Bryant J, Costantino JP, Geyer CE Jr, Wickerham DL, Wolmark N. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol. 2006; 24(23): 3207-14.

73 Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, Glas AM, d'Assignies MS, Bergh J, Lidereau R, Ellis P, Harris A, Bogaerts J, Therasse P, Floore A, Amakrane M, Piette F, Rutgers E, Sotiriou C, Cardoso F, Piccart MJ; TRANSBIG Consortium. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst. 2006; 98(17):1183-92.

74 Desmedt C, Piette F, Loi S, Wang Y, Lallemand F, Haibe-Kains B, Viale G, Delorenzi M, Zhang Y, d'Assignies MS, Bergh J, Lidereau R, Ellis P, Harris AL, Klijn JG, Foekens JA, Cardoso F, Piccart MJ, Buyse M, Sotiriou C; TRANSBIG Consortium. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. Clin Cancer Res. 2007; 13(11):3207-14.

75 Tewari M, Krishnamurthy A, Shukla HS. Predictive markers of response to neoadjuvant chemotherapy in breast cancer. Surg Oncol. 2008; 17(4):301-11.

76 Greenberg PAC, Hortobagyi GN. The importance of chemotherapy in locally advanced breast cancer. In: Wise L, Johnson Jr H, editors. Breast cancer: controversies in management. Armonk, NY: Futura publishing company Inc.; 1994. p. 439-58.

77 Portera CC, Swain SM. Neoadjuvant chemotherapy: a step closer to individualized therapy. In: Govindan R, editor. ASCO educational book. Alexandria, VA: ASCO; 2007. p. 51-5.

78 Ross AA, Cooper BW, Lazarus HM, Mackay W, Moss TJ, Ciobanu N, Tallman MS, Kennedy MJ, Davidson NE, Sweet D. Detection and viability of tumor cells in peripheral blood stem cell collections from breast cancer patients using immunohistochemical and clonogenic assay techniques. Blood. 1993; 82(9):2605-10.

79 Retsky M, Bonadonna G, Demicheli R, Folkman J, Hrushesky W, Valagussa P. Hypothesis: Induced angiogenesis after surgery in premenopausal node positive breast cancer patients is a major underlying reason why adjuvant chemotherapy works particularly well for those patients. Breast Cancer Research. 2004; 6(4):372-4.

80 Goldie JH, Coldman AJ. A mathematical model for relating thedrug sensitivity of tumors to their spontaneous mutation rate. Cancer Treatment Reports. 1979; 63(11-12):1727-33.

81 Norton L, Simon R. Tumor size, sensitivity to therapy and design of treatment schedules. Cancer Treatment Reports. 1977; 61(7):1307-17.

82 Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, van Oosterom AT, Christian MC, Gwyther SG. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst. 2000; 92(3):205-16.

83 Hayward JL, Carbone PP, Heuson JC, Kumaoka S, Segaloff A, Rubens RD. Assessment of response to therapy in advanced breast cancer: a project of the Programme on Clinical Oncology of the International Union Against Cancer, Geneva, Switzerland. Cancer. 1977; 39(3):1289-94.

84 Jones RL, Smith IE. Neoadjuvant treatment for early-stage breast cancer: opportunities to assess tumour response. Lancet Oncology. 2006; 7(10):869-74.

85 Bear HD, Anderson S, Smith RE, Geyer CE Jr, Mamounas EP, Fisher B, Brown AM, Robidoux A, Margolese R, Kahlenberg MS, Paik S, Soran A, Wickerham DL, Wolmark N. Sequential preoperative or postoperative docetaxel added to preoperative doxorubicin plus cyclophosphamide for operable breast cancer:National Surgical Adjuvant Breast and Bowel Project Protocol B-27. J Clin Oncol. 2006; 24(13):2019-27.

86 Guarneri V, Broglio K, Kau SW, Cristofanilli M, Buzdar AU, Valero V, Buchholz T, Meric F, Middleton L, Hortobagyi GN, Gonzalez-Angulo AM.. Prognostic value of pathologic complete response after primary chemotherapy in relation to hormone receptor status and other factors. Journal of Clinical Oncology. 2006; 24(7):1037-44.

87 Hannemann J. Gene expression profiling in breast cancer: a link between biology and clinical decision making. Academisch Proefscrift (http://dare.uva.nl/document/104758). 2008; Chapter 2: 5-23.

88 Sparano JA. Taxanes for breast cancer: an evidence-based review of randomized phase II and phase III trials. Clin Breast Cancer. 2000; 1(1): 32-40.

89 Cristofanilli M, Gonzalez-Angulo A, Sneige N, Kau SW, Broglio K, Theriault RL, Valero V, Buzdar AU, Kuerer H, Buccholz TA, Hortobagyi GN.. Invasive lobular carcinoma classic type: response to primary chemotherapy and survival outcomes. Journal of Clinical Oncology. 2005; 23(1):41-8.

90 Yarden Y, Sliwkowski MX. Untangling the the ErbB signalling network. Nat Rev Mol Cell Biol. 2001; 2(2):127-137.

91 Sullivan DM, Latham MD, Ross WE. Proliferation dependent topoisomerase II content as a determinant of antineoplastic drug action in human, mouse, and Chinese hamster ovary cells. Cancer Research. 1987; 47(15):3973-9.

92 Kariya S, Ogawa Y, Nishioka A, Moriki T, Ohnishi T, Ito S, Murata Y, Yoshida S. Relationship between hormonal receptors, HER-2, p53 protein, Bcl-2, and MIB-1 status and the antitumor effects of neoadjuvant anthracycline-based chemotherapy in invasive breast cancer patients. Radiation Medicine. 2005; 23(3):189-94.

93 Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, Scherf U, Lee JK, Reinhold WO, Weinstein JN, Mesirov JP, Lander ES, Golub TR. Chemosensitivity prediction by transcriptional profiling. Proc Natl Acad Sci U S A. 2001; 98(19):10787–92.

94 Ayers M, Symmans WF, Stec J, Damokosh AI, Clark E, Hess K, Lecocke M, Metivier J, Booser D, Ibrahim N, Valero V, Royce M, Arun B, Whitman G, Ross J, Sneige N, Hortobagyi GN, Pusztai L. Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer. J Clin Oncol. 2004; 22(12):2284–93.

95 Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Elledge R, Mohsin S, Osborne CK, Chamness GC, Allred DC, O'Connell P. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. Lancet. 2003; 362(9381):362–9.

96 Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Tham YL, Kalidas M, Elledge R, Mohsin S, Osborne CK, Chamness GC, Allred DC, Lewis MT, Wong H, O'Connell P. Patterns of resistance and incomplete response to docetaxel by gene expression profiling in breast cancer patients. J Clin Oncol. 2005; 23(6):1169–77.

97 Hannemann J, Oosterkamp HM, Bosch CA, Velds A, Wessels LF, Loo C, Rutgers EJ, Rodenhuis S, van de Vijver MJ. Changes in gene expression associated with response to neoadjuvant chemotherapy in breast cancer. J Clin Oncol. 2005; 23(15):3331–42.

98 Gianni L, Zambetti M, Clark K, Baker J, Cronin M, Wu J, Mariani G, Rodriguez J, Carcangiu M, Watson D, Valagussa P, Rouzier R, Symmans WF, Ross JS, Hortobagyi GN, Pusztai L, Shak S. Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer. Journal of Clinical Oncology. 2005; 23(29):7265-77.

99 Thuerigen O, Schneeweiss A, Toedt G, Warnat P, Hahn M, Kramer H, Brors B, Rudlowski C, Benner A, Schuetz F, Tews B, Eils R, Sinn HP, Sohn C, Lichter P. Gene expression signature predicting pathologic complete response with gemcitabine, epirubicin, and docetaxel in primary breast cancer. Journal of Clinical Oncology. 2006; 24(12):1839-45.

100 Dressman HK, Hans C, Bild A, Olson JA, Rosen E, Marcom PK, Liotcheva VB, Jones EL, Vujaskovic Z, Marks J, Dewhirst MW, West M, Nevins JR, Blackwell K. Gene expression profiles of multiple breast cancer phenotypes and response to neoadjuvant chemotherapy. Clin Cancer Res. 2006; 2(3 Pt 1):819-26.

101 Hess KR, Anderson K, Symmans WF, Valero V, Ibrahim N, Mejia JA, Booser D, Theriault RL, Buzdar AU, Dempsey PJ, Rouzier R, Sneige N, Ross JS, Vidaurre T, Gómez HL, Hortobagyi GN, Pusztai L. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. Journal of Clinical Oncology. 2006; 24(26):4236-44.

102 Györffy B, Serra V, Jürchott K, Abdul-Ghani R, Garber M, Stein U, Petersen I, Lage H, Dietel M, Schäfer R. Prediction of doxorubicin sensitivity in breast tumors based on gene expression profiles of drug-resistant cell lines correlates with patient survival. Oncogene. 2005; 24(51):7542-51.

103 Bonnefoi H, Potti A, Delorenzi M, Mauriac L, Campone M, Tubiana-Hulin M, Petit T, Rouanet P, Jassem J, Blot E, Becette V, Farmer P, André S, Acharya CR, Mukherjee S, Cameron D, Bergh J, Nevins JR, Iggo RD. Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial. Lancet Oncology. 2007; 8(12):1071-8.

104 Nagasaki K, Miki Y. Molecular prediction of therapeutic response to neoadjuvant chemotherapy in breast cancer. Breast Cancer. 2008; 15(2):117-120.

105 Jansen MP, Foekens JA, van Staveren IL, Dirkzwager-Kiel MM, Ritstier K, Look MP, Meijer-van Gelder ME, Sieuwerts AM, Portengen H, Dorssers LC, Klijn JG, Berns EM. Molecular classification of tamoxifenresistant breast carcinomas by gene expression profiling. J Clin Oncol. 2005; 23(4):732–40.

106 Schardt JA, Meyer M, Hartmann CH, Schubert F, Schmidt-Kittler O, Fuhrmann C, Polzer B, Petronio M, Eils R, Klein CA. Genomic analysis of single cytokeratin-positive cells from bone marrow reveals early mutational events in breast cancer. Cancer Cel. 2005; 8(3):227–39.

107 Nagrath S, Sequist LV, Maheswaran S, Bell DW, Irimia D, Ulkus L, Smith MR, Kwak EL, Digumarthy S, Muzikansky A, Ryan P, Balis UJ, Tompkins RG, Haber DA, Toner M. Isolation of rare circulating tumour cells in cancer patients by microchip technology. Nature. 2007; 450(7173):1235–9.

108 Pando MP, Kotraiah V, McGowan K, Bracco L, Einstein R. Alternative isoform discrimination by the next generation of expression profiling microarrays. Expert Opin Ther Targets 2006; 10(4):613–25.

109 Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. Journal of Clinical Oncology. 2005; 23(29): 7332-41.

110 Gonzalez-Angulo AM, Morales-Vasquez F, Hortobagyi GN. Overview of resistance to systemic therapy in patients with breast cancer. Adv Exp Med Biol. 2007; 608:1-22.

111 Gottesman MM, Fojo T, Bates SE. Multidrug resistance in cancer: role of ATP-dependent transporters. Nature Rev. 2002; 2(1):48-58.

112 Hahn WC, Weinberg RA. Modelling the molecular circuitry of cancer. Nat Rev Cancer. 2002; 2(5):331–41

113 Yague E, Raguz S. Drug resistance in cancer. British Journal of Cancer. 2005; 93(9): 973-76.

114 Ferguson LR, De Flora S. Multiple drug resistance, antimutagenesis and anticarcinogenesis. Mutat Res. 2005; 591(1-2):24-33.

115 Scheffer GL, Schroeijers AB., Izquierdo MA, Wiemer EA, Scheper RJ. Lung resistance-related protein/major vault protein and vaults in multidrug-resistant cancer. Curr. Opin. Oncol. 2000; 12(6):550–6.

116 Schuetz EG, Beck WT, Schuetz JD. Modulators and substrates of P-glycoprotein and cytochrome P4503A coordinately up-regulate these proteins in human colon carcinoma cells. Mol. Pharmacol. 1996; 49(2):311–8.

117 McCubrey JA, Steelman LS, Abrams SL, Lee JT, Chang F, Bertrand FE, Navolanic PM, Terrian DM, Franklin RA, D'Assoro AB, Salisbury JL, Mazzarino MC, Stivala F, Libra M. Roles of the RAF/MEK/ERK and PI3K/PTEN/AKT pathways in malignant transformation and drug resistance. Adv Enzyme Regul. 2006; 46:249-79.

118 Simpson D, Plosker GL. Paclitaxel as adjuvant or neoadjuvant therapy in early breast cancer. Drugs. 2004; 64(16): 1839-1847.

119 Rowinski EK, Donehower RC. Antimicrotubule agents. Pharmacology of Cancer Chemotherapy. Cancer: Principle and Practice of Oncology. 1997 Chapter 19.8: 467-82.

120 Rowinsky EK, Donehower RC. Drug Therapy-Paclitaxel. New Engl Journal of Med. 1995; 332 (15): 1004-14.

121 Jordan MA, Wilson L. Microtubules as a target for anticancer drugs. Nat Rev Cancer. 2004; 4(4):253-65.

122 Martello L, Verdier-Pinard P, Shen HJ, He L, Torres K, Orr GA, Horwitz SB. Elevated levels of microtubule destabilizing factors in a taxol-resistant/dependent A549 cell line with an alpha-tubulin mutation. Cancer Res. 2003; 63:1207-1213.

123 Rouzier R, Rajan R, Wagner P, Hess KR, Gold DL, Stec J, Ayers M, Ross JS, Zhang P, Buchholz TA, Kuerer H, Green M, Arun B, Hortobagyi GN, Symmans WF, Pusztai L. Microtubule-associated protein tau: a marker of paclitaxel sensitivity in breast cancer.Proc Natl Acad Sci U S A. 2005; 102(23):8315-20.

124 Stewart CF, Ratain MJ. Topoisomerase Interactive Agents. Pharmacology of Cancer Chemotherapy. Cancer: Principle and Practice of Oncology. 1997. Chapter 19.7: 452-66.

125 Binaschi M, Bigioni M, Cipollone A, Rossi C, Goso C, Maggi CA, Capranico G, Animati F. Anthracyclines: selected new developments. Curr Med Chem Anticancer Agents. 2001; 1(2):113-30.

126 Peng H, Wang MM, Jiang LY, Liu HT, Sun JZ. Paclitaxel-doxorubicin sequence is more effective in breast cancer cells with heat shock protein 27 overexpression. Chinese Med Journal. 2008; 121(20):1975-9.

127 Kubo A, Yoshikawa A, Hirashima T, Masuda N, Takada M, Takahara J, Fukuoka M, Nakagawa K. Point mutations of the topoisomerase IIalpha gene in patients with small cell lung cancer treated with etoposide. Cancer Res. 1996; 56(6):1232-6.

128 Rayson D, Richel D, Chia S, Jackisch C, van der Vegt S, Suter T. Athracycline-trastuzumab regimens for HER2/neu-overexpressing breast cancer: current experience and future strategies. Annals of Oncology. 2008; 19(9):1530-9.

129 Park K, Kim J, Lim S, Han S. Topoisomerase II-alpha (topoII) and HER2 amplification in breast cancers and response to preoperative doxorubicin chemotherapy. Eur J Cancer. 2003; 39(5):631-4.

130 Razis ED, Fountzilas G. Paclitaxel: Epirubicin in metastatic breast cancer- a review. 2001; 12(5):593-8.

131 Chomczynski P, Mackey K, Drews R, Wilfinger W. DNAzol: a reagent for the rapid isolation of genomic DNA. Biotechniques. 1997; 22(3):550-3.

132 Yue H, Eastman PS, Wang BB, Minor J, Doctolero MH, Nuttall RL, Stack R, Becker JW, Montgomery JR, Vainer M, Johnston R. An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. Nucleic Acids Res. 2001; 29(8):E41-1.

133 Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH. Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proc Natl Acad Sci U S A. 1990; 87(5):1663-7.

134 Feldman AL, Costouros NG, Wang E, Qian M, Marincola FM, Alexander HR, Libutti SK. Advantages of mRNA amplification for microarray analysis. Biotechniques. 2002; 33(4):906-12, 914.

135 Polacek DC, Passerini AG, Shi C, Francesco NM, Manduchi E, Grant GR, Powell S, Bischof H, Winkler H, Stoeckert CJ Jr, Davies PF. Fidelity and enhanced sensitivity of differential transcription profiles following linear amplification of nanogram amounts of endothelial mRNA. Physiol Genomics. 2003; 13(2):147-56.

136 't Hoen PA, de Kort F, van Ommen GJ, den Dunnen JT. Fluorescent labelling of cRNA for microarray applications. Nucleic Acids Res. 2003; 31(5):e20.

137 Rosati P and Colombo R. Microscopia confocale. Tecniche per lo studio dei campioni biologici. La Cellula -Seconda edizione- 1999; 2: 50-1.

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

139 Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, Quackenbush J. Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biol. 2002; 3(11):research0062.

140 Quackenbush J. Microarray data normalization and transformation. Nat Genet. 2002; 32 Suppl:496-501.

141 Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001; 17(6):520-5.

142 Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000; 403(6769):503-11.

143 Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A; 99(10):6567-72.

144 Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998; 95(25):14863-8.

145 Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A. 2002; 99(10):6567-72.

146 Mukherjee S. Classifyng microarray data using support vector machines. A Practical Approach to Microarray Data Analysis. 2003; Chapter 9:166-185.

147 Vapnik V. Statistical Learning Theory. Wiley, 1998.

148 Guyon I, Weston J, Barnhill S. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning. 2002, 46: 389-422.

149 Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A. 2002;99(10):6562-6.

150 Zhang X, Lu X, Shi Q, Xu XQ, Leung HC, Harris LN, Iglehart JD, Miron A, Liu JS, Wong WH. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. BMC Bioinformatics. 2006;7:197.

151 Li F, Yang Y. Analysis of recursive gene selection approaches from microarray data. Bioinformatics. 2005;21(19):3741-7.

152 Ivan Kojadinovic, Thomas Wottka. Comparison between a filter and a wrapper approach to variable subset selection in regression problems. European Symposium on Intelligent Techniques (ESIT) 2000.

153 Platt J. Probabilistic Outputs for Support Vector Machines and Comparisons to regularized likelihood methods. Advances in Large Margin Classifiers. 1999.

154 Armitage P, Berry G, Matthews JNS Analysing non-normal data (Rank-correlation). Statistical methods in medical research (Blackwell Science) fourth edition; cap. 10 par. 5: 288-292.

155 Armitage P, Berry G, Matthews JNS General contingency tables (Comparison of several groups). Statistical methods in medical research (Blackwell Science) fourth edition; cap. 8 par. 6: 231-232.

156 Clark JI, Brooksbank C, Lomax J. It's all GO for plant scientists. Plant Physiol. 2005; 138(3):1268-79.

157 Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. 2003;4(4):R28.

158 Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JC, Lashkari D, Shalon D, Brown PO, Botstein D. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci U S A. 1999; 96(16):9212-7.

159 Bertucci F, Nasser V, Granjeaud S, Eisinger F, Adelaïde J, Tagett R, Loriod B, Giaconia A, Benziane A, Devilard E, Jacquemier J, Viens P, Nguyen C, Birnbaum D, Houlgatte R. Gene expression profiles of poor-prognosis primary breast cancer correlate with survival. Hum Mol Genet. 2002; 11(8):863-72.

160 Ginestier C, Charafe-Jauffret E, Bertucci F, Eisinger F, Geneix J, Bechlian D, Conte N, Adélaïde J, Toiron Y, Nguyen C, Viens P, Mozziconacci MJ, Houlgatte R, Birnbaum D, Jacquemier J. Distinct and complementary information provided by use of tissue and DNA microarrays in the study of breast tumor markers. Am J Pathol. 2002; 161(4):1223-33.

161 Urruticoechea A, Smith IE, Dowsett M. Proliferation marker Ki-67 in early breast cancer. J Clin Oncol. 2005; 23(28):7212-20.

162 Assersohn L, Salter J, Powles TJ, A'hern R, Makris A, Gregory RK, Chang J, Dowsett M. Studies of the potential utility of Ki67 as a predictive molecular marker of clinical response in primary breast cancer. Breast Cancer Res Treat. 2003; 82(2):113-23.

163 Pusztai L, Krishnamurti S, Perez Cardona J, Sneige N, Esteva FJ, Volchenok M, Breitenfelder P, Kau SW, Takayama S, Krajewski S, Reed JC, Bast RC Jr, Hortobagyi GN. Expression of BAG-1 and BcL-2 proteins before and after neoadjuvant chemotherapy of locally advanced breast cancer. Cancer Invest. 2004; 22(2):248-56.

164 Sullivan R, Paré GC, Frederiksen LJ, Semenza GL, Graham CH. Hypoxia-induced resistance to anticancer drugs is associated with decreased senescence and requires hypoxia-inducible factor-1 activity. Mol Cancer Ther. 2008; 7(7):1961-73.

165 Liao DJ, Thakur A, Wu J, Biliran H, Sarkar FH. Perspectives on c-Myc, Cyclin D1, and their interaction in cancer formation, progression, and response to chemotherapy. Crit Rev Oncog. 2007; 13(2):93-158.

166 Salter KH, Acharya CR, Walters KS, Redman R, Anguiano A, Garman KS, Anders CK, Mukherjee S, Dressman HK, Barry WT, Marcom KP, Olson J, Nevins JR, Potti A. An integrated approach to the prediction of chemotherapeutic response in patients with breast cancer. PLoS ONE. 2008; 3(4):e1908.

167 DeLuca JG, Dong Y, Hergert P, Strauss J, Hickey JM, Salmon ED, McEwen BF. Hec1 and nuf2 are core components of the kinetochore outer plate essential for organizing microtubule attachment sites. Mol Biol Cell. 2005; 16(2):519-31.

168 Ciferri C, De Luca J, Monzani S, Ferrari KJ, Ristic D, Wyman C, Stark H, Kilmartin J, Salmon ED, Musacchio A. Architecture of the human ndc80-hec1 complex, a critical constituent of the outer kinetochore. J Biol Chem. 2005; 280(32):29088-95.

169 Zhu N, Gu L, Findley HW, Chen C, Dong JT, Yang L, Zhou M.J. KLF5 Interacts with p53 in regulating survivin expression in acute lymphoblastic leukemia. Biol Chem. 2006; 281(21):14711-8.

170 **.** Weng D, Song X, Xing H, Ma X, Xia X, Weng Y, Zhou J, Xu G, Meng L, Zhu T, Wang S, Ma D. Implication of the Akt2/survivin pathway as a critical target in paclitaxel treatment in human ovarian cancer cells. Cancer Lett. 2009; 273(2):257-65.

171 Tong D, Czerwenka K, Heinze G, Ryffel M, Schuster E, Witt A, Leodolter S, Zeillinger R. Expression of KLF5 is a prognostic factor for disease-free survival and overall survival in patients with breast cancer. Clin Cancer Res. 2006; 12(8):2442-8.

172 Yang Q, Sakurai T, Yoshimura G, Takashi Y, Suzuma T, Tamaki T, Umemura T, Nakamura Y, Nakamura M, Utsunomiya H, Mori I, Kakudo K. Overexpression of p27 protein in human breast cancer correlates with in vitro resistance to doxorubicin and mitomycin C. Anticancer Res. 2000; 20(6B):4319-22.

173 Bagui TK, Cui D, Roy S, Mohapatra S, Shor AC, Ma L, Pledger WJ. Inhibition of p27(Kip1) gene transcription by mitogens. Cell Cycle. 2009 Jan;8(1).

174 Liang Y, Meleady P, Cleary I, McDonnell S, Connolly L, Clynes M. Selection with melphalan or paclitaxel (Taxol) yields variants with different patterns of multidrug resistance, integrin expression and in vitro invasiveness. Eur J Cancer. 2001; 37(8):1041-52.

175 Narita T, Kimura N, Sato M, Matsuura N, Kannagi R. Altered expression of integrins in doxorubicin-resistant human breast cancer cells. Anticancer Res. 1998; 18(1A):257-62.

176 Quaresima B, Romeo F, Faniello MC, Di Sanzo M, Liu CG, Lavecchia A, Taccioli C, Gaudio E, Baudi F, Trapasso F, Croce CM, Cuda G, Costanzo F. BRCA1 5083del19 mutant allele selectively up-regulates periostin expression in vitro and in vivo. Clin Cancer Res. 2008 Nov; 14(21):6797-803.

177 Byrski T, Gronwald J, Huzarski T, Grzybowska E, Budryk M, Stawicka M, Mierzwa T, Szwiec M, Wiśniowski R, Siolek M, Narod SA, Lubinski J; Polish Hereditary Breast Cancer Consortium. Response to neo-adjuvant chemotherapy in women with BRCA1-positive breast cancers. Breast Cancer Res Treat. 2008; 108(2):289-96.

Acknowledgements

CRIBI Biotech centre

Giorgio Valle
Stefano Campanaro
Elisa Caniato
Simone Picelli

Department of Pure and Applied Mathematics

Riccardo Cardin
Alessandro Sperduti

The Netherlands Cancer Institute

Laura van 't Veer and group

AB ANALITICA s.r.l.

Rossana Pelizzaro
Dino Paladin

Oncology and Surgery sections of Dolo, Mirano and Noale Hospitals

Donata Sartori
Mario Bari
Giovanni Tacchetti

Special thanks to

My parents and Eva
Hans
Chiara