UNIVERSITÀ
DEGLI STUDI
DI PADOVA

UNIVERSITA' DEGLI STUDI DI PADOVA

Dipartimento di Biologia

Scuola di Dottorato di Ricerca in Bioscienze e Biotecnologie
Indirizzo: Genetica e Biologia Molecolare dello Sviluppo
Ciclo XXVI

A computational approach to identify predictive gene signatures in Triple Negative Breast Cancer

Direttore della Scuola: Prof. Giuseppe Zanotti
Coordinatore indirizzo: Prof. Rodolfo Costa
Supervisore: Prof. Paolo Bonaldo
Co-supervisore: Prof. Silvio Bicciato

Dottoranda: Simona Nuzzo

Abstract

Microarray technology has been extensively used to detect patterns in gene expression that stem from regulatory interactions. Seminal studies demonstrated that the synergistic use of microarray-based techniques and bioinformatics analysis of genomic data might not only further the understanding of pathological phenotypes, but also provide lists of genes to dissect a disease into distinct groups, with different diagnostic or prognostic characteristics. Nonetheless, optimism for microarray-based technologies as clinical tools has suffered of both perceptual and real setbacks. Criticism is largely on the ground of general non-reproducibility of gene signatures and the inability to replicate results.

The research activity illustrated in this thesis aimed at fulfilling methodological gaps still hampering the identification of gene signatures with proved prognostic and predictive value and, finally, affecting their reliability, reproducibility, and applicability. Specifically, we developed computational methods to efficiently merge gene expression profiles of tumors from multiple, independent, retrospective studies and to construct meta-datasets storing high throughput gene expression profiles and clinical information from thousands cancer patients. Moreover, we expanded on the concept of *gene signature* and derived *consensus signatures*, i.e. linear weighted combinations of gene signatures that, singularly, recapitulate independent signaling pathways or specific molecular mechanisms, while intertwined together render a more comprehensive molecular model of tumor progression or chemo-resistance.

This approach has been applied to breast cancer, in general, and to triple negative breast cancer (TNBC), in particular, and resulted in the identification of gene signature combinations with increased robustness and power to predict cancer progression or response to therapy over the use of single signatures.

Riassunto

Tra le varie tecnologie high-throughput, i microarray, unitamente agli strumenti bioinformatici per l'analisi dei relativi segnali, rappresentano una risorsa preziosissima per lo studio dei meccanismi di regolazione trascrizionale che contribuiscono a determinare gli stati fisiologici e patologici delle cellule. In ambito oncologico, molti studi hanno dimostrato che l'utilizzo sinergico dei microarray e della bioinformatica può contribuire, non solo a una maggiore comprensione dei meccanismi coinvolti nel cancro, ma anche alla definizione di liste di geni con i quali identificare gruppi patologici con diverse caratteristiche diagnostiche o prognostiche. Tuttavia, l'ottimismo per le tecnologie basate sui microarray come strumenti clinici ha subito delle battute d'arresto sia percettive che reali . La critica è in gran parte dovuta alla non riproducibilità delle firme geniche e all'incapacità di replicare i risultati. L'attività di ricerca illustrata in questa tesi ha avuto l'obiettivo di colmare lacune metodologiche che ancora ostacolano l'identificazione di marcatori prognostici e predittivi e che, infine, inficiano affidabilità, riproducibilità ed applicabilità. In particolare, sono stati sviluppati metodi computazionali per integrare set multipli di dati di profili di espressione genica di tumori provenienti da studi indipendenti gli uni dagli altri al fine di costruire un meta-dataset di profili di espressione genica con associate le informazioni cliniche dei pazienti. Inoltre, è stato ampliato il concetto di firma genica e di firme consenso derivate, cioè combinazioni lineari di firme geniche che, singolarmente, ricapitolano vie di segnalazione indipendenti o meccanismi molecolari specifici, mentre unite insieme rendono un modello molecolare di progressione del tumore o chemio-resistenza più completo. Questo approccio è stato applicato al tumore al seno, in generale , e al tumore triplo negativo (TNBC), in particolare , e ha portato all'identificazione di combinazioni di firme geniche con maggiore robustezza e potere di predire la progressione del tumore o la risposta alla terapia rispetto all'uso delle firme singole.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations

The completion of genome sequencing of specific organisms, combined with technological advances in the capability to detect genome-wide changes, has heralded a new era in the quest for understanding the complex molecular mechanisms of living organisms. The exploration of all genes at once, in a systematic fashion, represented a sort of revolution that shifted molecular biology from a reductionist, hypothesis-driven approach towards deciphering the signaling networks that operate in the cell and the molecular basis of physiological states. Microarray technology for the parallel quantification of large numbers of messenger RNAs is one of the technical cornerstones of a new approach to molecular biology, the so-called *omics revolution*, in which an organism is viewed as an integrated and interacting network of genes, proteins and biochemical reactions. According to the central dogma of molecular biology, genomic DNA is first transcribed into mRNA, which thereafter is translated into protein. Proteins play critical roles in most intra- and extra-cellular activities, including enzymatic, regulatory and structural functions. However, relative difficulties of expression measurement capabilities at the protein level and availability of high-throughput technologies for detection of individual mRNAs have led to the wide use of microarrays to simultaneously measure the sum of all mRNA expression in a sample. Like most classical methods for analysis of gene expression at the mRNA level, the basic principle of microarray technology is complementary hybridization of nucleotides, as explained by the Watson–Crick double helical model of DNA. Microarrays measure transcriptomic modifications that, either at the single gene level or collectively in multiple genes, are supposed to induce or capture changes in protein expression. As opposed to the classical northern-blotting analysis, the mRNA from a given cell line or tissue is used to generate a labeled sample, sometimes termed the *target*, which is hybridized in parallel to a large number of DNA sequences, immobilized on a solid surface in an ordered topology. Tens of thousands of transcript species can be detected and quantified simultaneously. In the last two decades, DNA microarray technology has been advancing rapidly. The development of more powerful robots for arraying, new surface technology for glass and silicon slides, and new labeling protocols and dyes, together with increasing genome-sequence information for different organisms, including humans, allowed extending the quality and complexity of microarray experiments. Although academic groups and commercial suppliers have developed many different microarray systems, in the most commonly used technology the arrayed material, generally termed the *probe* (being the equivalent to the probe used in a northern blot analysis), is an oligonucleotide sequence. In oligonucleotide arrays, short 20–25mers are synthesized in situ, either by photolithography onto silicon wafers (high-density-oligonucleotide arrays from Affymetrix) or by ink-jet technology (developed by Rosetta Inpharmatics and licensed to Agilent Technologies) or the BeadArray technology (from Illumina company). Since the late 1990s, the power and potential of microarray technology have been fully appreciated and applied to develop novel descriptions of complex diseases as cancer. The underlying hypothesis is that novel genes and pathways, previously not implicated in the patho-physiology of a certain tumor, might emerge from microarray studies to provide new theories regarding the disease process and potential therapeutic drug targets. The spread in use of the technology was unprecedented, with exponential growth in the number of publications reporting results from its application. Seminal studies demonstrated that the synergistic use of microarray-based techniques and bioinformatics analysis of genomic data might not only further the understanding of cancer taxonomy, but also provide lists of genes that can dissect a tumor into distinct groups, with different diagnostic or prognostic characteristics. The identification of these gene expression signatures held promise for being more effective than standard prognostic and predictive factors. A demonstrable success occurred in early 2007 when the U.S. Food and Drug Administration approved MammaPrint, the first microarray-based commercial molecular prognostic test for breast cancer. Tumor profiling has also changed the perception of metastatic propensity suggesting that a metastasis trait may be encoded within the genome. The fact that specific collection of genes expressed in primary tumors could be predictive for metastasis allowed inferring that metastatic proclivity might be intimately wired to the same aberrant genetic pathways that control malignant progression at the primary tumor site. Although the invasion-metastasis cascade may originate at a level unobservable using microarray-based technology (e.g., molecular interactions occurring at the protein scale), some genes or functional classes of genes are invariably altered when tumor cells acquire malignant properties, including genes involved in cell-cycle control, adhesion, motility, apoptosis and angiogenesis. Nonetheless, optimism for microarray-based technologies as predictive tests of cancer recurrence has suffered both perceptual and real setbacks. Criticism is largely on the grounds of general non-reproducibility of gene signatures and the inability to replicate results in terms of significant genes identified from experiments in different laboratories and from different experimental platforms. Skepticism regarding reliability and reproducibility reflects the complexity of the analytical methods and the peculiar nature of the data generated by high-throughput technologies. Different microarray studies led to the identification of different gene expression signatures able to predict the clinical outcome but characterized by a minimal, if not null, number of overlapping genes. Again, several technical, analytical and biological reasons may, at least partially, explain these seemingly discrepant results. These include the use of different microarray platforms with different sets of probe and data normalization methods, as well as differences in the study populations. Two other major explanations are the lack of independent measurements between the expressed genes and the limited statistical power applied to select individual genes associated with clinical outcome. The intrinsic dependency of gene expression signals implies that if the expression of a particular gene is associated with clinical outcome, all other genes, whose expression is closely correlated with that gene, will also correlate with clinical outcome. Since the strength of correlation between the genes and clinical outcome varies from data set to data set, the rank order of these informative genes in the prognostic signatures is highly unstable, thus leading to different gene lists with a small overlap. The low statistical power of prognostic or predictive signatures is mostly due to the limited number of samples included in the different data sets used for the development of classifiers. In retrospective studies, an adequately powered sample size is the most important, and most overlooked, aspect of successful expression array analysis. The small number of samples in individual studies, particularly for human studies where there is a high degree of both intra- and inter-population variability, represents a major limitation for the detection of gene expression signatures and ultimately results in disease biomarkers that are population dependent, rather than having global applicability (Bhattacharya and Mariani, 2009). Regarding the computational approach, concepts inspiring the marker discovery process can be classified in top-down or bottom-up approaches (Sotiriou and Piccart, 2007). In the top-down approach, a prognostic model is derived simply by looking for gene expression patterns associated with clinical outcome without any a priori biological assumption, whereas in the bottom-up, gene expression profiles linked with a specific biological phenotype are first identified and subsequently correlated to survival. Both strategies rely only on gene expression data and/or clinical information and none of them include mechanistic insights in the discovery process. It is most likely that the selection of predictive genes on the basis of mechanistic insights, rather than solely on the basis of expression levels and outcome data, will dramatically improve reliability, robustness, and, ultimately, biological significance of prognostic and predictive signatures.

1.2 Breast cancer

In recent decades, we have witnessed an increased incidence of cancer, rendering cancer one of today's major public health issues. Currently, breast cancer is the most frequently diagnosed malignancy in women and it causes death mainly in European and American women. The use of screening mammograms in developed countries brought to the identification of more and more women diagnosed with breast cancer at an early stage (i.e., small size tumors and no invasion of regional lymph node). In the majority of these cases, surgery is the primary treatment, alone or in combination with radiotherapy. Unfortunately, despite early detection, up to 50% of these women will develop distant metastasis, i.e. development of new tumors in different organs. Metastatic breast cancer is unfortunately incurable. As a result, since the mid 1980s, randomized trials of adjuvant systemic therapy (i.e., after surgery) have been conducted in an effort to reduce the rate of recurrence and to prolong the survival of patients with operable disease (EBCTCG, 2005). Due to the importance of breast cancer for public health, this disease has been the subject of intense research for decades. Moreover, the introduction of high throughput technologies, such as gene expression profiling, has provided powerful tools to study and fight this disease. The use of high-throughput technologies for the analysis of cancers has provided new hints for understanding the diversity and heterogeneity of cancers and to devise classification methods that better recapitulate the biology and clinical behavior of human tumors. Microarray-based gene expression profiling has highlighted the existence of breast cancer subtypes with distinct biology and clinical behavior (Sotiriou and Pusztai, 2009; Weigelt et al., 2010). Instead, traditional histo-pathological characteristics, i.e. microscopic examination of the diseased tissues anatomy, are unable to capture the biologic heterogeneity of these tumors. Expression profiling class discovery studies have led to a working model for a breast cancer molecular taxonomy (Perou et al., 2000 ; Sorlie et al., 2003 ; Hu et al., 2006; Parker et al., 2009), which has become widely used and recently adopted for the design of clinical trials. Briefly, breast cancers were classified by hierarchical cluster analysis using an "intrinsic" gene list, i.e., list of "genes with significantly greater variation in expression between different tumors than between paired samples from the same tumour" (first described in Perou et al.,

2000) into at least one of different molecular subtype classes: the luminal (often differentiated into two subgroups, for example, luminal A and B), HER2-enriched, basal-like, and normal breast-like (Perou et al., 2000; Sorlie et al., 2001-2003, Hu et al., 2006, Parker et al., 2009). Luminal A tumors usually have intermediate to high expression of ESR1 (or ER, estrogen receptor gene) and ER-regulated genes and rarely have high ERBB2 expression (HER2 gene). Luminal B tumors usually have intermediate to high expression of ESR1 and ER-regulated genes and often have higher proliferation than luminal A tumors. Together, luminal tumors constitute the most common molecular subtype of breast tumor and represent approximately 50% of all tumors in most series. Tumors of this subgroup are associated with a good prognosis and can be treated with targeted therapies, e.g. selective oestrogen receptor modulators (SERMS), such as tamoxifen or, in post-menopausal women, aromatase inhibitors such as anastrozole. HER2-enriched tumors usually have intermediate to high expression of the ERBB2 gene and intermediate to low expression of ER gene and estrogen-regulated genes; this subtype comprises approximately 10% of all breast tumors. Before the introduction of Trastuzumab into breast cancer treatment in 2001 for metastatic disease (Slamon et al., 2001) and in 2005 for early breast cancer (Piccart-Gebhart et al., 2005), tumors of HER2-overexpressing type were associated with a poor prognosis (Sorlie et al., 2001-2003, Hu et al., 2006, Parker et al., 2009). Basal-like tumors usually have low expression of ER, PgR (or PR, progesterone receptor gene), and HER2, but have high proliferation rate. This particular subtype represents approximately 15-20% of breast cancers. They occur in younger women than other subgroups (Carey et al., 2006 ; Foulkes et al., 2004 ; Calza et al., 2006 ; Rakha et al., 2006). They have been shown to have worse overall and relapse-free survival rates than luminal and HER2-overexpressing subtypes (Sorlie et al., 2001) and they are more likely than other subtypes to metastasize to lung and brain, sites that are known to be associated with poor survival (Tsuda et al., 2000; Banerjee et al., 2006 ; Rodríguez-Pinilla et al., 2006; Hicks et al., 2006; Fulford et al., 2007). Basal-like tumours are highly proliferative tumours, which is thought to be largely due to their deficiencies in both p53 and retinoblastoma 1 (RB1) protein function (Perou, 2010). This subtype is also called *triple negative breast cancer (TNBC).* The main characteristics of triple-negative cancers that have emerged from the literature illustrate their similarities to basal-like cancers (Badve et al., 2011). TNBC is diagnosed by immunohistochemistry (IHC) methodologies to detect ER, PR and HER2 expression and joint guidelines by the American Society of Clinical Oncology (ASCO) and the College of American Pathologists (CAP) helped to standardize the techniques in the hope of improving their reliability and reproducibility (Penault-Llorca et al., 2009). To date, there is no targeted therapy available for treatment of this specific subtype and no preferred standard form of treatment for this group, and treatment should be selected as it is for other subtypes (Foulkes et al., 2010).

In cancer behavior, prognostication and prediction of therapy benefit are two main issues. The goal of prognostication is to predict the survival of a patient, or her risk to develop metastases without treatment. Specifically, prognosis attempts to predict the prospect of remission of a breast cancer patient from the usual course of disease after the initial surgery. This information is extremely important because it assists oncologists in determining which breast cancer patients require chemo-, hormono- or other systemic therapies, and which women can safely be treated with radiotherapy alone. There are several clinical variables commonly used for breast cancer prognosis. The risk of recurrence is primarily determined by the age of the patient, nodal status, tumor size, histological grade, the expression status of the hormonal receptors, i.e. the estrogen (ER) and the progesterone receptors (PgR) as quantified by immunohistochemistry (IHC), and the expression (IHC) or the gene amplification (fluorescence In situ hybridization, FISH) status of the HER2 oncogene. These clinical variables can provide prognostic information and are summarized in clinical guidelines, such as the National Institute of Health (Eifel et al., 2001) in the USA or the St. Gallen consensus criteria (Goldhirsch et al., 2003) in Europe in order to assist clinicians and patients in adjuvant therapy decision-making. Histological grade (Scarff and Torloni, 1968) is a well-known histo-pathological parameter routinely used in the clinic to measure tumor differentiation, i.e. how much tumor cells look like the normal tissue from which they originated. Histological grade is known to be highly prognostic in breast cancer (Elston and Ellis, 1991). Patients having a histological grade 1 tumor exhibit better survival than patients having a histological grade 3 tumor. So, the use of histological grade is not sufficient to predict precisely the clinical outcome of a breast cancer patient. To reduce uncertainty in prognosis, these clinical variables can also be combined into multivariable outcome prediction models, like Adjuvant! Online (Olivotto et al., 2005) and the Nottingham Prognostic Index (Galea et al., 1992). These tools use age, nodal status, tumor size, histological grade, and ER status as clinical variables to estimate the risk of recurrence of breast cancer patients. However, risk estimation based on these guidelines or prognostic models is far from perfect and much progress is needed before it will be possible to clearly identify those patients, especially with early (node-negative, i.e. nodal status equal to 0) breast cancer, who would really need adjuvant systemic therapy (Isaacs et al., 2001 ; Sotiriou and Piccart, 2007). As a result, many women are prescribed adjuvant chemotherapy that probably would have had excellent long-term outcomes without it, exposing them to the potential adverse effects of chemotherapy such as cardiac dysfunction, second malignancies and premature menopause. Therefore, better prognostic tools could avoid the adverse side effects of adjuvant therapies, as well as the high costs of such treatments. During the last two decades, several clinical and pathological parameters have been used to evaluate the prognosis of breast cancer patients, but it still remains a challenge to distinguish those patients who would really need adjuvant systemic therapy from those who could be spared such treatment. Clinical investigators rapidly harnessed the great potential of gene expression profiling, not only for gaining new insights into cancer biology, but also as a powerful prognostic tool. Unlike the traditional clinical variables routinely measured in the clinic, which are limited to few, sometimes subjective, measurements, this technology enables the quantitative measurement of thousands of gene expressions in parallel, making possible the development of prognostic models with numerous molecular markers. In order to develop a more accurate tool for early breast cancer prognosis, the Netherlands Cancer Institute (NKI) conducted a comprehensive, genome-wide assessment of gene expression profiling (van 't Veer et al., 2002). They identified the genes differentially expressed between two groups of patients that differ in their survival. The low-risk group included patients who had not developed distant metastases within the first five years after diagnosis, a result that contrasted with the high-risk group. The NKI group refined the set of relevant genes and built a risk prediction model with 70 prognostic genes (denoted by GENE70). This set of genes (*gene signature*) included mainly genes involved in the cell cycle, invasion, metastasis, angiogenesis and signal transduction. This gene signature was then validated on a larger set of patients, including both node-negative and node-positive breast tumors in treated and untreated patients from the same institution(van 't Veer et al., 2002), and consequently proved to be predictor for distant metastasis-free survival, independently of several clinical prognostic indicators described above. To assess the clinical relevance of the GENE70 signature, the authors compared its performance to the National Institute of Health (NIH) consensus and the St Gallen guidelines. The NKI group found that the GENE70 signature, compared to the NIH and St Gallen classifications, was better at predicting which patients should have been spared adjuvant chemotherapy (low risk) and which patients should have been prescribed adjuvant chemotherapy (high-risk). The authors concluded that the GENE70 signature could outperform current clinical risk classifications and therefore could significantly impact on breast cancer management by sparing some women from over-treatment and the unnecessary toxicity of chemotherapy. Using a similar approach, Erasmus Medical Center and Veridex identified a prognostic gene signature (denoted by GENE76) that could be used to predict the development of distant metastases within the first five years after diagnosis in early (node-negative) breast cancer patients who did not receive systemic treatment (Wang et al., 2005). In contrast to van't Veer, this study considered ER-positive patients separately from ER-negative patients. This decision was based on the assumption that the mechanisms for disease progression could differ for these two ER-based subgroups of breast cancer patients. Similarly to the GENE70 signature, when compared to the classification results of St. Gallen and NIH, the GENE76 signature better identified the low-risk patients not needing treatment. By using gene expression profiling to develop gene signatures that are advantageous when compared to clinical guidelines, we could therefore significantly reduce the number of patients subject to unnecessary treatment. This would ultimately also translate into savings in cost and health resources, without sacrificing long-term clinical outcome. However, a careful validation of the gene expression profiling technology and prognostic gene signatures is required before bringing this predictive tool into day-to-day clinical practice. The use of systemic adjuvant treatments has increased in the last twenty years, with the objective of prolonging the survival of breast cancer patients. New treatments are continually being developed in order to target specifically the cancer cells and to reduce toxicity for the individual. The goal of prediction is to predict the response of a breast cancer patient to a treatment. There exist two settings for breast cancer prediction: the adjuvant and the neo-adjuvant settings (Mauri et al., 2005). The adjuvant setting is similar to the prognostication, except that the patients are prescribed a therapy. In the neo-adjuvant setting, the situation is more complex. First, a biopsy of the breast tumor is taken at diagnosis, before the neo-adjuvant therapy. Second, breast surgery is carried out to remove the tumor and to assess whether the tumor was affected by the treatment (e.g. decrease in tumor size). A pathological complete response (pCR) is then defined as the complete disappearance of tumor cells in the breast and the axillary lymph nodes and it has been shown that a pathological complete response is associated with excellent long-term survival. In this case, only the response or the resistance to the treatment is analyzed, leaving aside the issue of the survival of the patients. Currently, there exist few tools for prediction. For instance, the expression status of the hormonal receptors (ER and PR) and the expression/gene amplification status of the HER2 oncogene are used to define the subset of individuals who may benefit from hormono- and chemo- therapy, respectively. Despite the existence of the tools described above, current prediction models need to be improved, since the accuracy of these tools is poor (Sotiriou and Piccart, 2007; Lønning et al., 2007). Numerous attempts have been made to identify prognostic groups based on other pathological characteristics, mainly lymph vascular invasion or proliferation markers such as S-phase fraction, which might better reflect tumor biology and serve as prognostic and/or predictive markers that may aid in treatment decision making in the adjuvant setting (Colozza et al., 2005). In addition, a variety of molecular tumor markers have been studied both in the laboratory and in the clinical settings for their ability to predict response to treatment, but unfortunately, the studies examining the clinical utility of these tumor markers have usually used small, heterogeneous, retrospective patient series, often with insufficient power to draw robust conclusions; moreover, they have not been reported in a detailed enough fashion to provide information for the reproduction and external validation of results (McShane et al., 2005). There is also a lack of well-designed, prospective clinical trials addressing the clinical utility of such markers. Given the complexity of breast cancer and the huge diversity in molecular pathways dissected by basic research scientists (Konecny et al., 2004), isolated markers might not be sufficient to predict response or resistance to treatment, and a comprehensive view of the disease is needed. These limitations have driven breast cancer research to develop more accurate molecular predictors of clinical outcome and response to various anti-cancer therapies using a multi-marker approach with the help of the quantitative gene expression profiling technologies.

1.3 Contribution

The introduction of gene-expression tests have ushered in a new era in which many conventional clinical markers and predictors may be seen merely as surrogates for more fundamental genetic and physiologic processes. However, the multidimensional nature of these predictors demands both large numbers of clinically homogeneous patients to the used in the validation process, and exceptional rigor and discipline. Every study provides an opportunity to tweak a genetic signature, but the development of scientifically robust and clinically reliable tools require study designs and computational procedures. If gene-expression signatures are to reach the clinical setting, several outstanding issues will need to be addressed. First, researchers in this area will now need to turn their attention to methods of sample acquisition and the effect these methods have on the prognostic and predictive power of microarray data. Secondly, standardization of protocols and platforms for the measurement of gene-expression signatures in a robust and reproducible manner will have to be adopted. Thirdly, prior to commercialization of these signatures, a significant amount of validation will be required. Lastly, statistically powered studies with large, independent patient cohorts will be a prerequisite for acceptance. The research activity illustrated in this thesis aimed at fulfilling these methodological gaps that still hamper the identification of prognostic and predictive markers and affecting their reliability and reproducibility. Specifically, we addressed aspects related to i) the sample size of analyzed studies (*dataset*) and ii) the computational approaches applied in the discovery process. We developed a bioinformatics strategy to i) integrate multiple, independently generated datasets of tumor specimens with well-annotated clinical data, ii) to exploit this large-scale genomic data, in a retrospective behavior, for elucidating mechanisms of cancer progression and iii) to derive *gene signatures* as models for predicting neo-adjuvant chemotherapy sensitivity or resistance. This approach was tested and applied to breast cancer. These computational methods contribute fulfilling gaps in the bioinformatics analysis of microarray data where probe selection, annotation and specificity, comparability of different microarray platforms and signal normalization strategies, still represent a major, and partially unresolved, computational issue when analyzing multiple gene expression datasets. The problem of the limited sample size characterizing most functional genomics studies was addressed taking advantage of the increasing number of microarray datasets being deposited in public domains as Gene Expression Omnibus (GEO). Real opportunity exists for more reliable information to be generated through the integration of multiple, independently generated data focusing on the same tumor type, i.e., through *meta-analysis*. Meta-analysis strategies can be divided into two broad classes: data integration and data combination. In meta-analyses based on data integration each dataset is analyzed by itself and then results are combined with statistical techniques. Instead, data combination requires an ad-hoc normalization step of the raw data files and is applicable only when the expression profiles have been obtained using the same array technology (e.g. Affymetrix, Agilent, Illumina, etc.). Despites numerous efforts, mining and analyzing publicly available microarray data still represents a bioinformatics challenge and the lack of appropriate tools able to overcome critical issues, as annotation, cross-platform comparison and handling of metadata, is still hampering the potentialities of large-scale meta-analyses. Performing a meta-analysis of independent microarray studies requires to carefully handling the heterogeneity of array designs, which complicates cross-platform integration, and of sample descriptions, which impacts the correct characterization of specimens. At least for the case of Affymetrix arrays, cross-platform comparison has partially been solved by the adoption of custom Chip Definition Files (custom-CDF) which, linking probe sequences to annotated entities as genes or transcripts, allow matching expression profiles across subsequent generations of microarrays (Gautier et al., 2004 ; Dai et al., 2005; Ferrari et al., 2007). In custom CDFs, probes matching the same transcript, but belonging to different probes sets, are aggregated into putative custom-probe sets, each one including only those probes with a unique and exclusive correspondence with a single transcript. Similarly, probes matching the same transcript but located at different coordinates on different type of arrays may be merged in custom-probe sets and arranged in a virtual platform grid. As for any other microarray geometry, this virtual grid may be used as a reference to create i) the virtual-CDF file, containing the probes, shared among the platforms of interest, and their coordinates on the virtual platform, and ii) the virtual-CEL files containing the intensity data of the original CEL files properly re-mapped on the virtual grid. Once defined the virtual platform through the creation of its custom-CDF and transformed the CEL files into virtual-CELs, raw data, originally obtained from different platform, are homogeneous in terms of platform and can be preprocessed and normalized adopting standard approaches, as RMA (Robust Multiarray Analysis; Irizarry et al., 2003). Instead, retrieval, organization and utilization of meta-information is still an extremely critical step which affects the correct match between raw data files and sample IDs and the organization of samples into meaningful, homogeneous groups. This task is further complicated by the fact that i) datasets may be incompletely annotated, ii) the relationship between specimen, biological sample, phenotypic characteristics and raw data files, the most granular object in repositories, may be not sufficiently explicit, and iii) the procedures for managing large numbers of data files and related meta-information are tedious and error prone (Ioannidis et al., 2009).

Considering as model breast cancer, we collected 27 datasets, all hybridized on the same Affymetrix platform and with available raw data. Thus we used a combination approach that involved not only the expression signals but also sample meta-information. As a result, we constructed a meta-dataset comprising 3661 unique breast cancer samples with associated detailed clinical and outcome information and response to neo-adjuvant chemotherapy that allowed a statistically robust investigation of cancer subpopulations. In fact, microarray-based gene expression profiling allows the stratification of breast cancers into molecularly and clinically different subtypes with distinct gene expression patterns based on the activity of specific signaling cascades. In basic and translational research, this technique has become a working model for breast cancer molecular classification and for the definition of effective predictive and prognostic tools. Both these issues are critical in Triple Negative Breast Cancer (TNBC), a particular molecular subtype also known as *basal-like*, which still lacks not only of prognostic and therapeutic options, but also of a solid understanding of the molecular mechanisms at the base of its metastatic proclivity. Moreover, selecting markers extracted from gene signatures with *biological insights*, rather than solely on the basis of gene expression and phenotypic data, without taking into account *a priori* biological knowledge, could dramatically improve the reliability and robustness of prognostic and predictive signatures. Prediction of response to therapy is a clinically relevant need to improve patient selection for drug administration. An option would be the use of predictive markers of response to distinguish patients who are likely to receive benefits from those who are not, thus sparing predicted poor responders from the significant associated toxicities. Unfortunately, although this is an attractive strategy, suitable biomarkers predicting response to specific chemotherapy agents have, on the whole, remained elusive. Recently, it has been suggested that a single biomarker may not be sufficient for predicting anthracycline response, rather that a multifactorial approach might be better. Based on this, we exploited genetic data and clinical characteristics and responses of the retrospective cohort study to expand on the concept of multifactorial scoring and to derive *Consensus Signatures* as models for predicting neo-adjuvant chemotherapy sensitivity or resistance in triple negative breast cancer (TNBC). We designed *Consensus Signatures* as linear weighted combinations of gene signatures that, singularly, recapitulate independent signaling pathways (e.g., YAP/TAZ, mutp53/p63) or specific molecular mechanisms (i.e., hypoxia, immune function, induction of apoptosis), while intertwined together render a more comprehensive molecular model of chemo-resistance. As such, combinations of gene signatures resulted in a substantial improvement of the power to predict response to therapy over the use of single signatures.

Finally, in collaboration with the group headed by Giannino Del Sal at the University of Trieste, we investigated the breast cancer meta-dataset to gain new insights into the molecular bases of breast cancer stem cell (CSC) malignant properties which are implicated in both treatment resistance and disease relapse (Rustighi et al., 2014).

Chapter 2

Materials and methods

This chapter contains a description of breast cancer studies (datasets), after searching public databases (Gene Expression Omnibus (GEO), ArrayExpress), published in peer-reviewed journals analyzing gene expression profiling data from tumor tissues or biopsies from patients. Each dataset is fully reviewed and detailed. Following paragraphs describe how different datasets were combined together, the subtype molecular classification models and the methods used to derive predictive signatures. Finally, theoretical aspects of survival and statistical analyses applied for the meta-analysis of gene expression data are presented.

2.1 Breast cancer datasets

Public repositories have been inspected to retrieve gene expression data from tissue of breast cancer patients that were produced using Affymetrix technology and for which clinical annotations were publicly available.

The huge amount of gene expression profiles produced using microarray technology induced the creation of public repositories where storing and make publicly available to the scientific community this ocean of genomic data. Gene expression profiling, obtained during experiments designed to study a particular biological pathway, contains indeed a wealth of information not necessarily used in the original study and therefore available to other researchers for validating and confirming biological hypotheses. To enable the storage and exchange of gene expression data produced by high-throughput technologies, in 2001 a standard for describing all the information characterizing an experiment has been developed.

This standard is called MIAME (Minimum Information About a Microarray Experiment; Brazma et al., 2001) and it provides guidelines for the storage of data produced with microarrays and specify all the information that has to complement a genome-wide gene expression experiment. According to MIAME standards, public databases have been created with the purpose of maintaining, coordinating and distributing data from experiments involving the microarray technology. The major repositories of gene expression profiles are:

- Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) at the National Center for Biotechnology Information (NCBI, Bethesda, MD, USA);
- ArrayExpress at the European Bioinformatics Institute (http://www.ebi.ac.uk/microarray-as/ae/);
- caArray (Cancer Array Informatics Project), a database dedicated to the study of gene expression profiles in tumor cells and developed by the NCI Center for Bioinformatics https://cabig.nci.nih.gov/tools/caArray);
- Stanford Microarray Database (SMD; http://smd.stanford.edu/), collecting data mostly derived from spotted microarrays.

All these databases have been inspected to retrieve gene expression data from tissue of non-hereditary breast cancer patients that were produced using Affymetrix arrays and annotated with information on the clinical outcome. This survey returned a total of 4640 samples referring to 27 major studies listed in Table 2.1 and described in the following paragraphs.

Additional datasets produced using different type of microarrays are described in paragraph 2.3 and was used as a validation set.

Table 2.1: Breast cancer datasets analyzed in this thesis.

Study	Affymetrix platform	Samples	Data source	References
Stockholm	HG-U133A	159	GSE1456	Pawitan et al., 2005
EMC-286	HG-U133A	286	GSE2034	Wang et al., 2005
EMC-58	HG-U133A	58	GSE5327	Minn et al., 2007
MSK	HG-U133A	82	GSE2603	Minn et al., 2005
Uppsala-Miller	HG-U133A	236	GSE3494	Miller et al., 2005
Ivshina-Miller	HG-U133A	249	GSE4922	Ivshina et al., 2006
Loi	HG-U133A HG-U133 Plus 2.0	414	GSE6532	Loi et al., 2007; Loi et al., 2008; Loi et al., 2010
Sotiriou	HG-U133A	187	GSE2990	Sotiriou et al., 2006
Tamoxifen	HG-U133 Plus 2.0	77	GSE9195	Loi et al., 2008; Loi et al., 2010
Desmedt	HG-U133A	198	GSE7390	Desmedt et al., 2007
Schmidt	HG-U133A	200	GSE11121	Schmidt et al., 2008

Veridex	HG-U133A	136	GSE12093	Zhang et al., 2009
Chin	HG-U133AAofAV2	129	E-TABM-158	Merritt et al., 2008
Zhou	HG-U133AAofAV2	54	GSE7378	Zhou T et al., 2007; Yau C et al., 2008
TOP trial	HG-U133 Plus2.0	120	GSE16446	Desmedt Cet al., 2011; Li Y et al., 2010;Juul N et al., 2010
GSE19615	HG-U133 Plus2.0	115	GSE19615	Li Y et al., 2010
IPC	HG-U133 Plus2.0	266	GSE21653	Sabatier R et al., 2011
KFSYSCC	HG-U133 Plus2.0	327	GSE20685	Kao KJ et al., 2011
GSE31519	HG-U133 Plus2.0	67	GSE31519	Rody A. et al., 2011; Karn T. et al., 2011; Karn T. et al., 2012
GSE22093	HG-U133A	103	GSE22093	Iwamoto T et al., 2011
Hatzis	HG-U133A	508	GSE25066	Hatzis C. et al., 2011
GSE23988	HG-U133A	61	GSE23988	Iwamoto T et al., 2011
GSE20271	HG-U133A	178	GSE20271	Tabchy A. et al., 2010
GSE20194	HG-U133A	230	GSE20194	Popovici V. at al., 2010; Shi L. et al., 2010
Miyake	HG-U133 Plus2.0	115	GSE32646	Miyake T et al., 2012
GSE18728	HG-U133 Plus2.0	24	GSE18728	Lin Y et al., 2010
GSE19697	HG-U133 Plus2.0	61	GSE19697	Korde LA et al., 2010

2.1.1 Stockholm

The Stockholm dataset derives from the analysis of 524 breast cancer patients that have been operated at the Karolinska Hospital from January 1 1994 to December 31 1996 and identified from the population-based Stockholm–Gotland breast cancer registry established in 1976 (Pawitan et al., 2005; Table 2.2). Available tumor material was frozen on dry ice or in liquid nitrogen and stored in -70°C freezers. Out of the 524 tumors, 231 samples were excluded from gene expression profiling because of insufficient quantity of frozen tissue, 89 for various technical reasons (as degraded tumors and insufficient amount of RNA), 7 because the tissue was from patients living abroad and 6 because the patient refused participation in the study. Of the remaining 191 samples, 159 were profiled using Affymetrix HG-U133A and HG-U133B platforms, 17 were excluded because from neo-adjuvant therapy-treated patients, and RNA from the remaining 14 samples was hybridized on HG-U95 arrays. The remaining 159 tumors were divided in two groups according the different treatment: 126 patients received systemic adjuvant therapy and 33 no systemic adjuvant therapy. In the first group, 104 patients received Tamoxifen and its combinations.

Table 2.2: Characteristics of patients operated for breast cancer at the Karolinska Hospital 1994–1996 and considered in the Stockholm dataset (Pawitan et al., 2005).

Patient categories	**All patients (n=524)**	**No available tissue (n=231)**	**Excluded for other reasons (n=134)**	**Included for analysis (n=159)**

Mean age at breast cancer diagnosis (years)	58	57	58	58
Mean tumor size (mm)	20	16	24	22
Proportion of patients with tumor size <21 mm (%)	68	77	57	62
Proportion of patients with positive lymph nodes (%)	26	16	32	38
Proportion deceased (%)	20	12	26	24

Tumors sections from the primary tumors from patients with array profiles were classified using the Elston–Ellis grading (Elston and Ellis, 1991). These tumors were characterized with ER, PR and HER2 status. In the adjuvant treatment Tamoxifen and/or goserelin is normally used for hormonal treatment, but mostly intravenous cyclophosphamide, methotrexate and 5-fluorouracil (CMF) on days 1 and 8 was used as adjuvant chemotherapy. After primary therapy, patients were recommended to have regular clinical examinations and yearly mammograms, in addition to laboratory and X-ray tests guided by clinical signs and symptoms. Patients were normally followed for 5 years. There was no loss to follow-up. The relapse site, date of relapse, relapse therapy and date of death were ascertained in May 2002. The average follow-up was 6.1 years. Cause of death has been coded as death due to breast cancer (DEATH_BC; 1 = dead from breast cancer, 0 = alive or censored), including those with distant metastases or for other related causes, death due to other malignancies and nonmalignant disorders (DEATH; 1 = dead, 0 = alive or censored). During the follow-up patients developing breast cancer relapse were coded as RELAPSE (1 = relapse, 0 = no relapse or censored). Data are publicly available in the form of raw files at Gene Expression Omnibus GSE1456 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1456).

2.1.2 EMC-286

The Erasmus Medical Centre (EMC) dataset includes 286 samples derived from the Rotterdam tissue bank (Wang et al., 2005). These tumor samples come from patients with lymph node-negative breast cancer who were treated during 1980-95, but who did not receive any systemic neoadjuvant or adjuvant therapy. Tumors samples have been submitted to EMC from 25 regional hospitals for measurements of steroid-hormone receptors and a total of 436 samples of invasive tumors have been processed. Patients with poor, intermediate and good clinical outcome have been included. Out of the original 436 samples, 150 have been rejected on the basis of insufficient tumor content (53), poor RNA quality (77) and poor chip quality (20); thus, 286 samples were considered eligible for gene expression analysis. Clinical and pathological features of the EMC patients are summarized in Table 2.3. The median age of the patients at surgery was 52 years (range 26-83); 219 had undergone breast-conserving surgery and 67 modified radical mastectomy. Radiotherapy was given to 248 patients (87%) according to the institutional protocol. All involved patients were lymph-node-negative, based on pathological examination by regional pathologists. Amounts of estrogen receptors (ER) and progesterone receptors (PR) have been measured by ligand-binding assay, enzyme immunoassay or immunohistochemistry. The post-operative follow-up involved examinations every 6 months for 2 years, every 6 months for 3-5 years and every 12 months for 5 years. The date of metastasis was defined as the time of metastasis confirmation after symptoms reported by the patients, detection of clinical signs or at regular follow-up. Tumor samples have been hybridized to the Affymetrix oligonucleotide microarray U133A GeneChip. The median follow-up for the 198 patients who survived was 101 months (range 20-171). Of the 286 patients included, 93 (33%) showed evidence of distant metastasis within 5 years. Data are publicly available in the form of raw files at Gene Expression Omnibus GSE2034 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2034).

Table 2.3: Characteristics of patients from the EMC-286 dataset.

Clinical variables	Patients n=286
Age, years	
<40	29 (10%)
40-60	159 (56%)
>60	87 (30%)
Menopausal status	
Pre	139 (49%)
Post	147 (51%)
T stage	
T1	146 (51%)
T2	132 (46%)
T3/4	8 (3%)
Grade	
Poor	148 (52%)
Moderate	42 (15%)
Good	7 (2%)
Unknown	89 (31%)
ER status	
Positive	209 (73%)
Negative	77 (27%)
PR status	
Positive	166 (58%)
Negative	107 (37%)
LN status	
Positive	0
Negative	286 (100%)
Metastasis within 5 years	
Yes	93 (33%)
No	183 (64%)
Type of metastasis	
Brain	10 (3%)
Therapy	
Radio	248 (87%)
Surgery	

Breast conserving	219 (77%)
Radical mastectomy	67 (23%)

2.1.3 EMC-58

The EMC-58 cohort consists of 58 estrogen receptor-negative samples available at GEO GSE5327 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5327). Samples, obtained from the Erasmus Medical Center, were hybridized on HG-U133A Affymetrix arrays and, together with the 286 samples previously described (EMC-286), constitute the so-called EMC-344 dataset described in Minn et al., 2007. Patients' characteristics are reported in Table 2.4.

Table 2.4: Characteristics of patients from the EMC-58 dataset.

Clinical variables	Patients n=58
Age, years	
<40	13 (22%)
40-60	32 (55%)
>60	13 (22%)
T stage	
T1	21 (36%)
T2	33 (57%)
T3/4	4 (7%)
ER status	
Positive	0
Negative	58 (100%)
PR status	
Positive	15 (26%)
Negative	41 (71%)
Unknown	2 (3%)
Metastasis within 5 years	
Yes	10 (17%)
No	1 (2%)
Type of metastasis	
Lung	7 (12%)
Other	4 (7%)

2.1.4 MSK

This study includes 121 samples of which 99 are derived from primary breast cancers surgically resected at the Memorial Sloan-Kettering Cancer Center (MSKCC; Minn et al., 2005). Clinical information as age at diagnosis, tumor size (cm), Lymph Nodes status (LN), ER status, PR status and Her2 status are available for 82 patients. Moreover, the samples are annotated in terms of clinical outcome as MFS (metastasis free survival), LMFS (lung metastasis free survival) and BMFS (bone metastasis free survival), defined as the interval between the date of breast surgery and the date of diagnosed all metastasis, lung metastasis and bone metastasis of breast cancer, respectively (Table 2.5). The MSK dataset is available at GEO GSE2603 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2603).

Table 2.5: Characteristics of patients from the MSK dataset.

Clinical variables	Patients n=82
Age, years	
<40	8 (10%)
40-60	44 (54%)
>60	30 (37%)
T stage	
T1	12 (15%)
T2	58 (71%)
T3/4	12 (15%)
ER status	
Positive	46 (56%)
Negative	36 (44%)
PR status	
Positive	46 (56%)
Negative	36 (44%)
HER2 status	
Positive	58 (71%)
Negative	18 (22%)
Unknown	6 (7%)
LN status	
Positive	54 (66%)
Negative	28 (34%)
Metastasis within 5 years	
Yes	22 (27%)
No	5 (6%)
Type of metastasis	
Lung	14 (17%)
Bone	14 (17%)
Brain	NA
Therapy	
Chemo	69 (84%)
Hormonal	53 (65%)

2.1.5 Uppsala-Miller

The original patient material are freshly frozen breast cancers from a population-based cohort of 315 women that represented 65% of all breast cancers operated in Uppsala County during the time period from January 1, 1987 to December 31, 1989. In the Uppsala-Miller dataset, frozen tumor tissues are available in 293 out of original 315 patients (Miller et al., 2005). Of these, 251 had RNA of sufficient quantity and quality for microarray experiments passed Affymetrix quality controls. Survival data were based on the information from the Swedish population registry and cause of death was obtained from a review of the patient records last completed in 1999. Outcome information was available for 236 tumor samples whose clinical characteristics are reported in Table 2.6. Among the 251 tumors included in the present study, 58 had p53 mutations found by cDNA sequence analysis of exons 2-11 of the p53 gene. Clinic pathological variables were derived from the patient records and from routine clinical measurements at the time of diagnosis. Estrogen

(ER) and progesterone receptor (PR) status was determined by biochemical assay as part of the standard clinical procedure. An experienced pathologist determined the Elston-Ellis histological grade, classifying tumors into low-, medium-, and high-grade. Axillary lymph node status was positive in 78 patients. Nine patients had unknown node status, because no axillary examination was performed due to advanced age or concomitant serious disease. Systemic adjuvant therapy was offered to all node-positive patients. In general, premenopausal women were offered chemotherapy and postmenopausal women received endocrine treatment. The samples were hybridized on Affymetrix HG-U133A and B platforms. Data are publicly available at GEO GSE3494 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3494).

Table 2.6: Characteristics of the 236 patients from the Uppsala-Miller dataset.

Clinical variables	Patients n=236
Age, years	
<40	15 (6%)
40-60	79 (33%)
>60	140 (59%)
Size (mm)	
Median	20 (2-130)
Grade	
G1	62 (26%)
G2	121 (51%)
G3	51 (22%)
Unknown	2 (1%)
ER status	
Positive	201 (85%)
Negative	31 (13%)
Unknown	4 (2%)
PR status	
Positive	179 (76%)
Negative	57 (24%)
p53 status	
Mutation	55 (23%)
Normal	181 (77%)
LN status	
Positive	78 (33%)
Negative	149 (63%)
Unknown	9 (4%)
Survival time (years)	
Median	10.17 (0.25-12.75)
Disease specific-death	55 (23%)

2.1.6 Ivshina-Miller

The Ivshina-Miller dataset consists of 289 patients profiled on Affymetrix HG-U133A and B arrays (Ivshina et al., 2006). These samples were divided into two groups, depending on the site where samples were collected, i.e. the Uppsala cohort

(249 samples) and the Singapore cohort (40 samples). The Uppsala cohort originally was composed of 315 women representing 65% of all breast cancers resected in Uppsala County, Sweden, from January 1 1987 to December 31, 1989. For histological grading, new tumor sections were prepared from the original paraffin blocks and stained with eosin. All sections were graded in a blinded fashion according to the Nottingham Grading system. Estrogen receptors was assessed by Abbott's quantitative enzyme immunoassay and deemed positive if >0.05 fmol/μg DNA. After exclusions based on tissue availability, RNA amount, RNA integrity, clinical annotation, and microarray quality control, expression profiles of 249 and 40 tumors from the Uppsala and Singapore cohorts, respectively, were considered suitable for further analysis. Clinical outcomes were available only in Uppsala cohort. Disease free survival event (DFS) was defined as 0 if censored and as 1 in case of event defined as any type of recurrence (local, regional or distant) or death from breast cancer. In Table 2.7 are described all clinical and pathological features of the Uppsala cohort. Data are publicly available at GEO GSE4922 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4922).

Table 2.7: Characteristics of patients from the Uppsala cohort of the Ivshina-Miller dataset.

Clinical variables	Patients n=249
Age, years	
<40	16 (6%)
40-60	90 (36%)
>60	143 (57%)
Size (mm)	
Median	20 (2-130)
Grade	
G1	49 (20%)
G2	68 (27%)
G3	9 (4%)
ER status	
Positive	211 (85%)
Negative	34 (14%)
Unknown	4 (2%)
p53 status	
Mutation	58 (23%)
Normal	189 (76%)
Unknown	2 (1%)
LN status	
Positive	81 (33%)
Negative	159 (64%)
Unknown	9 (4%)
Metastasis within 5 years	
Yes	47 (19%)
No	16 (6%)

2.1.7 Loi

This dataset contains 414 samples obtained from the John Radcliffe Hospital (OXFT, Oxford, UK), the Guys Hospital (GUYT, London, UK) and the Uppsala University Hospital (KIT, Uppsala, Sweden). All samples had been hybridized using Affymetrix U133 arrays and specifically the OXFT and KIT samples have been analyzed using HG-U133A and HG-U133B arrays while the GUYT cohort was hybridized on HG-U133 Plus 2.0 chips (Loi et al., 2007; Loi et al., 2008; Loi et al., 2010). All samples were required to be estrogen (ER) and/or progesterone receptor (PR) positive by ligand-binding assay. The cut-off value for classification of patients as positive or negative for ER and PR was 10 fmol/mg of protein. OXFT and KIT groups are formed of 178 and 149 samples, respectively, while the GUYT group is composed of 87 samples. Table 2.8 reports all clinical and pathological characteristics of patients while gene expression data are publicly available at GEO GSE6532 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6532).

Table 2.8: Characteristics of patients from the Loi dataset.

Clinical variables	Patients n=414
Age, years	
<40	16 (4%)
40-60	177 (43%)
>60	208 (50%)
Size	
Median	2.1 (0-8.2)
Grade	
G1	82 (20%)
G2	182 (44%)
G3	76 (18%)
Unknown	74 (18%)
ER status	
Positive	349 (84%)
Negative	45 (11%)
Unknown	20 (5%)
PR status	
Positive	185 (45%)
Negative	32 (8%)
Unknown	192 (46%)
LN status	
Positive	143 (35%)
Negative	250 (60%)
Unknown	21 (5%)
Therapy	
Tamoxifen	277 (67%)
None	137 (33%)
Metastasis within 5 years	
Yes	96 (23%)
No	284 (69%)

2.1.8 Sotiriou

This dataset is divided in two subsets, i.e. KJX64 and KJ125, and consists of information obtained from a total of 189 patients with primary operable invasive breast cancer, whose frozen tumor specimens were archived at the John Radcliffe Hospital (Oxford, UK) and at the Uppsala University Hospital (Uppsala, Sweden). The set KJX64 contains data from 64 ER-positive primary breast tumor samples and the set KJ125 contains data from 125 breast tumor samples (Sotiriou et al., 2006). No patient in the KJ125 dataset received any adjuvant systemic therapy. Histological tumor grade was based on the Elston–Ellis grading system and determined from data extracted from the pathology reports and reviewed separately by one pathologist for the Oxford population and another pathologist for the Swedish population. A total of 187 samples are characterized by clinical outcome in terms of relapse free survival (RFS) defined as the interval between the date of breast surgery and the date of diagnosis of any type of relapse (local, regional or distant), while 179 have clinical information about distant metastasis free survival (DMFS) defined as the interval between the date of breast surgery and the date of diagnosed distant relapse of breast cancer. Microarray analysis was performed on Affymetrix U133A GeneChip. Table 2.9 reports all clinical and pathological characteristics of patients and gene expression data are publicly available at GEO GSE2990 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2990).

Table 2.9: Characteristics of patients from the Sotiriou dataset.

Clinical variables	Patients n=187
Age, years	
<40	14 (7%)
40-60	97 (52%)
>60	76 (41%)
Size (cm)	
≤2	104 (56%)
<2 ≤5	81 (43%)
>5	4 (2%)
Grade	
G1	64 (34%)
G2	48 (26%)
G3	55 (29%)
Unknown	20 (11%)
ER status	
Positive	147 (79%)
Negative	34 (18%)
Unknown	6 (3%)
LN status	
Positive	30 (16%)
Negative	153 (82%)
Unknown	4 (2%)
Metastasis within 5 years	
Yes	28 (15%)
No	35 (19%)
Therapy	
Tamoxifen	64 (34%)

2.1.9 Tamoxifen

This dataset contains 77 patients diagnosed at the Guy's Hospital (London, UK) with early stage breast cancer and treated with adjuvant Tamoxifen monotherapy (Loi et al., 2008; Loi et al., 2010). Samples were hybridized on Affymetrix HG-U133 Plus 2.0 microarrays according to standard Affymetrix protocols. Clinical characteristic such as age, tumor size, PR, ER, lymph node status and histological grade are in Table 2.10. The median follow-up is 12.5 years. Data are available at GSE9195 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9195).

Table 2.10: Characteristics of patients from the Tamoxifen dataset.

Clinical variables	Patients n=77
Age, years	
<40	0 (0%)
40-60	29 (38%)
>60	48 (62%)
Size (cm)	
Median	2.1 (1.09- 6)
Grade	
G1	14 (18%)
G2	20 (26%)
G3	24 (31%)
Unknown	19 (2%5)
ER status	
Positive	36 (47%)
Negative	41 (53%)
PR status	
Positive	36 (47%)
Negative	41 (53%)
LN status	
Positive	59 (%)
Negative	18 (%)
Metastasis within 5 years	
Yes	10 (13%)
No	0
Therapy	
Tamoxifen	77 (100%)

2.1.10 Desmedt

The TRANSBIG consortium constructed the Desmedt dataset to validate the 70-gene signature of the MammaPrint (van't Veer et al., 2002). The consortium analyzed, blinded to clinical data, the gene expression profiles of 198 lymph node negative, systemically untreated patients, of the Bordet Institute. These patients were younger than the age of 61 years (median age 47 years) and affected by lymph node-negative, T1-T2 (≤5 cm) tumors. Patients in this series have been diagnosed between 1980 and 1998 (median follow-up, 13.6 years) in 6 different centers, i.e., Institut Gustave Roussy, (IGR, Villejuif, France), Karolinska Institute and Uppsala

University Hospital, (KI, Stockholm and Uppsala, Sweden), Centre René Huguenin (CRH, Saint-Cloud, France); Guy's Hospital (GH, London, UK), and John Radcliffe Hospital ((JRH, Oxford, UK). Patients with previous malignancies (except basal cell carcinoma) and bilateral synchronous breast tumors were excluded. ER status (by immunohistochemistry) and histological grade (by the Elston and Ellis method) were determined by the same pathologist, blinded to the clinical and genomic data, from the corresponding paraffin-embedded tumor samples at the Department of Pathology at the European Institute of Oncology (Milan, Italy). Tumor characteristics (age, tumor size and grade, ER status, and proportion of patients alive at 10 years) are all available. The median follow-up is 14.0 years and distant metastases are found in 51 (26%) patients, with 35 patients showing progression within 5 years (18%). Time from diagnosis to distant metastases (TDM) and overall survival (OS), defined as time from diagnosis to death from any cause, are the clinical outcomes. Table 2.11 contains the clinical characteristics. Data are available at GSE7390 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7390).

Table 2.11: Characteristics of patients from the Desmedt dataset.

Clinical variables	Patients n=198
Age, years	
<40	0 (%)
40-60	162 (82%)
>60	36 (18%)
Size (cm)	
Median	2 (0.6-5)
Grade	
G1	30 (15%)
G2	83 (42%)
G3	83 (42%)
Unknown	2 (1%)
ER status	
Positive	134 (68%)
Negative	64 (32%)
LN status	100
Positive	
Negative	198 (%)
Metastasis within 5 years	
Yes	36 (18%)
No	8 (4%)

2.1.11 Schmidt

The Schmidt study consists of 200 lymph node-negative breast cancer patients treated at the Department of Obstetrics and Gynecology of the Johannes Gutenberg University in Mainz between 1988 and 1998 (Schmidt et al., 2008). Patients were all treated with surgery and did not receive any systemic therapy in the adjuvant setting. The established prognostic factors as histological grade, tumor size, age at diagnosis, and steroid receptor status were collected from the original pathology reports of the gynecologic pathology division. 75 patients were treated with modified radical mastectomy and 125, without evidence of regional lymph node and distant metastasis at the time of surgery, with breast-conserving surgery followed by irradiation. The median age of the patients at surgery was 60 years (range, 34–89 years) and the median time of follow up was 92 months (Table 2.12). For all tumors, samples were snap frozen and stored at 80°C and RNA hybridized on Affymetrix HG-U133A arrays. Data are available at Gene Expression Omnibus GSE11121 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11121).

Table 2.12: Characteristics of patients from the Schmidt dataset.

Clinical variables	Patients n=200
Size (cm)	
Median	2.1 (0.1-6)
Grade	
G1	29 (15%)
G2	136 (68%)
G3	35 (18%)
LN status	
Positive	0
Negative	200 (100%)
Metastasis within 5 ys	
Yes	28 (14%)
No	18 (9%)

2.1.12 Veridex

This dataset has been collected to validate a 76-gene signature identified by Veridex (a Johnson & Johnson Company) to predict high-risk patients that benefit from adjuvant Tamoxifen therapy (Zhang et al., 2009). After defining the signature in an independent cohort of untreated patients, Veridex selected frozen tumor specimens of patients treated with adjuvant Tamoxifen (n=136) from the tumor banks in three of the four European institutions that provided the samples of patients without systemic therapy, i.e., Institute of Oncology of Ljubljana (36 samples), National Cancer Institute of Bari (28 samples) and Technical University of Munich (9 samples), and from one US institution, i.e., the Cleveland Clinic Foundation (63 samples; period 1981–2000). Routine postsurgical follow-up was similar among the participating institutions and involved examination every 3 months during the first 2 years, every 6 months for years 2–5, and annually after year 5 of the follow-up period. Date of diagnosis of metastasis was defined as the date of imaging or histological confirmation of metastasis after complaints and/or clinical symptoms, or at regular follow-up. The surviving patients (n=119) had a median follow-up time of 90 months (range 29–193 months). Twenty patients (15%) showed evidence of distant metastases with 12 (9%) having metastases within 5 years. A total of 17 patients died, with 6 dying without evidence of metastasis. These patients were censored at last follow-up in the analysis of distant metastasis free survival. RNA has been hybridized on Affymetrix HG-U133A arrays and the raw data submitted to GSE12093 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12093). Table 2.13 contains the clinical characteristics of the patients included in the Veridex study.

Table 2.13: Characteristics of patients from the Veridex dataset.

Clinical variables	Patients n=136
Age, years	
Median	64 (9%)
<40	4 (3%)
41-55	23 (17%)
56-70	80(59%)
>70	29 (21%)
T stage	
T1	63 (48%)
T2	65 (48%)
T3/T4	7 (5%)
Unknown	1 ((1%)
ER status	
Positive	136 (100%)
Negative	0
Grade	
Poor	30 (22%)
Moderate	43 (32%)
Good	8 (6%)
Unknown	55 (40%)
Therapy	
Hormonal	136 (100%)
Metastasis within 5 ys	
Positive	12 (9%)
Negative	124 (91%)

2.1.13 Chin

Frozen tissue from UC San Francisco and the California Pacific Medical Center collected between 1989 and 1997 was used for this study (Chin et al., 2006). Tissues were collected under IRB-approved protocols with patient consent. Tissues were collected, frozen over dry ice within 20 min of resection, and stored at −80°C. An H&E section of each tumor sample was reviewed, and the frozen block was manually trimmed to remove normal and necrotic tissue from the periphery. Clinical follow-up was available with a median time of 6.6 years. Tumors were predominantly early stage (83% stage I and II) with an average diameter of 2.6 cm. About half of the tumors were node positive, 67% were estrogen receptor positive, 60% rceived tamoxifen, and half received adjuvant chemotherapy (typically adriamycin and cytoxan). Samples were hybridized to Affymetrix HT-HG_U133A

(U133AAofAV2) GeneChip. Table 2.14 contains the clinical characteristics of the patients included in the E-TABM-154 study. The raw data for expression profiling are available at ArrayExpress repository with accession number E-TABM-158 (http://www.ebi.ac.uk/arrayexpress/experiments/E-TABM-158/).

Table 2.14: Characteristics of patients from the E-TABM-158 dataset.

Clinical variables	Patients n=129
Age, years	
<40	19 (15%)
40-60	70 (54%)
>60	39 (30%)
Unknown	1 (1%)
Size (cm)	
≤2	54 (42%)
>2 ≤5	66 (51%)
>5	7 (5%)
Grade	
G1	14 (11%)
G2	46 (36%)
G3	64 (50%)
Unknown	5 (4%)
ER status	
Positive	83 (64%)
Negative	46 (36%)
PR status	
Positive	73 (57%)
Negative	54 (42%)
Unknown	2 (2%)
HER2 status	
Positive	11 (9%)
Negative	78 (60%)
Unknown	40 (31%)
LN status	
Positive	71 (55%)
Negative	58 (45%)
Metastasis within 5 ys	
Yes	36 (28%)
No	25 (19%)
Therapy	
Hormonal	74 (57%)

2.2.14 Zhou

Cryobanked breast cancer specimens were obtained from the University of California San Francisco (UCSF) Comprehensive Cancer Center Breast Oncology Program Tissue Core, and collected under UCSF approved protocols following patient consent (Zhou et al., 2007; Yau et al., 2008). From an archive of over 1,000 liquid nitrogen frozen breast cancer specimens, 54 primary breast cancer samples (UCSF cases) had been identified, using the following criteria: early clinical stage

(T1/2, N0, M0) invasive breast cancer, ER-positive status (>10% nuclear immunohistochemical staining), known clinical outcome (relapse-free survival, RFS), and stratification into young (≤ 45 years, n = 29) or old (≥ 70 years, n = 25) age-at-diagnosis. Clinical characteristics are illustrated in Table 2.15. Total RNA was labeled and analyzed using Affymetrix HT-HG_U133A (U133AAofAV2) GeneChip. Raw data files have been entered into the NCBI Gene Expression Omnibus (GEO) repository with accession number GSE7378 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7378).

Table 2.15: Characteristics of patients from the Zhou dataset.

Clinical variables	Patients n= 54
Age, years	
<40	10 (19%)
40-60	19 (35%)
>60	25 (46%)
ER status	
Positive	54 (100%)
Negative	0 (0%)
LN status	
Positive	0 (0%)
Negative	54 (100%)
Metastasis within 5 ys	
Yes	6 (11%)
No	17 (31%)

2.2.15 TOP trial

The prospective multicentric TOP trial enrolled 149 patients between January 2003 and June 2008 (Desmedt et al., 2011). One patient was excluded because of concomitant contralateral breast cancer. Of these 148 patients, nine were excluded from further analysis, leading to a total of 139 evaluable patients. Epirubicin monotherapy (100 mg/m^2) was administered as neo adjuvant chemotherapy, with four cycles every 3 weeks for patients with early breast cancer and a dose-dense schedule of six cycles every 2 weeks for patients with locally advanced and inflammatory disease. All patients underwent pre-treatment biopsies of the primary breast tumor before starting chemotherapy. Pathological complete response (pCR) was defined as the absence of residual invasive breast carcinoma in the breast and in the axillary nodes after completion of chemotherapy. Persistence of in situ carcinoma without an invasive component was also considered pCR. Clinical characteristics are illustrated in Table 2.16. One hundred twenty samples' RNA have been hybridized to HG-U133 Plus 2.0 microarrays. Raw data files have been entered into the NCBI Gene Expression Omnibus (GEO) repository with accession number GSE16446 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16446).

Table 2.16: Characteristics of patients from the TOP trial dataset.

Clinical variables	Patients n=120
Age, years	
≤50	70 (58%)
>50	50 (42%)
Size (cm)	
T1	17 (14%)
T2	83 (69%)
T3	5 (4%)
T4	15 (13%)
Node status	
N0	55 (46%)
N1	60 (50%)
N2	3 (2%)
N3	2 (2%)
Grade	
G1	2 (2%)
G2	20 (17%)
G3	92 (76%)
Unknown	6 (5%)
ER status	
Positive	0 (0%)
Negative	120 (100%)
HER2 status	
Amplified	31 (26%)
Not amplified	62 (52%)
Unknown	27 (22%)
Metastasis within 5 ys	
Yes	23 (19%)
No	70 (58%)
Response neoadjuvant therapy	
pathological complete response (pCR)	16 (13%)
residual disease (RD)	98 (82%)
Unknown	6 (5%)

2.2.16 GSE19615

One hundred fifteen primary breast tumors were recruited from the US National Cancer Institute–Harvard Breast Specialized Program Of Research Excellence blood and tissue repository under protocols approved by the DF/HCC Institutional Review Board, with informed consent from subjects (Li et al., 2010). Table 2.17 contains the clinical characteristics of the patients included in the GSE16915. For all tumors RNA were hybridized on Affymetrix HG-U133 Plus 2.0 arrays. Raw data are deposited in the NCBI Gene Expression Omnibus (GEO) database under accession number GSE19615 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19615).

Table 2.17: Characteristics of patients from the GSE19615 dataset.

Clinical variables	Patients n= 115
Age, years	

<40	8 (7%)
40-60	80 (70%)
>60	27 (23%)
Size (cm)	
≤2	51 (44%)
<2 ≤5	62 (54%)
>5	2 (2%)
Grade	
G1	23 (20%)
G2	28 (24%)
G3	64 (56%)
ER status	
Positive	66 (57%)
Negative	45 (39%)
Unknown	4 (3%)
HER2 status	
Positive	30 (26%)
Negative	79 (69%)
Unknown	6 (5%)
LN status	
Positive	51 (44%)
Negative	62 (54%)
Unknown	2 (2%)
Metastasis within 5 ys	
Yes	14 (12%)
No	41 (36%)
Therapy	
Hormonal	64 (56%)

2.2.17 GSE21653

The IPC (Institut Paoli-Calmettes) series contained frozen tumor samples obtained from 266 early breast cancer patients who underwent initial surgery in our institution between 1992 and 2004 (Sabatier et al., 2011). The study was approved by the IPC review board, and informed consent was available for each case. Inclusion criteria included: pre-treatment sample of an invasive adenocarcinoma, non-inflammatory and non-metastatic, with available histoclinical data. All samples were similarly profiled using Affymetrix U133 Plus 2.0 human oligonucleotide DNA microarrays. Clinical characteristics are illustrated in Table 2.18. Raw data files have been entered into the NCBI Gene Expression Omnibus (GEO) repository with accession number GSE21653 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21653).

Table 2.18: Characteristics of patients from the GSE21653 dataset.

Clinical variables	Patients n= 266
Age, years	
<40	48 (18%)
40-60	121 (45%)

>60	96 (36%)
Unknown	1 (0.4%)
Size (cm)	
≤2	59 (22%)
<2 ≤5	126 (47%)
>5	69 (26%)
Unknown	13 (5%)
LN status	
Positive	140 (53%)
Negative	120 (45%)
Unknown	6 (2%)
Grade	
G1	45 (17%)
G2	89 (34%)
G3	125 (47%)
Unknown	7 (3%)
ER status	
Positive	150 (56%)
Negative	113 (43%)
Unknown	3 (1%)
HER2 status	
Positive	29 (11%)
Negative	216 (81%)
Unknown	21 (8%)
P53 status	
Mutant	69 (26%)
Wild-type	125 (47%)
Unknown	72 (27%)
Metastasis within 5 ys	
Yes	69 (26%)
No	78 (29%)

2.2.18 GSE20685

Fresh frozen breast cancer tissue from every patient diagnosed and treated between 1991 and 2004 at the Koo Foundation Sun-Yat-Sen Cancer Center (KFSYSCC) were randomly selected for the study (Kao et al., 2011). Patients with follow-up periods shorter than three years were excluded, with the exception of those who died of the disease within three years of the initial treatment. In cases of ineligibility, the following sample was selected. The selected tissue samples spanned the major transition periods of adjuvant chemotherapy from CMF (cyclophosphamide, methotrexate and fluorouracil) to CAF (cyclophosphamide, doxorubicin, fluorouracil) and to taxane-based regimens. Four hundred forty seven samples were obtained, but 135 samples were excluded due to insufficient RNA (n = 1), poor RNA quality (n = 116), or unacceptable microarray quality (n = 18). A total of 312 samples were eligible for the study plus an additional 15 lobular breast carcinoma samples, collected between 1999 and 2004, were also included. All patients were treated by a multidisciplinary team according to the guidelines consistent with the

National Comprehensive Cancer Network. Following modified radical mastectomy or breast-conserving surgery plus dissection of axillary nodes, patients received radiotherapy, adjuvant chemotherapy, and/or hormonal therapy, if indicated. Neoadjuvant chemotherapy was administered to patients with locally advanced disease. The study was approved by the institutional review board and ethical approval was obtained from the same board for samples without obtainable informed consent. Clinical characteristics are illustrated in Table 2.19. Samples' RNA have been hybridized to HG-U133 Plus 2.0 microarrays. Raw data files have been entered into the NCBI Gene Expression Omnibus (GEO) repository with accession number GSE20685 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20685).

Table 2.19: Characteristics of patients from the GSE20685 dataset.

Clinical variables	Patients n= 327
Age, years	
<40	71 (22%)
40-60	211 (65%)
>60	45 (14%)
Unknown	0 (0%)
ER status	
Positive	41 (13%)
Negative	37 (11%)
Unknown	249 (76%)
HER2 status	
Positive	0 (0%)
Negative	41 (13%)
Unknown	286 (87%)
Metastasis within 5 ys	
Yes	67 (20%)
No	15 (5%)

2.2.19 GSE31519

Tissue samples of invasive breast cancer cases were obtained with IRB approval and informed consent from consecutive patients undergoing surgical resection between December 1996 and July 2007 at the Department of Gynecology and Obstetrics at the Goethe-University in Frankfurt (Rody et al., 2011; Karn et al., 2011). Primary breast cancer biopsies were obtained from patients before treatment. Table 2.20 reports all clinical and pathological characteristics of patients. Gene expression data have been deposited into the GEO database with accession number GSE31519 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31519).

Table 2.20: Characteristics of patients from the GSE31519 dataset.

Clinical variables	Patients n= 67
Size (cm)	

≤1	16 (24%)
≥1	48 (72%)
Unknown	3 (4%)
Grade	
G1/G2	18 (27%)
G3	45 (67%)
Unknown	4 (6%)
ER status	
Positive	0 (0%)
Negative	67 (100%)
LN status	
Positive	21 (31%)
Negative	44 (66%)
Unknown	2 (3%)
Metastasis within 5 ys	
Yes	21 (31%)
No	32 (48%)
Biopsy type	
Core needle	19 (28%)
Surgical	48 (72%)

2.2.20 GSE22093

The chemotherapy sensitivity analysis was performed on the USO-02103 included 103 patients (42 ER-positive and 56 ER-negative patients) who received four courses of 5-fluorouracil (500 mg/m^2), eprirubicin (100 mg/m^2), and cyclophosphamide (500 mg/m^2), given once every 21 days, followed by 12 weeks of docetaxel (35 mg/m^2), given once weekly concomitant with capecitabine (850 mg/m^2 given twice daily for 14 days, repeated every 21 days) (FEC/wTX) (Iwamoto T et al., 2011). Clinical characteristics are illustrated in Table 2.21. Samples' RNA have been hybridized to HG-U133A microarrays. Raw data files have been entered into the NCBI Gene Expression Omnibus (GEO) repository with accession number GSE22093 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22093).

Table 2.21: Characteristics of patients from the GSE22093 dataset.

Clinical variables	Patients n= 103
Age, years	
<40	22 (21%)
40-60	59 (57%)
>60	16 (16%)
Unknown	6 (6%)
Size (cm)	
T0/T1	3 (3%)
T2	51 (50%)
T3	26 (25%)
T4	18 (17%)
Unknown	5 (5%)

Grade	
G1	3 (3%)
G2	29 (28%)
G3	47 (46%)
Unknown	24 (23%)
ER status	
Positive	42 (41%)
Negative	56 (54%)
Unknown	5 (5%)
P53 status	
Wild-type	42 (41%)
Mutant	58 (56%)
Unknown	3 (3%)
Neoadjuvant chemotherapy response	
pathological complete response (pCR)	28 (27%)
residual disease (RD)	69 (67%)
Unknown	6 (6%)
Biopsy type	
Fine needle	103 (100%)
Core needle	0 (0%)

2.2.21 GSE20271

Patients with clinical stage I to III breast cancer were eligible (Tabchy et al., 2010). Histologic diagnosis of invasive cancer and estrogen receptor (ER), progesterone receptor (PR), and HER2 receptor status were determined from a diagnostic core needle or incisional biopsy before therapy. All patients had to agree to a separate, pretreatment research fine-needle aspiration (FNA) of the cancer for gene expression analysis. Patients were accrued at six international sites including The University of Texas M.D. Anderson Cancer Center (MDACC; *n* = 96) and the Lyndon B Johnson General Hospital (*n* = 19) in Houston, Texas; the Instituto Nacional de Enfermedades Neoplasicas in Lima, Peru (*n* = 79); the Centro Medico Nacional de Occidente in Guadalajara, Mexico (*n* = 19); and the clinical trial group Grupo Español de Investigacion en Cancer de Mama in Spain (*n* = 60). This study was approved by the institutional review boards of each participating institution, and all patients signed an informed consent for voluntary participation. The study was conducted between October 2003 and October 2006. Two hundred and seventy-three patients were enrolled: 138 were randomized to T/FAC and 135 to FAC chemotherapy. Twenty (7%) and 16 (6%) patients were excluded from genomic response analysis in each treatment arm, respectively, due to eligibility violations including nonstudy treatment regimen, patient withdrawal, or lack of pathologic assessment of response. Of the 118 patients who received T/FAC, 9 patients progressed clinically, and these were considered as RD for response prediction analysis. Of the 119 patients who were assigned to receive FAC chemotherapy, 11 received T/FAC treatment (to maximize response or due to progression on FAC), and these cases were assigned to the T/FAC treatment group for genomic response prediction analysis. The remaining 108 cases, including 5 cases that progressed, comprised the FAC treatment cohort for the final response prediction analysis. Two hundred and four FNA samples (75%) yielded sufficient quality and quantity of RNA to do gene expression analysis. The main reasons for failure were acellular aspirates and low RNA yield; five profiles (2.5%) failed array QC after hybridization. After excluding the patients who had no response information available, 178 cases remained with complete pathologic response and genomic prediction results for final analysis. Of these, 91 received T/FAC and 87 received FAC chemotherapy. Clinical characteristics of these patients are presented in Table 2.22. Gene expression data have been deposited into the GEO database with accession number GSE20271 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20271).

Table 2.22: Characteristics of patients from the GSE20271 dataset.

Clinical variables	Patients n=178
Age, years	
<40	25 (14%)
40-60	114 (64%)
>60	39 (22%)
Size (cm)	
T0/T1	13 (7%)
T2	76 (42%)
T3	37 (21%)
T4	51 (29%)
Unknown	51 (1%)
Grade	
G1	15 (9%)
G2	61 (34%)
G3	72 (40%)
Unknown	30 (17%)
ER status (IHC)	
Positive	98 (55%)
Negative	80 (45%)
HER2 status (FISH, IHC)	
Not overexpressed	152 (85%)
Overexpressed	26 (15%)
Race	
White	81 (46%)
Black	13 (7%)
Hispanic	83 (46%)
Asian	1 (1%)
Neoadjuvant chemotherapy response	
pathological complete response (pCR)	26 (15%)
residual disease (RD)	152 (85%)

2.2.22 GSE20194

Gene expression data from microarrays are being applied to predict preclinical and clinical endpoints, but the reliability of these predictions has not been established. In the MAQC-II project (Popovici et al., 2010; Shi et al., 2010), 36 independent teams analyzed six microarray data sets to generate predictive models for classifying a sample with respect to one of 13 endpoints indicative of lung or liver toxicity in rodents, or of breast cancer, multiple myeloma or neuroblastoma in humans. The human breast cancer data set was contributed by the University of Texas M.D. Anderson Cancer Center. Gene expression data from 230 stage I–III breast cancers were generated from fine needle aspiration specimens of newly diagnosed breast cancers before any therapy. The biopsy specimens were collected sequentially during a prospective pharmacogenomic marker discovery study between 2000 and 2008. These specimens represent 70–90% pure neoplastic cells with minimal stromal contamination. Patients received 6 months of preoperative (neoadjuvant) chemotherapy including paclitaxel (Taxol), 5-fluorouracil, cyclophosphamide and doxorubicin (Adriamycin) followed by surgical resection of the cancer. Response to preoperative chemotherapy was categorized as a pathological complete response (pCR = no residual invasive cancer in the breast or lymph nodes) or residual invasive cancer (RD). RNA extraction and gene expression profiling were performed using Affymetrix HG-U133A microarrays. Clinical characteristics of patients are presented in Table 2.23. Gene expression data have been deposited into the GEO database with accession number GSE20194 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20194).

Table 2.23: Characteristics of patients from the GSE20194 dataset.

Clinical variables	Patients n=230
Age, years	
<40	27 (12%)
40-60	151 (66%)
>60	52 (22%)
Size (cm)	
T0/T1	23 (10%)
T2	132 (57%)
T3	34 (15%)
T4	41 (18%)
Grade	
G1	13 (6%)
G2	94 (41%)
G3	123 (53%)
ER status (IHC)	
Positive	141 (61%)
Negative	89 (39%)
HER2 status (FISH, IHC)	
Not overexpressed	190 (83%)
Overexpressed	40 (17%)
Race	
White	153 (66%)
Black	25 (11%)

Hispanic	34 (15%)
Asian	16 (7%)
Neoadjuvant chemotherapy response	
pathological complete response (pCR)	48 (21%)
residual disease (RD)	182 (79%)
Biopsy type	
Fine needle	230 (100%)

2.2.23 GSE25066

Five hundred eight patients prospectively provided written informed consent to participate in an institutional review board–approved research protocol (LAB99-402, USO-02-103, 2003-0321, I-SPY-1) to obtain a tumor biopsy sample by fine-needle aspiration or core biopsy prior to any systemic therapy for genomic studies to develop and test predictors of treatment outcome (Hatzis et al., 2011). Clinical nodal status was determined before treatment from physical examination, with or without axillary ultrasound, with diagnostic fine-needle aspiration as required. Pathologic ERBB2 status was defined as negative according to American Society of Clinical Oncology/College of American Pathologists guidelines. Patients with any nuclear immunostaining of estrogen receptor (ER) in the tumor cells were considered eligible for adjuvant endocrine therapy. In the discovery cohort, biopsy samples were obtained from June 2000 to December 2006; 227 were obtained by fine-needle aspiration (MDACC) and 83 by core biopsy (I-SPY), and all chemotherapy was administered as neo-adjuvant with taxane-anthracycline pre-operative chemotherapy treatment. In the validation cohort, biopsy samples were obtained from April 2002 to January 2009; 157 were obtained by fine-needle aspiration (MDACC, Peru, US Oncology) and 41 by core biopsy (MDACC, Lyndon B. Johnson Hospital, Spain), and all chemotherapy was administered as neo-adjuvant with taxane-anthracycline pre-operative chemotherapy treatment. Response was assessed at the end of neo-adjuvant treatment and distant-relapse-free survival was followed for at least 3 years post-surgery. Clinical characteristics are illustrated in Table 2.24. All gene expression microarrays were profiled in the Department of Pathology at the M. D. Anderson Cancer Center (MDACC), Houston, Texas. Samples' RNA have been hybridized to HG-U133A microarrays. Raw data files have been entered into the NCBI Gene Expression Omnibus (GEO) repository with accession number GSE25066 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25066).

Table 2.24: Characteristics of patients from the GSE25066 dataset.

Clinical variables	Patients n=508
Age, years	
<40	97 (19%)
40-60	307 (60%)
>60	104 (21%)

Size (cm)	
T0/T1	33 (6%)
T2	255 (50%)
T3	145 (29%)
T4	75 (15%)
Grade	
G1	32 (6%)
G2	180 (35%)
G3	258 (51%)
Unknown	38 (8%)
ER status (IHC)	
Positive	297 (59%)
Negative	205 (40%)
Unknown	6 (1%)
HER2 status	
Positive	6 (1%)
Negative	485 (96%)
Unknown	17 (3%)
Metastasis within 5 years	
Yes	110 (23%)
No	337 (67%)
Neoadjuvant chemotherapy response	
pathological complete response (pCR)	99 (19%)
residual disease (RD)	389 (77%)
Unknown	20 (4%)

2.2.24 GSE23988

This is Phase II Trial of four courses of 5-fluorouracil, doxorubicin and cyclophosphamide followed by four additional courses of weekly docetaxel and capecitabine administered as Preoperative Therapy for Patients with Locally Advanced Breast Cancer, Stages II and III by US oncology (PROTOCOL 02-103) (Iwamoto et al., 2011). Table 2.25 reports all clinical and pathological characteristics of patients. Pre-treatment FNA from primary tumors were obtained and RNA extracted and hybridized to Affymetrix HG-U133A microarrays. Raw data files have been entered into the NCBI Gene Expression Omnibus (GEO) repository with accession number GSE23988 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23988).

Table 2.25: Characteristics of patients from the GSE23988 dataset.

Clinical variables	Patients n=61
Age, years	
<40	8 (13%)
40-60	45 (74%)
>60	8 (13%)
Size (cm)	
≤2	1 (2%)
<2 ≤5	20 (33%)
>5	40 (65%)

Nodal status	
N0	21 (34%)
N1	32 (53%)
N2	5 (8%)
N3	3 (5%)
Grade	
G1	1 (2%)
G2	19 (31%)
G3	37 (60%)
Unknown	4 (7%)
ER status	
Positive	32 (53%)
Negative	29 (47%)
Neoadjuvant chemotherapy response	
pathological complete response (pCR)	20 (33%)
residual disease (RD)	41 (67%)

2.2.25 GSE32646

Primary breast cancer patients (n = 123, T1-4b N0-1 M0) who were consecutively recruited for the present study (Miyake T et al., 2012) had been treated with neoadjuvant chemotherapy (NAC) consisting of paclitaxel (80 mg/m^2) weekly for 12 cycles followed by 5-FU (500 mg/m^2), epirubicin (75 mg/m^2) and cyclophosphamide (500 mg/m^2) every 3 weeks for four cycles (paclitaxel followed by 5-fluorouracil/epirubicin/cyclophosphamide [P-FEC]) at Osaka University Hospital between 2004 and 2010. The NAC was indicated for stage IIA–IIIB breast cancer patients. Prior to NAC, every patient underwent vacuum-assisted core biopsy of tumors (Mammotome 8G; Ethicon Endosurgery, Johnson & Johnson, Cincinnati, OH, USA) under ultrasonographic guidance. The tumor samples obtained were then subjected to histological examination and DNA and RNA extraction. Tumor samples for extraction of DNA and RNA were snap frozen in liquid nitrogen and kept at −80°C until use. Inclusion of tumor cells in the biopsy samples for extraction of DNA and RNA was estimated using histological confirmation of tumor cells in the adjacent biopsy samples. The present study was approved by the Ethics Review Committee at Osaka University Hospital (Osaka, Japan) and informed consent was obtained from each patient before the core biopsy of tumors. Clinical characteristics are reported in Table 2.26. Gene expression data have been deposited into the GEO database with accession number GSE32646 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32646).

Table 2.26: Characteristics of patients from the GSE32646 dataset.

Clinical variables	Patients n=115
Age, years	
<40	15 (13%)
40-60	75 (65%)
>60	25 (22%)

Size (cm)	
T0/T1	5 (4%)
T2	87 (76%)
T3	18 (16%)
T4	5 (4%)
Grade	
G1	16 (14%)
G2	78 (68%)
G3	21 (18%)
LN status	
Positive	83 (72%)
Negative	32 (28%)
ER status (IHC)	
Positive	83 (62%)
Negative	32 (38%)
HER2 status (FISH, IHC)	
Not overexpressed	71 (70%)
Overexpressed	44 (30%)
Neoadjuvant chemotherapy response	
pathological complete response (pCR)	27 (23%)
residual disease (RD)	88 (77%)
Biopsy type	
Fine needle	0 (0%)
Core needle	115 (100%)

2.2.26 GSE19697

Core biopsies were obtained from 86 patients prior to neoadjuvant therapy out of which 70 fulfilled the requirements to undergo expression analysis (24 of these 70 were used in the published analysis) (Lin et al., 2010). pCR was defined as no residual invasive disease in the breast or lymph nodes. Residual in situ carcinoma was also considered as pCR. RNA was extracted from snap frozen 14-gauge core samples obtained from pre-treatment tumors. Specimens containing more than 40% of tumor on histological examination were analyzed. Table 2.27 reports all clinical and pathological characteristics of patients. Samples' RNA have been hybridized to HG-U133 Plus 2.0 microarrays. Raw data files have been entered into the NCBI Gene Expression Omnibus (GEO) repository with accession number GSE19697 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19697).

Table 2.27: Characteristics of patients from the GSE19697 dataset.

Clinical variables	Patients n=24
Size (cm)	
T1	1 (4%)
T2	20 (83%)
T3	3 (13%)
Grade	
G1	0 (0%)

G2	2 (8%)
G3	22 (92%)
ER status	
Positive	1 (4%)
Negative	23 (96%)
HER2 status	
Positive	1 (4%)
Negative	23 (96%)
Nodal status	
Positive	9 (38%)
Negative	15 (62%)
Race	
African American	10 (42%)
Caucasian	14 (58%)

2.2.27 GSE18728

Patients referred to the National Cancer Institute with newly diagnosed stage 2 or 3 breast cancer (American Joint Commission on Cancer, fifth version) with a tumor size of >2 cm were eligible. Eligibility criteria included an absolute neutrophil count > 1200/mm^3, platelet count > 100,000, creatinine < 1.5 mg/dL, calculated creatinine clearance > 50 mL/min, total bilirubin < 1.4, aspartate aminotransferase/alanine aminotransferase < 1.5× upper limit of normal, alkaline phosphatase < 2.5 upper limit of normal. Patients were excluded if they had a bleeding disorder, a cardiac ejection fraction below normal limits, serious cardiac events within the past 12 months, or prior treatment of breast cancer. Pregnant or lactating women were excluded. The protocol was approved by the Institutional Review Board of the National Cancer Institute and written informed consent was obtained. From January 2001 to August 2003, 30 patients were enrolled and treated with 116 total courses of docetaxel and capecitabine. One patient voluntarily withdrew from the trial after one cycle of therapy; the remainder of the patients completed the treatment phase of the trial (Korde et al., 2010). Twenty-one patients had baseline tumor biopsies that contained malignant cells, and are included in the analysis of baseline gene expression in responders vs. non-responders. Of these, 14 patients had evaluable tumor samples at baseline and after one cycle of chemotherapy, and are included in subsequent analyses. Demographic and tumor characteristics for the entire study population are reported in Table 2.28. Gene expression data have been deposited into the GEO database with accession number GSE18728 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18728).

Table 2.28: Characteristics of patients from the GSE18728 dataset.

Clinical variables	Patients n=61
ER status	
Positive	32 (53%)

Negative	29 (47%)
HER2 status	
Positive	17 (28%)
Negative	44 (72%)

2.2 Validation datasets

Independent cohorts of breast cancer samples hybridized on different types of array platform were used in this thesis: a proprietary AIRC 5X1000 dataset (see 2.2.1 paragraph) and GSE6861 (see 2.2.2 paragraph).

2.2.1 AIRC 5X1000 cohort

A sample size of 48 breast cancer samples provides an independent *in-house* clinical dataset of gene expression profiles. Samples were obtained before any type of treatment. RNA gene expression profiles from FFPE (formalin fixed, paraffin-embedded) tissues were hybridized on Illumina Whole Genome DASL HT-12 array. Being an *in-house* dataset, available cancer tissues allowed obtaining the correct molecular subtypes by immunohistochemistry (IHC) assays (Figure 2.1).

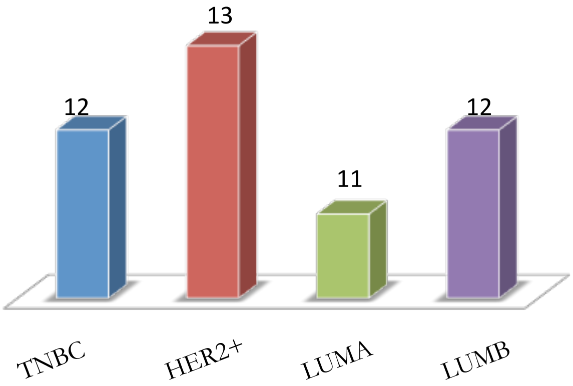


Figure 2.1: Molecular subtypes of patients from the AIRC 5X1000 dataset.

2.2.2 GSE6861

Samples of this dataset derived from the EORTC 10994 phase III breast cancer clinical trial, in which FEC activity (5-fluorouracil, cyclophosphamide, epirubicin) was compared with ET (epirubicin, docetaxel). 161 needle biopsies of locally advanced or large operable breast tumours were hybridised to Affymetrix X3P chips. The array data from the ER negative tumours (28/65 pathological CR in the FEC arm, 27/59 pathological CR in the ET arm) were used to validate the cell line-based chemotherapy response predictors developed at Duke University predictors developed at Duke University. Tumor characteristics for the entire study population are reported in Table 2.29. Gene expression data have been deposited into the GEO database with accession number GSE6861 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6861).

Table 2.29: Characteristics of patients from the GSE6861 dataset.

Clinical variables	Patients n=161
Size (cm)	
T0/T1	2 (1%)
T2	63 (39%)
T3	34 (21%)
Unknown	62 (39%)
Lymph nodes status	
N0	37 (23%)
N1	55 (34%)
N2	7 (4%)
N3	62 (39%)
ER status (IHC)	
Positive	37 (23%)
Negative	65 (40%)
Unknown	59 (37%)
Neoadjuvant chemotherapy response	
pathological complete response (pCR)	66 (41%)
residual disease (RD)	95 (59%)
Neoadjuvant chemotherapy type	
Anthracycline-based	102 (63%)
Taxane - anthracycline-based	59 (37%)

2.3 Meta-analysis of gene expression data

A common criticism about gene expression-based prognostic/predictive signatures regards the limited number of samples used for their development and validation. No doubt that an adequately powered sample size is one of the most important, and most overlooked, aspects in the definition of a prognostic/predictive signature. The small number of samples in individual studies, particularly for human studies where there is a high degree of both intra- and inter-population variability, represents a major limitation for the detection of gene-expression signatures and, ultimately, results in disease biomarkers that are population dependent, rather than having global applicability (Bhattacharya and Mariani, 2009). However, the datasets deposited in public gene expression data repositories, as GEO, represent a real opportunity to strengthen the validation of prognostic/predictive signatures through the meta-analysis of multiple, independently generated data focusing on the same tumor type. Meta-analysis strategies can be divided into data integration and data combination. Data integration aims at merging results obtained from the analyses of independent studies through statistical techniques. Instead, data combination integrates multiple datasets directly at the level of raw data and generates a unique matrix of gene expression signals. The direct merging of raw data from different studies is applicable only when expression profiles have been obtained using the same array technology (e.g. Affymetrix, Agilent, Illumina, etc.) and requires an *ad-hoc* normalization step. The computational pipeline for the combination of multiple datasets is composed of three major steps:

- re-definition of outcome descriptions;
- probe re-mapping and selection;
- integration and normalization of different datasets.

Re-definition of event descriptions has been conducted carefully considering the clinical annotations of any single study and defining two major types of events, one associated to the metastatic spread and one to overall survival (see 2.2.1 paragraph). Probe re-mapping and selection is based on the adoption of modified custom Chip Definition Files (custom-CDF) (see 2.2.2 paragraph) while the integration and normalization of gene expression signals has been obtained applying the *virtual chip* procedure (Bisognin et al., 2010; Fallarino et al., 2010) (see 2.2.3 paragraph). The application of this approach allowed constructing a meta-dataset of 3661 breast cancer samples derived from the combination of the 27 gene expression datasets described in Table 2.1 (see Results for details).

2.3.1 Re-definition of outcome and clinic-pathological variables descriptions for breast cancer meta-dataset

Retrieval, organization and utilization of meta-information is still an extremely critical step which impacts the correct match between raw data files and sample IDs and the organization of samples into meaningful, homogeneous groups. This task is further complicated by the fact that datasets may be incompletely annotated, the relationship between specimen, biological sample, phenotypic characteristics and raw data files, the most granular object in repositories, may be not sufficiently explicit, and the procedures for managing large numbers of data files and related meta-information are tedious and error prone (Ioannidis et al., 2009). To homogenize the clinical information of the various cancer datasets, the outcome descriptions have been re-defined based on the clinical annotations of any single study. Specifically, the description the various authors used to indicate the type of outcome have been surveyed in all breast cancer datasets and used to define two major types of events, i.e., *metastasis* and *survival. Metastasis* is associated to the metastatic spread and includes the following descriptions in the original studies:

- recurrence free survival;
- metastasis free survival;

- distant metastasis free survival or distant recurrence (and subtypes at different districts as lung, bone, and brain distant metastasis free survival/distant recurrence);
- time to distant metastasis.

Survival is associated to death because of cancer and includes the following descriptions in the original studies:

- overall survival;
- disease free survival;
- disease specific survival.

In neo-adjuvant chemotherapy studies, patients were labeled according their tumor sensibility or resistance to specific chemotherapy treatment; we used to define two types of events, i.e., *pathological complete response* and *residual disease*. *Pathological complete response* (pCR) is defined as no invasive and no in situ residuals in breast and nodes after chemotherapy. Patients with noninvasive or focal-invasive residues or involved lymph nodes should be considered as having *residual disease*. The most used of systemic neo-adjuvant chemotherapies included in our studies of interest comprised anthracycline- or taxane-anthracycline based protocols. The anthracyclines are among the most effective anticancer treatments ever developed and are effective against more types of cancer than any other class of chemotherapeutic agents. Their main adverse effect is several cardiotoxicity, which considerably limits their usefulness. They inhibit i) DNA and RNA synthesis by intercalating between base pairs of the DNA/RNA strand, thus preventing the replication of rapidly-growing cancer cells (Takimoto et al., 2008) and ii) topoisomerase II (TOP2A) enzyme, preventing the relaxing of supercoiled DNA and thus blocking DNA transcription and replication; topoisomerase II stabilizes the topoisomerase II complex after it has broken the DNA chain. This leads to topoisomerase II mediated DNA-cleavage, producing DNA breaks (Pommier et al., 2010). The binding of topoisomerase II inhibitor prevents DNA repair by ligase (Buhl et al., 1993). Available agents include for examples epirubicin, doxorubicin, daunorubicin, etc. Taxanes are diterpenes produced by the plants of the genus Taxus (yews), and are widely used as chemotherapy agents (Hagiwara et al., 2004). The principal mechanism of action of the taxanes is the disruption of microtubule function. Microtubules are essential to cell division, and taxanes stabilize GDP-bound tubulin in the microtubule, thereby inhibiting the process of cell division, "frozen mitosis". Thus, in essence, taxanes are mitotic inhibitors. Taxane agents include paclitaxel (Taxol) and docetaxel (Taxotere). In this thesis, since we have different types of anthracyclines or taxane, we classified two main chemotherapy protocols: the anthracycline-based (labeled as A) and taxane-plus anthracycline-based (labeled as AT) chemotherapy.

Patients were also characterized by several clinic-pathological variables such as:

- Age of patients
- tumor size (T): dimension of tumor (mm)
- lymph nodes status (N): involvement of regional and axillar lymph nodes
- stage of tumors: cancer staging can be divided into a clinical stage and a pathologic stage. In the TNM (Tumor, Node, Metastasis) system, clinical stage and pathologic stage are denoted by a small "c" or "p" before the stage (e.g., cT3N1M0 or pT2N0). Clinical stage is based on all of the available information obtained before a surgery to remove the tumor. Thus, it may include information about the tumor obtained by physical examination, radiologic examination, and endoscopy. Pathologic stage adds additional information gained by examination of the tumor microscopically by a pathologist.
- histological grade: histological study of the tumor tissue removed during after a biopsy to check: 1) how much the cancer cells look like normal cells (the more the cancer cells look like normal cells, the lower the tumor grade tends to be) and 2) how many of the cancer cells are in the process of dividing (the fewer cancer cells that are in the process of dividing, the more likely it is that the tumor is slow-growing slowly and the lower the tumor grade tends to be).

All these clinical variables were homogenized in order to have the same labeled for all patients (see Results, Table 3.2).

2.3.2 Probe re-mapping and selection

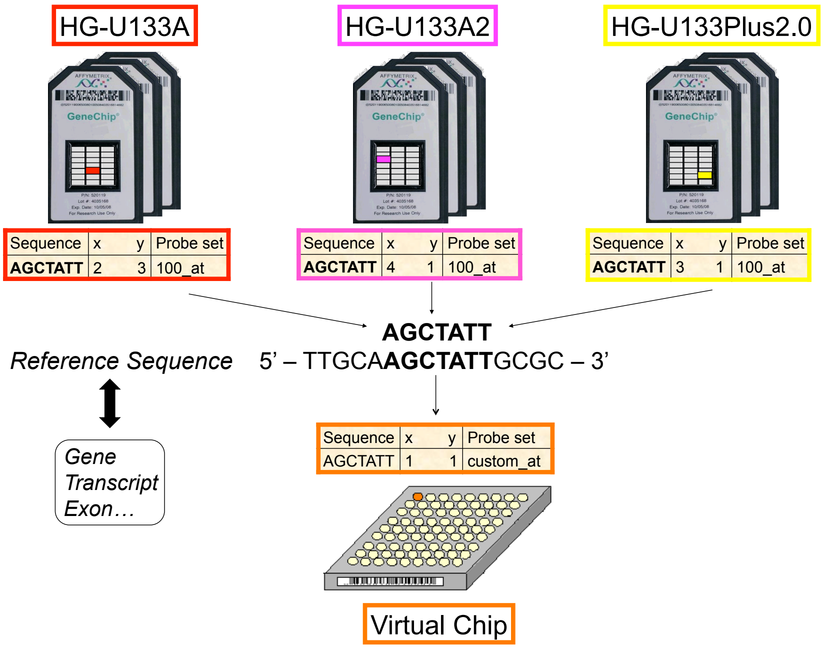
Performing a meta-analysis of independent microarray studies requires to carefully handling the heterogeneity of array designs, which complicates cross-platform integration, and of sample descriptions, which impacts the correct characterization of specimens. At least for the case of Affymetrix arrays, cross-platform comparison has partially been solved by the adoption of custom Chip Definition Files (custom-CDF) that allow matching expression profiles across subsequent generations of microarrays (Gautier et al., 2004; Dai et al., 2005; Ferrari et al., 2007). Despite the computational differences, all methods for signal quantification rely on the correspondence between probes and genomic sequences. The Affymetrix Chip Definition Files (CDFs) encode the physical design of the microarray and contain the sequence details to link the oligonucleotide probes of the chip to the interrogated transcripts. The information of a CDF file relies so deeply on the genome annotation contained in the databases that the same name of the chip reflects the version of the UniGene Build used for probe design (e.g., the HG-U133 expression set and the human UniGene Build 133). The evolution of genome sequence annotation from the time when probe sets were designed caused a massive deviation from the original one-to-one probe set/transcription locus (i.e. UniGene entry) assignment. Affymetrix continuously updates probe sets annotations and redefines the links between probe sets and genes indicating the UniGene cluster that contains the probe set representative sequences and linking them to the corresponding EntrezGene ID. Similarly, the Bioconductor Biocore team quarterly releases CDFs and annotation libraries at the Bioconductor website, which can be used for analysis of gene expression data in R environment. However, these update actions simply affect the qualitative attributes of probe sets without any degree of control on the effective matching between probes and genome sequences. As such, Dai et al. developed a novel system for associating probes to genomic information, based on custom-probe sets which are composed of at least four probes specifically matching the same sequence (Dai et al., 2005). They defined custom probesets based on updated versions of various datasets entries (e.g. Entrez, RefSeq) and generated custom CDFs for the most popular Affymetrix microarrays. The development of custom CDF deeply improves the analysis outcome when the focus of the experiment is the identification of differentially expressed genes. In this thesis, probes designed by Affymetrix have been re-mapped and re-defined on Affymetrix and Entrez CDFs.

2.3.3 Quantification of combined gene expression signals

The integration and normalization of different datasets can be obtained, first, generating the gene expression signals in each dataset, then, combining the expression levels in a single data matrix, and, finally, applying a meta-normalization step to the combined data matrix.

Briefly, the final normalization step of the combined expression set is a crucial issue, since the direct integration of different datasets may result in misleading outcomes, due to different experimental conditions, laboratory-dependent bias, etc. In alternative, RMA-quantile normalization could be directly used on the entire dataset as far as all data refer to a unique platform. Although RMA-quantile is the most effective normalization method, it cannot be applied to data obtained from different platforms (e.g., the HG-U133A, the HG-U133 Plus 2.0, and HG-U133A2 arrays), due to differences in number, type, and physical localization of probes. As such, Bisognin and collaborators implemented a procedure, the *Virtual Chip*, to create a custom and virtual microarray grid that integrates the geometry and probe content of two or more types of Affymetrix arrays (Bisognin et al., 2010; Fallarino et al., 2010). Once defined the virtual grid, all CEL files, obtained from different platforms, are re-organized to match a single platform, i.e., the virtual chip. As such, raw data, originally from different types of microarrays, become homogeneous in terms of platform and can be preprocessed and normalized adopting standard approaches, as RMA or GCRMA (see RMA algorithm). The *Virtual Chip* approach allows combining data directly at the level of probe fluorescence intensity and presents the advantage that gene expression signals are generated with a single step of background correction, normalization and summarization.

The construction of the virtual grid is inspired by the generation of custom Chip Definition Files (CDFs). In custom CDFs, probes matching the same transcript, but belonging to different probe sets, are aggregated into putative custom-probe sets, each one including only those probes with a unique and exclusive correspondence with a single transcript. Similarly, probes matching the same transcript but located at different coordinates on different type of arrays may be merged in custom-probe sets and arranged in a virtual platform grid, whose geometry can be arbitrarily set (Figure 2.2).



Sequence	x	y	Probe set
AGCTATT	2	3	100_at

Sequence	x	y	Probe set
AGCTATT	4	1	100_at

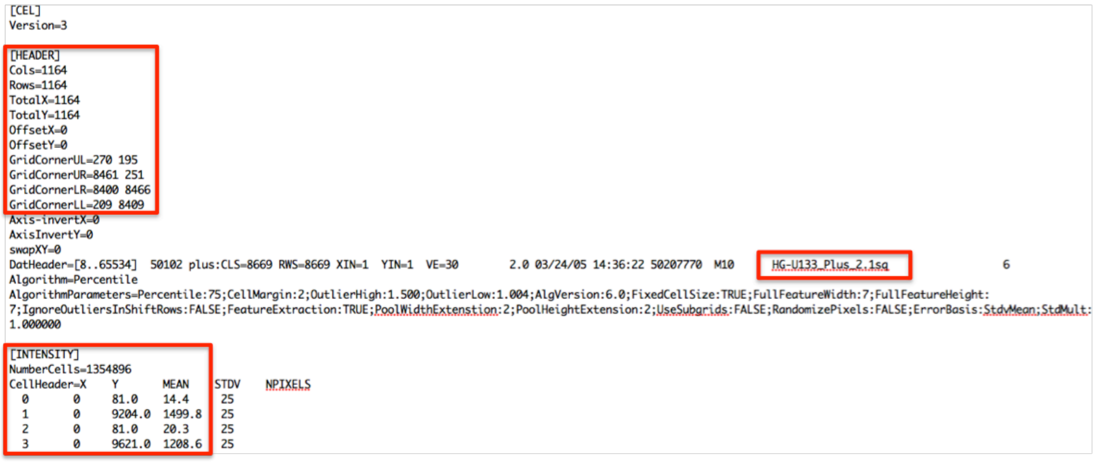
Sequence	x	y	Probe set
AGCTATT	3	1	100_at

Sequence	x	y	Probe set
AGCTATT	1	1	custom_at

Figure 2.2: Construction of the *Virtual Chip*.

As for any other microarray geometry, this virtual grid may be used as a reference to create a *virtual CDF* file containing the probes of the *Virtual Chip* and their coordinates on the virtual platform. The probes included in the *virtual CDF* are those shared among the platforms of interest, with the additional condition of generating *custom probe set* of at least 4 probes. The *virtual CDF* can be derived from any *custom CDF*, e.g., those developed by Dai and publicly accessible at the Molecular and Behavioral Neuroscience Institute Microarray Lab. Finally, the *virtual CDF* can be used as the geometry file in RMA as far as the original CEL files are properly re-mapped to match the topology described in the *virtual CDF*. Re-mapped CEL files, called *virtual CEL file*, are homogeneous in terms of platform and gene expression data can be generated with a single step of background correction, normalization and summarization directly from the fluorescence signals of all microarrays composing the meta-dataset. CEL file re-mapping requires re-defining:

- the content of the [HEADER] field of Figure 2.3, i.e., all physical coordinates (total number of cells containing the probes, indicated by *Cols, Rows*, *TotalX*, and *TotalY*, and localization of the 4 border cells) and the name of the platform (*HG-133_Plus_2.1sq*);
- all data contained in the [INTENSITY] field, i.e., physical localization (*X* e *Y*) and fluorescence intensity (*MEAN*) of any probe.



```
[CEL]
Version=3

[HEADER]
Cols=1164
Rows=1164
TotalX=1164
TotalY=1164
OffsetX=0
OffsetY=0
GridCornerUL=270 195
GridCornerUR=8461 251
GridCornerLR=8400 8466
GridCornerLL=209 8409
Axis-invertX=0
AxisInvertY=0
swapXY=0
DatHeader=[8..65534]  50102 plus:CLS=8669 RWS=8669 XIN=1  YIN=1  VE=30        2.0 03/24/05 14:36:22 50207770  M10     HG-U133_Plus_2.1sq          6
Algorithm=Percentile
AlgorithmParameters=Percentile:75;CellMargin:2;OutlierHigh:1.500;OutlierLow:1.004;AlgVersion:6.0;FixedCellSize:TRUE;FullFeatureWidth:7;FullFeatureHeight:
7;IgnoreOutliersInShiftRows:FALSE;FeatureExtraction:TRUE;PoolWidthExtenstion:2;PoolHeightExtension:2;UseSubgrids:FALSE;RandomizePixels:FALSE;ErrorBasis:StdvMean;StdMult:
1.000000

[INTENSITY]
NumberCells=1354896
CellHeader=X     Y       MEAN     STDV    NPIXELS
  0       0      81.0    14.4      25
  1       0      9204.0  1499.8    25
  2       0      81.0    20.3      25
  3       0      9621.0  1208.6    25
```

Figure 2.3: Fields modified in a *virtual CEL file*.

RMA algorithm

In Affymetrix microarrays, the expression signal of each gene is quantified summarizing the intensities of all the oligonucleotides, i.e. the probes (e.g., 11 or 16), of a probe set matching a target gene or transcript. Each probeset is composed by a set of Perfect Match (PM) and Mis-Match (MM) probes, that contains mismatches and should measure non-specific hybridization. The signal can be generated using a series of statistical or model-based algorithms (i.e., MAS5.0, MBEI, RMA, GCRMA, PLIER, PDNN). In this thesis expression levels were quantified using RMA algorithm (Robust Multichip Average; Irizarry et al., 2003). The RMA method for computing an expression measure begins by computing background- corrected perfect match (PM) intensities for each perfect match cell on every array. The background corrected intensities are computed in such a way that all background- corrected values must be positive. After background correction, the log-2 of each background-corrected PM intensity is obtained. These background-corrected and log- transformed PM intensities are normalized using the quantile normalization method developed by Bolstad et al. (Bolstad et al., 2003). In the quantile normalization method, the highest background-corrected and log-transformed PM intensity on each array is determined. These values are averaged, and the individual values are replaced by the average. This process is repeated with what were originally the second highest background-corrected and log-transformed PM intensities on each array, the third highest, etc. Following quantile normalization, an additive linear model is fit to the normalized data to obtain an expression measure for each probe on each array. The linear model for a particular probeset can be written as $Y_{ij}=m_i+a_j+e_{ij}$ where Y_{ij} denotes the normalized probe value corresponding to the *i*-th array and the *j*-th probe within the probeset, m_i denotes the log-scale expression for the probeset in the sample hybridized to the *i*-th array, *aj* denotes the probe affinity effect for the *j*-th probe within the probeset, and e_{ij} is a random error term. Tukey's median polish is used to obtain estimates of the m_i values. These estimates serve as the log-scale expression measures associated with the particular probeset.

2.4 Molecular subtype classification models

Breast tumors are biologically heterogeneous and exhibit different clinical outcomes and an accurate identification of molecular subtypes would make it possible to better understand breast cancer biology and to test the prognostic/predictive value of molecular markers with respect to these subtypes. In this thesis, we used PAM50 Single Sample Prediction (see 2.4.1 paragraph) and Subtype clustering models (see 2.4.2 paragraph). PAM50 and Subtype clustering models are the most widely used classifiers in microarray studies.

2.4.1 Clustering

Clustering is a group of methodologies that assigns a set of objects into groups (clusters) that have a high internal homogeneity and a strong dishomogeneity with other clusters. In the microarray setting, clustering can be applied to genes or to samples. In the latter case, the focus is on clustering experiments (samples) according to a list of *n* genes, i.e., grouping together samples with a similar transcriptional profile. Among the various algorithms, the most commonly used is hierarchical clustering. Hierarchical clustering methods are non-parametric method, i.e. they do not rely on a probabilistic model that generates the observed data. However, they require that the analyst specifies the strategy for building the dendrogram, the measure of dissimilarity (distance) and the linkage, i.e. the measure The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables. For a sample of size *n*, the *n* raw scores X_i, Y_i are converted to ranks x_i, y_i and ρ is computed from equation (1):

$$\rho = \frac{\sum_i (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_i (r_i - \bar{r})^2}\sqrt{\sum_i (s_i - \bar{s})^2}}$$

Another important parameter to define hierarchical clustering is linkage, which determines the distance between sets of objects as a function of the pairwise distances between objects. The result of hierarchical clustering is a tree (dendrogram) that starts from the leaves (samples) and iteratively groups them together until forming a super-cluster containing all elements. To obtain a specific number of groups of samples, the tree has to be "cut" at the appropriate level.

2.4.2 Single sample predictor (SSP)

Hierarchical clustering has been widely used to identify molecular subtypes of breast cancer, but this approach can only be applied retrospectively to sufficiently sized cohorts of patients (Weigelt et al., 2010, Pusztai et al., 2006), but not prospectively to individual samples. Moreover, the hierarchical clustering model fitted onto the training set could not be used directly to identify the subtype of a tumor of a new breast cancer patient. Indeed, any new case should be added to the training set and the hierarchical clustering model should be fitted again, leading to a potentially different dendrogram. To avoid this difficulty, it was developed methods based on nearest centroid (Dudoit et al., 2002), called SSP (Single Sample Predictor). To define the SSPs, each molecular subtype was initially identified by hierarchical clustering based on several "intrinsic" gene lists (Table 2.30), and then the centroids (i.e., mean gene expression profile) of each molecular subtype (ie, luminal A, luminal B, HER2-enriched, basal-like, and normal breast-like) were derived; after, the gene expression profile of the new individual sample was compared to each centroid and assigned by the SSP to the nearest subtype centroid as determined by Spearman correlation. The method used is explained in Figure 2.4.

Table 2.30: "Intrinsic" gene lists and Single sample predictor genes.

Study	Intrinsic gene lists	Single sample predictor (SSP) genes
Sorlie T et al., PNAS 2003	534	500
Hu Z et al., BMC Genomics 2006	1300	306
Parker JS et al., J Clin Oncol 2009	1906	50

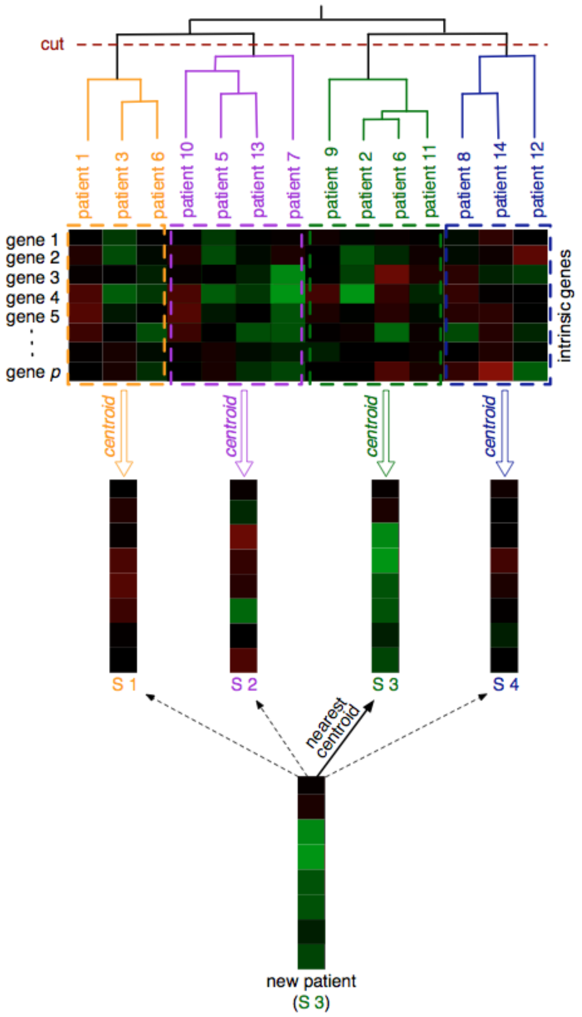


Figure 2.4: Illustration of the SSP method used to identify breast cancer molecular subtypes. A hierarchical clustering is performed by using the intrinsic gene list to generate a dendrogram of patients' tumors. The dendrogram is then cut to identify the different subtypes (in this case, S1 to S4). A centroid is computed for each subtype. A nearest centroid approach is used to classify a new patient's tumor. In this case, the new tumor is highly correlated with centroid S3, making this the nearest centroid. So the new tumor is predicted to be of the subtype 3.

In this thesis, intrinsic molecular subtypes were assigned using the *intrinsic.cluster.predict* function of *genefu* R package using the "50 intrinsic gene list" as proposed by Parker and colleagues: the 50-gene set classifier (henceforth called PAM50). PAM50 consisted of centroids constructed using the PAM (Prediction Analysis of Microarray) algorithm (Tibshirani et al., 2002) and distances calculated using Spearman's rank correlation. The genes used for subtyping are provided in Table 2.31.

Table 2.31: PAM50 gene list.

Gene symbols	Entrez gene ID
ACTR3B	57180
ANLN	54443
BAG1	573
BCL2	596
BIRC5	332
BLVRA	644
CCNB1	891
CCNE1	898
CDC20	991
CDC6	990
CDCA1	83540
CDH3	1001
CENPF	1063
CEP55	55165
CXXC5	51523
EGFR	1956
ERBB2	2064
ESR1	2099
EXO1	9156
FGFR4	2264
FOXA1	3169
FOXC1	2296
GPR160	26996
GRB7	2886
KIF2C	11004
KNTC2	10403
KRT14	3861
KRT17	3872
KRT5	3852
MAPT	4137
MDM2	4193
MELK	9833
MIA	8190
MKI67	4288
MLPH	79083
MMP11	4320
MYBL2	4605
MYC	4609
NAT1	9
ORC6L	23594
PGR	5241
PHGDH	26227
PTTG1	9232
RRM2	6241
SFRP1	6422
SLC39A6	25800
TMEM45B	120224
TYMS	7298
UBE2C	11065
UBE2T	29089

2.4.2 Subtype clustering models (SCMs)

Subtype Clustering Model is an unsupervised method able to robustly identify the breast cancer molecular subtypes. It returns an accurate estimate of the classification uncertainty, i.e. for each subtype, the probability for a patient to have a tumor of this subtype. It is based on a mixture model; it assumes that the data is an independent and identically-distributed sample from a population described by a probability density function. This density function is characterized by a parametrized model, taken to be a mixture of component density functions, where each component density describes one of the clusters. The population B of objects b is described by a finite mixture distribution of the form (2)

$$\Pr(b) = \sum_{r=1}^{u} \pi_r \Pr(b|r)$$

where *u* is the number of clusters in the population, π_r are the mixing proportions such that $\sum_{r=1}^{u} \pi_r = 1$, and Pr(b | r) is the r^th probability density function of b. The quantity π_r is typically interpreted as the prior probability that a data point is generated by the r^th component of the mixture. There are three sets of parameters to estimate: the values of π_r, the parameters of the probability distribution of each of the components, and the value of *u*. The usual approach to clustering using finite mixture distributions is first to specify the form of the component distributions, Pr(b | r). Then the number of clusters, u, is prescribed. The parameters of the model are estimated and the objects are grouped on the basis on their estimated posterior probabilities of cluster membership. Using Bayes' theorem, the object b is assigned to cluster r if

$$\Pr(r|b) = \geq \Pr(s|b)\ \forall r \neq s \text{ with } r, s\ \in \{1, \dots, u\}$$

where (3)

$$\Pr(r|b) = \frac{\pi_r \Pr(b|r)}{\sum_{s=1}^{u} \pi_s \Pr(b|s)}$$

The analyst could easily use the probabilities of an object b to belong to each cluster. The most widely available form of mixture distribution for continuous variables is the mixture of normal (Gaussians) distributions, where the r^th component Pr(b | r) ~ N(μ_r, Σ_r), where μ_r, and Σ_r are the means and covariance matrix of a multivariate normal distribution. So (4)

$$\Pr(b) = \sum_{r=1}^{u} \pi_r \, \mathrm{N}(\mathrm{b};\ \mu_r, \Sigma_r)$$

The estimation of the parameters of a normal mixture model can be achieved by the maximum likelihood procedure through the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). The Bayesian information criterion (BIC) (Schwarz et al., 1978) can be used to estimate the likelihood of a mixture model with u clusters. The BIC is the value of the maximized log- likelihood with a penalty for the number of parameters in the model, and allows comparison of models with different parameterizations and/or different numbers of clusters.

The Subtype Clustering Model is composed of two steps:

- Prototype-Based Feature Transformation: the genome-wide microarray data are transformed into few features quantifying the activity of key biological processes in breast cancer. It uses a robust estimation of gene co-expression and a priori knowledge about the biological processes of interest. These features should be specific to the biological process they represent, i.e. a feature representative to a biological process should not be also representative to another biological process. So, the genes sharing a biological affinity to the same biological process are before clustered together and then each cluster of genes is summarized by a single feature quantifying the activity at the gene expression level of the corresponding biological process (*gene module*). In this case, the method uses three gene modules: ER and HER2 signaling, and proliferation signaling (aurora kinase A, AURKA).
- Subtypes identification: represent the patients in a low dimensional space defined by gene module scores quantifying the activity of the three gene modules.

The model-based clustering is a mixture of Gaussians in a low dimensional space. The input space is defined by the three gene module scores computed through the prototype-based feature transformation. Let $X_{n \times p}$ be the matrix of *p* gene module scores for *n* patients and x_i be the profile of the ith patient. A mixture of Gaussians model can be written as (5)

$$\Pr(x_i) = \sum_{r=1}^{u} \pi_r \, N(x_i; \mu_r, \Sigma_r)$$

where u is the number of Gaussians, π_r is the prior probability of x_i to be generated by the rth Gaussian $N(x_i\ ;\ \mu r\ ,\ \Sigma_r\ )$ of mean μ_r and covariance matrix Σ_r. From equation (3), we define the probabilities to belong to each subtype r as (6)

$$\Pr(r|x_i) = \frac{\pi_r N(x_i; \mu_r, \Sigma_r)}{\sum_{s=1}^{u} \pi_s N(x_i; \mu_s, \Sigma_s)}$$

So, Pr(r | x_i) is the probability that the patient having the profile x_i has a breast tumor of subtype r.

In this thesis, we used the *SCMGENE* (Desmedt et al., 2008), *SCMOD1* (Desmedt C., 2008) and *SCMOD2* methods (Warapati et al., 2008). *SCMGENE*, a simplified *SCM*, is based on the expression of ER, HER2, and AURKA genes; whereas, *SCMOD1* is based on 726 genes (Desmedt et al., 2008), and *SCMOD2* on 663 genes (Wirapati et al, 2008) both related to ER, HER2, and AURKA gene module. All methods use the Perou's subtype nomenclature (Perou et al., 2000): basal-like, HER2-enriched, and luminal A and B, which correspond, respectively, to the ER-/HER2-, HER2+, and ER+/HER2- low and high proliferation tumors. The function *subtype.cluster.predict* of *genefu* R package fits the Subtype Clustering Models. This function includes MCLUST function (*mclust* package) that combines hierarchical clustering, Expectation-Maximization (EM) algorithm (Dempster et al., 1977) and the Bayesian Information Criterion (BIC). There are several models to fit mixture of Gaussians; in this case, we used EEI models with diagonal variance, equal volume, equal shape and identical orientation of distributions.

2.5 Gene signatures

This paragraph describes the gene or signatures used in this thesis. The gene/ gene sets have been applied in the study for evaluation of a multifactorial approach by in-silico analysis for predicting response to neo adjuvant anthracycline-based chemotherapy in triple negative breast cancer patients. Paragraph 2.5.6 describes a potential predictive test, A-score (anthracycline-based score).

2.5.1 Penetration of drug into the cancer cell: HIF and SHARP1 signatures

Hypoxia, promoted by HIFs, is a well-known contributor to decreased drug penetration, and chemo resistance (Teicher et al., 1994). Montagner et al recently described a hypoxia signature of 22 genes (Table 2.32), with increased expression correlated with increased HIF activity (Montagner et al., 2012). A direct interaction between SHARP1 (a downstream target of the tumor suppression gene p63) and HIF1α and HIF2α was demonstrated, with a signature of low SHARP1 activity in TNBC conferring increased HIF function and increased hypoxia (Montagner et al., 2012). The SHARP1 signature (Table 2.33) measures low SHARP activity and thus increased HIF function.

Table 2.32: HIF gene signature.

Symbol	EntrezGene ID
ADM	133
ALDOC	230
BHLHE40	8553
BIRC2	329
BNIP3	664
CENPF	1063
DDIT4	54541
ENO2	2026
GLRX	2745
HK2	3099
INSIG1	3638
MAFF	23764
NDRG1	10397
PDK1	5163
PDPK1	5170
PFKP	5214
PGK1	5230
SAP30	8819
SLC2A1	6513
SLC2A3	6515
VEGFA	7422
WSB1	26118

Table 2.33: SHARP1 gene signature.

Symbol	EntrezGene ID
AGTPBP1	23287
CHN2	1124
COBL	23242
DSC2	1824
EPS8L2	64787
F2RL1	2150
FSCN1	6624
GBE1	2632
GPR56	9289
HSF2	3298
IFIT3	3437
IGF2BP3	10643
IMPA2	3613
ITGB2	3689
LPIN1	23175
LYZ	4069
ME1	4199
NOX5	79400
PLCE1	51196
RAPGEF5	9771
S100A3	6274
SH2D3A	10045
SLC5A3	6526
SLCO4A1	28231

SREBF1	6720
TGFA	7039
WWTR1	25937

2.5.2 Location of TOP2A protein within the nucleus: LAPTM4B mRNA

In order to work effectively, the target of anthracyclines, TOP2A protein, must have access to nuclear DNA; thus, it must be located in the nucleus. Nuclear export of TOP2A protein may contribute to anthracycline resistance (Oloumi et al., 2000, Turner et al., 2004). TOP2A protein nuclear location might be inferred using expression level of LAPTM4B (Lysosomal Associated Protein Transmembrane 4B gene) (Li et al., 2010). LAPTM4B gene resides on chromosome 8q22, with overexpression shown to increase sequestration of anthracyclines in the cytoplasm. Increased levels of LAPTM4B mRNA have been correlated with increased anthracycline resistance, while selective depletion of LAPTM4B significantly increased sensitivity to anthracycline, but not cisplatin or taxane, chemotherapy (Li et al., 2010).

2.5.3 Increased expression of TOP2A: TOP2A mRNA

Topoisomerase II (TOP2A) is a key enzyme in DNA replication, one of the molecular targets of anthracyclines, and it is amplified in 24% to 54% of HER2-amplified tumors (Slamon et al., 2009). TOP2A gene amplification has been shown to predict increased sensitivity to anthracyclines in several studies (Di Leo et al., 2011; Di Leo et al., 2002; Press et al., 2011; Slamon et al., 2011, Arriola et al., 2007; Desmedt et al., 2011). However this finding has not been entirely consistent across all trials (Bartlett et al., 2008; Martin et al., 2011), and further research is needed to clarify the role of TOP2A gene status as a predictive biomarker of anthracycline sensitivity.

2.5.4 Induction of apoptosis: YWHAZ and Minimal signature

The anti-apoptotic gene YWHAZ (coding for 14-3-3ζ) resides on chromosome 8q22 close to LAPTM4B gene and may promote *de novo* anthracycline resistance (Li, 2010). Increased expression has been associated with increased doxorubicin resistance in breast cancer cell lines, and early relapses after anthracycline chemotherapy. siRNA knockdown of YWHAZ in breast cancer cell lines significantly increased doxorubicin-induced apoptosis (Li et al., 2010). An alternate marker of apoptosis is the Minimal Signature, MS, (Adorno et al., 2009), comprising two genes, SHARP1 and CCNG2. As with SHARP1, CCNG2 is a downstream target of p63. As p63 is inhibited by mutant p53, lack of MS expression implies dysfunction in the p53 pathway, the major apoptotic pathway in the presence of oncogenic stress, and may be a suitable surrogate for lack of apoptosis. YWHAZ and the MS were both selected for evaluation as markers of apoptosis

2.5.5 Active immune and stromal function: STAT1 and PLAU signatures

Both innate and adaptive immune responses are important in anthracycline toxicity (Mattarolo et al., 2011, Zitvogel et al., 2008). Anthracyclines trigger immunogenic cell death by eliciting tumor-specific IFNγ CD8+ cytotoxic T lymphocytes, thus an anthracycline-induced anticancer immune response can help eradicate residual cancer cells, or maintain residual cells in state of dormancy. Moreover, immune module scores (Teschendorff et al., 2007, Desmedt, 2008) (STAT1, Table 2.34) have been associated with higher probability of achieving pCR after anthracycline +/- taxane chemotherapy among all breast cancer subtypes when defined by immunohistochemistry (Ignitiadis et al., 2012). Closely related to immune function, stromal signatures (PLAU, Table 2.35) may also be useful in predicting anthracycline sensitivity or resistance (Desmedtet al., 2008, Farmer et al., 2009).

Table 2.34: Immune gene signature (STAT1).

Symbol	EntrezGene ID
STAT1	6772
CXCL10	3627
TAP1	6890
CXCL11	6373
INDO	3620
CXCL9	4283
MX1	4599
LAMP3	27074
ISG15	9636
RTP4	64108
HERC6	55008
IFI44L	10964
MX2	4600
IFIT3	3437
HERC5	51191
RSAD2	91543
DDX58	23586
CCL5	6352
ADAMDEC1	27299
CD2	914
NA	55601
HCP5	10866

NMI	9111
SPOCK2	9806
CCL8	6355
TRIM22	10346
LYZ	4069
IRF1	3659
LAG3	3902
PSCDBP	9595
TFEC	22797
UBD	10537
SP140	11262
CTSC	1075
IFI6	2537
PLA2G7	7941
CD3G	917
ECGF1	1890
PLAC8	51316
FGL2	10875
GZMK	3003
CD48	962
STAT4	6775
GPR18	2841
P2RX5	5026
IFI30	10437
SH2D1A	4068
LAPTM5	7805
CD69	969
PTPN7	5778
IRF8	3394
PIM2	11040
ETV7	51513
GPR171	29909
PSME1	5720
BIRC3	330
FASLG	356
IFITM1	8519
IFIT5	24138
ITGB2	3689
BTN3A2	11118
HCLS1	3059
SECTM1	6398

ARHGAP15	55843
KLRK1	22914
IGSF6	10261
EBI2	1880
NA	26034
SNX10	29887
NA	79132
BST2	684
NA	55337
APOC1	341
NA	51237
NA	445347
ZC3HAV1	56829
DDAH2	23564
LILRA4	23547
EBI3	10148
KLRC3	3823
CLEC4A	50856
CD40LG	959
VAV1	7409
GLRX	2745
ACP5	54
RFX5	5993
CECR1	51816
TRAF3	7187
RAB8A	4218
IL18	3606
EFNA1	1942
RASGRP1	10125
REC8L1	9985
CCRL2	9034
DNAL4	10126

Table 2.35: Stromal gene signature (PLAU).

Symbols	EntrezGene ID
PLAU	5328
BMP1	649
MMP14	4323
THY1	7070
COL5A2	1290

ADAM12	8038
ANGPTL2	23452
MFAP2	4237
SERPINH1	871
COL6A1	1291
ISLR	3671
PDLIM7	9260
PARVA	55742
OLFML2B	25903
TAGLN	6876
CTSA	5476
PDGFRB	5159
MXRA8	54587
OSMR	9180
COL3A1	1281
GREM1	26585
FAP	2191
DBN1	1627
BICD2	23299
TNFRSF12A	51330
VDR	7421
SNAI2	6591
EPB41L2	2037
FKBP14	55033
NBL1	4681
CAP1	10487
ATP6V1B2	526
EPHB4	2050
TRAM2	9697
DDR2	4921
GFPT2	9945
NID1	4811
OFD1	8481
CADM1	23705
STAB1	23166
TPST2	8459
PPP1R15A	23645
PDLIM3	27295
ATPIF1	93974
TRIM33	51592
MMP3	4314

EPYC	1833
ANKRD46	157567
CPNE1	8904
BCL3	602
GLB1	2720
UBL5	59286
ULK1	8408
NOL8	55035
TGFB2	7042
PDGFB	5155
BASP1	10409
SDS	10993
RPS27A	6233
ENC1	8507
ACAN	176
ZNF518A	9849
RPL18	6141
MEF2A	4205
DNASE1L1	1774
MYO1B	4430
JPH2	57158

2.5.6 A-score

Desmedt et al. developed a gene expression signature to identify patients who would not benefit from anthracyclines and could thus be spared the non-negligible risks of this type of chemotherapy. The anthracycline-based score (A-Score) was developed integrating three biologically different expression signatures associated with the efficacy of anthracyclines: TOP2A signature, stroma and immune response signatures. TOP2A signature is composed by TOP2A and several additional genes based on a genomic region of genes that were reported to be co-amplified with TOP2A, but that are not part of the smallest region of amplification of HER2 as defined by Marchio et al., 2008. So, TOP2A signature is the averaged sum of all the genes annotated in the region ranging from 35.37 Mb to 36.06 Mb of chromosome 17 (suppl. Data in Desmedt, 2008). The TOP2A signature was significantly associated with pCR in the ER-negative/HER2-positive tumors of patients receiving anthracycline-based treatment, but not in those of patients receiving combined taxane plus anthracycline treatment. The last two signatures have been previously described in 2.5.5. This model takes into consideration the heterogeneity of ER-negative tumors in terms of HER2 status by assessing their probability of belonging to the ER-negative/HER2-negative and the ER-negative/HER2-positive subtypes and by only considering the TOP2A signature for the latter, given the fact that the amplification of TOP2A and its predictive value was observed only in HER2-positive samples.

2.6 Statistical analysis

This paragraph illustrates common analyses of gene expression profiles such as differential expression (see 2.61 paragraph), the methods to classify samples and gene signatures (see 2.6.2 paragraph), survival (see 2.6.3 paragraph) and ROC analysis (see 2.6.4 paragraph).

2.6.1 Differential expression analysis

Identification of differentially expressed genes is a high level analysis and consists in finding genes that are differentially expressed between different conditions or phenotypes, e.g. two different tumor types. Differentially expressed genes are genes whose expression levels are associated with a phenotype of interest. There are different methods to identify such genes, including statistical tests. In this thesis a statistical test was used: Significance Analysis of Microarrays (SAM). The input for test is N gene expression measurements from a set of M microarray experiments, as well as a response variable from each experiment. The response variable is usually a label like "untreated", "treated" (either unpaired or paired). The null hypothesis is that the average gene expression is the same in the two populations.

Significance Analysis of Microarrays (SAM)

Tusher et al. (2001) introduced Significance Analysis of Microarrays (SAM) as a statistical technique for finding significant genes in microarrays. This technique aims to control the False Discovery Rate (FDR), which is the proportion falsely rejected null hypothesis among all rejected null hypotheses. SAM is available in the *samr* R package. Given two populations with *m1* and *m2* samples, SAM computes a statistic t_g for each gene *g*, measuring the strength of the relationship between gene expression and the response variable:

$$t_g = \frac{\mu_g^1 - \mu_g^2}{\sqrt{s_{gp}^2\left(\frac{1}{m_1} + \frac{1}{m_2}\right) + s_0}}$$

where μ_g^1 and μ_g^2 are the mean of gene *g* in population 1 and 2, respectively and s_{gp}^2 is the pooled variance of the gene in the two populations defined as:

$$s_{gp}^2 = \frac{(m_1 - 1)s_{g1}^2 + (m_2 - 1)s_{g2}^2}{m_1 + m_2 - 2}$$

where $s_{g1}^2$and s_{g2}^2 are the variances of gene *g* in population 1 and 2. SAM uses a regularized version of the t-statistics where it adds a positive constant s_0 to make the coefficient of variation of t_g approximately constant as a function of s_g. It then uses repeated permutations of the data to determine if the expression level of any genes is significantly related to the response. The cutoff for significance is determined by a tuning parameter Δ, chosen by the user based on the false positive rate. This tuning is achieved controlling the q-value or the false discovery rate for the gene list that includes that gene and all genes that are more significant. The user can also choose a fold change parameter, to ensure that called genes change at least of a pre-specified amount.

2.6.2 Signature quantification

These techniques have been used to calculate the continuous score of the signatures described in paragraph 2.5. In this thesis, we used two approaches to define a continuous signature score: the combined Z-score and Module score.

Combined Z-score
Given a list containing *n* genes and a sample *j*, a score can be defined as the sum of the standardized expression values (z-scores) of the *n* genes in the list:

$$Score_j = \sum_{i=1}^{n} \frac{x_{i,j} - \mu_i}{s_i}$$

Two groups of objects can be obtained setting a threshold on the score value. Since the score is centered on the mean, a common threshold is zero. If the list contains genes that represent the read-out of a pathway, a positive score will indicate that the pathway is active, while a negative will mean that the pathway is inactive (Fig 2.5).

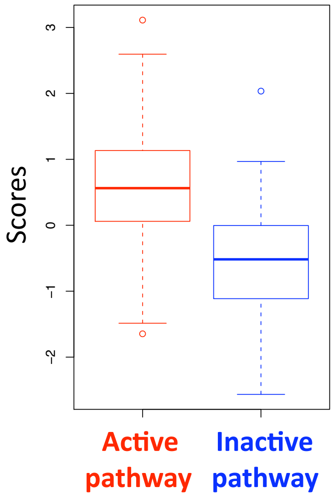


Figure 2.5: Distributions of score values in samples with activation/inactivation of a given pathway.

Module score

The other way to calculate a continuous signature score is the *module score* as described in Desmedt et al., 2008. For each sample, the signature was quantified as:

$$Score_j = \Sigma_i \omega_i x_i / \Sigma_i |\omega_i|$$

where x_i is the expression of a gene included in the set of genes of interest and ω_i is either +1 or -1 depending on the sign of the association under study. The signature scores have been assessed with the *sig.score* function of R (*genefu* package).

2.6.3 Survival analysis

The importance of statistics in biomedical research is that it allows drawing conclusions or making inferences based on a sample from a population rather than from a total population. Once a sample from a population is extracted, one or more hypotheses can be tested on it through the combined use of descriptive statistics and inferential statistics. The descriptive statistics describes the samples. For example, a sample of population can be described on the basis of its characteristics as age, sex or smoking habit. These factors can be next summarized, generally through the media or median or by measures of frequency (proportion of males, smoking rates, etc). Survival analysis is a part of inferential statistics and describes and quantifies time to event data. Once times to event data have been collected, the first task is to describe them and usually this is done graphically through a survival curve. There are several methods to estimate the survival curve. In epidemiology the most frequently used methods make no assumption on the distribution of the data (non-parametric methods). There are three non-parametric methods for describing time to event data, i.e., the Kaplan-Meier, the life table, and the Nelson-Aalen method. All survival analyses performed in this thesis are based on the Kaplan-

Meier method (Kalbfleisch and Prentice, 1980). The Kaplan-Meier method is a statistical tool that allows to construct the survival curves (the relationship between the probability of survival, on y-axis, and the observation time on the x-axis) and to measure the hazard rate. In cohort study each patient is described through two terms: *failure* is used to define the occurrence of the event of interest and *survival time* specifies the length of time taken for failure to occur, i.e., the survival of tumor patients after surgery. Obviously, the survival time for patients without the event of interest corresponds to that between the beginning of study and the end of the observation. Censored patients are those who don't incur in the event during the observation period or those who survive until the end of the observation or leaving the study before the end of it for various reasons (e.g., patients lost to follow-up, moved to another center, or died from other causes). In all cases, censored patients remain in the analysis until fixed data on their health and their presence are available. In the survival curve, the curve has a step down whenever the patient has the event of interest, while it continues when a patient is censored, and has a sign (normally + sign) when the patient was still alive at the end of his follow-up (Figure 2.6). The part of the curve after censoring of the first patient is only an estimate of survival for the group rather than the actual survival, which is not yet known since the censored patients are still alive at the time of analysis. If the analysis were done later (and often results from clinical studies are updated with increased follow-up), then the information that a previously censored patient had continued to survive or had died at some point would be incorporated into the curve. Since everyone eventually dies and all patients of the study will have died, the survival curve for the group will be precisely known. Until that, the curve is only an estimate.

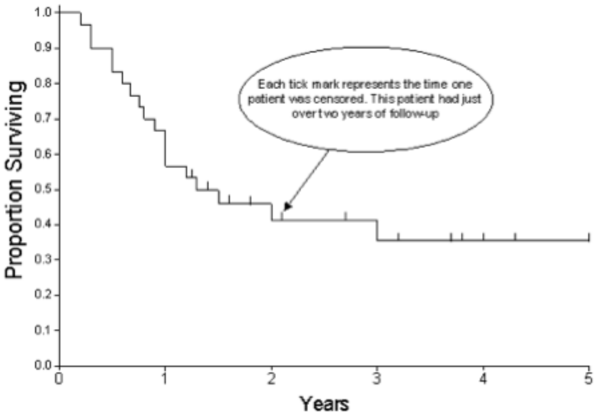


Figure 2.6: Example of a censored curve with tick marks

When a patient is censored the sample size of patients at risk is reduced by one after the time of censorship; this always reduces the reliability of the curve, so the more patients are censored and at earlier time they are censored the more unreliable the curve is. The end of the curve is most affected, because each censored patient reduces the reliability of the curve from that point forward. In these curves the number of intervals is dictated by times when the event of interest takes place, showing the number remaining *at risk* at each interval. The number at risk at any interval is the number of patients who are still alive and whose follow-up extends at least that far into the curve. The percentage surviving at the start of any interval is equal to the probability of survival multiplied for each one of the preceding intervals. In fact, the aim is to find a way to account for censored patients and to remove them from the curve at the time their follow-up ends. So when a patient dies, the survival for the interval ending with his death is calculated according to the number remaining at risk at the time of death.

It is frequently of interest to compare the survival of two group of study. The most commonly used test for comparing survival distributions is the *log-rank test* (Harrington and Fleming, 1982), also known as the *Mantel-Cox test*. This test takes each time point when a failure event occurs and creates a 2×2-table showing the number of deaths and the total number of subjects under follow-up. For each table, the observed deaths in each group, the expected deaths and the variance of the expected number are calculated. These quantities are summed over all tables to yield a χ^2 statistic with one degree of freedom. This test also produces the observed to expected ratio of each group, i.e., the ratio between the number of deaths observed during the follow-up and the expected number of deaths under the null hypothesis that the survival curve for that group would be the same as that for the combined data. Survival analysis, Kaplan-Meier curves, and the multivariate analysis have been performed using *survival* package of R. Kaplan-Meier plots were drawn using the *km.coxph.plot* function of *survcomp* package. Log-rank test has been performed with *surv_test* function of *coin* package.

2.6.4 ROC analysis

The receiver operating characteristic (ROC) curve is a standard technique for visualizing, organizing and selecting classifiers based on their performance (Swets et al, 2000). Given a classifier and a condition, there are four possible outcomes. If the condition is positive and it is classified as positive, it is counted as a *true positive*; if it is classified as negative, it is counted as a *false negative*. If the condition is negative and it is classified as negative, it is counted as a *true negative*; if it is classified as positive, it is counted as a *false positive*. Given a classifier and a set of conditions (the test set), a two-by-two confusion matrix (also called a *contingency table*) can be constructed representing the dispositions of the set of instances. This matrix forms the basis for many common metrics. Figure 2.7 shows a *contingency table* and equations of several common metrics that can be calculated from it.

		Condition		
		Condition positive (P)	Condition negative (N)	
Classifier outcome	Classifier outcome positive	True positive (TP)	False positive (FP)	Positive predictive value (Precision) = $\frac{TP}{TP + FP}$
	Classifier outcome negative	False negative (FN)	True negative (TN)	Negative predictive value = $\frac{TN}{FN + TN}$
		Sensitivity = $\frac{TP}{P}$	Specificity = $\frac{TN}{N}$	

Figure 2.7: *Contingency table* and common performance metrics calculated from it.

A ROC curve is a two-dimensional graph in which the sensitivity is plotted on the Y axis and 1 - specificity is plotted on X axis. The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). The (0,1) point is also called a *perfect classification*. A completely random guess would give a point along a diagonal line (the so-called *line of no-discrimination*) from the left bottom to the top right corners (Figure 2.8).

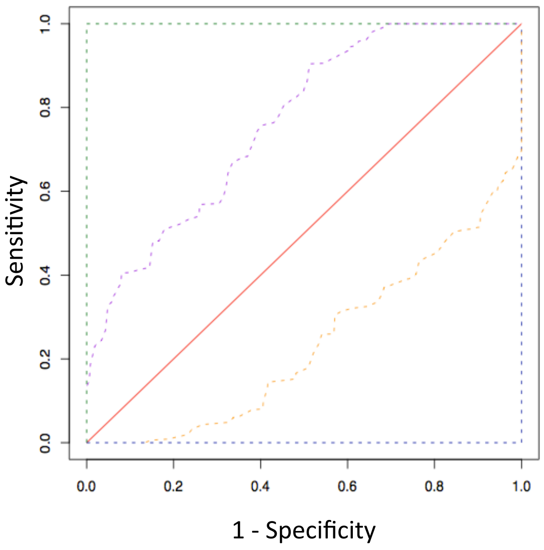


Figure 2.8: ROC curve graph.

To compare classifiers we may want to reduce ROC performance to a single scalar value representing expected performance. A common method is to calculate the area under the ROC curve, abbreviated AUC (Bradley et al., 1997; Hanley and McNeil, 1982). Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1. The concordance-index is a generalization of the area under the ROC curve (AUC), therefore it measures how well the classifier discriminates between different responses, i.e., is your predicted response low for low observed responses and high for high observed responses. So concordance-index > 0.5 implies a good prediction ability, equal to 0.5 implies no predictive ability (no better than random guessing), and < 0.5 implies "good" anti-prediction (worse than random, but if you flip the prediction direction it becomes a good prediction). In this thesis, using the receiver operating characteristic (ROC) analysis, we assessed the ability of any single gene/gene signature score or their linear combination to discriminate patients with pathologic complete response from patients with residual disease. We calculated the area under the curve (AUC) to assess the prediction performance of any score (*rocr* package). AUC was estimated through the *concordance.index* (*survcomp* R package).

Performance of a classifier

To classify a patient as a putative responder or as a resistant one, we first need to determine the appropriate threshold for the continuous score of the classifier. Metrics for the classifier score as the positive (PPV) and negative predictive values (NPV), sensitivity (SENS), specificity (SPEC) (see Figure 2.7) were determined at the threshold that maximizes the Youden Index (SPEC + SENS − 1) in a cohort of interest. Youden's index (*J*), defined as the maximum vertical distance between the ROC curve and the diagonal line, serves as another global measure of overall classifier accuracy and can be used in choosing an optimal cut-point. These metrics can be represented using forest plot. The word originated from the idea that graph had a forest of lines. The horizontal line corresponds to exact 95% confidence intervals (CIs) of each classifier. Point estimates are displayed as blue squares (Figure 2.9).

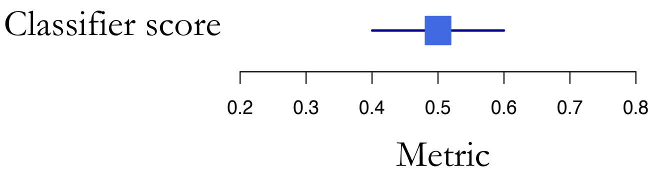


Figure 2.9: Forest plot.

2.7 Enrichment analysis

Enrichment analysis consists in identifying statistically significant associations between microarray experiments and gene sets. This approach requires expression data that represent the transcriptional effect of an event occurring at the molecular level (e.g., overexpression/silencing of a gene, treatment with pharmaceutical compounds, activation of a regulatory axis) and the use of statistical tools to evaluate the significance of the overlap between this effect and gene sets representing signaling pathways, gene ontologies or other transcriptional characteristics. Approach to perform enrichment analysis: i) first identify a list of differentially expressed genes (for instance using as SAM) and then apply a Fisher's exact test to determine the enrichment of specific gene sets among the differentially expressed genes.

2.7.1 Fisher's exact test

The Fisher's exact test (Fisher, 1925) can be used to conduct an over-representation analysis to assess the statistical association between two nominal variables that result from classifying objects in two different ways. The test uses a 2×2 contingency table and verifies if the data observed in the table are compatible with the null hypothesis that the relative proportions of one variable are independent of the second variable. The probability of getting the observed data under the null hypothesis that the proportions are the same follows the hyper geometric distribution. The hyper geometric distribution is a discrete probability distribution that describes the probability of k successes in n draws from a finite population without replacement. Given an urn with two types of balls, black ones and white ones, if the variable N describes the number of all balls in the urn and m describes the number of white balls, then N−*m* corresponds to the number of black balls. Given the contingency table indicated in Table 2.36, X is the random variable whose outcome is *k*, the number of white balls drawn in the experiment.

Table 2.36: example of a contingency table.

	drawn	**not drawn**	**total**
White balls	*k*	*m* - *k*	*m*
Black balls	*n* - *k*	N + *k* - *n* - *m*	N - *m*
Total	*n*	N - *n*	N

The probability of drawing exactly k white balls can be calculated as:

$$P = (X = k) = f(k; N, m, n) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

In order to calculate the significance of the observed data, i.e. the total probability of observing data as extreme or more extreme if the null hypothesis is true, the values of *p* for this table and all tables with a more extreme configuration have to be calculated and added together. This gives a one-tailed test; for a two-tailed test we must also consider tables that are equally extreme but in the opposite direction. In this thesis, the one-sided Fisher's exact test was used to assess the significance of the overlap between two lists of genes, e.g. overexpressed genes and genes belonging to a signaling pathway. Given *m* up-regulated probesets and *n* probesets in a predefined pathway signature, the probability of observing an overlap of k probesets for that signature, under the hypothesis that the probesets were picked out randomly from the N total probesets of the microarray, is given by the hypergeometric distribution. Testing more signaling pathways, a 2×2 contingency table was built for each pathway recording the relation between genes in the pathway signature and overexpressed genes. For any given signature, the significance of the observed overlap *k* (p-value) is computed as the sum of the probabilities for all possible contingency tables with an overlap greater than or equal to *k*. The null hypothesis is then rejected if the p-value is smaller than a predetermined threshold. Considering that multiple signatures were tested, p-values were finally adjusted for false discovery rate (FDR) using Benjamini-Hochberg (BH) correction. The over-representation analysis has been conducted using the *phyper* function of the R stats package. The p-value threshold has been set to 0.05 and p-values adjusted using the *p.adjust* function of the R *stats* package.

2.7.2 Gene Sets

The Molecular Signature Database (MSigDb) is a publicly accessible collection of curated gene sets that is maintained by the GSEA team (http://www.broadinstitute.org/gsea/msigdb/index.jsp; Subramanian et al., 2005).
The MSigDB gene sets are divided into seven major collections:
C1: positional gene sets for each human chromosome and each cytogenetic band;
C2: curated gene sets from online pathway databases, publications in PubMed, any knowledge of domain experts;
C3: motif gene sets based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat and dog genomes;
C4: computational gene sets defined by mining large collections of cancer-oriented microarray data;
C5: GO gene sets consist of genes annotated by the same GO terms;
C6: oncogenic signatures defined directly from microarray gene expression data from cancer gene perturbations;
C7: immunologic signatures defined directly from microarray gene expression data from immunologic studies.

In this thesis, we used a subset of the C2 collection (version 4.0), i.e. those gene sets derived from the BioCarta pathway database (http://www.biocarta.com/genes/index.asp). This collection contains 217 gene sets describing metabolic and signaling pathways. Moreover, we added to the BioCarta gene sets several other pathways derived from the literature (Table 2.37) and signatures characterizing normal mammary (hNMSC), breast cancer (IGS, CD44high), and embryonic (ES1, ES2, ES-like) stem cells listed in Table 2.38.

Table 2.37: List of gene sets representing signaling pathways obtained from literature.

Signature name	# probesets	Reference
β-catenin	13	Bild et al., 2006
E2F3	258	Bild et al., 2006
ERBB2	30	Mackay et al., 2003
HIFs	54	Montagner et al., 2011
H-Ras	262	Bild et al., 2006
Mutant-p53	165	Miller et al., 2005
NCID	135	Mazzone et al., 2010
NF-KB	223	Park et al., 2007
Notch	418	Mazzone et al., 2010
Src	73	Bild et al., 2006
STAT3	12	Alvarez et al., 2005
TGF-β-a	120	Padua et al., 2008
TGF-β-b	51	Adorno et al., 2009
TGF-β-c	170	Adorno et al., 2009
WNT	13	Di Meo et al., 2009
WNT/TCF4	47	van de Wetering et al., 2002
YAP.TAZ	619	Zhang H et al., 2009
YAP.TAZ conserved	93	Zhao et al., 2008 Dong et al., 2007 Ota and Sasaki, 2008
Core_Human	126	Bild et al., 2006
Myc_Human	605	Bild et al., 2006
MYC	53	Bild et al., 2006
SHARP1	28	Montagner et al., 2012
TenGene	29	Girardini et al., 2011

Table 2.38: List of staminal signatures.

Signature name	Description	# probesets	Reference
hNMSC (Staminal)	Human normal mammary stem cells	294	Pece et al., 2010
CD44high	CD44+ breast cancer cells	39	Shipitsin et al., 2007
ES1	Embryonic stem cells 1	328	Ben-Porath et al., 2008
ES2	Embryonic stem cells 2	40	Ben-Porath et al., 2008
ES-like	Embryonic stem cells-like	731	kim et al., 2010
IGS	Invasive gene signature	97	Liu et al., 2007

Chapter 3

Results

Paragraph 3.1 contains a description of the re-organization of the breast cancer datasets; paragraph 3.1.1 describes the clinical covariates of meta dataset; paragraph 3.2 describes the used methods for classification in the different molecular subtypes and finally, paragraph 3.3 describes the definition of two large meta-cohorts; in particular, the *prognostic meta-cohort* and its application are described in paragraph 3.3.1 and the *predictive meta-cohort* and its application are described in paragraphs 3.3.2.

3.1 Breast cancer data collection

The survey for gene expression profiles of breast cancer samples analyzed using Affymetrix microarrays and for which raw data and patients' clinical annotations were publicly available returned 4640 samples collected in 27 major datasets (see Table 2.1 and paragraphs 2.2.1-2.2.27 for details). Prior to analyzing the gene expression data, the content of all breast cancer datasets has been manually verified and up-dated. Indeed, a finer inspection of cohort descriptions, in primary article and supplementary information, and of GEO meta-data seemed to indicate that, in some cases, the same sample was included in more than a study and submitted in more than one GSE. As such, we re-organized all datasets eliminating multiple copies of the same sample and designing a final database in which any sample appears only once. Moreover, we changed the name of the original dataset and named the new datasets after the medical center where patients were recruited. In summary, the studies of Table 2.1 have been changed as follows:

- *Stockholm* dataset has been fully confirmed, except for the name changed to *KI_Stockholm* (Karolinska Institutet Stockholm);
- *EMC-286* and *EMC-58* were merged to create *EMC-344* (Erasmus Medical Center);
- *MSK* dataset has been fully confirmed, except for the name changed to *MSKCC* (Memorial Sloan-Kettering Cancer Center);
- *Uppsala-Miller*, *Ivshina-Miller*, and *Loi* datasets (GSE3494, GSE4922, and GSE6532) includes samples derived from 3 cancer centers, i.e., Uppsala University Hospital, for GSE3494 and GSE4922, and John Radcliffe Hospital in Oxford, Guys Hospital in London and, again, Uppsala University Hospital for GSE6532. Moreover, a comparison of the hybridization dates on the CEL files of GSE3494 and GSE4922 and of the patients' clinical information revealed that, despite being deposited twice, the two series are identical. As such, these 3 datasets have been split into:
  - *KI_Uppsala* comprising all 258 unique patients of the Uppsala University Hospital;
  - *OXF* composed of the 178 samples collected at the John Radcliffe Hospital in Oxford and formerly part of GSE6532;
  - *GUY* composed of the 87 samples from the Guys Hospital in London and formerly part of GSE6532 and, as explained later, of 77 samples from the former *Tamoxifen* study;
- *Sotiriou* dataset has been eliminated since samples of this series are all included in GSE6532;
- *Tamoxifen* dataset has been added to *GUY* cohort since all patients were recruited at the Guys Hospital in London;
- *Desmedt* dataset has been fully confirmed, except for the name changed to *TRANSBIG* (after the consortium of cancer centers where samples have been collected);
- *Schmidt* datasets has been fully confirmed, except for the name changed to *Mainz* (Johannes Gutenberg University in Mainz);
- *Veridex* dataset has been fully confirmed;
- E-TABM-158 and GSE7378 were merged to create UCSF since all patients were recruited at the University of California, San Francisco (173 samples). Moreover, a comparison of the hybridization dates on the CEL files of E-TABM-154 and GSE7378 and of the patients' clinical information revealed that, 17 samples were deposited twice for a total of 166 unique samples out of 173 samples;

- GSE16446 was re-named as IJB_TOP *(*Institut Jules Bordet /Trial of Principle) since all patients were recruited at Institut Jules Bordet;
- *GSE19615* was re-named as *US_NCI* since all patients were recruited at US National Cancer Institute;
- *IPC-GSE21653* was re-named as *CRCM* since all patients were recruited at Centre de cancérologie de Marseille;
- *GSE20685* was re-named as *KOOF* since all patients were recruited at Koo Foundation SYS Cancer Center;
- *GSE31519* was re-named as *Goethe* since all patients were recruited at Goethe-University, Frankfurt;
- *GSE22093* was re-named as *MDACC/IGR* (M.D. Anderson Cancer Center/Institut Goustave Russy) and comprises 103 samples, 36 of which are included in *GSE20271*;
- *GSE20271* was re-named *MDACC* and comprised 178 samples, 78 of which are included in GSE25066;
- *GSE20194* is largely included in *MDACC* cohort of GSE25066 (187 out of 230 samples; other four samples are included in GSE20271); the remaining 39 samples compose the cohort and it was re-named as *MDACC MAQC-II;*
- *GSE25066* includes samples derived from 4 cancer centers, i.e., *I-SPY-1* (Investigation of Serial Studies to Predict Your Therapeutic Response With Imaging and Molecular Analysis), *LBJ_INEN_GEICAM (*Lyndon B. Johnson Hospital, Instituto Nacional de Enfermedades Neoplásicas, and Grupo Español de Investigación en Cáncer de Mama), *USO-02103* (US Oncology) and *MDACC* (M. D. Anderson Cancer Center, Houston). Moreover, a comparison of the hybridization dates on the CEL files of GSE25066 and GSE20271 and GSE20194 and also of the patients' clinical information revealed that, despite being deposited twice, some samples are identical. As such, these four datasets have been split into:
  - *I-SPY-1* comprising 83 samples;
  - *LBJ_INEN_GEICAM* comprising 58 samples;
  - *MDACC* comprising 313 samples;
  - *USO-02103* was entirely included in GSE23988;
- *GSE23988* was re-named *USO-02103* and it is composed of 54 samples included in *USO-02103* cohort of GSE25066 and 61 from *GSE23988.* Twenty samples from *GSE23988* were removed, because they were deposited twice in the two series;
- *GSE32646*, *GSE19697* and *GSE18728* were re-named as *Osaka, St. Louis* and *UW* respectively since all patients were recruited at the Osaka University,

Washington University School of Medicine (St. Louis) and University of Washington (Seattle), respectively.

The re-organization of the downloaded datasets returned 3661 unique samples distributed in 27 cohorts (Table 3.1).

Table 3.1: Re-organized datasets of breast cancer expression profiles.

Cohort	Affymetrix platform	Samples	Data source	References
KI_Stockholm	HG-U133 A	159	GSE1456	Pawitan et al., 2005
EMC-344	HG-U133A	344	GSE2034 GSE5327	Wang et al., 2005; Minn et al., 2007
MSKCC	HG-U133A	82	GSE2603	Minn et al., 2005

KI_Uppsala	HG-U133A	253	GSE3494
			GSE4922
			GSE6532
				Loi et al, 2008;
				Ivshina et al, 2006;
				Miller et al, 2005

OXF	HG-U133A	178	GSE6532	Ivshina et al., 2006
TransBIG	HG-U133A	198	GSE7390	Desmedt et al., 2007
Mainz	HG-U133A	200	GSE11121	Schmidt et al., 2008

Veridex	HG-U133A	136	GSE12093	Zhang et al., 2009;
				Loi et al., 2007;

GUY	HG-U133 Plus2.0	164	GSE6532 GSE9195	Loi et al., 2008; Loi et al., 2010

UCSF	HG-U133AAofAV2	166	E-TABM-158
			GSE7378
				Merritt et al., 2008;
				Zhou T et al., 2007;
				Yau C et al., 2008

IJB_TOP	HG-U133 Plus2.0	114	GSE16446	Desmedt Cet al., 2011;
				Li Y et al., 2010;
				Juul N et al., 2010

US_NCI	HG-U133 Plus2.0	115	GSE19615	Li Y t al., 2010
CRCM	HG-U133 Plus2.0	252	GSE21653	Sabatier R et al., 2011
KOOF	HG-U133 Plus2.0	327	GSE20685	Kao KJ et al., 2011

Goethe	HG-U133A	64	GSE31519	Rody A. et al., 2011;
				Karn T. et al., 2011;

MDACC_IGR	HG-U133A	61	GSE22093	Iwamoto T et al., 2011;
MDACC_GSE25066	HG-U133A	313	GSE25066 GSE20194	Hatzis C. et al., 2011; Popovici V. at al., 2010; Shi L. et al., 2010
I-SPY-1	HG-U133A	83	GSE25066	Hatzis C. et al., 2012
LBJ_INEN_GEICAM	HG-U133A	58	GSE25066	Hatzis C. et al., 2012
USO-02103	HG-U133A	95	GSE23988	Iwamoto T et al., 2011; Hatzis C. et al., 2012
MDACC_GSE20271	HG-U133A	100	GSE20271	Tabchy A. et al., 2010
MDACC MAQC-II	HG-U133A	39	GSE20194	Popovici V. at al., 2010; Shi L. et al., 2010
Osaka	HG-U133 Plus2.0	115	GSE32646	Miyake T et al., 2012
UW	HG-U133 Plus2.0	21	GSE18728	Lin Y et al., 2010
St.Louis	HG-U133 Plus2.0	24	GSE19697	Korde LA et al., 2010

Since raw data (.CEL files) were available for all samples, the integration and normalization of gene expression signals has been obtained applying the *virtual chip* procedure (see paragraph 2.2.3). Gene expression signal have been generated using the robust multi-array average procedure RMA. Specifically, intensity levels have been background adjusted, normalized using quantile normalization, and log2 expression values calculated using median polish summarization.

3.1.1 Analyses of outcome and clinic-pathological covariates of meta-dataset

Combining data from independent but related studies is at the base of *meta-analysis*. Meta-analysis strategies include an important step: data combination. To homogenize the clinic-pathological and outcome information of the various cancer datasets in order to combine the data in a unique meta-dataset, the variables have been carefully re-defined considering the clinical annotations of any re-organized study (Tables 2.2 - 2.28). The two types of events, i.e., *metastasis* and *survival* were defined. *Metastasis* is associated to the metastatic spread and includes the following descriptions in the studies of Table 3.1:

- recurrence free survival;
- metastasis free survival;
- distant metastasis free survival or distant recurrence (and subtypes at different districts as lung, bone, and brain distant metastasis free survival/distant recurrence);
- time to distant metastasis.

Survival is associated to death because of cancer and includes the following descriptions in the studies of Table 3.1:

- overall survival;
- disease free survival;
- disease specific survival.

All clinic-pathological variables such as tumor size (T), pathologic lymph nodes status (N), stage of tumor, and grade of tumor were re-organized as the American Joint Committee on Cancer (AJCC) staging system; moreover, ER, PR, and HER2 status (IHC), p53 status, response to neo-adjuvant chemotherapy, type of chemotherapy were labeled as dichotomous variables as described in Table 3.2.

Table 3.2: Re-defined clinic-pathological covariates of meta-cohort of breast cancer samples.

Clinic-pathological covariates	Class description
Age	<40 40-60 >60

Tumor size (T)	T1: Tumor 2.0 cm or less in greatest dimension T2: Tumor more than 2.0 cm but not more than 5.0 cm in greatest dimension T3: Tumor more than 5.0 cm in greatest dimension T4: Tumor is any size, but has spread beyond the breast tissue to the chest wall and/or skin
Pathologic lymph nodes status (N)	N0 : Axillary and other nearby lymph nodes do not have cancer N1 : Micrometasases (very small clusters of cancer) or 1–3 axillary lymph nodes have cancer N2 : 4–9 axillary lymph nodes have cancer or internal mammary nodes have cancer, but axillary lymph nodes do not have cancer N3 : 10 or more axillary lymph nodes have cancer or infra- or supra-clavicular nodes have cancer
Stage of tumor	1 = Early breast cancer 2 = Early/locally breast cancer 3 = Locally advanced breast cancer 4 = Metastatic breast cancer
Estrogen receptor - ER (IHC)	+ = positive - = negative
Progesteron receptor - PR (IHC)	+ = positive - = negative
Epidermal growth factor receptor - HER2 (IHC or FISH)	+ = positive - = negative
Lymph node status (LN)	+ = positive - = negative
Grade of tumor	G1 = well differentiated state, like normal breast tissue G2 = moderate differentiated state G3 = poor differentiated state, few similarities to normal breast tissue

P53 status	WT = Wild-type status
	MUT = Mutant status

Metastasis *Survival*	associated to the metastatic spread associated to death because of cancer

Neo-adjuvant chemotherapy type	A = Anthracycline - based
	AT = Taxane plus anthracycline - based

Response neo-adjuvant chemotherapy	pCR = complete response
	RD= residual disease

Since samples were derived from different studies, the samples in a specific dataset have clinic-pathologic variables of interest for the objective of study, whereas samples of other datasets could have missing information (*unknown data*) about these considered variables. Unfortunately, according the variables derived from IHC assay (e.g., estrogen and progesterone receptors, or epidermal growth factor receptor) we have sometime only qualitative and not quantitative information; indeed, IHC assay could change among different studies and experimental conditions; so, we can know only the positivity or negativity of a sample about a specific variables, but not how many is positive or negative than others. About p53 status, we have only a dichotomous variable (wild-type or mutant protein), but we don't have information about the mutation type (i.e. missense, non-sense, or frame shift). We don't have information about race of patients, but only the name of hospitals where they were recruited. Table 3.3 describes the clinical and pathological characteristics of the samples included in the meta-dataset.

Table 3.3: Distribution of clinical and pathological characteristics of the samples included in the meta-dataset.

Clinical variables	Patients n=3661
Age, years	
<40	385 (10%)
40-60	1686 (46%)
>60	870 (24%)
T stage	
T0/T1	919 (25%)
T2	1486 (41%)
T3/4	370 (10%)
Unknown	150 (4%)
Grade	
1	351 (10%)
2	1047 (29%)
3	1128 (31%)
ER status	
Positive	1839 (50%)
Negative	1073 (29%)
Unknown	749 (21%)
p53 status	
Mutated	187 (5%)
Wild type	409 (11%)
Unknown	3065 (84%)
PR status	
Positive	1127 (31%)
Negative	1041 (28%)
Unknown	1493 (41%)
LN status	
0	241 (7%)
1	327 (9%)
2	97 (3%)
3	57 (2%)

HER2 status	
Positive	275 (8%)
Negative	1482 (40%)
Unknown	1904 (52%)
Response to neo-adjuvant therapy	
pCR	209 (6%)
RD	770 (21%)
Type of neo-adjuvant therapy	
Anthracycline	244 (7%)
Anthracycline + Taxane	758 (21%)
Metastasis within 5 years	
Yes	652 (18%)
No	756 (21%)
Follow-up all cases (years)	
Median (CI)	7.78 (4.7, 10.7)
Follow-up still living (years)	
Median (CI)	9.07 (6.2, 11.2)

3.2 Molecular subtypes identification

We used the proprietary AIRC 5X1000 dataset as a "gold standard", since we have the correct molecular subtypes characterized by IHC assay, to chose the best molecular subtype classifier among *PAM50* model (see 2.5.1 paragraph) and the three SCM models (*SCMGENE*, *SCMOD1*, and *SCMOD2*) described in 2.5.2 paragraphs.

AIRC 5X1000 dataset's samples were hybridized on Illumina Whole Genome DASL HT-12 arrays, while the meta-cohort's samples on Affymetrix platforms. Prior the analysis, the PAM50 gene-list was refined so that only genes with a corresponding probe on these two types of platform were used for classification, in order to compare the concordance of the models on different platform. As a result, ANLN, CDCA1, CXXC5, GPR160, TMEM45B and UBE2T were not included in the classification. PAM50 gene list was reduced to a list of 43 common genes out of 50 (Table 3.4). For genes with more than one probe set, all probe sets were median gene centered prior to classification.

Table 3.4: PAM50 reduced gene list.

Gene symbols	Entrez gene ID
ACTR3B	57180
BAG1	573
BCL2	596

BIRC5	332
BLVRA	644
CCNB1	891
CCNE1	898
CDC20	991
CDC6	990
CDH3	1001
CENPF	1063
CEP55	55165
EGFR	1956
ERBB2	2064
ESR1	2099
EXO1	9156
FGFR4	2264
FOXA1	3169
FOXC1	2296
GRB7	2886
KIF2C	11004
KNTC2	10403
KRT14	3861
KRT17	3872
KRT5	3852
MAPT	4137
MDM2	4193
MELK	9833
MIA	8190
MKI67	4288
MLPH	79083
MMP11	4320
MYBL2	4605
MYC	4609
NAT1	9
PGR	5241
PHGDH	26227
PTTG1	9232
RRM2	6241
SFRP1	6422
SLC39A6	25800
TYMS	7298
UBE2C	11065

The same approach was applied for the *SCMOD1* and *SCMOD2* models gene lists. *SCMOD1* is based on 726 genes related to ER, HER2, and AURKA signalings (Desmedt, 2008) divided in 469, 28, and 229 gene modules, respectively; *SCMOD2* on 663 genes (Wirapati et al., 2008) divided in 288, 20 and 353 gene modules, respectively (Table 3.5).

Table 3.5: *SCMOD1* and *SCMOD2* reduced gene modules.

Models name	**Original gene modules**			**Reduced gene modules**		
	ESR1	**ERBB2**	**AURKA**	**ESR1**	**ERBB2**	**AURKA**
SCMOD1	469	28	229	284	24	216
SCMOD2	288	20	353	193	11	255

AIRC 5X1000 dataset comprises a *pilot study* of 48 breast cancer samples, subdivided among well characterized 11 luminal A, 12 luminal B, 13 HER2+ and 12 basal-like subtypes. Finally, we assessed the best molecular subtype classifier calculating its specificity, sensitivity, accurancy, and the negative and positive predictive value to classify basal-like samples. Comparing the different models, *SCMOD2* resulted the best classifier with almost 98% of specificity, 92% of sensitivity, 96% of accurancy, and 98% of negative predictive value.

Starting from these results, we classified the meta-cohort of breast cancer samples using *SCMOD2* model. This model uses the same molecular subtype nomenclature described for the first time by Perou and collaborators (Perou et al., 2000): ER-/HER2-, HER2+, and ER+/HER2- low and high proliferation tumors which correspond, respectively, to basal-like, HER2-enriched, and luminal A and B (Figure 3.1 and 3.2).

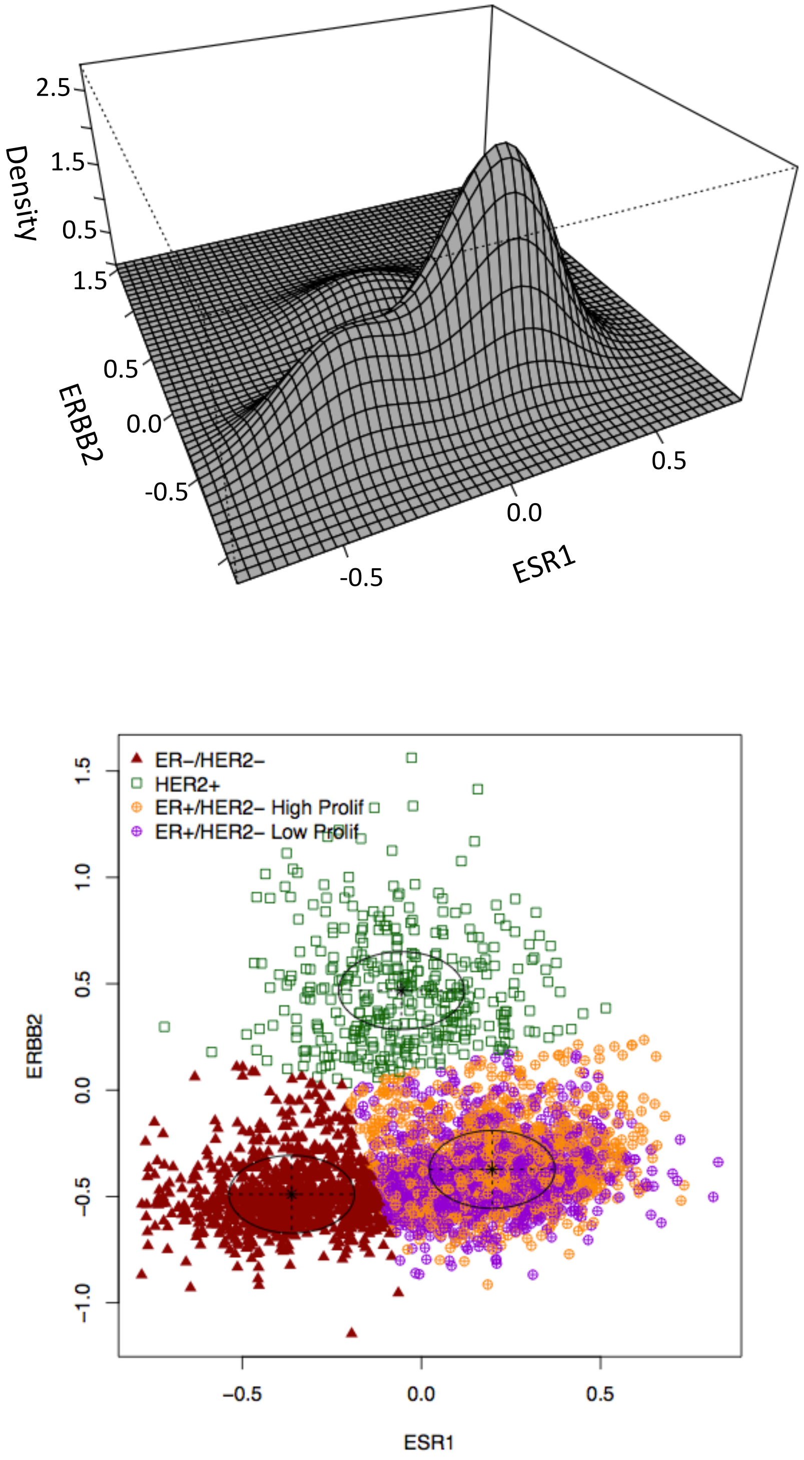


Figure 3.1: Molecular subtypes in meta-cohort using *SCMOD2*. Upper panel: Density distribution of the mixture of three gaussians fitted for the subtype clustering model. Bottom panel: Scatter-

plot. Each subtype is represented by a different color and symbol. The ellipses shown are the multivariate analogs of the SDs of the Gaussian of each cluster.

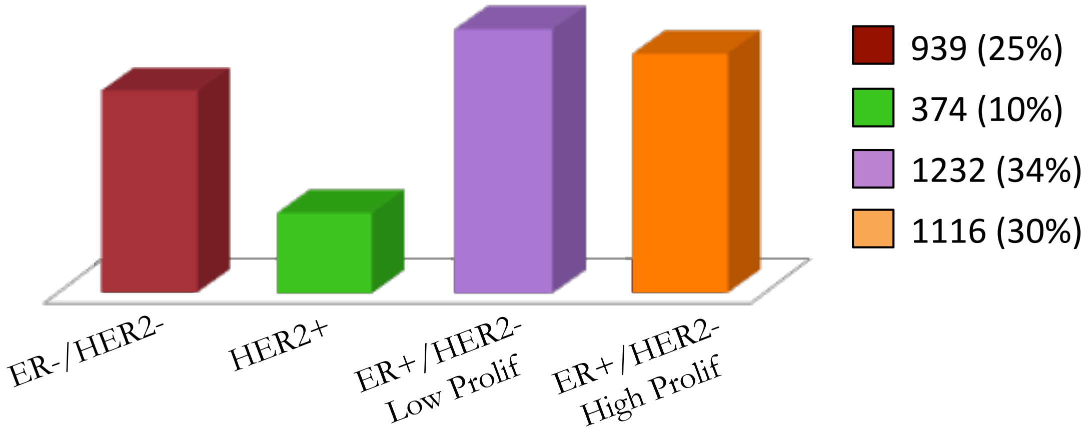


Figure 3.2: Distribution of molecular subtypes in breast cancer meta-cohort.

3.3 Definition of meta-cohorts

The *meta-dataset* allowed the definition of two large meta-cohorts (Figure 3.3): one includes samples with clinical outcome information, called *prognostic meta-cohort* (see 3.3.1 paragraph) and the other ones is composed by neo-adjuvant chemotherapy (NAC) treated samples with available chemotherapy response called *predictive meta-cohort* (see 3.3.2 paragraph).

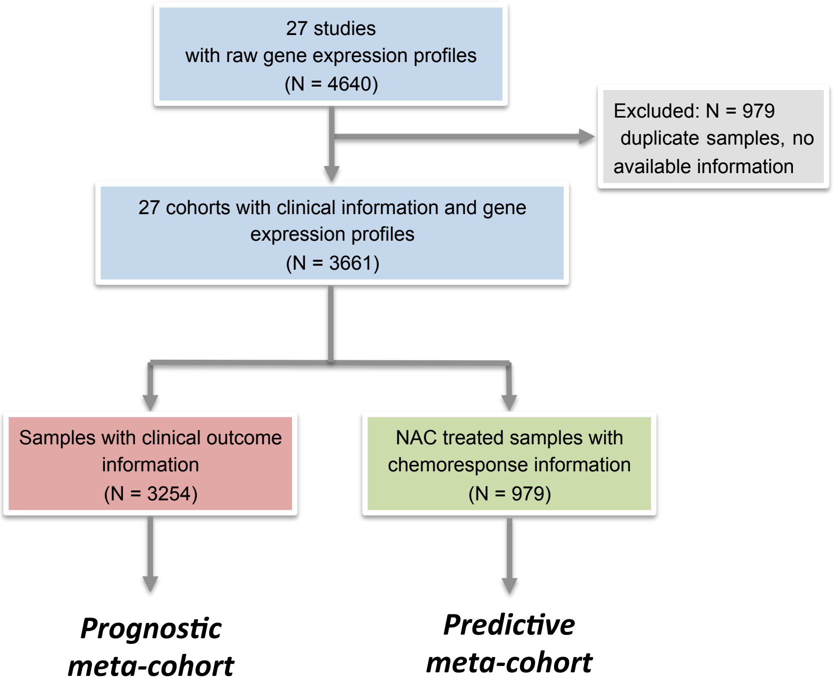


Figure 3.3: Flow chart of meta-analysis

3.3.1 Prognostic meta-cohort

In the *prognostic meta-cohort* we included 21 datasets with available clinical outcome information derived by the complete meta-cohort of Table 3.1 (i.e., CRCM, EMC-344, Goethe, GUY, I-SPY-1, LBJ_INEN_GEICAM, USO-02103, MDACC, IJB-TOP, KI_Stockholm, KI_Uppsala, KOOF, Mainz, MSKCC, OXF, TransBIG, UCSF, US_NCI, Veridex) for a total of 3254 primary tumors.

Application of prognostic meta-cohort: *Control of Prolyl-isomerase Pin1 in breast normal and cancer stem cells*

Breast cancer is the most frequently diagnosed cancer and the leading cause of cancer mortality in females worldwide (Siegel et al, 2011). Despite advances in diagnosis and treatment, a significant percentage of breast cancer patients still die, due to the development and dissemination of metastases (Steeg & Theodorescu, 2008). It is increasingly acknowledged that a subpopulation of cancer cells, termed cancer stem cells (CSCs) play a major role in cancer growth, metastasis formation and chemo resistance (Dean et al, 2005; Stingl & Caldas, 2007; Visvader & Lindeman, 2012). Like their normal counterpart, CSCs are able to self-renew and maintain a reservoir of cancer-initiating cells that may produce a more differentiated progeny of cells and contribute to intratumor heterogeneity (Stingl & Caldas, 2007). This evidence has been observed for breast cancers, where it has been shown that poorly differentiated, more aggressive tumors (histological grade 3) have an increased number of CSCs than well-differentiated (histological grade 1) tumors (Pece et al, 2010). Considerable similarities are found between normal and CSCs regarding the molecular pathways and stem cell factors that determine the undifferentiated state of these cells, which suggested that CSCs originate from the transformation of adult tissue stem cells or from more differentiated progenitors that have acquired self-renewal ability (Reya et al, 2001; Ben-Porath et al, 2008; Visvader & Lindeman, 2012). Several studies indicated that oncogenic activation of pathways involved in the regulation of normal stem cells, such as Notch, WNT, SHH, RTKs, and PI3K/AKT among others, might be involved in self-renewal properties and aggressive features of CSCs (Polyak & Weinberg, 2009; Thiery et al, 2009; Visvader & Lindeman, 2012). However, how these signaling networks govern CSCs still remains to be elucidated. One appealing candidate as a fine-tuner of stem cell traits might be the prolyl-isomerase Pin1. This unique enzyme catalyzes the *cis/trans* conversion of specific motifs composed by phosphorylated Serines or Threonines preceding a Proline in certain proteins, thereby inducing conformational changes required for the full activity and cross-talk of a plethora of signaling pathways (Liou et al, 2011). The discovery of Pin1-catalyzed *cis/trans* isomerization of phospho-Ser/Thr-Pro motifs revealed a post-phosphorylation mechanism critical for several biological processes involved in physiology and disease (Lu & Zhou, 2007; Yeh & Means, 2007). In particular, Pin1 is required for full activity and cross-talk of a variety of oncogenic pathways in breast and other cancers (Wulf et al, 2005), acting as an amplifier of phosphorylation signals. Of note, deregulated levels of Pin1 have been shown to disrupt cellular polarity of breast epithelial cells (Ryo et al, 2002) and found associated to high tumor grade and aggressiveness in breast cancer (Wulf et al, 2001; Girardini et al, 2011). However, so far Pin1-dependent signaling mechanisms have not been linked to breast CSCs' biology. The aim of this work was to show, by performing in vivo and in vitro functional studies, that Pin1 acts as a fundamental regulator of stem cell features both in normal stem cells and CSCs of the mammary gland. Pin1 knock-out mice show a number of developmental defects (Atchison & Means, 2004) affecting among others mammary epithelium, that fails to undergo the dynamic changes required to its expansion during pregnancy (Liou et al, 2002). Based on this, our collaborators hypothesized a possible function of Pin1 in governing the functions of mammary stem cells and thus it was evaluated the stem cell activity of mammary epithelial cells from wild-type (Pin1+/+) and knock-out (Pin1−/−) mice. To this aim, mammary tissues from 8 to 10 weeks old virgin female mice were dissociated, prepared as single cell suspensions of purified, lineage-depleted epithelial cells (Sleeman et al, 2006; Stingl et al, 2006) and grown in suspension cultures to form secondary mammospheres (M2) (Dontu et al, 2003). Whereas cells obtained from Pin1+/+ mice formed an average of 22.9 (±1.44) M2 mammospheres per 100.000 seeded cells, a 40% reduction of M2 formation from Pin1−/− cells was observed (Figure 3.4 A). In addition, to assess the impact of Pin1 on the replicative potential of mammary stem cells, wild-type cells were serially replated from primary mammospheres (M1) for four more times (M2–M5). As expected in these conditions, a progressive decrease was observed in mammosphere formation at each passage, due to exhaustion of adult stem cells (Cicalese et al, 2009). Notably, this effect was significantly exacerbated by addition of the Pin1 small molecule inhibitor PiB (Uchida et al, 2003): mammosphere formation efficiency of Pin1+/+ shrunk progressively and was reduced by almost 50% at the stadium of quaternary mammospheres (M4) and did not reach the M5 level (Figure 3.4 B).

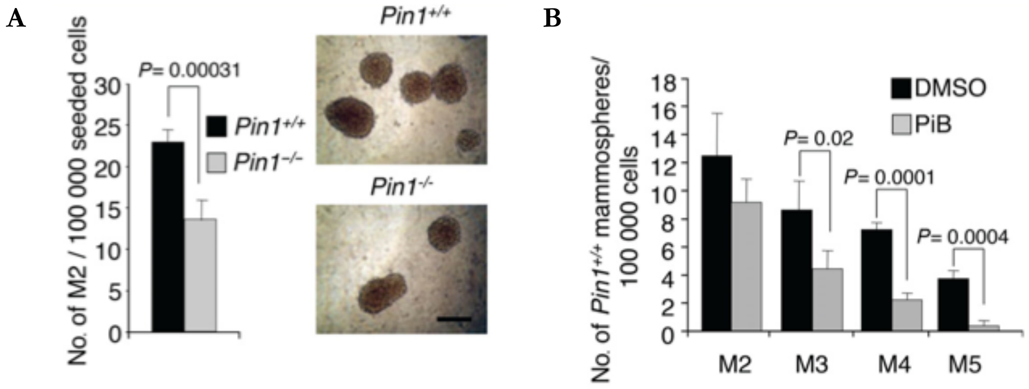
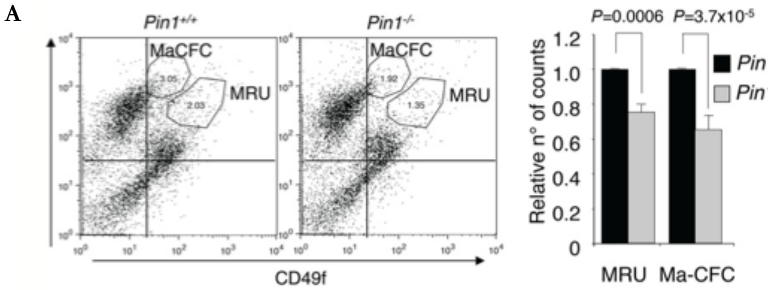


Figure 3.4. A) Pin1−/− mice display decreased self-renewal of mammary stem cells. Left panel: Number of secondary mammospheres (M2) generated from primary mammary epithelial cells of indicated mice. Means, standard deviations and *P*-values (*t*-test, *n* = 4) are indicated in the histogram. Right panel: representative M2 microscope images with 200 μm scale bar. B) Inhibition of Pin1 affects replicative potential of mammary stem cells. Serial replating of mammospheres (M1–M5) generated from Pin+/+ mice treated with DMSO or PiB (1.5 μM).

To better characterize this aspect, the proportion of stem cells and progenitors was analyzed by flow cytometric analyses and sorting (FACS) analysis using the surface markers CD24 and CD49f. These markers are widely used to identify two populations of cells functionally characterized as stem/bipotent progenitors (CD24^{med}/CD49f^{high} or mammary repopulating units, MRU) and luminal progenitors (CD24^{high}/CD49f^{low} or mammary colony forming cells, Ma-CFCs) (Stingl et al, 2006). In line with this hypothesis, the MRU and Ma-CFC cell populations from Pin1−/− mammary glands were present at lower proportion as compared to Pin1+/+ mice (Figure 3.5 A). In addition, Pin1 mRNA and protein levels in the MRU cell population were almost three times higher as compared to the total of mammary epithelial cells (Figure 3.5 B). This evidence confirmed the hypothesis and suggests a prominent role of Pin1 in sustaining the mammary stem cell compartment in vivo.



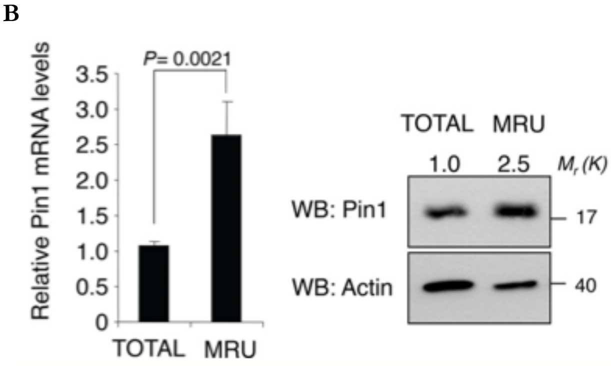


Figure 3.5. A) Decreased number of bipotent stem cell and luminal progenitor in Pin1−/− mammary tissue. Left panel: representative FACS analyses of mammary epithelial cells from indicated mice. CD24/CD49f plots and gatings for MRU and Ma-CFC populations are indicated. Right panel: histogram of mean counts of MRU and MA-CFC populations from Pin−/− normalized to Pin1+/+ mice. Means, standard deviations and P-values (t-test, n = 3) are indicated. B) Pin1 mRNA and protein levels are up-regulated in the mammary stem cell compartment. Left panel: qRT-PCR of endogenous Pin1 mRNA in MRU sorted populations relative to total population. Means, standard deviations and P-values (t-test, n = 3) are indicated. Right panel: Western blot analysis of the same cell populations as in the left panel. Fold change in Pin1 protein levels determined by Image J software (Rasband, 1997–2012) with respect to actin levels is indicated by a number, Molecular weights in kDa (Mr (K)) are shown on the right.

Stem cell traits in a subpopulation of mammary tumor cells are thought to be implicated in treatment resistance (Dean et al, 2005) and metastasis dissemination (Malanchi et al, 2012; Rosenthal et al, 2012; Visvader & Lindeman, 2012) and high levels of Pin1 correlate with high grade breast cancer and chemoresistance (Wulf et al, 2001; Ding et al, 2008; Kim et al, 2009; Girardini et al, 2011). Therefore it was next chosen to investigate whether Pin1 could also control mammary CSCs. NOP6 mouse mammary tumor cells, harboring the Her2/Neu amplification, were grown as mammospheres in presence or absence of the Pin1 inhibitor (Figure 3.6 A). NOP6 cells formed very fast growing spheres that did not decrease when propagated to M3 or M4, indicating that mammosphere-forming cells were self-renewing at a constant rate. Conversely, when cells were treated with Pin1 inhibitor, mammosphere formation efficiency (MFE) was strongly impaired already at the M2 level. It was also next tested whether Pin1 could be required for the maintenance of human breast CSCs. To address this question, the MDA-MB-231 breast cancer cell line expressing a doxycycline-inducible knockdown costruct for Pin1 (pLKO-TetO-shPin1) was generated and tested in mammosphere formation assays. As shown in Figure 3.6 B (left), in agreement with other reports (Harrison et al, 2010; Cordenonsi et al, 2011), non-induced cells had on average 0.6% of MFE, remaining constant throughout serial replating to M4. Instead, in Pin1 silenced (+DOX) cells, MFE decreased already at M2 stage and progressively at M3 and M4. The content of putative stem cells was lower following Pin1 silencing or inhibition, as confirmed by the Aldefluor assay, that evaluates the activity of Aldehyde dehydrogenase 1 (Aldh), a marker for breast CSCs (Ginestier et al, 2007) (Figure 3.6 B, right).

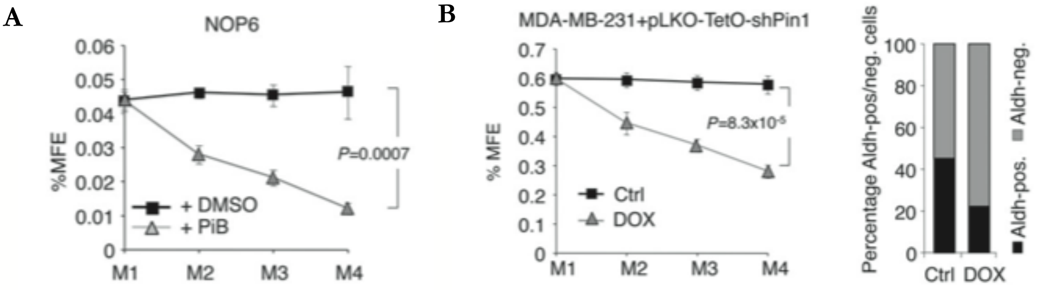


Figure 3.6. A) Pin1 inhibition decreases self-renewal of mouse mammary tumor cells. Serial replating of mammospheres (M1–M4) generated from NOP6 cells treated with DMSO or PiB (1.5 lM). Mammosphere formation efficiency (%MFE) was calculated as percentage of mammospheres divided by the number of plated cells. B) Pin1 knockdown decreases self-renewal of human breast cancer cells. Left panel: MFE of MDA-MB-231-pLKO-shPin1 control cells (Ctrl) compared to shPin1 inducedcells (DOX) upon serial passages. Right panel: Quantification of Aldh-positive and Aldh-negative cells from control- and shPin1 induced M4, as assessed by FACS.

Next it was evaluated the expression of several genes acting within pathways governing the stemness phenotypes of breast CSCs (Leong et al, 2007; Yu et al, 2007, 2011; Polyak & Weinberg, 2009; Cordenonsi et al, 2011; Visvader & Lindeman, 2012). As shown in Figure 3.7, the expression of tested factors (Hes1, HeyL, Birc5, CTGF, Slug, ABCG2, Ptch, Bmi-1, HMGA2 and Klf4) decreased by Pin1 knockdown. Epithelial-mesenchymal plasticity in breast carcinoma has recently been linked to acquisition of stem cell traits by tumor cells (Mani et al, 2008). It was therefore also analysed the impact of Pin1 modulation on this process by analyzing markers of epithelial-mesenchymal transition (EMT). Of note, Pin1 down-modulation caused enhanced mRNA expression of the epithelial marker E-cadherin (CDH1) while that of mesenchymal markers Vimentin and Fibronectin (VIM1, FN) was reduced (Figure 3.7).

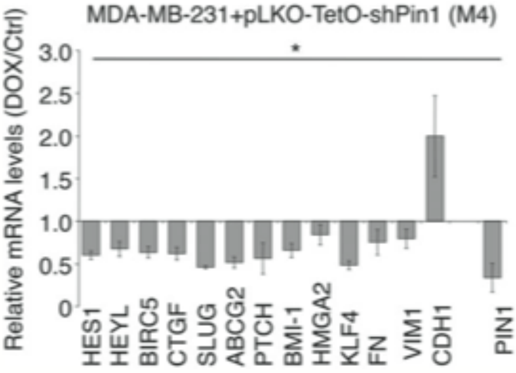


Figure 3.7: Pin1 knockdown affects expression of stem cell markers. qRT-PCR of the indicated stemness and EMT marker genes from MDA-MB-231-pLKO-shPin1 quaternary mammospheres (M4) upon shPin1 induction (DOX) with respect to control cells (Ctrl). Standard deviations are indicated, P-values * <0.02 (t-test, n = 3).

All together these results indicate that high Pin1 levels are required to sustain mesenchymal traits and to keep pro-stemness signaling constant. The majority of genes described above are controlled by the Notch pathway (Lee et al, 2008; Ranganathan et al, 2011; Li et al, 2012), which was shown to be required for EMT induction (Leong et al, 2007) and regulation of both normal stem cells of the mammary gland and breast CSC (Dontu et al, 2004; Bouras et al, 2008; Raouf et al, 2008; Harrison et al, 2010; Xing et al, 2012). It was investigated whether the action of Pin1 in breast CSCs maintenance is driven by Notch function. Notch proteins are membrane-bound receptors, that upon ligand binding are subjected to cleavage by gamma-secretase, releasing an intracellular domain (N-ICD) directly involved in transcriptional control (Ranganathan et al, 2011). In particular, two members of the family, Notch1 and Notch4, have been linked to induction and maintenance of breast CSC features (Farnie et al, 2007; Grudzien et al, 2010; Harrison et al, 2010).

Notably, the levels of their active forms (N1-ICD and N4-ICD) were strongly reduced (about five fold) by Pin1 knockdown in M4 mammospheres compared to control cells (Figure 3.8 A). To address the question whether the effect of Pin1 on breast CSC relies on its action on N1-ICD levels, we tested the ability of wild- type N1-ICD or of a constitutively stable N1-ICD mutant (dPEST) to rescue M2 formation following Pin1 knockdown. This mutant lacks the cdc4-phosphodegron constituting the consensus for the E3 ubiquitin-ligase Fbxw7a, the major negative regulator of the intracellular Notch signal (O'Neil et al, 2007; Thompson et al, 2007). As expected, M2FE of MDA-MB-231-pLKO-shPin1 cells decreased upon Pin1 silencing (+DOX) (Figure 3.8 B). Notably, M2FE did not further increase following ectopic expression of N1-ICD in control cells, since in these cells endogenous Notch pathway is already strongly activated (Harrison et al, 2010). Moreover, N1-ICD overexpression was not able to rescue M2FE in Pin1 silenced cells. By contrast, overexpression of N1-ICD-dPEST was able to rescue M2FE.

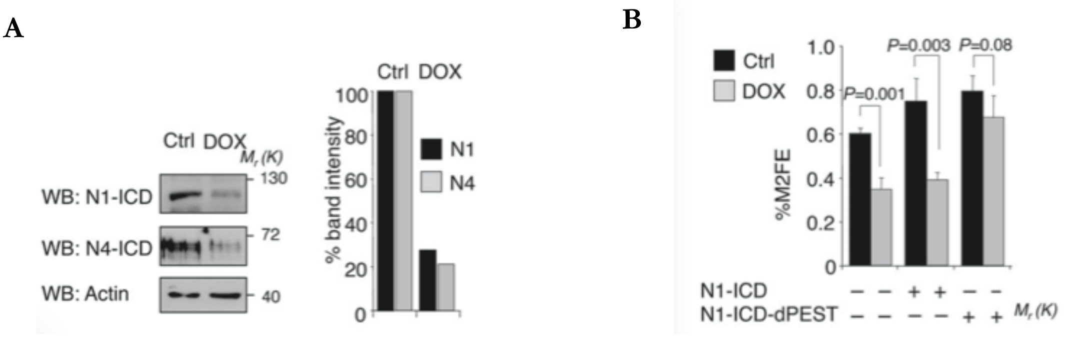


Figure 3.8: Pin1 controls breast CSC self-renewal through N1-ICD stabilization. A) Pin1 depletion causes reduced N1- and N4-ICD protein levels. Left panel: Western Blot analysis of N1- and N4-lCD protein from MDA-MB-231-pLKO-shPin1 M4 control cells (Ctrl) and shPin1 induced cells (DOX). Molecular weights (Mr) are indicated in kDa. Right panel: histogram representing the percentage of band intensity with respect to actin levels. B) Expression of N1-ICD-dPEST stable mutant rescues M2FE following Pin1 depletion. Upper panel: Percentage of secondary mammosphere formation efficiency (%M2FE) of control (Ctrl, black bars) or Pin1 silenced (DOX, grey bars) cells, transduced with empty (-), N1-ICD or N1-ICD-dPEST vectors (+). Means, standard deviations and P-values (t-test, n = 3) are indicated.

At the biochemical level, it was demonstrated that Notch1 and Notch4 escape from Fbxw7α-dependent proteasomal degradation following interaction with Pin1 and that phospho-specific prolyl-isomerization of Notch1 triggers de-phosphorylation by the PP2A phosphatase, preventing Fbxw7α interaction and subsequent poly-ubiquitination. While mouse xenograft experiments prove the relevance of Pin1 in tumor growth and metastasis formation in vivo, gene expression and immunohistochemical analyses of primary tumors from breast cancer patients show that Pin1 overexpression is significantly linked to activated Notch, irrespectively of the coexistance of functional Fbxw7α. In human patients with breast cancer high expression of Notch receptors and ligands is causally involved and has been linked to poor clinical outcomes (Han et al, 2011; Xu et al, 2012); in this context, in the absence of mutations, high Pin1 expression might contribute to sustain levels and function of nuclear N1- and N4-ICD by interfering with their degradation by Fbxw7a. To evaluate this hypothesis, serial sections of 38 TNBC samples of breast cancer tissues were stained with anti-N1-ICD, anti-Pin1 and anti-Fbxw7 antibodies. Among 22 patients with high intracellular Notch1 immunoreactivity, an high percentage of patients with a strong nuclear Fbxw7a signal (72.7%) was found; moreover, the majority of these samples (93.8%) also displayed high Pin1 levels, that might be responsible for the simultaneous presence of high N1-ICD and its ub-ligase. These results were finally confirmed in silico on the *prognostic meta-dataset*. The entire meta-cohort was classified according to high/low expression of FBXW7 and PIN1 mRNA with the combined Z-score. Considering that mRNA levels of Notch receptors are frequently not representative of the protein levels of N1-ICD, activated Notch1 pathway status in this cohort was inferred from expression levels of a Notch-dependent gene signature (Notch direct target gene signature, NDT), built up by selecting published Notch1 targets, for which Notch responsiveness and/or direct promoter binding as well as their expression in breast cancer was demonstrated (Table 3.6). The combined Z-score was used to identify two groups of tumors with either high or low NDT activity. Tumors were classified as NDT activity High if the combined score was positive and as NDT activity Low if the combined score was negative.

Table 3.6: Notch direct target (NDT) gene signature.

Gene symbol	Entrez Id	Reference
HES1	3280	Grabher et al., 2006
HEY1	23462	Grabher et al., 2006
HEY2	23493	Grabher et al., 2006
HEYL	26508	Grabher et al., 2006
MYC	4609	Palomero et al., 2006; Weng et al., 2006
BUB1B	701	Palomero et al., 2006
BUB3	9184	Palomero et al., 2006
CDC25A	993	Palomero et al., 2006
PHB	5245	Palomero et al., 2006
RBL1	5933	Palomero et al., 2006
RPL3	6122	Palomero et al., 2006
USP5	8078	Palomero et al., 2006
PLAU	5328	Shimizu et al., 2011
SHQ	55164	Chadwick et al., 2009
CCND1	595	Ronchini and Capobianco, 2001
GATA3	2625	Amsen et al., 2007
SKP2	6502	Sarmen et al., 2005
ERBB2	2064	Chen et al., 1997
CDKN1A	1026	Rangarajan et al., 2001
SNAI1	6615	Sahlgren et al., 2008
SNAI2	6591	Leong et al., 2007
NFKB2	4791	Oswald et al., 1998
BIRC5	332	Lee et al., 2008
NOTCH1	4851	Weng et al., 2006; Hamidi et al., 2011
NOTCH3	4854	Weng et al., 2006; Hamidi et al., 2011
NOTCH4	4855	Hamidi et al., 2011
IFRD2	7866	Palomero et al., 2006
ING3	54556	Palomero et al., 2006
PTCRA	171558	Grabher et al., 2006
CD3D	915	Palomero et al., 2006

More than 48% of all samples expressed high levels of NDT signature genes and this correlated with poorer overall survival; the two NDT signature groups were compared by univariate Kaplan-Meier and depicted in Figure 3.9. The group with high-NDT expression signature displayed a significantly higher probability to reduced survival (p-value=5e-04). This is consistent with previously published analyses (Farnie et al, 2007) and confirms the usefulness of this signature as a surrogate of activated Notch1 pathway.

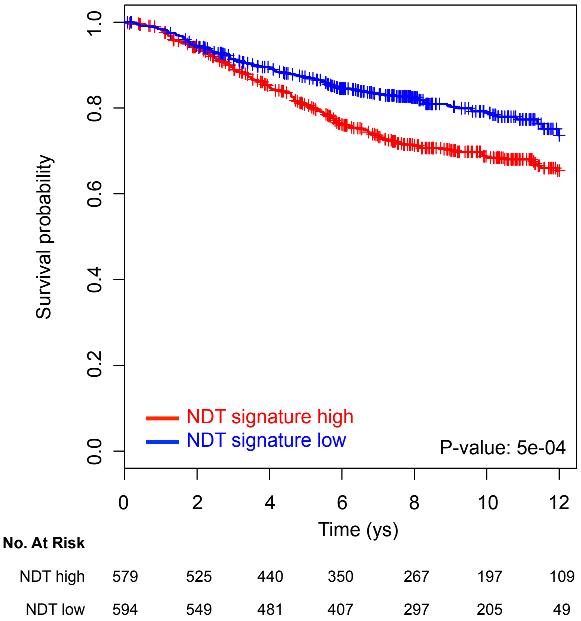
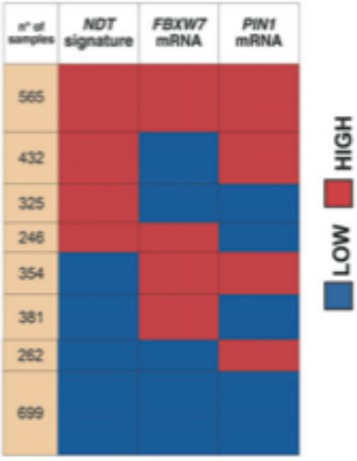


Figure 3.9: Survival analysis of patients in function of NDT signature expression. Kaplan-Meier graphs representing the probability of overall survival in breast cancer patients of prognostic meta-cohort stratified according to high or low expression levels of the NDT signature. The log-rank test p value reflects the significance of the association between NDT gene signature low and longer survival.

Interestingly, high FBXW7 levels expression was found in 51.7% of patients with hyperactive Notch (N1-ICD) (Fig 3.10). Similarly to the findings in the cohort of 38 TNBC samples of breast cancer tissues, high correlation with PIN1 overexpression was found in the great majority of these cases, while those with low PIN1 expression were underrepresented with respect to any other possible category of patients with high levels of FBXW7 mRNA (Fig 3.10).



			N1-ICD		PIN1	
FBXW7	high	1546 (47.5%)	high	811 (51.5%)	high	565 (69.7%)
					low	246 (30.3%)
			low	735 (47.5%)	high	354 (48.2%)
					low	381 (51.8%)
	low	1708 (52.5%)	high	757 (44.3%)	high	432 (57.1%)
					low	325 (42.9%)
			low	951 (55.7%)	high	262 (27.5%)
					low	689 (72.5%)

Figure 3.10: NDT expression analysis in the *prognostic meta-dataset*. Left panel: heat map representing the contingency table frequencies of samples classified as having high or low levels of FBXW7, of PIN1 and of the NDT gene signature. Number of samples in each category is indicated on the left. The association among high levels of NDT gene signature, PIN1, and FBXW7 resulted statistically significant (P < 0.001; chi-square test). Right panel: Contingency table showing percentage of each category calculated on the precedent category of patients.

Notably, the average expression value of NDT gene signature was contingent on PIN1 mRNA levels (Fig 3.11), while FBXW7 was non influential, therefore highlighting the biological dominance of Pin1 over Fbxw7a in regulating Notch signaling in breast cancer.

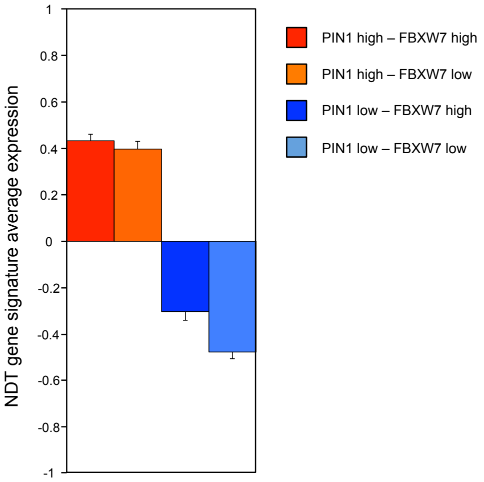


Figure 3.11: Expression correlation between NDT gene signature, PIN1 and FBXW7 mRNA levels. Average expression of NDT gene signature in breast cancer samples stratified according to high or low expression of PIN1 and FBXW7 mRNA. Data are shown as mean standard error of the mean (s.e.m.).

To evaluate the clinical significance of this finding, the effect of high or low Pin1 expression levels was searched on the survival of patients with high or low NDT signature. While Pin1 levels did not affect the clinical outcome in all patients, we found that in grade 3 breast cancer high Pin1 levels correlate with a worse outcome in patients with activated Notch1 signature (high NDT-high PIN1, Fig 3.12).

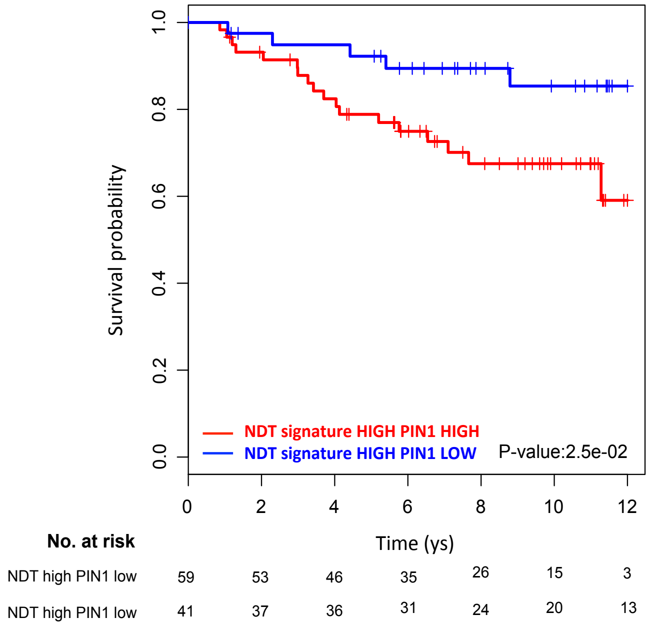


Figure 3.12: Survival analysis of grade 3 high NDT expressing patients in function of PIN1 expression. Kaplan–Meier survival curve is indicated for high NDT signature, grade 3 breast cancer patients in function of high or low PIN1 mRNA levels. P-value and the number of subjects at risk at each time point are indicated below.

Moreover, pathway enrichment showed association with several stem cell pathway signature genes in high NDT signature-high PIN1 group of patients with grade 3 (Table 3.7). The over-representation of signaling pathways (see Tables 2.37-38) in genes up-regulated in samples with high NDT signature-high PIN1 was tested using a one-sided Fischer's exact test on signaling pathways as described in 2.8.2 paragraph. Genes up-regulated in grade 3 high NDT signature-high PIN1 patients have been identified using SAM algorithm with 1,000 permutations and setting the q-value threshold at 0.01 and fold change higher than 1.5. SAM comparison of gene expression profiles in high NDT signature-high PIN1 and high NDT signature – low PIN1 grade 3 samples resulted in 1460 differentially expressed genes up-regulated in high NDT signature-high PIN1 grade 3 samples.

Table 3.7: Fisher's test's results. Signaling pathways significantly enriched in genes up-regulated in high NDT signature-high PIN1 grade 3 samples as determined by one-sided Fisher's exact test. Enrichment was considered significant with a BH adjusted p-value < 0.05.

Pathway	Adjusted p-value
ES1	0.008
ES.like	0.021
IGS	0.029
Mutant p53	0.034

3.3.2 Predictive meta-cohort

A *predictive meta-cohort* was composed by breast cancer samples from neo-adiuvant chemotherapy (NAC) treated patients with available chemotherapy response information in terms of pCR (pathological complete response) and RD (residual disease). All patients underwent pretreatment biopsies of the primary breast tumor. This cohort included 10 datasets (i.e., MDACC/IGR, USO-02103, I-SPY-1, LBJ_INEN_GEICAM, Osaka, TOP, MDACC, MDACC_MACQ_GSE25066, MDACC_MAQC_GSE20194, UW) for a total of 979 primary tumors.

Use of predictive meta-cohort: a multifactorial tool for predicting response to neo-adjuvant anthracycline-based chemotherapy in Triple-Negative Breast Cancers

A vast majority of patients considered to be at moderate or high risk of relapse is treated with cytotoxic agents, most of them are anthracyclines. Across all anthracycline-treated patients, only a small percentage actually receives benefit, while these agents are associated with significant toxicities. Breast cancer is well recognized as a heterogeneous disease and therefore treating all breast cancers with the same chemotherapeutic agents could be considered illogical. Little progress has been made in the field of biomarkers predictive of chemotherapy benefit in breast cancer. In this thesis, we focused on identifying molecular markers that predict response or resistance to anthracyclines in a specific molecular subtype of breast cancer: basal-like (also known as triple negative breast cancer). We therefore aimed to develop a gene expression signature to identify those patients who would not benefit from anthracyclines and could thus be spared the non-negligible risks of this type of chemotherapy.

The cohort of interest of TNBCs treated with anthracycline-based chemotherapy was derived from the meta-cohort (Table 3.1). Among the overall 939 basal-like samples (25.65%) classified using the *SCMOD2* subtype clustering classifier, 331 had information about neo-adjuvant chemotherapy: 67 samples were treated with anthracycline-based chemotherapy, 264 with a combination of anthracycline and taxane based chemotherapy. To mimic in-silico a sort of small clinical trial, we separated the 331 TNBC samples treated with neo-adjuvant chemotherapy between a *design cohort* (namely patients treated with anthracyclines (A)) and a *control cohort* comprising samples treated with another neo-adjuvant therapy (i.e., a mix of taxanes and anthracyclines (AT)). Of course a cohort of patients treated without anthracyclines would have been desirable as *control cohort*, but unfortunately no such data are available at all. Finally, for validation we used data from a third, external cohort (*validation cohort*), not previously comprised in the meta-cohort because data have been obtained from paraffin fixed and not fresh frozen tissues and hybridized on a different type of platform (GSE6861, see paragraph 2.3.2) (Figure 3.13).

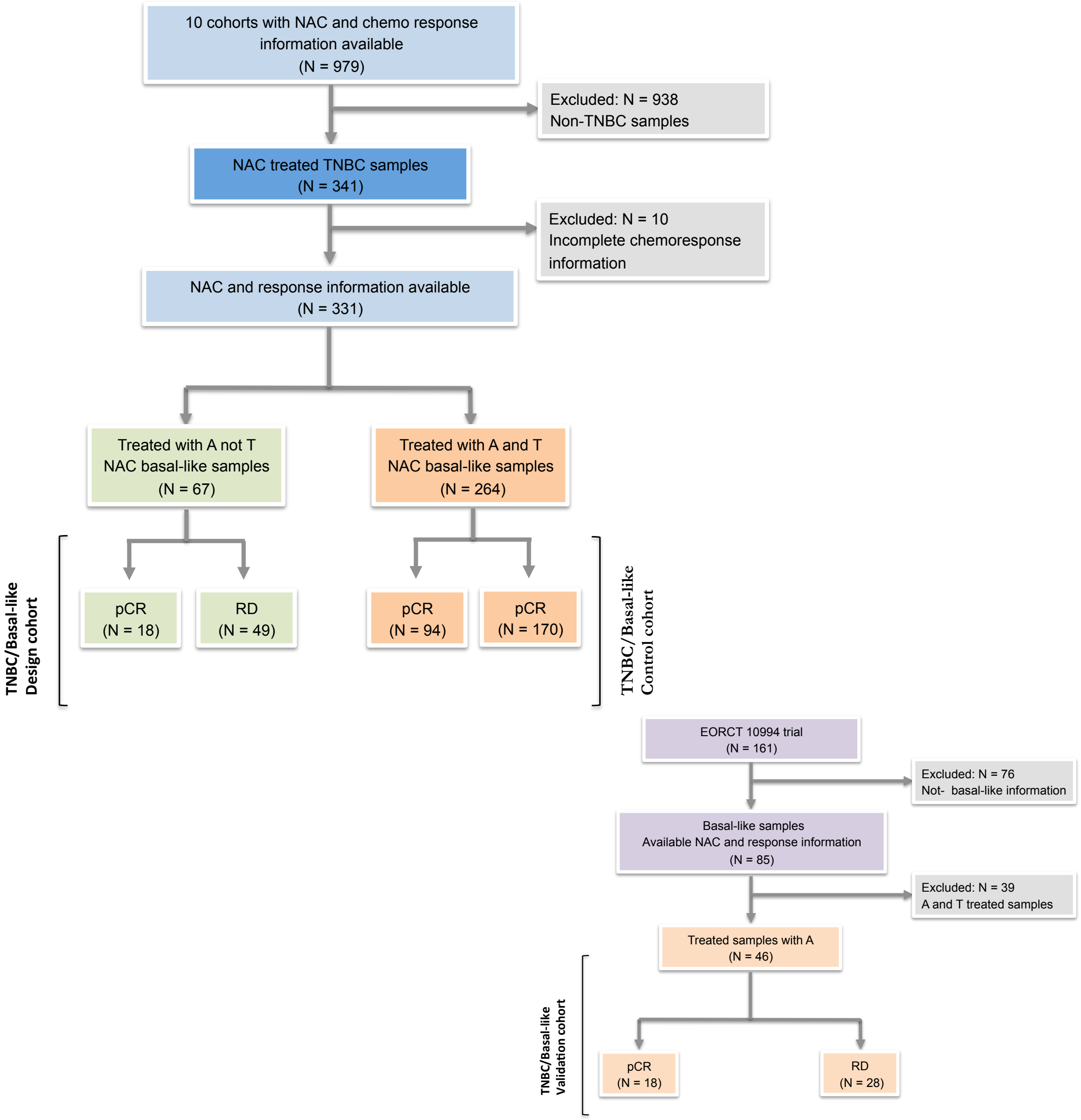

10 cohorts with NAC and chemo response information available (N = 979)
Excluded: N = 938 Non-TNBC samples
NAC treated TNBC samples (N = 341)
Excluded: N = 10 Incomplete chemoresponse information
NAC and response information available (N = 331)
Treated with A not T NAC basal-like samples (N = 67)
Treated with A and T NAC basal-like samples (N = 264)
pCR (N = 18)
RD (N = 49)
pCR (N = 94)
pCR (N = 170)
TNBC/Basal-like Design cohort
TNBC/Basal-like Control cohort
EORCT 10994 trial (N = 161)
Excluded: N = 76 Not- basal-like information
Basal-like samples Available NAC and response information (N = 85)
Excluded: N = 39 A and T treated samples
Treated samples with A (N = 46)
pCR (N = 18)
RD (N = 28)
TNBC/Basal-like Validation cohort

Figure 3.13. Flow-chart to derive cohorts of interest from the *in silico* predictive breast cancer meta-cohort.

Clinical and tumor characteristics for patients of *design cohort* are listed in Table 3.8. These samples were originally contained in 4 different datasets (Table 3.9), i.e., I-SPY-1, MDACC, MDACC/IGR and TOP and 18 samples (26.87%) achieved complete pathological response (pCR) to therapy while 49 had residual disease (RD).

Table 3.8. Clinical characteristics of TNBC patients treated with anthracycline based-chemotherapy (*design cohort*) (n = 67).

Clinical variables	Patients = 67
Age (years)	
<40	12 (18 %)
40-60	33 (49 %)
>60	4 (6 %)
Unknown	18 (27 %)
Tumor size	
T1	3 (4 %)
T2	40 (60 %)
T3	16 (24 %)
T4	8 (12 %)
Lymph nodes status	
N0	11 (16 %)
N1	9 (13 %)
N2	5 (8 %)
Unknown	42 (63 %)
Grade of tumor	
G1	0 (0 %)
G2	9 (13 %)
G3	53 (79 %)
Unknown	5 (8 %)
pCR	
Yes	18 (27 %)
No	49 (73 %)

Table 3.9. Distribution of samples treated with anthracycline-based chemotherapy among original single datasets.

Response	Single datasets				Total
	I-SPY-1	MDACC	MDACC/IGR	TOP	
pCR	0	3	15	0	**18**
RD	1	18	12	18	**49**
Total	**1**	**21**	**27**	**18**	**67**

All clinical variables were tested for their ability to predict pCR. Odd-Ratios (ORs) and their associated p-values were used to compare response to treatment (pathological complete response, pCR) between groups defined by different clinical and molecular characteristics (age, tumor size, nodal status, histologic grade, p53 status). No significant association between any clinical characteristics and pCR was found (Table 3.10).

Table 3.10: Odds ratios (OR) for response to treatment defined as pCR (pathological complete response) according to the clinical characteristics of the *design cohort* (n = 67).

Characteristic	No. of patients	Patients with pCR (%)	OR	95% CI	p value
Age (years)					
<40	12	33.3			
40-60	33	39.4	1.3	0.33 - 5.69	0.711
>60	4	25	0.67	0.03 – 7.45	0.756
Tumor size (T)					
T1-T2	43	30.2	0.61	0.17 - 1.90	0.408
T3-T4	24	20.8			
Nodal status (N)					
N0	11	18.2	0.75	0.08 - 7.25	0.792
N1-N2	14	14.3			
Histologic grade					
G1-G2	9	11.1	3.78	0.62 - 72.98	0.227
G3	53	32.1			

Design of the Consensus Signature

In order for anthracycline-based chemotherapy to be effective we postulated the following steps (Figure 3.14).

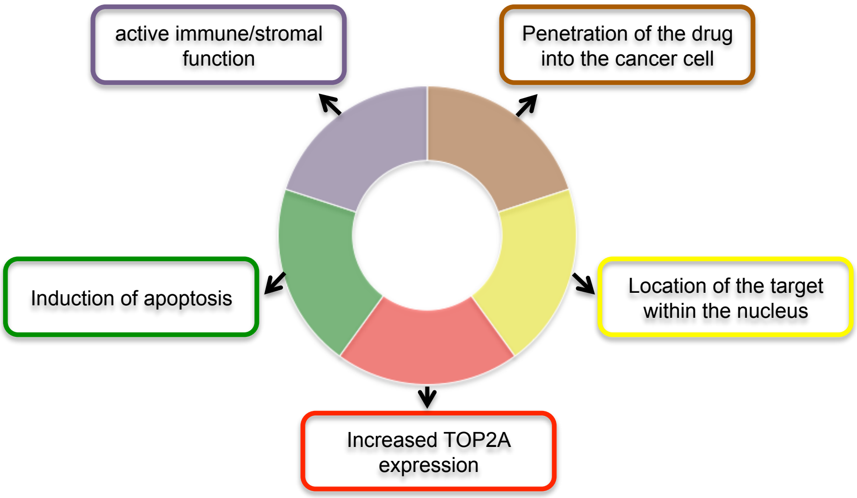

active immune/stromal function
Penetration of the drug into the cancer cell
Induction of apoptosis
Location of the target within the nucleus
Increased TOP2A expression

Figure 3.14: Postulated components in order to be effective for anthracycline-based chemotherapy.

The genes or gene signatures, described in paragraph 2.5, have been applied in the study for evaluation of a multifactorial approach, by *in silico* analysis, for predicting response to neo-adjuvant anthracycline-based chemotherapy in triple negative breast cancer patients (Table 3.11). All genes and gene lists have a direct biological meaning in the contest of breast cancer behavior; in particular, Minimal signature, Sharp1 and HIF signatures are the read-outs of specific molecular mechanism involved in breast cancer (see paragraph 2.5).

Table 3.11: Postulated steps required for effective anthracycline-induced cytotoxicity.

Step required for anthracycline sensitivity	Surrogate marker	Association with pCR
1. Penetration of drug into the cancer cell	SHARP1 signature	Negative
	Hypoxia signature (HIF)	Negative
2. Location of topoIIα protein within the nucleus	LAPTM4B mRNA	Negative
3. Increased expression of TOP2A mRNA	TOP2A mRNA	Positive
4. Induction of apoptosis	YWHAZ	Negative
	Minimal signature (MS)	Positive
5. Active immune function	Immune signature (STAT1)	Positive
	Stromal signature (PLAU)	Negative

As shown in table 3.11, we linked the high expression of genes or gene signatures with a positive or negative pCR as highlighted in biological experiments. Increased hypoxia function and high expression of LAPTM4B have a negative association with pCR. Moreover, increased expression of TOP2A mRNA has a positive association with pCR. YWHAZ and the MS were both selected for evaluation as markers of apoptosis and their high activity has a negative and positive association with pCR, respectively. Finally, high immune (STAT1) and stromal (PLAU) activities have respectively positive and negative association with pCR.

Quantification of genes and gene signatures

SHARP1, HIF signatures and Minimal Signature were quantified using a continuous combined Z-score as previously described in paragraph 2.6. Briefly, each signature was calculated by summarizing the standardized expression levels of the genes in the signature into a combined score with zero mean. STAT1 and PLAU signatures were quantified in a continuous score (module score) as previously described in paragraph 2.6. LAPTM4B, TOP2A, and YWHAZ mRNA expression levels have been calculated using the correspondent probe sets or the median expression if multiple probe sets were available for each gene.

Predictive power of single gene/gene signatures

We first assessed, using the receiver operating characteristic (ROC) analysis, the ability of any single gene/gene signature to discriminate patients with pathologic complete response from patients with residual disease. The ROC analysis requires defining the sign of the association between any signature and pCR. Given the postulated role of each component in determining anthracycline sensitivity, we defined the association of any gene/gene signature with pCR as reported in Table 3.12. The area under the curve (AUC) and the associated p-value highlighted that STAT1, HIF signatures, and TOP2A mRNA were significantly associated with pCR status, with a positive association for STAT1 and TOP2A mRNA and negative association for HIF, again consistent with their putative roles in anthracycline function. All other genes/gene signatures, when considered individually, were not significantly correlated with pCR.

Table 3.12. Predictive power of single component signatures in TNBC patients treated with anthracycline (*design cohort*).

Marker of activity of specific component	**Association with pCR**	**AUC**	**95% CI**	**p value**
SHARP1 signature	Negative	0.41	0.25 - 0.57	0.853
Hypoxia signature (HIF)	Negative	0.63	0.51 - 0.76	**0.018**
LAPT4MB mRNA	Negative	0.48	0.33 - 0.64	0.581
TOP2A mRNA	Positive	0.63	0.49 - 0.77	**0.035**
Minimal signature (MS)	Positive	0.53	0.39 - 0.66	0.347
YWHAZ mRNA	Negative	0.55	0.41 - 0.69	0.251
Immune signature (STAT1)	Positive	0.69	0.54 - 0.83	**0.006**
Stromal signature (PLAU)	Negative	0.52	0.38 - 0.67	0.371

Predictive power of the Consensus Signature

The use of single predictive markers of chemotherapy response to distinguish patients who are likely to receive benefits from those who are not is a clinically relevant need to improve patient selection for drug administration. It has been recently suggested that a multifactorial approach might be a more efficient alternative (Desmedt, 2011). In this thesis, starting from these hypotheses and previously described results, we derive *Consensus Signatures* as models for predicting neo-adjuvant chemotherapy sensitivity or resistance in TNBCs. *Consensus Signatures* are designed as linear combinations of the gene or gene signature scores highlighted above (Table 3.11). In particular, *Consensus Signatures* were designed by considering the core set of components comprising the genes or gene signatures shown to have significant predictive capability when used alone; that is, HIF, STAT1, TOP2A mRNA, and then, adding other components. Using a continuous score to quantify *Consensus Signature* expression level, combinations of core components demonstrated a significant correlation with pCR, with the HIF + STAT1 + TOP2A mRNA combination (*ConSig1*) being the most predictive, with AUC 0.79, 95% CI, 0.51 to

0.76, p = 9.1 x 10^{-9} (Table 3.13 and Figure 3.15). Interestingly, the addition of further component genes or gene signatures to the *ConSig1* did not increase overall predictive power, although each combination still retained excellent correlation with pCR.

Table 3.13 Predictive power of *Consensus Signatures* in TNBC patients treated with anthracycline (design cohort) using continuous score. In bold are the best performing consensus signature that includes components for 3, 4 or 5 of the steps required for anthracycline function (Table 3.11).

Combination of Consensus Signature components	AUC	95% CI	p value
STAT1	0.69	0.54 - 0.83	5.7 x 10^{-3}
HIF + STAT1	0.72	0.60 - 0.84	2.2 x 10^{-4}
HIF + STAT1 + TOP2A mRNA (*ConSig1*)	**0.79**	**0.69 - 0.90**	**9.1 x 10^{-9}**
HIF + STAT1 + TOP2A mRNA + LAPT4MB	0.76	0.65 - 0.87	1.6 x 10^{-6}
HIF + STAT1 + TOP2A mRNA + YWHAZ	0.78	0.68 - 0.88	1.3 x 10^{-8}
HIF + STAT1 + TOP2A mRNA + MS	0.76	0.65 - 0.87	1.1 x 10^{-6}
HIF + STAT1 + TOP2A mRNA + LAPT4MB + YWHAZ	0.73	0.62 - 0.85	2.7 x 10^{-5}
HIF + STAT1 + TOP2A mRNA + LAPT4MB + MS	0.74	0.62 - 0.85	1.6 x 10^{-5}
HIF + STAT1 + TOP2A mRNA + PLAU + LAPT4MB + YWHAZ	0.73	0.62 - 0.85	2.7 x 10^{-5}
HIF + STAT1 + TOP2A mRNA + PLAU + LAPT4MB + MS	0.75	0.64 - 0.86	2.7 x 10^{-6}

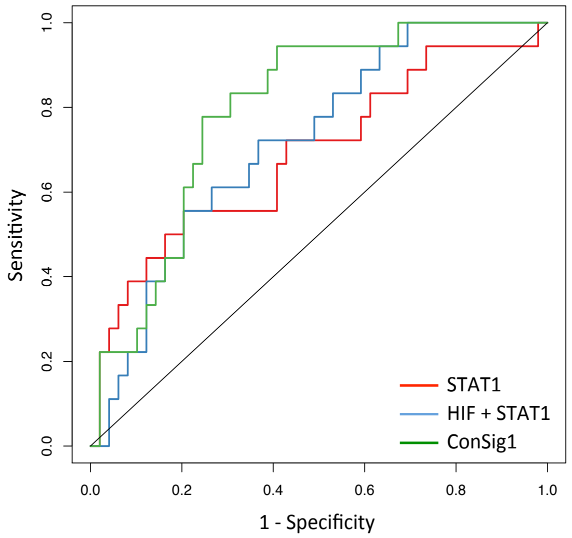


Figure 3.15: ROC analysis of ConSig1 in *design cohort*. Receiver operating characteristic (ROC) analysis of the ability of *ConSig1* to discriminate patients with pathologic complete response from patients with residual disease in the *design cohort*.

Moreover, to assess specificity of *ConSig1* for anthracycline-based chemotherapy response compared with other chemotherapy regimens, we analyzed its performance in a *control cohort* of patients who received taxanes in addition to anthracyclines (n=272), 264 of whom had information about response, and 94 with pCR (Figure 3.12). *ConSig1* was not predictive of response in this cohort, with no significant correlation with pCR seen on ROC analysis (AUC = 0.51, 95% CI, 0.44 to 0.58, p = 0.386) (Figure 3.16). These findings support that *ConSig1* has ability to discriminate patients with pathologic complete response from patients with residual disease in anthracycline-based chemotherapy treated patients.

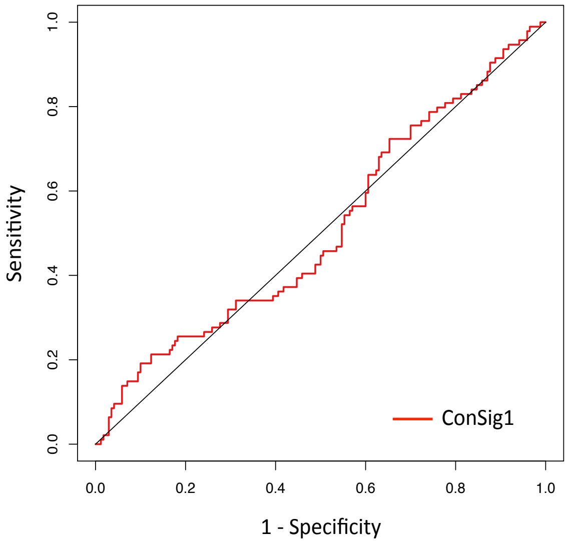


Figure 3.16: ROC analysis of ConSig1 in *control cohort*. Receiver operating characteristic (ROC) analyses of the ability of *ConSig1* to discriminate patients with pathologic complete response from patients with residual disease in the *control cohort*.

Performance of the Consensus Signature

As the best performing consensus signature, *ConSig1* was selected for subsequent analyses. To classify a patient as a putative responder or as resistant, a threshold for *ConSig1* score was determined by identifying the score value that maximized the Youden index (i.e. specificity + sensitivity – 1), with positive predictive value (PPV), negative predictive value (NPV), sensitivity (SENS), and specificity (SPEC) then calculated. When considering patients treated with anthracycline-based chemotherapy, NPV was higher (97%) than NPV considering patients treated with taxane plus anthracycline-based chemotherapy, while the ability to predict anthracycline sensitivity was moderate (PPV = 46%), but higher than PPV in *control cohort* (Figure 3.17).

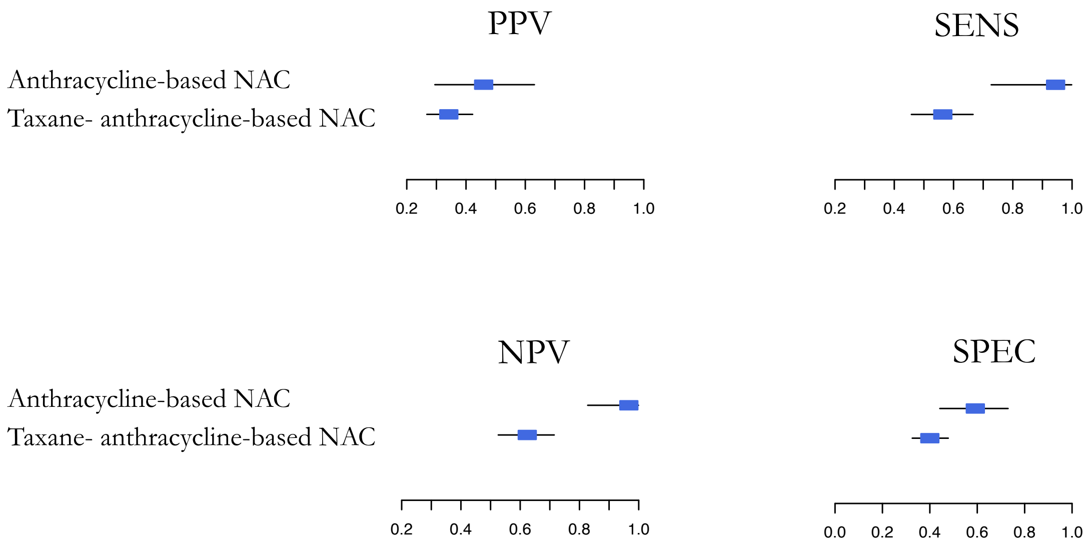


Figure 3.17. Performance of the *ConSig1* according to the cutoff defined by the maximal Youden Index in *design cohort*. The positive (PPV) and negative predictive values (NPV), sensitivity (SENS), specificity (SPEC) were determined at the threshold that maximizes the Youden Index (SPEC + SENS - 1) in the *design cohort*. Point estimates are displayed as squares. The horizontal lines correspond to exact 95% CIs.

We then also compared the predictive power of *ConSig1* with the A-SCORE (see paragraph 2.5.6), the only existing predictor of response to anthracycline chemotherapy. It resulted to have a predictive power 5 orders of magnitude lower than the *ConSig1* (AUC = 0.7, 95% CI, 0.58 to 0.83, p =6.5 x 10^{-4}). Moreover, the *ConSig1* has both negative predictive power and sensitivity 10% higher than the A-SCORE (Figure 3.18).

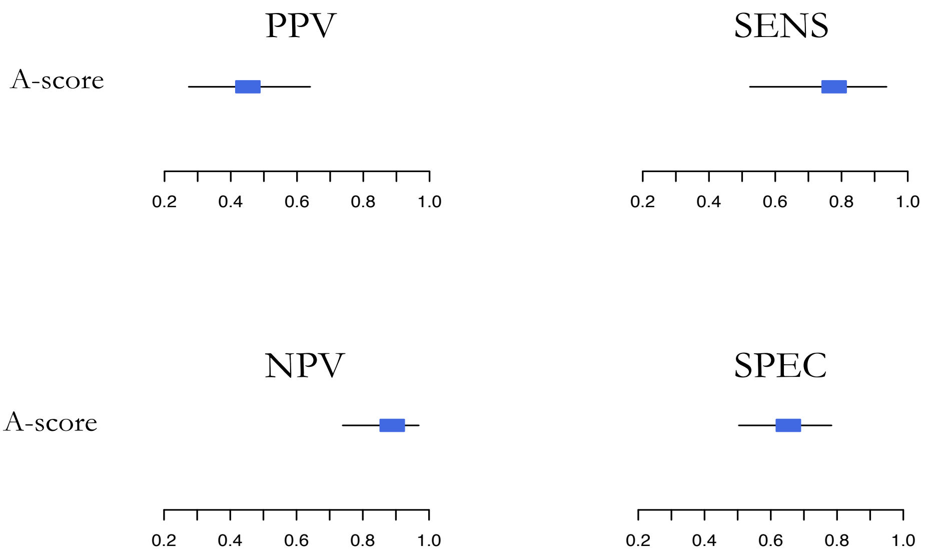


Figure 3.18: Performance of A-score according to the cutoff defined by the maximal Youden Index in *design study*. The positive (PPV) and negative predictive values (NPV), sensitivity (SENS), specificity (SPEC) were determined at the threshold that maximizes the Youden Index (SPEC + SENS - 1) in the *design cohort*. Point estimates are displayed as squares. The horizontal lines correspond to exact 95% CIs.

Predictive power and performance of the *ConSig1* in validation cohort

We assessed the predictive power and the performance of *ConSig1* using a *validation cohort* that contains samples from anthracycline-based treated patients. This cohort contained patients enrolled in the EORTC 10994 phase III breast cancer clinical trial (see 2.3.2 paragraph); we selected the basal-like samples using the *SCMOD2* classifier for a total of 85 samples of which 46 samples were treated with anthracycline based-chemotherapy. *ConSig1* was predictive of response also in this group, with significant correlation with pCR seen on ROC analysis (AUC = 0.65, 95% CI, 0.5 to 0.8, p = 0.024) (Figure 3.19).

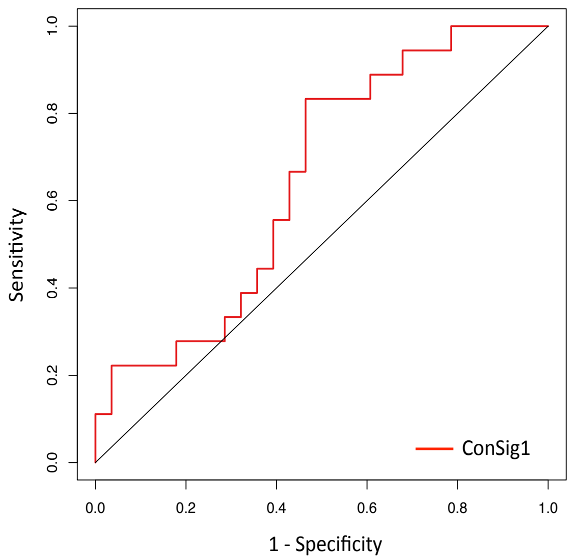


Figure 3.19: ROC analysis of ConSig1 in *validation cohort*. Receiver operating characteristic (ROC) analysis of the ability of *ConSig1* to discriminate patients with pathologic complete response from patients with residual disease in the *validation cohort*.

NPV was little lower than the NPV calculated on *design cohort* (71%), while the ability to predict anthracycline sensitivity was quite similar (PPV = 48%) (Figure 3.20).

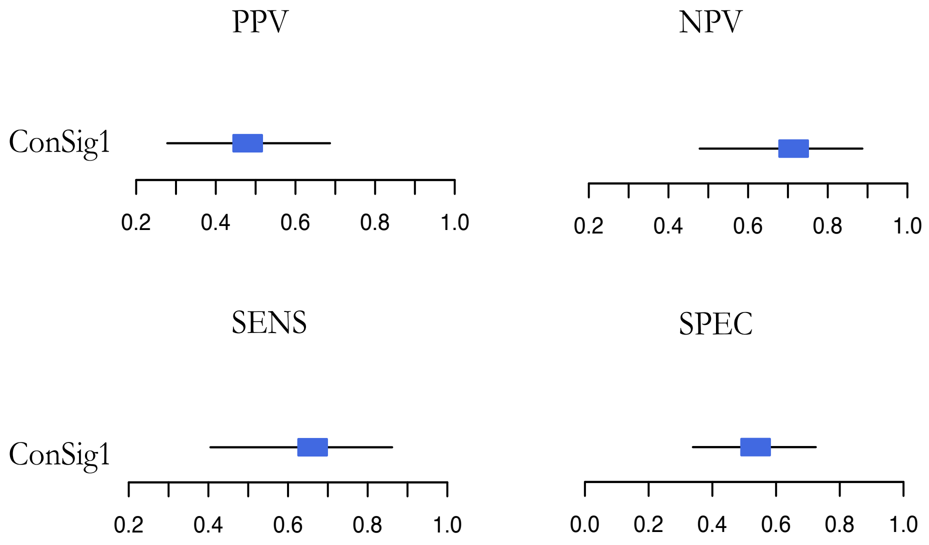


Figure 3.20. Performance of the *ConSig1* according to the cutoff defined by the maximal Youden Index in *validation cohort*. The positive (PPV) and negative predictive values (NPV), sensitivity (SENS), specificity (SPEC) were determined at the threshold that maximizes the Youden Index (SPEC + SENS - 1) in the *validation cohort*. Point estimates are displayed as squares. The horizontal lines correspond to exact 95% CIs.

Chapter 4

Conclusions

Since the completion of the human genome sequencing and the development of high throughput techniques, as DNA microarray, monitoring the expression of thousands of genes in a given tumor has become possible. These technological advances have been accompanied by the development of bioinformatics methods for the analysis and interpretation of an overwhelming mass of genomic data. The common objective of all these methods is the identification of statistically relevant genes sharing particular profiles from huge matrices bearing values for thousands of molecules. Seminal studies demonstrated that the synergistic use of microarray-based techniques and computational tools may not only further the understanding of cancer taxonomy, but also provide lists of genes that can classify tumors into distinct groups, with different diagnostic or prognostic characteristics. The identification of these gene expression signatures held promise for being more effective than standard prognostic and predictive factors. A demonstrable success occurred in early 2007 when the U.S. Food and Drug Administration approved MammaPrint, the first microarray-based commercial molecular prognostic test for breast cancer. Nonetheless, optimism for microarray-based technologies as predictive tests of cancer sensitive to therapy or recurrence has suffered both perceptual and real setbacks. Criticism is largely on the grounds of general non-reproducibility of gene signatures and the inability to replicate results in terms of significant genes identified from experiments in different laboratories and from different experimental platforms. Skepticism regarding reliability and reproducibility reflects the complexity of the analytical methods and the peculiar nature of the data generated by high-throughput technologies. Indeed, most of the discrepancies have to be ascribed to inconsistent probe annotation, low comparability of different microarray platforms, lack of probe specificity for different isoforms, or differences in the hybridization conditions, fluorescence measurement, normalization strategies and computational procedures adopted. Reliability concerns are further supported by results from several microarray studies that, investigating the same tumor type, identified different gene-expression signatures, all able to predict response to therapy or clinical outcome, but characterized by a minimal, if not null, number of overlapping genes. Again, several technical, analytical and biological reasons may, at least partially, explain these seemingly discrepant results. These include the use of different microarray platforms with different sets of probe and data normalization methods, as well as differences in the study populations. Two other major explanations are the lack of independent measurements between the expressed genes and the limited statistical power applied to select individual genes associated with response to therapy or clinical outcome. Since the strength of correlation between the genes and clinical outcome varies from data set to data set, the rank order of these informative genes in the prognostic and predictive signatures is highly unstable, thus leading to different gene lists with a small overlap. The low statistical power of prognostic and predictive signatures is mostly due to the limited number of samples included in the different data sets used for the development of classifiers. A final concern on the robustness of gene-expression signatures stems from the concepts inspiring the two different approaches applied so far for prognostic or predictive marker discovery: the top-down and the hypothesis-driven or bottom-up approaches. In the top-down approach a prognostic/predictive model is derived simply by looking for gene-expression patterns associated with clinical outcome without any a priori biological assumption, whereas in the bottom-up gene-expression profiles linked with a specific biological phenotype are first identified and subsequently correlated to survival or response to therapy. Although both valuable, the two strategies rely only on gene expression data and/or clinical information to derive the classification rules and none of the approaches include mechanistic insights in the discovery process. It is most likely that the selection of predictive genes on the basis of mechanistic insights, rather than solely on the basis of expression levels and outcome data, will dramatically improve the reliability and robustness of prognostic/predictive signatures.

The introduction of gene-expression tests have ushered in a new era in which many conventional clinical markers and predictors may be seen merely as surrogates for more fundamental genetic and physiologic processes. However, the multidimensional nature of these predictors demands both large numbers of clinically homogeneous patients to the used in the validation process, and exceptional rigor and discipline. Every study provides an opportunity to tweak a genetic signature, but the development of scientifically robust and clinically reliable tools require study designs and computational procedures. If gene-expression signatures are to reach the clinical setting, several outstanding issues will need to be addressed. First, researchers in this area will now need to turn their attention to methods of sample acquisition and the effect these methods have on the prognostic and predictive power of microarray data. Secondly, standardization of protocols and platforms for the measurement of gene-expression signatures in a robust and reproducible manner will have to be adopted. Thirdly, prior to commercialization of these signatures, a significant amount of validation will be required. Lastly, statistically powered studies with large, independent patient cohorts will be a prerequisite for acceptance.

The research activity illustrated in this thesis aimed at fulfilling these methodological gaps that still hamper the identification of prognostic and predictive markers and affecting their reliability and reproducibility. Specifically, we addressed aspects related to i) the sample size of analyzed studies (*dataset*) and ii) the computational approaches applied in the discovery process. We developed a bioinformatics strategy to i) integrate multiple, independently generated datasets of tumor specimens with well-annotated clinical data, ii) to exploit this large-scale genomic data, in a retrospective behavior, for elucidating mechanisms of cancer progression and iii) to derive *gene signatures* as models for predicting neo-adjuvant chemotherapy sensitivity or resistance. These computational methods contribute fulfilling gaps in the bioinformatics analysis of microarray data where probe selection, annotation and specificity, comparability of different microarray platforms and signal normalization strategies, still represent a major, and partially unresolved, computational issue when analyzing multiple gene expression datasets.

In summary, the computational pipeline for the combination of multiple datasets is composed of three major steps, i.e., i) re-definition of clinic-pathological and outcome descriptions, ii) probe re-mapping and selection; and iii) integration and normalization of different datasets.

Re-definition of clinic-pathological and outcome descriptions has been conducted carefully considering the clinical annotations of any single study and defining two major types of events, one associated to the metastatic spread (*metastasis*) and one to overall survival (*survival*) and also standardized clinic-pathological variables. Probe re-mapping and selection has been based on the adoption of modified custom Chip Definition Files (custom-CDF) while the integration and normalization of gene expression signals has been obtained applying the *virtual chip* procedure (Bisognin et al., 2010; Fallarino et al., 2010).

The application of this approach allowed constructing a meta-dataset of 3661 gene expression profiles (*samples*) derived from sporadic breast cancer patients' tissues (all hybridized on Affymetrix platforms and with available raw data) arising from the combination of the 27 gene expression datasets (a collection of *samples* deriving from the same experiment). Detailed clinical and outcome information and response to neo-adjuvant chemotherapy were available. To date, this meta-dataset represents the largest collection of integrated gene expression data from fully annotated sporadic breast cancer specimens. Moreover, this meta-dataset allows a statistically robust investigation of cancer subpopulations (i.e., triple negative breast cancer). This specific subtype, also known *basal-like*, has a small incidence (~20%) in overall population and it is an aggressive subtype with early death in younger women. Microarray-based gene expression profiling allows the stratification of breast cancers into molecularly and clinically different subtypes with distinct gene expression patterns based on the activity of specific signaling cascades. In basic and translational research, this technique has become a working model for breast cancer molecular classification and for the definition of effective predictive and prognostic tools. Both these issues are particularly critical in Triple Negative Breast Cancer (TNBC), which still lacks not only of prognostic and therapeutic options, but also of a solid understanding of the molecular mechanisms at the base of its metastatic proclivity. A focus of this research was to identify the TNBCs using molecular subtype classification models based on gene expression data from the breast cancers collection.

Moreover, this thesis addressed to identify predictive gene signatures in triple negative breast cancers. Prediction of response to chemotherapy is a clinically relevant need to improve patient selection for drug administration. An option would be the use of predictive markers of response to distinguish patients who are likely to receive benefits from those who are not, thus sparing predicted poor responders from the significant associated toxicities. Unfortunately, although this is an attractive strategy, suitable biomarkers predicting response to specific chemotherapy agents have, on the whole, remained elusive. Recently, it has been suggested that a single biomarker may not be sufficient for predicting anthracycline response, rather that a multifactorial approach might be better (Desmedt et al., 2011; Di Leo et al, 2011). In this thesis work it was constructed a computational approach to derive *gene signatures* as models for predicting neo-adjuvant chemotherapy sensitivity or resistance in anthracycline treated TNBC patients. *A Consensus Signature* was designed as linear weighted combinations of gene signatures that, singularly, recapitulate independent signaling pathways (e.g., mutp53/p63) or specific molecular mechanisms (i.e., hypoxia, immune function), while, intertwined together, render a more comprehensive molecular model of chemo resistance. The selection of markers extracted from gene signatures with *biological insights*, rather than solely on the basis of gene expression and phenotypic data, without taking into account *a priori* biological knowledge, could dramatically improve the reliability and robustness of prognostic and predictive signatures. Specifically, a *Consensus Signature* was constructed based on five biologically relevant steps required for anthracycline-induced cytotoxicity: i) penetration of the drug into the cancer cell; ii) location of TOP2A, target of anthracycline, within the nucleus; iii) increased TOP2A expression; iv) induction of apoptosis; v) active stromal and immune function. Genes/gene signatures were selected as surrogate measures of each of these components and various combinations of these signatures were assessed for correlation with pathological complete response (pCR) to anthracycline-based chemotherapy without taxane using the cohort of TNBC patients (*design cohort*). The most powerful combination (*ConSig1*) included HIF signature, immune response signature (STAT1), and TOP2A mRNA expression. *ConSig1* demonstrated high correlation with pCR in ROC analyses (AUC = 0.79, p=9.05x10^{-9}), while no correlation with response was seen in a cohort of patients treated with anthracycline plus taxane (*control cohort*), supporting the *ConSig1*'s ability to discriminate sensitive patients from those resistant in a anthracycline-based regime. Testing *ConSig1* in another cohort of TNBC samples (hybridized on different type of microarray platform) from patients treated always with anthracycline-based chemotherapy, it had still predictive power (AUC = 0.65, p=2.4x10^{-2}).

Lastly, in collaboration with the group headed by Giannino Del Sal at the University of Trieste, new insights were gained in the molecular bases of breast cancer stem cell (CSC) malignant properties which are implicated in both treatment resistance and disease relapse. Rustighi and collaborators show that both normal stem cells and CSCs of the breast are controlled by the propyl-isomerase Pin1. Mechanistically, following interaction with Pin1, Notch1 and Notch4, key regulators of cell fate, escape from proteasomal degradation by their major ubiquitin-ligase Fbxw7α. Functionally, we show that Fbxw7α acts as an essential negative regulator of breast CSCs' expansion by restraining Notch activity, but the establishment of a Notch/Pin1 active circuitry opposes this effect, thus promoting breast cancer CSCs self-rewenal, tumor growth and metastasis in vivo. In human breast cancers, despite Fbxw7α expression, high levels of Pin1 sustain Notch signaling, which correlates with poor prognosis. Suppression of Pin1 holds promise in reverting aggressive phenotypes, through CSC exhaustion as well as recovered drug sensitivity carrying relevant implications for therapy of breast cancers.

Chapter 5

References

Adorno, M., Cordenonsi, M., Montagner, M., Dupont, S., Wong, C., Hann, B., Solari, A., Bobisse, S., Rondina, M.B., Guzzardo, V., et al. (2009). A Mutant-p53/Smad Complex Opposes p63 to Empower TGFβ-Induced Metastasis. Cell *137*, 87–98.

Alvarez, J.V., Febbo, P.G., Ramaswamy, S., Loda, M., Richardson, A., and Frank, D.A. (2005). Identification of a genetic signature of activated signal transducer and activator of transcription 3 in human tumors. Cancer Res. *65*, 5054–5062.

Amsen, D., Antov, A., Jankovic, D., Sher, A., Radtke, F., Souabni, A., Busslinger, M., McCright, B., Gridley, T., and Flavell, R.A. (2007). Direct regulation of Gata3 expression determines the T helper differentiation potential of Notch. Immunity *27*, 89–99.

Arriola, E., Rodriguez-Pinilla, S.M., Lambros, M.B.K., Jones, R.L., James, M., Savage, K., Smith, I.E., Dowsett, M., and Reis-Filho, J.S. (2007). Topoisomerase II alpha amplification may predict benefit from adjuvant anthracyclines in HER2 positive early breast cancer. Breast Cancer Res. Treat. *106*, 181–189.

Atchison, F.W., and Means, A.R. (2004). A role for Pin1 in mammalian germ cell development and spermatogenesis. Front. Biosci. J. Virtual Libr. *9*, 3248–3256.

Badve, S., Dabbs, D.J., Schnitt, S.J., Baehner, F.L., Decker, T., Eusebi, V., Fox, S.B., Ichihara, S., Jacquemier, J., Lakhani, S.R., et al. (2011). Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc *24*, 157–167.

Banerjee, S., Reis-Filho, J.S., Ashley, S., Steele, D., Ashworth, A., Lakhani, S.R., and Smith, I.E. (2006). Basal-like breast carcinomas: clinical outcome and response to chemotherapy. J. Clin. Pathol. *59*, 729–735.

Bartlett, J.M.S., Munro, A., Cameron, D.A., Thomas, J., Prescott, R., and Twelves, C.J. (2008). Type 1 receptor tyrosine kinase profiles identify patients with enhanced benefit from anthracyclines in the BR9601 adjuvant breast cancer chemotherapy trial. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. *26*, 5027–5035.

Bhattacharya, S., and Mariani, T.J. (2009). Array of hope: expression profiling identifies disease biomarkers and mechanism. Biochem. Soc. Trans. *37*, 855–862.

Bild, A.H., Yao, G., Chang, J.T., Wang, Q., Potti, A., Chasse, D., Joshi, M.-B., Harpole, D., Lancaster, J.M., Berchuck, A., et al. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature *439*, 353–357.

Bisognin A, Mazza EMC, Ferrari F, Forcato M, Pizzini S, Bortoluzzi S, Bicciato S. The Virtual Chip: a new approach for the integration of different oligonucleotide arrays. RECOMB 2010 – Fourteenth International Conference on Research in Computational Molecular Biology, August 12-15 2010, Lisbon, Portugal

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinforma. Oxf. Engl. *19*, 185–193.

Bouras, T., Pal, B., Vaillant, F., Harburg, G., Asselin-Labat, M.-L., Oakes, S.R., Lindeman, G.J., and Visvader, J.E. (2008). Notch signaling regulates mammary stem cell function and luminal cell-fate commitment. Cell Stem Cell *3*, 429–441.

Bradley A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. (1997).

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., et al. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat. Genet. *29*, 365–371.

Calza, S., Hall, P., Auer, G., Bjöhle, J., Klaar, S., Kronenwett, U., Liu, E.T., Miller, L., Ploner, A., Smeds, J., et al. (2006). Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. Breast Cancer Res. BCR *8*, R34.

Carey, L.A., Perou, C.M., Livasy, C.A., Dressler, L.G., Cowan, D., Conway, K., Karaca, G., Troester, M.A., Tse, C.K., Edmiston, S., et al. (2006). Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. JAMA J. Am. Med. Assoc. *295*, 2492–2502.

Chadwick, N., Zeef, L., Portillo, V., Fennessy, C., Warrander, F., Hoyle, S., and Buckle, A.-M. (2009). Identification of novel Notch target genes in T cell leukaemia. Mol. Cancer *8*, 35.

Chen, Y., Fischer, W.H., and Gill, G.N. (1997). Regulation of the ERBB-2 promoter by RBPJkappa and NOTCH. J. Biol. Chem. *272*, 14110–14114.

Cicalese, A., Bonizzi, G., Pasi, C.E., Faretta, M., Ronzoni, S., Giulini, B., Brisken, C., Minucci, S., Di Fiore, P.P., and Pelicci, P.G. (2009). The tumor suppressor p53 regulates polarity of self-renewing divisions in mammary stem cells. Cell *138*, 1083–1095.

Colozza, M., Azambuja, E., Cardoso, F., Sotiriou, C., Larsimont, D., and Piccart, M.J. (2005). Proliferative markers as prognostic and predictive tools in early breast cancer: where are we now? Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. ESMO *16*, 1723–1739.

Cordenonsi, M., Zanconato, F., Azzolin, L., Forcato, M., Rosato, A., Frasson, C., Inui, M., Montagner, M., Parenti, A.R., Poletti, A., et al. (2011). The Hippo transducer TAZ confers cancer stem cell-related traits on breast cancer cells. Cell *147*, 759–772.

Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H., et al. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. *33*, e175.

Dean, M., Fojo, T., and Bates, S. (2005). Tumour stem cells and drug resistance. Nat. Rev. Cancer *5*, 275–284.

Dempster, A.P., Laird, N.M., and Rubin, D.B. Maximum likelihood from incomplete data via the algorithm. (1977). Journal of the Royal Statistical Society. 39(1):1-38.

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d' Assignies, M.S., et al. (2007). 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. Off. J. Am. Assoc. Cancer Res. *13*, 3207–3214.

Desmedt, C., Haibe-Kains, B., Wirapati, P., Buyse, M., Larsimont, D., Bontempi, G., Delorenzi, M., Piccart, M., and Sotiriou, C. (2008). Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *14*, 5158–5165.

Desmedt, C., Di Leo, A., de Azambuja, E., Larsimont, D., Haibe-Kains, B., Selleslags, J., Delaloge, S., Duhem, C., Kains, J.-P., Carly, B., et al. (2011). Multifactorial approach to predicting resistance to anthracyclines. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. *29*, 1578–1586.

DiMeo, T.A., Anderson, K., Phadke, P., Fan, C., Feng, C., Perou, C.M., Naber, S., and Kuperwasser, C. (2009). A novel lung metastasis signature links Wnt signaling with cancer cell self-renewal and epithelial-mesenchymal transition in basal-like breast cancer. Cancer Res. *69*, 5364–5373.

Ding, Q., Huo, L., Yang, J.-Y., Xia, W., Wei, Y., Liao, Y., Chang, C.-J., Yang, Y., Lai, C.-C., Lee, D.-F., et al. (2008). Down-regulation of myeloid cell leukemia-1 through inhibiting Erk/Pin 1 pathway by sorafenib facilitates chemosensitization in breast cancer. Cancer Res. *68*, 6109–6117.

Dong, J., Feldmann, G., Huang, J., Wu, S., Zhang, N., Comerford, S.A., Gayyed, M.F., Anders, R.A., Maitra, A., and Pan, D. (2007). Elucidation of a universal size-control mechanism in Drosophila and mammals. Cell *130*, 1120–1133.

Dontu, G., Abdallah, W.M., Foley, J.M., Jackson, K.W., Clarke, M.F., Kawamura, M.J., and Wicha, M.S. (2003). In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. Genes Dev. *17*, 1253–1270.

Dontu, G., Jackson, K.W., McNicholas, E., Kawamura, M.J., Abdallah, W.M., and Wicha, M.S. (2004). Role of Notch signaling in cell-fate determination of human mammary stem/progenitor cells. Breast Cancer Res. BCR *6*, R605–615.

Early Breast Cancer Trialists' Collaborative Group (EBCTCG) (2005). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. Lancet *365*, 1687–1717.

Eifel, P., Axelson, J.A., Costa, J., Crowley, J., Curran, W.J., Jr, Deshler, A., Fulton, S., Hendricks, C.B., Kemeny, M., Kornblith, A.B., et al. (2001). National Institutes of Health Consensus Development Conference Statement: adjuvant therapy for breast cancer, November 1-3, 2000. J. Natl. Cancer Inst. *93*, 979–989.

Elston CW, Ellis IO.Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology. 1991 Nov;19(5):403-10

Fallarino, F., Volpi, C., Fazio, F., Notartomaso, S., Vacca, C., Busceti, C., Bicciato, S., Battaglia, G., Bruno, V., Puccetti, P., et al. (2010). Metabotropic glutamate receptor-4 modulates adaptive immunity and restrains neuroinflammation. Nat. Med. *16*, 897–902.

Farmer, P., Bonnefoi, H., Anderle, P., Cameron, D., Wirapati, P., Wirapati, P., Becette, V., André, S., Piccart, M., Campone, M., et al. (2009). A stroma-related gene signature predicts resistance to neoadjuvant chemotherapy in breast cancer. Nat. Med. *15*, 68–74.

Farnie, G., Clarke, R.B., Spence, K., Pinnock, N., Brennan, K., Anderson, N.G., and Bundred, N.J. (2007). Novel cell culture technique for primary ductal carcinoma in situ: role of Notch and epidermal growth factor receptor signaling pathways. J. Natl. Cancer Inst. *99*, 616–627.

Ferrari, F., Bortoluzzi, S., Coppe, A., Sirota, A., Safran, M., Shmoish, M., Ferrari, S., Lancet, D., Danieli, G.A., and Bicciato, S. (2007). Novel definition files for human GeneChips based on GeneAnnot. BMC Bioinformatics *8*, 446.

Fisher RA. Statistical methods for research workers. London: Oliver and Boyd; 1925

Foulkes, W.D., Brunet, J.-S., Stefansson, I.M., Straume, O., Chappuis, P.O., Bégin, L.R., Hamel, N., Goffin, J.R., Wong, N., Trudel, M., et al. (2004). The prognostic implication of the basal-like (cyclin E high/p27 low/p53+/glomeruloid-microvascular-proliferation+) phenotype of BRCA1-related breast cancer. Cancer Res. *64*, 830–835.

Foulkes, W.D., Smith, I.E., and Reis-Filho, J.S. (2010). Triple-negative breast cancer. N. Engl. J. Med. *363*, 1938–1948.

Fulford, L.G., Reis-Filho, J.S., Ryder, K., Jones, C., Gillett, C.E., Hanby, A., Easton, D., and Lakhani, S.R. (2007). Basal-like grade III invasive ductal carcinoma of the breast: patterns of metastasis and long-term survival. Breast Cancer Res. BCR *9*, R4.

Galea, M.H., Blamey, R.W., Elston, C.E., and Ellis, I.O. (1992). The Nottingham Prognostic Index in primary breast cancer. Breast Cancer Res. Treat. *22*, 207–219.

Gautier, L., Cope, L., Bolstad, B.M., and Irizarry, R.A. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. Bioinforma. Oxf. Engl. *20*, 307–315.

Ginestier, C., Hur, M.H., Charafe-Jauffret, E., Monville, F., Dutcher, J., Brown, M., Jacquemier, J., Viens, P., Kleer, C.G., Liu, S., et al. (2007). ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. Cell Stem Cell *1*, 555–567.

Girardini, J.E., Napoli, M., Piazza, S., Rustighi, A., Marotta, C., Radaelli, E., Capaci, V., Jordan, L., Quinlan, P., Thompson, A., et al. (2011). A Pin1/mutant p53 axis promotes aggressiveness in breast cancer. Cancer Cell *20*, 79–91.

Goldhirsch, A., Wood, W.C., Gelber, R.D., Coates, A.S., Thürlimann, B., and Senn, H.-J. (2003). Meeting highlights: updated international expert consensus on the primary therapy of early breast cancer. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. *21*, 3357–3365.

Grudzien, P., Lo, S., Albain, K.S., Robinson, P., Rajan, P., Strack, P.R., Golde, T.E., Miele, L., and Foreman, K.E. (2010). Inhibition of Notch signaling reduces the stem-like population of breast cancer cells and prevents mammosphere formation. Anticancer Res. *30*, 3853–3867.

Hagiwara, H., and Sunada, Y. (2004). Mechanism of taxane neurotoxicity. Breast Cancer Tokyo Jpn. *11*, 82–85.

Hamidi, H., Gustafason, D., Pellegrini, M., and Gasson, J. (2011). Identification of novel targets of CSL-dependent Notch signaling in hematopoiesis. PloS One *6*, e20022.

Han, J., Hendzel, M.J., and Allalunis-Turner, J. (2011). Notch signaling as a therapeutic target for breast cancer treatment? Breast Cancer Res. BCR *13*, 210.

Hanley, J.A., and McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology *143*, 29–36.

Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. Biometrika. 1982; 69:4

Hatzis, C., Pusztai, L., Valero, V., Booser, D.J., Esserman, L., Lluch, A., Vidaurre, T., Holmes, F., Souchon, E., Wang, H., et al. (2011). A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. JAMA J. Am. Med. Assoc. *305*, 1873–1881.

Hicks, D.G., Short, S.M., Prescott, N.L., Tarr, S.M., Coleman, K.A., Yoder, B.J., Crowe, J.P., Choueiri, T.K., Dawson, A.E., Budd, G.T., et al. (2006). Breast cancers with brain metastases are more likely to be estrogen receptor negative, express the basal cytokeratin CK5/6, and overexpress HER2 or EGFR. Am. J. Surg. Pathol. *30*, 1097–1104.

Hu, Z., Fan, C., Oh, D.S., Marron, J.S., He, X., Qaqish, B.F., Livasy, C., Carey, L.A., Reynolds, E., Dressler, L., et al. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics *7*, 96.

Ignatiadis, M., Singhal, S.K., Desmedt, C., Haibe-Kains, B., Criscitiello, C., Andre, F., Loi, S., Piccart, M., Michiels, S., and Sotiriou, C. (2012). Gene Modules and Response to Neoadjuvant Chemotherapy in Breast Cancer Subtypes: A Pooled Analysis. J. Clin. Oncol. *30*, 1996–2004.

Ioannidis, J.P.A., Allison, D.B., Ball, C.A., Coulibaly, I., Cui, X., Culhane, A.C., Falchi, M., Furlanello, C., Game, L., Jurman, G., et al. (2009). Repeatability of published microarray gene expression analyses. Nat. Genet. *41*, 149–155.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostat. Oxf. Engl. *4*, 249–264.

Isaacs, C., Stearns, V., and Hayes, D.F. (2001). New prognostic factors for breast cancer recurrence. Semin. Oncol. *28*, 53–67.

Ivshina, A.V., George, J., Senko, O., Mow, B., Putti, T.C., Smeds, J., Lindahl, T., Pawitan, Y., Hall, P., Nordgren, H., et al. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res. *66*, 10292–10301.

Iwamoto, T., Bianchini, G., Booser, D., Qi, Y., Coutant, C., Shiang, C.Y.-H., Santarpia, L., Matsuoka, J., Hortobagyi, G.N., Symmans, W.F., et al. (2011). Gene pathways associated with prognosis and chemotherapy sensitivity in molecular subtypes of breast cancer. J. Natl. Cancer Inst. *103*, 264–272.

Jensen, P.B., Sørensen, B.S., Sehested, M., Demant, E.J., Kjeldsen, E., Friche, E., and Hansen, H.H. (1993). Different modes of anthracycline interaction with topoisomerase II. Separate structures critical for DNA-cleavage, and for overcoming topoisomerase II-related drug resistance. Biochem. Pharmacol. *45*, 2025–2035.

Juul, N., Szallasi, Z., Eklund, A.C., Li, Q., Burrell, R.A., Gerlinger, M., Valero, V., Andreopoulou, E., Esteva, F.J., Symmans, W.F., et al. (2010). Assessment of an RNA interference screen-derived mitotic and ceramide pathway metagene as a predictor of response to neoadjuvant paclitaxel for primary triple-negative breast cancer: a retrospective analysis of five clinical trials. Lancet Oncol. *11*, 358–365.

Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. 1980. New York; Chichester, Wiley

Kao, K.-J., Chang, K.-M., Hsu, H.-C., and Huang, A.T. (2011). Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. BMC Cancer *11*, 143.

Karn, T., Pusztai, L., Holtrich, U., Iwamoto, T., Shiang, C.Y., Schmidt, M., Müller, V., Solbach, C., Gaetje, R., Hanker, L., et al. (2011). Homogeneous datasets of triple negative breast cancers enable the identification of novel prognostic and predictive signatures. PloS One *6*, e28403.

Karn, T., Pusztai, L., Ruckhäberle, E., Liedtke, C., Müller, V., Schmidt, M., Metzler, D., Wang, J., Coombes, K.R., Gätje, R., et al. (2012). Melanoma antigen family A identified by the bimodality index defines a subset of triple negative breast cancers as candidates for immune response augmentation. Eur. J. Cancer Oxf. Engl. 1990 *48*, 12–23.

Kim, J., Woo, A.J., Chu, J., Snow, J.W., Fujiwara, Y., Kim, C.G., Cantor, A.B., and Orkin, S.H. (2010). A Myc network accounts for similarities between embryonic stem and cancer cell transcription programs. Cell *143*, 313–324.

Kim, M.R., Choi, H.-K., Cho, K.B., Kim, H.S., and Kang, K.W. (2009). Involvement of Pin1 induction in epithelial-mesenchymal transition of tamoxifen-resistant breast cancer cells. Cancer Sci. *100*, 1834–1841.

Konecny, G.E., Meng, Y.G., Untch, M., Wang, H.-J., Bauerfeind, I., Epstein, M., Stieber, P., Vernes, J.-M., Gutierrez, J., Hong, K., et al. (2004). Association between HER-2/neu and vascular endothelial growth factor expression predicts clinical outcome in primary breast cancer patients. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *10*, 1706–1716.

Korde, L.A., Lusa, L., McShane, L., Lebowitz, P.F., Lukes, L., Camphausen, K., Parker, J.S., Swain, S.M., Hunter, K., and Zujewski, J.A. (2010). Gene expression pathway analysis to predict response to neoadjuvant docetaxel and capecitabine for breast cancer. Breast Cancer Res. Treat. *119*, 685–699.

Lee, C.W., Simin, K., Liu, Q., Plescia, J., Guha, M., Khan, A., Hsieh, C.-C., and Altieri, D.C. (2008). A functional Notch-survivin gene signature in basal breast cancer. Breast Cancer Res. BCR *10*, R97.

Di Leo, A., Gancberg, D., Larsimont, D., Tanner, M., Jarvinen, T., Rouas, G., Dolci, S., Leroy, J.-Y., Paesmans, M., Isola, J., et al. (2002). HER-2 amplification and topoisomerase IIalpha gene aberrations as predictive markers in node-positive breast cancer patients randomly treated either with an anthracycline-based therapy or with cyclophosphamide, methotrexate, and 5-fluorouracil. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *8*, 1107–1116.

Di Leo, A., Desmedt, C., Bartlett, J.M.S., Piette, F., Ejlertsen, B., Pritchard, K.I., Larsimont, D., Poole, C., Isola, J., Earl, H., et al. (2011). HER2 and TOP2A as predictive markers for anthracycline-containing chemotherapy regimens as adjuvant treatment of breast cancer: a meta-analysis of individual patient data. Lancet Oncol. *12*, 1134–1142.

Leong, K.G., Niessen, K., Kulic, I., Raouf, A., Eaves, C., Pollet, I., and Karsan, A. (2007). Jagged1-mediated Notch activation induces epithelial-to-mesenchymal transition through Slug-induced repression of E-cadherin. J. Exp. Med. *204*, 2935–2948.

Li, Y., Zou, L., Li, Q., Haibe-Kains, B., Tian, R., Li, Y., Desmedt, C., Sotiriou, C., Szallasi, Z., Iglehart, J.D., et al. (2010). Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer. Nat. Med. *16*, 214–218.

Li, Y., Hibbs, M.A., Gard, A.L., Shylo, N.A., and Yun, K. (2012). Genome-wide analysis of N1ICD/RBPJ targets in vivo reveals direct transcriptional regulation of Wnt, SHH, and hippo pathway effectors by Notch1. Stem Cells Dayt. Ohio *30*, 741–752.

Lin, Y., Lin, S., Watson, M., Trinkaus, K.M., Kuo, S., Naughton, M.J., Weilbaecher, K., Fleming, T.P., and Aft, R.L. (2010). A gene expression signature that predicts the therapeutic response of the basal-like breast cancer to neoadjuvant chemotherapy. Breast Cancer Res. Treat. *123*, 691–699.

Liou, Y.-C., Ryo, A., Huang, H.-K., Lu, P.-J., Bronson, R., Fujimori, F., Uchida, T., Hunter, T., and Lu, K.P. (2002). Loss of Pin1 function in the mouse causes phenotypes resembling cyclin D1-null phenotypes. Proc. Natl. Acad. Sci. U. S. A. *99*, 1335–1340.

Liu, R., Wang, X., Chen, G.Y., Dalerba, P., Gurney, A., Hoey, T., Sherlock, G., Lewicki, J., Shedden, K., and Clarke, M.F. (2007). The prognostic role of a gene signature from tumorigenic breast-cancer cells. N. Engl. J. Med. *356*, 217–226.

Loi, S., Haibe-Kains, B., Desmedt, C., Lallemand, F., Tutt, A.M., Gillet, C., Ellis, P., Harris, A., Bergh, J., Foekens, J.A., et al. (2007). Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. *25*, 1239–1246.

Loi, S., Haibe-Kains, B., Desmedt, C., Wirapati, P., Lallemand, F., Tutt, A.M., Gillet, C., Ellis, P., Ryder, K., Reid, J.F., et al. (2008). Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. BMC Genomics *9*, 239.

Loi, S., Haibe-Kains, B., Majjaj, S., Lallemand, F., Durbecq, V., Larsimont, D., Gonzalez-Angulo, A.M., Pusztai, L., Symmans, W.F., Bardelli, A., et al. (2010). PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor-positive breast cancer. Proc. Natl. Acad. Sci. U. S. A. *107*, 10208–10213.

Lønning, P.E., Knappskog, S., Staalesen, V., Chrisanthar, R., and Lillehaug, J.R. (2007). Breast cancer prognostication and prediction in the postgenomic era. Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. ESMO *18*, 1293–1306.

Lu, K.P., and Zhou, X.Z. (2007). The prolyl isomerase PIN1: a pivotal new twist in phosphorylation signalling and disease. Nat. Rev. Mol. Cell Biol. *8*, 904–916.

Mackay, A., Jones, C., Dexter, T., Silva, R.L.A., Bulmer, K., Jones, A., Simpson, P., Harris, R.A., Jat, P.S., Neville, A.M., et al. (2003). cDNA microarray analysis of genes associated with ERBB2 (HER2/neu) overexpression in human mammary luminal epithelial cells. Oncogene *22*, 2680–2688.

Mackay, A., Urruticoechea, A., Dixon, J.M., Dexter, T., Fenwick, K., Ashworth, A., Drury, S., Larionov, A., Young, O., White, S., et al. (2007). Molecular response to aromatase inhibitor treatment in primary breast cancer. Breast Cancer Res. BCR *9*, R37.

Malanchi, I., Santamaria-Martínez, A., Susanto, E., Peng, H., Lehr, H.-A., Delaloye, J.-F., and Huelsken, J. (2012). Interactions between cancer stem cells and their niche govern metastatic colonization. Nature *481*, 85–89.

Mani, K.M., Lefebvre, C., Wang, K., Lim, W.K., Basso, K., Dalla-Favera, R., and Califano, A. (2008). A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. Mol. Syst. Biol. *4*, 169.

Marchiò, C., Natrajan, R., Shiu, K., Lambros, M., Rodriguez-Pinilla, S., Tan, D., Lord, C., Hungermann, D., Fenwick, K., Tamber, N., et al. (2008). The genomic profile of HER2-amplified breast cancers: the influence of ER status. J. Pathol. *216*, 399–407.

Martin, M., Romero, A., Cheang, M.C.U., López García-Asenjo, J.A., García-Saenz, J.A., Oliva, B., Román, J.M., He, X., Casado, A., de la Torre, J., et al. (2011). Genomic predictors of response to doxorubicin versus docetaxel in primary breast cancer. Breast Cancer Res. Treat. *128*, 127–136.

Mattarollo, S.R., Loi, S., Duret, H., Ma, Y., Zitvogel, L., and Smyth, M.J. (2011). Pivotal Role of Innate and Adaptive Immunity in Anthracycline Chemotherapy of Established Tumors. Cancer Res. *71*, 4809–4820.

Mauri, D., Pavlidis, N., and Ioannidis, J.P.A. (2005). Neoadjuvant versus adjuvant systemic treatment in breast cancer: a meta-analysis. J. Natl. Cancer Inst. *97*, 188–194.

Mazzone, M., Selfors, L.M., Albeck, J., Overholtzer, M., Sale, S., Carroll, D.L., Pandya, D., Lu, Y., Mills, G.B., Aster, J.C., et al. (2010). Dose-dependent induction of distinct phenotypic responses to Notch pathway activation in mammary epithelial cells. Proc. Natl. Acad. Sci. U. S. A. *107*, 5012–5017.

McShane, L.M., Altman, D.G., Sauerbrei, W., Taube, S.E., Gion, M., Clark, G.M., and Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics (2005). REporting recommendations for tumor MARKer prognostic studies (REMARK). Nat. Clin. Pract. Urol. *2*, 416–422.

Merritt, W.M., Lin, Y.G., Han, L.Y., Kamat, A.A., Spannuth, W.A., Schmandt, R., Urbauer, D., Pennacchio, L.A., Cheng, J.-F., Nick, A.M., et al. (2008). Dicer, Drosha, and outcomes in patients with ovarian cancer. N. Engl. J. Med. *359*, 2641–2650.

Miller, L.D., Smeds, J., George, J., Vega, V.B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E.T., et al. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc. Natl. Acad. Sci. U. S. A. *102*, 13550–13555.

Minn, A.J., Gupta, G.P., Siegel, P.M., Bos, P.D., Shu, W., Giri, D.D., Viale, A., Olshen, A.B., Gerald, W.L., and Massagué, J. (2005). Genes that mediate breast cancer metastasis to lung. Nature *436*, 518–524.

Minn, A.J., Gupta, G.P., Padua, D., Bos, P., Nguyen, D.X., Nuyten, D., Kreike, B., Zhang, Y., Wang, Y., Ishwaran, H., et al. (2007). Lung metastasis genes couple breast tumor size and metastatic spread. Proc. Natl. Acad. Sci. U. S. A. *104*, 6740–6745.

Miyake, T., Nakayama, T., Naoi, Y., Yamamoto, N., Otani, Y., Kim, S.J., Shimazu, K., Shimomura, A., Maruyama, N., Tamaki, Y., et al. (2012). GSTP1 expression predicts poor pathological complete response to neoadjuvant chemotherapy in ER-negative breast cancer. Cancer Sci. *103*, 913–920.

Montagner, M., Enzo, E., Forcato, M., Zanconato, F., Parenti, A., Rampazzo, E., Basso, G., Leo, G., Rosato, A., Bicciato, S., et al. (2012). SHARP1 suppresses breast cancer metastasis by promoting degradation of hypoxia-inducible factors. Nature *487*, 380–384.

Olivotto, I.A., Bajdik, C.D., Ravdin, P.M., Speers, C.H., Coldman, A.J., Norris, B.D., Davis, G.J., Chia, S.K., and Gelmon, K.A. (2005). Population-based validation of the prognostic model ADJUVANT! for early breast cancer. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. *23*, 2716–2725.

Oswald, F., Liptay, S., Adler, G., and Schmid, R.M. (1998). NF-kappaB2 is a putative target gene of activated Notch-1 via RBP-Jkappa. Mol. Cell. Biol. *18*, 2077–2088.

Ota, M., and Sasaki, H. (2008). Mammalian Tead proteins regulate cell proliferation and contact inhibition as transcriptional mediators of Hippo signaling. Dev. Camb. Engl. *135*, 4059–4069.

Padua, D., Zhang, X.H.-F., Wang, Q., Nadal, C., Gerald, W.L., Gomis, R.R., and Massagué, J. (2008). TGFbeta primes breast tumors for lung metastasis seeding through angiopoietin-like 4. Cell *133*, 66–77.

Palomero, T., Lim, W.K., Odom, D.T., Sulis, M.L., Real, P.J., Margolin, A., Barnes, K.C., O'Neil, J., Neuberg, D., Weng, A.P., et al. (2006). NOTCH1 directly regulates c-MYC and activates a feed-forward-loop transcriptional network promoting leukemic cell growth. Proc. Natl. Acad. Sci. U. S. A. *103*, 18261–18266.

Park, B.K., Zhang, H., Zeng, Q., Dai, J., Keller, E.T., Giordano, T., Gu, K., Shah, V., Pei, L., Zarbo, R.J., et al. (2007). NF-kappaB in breast cancer cells promotes osteolytic bone metastasis by inducing osteoclastogenesis via GM-CSF. Nat. Med. *13*, 62–69.

Parker, J.S., Mullins, M., Cheang, M.C.U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. J. Clin. Oncol. *27*, 1160–1167.

Pawitan, Y., Bjöhle, J., Amler, L., Borg, A.-L., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., et al. (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res. BCR *7*, R953–964.

Pece, S., Tosoni, D., Confalonieri, S., Mazzarol, G., Vecchi, M., Ronzoni, S., Bernard, L., Viale, G., Pelicci, P.G., and Di Fiore, P.P. (2010). Biological and molecular heterogeneity of breast cancers correlates with their cancer stem cell content. Cell *140*, 62–73.

Penault-Llorca, F., Bilous, M., Dowsett, M., Hanna, W., Osamura, R.Y., Rüschoff, J., and van de Vijver, M. (2009). Emerging technologies for assessing HER2 amplification. Am. J. Clin. Pathol. *132*, 539–548.

Perou, C.M. (2010). Molecular stratification of triple-negative breast cancers. The Oncologist *15 Suppl 5*, 39–48.

Perou, C.M., Sørlie, T., Eisen, M.B., Rijn, M. van de, Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747–752.

Piccart-Gebhart, M.J., Procter, M., Leyland-Jones, B., Goldhirsch, A., Untch, M., Smith, I., Gianni, L., Baselga, J., Bell, R., Jackisch, C., et al. (2005). Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. N. Engl. J. Med. *353*, 1659–1672.

Polyak, K., and Weinberg, R.A. (2009). Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. Nat. Rev. Cancer *9*, 265–273.

Pommier, Y., Leo, E., Zhang, H., and Marchand, C. (2010). DNA topoisomerases and their poisoning by anticancer and antibacterial drugs. Chem. Biol. *17*, 421–433.

Popovici, V., Chen, W., Gallas, B.G., Hatzis, C., Shi, W., Samuelson, F.W., Nikolsky, Y., Tsyganova, M., Ishkin, A., Nikolskaya, T., et al. (2010). Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. Breast Cancer Res. BCR *12*, R5.

Ben-Porath, I., Thomson, M.W., Carey, V.J., Ge, R., Bell, G.W., Regev, A., and Weinberg, R.A. (2008). An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nat. Genet. *40*, 499–507.

Press, M.F., Sauter, G., Buyse, M., Bernstein, L., Guzman, R., Santiago, A., Villalobos, I.E., Eiermann, W., Pienkowski, T., Martin, M., et al. (2011). Alteration of topoisomerase II-alpha gene in human breast cancer: association with responsiveness to anthracycline-based chemotherapy. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. *29*, 859–867.

Pusztai, L., Mazouni, C., Anderson, K., Wu, Y., and Symmans, W.F. (2006). Molecular classification of breast cancer: limitations and potential. The Oncologist *11*, 868–877.

Rakha, E.A., Putti, T.C., Abd El-Rehim, D.M., Paish, C., Green, A.R., Powe, D.G., Lee, A.H., Robertson, J.F., and Ellis, I.O. (2006). Morphological and immunophenotypic analysis of breast carcinomas with basal and myoepithelial differentiation. J. Pathol. *208*, 495–506.

Ranganathan, P., Weaver, K.L., and Capobianco, A.J. (2011). Notch signalling in solid tumours: a little bit of everything but not all the time. Nat. Rev. Cancer *11*, 338–351.

Rangarajan, A., Talora, C., Okuyama, R., Nicolas, M., Mammucari, C., Oh, H., Aster, J.C., Krishna, S., Metzger, D., Chambon, P., et al. (2001). Notch signaling is a direct determinant of keratinocyte growth arrest and entry into differentiation. EMBO J. *20*, 3427–3436.

Raouf, A., Zhao, Y., To, K., Stingl, J., Delaney, A., Barbara, M., Iscove, N., Jones, S., McKinney, S., Emerman, J., et al. (2008). Transcriptome analysis of the normal human mammary cell commitment and differentiation process. Cell Stem Cell *3*, 109–118.

Reya, T., Morrison, S.J., Clarke, M.F., and Weissman, I.L. (2001). Stem cells, cancer, and cancer stem cells. Nature *414*, 105–111.

Rodríguez-Pinilla, S.M., Sarrió, D., Honrado, E., Hardisson, D., Calero, F., Benitez, J., and Palacios, J. (2006). Prognostic significance of basal-like phenotype and fascin expression in node-negative invasive breast carcinomas. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *12*, 1533–1539.

Rody, A., Karn, T., Liedtke, C., Pusztai, L., Ruckhaeberle, E., Hanker, L., Gaetje, R., Solbach, C., Ahr, A., Metzler, D., et al. (2011). A clinically relevant gene signature in triple negative and basal-like breast cancer. Breast Cancer Res. BCR *13*, R97.

Ronchini, C., and Capobianco, A.J. (2001). Induction of cyclin D1 transcription and CDK2 activity by Notch(ic): implication for cell cycle disruption in transformation by Notch(ic). Mol. Cell. Biol. *21*, 5925–5934.

Rosenthal, D.T., Zhang, J., Bao, L., Zhu, L., Wu, Z., Toy, K., Kleer, C.G., and Merajver, S.D. (2012). RhoC impacts the metastatic potential and abundance of breast cancer stem cells. PloS One *7*, e40979.

Rustighi, A., Zannini, A., Tiberi, L., Sommaggio, R., Piazza, S., Sorrentino, G., Nuzzo, S., Tuscano, A., Eterno, V., Benvenuti, F., et al. (2014). Prolyl-isomerase Pin1 controls normal and cancer stem cells of the breast. EMBO Mol. Med. *6*, 99–119.

Ryo, A., Liou, Y.-C., Wulf, G., Nakamura, M., Lee, S.W., and Lu, K.P. (2002). PIN1 is an E2F target gene essential for Neu/Ras-induced transformation of mammary epithelial cells. Mol. Cell. Biol. *22*, 5281–5295.

Sabatier, R., Finetti, P., Cervera, N., Lambaudie, E., Esterni, B., Mamessier, E., Tallet, A., Chabannon, C., Extra, J.-M., Jacquemier, J., et al. (2011a). A gene expression signature identifies two prognostic subgroups of basal breast cancer. Breast Cancer Res. Treat. *126*, 407–420.

Sabatier, R., Finetti, P., Adelaide, J., Guille, A., Borg, J.-P., Chaffanet, M., Lane, L., Birnbaum, D., and Bertucci, F. (2011b). Down-regulation of ECRG4, a candidate tumor suppressor gene, in human breast cancer. PloS One *6*, e27656.

Sahlgren, C., Gustafsson, M.V., Jin, S., Poellinger, L., and Lendahl, U. (2008). Notch signaling mediates hypoxia-induced tumor cell migration and invasion. Proc. Natl. Acad. Sci. U. S. A. *105*, 6392–6397.

Sarmento, L.M., Huang, H., Limon, A., Gordon, W., Fernandes, J., Tavares, M.J., Miele, L., Cardoso, A.A., Classon, M., and Carlesso, N. (2005). Notch1 modulates timing of G1-S progression by inducing SKP2 transcription and p27 Kip1 degradation. J. Exp. Med. *202*, 157–168.

Scarff, R.W. and Torloni, H. Histological typing of breast tumors. (1968).

International histological classification of tumors. 2(2):13-20,1968

Schmidt, M., Böhm, D., von Törne, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H.-A., Hengstler, J.G., Kölbl, H., and Gehrmann, M. (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. Cancer Res. *68*, 5405–5413.

Schwarz, G. Estimating the dimension of a model. (1978). Annals of Statistics, 6, 461-464

Shi, L., Campbell, G., Jones, W.D., Campagne, F., Wen, Z., Walker, S.J., Su, Z., Chu, T.-M., Goodsaid, F.M., Pusztai, L., et al. (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat. Biotechnol. *28*, 827–838.

Shipitsin, M., Campbell, L.L., Argani, P., Weremowicz, S., Bloushtain-Qimron, N., Yao, J., Nikolskaya, T., Serebryiskaya, T., Beroukhim, R., Hu, M., et al. (2007). Molecular definition of breast tumor heterogeneity. Cancer Cell *11*, 259–273.

Siegel, R., Ward, E., Brawley, O., and Jemal, A. (2011). Cancer statistics, 2011: the impact of eliminating socioeconomic and racial disparities on premature cancer deaths. CA. Cancer J. Clin. *61*, 212–236.

Slamon, D., Eiermann, W., Robert, N., Pienkowski, T., Martin, M., Press, M., Mackey, J., Glaspy, J., Chan, A., Pawlicki, M., et al. (2011). Adjuvant trastuzumab in HER2-positive breast cancer. N. Engl. J. Med. *365*, 1273–1283.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc. Natl. Acad. Sci. U. S. A. *100*, 8418–8423.

Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. U. S. A. *98*, 10869–10874.

Sotiriou, C., and Desmedt, C. (2006). Gene expression profiling in breast cancer. Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. ESMO *17 Suppl 10*, x259–262.

Sotiriou, C., and Piccart, M.J. (2007). Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nat. Rev. Cancer *7*, 545–553.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., et al. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J. Natl. Cancer Inst. *98*, 262–272.

Steeg, P.S., and Theodorescu, D. (2008). Metastasis: a therapeutic target for cancer. Nat. Clin. Pract. Oncol. *5*, 206–219.

Stingl, J., and Caldas, C. (2007). Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. Nat. Rev. Cancer *7*, 791–799.

Stingl, J., Eirew, P., Ricketson, I., Shackleton, M., Vaillant, F., Choi, D., Li, H.I., and Eaves, C.J. (2006). Purification and unique properties of mammary epithelial stem cells. Nature *439*, 993–997.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550.

Swets, J.A. Signal detection theory and ROC analysis in psychology and diagnostics: collected papers. (1996). Lawrence Erlbaum Associates, Mahwah, N.J.

Tabchy, A., Valero, V., Vidaurre, T., Lluch, A., Gomez, H., Martin, M., Qi, Y., Barajas-Figueroa, L.J., Souchon, E., Coutant, C., et al. (2010). Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *16*, 5351–5361.

Teicher, B.A. (1994). Hypoxia and drug resistance. Cancer Metastasis Rev. *13*, 139–168.

Teschendorff, A.E., Miremadi, A., Pinder, S.E., Ellis, I.O., and Caldas, C. (2007). An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biol. *8*, R157.

Thiery, J.P., Acloque, H., Huang, R.Y.J., and Nieto, M.A. (2009). Epithelial-mesenchymal transitions in development and disease. Cell *139*, 871–890.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. U. S. A. *99*, 6567–6572.

Tsuda, H., Takarabe, T., Hasegawa, F., Fukutomi, T., and Hirohashi, S. (2000). Large, central acellular zones indicating myoepithelial tumor differentiation in high-grade invasive ductal carcinomas as markers of predisposition to lung and brain metastases. Am. J. Surg. Pathol. *24*, 197–202.

Turner, J.G., Engel, R., Derderian, J.A., Jove, R., and Sullivan, D.M. (2004). Human topoisomerase IIalpha nuclear export is mediated by two CRM-1-dependent nuclear export signals. J. Cell Sci. *117*, 3061–3071.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U. S. A. *98*, 5116–5121.

Uchida, T., Takamiya, M., Takahashi, M., Miyashita, H., Ikeda, H., Terada, T., Matsuo, Y., Shirouzu, M., Yokoyama, S., Fujimori, F., et al. (2003). Pin1 and Par14 peptidyl prolyl isomerase inhibitors block cell proliferation. Chem. Biol. *10*, 15–24.

Van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*, 530–536.

Van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A.M., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. N. Engl. J. Med. *347*, 1999–2009.

Visvader, J.E., and Lindeman, G.J. (2012). Cancer stem cells: current status and evolving complexities. Cell Stem Cell *10*, 717–728.

Wang, Y., Klijn, J.G.M., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J., et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet *365*, 671–679.

Weigelt, B., Baehner, F.L., and Reis-Filho, J.S. (2010). The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. J. Pathol. *220*, 263–280.

Weng, A.P., Millholland, J.M., Yashiro-Ohtani, Y., Arcangeli, M.L., Lau, A., Wai, C., Del Bianco, C., Rodriguez, C.G., Sai, H., Tobias, J., et al. (2006). c-Myc is an important direct target of Notch1 in T-cell acute lymphoblastic leukemia/lymphoma. Genes Dev. *20*, 2096–2109.

Van de Wetering, M., Sancho, E., Verweij, C., de Lau, W., Oving, I., Hurlstone, A., van der Horn, K., Batlle, E., Coudreuse, D., Haramis, A.P., et al. (2002). The beta-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. Cell *111*, 241–250.

Wirapati, P., Sotiriou, C., Kunkel, S., Farmer, P., Pradervand, S., Haibe-Kains, B., Desmedt, C., Ignatiadis, M., Sengstag, T., Schütz, F., et al. (2008). Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Res. BCR *10*, R65.

Wulf, G., Finn, G., Suizu, F., and Lu, K.P. (2005). Phosphorylation-specific prolyl isomerization: is there an underlying theme? Nat. Cell Biol. *7*, 435–441.

Wulf, G.M., Ryo, A., Wulf, G.G., Lee, S.W., Niu, T., Petkova, V., and Lu, K.P. (2001). Pin1 is overexpressed in breast cancer and cooperates with Ras signaling in increasing the transcriptional activity of c-Jun towards cyclin D1. EMBO J. *20*, 3459–3472.

Xing, F., Kobayashi, A., Okuda, H., Watabe, M., Pai, S.K., Pandey, P.R., Hirota, S., Wilber, A., Mo, Y.-Y., Moore, B.E., et al. (2013). Reactive astrocytes promote the metastatic growth of breast cancer stem-like cells by activating Notch signalling in brain. EMBO Mol. Med. *5*, 384–396.

Xu, K., Usary, J., Kousis, P.C., Prat, A., Wang, D.-Y., Adams, J.R., Wang, W., Loch, A.J., Deng, T., Zhao, W., et al. (2012). Lunatic fringe deficiency cooperates with the Met/Caveolin gene amplicon to induce basal-like breast cancer. Cancer Cell *21*, 626–641.

Yau, C., and Benz, C.C. (2008). Genes responsive to both oxidant stress and loss of estrogen receptor function identify a poor prognosis group of estrogen receptor positive primary breast cancers. Breast Cancer Res. BCR *10*, R61.

Yeh, E.S., and Means, A.R. (2007). PIN1, the cell cycle and cancer. Nat. Rev. Cancer *7*, 381–388.

Yu, F., Yao, H., Zhu, P., Zhang, X., Pan, Q., Gong, C., Huang, Y., Hu, X., Su, F., Lieberman, J., et al. (2007). let-7 regulates self renewal and tumorigenicity of breast cancer cells. Cell *131*, 1109–1123.

Zhang, H., Liu, C.-Y., Zha, Z.-Y., Zhao, B., Yao, J., Zhao, S., Xiong, Y., Lei, Q.-Y., and Guan, K.-L. (2009a). TEAD transcription factors mediate the function of TAZ in cell growth and epithelial-mesenchymal transition. J. Biol. Chem. *284*, 13355–13362.

Zhang, Y., Sieuwerts, A.M., McGreevy, M., Casey, G., Cufer, T., Paradiso, A., Harbeck, N., Span, P.N., Hicks, D.G., Crowe, J., et al. (2009b). The 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy. Breast Cancer Res. Treat. *116*, 303–309.

Zhao, B., Ye, X., Yu, J., Li, L., Li, W., Li, S., Yu, J., Lin, J.D., Wang, C.-Y., Chinnaiyan, A.M., et al. (2008). TEAD mediates YAP-dependent gene induction and growth control. Genes Dev. *22*, 1962–1971.

Zhou, Y., Yau, C., Gray, J.W., Chew, K., Dairkee, S.H., Moore, D.H., Eppenberger, U., Eppenberger-Castori, S., and Benz, C.C. (2007). Enhanced NF kappa B and

AP-1 transcriptional activity associated with antiestrogen resistant breast cancer. BMC Cancer *7*, 59.

Zitvogel, L., Apetoh, L., Ghiringhelli, F., André, F., Tesniere, A., and Kroemer, G. (2008a). The anticancer immune response: indispensable for therapeutic success? J. Clin. Invest. *118*, 1991–2001.

Zitvogel, L., Apetoh, L., Ghiringhelli, F., and Kroemer, G. (2008b). Immunological aspects of cancer chemotherapy. Nat. Rev. Immunol. *8*, 59–73.