Inaugural Dissertation

Submitted to the

Combined Faculties for the Natural Sciences and for

Mathematics

of the

Ruperto-Carola University of Heidelberg,

Germany

for the degree of

Doctor of Natural Sciences

Presented by

Dipl.-Biol. Olaf Thürigen

Born in Ludwigshafen

Oral Examination: January 30, 2008

Development of a Procedure for Genome-wide Expression Profiling from Minute Tissue Samples and Application in Mammary Carcinoma: Gene Activity Patterns Unveiling Molecular Pathways and Predicting Clinical Response

Referees:

Prof. Dr. Werner Buselmaier

Prof. Dr. Peter Lichter

"The harder you work, the luckier you get."

Gary Player

Acknowledgements **(in order of appearance)**

- My family, especially my parents, for building and being my solid foundation,
- Jörg Schlingemann for introduction to Peter, our fruitful collaboration succeeding in TAcKLE, and good friendship,
- Peter Lichter for giving me the opportunity to earn my doctorate and his persevering attitude in the slopes of this (or any) endeavor,
- Meinhard Hahn, for schooling my eye for the detail and using his own copiously,
- Gunnar Wrobel and Felix Kokocinski for profound introduction into "R" and valuable help with management of large data,
- Heidi Schlingemann (then Kramer) for great help and good fun in the laboratory,
- Grischa Tödt for his broad and sophisticated (bio)informatics deployment, concerning both our collaborative work and home cinema systems,
- Björn Tews for his superb introduction into RQ-PCR and an example in sportsmanship,
- Daniel Haag for supporting my work and his wonderful cocktail recipes,
- Andreas Schneeweiß and Hans-Peter Sinn for providing excellent material and data, resulting in successful collaborations,
- Benedikt Brors and Patrick Warnat for help in identifying relevant and predictive genes and pathways,
- Magdalena Schlotter for patiently answering my newer ending "Bambi questions",
- Prof. Dr. Werner Buselmaier, for taking my supervision, being my first referee and the chairperson of the examining committee,
- All my former and current colleagues in the expression lab, in the Dept. of Molecular Genetics, and of course all my friends at the DKFZ, and
- Astrid Riehl for discussion of this thesis, and our future.

Summary

In this thesis, a novel procedure for linear amplification of messenger RNA (mRNA) molecules and labeling with fluorescently modified nucleotides was developed, that can be used to perform genome-wide expression analysis from minute tissue samples using microarrays of long gene-specific oligonucleotide DNA probes. The procedure was then applied to analyze core needle biopsies taken at time of diagnosis from tumors of female primary breast carcinoma patients. Upon receiving chemotherapy consisting of gemcitabine, epirubicin and docetaxel, the patients were classified according to their response to the chemotherapy into responders, defined as patients with a pathological complete remission of the tumor, and non-responders, defined as patients with no change or pathological partial remission.

The gene expression profiles of the tumors from these patients were then bioinformatically processed and analyzed to identify a gene expression signature, which could be used to predict the response of the patients. Additionally, this gene signature was inspected for the significantly enriched pathways and biological processes, and a subset of genes was analyzed in the patient's biopsies with respect to RNA expression as validated by real-time quantitative polymerase chain reaction and protein expression as measured by immuno-histochemistry.

The gene expression signature contained 512 genes, which allow a prediction of the patient response with an overall accuracy of 88%, a sensitivity of 78% and a specificity of 90%. Signaling pathways and biological processes identified with significant enrichment in the gene set were the Ras pathway, TGF β signaling, DNA damage response and apoptosis. From these pathways, the genes *DAPK2*, *BAMBI*, *LMO4* and *SMAD3* could be validated by RQ-PCR, but not *SRC*. In protein analysis by IHC, BAMBI was strongly associated with the patient's outcome, while BMP4, LMO4, SMAD3 and SRC were not directly associated. Additionally, BAMBI protein expression showed strong relationship with BRCA1 expression in the primary female breast carcinoma.

Taken together, these results show the applicability of the novel developed procedure for amplification and labeling of mRNA for genome-wide gene expression analysis with the long oligonucleotide microarray technique and the successful use in biological and clinical investigations. The analysis of gene expression profiles of the primary breast tumors revealed an association of the Ras pathway, TGF β signaling, DNA damage response and apoptosis with the outcome of the patients after chemotherapy, as well as associations of several genes within these pathways and biological processes.

Zusammenfassung

In der vorliegenden Dissertation wurde eine neue Methode zur Amplifikation von Boten-RNS und der Markierung mit fluoreszierenden Nukleotiden entwickelt, die sich zur Erstellung von Genexpressions-Profilen aus sehr kleinen Gewebeproben mit Hilfe genspezifischer Proben aus langen DNS-Oligonukleotiden auf Microarrays eignet. Nachfolgend wurde diese Methode angewendet, um Feinnadel-Biopsien aus Mamma-Karzinomen zu untersuchen, die den Patientinnen bei Diagnose entnommen worden waren. Nachdem die Patientinnen eine Kombinations-Chemotherapie aus Gemcitabin, Epirubicin und Docetaxel erhalten hatten, wurden sie je nach Ansprechen in "Responder", definiert als Patientinnen mit pathologisch gesicherter kompletter Remission, oder "Non-Responder", definiert als Patientinnen ohne Veränderung oder mit partialem Rückgang des Tumors, klassifiziert.

Die Genexpressions-Profile dieser Tumoren wurden mit Hilfe bioinformatischer Methoden verarbeitet und analysiert, um eine Gensignatur zu identifzieren, die eine Vorhersage des Therapieansprechens erlaubt. Zusätzlich wurde diese Gensignatur auf signifikant überrepräsentierte Signalwege und biologische Prozesse hin untersucht. Ein Teil der Signaturgene wurde in den Biopsien der Patientinnen bezüglich der RNS- und Protein-Expression mit Hilfe von quantitativer Echtzeit-PCR bzw. immunhistochemischer Färbungen analysiert.

Die ermittelte Genexpressions-Signatur enthält 512 Gene, und ermöglicht die Vorhersage des Therapieansprechens mit einer Gesamtgenauigkeit von 88%, einer Sensitivität von 78% und einer Spezifität von 90%. Als für das Ansprechen relevante Signalwege und biologische Prozesse wurden der Ras-Signalweg, die TGF-β-Kaskade, Antwortprozesse bei DNS-Schädigungen sowie der Apoptosemechanismus identifiziert. Aus diesen Signalwegen konnten die Gene *DAPK2*, *BAMBI*, *LMO4* und *SMAD3* durch qEZ-PCR validiert werden, nicht jedoch die Expression von *SRC*. Die Proteinanalyse zeigte eine starke Assoziation von BAMBI mit dem Therapieansprechen, während BMP4, LMO4, SMAD3 sowie SRC nicht direkt assoziiert waren. Zudem wurde ein starker Zusammenhang zwischen der Proteinexpression von BAMBI und BRCA1 in den primären Brusttumoren festgestellt.

Zusammengefaßt zeigen die Ergebnisse die Einsetzbarkeit der neu entwickelten Methode zur Amplifikation und Markierung von Boten-RNS für die genomweite Expressionanalyse mit der verwendeten Microarray-Technik, sowie die erfolgreiche Anwendung der Methode zur Untersuchung biologischer und klinischer Fragestellungen. Die Analyse der Genexpressions-Profile der Primärtumoren von Brustkrebspatientinnen zeigte Assoziationen des Ras-Signalwegs, der TGF-β-Kaskade, der Antwortprozesse bei DNS-Schädigungen sowie des Apoptosemechanismus mit dem Ansprechen der Patientinnen auf die Chemotherapie. Zudem wurden Abhängigkeiten zwischen diesen Signalwegen und biologischen Prozessen anhand verschiedener Gene nachgewiesen.

Abbreviations

'	minute(s)
"	second(s)
@	at
µg	microgram
µl	microliter
µM	micromolar
aRNA	antisense ribonucleic acid
BSA	bovine serum albumine
cDNA	complementary DNA
cRNA	complementary RNA
Da	dalton
ddH_2O	aqua bidestillata
DKFZ	Deutsches Krebsforschungszentrum, German Cancer Research Center
DMEM	Dulbecco's Modified Eagle Medium
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
dNTP	deoxy-nucleotide triphosphate
dsDNA	double strand DNA
DTT	dithiothreitol (Cleland's reagent)
dUTP	deoxy-uracil triphosphate
EDTA	ethylendiamine tetraacetate
FBNC	formamide, betaine, nitrocellulose (buffer)
FCS	fetal calf serum
g	gram
GFP	green fluorescent protein
h	hour
IVT	*in-vitro* transcription
kDa	kilodalton
l	liter
MDA	mixture discriminant analysis
mg	milligram
ml	milliliter
mM	millimolar
M-MLV	Moloney Murine Leukemia Virus
mRNA	messenger RNA
ng	nanogram
nl	nanoliter
nM	nanomolar
nm	nanometer
o/n	over night
p.A.	per analysis, high purity grade for chemicals
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
RNA	ribonucleic acid
rpm	rotations per minute
RPMI	RPMI-1640 was developed by Moore *et al.* at Roswell Park Memorial Institute, hence the acronym RPMI
RQ-PCR	realtime-quantitative PCR
rRNA	ribosomal RNA
RT	reverse transcription
SDS	sodium dodecylsulfate
SPA	single primer amplification
SSC	saline-buffered sodium chloride
ssDNA	second strand DNA

TE	Tris-EDTA buffer
temp.	temperature
Tris	trishydroxymethyl-aminomethane
tRNA	transfer RNA
TS	template switch
v/v	volume/volume (dilutions)
w/v	weight/volume (suspensions)
w/w	weight/weight (mixtures)
WHO	World Health Organization

1. Introduction

1.1. Cancer

Cancer is a very heterogeneous disease, comprising more than 100 different types of malignant tumors. Concurrently, it is the second leading cause of death, with a rate of 22.7% of all deaths worldwide in 2003.[1] Only the cardiovascular diseases, with a share of 28%, have a larger percentage.

1.1.1. Development of Cancer

The formation of a tumor depends on the transformation of at least one cell within the organism. The transformation can be fostered by cancerous agents, which due to their DNA mutating effect are also called mutagens. Such substances or media include different toxins, like those contained in tobacco smoke, free radicals like reactive oxygen or nitric oxide species, but they also include physically damaging sources like UV light or ionizing irradiation. Other sources of degeneration on the level of DNA include different viruses, like hepatitis B or C viruses (HBV, HCV) or human papilloma viruses (HPV). DNA damage can also occur on during chromosome segregation, leading to aneuploidy or translocations of chromosome parts.

Many of these events happen often during the life time of a cell. Even in an environment that is free of mutagens, mutations will occur spontaneously at an estimated rate of about 10^{-6} mutations per gene per cell division. Compared to the total number of cell divisions, estimated as 10^{16} in the course of a lifetime, this equates to approximately 10^{10} mutation events per gene in the whole human body.[2]

Nonetheless, most of these events do not lead to a cancerous cell. First of all, cells possess DNA repair mechanisms that check and repair single nucleotide mutations, e.g. during replication. Secondly, not all mutations actually lead to an amino-acid change in the protein, or the change translates but does not lead to a functional change. Thirdly, if the function of the protein or even the cell is severely restricted, it usually leads to a cell death program called apoptosis. And lastly, few cells actually live for the entire time span of the organism, as most somatic cells have a turnover rate and also stem cells are

limited in the number of cell divisions they are allowed to make by restriction mechanisms, e.g. through the length of their telomeres.
Only if the deteriorations are severe, like chromosomal translocations, genetic mutations that lead to a defect in the above-mentioned safeguard mechanisms themselves or several mutations that happen in a short interval, will they likely cause transformation and subsequently can lead to tumor formation.
In the year 1971, Alfred G. Knudsen proposed his model based on statistical analysis of retinoblastoma, which is today called the "two-hit" model.[3,4] In short, his hypothesis implies that dominantly inherited predisposition to cancer entails a germline mutation, while tumorigenesis requires a somatic mutation of the second copy of the respective gene. Only by the manifestation of both mutations, the early and frequent development retinoblastoma could be explained. This very specific finding is still seen as a basic but key concept in tumor genetics, even if certain modifications are necessary. As explained before, a single somatic mutation mostly does not lead to cancer. Conversely, if for example the DNA repair mechanisms are disabled by mutations in the respective repair genes, other mutations can easily manifest and lead to a degeneration of the cell, e.g. resulting in its micro-environmental survival advantage. Another example is a mutation leading to the activation of the *hTERT* gene, which encodes the human telomerase protein. The telomerase is capable of lengthening the telomeres, the ends of chromosomes, which normally are gradually lost by cell division and finally initiate the death of the cells after their complete breakdown. An activation of the telomerase protein in somatic cells leads to their immortalization, allowing other mutations in the affected cells to accumulate over time. These mutations then have a much higher probability to manifest and in effect cause such cells to transform.

1.1.2. Cancer Progression Models

Following the transformation of a cell to gain tumorigenic potential, for example by two or more mutation events, a clonal outgrowth may occur, if the cell has a survival advantage over those in its neighboring tissue environment.
In 1993, Bert Vogelstein and Kenneth Kinzler proposed a model that also explained the occurrence of sporadic tumors, in which they argued that for a cell or small group of cells to become a tumor, many subsequent steps are

necessary.[5] This multistep process involves several pathways and interactions, both within the tumor and its surrounding stroma, including inflammation, invasion, metastasis and vascularization. Vogelstein and Kinzler worked on colon carcinoma, which develop in well-defined morphological stages, a fact that could not be explained with the models existing at the time. They demonstrated that certain subsequent mutations, which happen rather in preferential than in a fixed order, could be associated with the disruption or over-activation of certain pathways and consequently lead from benign to pre-cancerous lesions, then to malignant carcinoma and finally to invasive carcinoma.

In recent research, another aspect of tumorigenesis has come into focus, namely the emergence of cancer stem cells.[6-9] There are two major questions in this respect to be answered: (i) Do tumors (and metastases) develop from a single or few progenitor cell(s), analogous to tissues deriving from one or few stem cells? (ii) Do tumors develop from mutations that had already occured in natural stem or progenitor cells? Of course, many more questions are connected to this concept, e.g. whether there is an asymmetric division of the tumor stem cell and a progression of its progenitor cells. However, the existence of cancer stem cells or tumor initiating cells, as they are sometimes more carefully referred to, seems to provide a valuable idea for understanding the progressive behavior of tumors. Nonetheless, some refinements to the very simple idea have to be taken into consideration as well, like the influence of tumor-stroma interaction, cross-talking processes involved in tumor invasion and vascularisazion, and the existence of so-called "dormant" cells. The latter appear for example in the bone marrow of breast cancer patients, but clearly show properties they inherited from the primary breast tumor.[10-13] The cancer stem cell idea seems also very valuable in the explanation of tumor relapse and the formation of distant metastases.

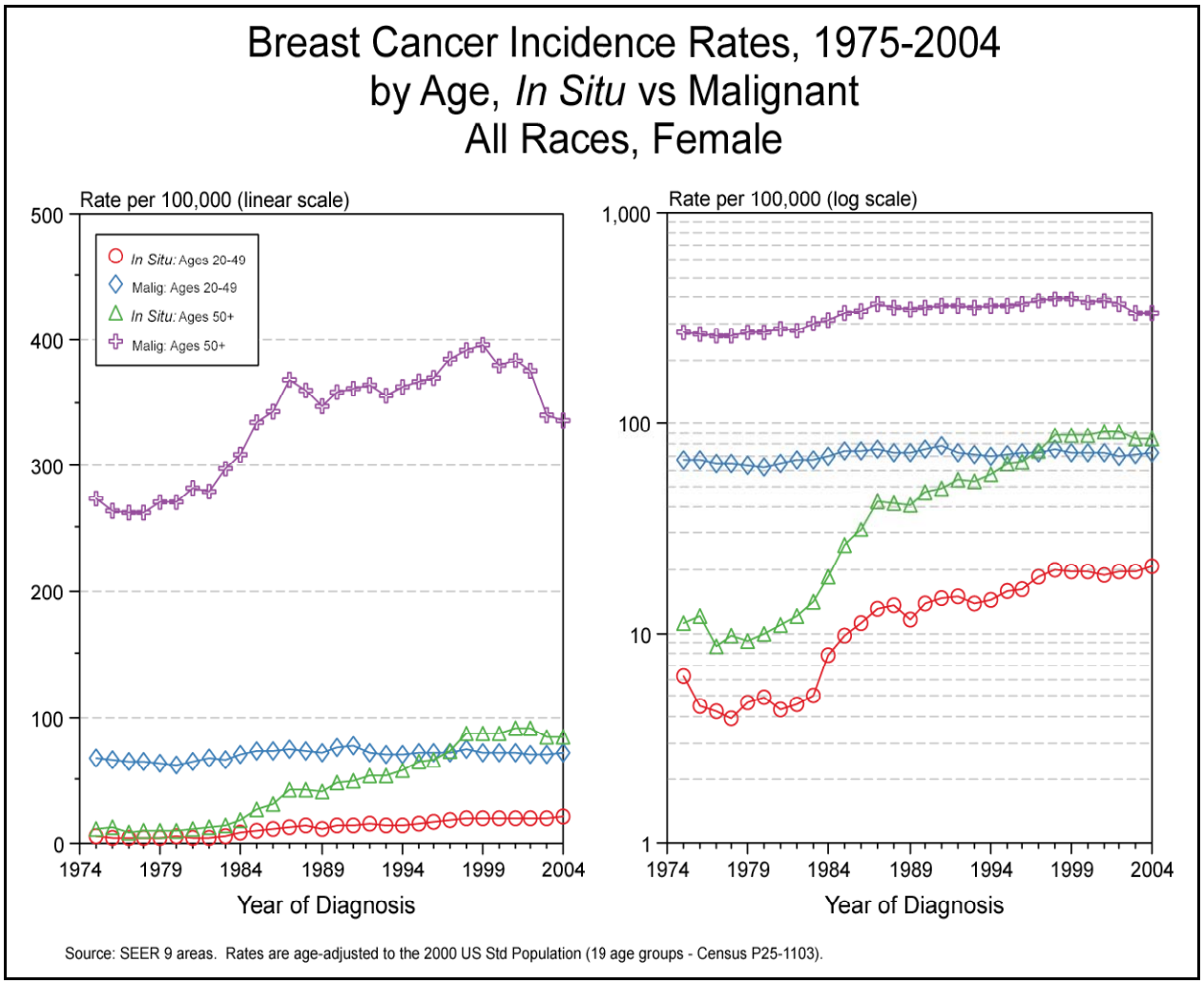
1.2. Breast Cancer

Breast cancer is the most common malignant tumor in women, both world-wide and in the high-income countries, as defined by the World Bank. It leads to more than 500,000 deaths per year in the world, and belongs to the top ten mortal diseases in the high-income countries, with a mortality rate of 1.9% for 2002.[14] In order to put these absolute numbers into a more substantial measure, the lifetime risk of a woman living in the USA to develop cancer is estimated to be between 33% and 43%, while her lifetime risk to develop breast cancer dropped from one in eight to one in 13 individuals in the last five years.[1,15]

According to the American Society of Cancer, in the USA there were an estimated 270,000 new cases of breast cancer in the year 2005, and 40,410 deaths caused by the disease in the same time period.[15] Even though breast cancers display a relatively high survival rate compared to cancers of e.g. lung, stomach or colorectum, the vast numbers of cases and high incidence rates of approximately 128 invasive breast cancers per 100,000 US women plus approximately 30 non-invasive cases per 100,000 for the years 2000-2004, make breast cancer a clinically very important and highly investigated disease.

Besides the large number of cases, breast carcinoma is also among the most heterogeneous types of cancer: firstly, in terms of the clinical course and classification; secondly, in terms of the cellular and genetic background of the actual tumor mass. A successful treatment of patients with primary breast carcinoma is therefore highly dependent on an in-depth characterization of each individual case. This comprises not only acquiring standard clinical data like age, menopausal status or histopathological staging of the cancer. A more profound examination, e.g. concerning the local spread, the tissue origin (ductal, lobular, and others), the estrogen and progesterone hormone receptor status, as well as a detailed histochemical characterization of expressed proteins like HER2/NEU, P53, BCL-2 or the proliferation marker KI67, is today's clinical standard.[16]

Figure 1



US Incidence rates of Primary Breast Cancer. Depicted are females of all ethnities, separated for age and malignancy. From the National Cancer Institute (NIH), 2007.[15]

1.2.1. Breast Cancer Types

As the incidence rates demonstrate, the large majority of breast cancer cases are comprised of malignant or invasive forms of breast cancer. However, while the incidence rate of these has not changed significantly over the past three decades, the incidence rate of the non-invasive *in situ* lesions has increased considerably from the early 1980s (up to five per 100,000) to the late 1990s (more than 11 per 100,000; Fig. 1). This is due to the fact that the introduction of mammography screening, at least in high-income countries like the USA represented here, has lead to a great improvement of the early diagnosis.

Almost all invasive breast tumors are adenocarcinoma (96.9%), with the only other histology worth mentioning being the sarcoma (0.3%), the rest are of mixed histologies (2.7%). Among the adenocarcinoma, the largest subgroup is

comprised of the invasive or infiltrating ductal carcinoma with 67.3% of all breast carcinoma, followed by the infiltrating ductal and lobular carcinoma (12.7%) and the infiltrating lobular carcinoma (8.0%). Other adenocarcinoma subtypes, like mucinous, tubular, papillary, medullary or those not otherwise specified (NOS) arise only to very low percentages (2.6%, 1.6%, 0.4%, 0.7%, and 1.1%, respectively).[15]

Following the consistent screening for breast cancer since the 1980s, the percentage of non-invasive lesions has increased from 3% to currently 20 - 35%.[17,18] The largest proportion of non-invasive breast cancer cases in high income countries is comprised of the ductal carcinoma *in situ* (DCIS) with approximately 85%, followed by the lobular carcinoma *in situ* (LCIS) with 12% (numbers for USA, averaged for 1998-2002).

1.2.2. Hereditary Mammary Carcinoma

Albeit the immense number of cases, only a small proportion of patients presenting with mammary carcinoma could be associated with an inherited susceptibility to develop breast cancer. The hereditary breast tumors differ from the sporadic cases mostly by their incidence at an earlier age (mostly premenopausal), higher prevalence of bilateral manifestations and, of course, the significant number of associated tumors within families.[19] Genetic factors that have been directly associated with breast cancer comprise for approximately 5% of all patients, and the risk to develop breast cancer is significantly larger in families with a mammary carcinoma history.[20] On the other hand, in hereditary breast carcinoma carriers, general risk factors like late pregnancy, the number of pregnancies or the menopausal state do not alter the risk of developing the tumor significantly.[21]

In the middle of the 1990s two major susceptibility genes, *BRCA1* (Chromosome 17q21) and *BRCA2* (13q12), were discovered to be directly associated with the development of the disease.[22] These harbor autosomal dominant mutations, and have therefore found their way into clinical patient management in cases with a family history.[23]

However, since there are familial patterns that cannot be associated with *BRCA1/2* genes, the importance of other genetic factors in this context has been under constant investigation. Yet, whether these contribute only to small

subgroups of patients each, or if many genes have to be considered to form a polygenic model, is still not known.[24-26] Proposed genes to contribute to the predisposition to develop breast cancer are *TP53, PTEN, LKB1, ATM, PALB2* and *CHEK2*, but less than 1% of cases have been reported with a positive association.[27,28]

More than 60% of breast cancer patients with a *BRCA1/2* mutation develop the tumor before 50 years of age. These patients have a very high incidence of a tubular carcinoma, but their histopathology does not differ significantly from those of sporadic cases. The 5-year survival rate is also similar to that of sporadic cases, so currently, the therapeutic options remain the same as well. The poly(ADP-ribosyl)-transferases (PARP) 1 and 2, which are thought to be potential modulators of DNA-repair-mediated resistance to cytotoxic therapy, are targeted by novel PARP inhibitors, which are now investigated in clinical trials as therapeutic option for *BRCA*-positive cases of cancers.[29,30]

1.2.3. Clinical Treatment of Breast Cancer

Following the diagnosis of an invasive mammary carcinoma, the standard therapeutic approach is surgical removal of the tumor, either through local excision (breast-conserving) or by removal of the entire breast (mastectomy).
Of course, breast conserving strategies are favored; however, there are cases for which the mastectomy indisputably is the only option, namely those of an inflammatory carcinoma, multicentric carcinoma or an intraductal carcinoma *in situ* with a particular classification (*Van Nuys* score 7-9).[31]

Systemic therapies, consisting of either chemotherapy, endocrine (hormonal) treatment, or a combination of both, have been developed and new protocols are constantly under investigation in clinical studies.[32] In the adjuvant setting, the systemic treatment is given after surgery, to prevent relapse of the breast tumor. In the primary systemic (neo-adjuvant) setting, the treatment precedes surgical removal of the tumor, with the additional advantage of performing a systemic treatment and monitoring of the therapeutic effect on the tumor.

In order to investigate the long-term effect of adjuvant chemotherapy, a meta-analysis of several studies was conducted.[33,34] In summary, 33% of the patients investigated showed a relapse and 36% of the patients deceased (of which 5% not due to the treatment). In respect to no chemotherapy, the relative risk of death decreased by 14.9% with poly-chemotherapy, while the occurrence of relapse was reduced by 23.7% relatively. The number of deaths not directly associated with the breast cancer was not significantly different in the poly-chemotherapy treated patients.
The analysis also revealed that therapy protocols containing anthracyclines (e.g. doxorubicine, epirubicine) have a significant survival advantage for the patients in comparison with protocols of the CMF combination scheme (cyclophosphamide, methotrexate, and 5-fluorouracil), but the long-term toxicity has not been investigated well enough for a final conclusion.

Endocrine therapies are relevant for patients with a hormone receptor status of at least 10% of tumor cells being positive for the estrogen or progesterone receptors (ER, PR).[35,36] These patients are treated effectively with tamoxifen doses starting at 20 mg/day, and it could be shown that a 5-year treatment has a significant advantage for the patients *versus* no, only one or two years of treatment.
ER-positive patients treated with tamoxifen for 5 years showed a proportional recurrence reduction after 10 years of follow-up of 47%, and the relative risk of death was reduced by 26%. The proportional mortality reductions were similar for women with node-positive and node-negative breast cancer, but the absolute mortality reductions were greater in node-positive women: In the trials of about 5 years of adjuvant tamoxifen, the absolute improvements in 10-year survival were 10.9% for node-positive and 5.6% for node-negative patients.

The primary systemic (neo-adjuvant) treatment has become the standard therapy for inoperable or inflammatory mammary carcinoma.[37] Other than that, patients who are candidates for mastectomy but wish to have a breast-conserving therapy and patients participating in clinical studies are treated currently with primary systemic therapy protocols.[38] A major advantage of this method, is the possibility to monitor the effect of treatment on the tumor and

thus the sensitivity of the tumor to the applied drugs before its surgical removal.[39] It was also shown that for the use in primary systemic therapy, the third generation aromatase inhibitors, e.g. letrozole and anastrozole or exemestane show an improvement compared to tamoxifen[40-42], whereas the results for raloxifen are not conclusive.[43,44]
The NSABP-B-27 study shows an improvement of the response rate by sequentially adding 4 x Doc (docetaxel) to the standard neoadjuvant therapy of 4 x AC (doxorubicine, cyclophosphamide).[45,46]
The effectiveness in terms of disease-free and overall survival of the neoadjuvant therapies was shown to be the same as in the adjuvant setting, but there is an improvement in the number of breast conserving tumor surgeries.[39,47]

New developments in systemic therapy of breast cancer include mostly the use of trastuzumab (Herceptin) in addition to or as substitution of chemotherapy, since it has been approved both in combination to chemotherapy or as a mono-therapy.[48] The mode of action of this monoclonal antibody against the HER2 protein is not only given by blocking of the HER2 signaling pathway, but also through activation of cytotoxic lymphocytes and the inhibition of angiogenesis.[49] However, side effects to the cardiac system have been reported; therefore, the therapy is restricted to clinical studies and not in use as a primary therapy option.[50,51] The prerequisite is a standardized characterization of HER2 overexpression in the patients (HERCEP test). Besides, trastuzumab is becoming increasingly used in palliative therapy.
To overcome the problem of resistance and improve the tolerance to the trastuzumab treatment, current research in the field includes different kinds of combinations with other antibodies (e.g. against EGFR and VEGF proteins), as well as the development and testing of pertuzumab, an improved anti-HER2 antibody directed against the dimerization domain of the protein, that could be used in addition in case of resistance or as successor of the trastuzumab anti-HER2 antibody.[52-55]

Another emerging therapy is the adjuvant use of bisphosphonates. It has been shown that these decrease the risk of bone marrow metastasis.[56-59] However,

their therapeutic use is controversial, since they have also been reported to increase the rate of visceral metastases.[60] Long-term studies are currently ongoing to prove their therapeutic applicability.

The therapy of non-invasive mammary carcinoma has gained importance by the increasing early detection of non-invasive lesions.[61] However, the two major histologies, DCIS and LCIS, show little similarities, especially in respect to their tumorigenic potential. This results in two separate strategies for the therapy of these patients.

The ductal carcinoma *in situ* develops from cells within the ductal system, which show at this stage no infiltration of surrounding stroma tissue. The histopathology of DCIS is very heterogeneous, as well as the clinical course of the patients and their prognosis. Without any therapy, approximately 30% of patients develop an invasive breast carcinoma within 3 - 10 years. Patients undergoing a mastectomy have a 98% probability to be completely cured, while breast conserving surgery and excision of the lesion lead to a relapse rate of 50%.[62] The risk of relapse can be reduced by approximately 10% as a result of the application of radiotherapy after excision.
A breakthrough in reducing the rate of mastectomies in the therapy of DCIS was the development of the Van Nuys Prognostic Index by Silverstein *et al.* in 1996.[63,64] Depending on their risk group, patients can be cured by a more extensive excision of the lesion alone, additional radio therapy (NSABP-B-17 study) and additional tamoxifen treatment (NSABP-B-24), allowing to limit the need to perform a mastectomy to the cases with indisputably no other option.[65,66]

The lobular carcinoma *in situ* differs in its biology from the DCIS, as the lesion is formed by proliferation of relatively uniform cells in the lobuli and often in the terminal ducts. The LCIS is very difficult to detect early and often an incidental diagnostic finding, since there is no perceptibility of small tumors by palpation or mammography screening due to the lack of micro-calcifications.
The LCIS is relatively uncommon, its incidence amounts to 1 - 2% of all breast tumors. It leads to a mammary carcinoma in about 35% of the cases identified

even with a follow-up of 35 years.[67] Therefore, the current clinical management of patients with an LCIS is a regular examination of the lesion's spread by mammary sonography.

1.2.4. Diagnosis and Treatment Options

The attempt to achieve an individualized approach to cope with the heterogeneity of the clinical course and biology of breast cancer cases requires an exact and differentiated diagnosis of each patient, including e.g. the local spread of the lesion or carcinoma and the estimation of lymphatic metastases.[68]

Imaging techniques to facilitate diagnostics not only include mammography and sonography, but also newer and more detailed methods. Examples for these are Magnetic Resonance (MR) mammography and Sentinel Lymph Node Biopsy (SLNB).

MR mammography is used as an additional method for refinement or validation of conventional mammography and sonography findings. Indications for its use are in-breast relapse in previously surgically treated cases, axillary lymph nodes containing metastases, evaluation of response to primary systemic chemotherapy, screening of high risk populations (hereditary risk patients, *BRCA1/2*) or patients with silicone implants. A great advantage of the MR mammography is its high sensitivity, and the resolution of blood vessels; its disadvantage, however, is the high number of false positive findings in cases of DCIS.[69]

Sentinel Lymph Node Biopsy is used selectively as a minimally invasive staging of the nodal status. Its advantage is the additional opportunity to check the lymph nodes by immuno-histochemical examination. A large disadvantage of its use is the overestimation of very small tumor lesions or cell populations, such as micro-metastases, in respect to the therapeutic course of action for these patients. Therefore, the currently favored proceeding in such cases is to dissect the axillary lymph nodes.[70]

Beyond determination of the actual state of the carcinoma or lesion, the individual therapeutic plan of action is highly dependent on the estimation of the progression of the disease and the clinical course of the patient. Therefore,

the goal is to make a profound prognosis for each patient and give a prediction of each therapeutic action to be implemented.[71]

Generally, patients can be classified as those with a very good prognosis, showing little risk of a relapse; these can be treated locally, and do not need any chemotherapy. On the other end of the spectrum are patients with a poor prognosis, who definitely need a systemic treatment, including chemotherapy or more aggressive therapy. To assess each patients options and the optimal course of action, it is therefore most important to predict the therapy response *versus* resistance or relapse.

In order to estimate the prognostic value of certain parameters, Hayes *et al.* published in 1996 a list of criteria to be fulfilled.[72] It consists of (i) the understanding of the biological model, (ii) the quick and reliable estimation, including quality assurance, of the test, (iii) a prospective planning of the statistical analysis, e.g. the establishment of threshold values, (iv) the independent validation of the test and finally, (v) the clinical relevance for the decision of the therapy choice. Meeting all these criteria, there are currently the following clinically relevant prognostic factors for breast cancer.[73,74]

(a) Lymph node status, especially in axillary lymph nodes, displays the highest prognostic value: patients with a negative lymph node status can be cured by local treatment with a success rate of 70%.
(b) Tumor size: patients with tumors smaller than 1 cm have a very good prognosis.[75-77]
(c) Histological type of cancer: tubular, mucinous and medullary carcinoma show very good prognosis.[78]
(d) Grading: very well differentiated (WHO grade: G1) tumors have a significantly better prognosis than undifferentiated tumors (G3);[79] however, 70 - 80% of carcinoma are intermediately differentiated (G2).
(e) Hormone receptor status: 75% of patients are ER and/or PR positive and have a significantly better prognosis.[76,77] However, the hormone receptor status is more important as a predicitve factor for hormone treatment.
(f) Age: very young patients (<35 years) show a very bad prognosis, and have extremely aggressive tumors.[80] In contrast, menopause is more of a predictive factor for hormone treatment than a prognostic factor.

Newly developed prognostic factors, which are currently under validation, are the urokinase type plasminogen activator (uPA) and its inhibitor PAI-1.[81] They have been reported to play a key role in invasion and metastasis; however, classification is difficult because these factors show a heterogeneous expression in both tumor and stroma cells. Patients expressing low levels of both uPA and PAI-1 have a good prognosis, and patients who additionally have a negative nodal status do not need chemotherapy.[82] uPA and PAI-1 could therefore prove to be important prognostic factors for patients with intermediate grading (G2).

Another factor of high prognostic value is the growth factor receptor HER2.[76,77] However, the classification is still not uniform enough to make a reliable prognosis due to a lack of standardization. *ERBB2/HER2/NEU* gene amplification has shown to be of higher prognostic value than immuno-histochemical detection of HER2 protein.[83-85] Patients with a high expression or gene amplification have a bad prognosis.

Proliferation markers, like the mitotic index and the expression of marker proteins (KI67, MIB1, PCNA) as measured by IHC, have currently no prognostic value useful for the clinical routine. Nevertheless, they are continuously measured for later analyses. Other prognostic characteristics currently under investigation are invasion (e.g. laminin receptors), angiogenesis (VEGF), oncogenes (*TP53* or *NM23*), and apoptotic markers (BCL-2).

Predictive factors relevant for the clinical use are currently not sufficiently available for chemotherapeutic protocols. The steroid hormone receptors can predict the response to anti-hormonal therapy like tamoxifen: ER negativity is significantly correlated with no response.[34] *HER2* gene amplification or overexpression has been shown to predict the response to trastuzumab (Herceptin), either as systemic therapy or in palliative use during chemotherapy.[48] Again, FISH and RQ-PCR data correlate better than protein overexpression measured by IHC;[86,87] and additionally, it was shown that the serum level could also be correlated with response.[88] *HER2* positive patients have also been reported to respond poorly to CMF chemotherapy (cyclophosphamide, methotrexate, and 5-fluorouracil), but well to

chemotherapies containing anthracyclines (doxorubicine, epirubicine, and others) and taxanes (docetaxel, paclitaxel).[89-91] An overview of predictive factors currently in clinical use is given in Table 1.

Table 1 **Predictive Factors for Mammary Carcinoma with Clinical Relevance** [68]

Factor	**Class**	**Predictive for response to**
steroid hormone receptors	positive	endocrine therapy
menopausal status	premenopausal	ovary ablation
HER2 status	positive	Herceptin, palliative use
	positive	chemotherapy (Anthracyclines/Taxanes)§
	negative	chemotherapy (CMF)§
	negative	endocrine therapy§

§ Currently not recommended for selection of therapy, only retrospective data available

The promising results achieved by prognostic and predictive factors currently used for some therapies in the clinic show the importance of developing more and improving the existing ones. This is especially important in the case of chemotherapeutic treatment, both in the adjuvant and the primary systemic setting. From the presently available factors, a benefit could be achieved by combining some of these into multifactorial models, but this approach requires advanced mathematic models, which have to be developed and validated. To integrate many factors and limit the laboratory effort at the same time, there is a strong need to miniaturize and to integrate multiple measurements into a smaller number of experiments. Efforts in this direction have been and still are currently undertaken in many laboratories and clinical institutions, both in the proteome analysis[92] as well as in great numbers on the level of DNA and RNA analysis, e.g. by molecular profiling.[93]

1.2.5. Molecular Profiling in Prognosis and Therapy Response Prediction

The characterization of tumor patients as currently feasible in the clinical routine has not led to a reliable means of classification into tumor subtypes according to the patients prognosis after chemotherapy and does not allow predicting their response to chemotherapeutic treatment.[16,94,95] This is especially unfortunate, since in breast cancer the chemotherapy has a

substantial therapeutic impact on the clinical outcome of the patients. To successfully cure the patients with this method, there are many different therapeutic agents, combinations of these and protocols including delivery schedules, that are currently under investigation.[96,97] None of these has a proven general applicability to treat patients with a substantially better response compared to other protocols, yet their success may differ largely when applied to individual patients.

The measure taken into consideration for the success of any individual chemotherapy protocol is the rate of pathological complete remission (pCR), defined as the disappearance of all viable tumor cells in the tissue. Since the pCR rate is highly correlated with the disease-free and overall survival rates of treated patients, it can be used as an early and direct surrogate marker for treatment success.[98-101]

Current chemotherapy treatment protocols, as for example the neoadjuvant therapy administered in the study investigated in this dissertation, yield pCR rates of approximately 25 - 30%.[45,47,102-104] These rates could be substantially improved, if there was a more reliable way to predict the success of the treatment for each individual patient before application. Considering the multitude of treatment options just in the case of chemotherapy alone, this would hopefully result in an improvement of the overall treatment success.

The only way to substantially improve the pCR rates in the currently known chemotherapies seems to be an extensive and detailed multifactorial assessment of each patient's genetic or biochemical record, or likewise of the tumor to be treated, as a prerequisite to a tailored application of optimal therapy options.

To perform this multifactorial assessment, a number of methods are applicable. On the molecular biology level, array-based comparative genome hybridization (aCGH), expression profiling by means of microarrays and real-time quantitative polymerase chain reaction (RQ-PCR) provide suitable information. On the protein level, currently only advanced classical methods like two-dimensional electrophoresis followed by transfer to Western blots and immuno-histochemical staining, or IHC staining of tissue sections or tissue microarrays have the necessary precision and laboratory applicability. More

recent methods, like antibody or protein microarrays as well as small volume applications have not yet a proven reliability and lack necessary standardization.

All these procedures, however, are currently not feasible to be performed for each single patient on a daily clinical routine basis. Therefore, the genes or proteins used for determination of prognosis and prediction need to be identified first, and then narrowed down to a suitable number, in order to be effective both in regards to time and costs.

Recent developments to achieve prognosis of disease-free and overall survival after breast cancer have come from studies using cDNA microarrays, oligonucleotide microarrays or RQ-PCR (Table 2).[95,105] In order to find good prognostic markers, an unsupervised clustering of gene expression measurements in tumor samples from patients with breast cancer revealed different tumor subtypes than the clinically established ones. These groups show distinct gene expression patterns and different prognoses, as estimated by survival analysis in prospective studies.[93,106-109] The classification into groups distinguished by their molecular patterns, as represented in Figure 2, not only allows for a better subclassification, leading to more accurate prognosis of patients with primary breast cancer, but also includes patients who developed metastases.

Table 2 **Studies Investigating Clinical Potential of Multi-Gene Factors**

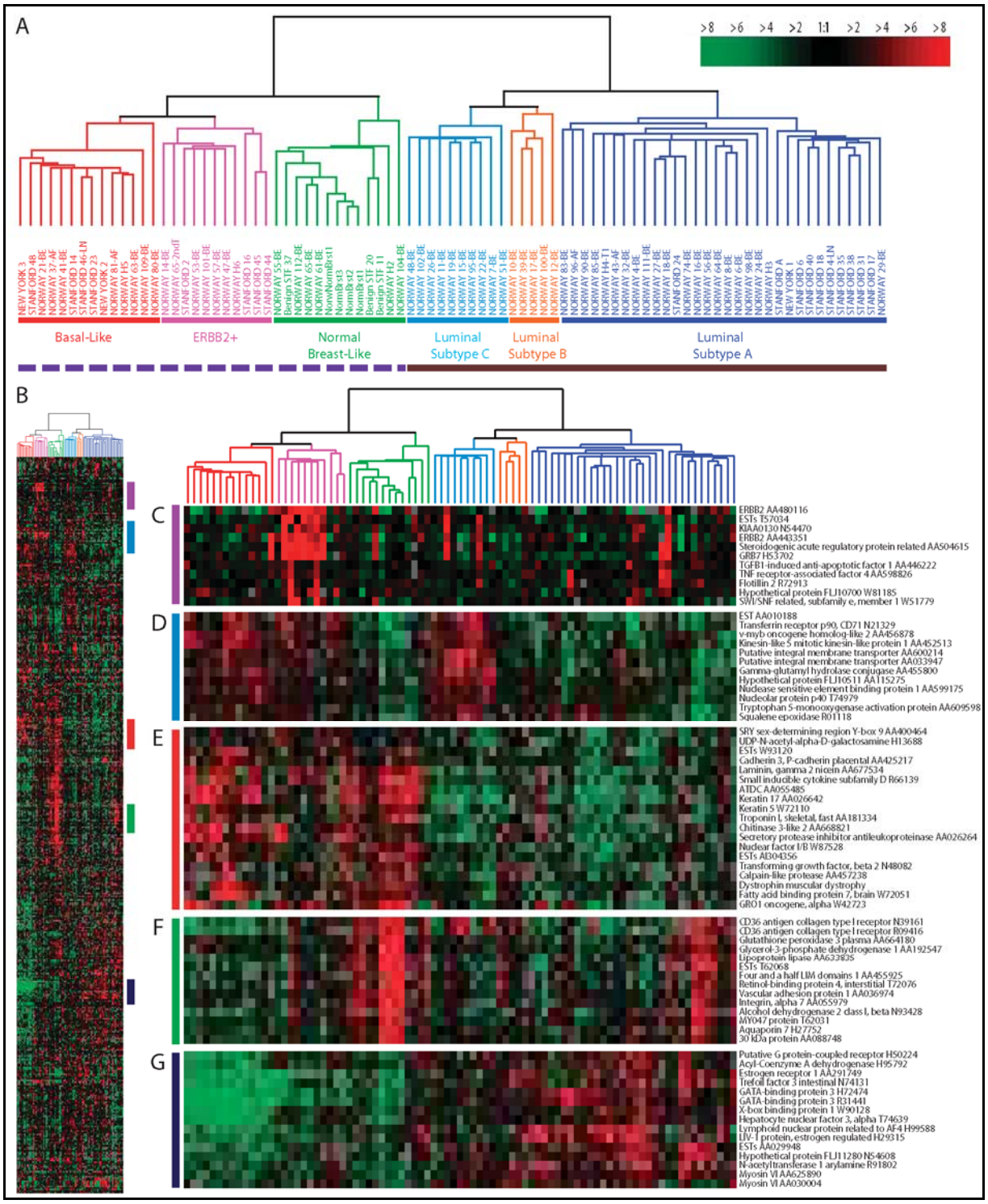
Authors	Tumors (n)	Primary Endpoint	Molecular Tool	Genes (n)	Year published
Sørlie T *et al.*	78	Classification to outcome	custom cDNA array (8,102)	427	09/2001,[106] PNAS
van 't Veer LJ & Dai H & van de Vijver M *et al.*	98	Prediction of distant metastases	Affymetrix, Hu25K	231	01/2002,[93] Nature
van de Vijver MJ *et al.*	295	Prediction of distant metastases	Affymetrix, Hu25K	70	12/2002,[109] N Engl J Med
Chang JC et al.	24	Prediction of reponse* to chemotherapy (A)	Affymetrix, HgU95-Av2 (12k)	92	08/2003,[110] Lancet
Ayers M *et al.*	42	Prediction of pCR in chemotherapy (T/FAC)	custom cDNA array (31k)	74	06/2004,[111] J Clin Oncol
Paik S *et al.*	668	Prediction of distant metastases	RQ-PCR	21	12/2004,[112] N Engl J Med
Wang Y *et al.*	286	Classification to outcome	Affymetrix, Hu133a (22k)	76	02/2005,[113] Lancet
Hannemann J *et al.*	48	Prediction of "near" pCR in chemotherapy (AD;AC)	custom cDNA (18k)	--	05/2005,[114] J Clin Oncol

* response defined as ≥75% regression of tumor

While these molecular classification models have been developed independently from the clinical parameters, they do reflect some of the clinical classifications. This includes, for example, the estrogen receptor positive and negative groups or the HER2 positive patients. However, the molecular patterns allow the identification of subgroups within these larger classes, and integrate a finer mapping of the biological setup. Prominent examples are luminal subtypes A, B and C, which together represent the ER positive patients, but show a varying prognosis and therefore benefit from different kinds of treatment.[106] The clinical group of HER2 positive patients, on the other hand, can also be further subdivided into those that have a prognosis similar to some of the luminal subtypes and those behaving differently. This demonstrates that while classifications based on single genomic or protein factors alone cannot be used in this example, the gene expression levels incorporated for several genes are able to identify distinct groups of patients with a high prognostic value. Additional groups of patients distinguished by molecular profiling include the basal-like subtype that includes *BRCA1/2* mutation or deregulation carriers, and the ER negative "normal breast-like" subtype.

These molecular subtypes allow for a better decision regarding their different treatment options; e.g. only the luminal subtype A with a low expression of proliferative genes shows a good prognosis, therefore suggesting a successful treatment with endocrine therapy alone.

Figure 2



Hierarchical clustering of patients using microarrays. Gene expression patterns of 85 samples (78 carcinomas, three benign tumors, four normal tissues) analyzed by hierarchical clustering using the 476 cDNA intrinsic clone set. **(A)** The tumor specimens were divided into subtypes based on differences in gene expression. The cluster dendrogram showing the subtypes of tumors are colored as: luminal subtype A, dark blue; luminal subtype B, orange; luminal subtype C, medium blue; normal breast-like, green; basal-like, red; and ERBB2+, pink. Estrogen receptor positive subtypes, solid brown; Estrogen receptor negative subtypes, dashed purple. **(B)** The full cluster diagram scaled down. The colored bars on the right represent the inserts presented in C-G. **(C)** ERBB2 amplicon cluster. **(D)** Novel unknown cluster. **(E)** Basal epithelial cell–enriched cluster. **(F)** Healthy breast-like cluster. **(G)** Luminal epithelial gene cluster containing ER. Adapted from Sorlie *et al.*, 2001.[106]

In order to estimate a predictive score, studies integrating the prognostic groups, e.g. for the luminal A group, to predict the recurrence of relapse after tamoxifen treatment identified a capable predictor of 21 genes.[112,115] In the case of predicting response to chemotherapy treatment protocols, several studies have been performed.[93,110,111,113] These studies yielded gene expression signatures of less than 100 genes, and outperformed other clinical parameters in their predictive power. However, the patient sets that were included in these studies were either limited in number or pre-selected in their patient cohorts (e.g. mean age below 50 years, node-negative patients).

The very urgent need to improve the pCR rate significantly is underlined by the continuous search for improvements to the existing chemotherapy protocols.[42,96,97] However, since these are yet to prove their ubiquitous applicability with a significant percentage of pCR patients, the gene expression signatures are currently the most promising approach to reach that goal without a long delay. In December of the year 2006, a large phase III trial (planned recruitment: 6,000 patients) was initiated to assess the clinical relevance of the 70-gene prognosis signature, and how it compares with common prognostic factors for assigning adjuvant chemotherapy for patients with node-negative breast cancer ("Microarray In Node-Negative Disease May Avoid Chemotherapy", MINDACT).[116,117]

1.3. DNA Microarrays

Over the past decade, the microarray technique has taken considerable steps forward. Originally, DNA microarrays had been developed from cDNA or genomic material, incorporated into plasmids or BACs (bacterial artificial chromosomes), respectively, then stored and sustained in *E.coli* libraries, and finally further amplified as PCR products before being spotted onto coated glass slides.[118,119] Since the information, which cDNA was contained within each clone of the library and thus contained in each feature of the array had to been obtained by sequencing of the cDNA, the annotation was insufficient for most of the collections.

The publication of the human genome sequences and those of other important mammalian species around the year 2001 has made detailed genomic information publicly available, giving commercial oligonucleotide manufacturers the opportunity to bioinformatically design and create specific oligonucleotides for each gene or genomic locus.[119,120] Today, the most common forms of the genomic and expression profiling microarrays contain either probes synthesized *in situ* on the support material as 20- to 60-mers, or oligonucleotides that were synthesized *in vitro,* e.g. as 70-mers, and then deposited onto glass slides as had been done with the PCR products.[120]

Advantages to the former, the *in situ* synthesis, are mass production with tight feature reproducibility, a much smaller feature size and therefore the possibility to analyze many samples on a vast number of DNA probes at once in a comparative manner. Their disadvantage is the higher production cost of the microarrays. Furthermore, Affymetrix' 20- to 25-mer oligonucleotide GeneChips require the addition of immobilized DNA probes containing single nucleotide mismatches to quantify unspecific hybridization events.

Advantages of the *in vitro* synthesized oligonucleotides are the much simpler and already highly standardized synthesis that leads to a dramatically lower cost per probe and the inclusion of quality control for the synthesized probes before actually depositing them on the array.

Common to both methods is the high consistency of the hybridization characteristics of the probes. This results from standardization of parameters like base content, length and melting temperature of the oligonucleotides, as

well as from prevention of loop structures, cross-reactivity and repetitive sequences, by bioinformatic design of the oligonucleotides. The uniformity is the major advantage of these types of arrays over their former cDNA counterparts. Additionally, changes in the annotation and mapping of the genome can be represented quickly and cost-effectively by adding new probes to the set. Furthermore, different splice variants of mRNAs from the same gene can be represented using specific oligonucleotides for common *versus* unique exons, if necessary.

In cDNA arrays, the variance in DNA content of the single features is large, as they are subject to the amplification efficiency, consequently resulting in varying hybridization requirements of the individual spots. Secondly, within each feature, different DNA molecules have to be expected as a result of full length and partial length PCR products. Thirdly, since they derive from full length cDNA molecules, the products to be PCR amplified range from 500 to 2,500 base pairs in length. However, longer initial cDNA molecules are more probable to be amplified partially.
Additionally, for their optimal hybridization performance, longer DNA molecules require different reaction buffers and/or temperatures than shorter ones, while a higher DNA molecule content dictates a different reaction time than lower density spots, respectively. As all the features are hybridized together on a single array, they can only be incubated in a certain buffer at a certain temperature for a certain time.
This heterogeneity of hybridization optima for the molecules between and within the features therefore leads to a deviation of results, making normalization a difficult yet very important process in the analysis of the raw data. Nevertheless, a direct comparison between single features can not be made without taking this aspect of heterogeneity into account.

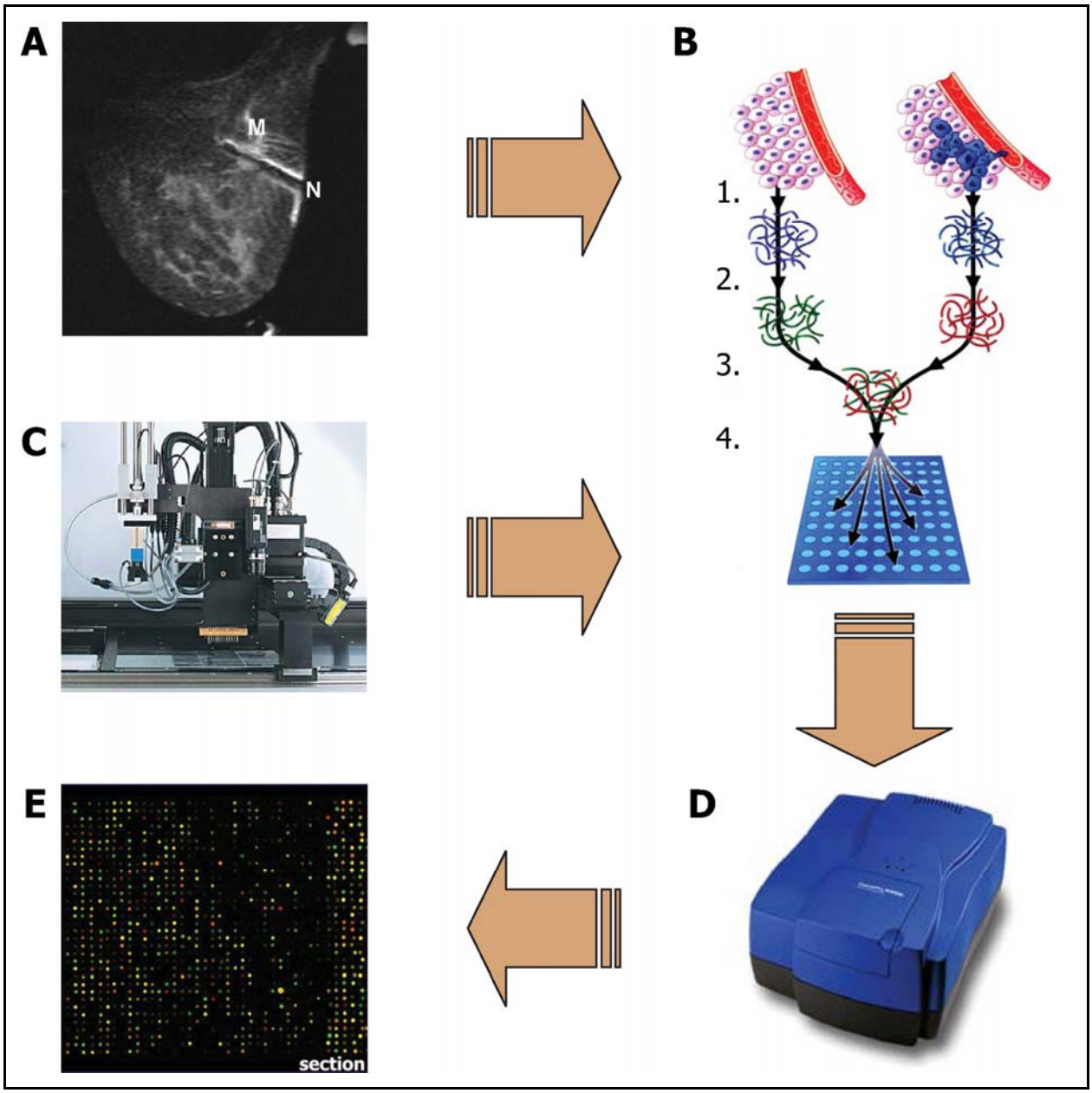
Oligonucleotide DNA microarrays have eliminated this problem almost entirely, since all molecules have a very tight distribution of hybridization properties, for example the T_m of the melting temperature usually varies only by ± 2 K. As the DNA quantity in each spot is the same for every feature, the hybridization

conditions are very homogeneous for the entire array, resulting in a much higher reproducibility of the gene expression measurements.

The advantages of the oligonucleotide-based microarrays have lead to an increase of their use and thus a much higher comparability of the results generated, e.g. experiments performed in different laboratories or even between different studies. A study comparing expression profiling experiments using different oligonucleotide microarray platforms on patient material was performed to show the consistency of the results.[121]

The workflow of obtaining tumor tissue, extracting nucleic acids from the tumor cells, amplifying the genetic material and generating labeled polynucleotides to hybridize onto microarrays is shown schematically in Figure 3.

Figure 3



Schematic workflow of gene expression profiling using microarrays. (A) Extraction of tissue from tumor mass (M) by needle biopsy (N) under sonographic surveillance. **(B)** 1. Extraction and isolation of DNA and RNA from reference and tumor tissues*, 2. amplification and labeling of the genetic material with fluorescent dyes; 3. mixing of tumor and reference samples appropriately labeled for 4. competitive hybridization on DNA microarrays. **(C)** Spotting of cDNA or oligonucleotide microarrays by deposition of DNA onto glass slides. **(D)** Scanning of microarrays to measure intensities of hybridized sample molecules using Axon Microarray Scanner Model 4000B. **(E)** Scanned microarray image showing individual DNA probes hybridized with Cy3- and Cy5-labeled sample DNA (green and red, respectively).

* Reference RNA can also be used from independent sources, e.g. cell lines.

1.4. Messenger RNA Amplification Methods

Even though the sensitivity and reproducibility of DNA microarray techniques for expression analysis have increased, there is a certain detection limit of these methods. This limit is given by the technical or biochemical variances of comparative hybridization, fluorescent labeling and detection of the molecules in respect to variations between the expression levels of different genes.

Such a limit, e.g. 1 µg of mRNA to be reversely transcribed and directly labeled with Cy-dye coupled nucleotides for hybridization to spotted cDNA or oligonucleotide microarrays, usually cannot be accomplished with small tumor samples, like biopsies or small cell cultures. In most applications, it is therefore necessary to enrich and specifically amplify the mRNA against other RNA types, since mRNA constitutes only 5 - 10% of the total RNA in cells on average.

Several different methods have been developed to achieve the necessary amount of DNA or RNA that can be successfully labeled, hybridized and detected.[122-129] These can generally be divided into two groups: Those that amplify linearly, mostly using *in vitro* transcription (IVT), and those amplifying exponentially, using polymerase chain reaction (PCR) based protocols. In both cases, the input nucleic acid is generated by reversely transcribing the mRNA into cDNA, firstly because the mRNA molecules can be selectively transcribed by making use of their poly(A) tails, secondly because they will then be transcribed into more stable and less digestion sensitive DNA molecules. There is an exception to the classification into these two groups, in that special Taq DNA polymerase based variants can also be used to amplify linearly.[127,128]

The advantage of the exponential amplification methods based on PCR is their rapid and effective usage. Since PCR is a standard method in scientific laboratories, they can be performed with widely used enzymes and materials, therefore resulting in great cost and labor efficiencies. Their disadvantage, however, lies in the exponential amplification itself: As the ratio of two different mRNA molecules in a cellular sample can exceed 1000-fold easily, the representation of the ratio would be greatly exaggerated by the exponentially amplification method. This can be explained by the probability of each single molecule to be processed by a polymerase, which is the limiting component in this reaction. Additionally, with increasing length of the mRNA molecules, the

probability to receive a full-length amplification product necessary to generate a signal decreases. Therefore, longer mRNAs are underrepresented. Another bias is introduced by the initial random selection of molecules to be amplified: as the number of DNA molecules exceeds the number of available DNA polymerase proteins, the selection of amplified sequences occurs randomly. At the initial steps of the PCR, this selection will introduce a bias, which will be exaggerated by the exponential amplification. The PCR-based methods are only suitable to significantly detect differences, if these are either occurring in highly abundant mRNA molecules or if the molecules differ only slightly in numbers or length. Otherwise, the results of the method do not represent the true situation within the cells, and can only be used as a qualitative result. As mentioned above, exceptions to this classification are the protocol variants in which the amplification occurs linearly despite usage of Taq DNA polymerase.
The advantage of linear amplification methods, mostly performed by IVT, is the preservation of cDNA molecule ratios independently of their original abundance. Their disadvantages, however, are the reintroduction of RNA molecules into the amplification procedure, which is less stable and prone to unintentional digestion, and the relatively high laborious effort.
Since the disadvantages of the exponential PCR-based methods outweigh the disadvantage of the more reliable linear methods, the standard method for amplification of mRNA and labeling onto cDNA microarrays has become the IVT-based protocol, as described by the laboratories of Eberwine and Baugh and later optimized by Kenzelmann and co-workers.[123-125]

With the introduction of single-stranded oligonucleotide microarrays into the laboratories, another disadvantage of the linear IVT-based method became obvious: Since the amplification step produces antisense-orientated RNA molecules, their labeled complementary DNA products are, of course, sense-orientated. However, these can not hybridize onto the oligonucleotide microarrays with sense-orientated DNA probes, which had been designed as such to be used with directly labeled antisense-orientated cDNA.
This incompatibility of IVT-based linear amplification and labeling with the use of sense-oriented oligonucleotide microarrays needs to be overcome in order to perform expression profiling studies with the oligonucleotide array technology.

Therefore, a novel protocol suitable for the amplification of mRNA yielding fluorescently labeled antisense nucleic acid and for the usage in expression profiling hybridization experiments with long oligonucleotide microarrays is required.

2. Aim and Procedure

The aim of this dissertation was to assess the applicability of the microarray expression profiling technology for finding a reliable predictive set of genes from small tumor biopsies of female primary breast cancer patients for the response to the tested neo-adjuvant chemotherapy comprised of gemcitabine, epirubicin and docetaxel.
Patients were considered as responders only if they had a pathological complete remission (pCR) after primary systemic chemotherapy. Patients with residual tumor cells at surgery, either resulting in pathological partial remission (pPR) or pathologically no change (pNC), were considered as non-responders.

For this purpose, the technique to generate 70-mer oligonucleotide microarrays representing transcripts of the whole human genome had to be established and optimized for the use with the given infrastructure in the laboratory. Additionally, a protocol to linearly amplify mRNA and fluorescently label the nucleic acids needed to be developed that could be used with spotted sense-orientated oligonucleotide microarrays and, at the same time, had the necessary fidelity to analyze small tumor biopsies.

After performing the genome-wide expression profiling of the tumors, an extensive bioinformatic analysis of the contained genes had to be performed to establish the gene signature predicting the classification of patients into responders and non-responders. For this purpose, the samples had to be split into two sets, one used as a training set to discover a predictive gene set, the other to validate its predictive power. Algorithms used to identify the genes were support vector machines and receiver-operator characteristic curve analysis.

In order to elucidate the biological mechanisms of response to the chemotherapy, it was of great interest to investigate the genes contained in this signature. Therefore, further pathway and immuno-histochemical analyses of some of these genes were the concluding objectives of this dissertation.

3. Material and Methods

3.1. Microarrays

DNA microarrays were generated using oligonucleotides which had been evaluated earlier.[130] Based upon these data, the Human Oligo Set 2.0 (Operon), containing 21,329 gene-specific 70-mer sequences plus controls, were obtained. After production and usage of microarrays with this set for the first group of patients, an upgrade set of 5,462 sequences was added. The new entire collection, containing 26,791 oligonucleotide probes (Human Oligo Set 2.1.1), was then used for the second group of patients, in the second patient group of the study (see Chapter 3.3).

3.1.1. Generation of Microarrays

The technique used here to deliver small spots of DNA onto coated microscope slides was split pin printing. For this method, oligonucleotides were diluted in an appropriate spotting buffer and distributed in 384-well plates. A robot equipped with steel pins, which have a fine slit, dipped these into the DNA solution and the pins were allowed take up a small but defined volume by capillary force. Subsequently, the pins were brought into contact with each of the slides to deliver a small drop on them. Afterwards, the pins were washed several times and dried. This cycle was repeated until all sequences of the entire set were successively deposited in spots, creating an array of the different DNA molecules, each with a defined position on the slide.

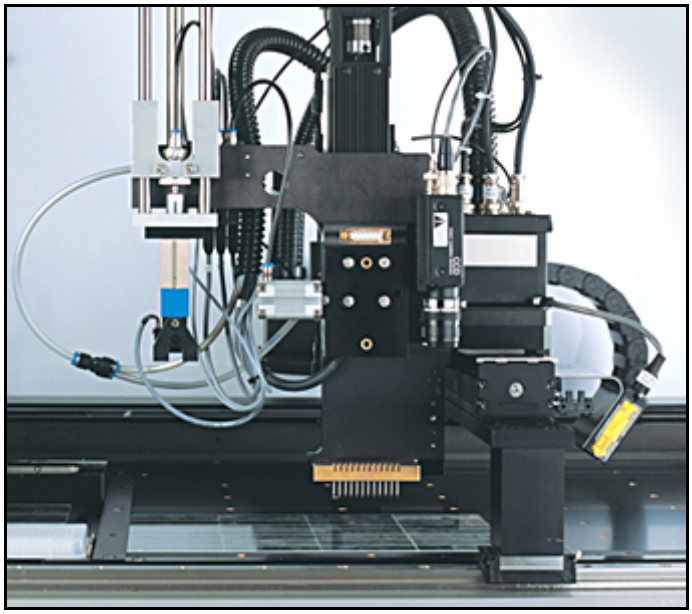
The oligonucleotides were delivered by the manufacturer in lyophilized form, 600 pmol of each DNA probe in 384-well plates. To obtain a concentration of 40 mM as recommended by the manufacturer, the sequences were dissolved in 15 µl of buffer. As seen during the evaluation of the oligonucleotides, the spotting buffer "FBNC", developed by Dr. Gunnar Wrobel, proved to be most useful for printing oligonucleotides on glass slides.[131] It contained formamide, aqueous betaine solution and nitrocellulose diluted in DMSO (Table 3). Along with its good spot *versus* background intensity ratio characteristic, it offered an important practical advantage over commonly used 3 × SSC or 3 × SSC / 1.5 M betaine spotting buffers, namely the minimized evaporation due to the components formamide and DMSO. The disadvantage of the buffer, a slightly

wider spread of the spots at delivery due to DMSO, was minimized by setting and the relative humidity of the air in the room to a maximum of 40%.

Table 3	**FBNC Spotting Buffer**, per 10 ml [130]
2.50 ml	formamide (p.A.; Merck)
0.25 ml	20 mg/ml nitrocellulose (Sigma-Aldrich) in DMSO (Merck)
2.00 ml	2.5 M betaine hydrochloride (pH 6.0; Sigma-Aldrich)
5.25 ml	H_2O (Milli-Q)

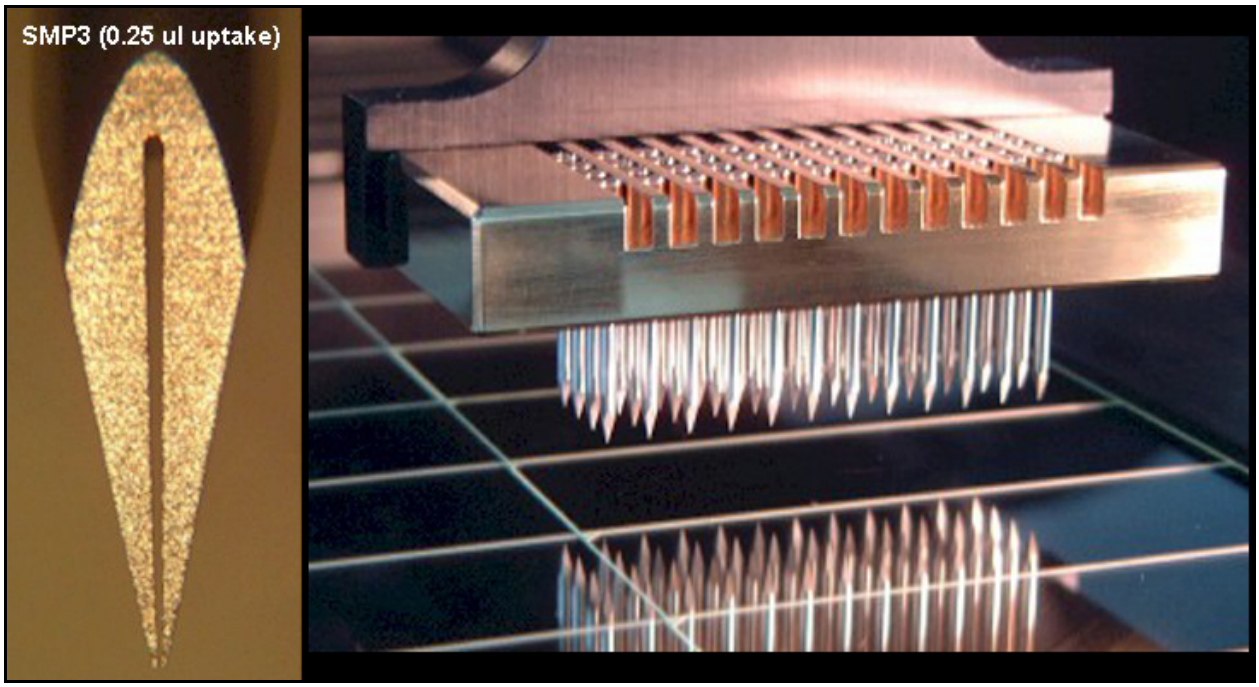
To print the DNA onto the epoxy-silane coated slides (Schott Nexterion), two spotting robots available in the laboratory were used, first the GeneMachines OmniGrid 100 (Genomic Solutions) with a capacity of 100 slides, later the VersArray ChipWriter Pro System (Bio-Rad, Figure 4), with a capacity of 108 slides. Both were equipped with a print head capable of carrying up to 48 SMP3 pins (TeleChem, Figure 5). The advantage of the VersArray System lay in its ability to process stacks of up to 4 × 13 plates, while the OmniGrid robot could only process one plate at a time, which required manually changing each plate of a set.

Figure 4



VersArray Microarray Spotting System. From Bio-Rad Laboratories, Inc.

Figure 5



Microarray SMP3 split pin needle (left) and pin head holding 48 pins (right). From Telechem Inc.

Another restriction of the OmniGrid was the incompatibility of its software to manage the required 4 × 6 pin setting, required for spotting array duplicates with the maximum distance between two repeat spots and maximum processing speed. The minimal possible distance of the spots with the FBNC buffer of 125 µm, limited the arrays to a maximum of 54 384-well plates in a 4 × 4 pin configuration with this robot. To spot arrays larger than 54 plates, e.g. the Operon Human Oligo Set 2.0 (57 plates), a 2 × 12 pin configuration, had to be used, which allowed only for array duplicates with a much lower distance of the repeat spots to each other. On the other hand, this setting had a more suitable spot-to-spot distance of 145 µm (961 spots / pin × 24 pins = 23,064 different spots) with up to 60 384-well plates. Spotting the entire Human Oligo Set 2.1.1, consisting of 72 plates including the update, was performed solely by using the VersArray system and a 4 × 6 pin configuration, with a spot-to-spot distance of 130 µm, creating 27,648 different spots in array duplicate.

For all spotting runs, SMP3 spotting pins were used (TeleChem, Figure 5, left panel), which have a take-up volume of 0.25 µl, and the robots set to a slide approach speed of 1 mm/s. On average, this generated spots with a diameter between 60 and 65 µm, so the spot-to-spot distance, e.g. for the entire Human

Oligo Set 2.1.1, was twice as large as the spots themselves. Each drop contained an estimated volume of 0.625 nl.
After spotting, arrays were post-processed by drying the slides for 60' at 60 °C in an oven and cross-linking of the DNA to the coated surface by UV-radiation (254 nm) for 2 × 2' in a Stratalinker 2400 (Stratagene), with "Auto-Crosslink" setting (maximum of 120 J/cm^2). Microarrays were sealed together with silica-gel in airtight packages for keeping them dry and stored at 4 °C.
Directly before usage, the microarrays were washed for 2' in 0.2% SDS (w/v) at room temperature, 2' in ddH_20 at room temperature, and 10" in boiling ddH_20. Right after that, the slides were immediately transferred to 50 ml-Falcon tubes and locked in to avoid evaporation of remaining water on the slides. To remove residual water, the arrays were centrifuged for 1' at 1000 rpm in a Heraeus Varifuge 3R (Kendro).

3.1.2. Hybridization and Post-Processing

Labeled and washed DNA or RNA samples (see Chapter 3.2) were diluted in UltraHyb buffer (Ambion), which had been pre-heated to 70 °C, to a final volume of 120 µl. The mix was pre-incubated for 30' (RNA) or 60' (DNA) at 60 °C while shaking at 1,200 rpm and shielded from light. Meanwhile, the microarrays were mounted in a GeneMachines HybStation (Genomic Solutions) and pre-heated for 5' at 60 °C. Finally, the samples were heated for 10' at 70 °C in the same conditions as before and spun down briefly to collect condensed solvent. The samples were then immediately injected into the HybStation chambers onto the slides.
Hybridization was performed for 16 h at 42 °C with agitation of the hybridization mix by the HybStation. Afterwards, each slide was washed with Medium Stringency Buffer (40" flow, 5' hold), High Stringency Buffer (40" flow, 3' hold) and Postwash Buffer (40" flow, 2' hold) at 36 °C on the HybStation (see Table 4 for composition of the buffers). Each microarray was then dismounted, immediately dipped into Postwash / Tween Buffer at room temperature and transferred to 50 ml-Falcon tubes which were immediately locked to avoid evaporation. Slides were centrifuged for 4' at room temperature in the Varifuge. Centrifugation was started at 500 rpm and the speed was increased every 30" by 500 rpm, resulting in a maximum centrifugation speed of 2,000 rpm, which

was kept for the remainder of the time (approximately 90" to 120"). The dried slides were then protected from light until scanning on the same day.

Table 4 **Microarray Hybridization Wash Buffers**

Medium Stringency	High Stringency	Postwash	Postwash / Tween	Component
0.5 x	0.05 x	0.05 x	0.05 x	SSC (150 mM NaCl, 15 mM Na_3-citrate, pH 7.0)
0.1%	0.1%	--	--	SDS (w/v)
--	--	--	0.05%	Tween-20 (v/v)

3.1.3. Scanning and Data Pre-Processing

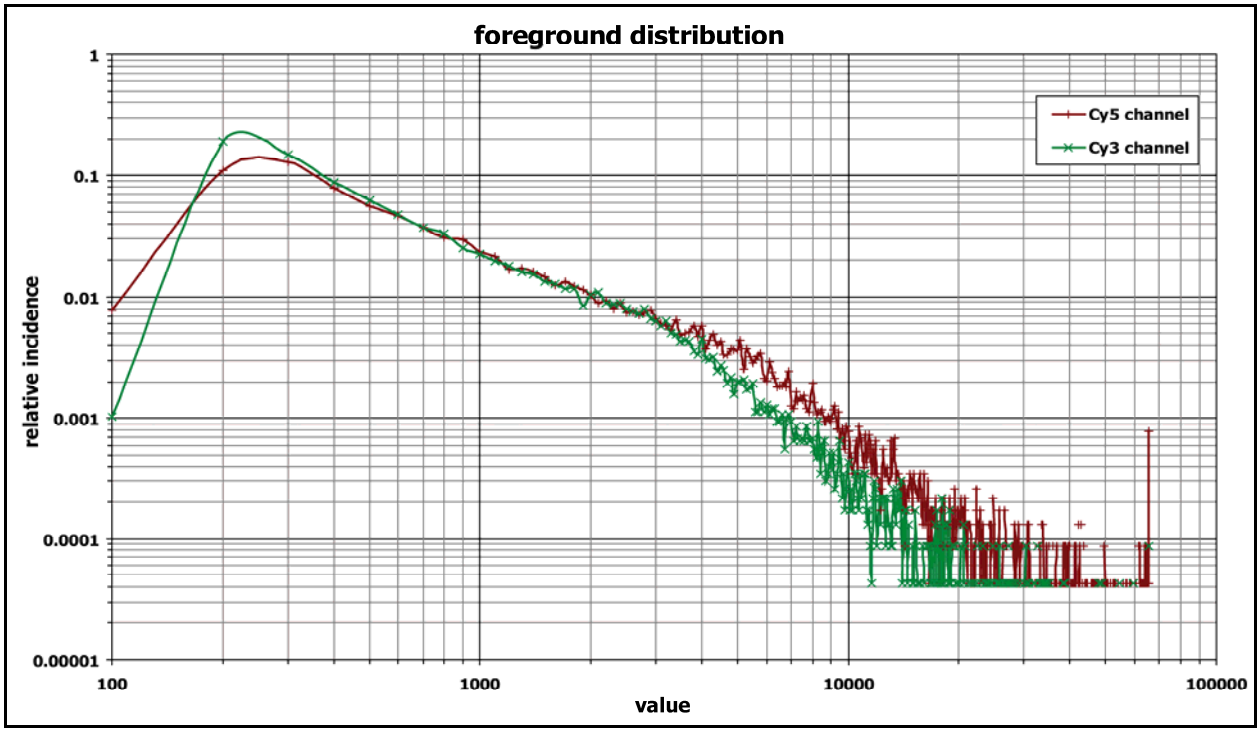
An Axon Microarray Scanner, Model 4000B (Molecular Devices), was used to document hybridization of the fluorescently labeled DNA or RNA samples to the gene-specific sequences immobilized on the array. For each channel, the fluorescent molecules were excited at their characteristic optimal wavelength with dedicated lasers, the locally emitted photons were specifically filtered by their wavelength and amplified via photo-multiplier tubes (PMTs) to be measured digitally on a 16-bit scale (maximum intensity = 65,536).

Scanning was performed at a resolution of 5 µm, and the voltage of the PMTs was adjusted so that the overall rate of pixels reaching saturation did not exceed 0.1%. At the same time, it was assured that the distribution of intensities for both channels was as similar to each other as possible, as seen in the histogram (Fig. 6). This was necessary to compensate for different incorporation rates of the labeled nucleotides as well as emission and bleaching specifics of the used fluorescent dyes Cy3 and Cy5.

The primary data generated from the measurement consisted of pixel intensity values, which had to be matched to the individual spots of DNA in the array. Therefore, a corresponding grid needed to be compiled, based on the table of the DNA sequences in the plates, using the software of the spotting robot. This grid was then overlaid in the scanner software GenePix Pro 5.0 (Molecular Devices) with the image representing each scan. This enabled averaging values of all pixels representing the individual DNA spots and the labeled samples

hybridized to them. For each spot, pixels from the surrounding area were taken as a background value. Spots which could not be considered as representative for a gene, e.g. neighboring spots that had accidentally joined or those significantly too small or large were marked manually as outliers. Spots near or below the background intensity, which could not be faithfully taken for a measurement, where marked automatically by the GenePix Pro software. The entire dataset was then exported and saved for each scanned slide. This raw data table contains values for each spot consisting of its position, the number of pixels, the fore- and background intensity values for each channel averaged as arithmetic mean and median, flags representing validity of the spot, and other data.

Figure 6



Histogram of pixel intensities. Scanned images were analyzed for pixel intensities in both dye channels in an overlay. Relative incidence gives ratio to sum of intensities for all pixels in the respective channel of the image. Value denotes pixel intensity measured in arbitrary units (max. intensity, 65,536).

3.2. Messenger RNA Amplification and Labeling Protocol

To successfully use microarrays generated from sense-orientated oligonucleotides, the sample RNA needed to be converted into labeled RNA or DNA with antisense-orientation. For this purpose, different protocols were developed and tested, partly in cooperation with Dr. Jörg Schlingemann.

3.2.1. Sample and Reference RNA

RNA used for the development of suitable amplification and labeling protocols for hybridization onto oligonucleotide arrays was generated from cell lines grown and harvested in the laboratory. Since the comparison of protocols included analyses concerning reproducibility and linearity of the amplification, two cell lines with well defined but limited genetic differences between them were chosen. The expression patterns of these cell lines was needed to include equally expressed genes as well as differentially expressed genes between the two, enabling analysis of various aspects for the suitability of the amplification protocols in question. Details of the chosen cell lines HL-60 and NU-DHL-1 are given in Appendix A.

Culture and Harvest of Cells

Both cell lines HL-60 and NU-DHL-1 are from myeloid origin and grow in suspension. Cells from frozen stocks (-80 °C or -196 °C) were quickly diluted in 5 ml 1640 RPMI medium (GibCo) containing 20% fetal calf serum (FCS, GibCo) and 1% 100 x Pen-Strep Solution (10,000 U/ml Penicillin, 10,000 µg/ml Streptomycin; GibCo), pre-incubated at 37 °C. After 4 h of incubation at 37 °C and 5% CO_2 (standard conditions) cells were pelleted at mild conditions (2' at 500 rpm) to remove residual DMSO from the freezing medium. Medium supernatant was removed and cells were again diluted in 5-10 ml of the same medium as before (containing 20% FCS) and incubated overnight at standard conditions. This procedure of pelletting and resuspension in medium containing 20% FCS was repeated every 12 h until the cells had grown into clusters for the first time, usually after 2-4 days. Cells were then diluted 1:2 to 1:2.5 and transferred to larger flasks, resuspending them in 20-25 ml medium containing 20% FCS, but changing the medium only every 24 h. When the cells formed clusters for the second time, they were again diluted 1:2 but now in

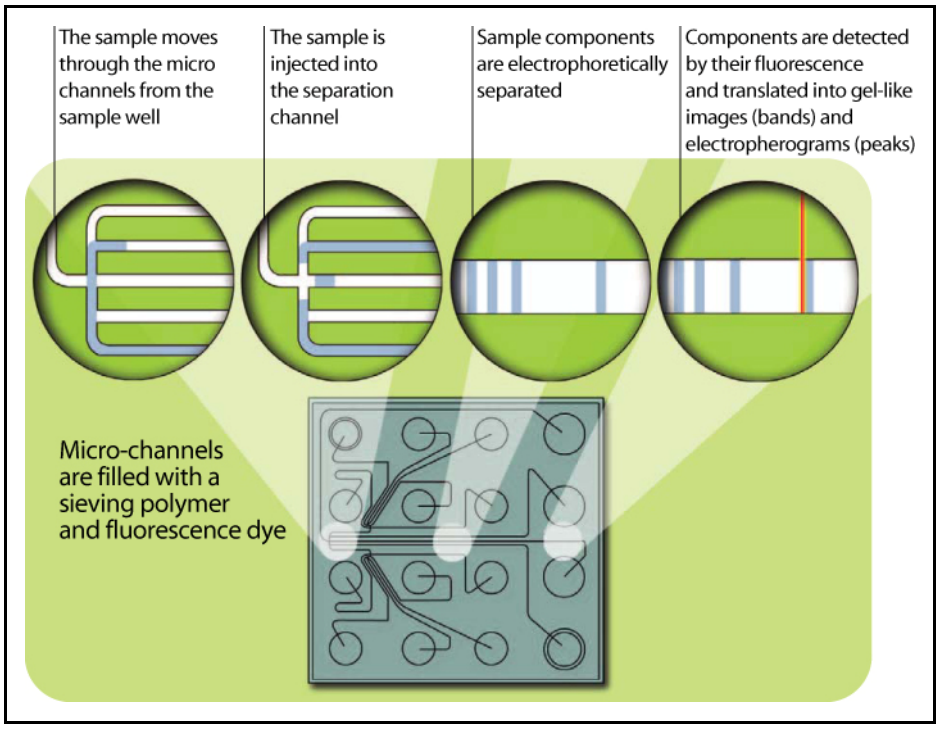
50 ml medium containing only 10% FCS. In this medium, the cell number doubled every 2-3 days on average and they were therefore diluted 1:2 to 1:4 every 2-4 days, as necessary. During resuspension, clusters were disintegrated by passing the cells from the pellet through the end of a glass pipette for several times.
Cells were harvested by centrifugation from 50-100 ml culture medium under harsh conditions (2' at 2,000 rpm). Medium supernatant was discarded and the cell pellet was resuspended in 10-15 ml TRIzol reagent (Invitrogen) at 4 °C.

RNA Extraction
Cells suspended in TRIzol reagent were incubated at room temperature for 5' and then mixed vigorously on a vortex. Chloroform, 1/5 of the volume of TRIzol used (2-3 ml), was added and the suspension was again mixed vigorously. To separate aqueous and organic phases, the tubes were then centrifuged for 30' - 60' at 3,000 rpm and 4 °C. The upper RNA containing aqueous phase (approximately 60% of the total volume) was collected with a pipet, thereby taking care not to take up any of the other two phases. The white intermediate phase contains DNA and proteins, while the pink organic phase contains membrane lipids, DNA and insoluble cell debris. When any amount of these phases was taken up into the pipet tip, this volume was discarded.
The aqueous phase was collected in a new falcon tube and mixed 1:2 with ethanol (p.A.). Immediately afterwards, this mixture was applied to RNeasy midi columns (Qiagen) at room temperature. After each loading step, columns were centrifuged at 3,750 rpm for 5' and the flow-through was discarded. The columns were washed with 4 ml buffer RW1 (Qiagen) and centrifuged for 5' at 3,000 rpm, followed by 2.5 ml buffer RPE (Qiagen) and centrifugation for 2' at 3,000 rpm and again with 2.5 ml buffer RPE but centrifuged for 5' at 3,000 rpm. Each flow-through was discarded. The RNA from the columns was then eluted twice with 250 µl RNase-free water as recommended by the manufacturer. Total RNA was stored at -80 °C.

Quality Control of Extracted Total RNA
Before the first usage or after several freeze-thaw cycles, extracted total RNA needed to be analyzed for yield and integrity or degradation. The yield was

determined by photometric measurements either with UV-spectrometer Cary 50 Bio (Varian Inc.), usually with 1:25 dilutions in RNase-free water, or undiluted in a ND-1000 spectrometer (NanoDrop Technologies). Measurements were taken at 260 and 280 nm wavelength and scans were taken from 230 to 400 nm wavelength. For integrity and degradation analysis, the 2100 BioAnalyzer (Agilent) with RNA 6000 Nano LabChip Kit was used as recommended by the manufacturer. The device works by application of high voltages to a current running through a matrix according to the principle of capillary electrophoresis. It requires only small amounts of RNA for a measurement (25-500 ng). The design of the RNA measurement kit is explained in Figure 7.

Figure 7



Schematic view of the Agilent BioAnalyzer RNA Nano 6000 electrophoresis chip. From Agilent Technologies.

3.2.2. Comparative Amplification and Labeling of RNA

Direct Labeling with Reverse Transcription (RT)

The commonly used protocol for creating fluorescently labeled cDNA from mRNA takes advantage of the Reverse Transcriptase, for example from Moloney Murine Leukemia Virus (M-MLV). Here, the enzyme SuperScript II (Invitrogen)

was used, which had been genetically engineered by the manufacturer to reduce RNase H activity and increase thermal stability. This improves overall yield and incorporation of the bulky nucleotides, as they are covalently coupled with fluorescent Cy-dyes (Amersham). The protocol was not only the starting point but also the benchmark for the testing of amplification procedures.
To ensure selective reverse transcription of messenger RNA, which contains the polyadenylation signal [poly(A)], an "anchored" oligo-d(T)$_{21}$-VN primer was used (Biospring; V = any except thymine, N = any nucleotide).

Table 5 **Direct Labeling Protocol**

SuperScript II RT mix, *on ice!*	**volume [µl]**
5 x 1st strand buffer (Invitrogen)	6.00
0.1 M DTT (Invitrogen)	3.00
RT dNTP-Mix (25 mM dATP, dCTP, dGTP; 10 mM dTTP)	0.60
Cy-dUTP (1 mM) (Amersham)	3.00
RNase Inhibitor (40 U/µl) (Promega)	1.50
SuperScript II RT (200 U/µl) (Invitrogen)	2.00
total volume RT mix	*16.10*
RNA (2-5 µg mRNA or 40-100 µg total RNA)	0.50 - 11.90
Oligo-d(T)$_{21}$ (1 µg/µl)	2.00
RNase-free water	ad 13.90
total volume RNA / primer	*13.90*
total reaction volume	***30.00***

Denature 13.9 µl ***RNA / primer*** 4' @ 70 °C and chill on ice
add 16.1 µl ***RT mix***
3' @ 25 °C
60' @ 42 °C
add 1 µl SuperScript II (200 U/µl)
60' @ 42 °C
add 15 µl 0.1 M NaOH, 2 mM EDTA
20' @ 70 °C
add 15 µl 0.1 M HCl

At least 40 µg of total RNA are necessary as input per channel and experiment to successfully hybridize the generated cDNA onto a microarray with 70-mer oligonucleotide DNA. Assuming an mRNA content of approximately 5% in total RNA extracted with the TRIzol procedure, this corresponds to 2 µg of mRNA.

Total RNA and oligo-d(T) primer were mixed and denatured for 4' at 70 °C and immediately chilled on ice. The RT reagents were mixed according to Table 5 for each fluorescent dye separately. Denatured RNA and primer were mixed with the RT reagents and pre-incubated for 3' at room temperature. The reaction was performed for two hours at 42 °C with addition of another 1 µl (200 U) SuperScript II after one hour. Next, RNA was selectively degraded by addition of 15 µl 0.1 M NaOH / 2 mM EDTA and incubation for 20' at 70 °C, and finally the mix was pH-neutralized by addition of 15 µl 0.1 M HCl.

For disposal of non-incorporated fluorescent nucleotides, very short products and degraded RNA, the reaction mix was passed through Microcon YM-30 columns (Millipore), which retain molecules of at least 30 kDa molecular weight. This corresponds to oligonucleotides with a minimal length of approximately 90 DNA bases, if no fluorescent dyes were incorporated, or 100 RNA bases. The enzymes were also retained, but denatured before by the incubation at 70 °C. For washing, the reaction mixes for both labelings (Cy3 or Cy5) of a hybridization experiment were mixed and diluted with TE buffer to a total volume of 450 µl, then passed through the Microcon columns by centrifugation at 13,000 rpm for 10' and the flow-through was discarded. This washing procedure was repeated twice. In the last cycle, the cDNA was washed with 450 µl TE containing 0.25 µg Cot-1 DNA (Roche), 0.25 µg poly(A) RNA and 0.75 µg bovine or yeast tRNA (both Sigma-Aldrich) per 1 µg total RNA input and centrifuged as above, but this time until the membrane started to become dry in the middle, though not entirely. In this manner, the residual volume was reduced to approximately 10-20 µl. The Cot-1 DNA, poly(A) RNA and tRNA were added as blocking mix to prevent unspecific hybridization events that would give background signals on the array. Although the 70-mer oligonucleotides spotted onto the arrays are said to be designed free of repeat elements by their manufacturer, this blocking procedure was kept as standard. After the last washing step, the columns with the residual volume were inverted into a collection tube and centrifuged for 1' at 13,000 rpm to collect the labeled cDNA and blocking mix. If not used for hybridization immediately, this mix was stored at -20 °C in the dark.

In Vitro Transcription Labeling with T7-RNA Polymerase

The most straightforward solution to the problem of limited RNA available for hybridization from small tissue samples included linear RNA amplification and simultaneously generating labeled antisense RNA. This was performed by first creating double-stranded cDNA from the mRNA introducing a T7 promotor, and then to perform the *in vitro* transcription (IVT) with simultaneous incorporation of fluorescently labeled nucleotides. If feasible, this would produce labeled antisense RNA, thereby amplifying the copy numbers of aRNA by repeatedly transcribing from the double-stranded DNA.

Similar protocols are used both by Affymetrix in the GeneChip technique and by Agilent for their Linear Amplification Kit PLUS.[43,132] The difference between the approaches proposed here or by Agilent and the protocol based on works by Lockhart *et al.* (Affymetrix) is that the latter recommend the usage of biotinylated RNA nucleotides. The cRNA containing these biotin labels are first hybridized onto the GeneChip Arrays, then in a second step detected by binding of streptavidin and thirdly anti-streptavidin antibodies. Consequently, the Affymetrix protocol allows only for detection of one channel per hybridization experiment, a competitive hybridization with two differently labeled samples is impossible. Therefore, a comparison between two tissues, e.g. tumor and reference, requires two different chips or arrays, and concentration or input deviations have to be addressed additionally.

The downside of the approach of incorporating fluorescently labeled nucleotides during IVT is that the T7-RNA polymerase, which is used for *in vitro* transcription, is barely permissive for bulky or modified nucleotides, and therefore has only a low incorporation rate for them. In the tested protocol, we tried to overcome this restraint by using a high concentration of fluorescently labeled nucleotides, which is of course a lavish solution.

Table 6 ***In vitro* Transcription (IVT) Labeling Protocol**

1. Reverse Transcription (RT)			
RNA/primer, *on ice!*	**vol [µl]**	**m [µg]**	**OK**
total RNA	4.0	0.02 - 2	
(dT)-T7 primer (100 ng/µl)	1.0	0.1	
RT-Mix, *on ice!*	**vol [µl]**	**m [µg]**	
1st strand buffer	2.0		
100 mM DTT	1.0		
10 mM dNTP-mix	0.5		
5-8 mg/ml $T4_{gp}32$	0.5	2.5 - 4	
RNase Inhibitor	0.5		
SuperScript II (200 U/µl)	0.5		
total	**5.0**		
denature RNA/primer 4' @ 70 °C, chill @ 4 °C			
add **ice cold** RT-Mix, mix well			
incubate 1 h @ 50 °C with heated lid			
inactivate 15' @ 65 °C			
chill on ice / 4 °C forever			

2. Second Strand Synthesis (SSS)			
SSS-mix, *on ice!*	**vol [µl]**	**m [µg]**	**OK**
5x 2nd strand buffer	15.00		
10 mM dNTP	1.50		
DNA Pol. I (9 U/µl)	2.22		
RNase H (10 U/µl)	0.10		
DNA Ligase (10 U/µl)	0.50		
RNase-free water	45.68		
total	**65.00**		
add ***ice cold(!)*** SSS-mix to RT reaction			
mix well			
incubate 2 h @ 14-16 °C in thermal cycler			
add 10 U T4 DNA-Polymerase (3 U/µl; 3.33 µl)			
mix by flicking and gentle vortexing			
incubate 15' @ 14-16 °C in thermal cycler			
heat inactivate 10' @ 70 °C			
add 75 µl phenol/chloroform/isoamylalcohol (pH 8)			
mix vigorously by pipetting			
transfer to pre-spun PLG Heavy 0.5 ml *			
spin 5' @ 13,000 rpm / RT			
transfer aqueous phase to prepared P-6 MicroSpin **			
spin 4' @ 1,000 x g (3,500 rpm), recover eluate			

3. *In vitro* Transcription (IVT)			
SSS-resuspension	**vol [µl]**	**m [µg]**	**OK**
transfer eluate to 0.6 ml PCR tube			
add LPA		5	
add 1/25 vol 5 M NaCl	3.5		
add 2.5 vol 100% EtOH	220.0		
mix well			
precipitate 30-60' @ -70 °C or 2h - o/n @ -20 °C			
spin tube 30' @ 13,000 rpm, remove s/n			
wash pellet with 500 µl 70% Et-OH			
spin tube 5' @ 13,000 rpm, remove s/n			
pulse spin, remove s/n completely			
allow pellet to dry 2-3' @ room temp.			
resuspend in **8.5** µl water			

Promega IVT#	**vol [µl]**	**Conc.**	**OK**
RNAse-free water	0.0		
5x T7 Transcr. Buffer	8.0		
100 mM ATP	2.6	6.5 mM	
100 mM CTP	2.6	6.5 mM	
5 mM Cy-UTP **(not dUTP)**	9.6	1.2 mM	
100 mM GTP	2.6	6.5 mM	
100 mM UTP	2.1	5.25 mM	
T7 Enzyme Mix	4.0		
SSS resuspension	8.5		
Total	**40.0**		
mix by pipetting and gentle vortexing			
incubate 6h @ 37 °C **in dark condition!!!**			
mix regularly (every 15-30') by gently flicking			

4. aRNA Cleanup			
RLT/β-ME - mix	**vol [µl]**	**m [µg]**	**OK**
β-mercaptoethanol	3.5		
water	76.5		
RLT (Qiagen)	350.0		
total	**430.0**		
aliquot into 1.5 ml tube	430.0		
add IVT product (aRNA)	40.0		
mix well			
add 100% EtOH	250.0		
apply to RNeasy mini column (Qiagen)			
spin 15" @ 8,000 x g (9,900 rpm)			
discard flow-through			
transfer column to new 2 ml tube			
wash with 500 µl RPE (contains ethanol)			
spin 15" @ 8,000 x g (9,900 rpm)			
discard flow-through			
wash with 500 µl RPE (contains ethanol)			
spin 2' @ 13,000 rpm			
discard flow-through			
wash with 500 µl RPE (contains ethanol)			
spin 2' @ 13,000 rpm			
discard flow-through			
transfer column to new 1.5 or 2 ml tube			
spin 1' @ 13,000 rpm			
transfer column to new 1.5 ml tube			
add 30 µl RNase-free water onto membrane			
spin 1' @ 8,000 x g (9,900 rpm)			
repeat eluting steps			
clean with Microcon YM-30			
wash with 450 µl TE			
wash with 500-50 µl TE§ / blocking mix			
concentrate to ~ 10 µl and "elute"			

***** 30'' @ 13,000 rpm

****** resuspend, drain by gravity, 2' @ 3,500 rpm

§ Optionally measure incorporation rate before adding mix (use 50 µl of 500 µl TE resuspension)

#All reagents for IVT mix, **except enzyme**, must be used @ room temp.

The first steps of the protocol, reverse transcription, second strand synthesis and cleanup of double-stranded DNA, were adapted from Kenzelmann *et al.* (Table 6).[125] In short, mRNA from 2 µg total RNA was reverse transcribed using a protocol modified from the Direct Labeling procedure, including a primer combining the promotor sequence for T7-RNA polymerase with the oligo-$d(T)_{21}VN$ sequence from above to 5'-GCA-TTA-GCG-GCC-GCG-AAA-TTA-ATA-CGA-CTC-ACT-ATA-GGG-AGA-$(T)_{21}$VN-3'.[123] The RNA in the resulting DNA-RNA heteroduplex was slowly digested using low concentrated RNase H (Epicentre); thus it could be used to prime the second strand synthesis using DNA polymerase I from *E. coli* (Promega). Joints resulting from different polymerized DNA stretches were closed with DNA ligase, also from *E. coli* (Amersham). Afterwards, overhanging ends were filled ("polished") by use of T4 DNA polymerase (New England Biolabs). Double-stranded DNA was extracted from the reaction mix by use of phenol:chloroform:isoamylalcohol (24:25:1, Sigma) and Phase-Lock-Gels (Eppendorf) and subsequently washed through P-6 MicroSpin columns (Bio-Rad) buffered with TE according to the manufacturer's recommendations. Finally, the DNA was precipitated, using Linear Polyacrylamide (LPA, Ambion) as nucleation agent, and resuspended in RNase-free water to the appropriate volume.

The double-stranded DNA, featuring the T7-RNA polymerase promotor sequence, was used for *in vitro* transcription, using the RiboMAX Large Scale RNA Production System (T7; Promega), modified from the manufacturer's protocol for 40 µl reaction volume and incorporation of Cy-UTPs. The resulting aRNA was cleaned using RNeasy mini columns (Qiagen), also modified slightly from the manufacturer's protocol to contain one more washing step. Analogous to the Direct Labeling procedure, RNA was again washed, blocking mix was added and the blend was concentrated using Microcon YM-30 columns. Before the addition of blocking mix, 10% of the TE-buffered RNA was withdrawn to measure dye-associated nucleotide incorporation rates for both channels. Fluorescently labeled and concentrated aRNA was used immediately for hybridization.

"Baugh Standard" and "Baugh + Klenow" (TAcKLE) Protocols

Based on protocols developed by Eberwine *et al.* and Baugh *et al.* for linear amplification of RNA by *in vitro* transcription (IVT), a protocol for amplification and labeling was developed that is useful in finally yielding antisense fluorescently labeled DNA.[123,124] To obtain labeled antisense DNA from aRNA, their protocols were consequently reduced to a reverse transcription without labeled nucleotides, and a different step to transcribe sense DNA into antisense DNA was appended that could be used for labeling. For the enzymatic reaction step, a DNA-dependent DNA polymerization, the Klenow-fragment of DNA polymerase I (BioPrime Kit, Invitrogen) was selected. It had already been proven useful for generating fluorescently labeled DNA from fractionated genomic DNA in protocols used for Matrix- or Array-CGH.[133] To compare the methods, both the original protocol, as adopted by Kenzelmann *et al.* including labeling during the 2nd round RT reaction, and the labeling procedure with the Klenow enzyme afterwards were performed and evaluated by hybridizing the products onto oligonucleotide microarrays.

The first two steps, from the first RT reaction to the beginning of the IVT, were the same as given in the IVT labeling (see Table 6), since both protocols were derived from the same sources. The IVT itself differed, since here there are no labeled nucleotides incorporated (Table 7). Therefore, it was only slightly modified from the manufacturer's recommendations, with respect to the reaction volume. Antisense RNA extraction and cleanup were again similar to the IVT labeling protocol, using RNeasy mini columns (Qiagen) as above. Afterwards, the aRNA was additionally precipitated, again using LPA as nucleation agent, but using ammonium acetate for RNA precipitation instead of NaCl for double-stranded DNA.

For the "Second Round RT" reaction in the "Baugh + Klenow" protocol (later termed TAcKLE), a slightly different procedure than for the first RT was applied, with respect to its aRNA concentration of 0.25 µg/µl, the use of random hexamer primers (N_6), SuperScript II reverse transcriptase and the reaction temperature profile, as given in Table 8.

For the labeling with Klenow fragment, the enzyme and random primer solutions from the BioPrime Labeling Kit (Invitrogen) were used with a slightly

modified protocol from the recommended one, again concerning the scale of the reaction.

Table 7 **Baugh Standard Protocol**

3. *In vitro* Transcription (IVT)			
SSS-eluate @ RT	**vol [µl]**	**m [µg]**	**OK**
transfer eluate to 0.6 ml PCR tube			
add LPA		5	
add 1/25 vol 5 M NaCl	3.5		
add 2.5 vol 100% EtOH	220.0		
mix well			
precipitate 60' @ -70 °C or 2 h - o/n @ -20 °C			
spin tube 30' @ 13,000 rpm, remove s/n			
wash pellet with 500 µl 70% EtOH			
spin tube 5' @ 13,000 rpm, remove s/n			
spin tube 2' @ 13,000 rpm, remove s/n completely			
allow pellet to dry 2-3' @ 20-40 °C			
resuspend in **10 µl** water			
Promega IVT[#]	**vol [µl]**	**m [µg]**	**OK**
RNase free water	6.0		
5x T7 Transcr. buffer	8.0		
100 mM ATP	3.0		
100 mM CTP	3.0		
100 mM GTP	3.0		
100 mM UTP	3.0		
T7 Enzyme Mix	4.0		
SSS eluate	10.0		
Total	**40.0**		
mix by pipetting and gentle vortexing @ **room temp**			
incubate 6 h @ 37 °C			
mix regularly (every 15-30') by gently flicking			
(freeze @ -20 °C or proceed)			

4. aRNA Cleanup			
RLT / β-ME - mix	**vol [µl]**	**m [µg]**	**OK**
β-mercapto-ethanol	3.5		
RNase-free water	76.5		
RLT buffer (Qiagen)	350.0		
Total	**430.0**		
aliquot into 1.5 ml tube	430.0		
add IVT product (aRNA)	40.0		
mix well			
add 100% EtOH	250.0		

[#] All reagents for IVT mix, **except enzyme**, must be @ room temp.

apply to RNeasy Mini column
spin 15" @ 8,000 x g (9,900 rpm)
discard flow-through
transfer column to new 2 ml tube
wash with 500 µl RPE (contains ethanol)
spin 15" @ 8,000 x g (9,900 rpm)
discard flow-through
wash with 500 µl RPE (contains ethanol)
spin 2' @ 13,000 rpm
discard flow-through
wash with 500 µl RPE (contains ethanol)
spin 2' @ 13,000 rpm
discard flow-through
transfer column to new 1.5 or 2 ml tube
spin 1' @ 13,000 rpm
transfer column to new 1.5 ml tube
add 30 µl RNase-free water onto membrane
spin 1' @ 8,000 x g (9,900 rpm)
repeat eluting steps, measure RNA conc. (5 µl)
precipitate aRNA with 1 µl LPA, 0.5 vol 7.5 M ammonium acetate and 2.5 vol 100% EtOH
wash pellet with 500 µl 70% EtOH
resuspend in 10 µl water

5. Labeling			
RT mix, *on ice!*	**vol [µl]**	**m [µg]**	**OK**
10x Buffer RT	2.0		
RNase Inhibitor	1.0		
RT dNTP-Mix	0.4		
N_6-primer (2 µg/µl)	3.0	6 µg	
Cy-dUTP (Amersham)	1.5		
Omniscript RT (Qiagen)	1.5		
aRNA	10.0	2 - 5 µg	
Water	ad 20.0		
Total	**20.0**		
clean with Microcon YM-30			
wash with 450 µl TE			
wash with 500-50 µl TE[§] / blocking mix			
concentrate to ~ 10 µl and "elute"			

[§] Optionally measure incorporation rate before adding mix (use 50 µl of 500 µl TE resuspension)

The incubation time of 16 h was used as proposed for Matrix-CGH experiments.[133]

Since the Klenow enzyme and random octamer primer in the reaction mix allowed copying one DNA strand from another regardless of sense or antisense orientation, this introduced an approximately 20-fold amplification of copy

numbers. But as the random octamers could prime anywhere in the template sequences, the resulting DNA strands became shorter with every copy cycle. In consequence, the Klenow labeling procedure resulted in both sense and antisense strands comprising fluorescent nucleotides, which differ in length. The cleanup method, however, excluded DNA strands of 90 nucleotides or less as an effect of the pore size of the Microcon YM-30 columns.

Table 8 **Baugh + Klenow Protocol (TAcKLE)**

5. 2nd Round RT

RNA/primer, *4 °C*	**vol [µl]**	**m [µg]**	**OK**
aRNA	4.0	1	
N_6-primer **(0.5 µg/µl)**	41.0	0.5	
RT-Mix*	**vol [µl]**	**m [µg]**	
1st strand buffer	2.0		
100 mM DTT	1.0		
10 mM dNTP-mix	0.5		
5-8 mg/ml $T4_{gp}32$	0.5	2.5 - 4	
RNase Inhibitor	0.5		
SuperScript II (200 U/µl)	0.5		
total	**5.0**		
denature RNA/primer 5' @ 70 °C			
snap cool on ice 2'			
hold 5' @ room temp.			
mix well			
incubation:			
20' @ 37 °C			
20' @ 42 °C			
10' @ 50 °C			
10' @ 55 °C			
15' @ 65 °C			
hold at 37 °C			
add 1U RNase H			
30' @ 37 °C			
2' @ 95 °C			
4 °C forever			

6. Klenow Labeling

Klenow mix, *4°C*	**vol [µl]**	**m [µg]**	**OK**
eluted cDNA	10.0	~ 1	
2.5x Random Primer	40.0		
10x dNTP (low dTTP)#	10.0		
Cy-labeled dUTP	3.0		
water ad 98 µl	35.0		
mix briefly			
Klenow fragment	2.0		
total	**100.0**		
mix gently but thoroughly			
centrifuge 15-30"			
incubate o/n @ 37 °C (~16 h)			
clean with Microcon YM-30			
wash with 450 µl TE			
wash with 500-25 µl TE§ / blocking mix			
concentrate to ~ 10 µl and "elute"			

* RT mix for 2nd Round RT must be @ room temp.!

§ Optionally measure incorporation rate before adding mix (use 25 µl of 500 µl TE resuspension)

#10x dNTP (low dTTP)	Final conc.	Stock conc.	µl
dTTP	0.5 mM	100 mM	5
ATP	2.0 mM	100 mM	20
CTP	2.0 mM	100 mM	20
GTP	2.0 mM	100 mM	20
TE buffer (pH 8.0)	ad	1000 µl	935

Primer-Assisted Linear DNA Amplification (PALDA)

In addition to the abovementioned protocols of IVT-based amplification, protocols based on PCR amplification were also tested. As the standard protocol of PCR uses two primers, one at each end of the sequence in question, each DNA strand is doubled in copy number per cycle. Consequently, the amplification occurs exponentially. However, since the mRNA sequences differ in length from a few hundred to a few thousand bases, there is in theory a bias introduced by an exponential PCR amplification: the length of each cycle determines how many nucleotides can be polymerized in the given time. With

two primers, only a full length sequence will be amplified in all following cycles. Consequently, long sequences have a lower probability to be amplified for the full set of cycles as shorter sequences.

Table 9 **PALDA Protocols (*Taq versus Pfu* exo$^-$)**

3. PCR Labeling *Taq* Polymerase			
SSS-eluate @ *RT*	**vol**	**m [µg]**	**OK**
transfer eluate to 0.6 ml PCR tube			
add LPA		5	
add vol 5 M NaCl	3.5		
add vol 100% EtOH	220.0		
mix well			
precipitate 60' @ -70 °C or 2h - o/n @ -20 °C			
spin tube 30' @ 13,000 rpm, remove s/n			
wash pellet with 500 µl 70% EtOH			
spin tube 5' @ 13,000 rpm, remove s/n			
spin tube 2' @ 13,000 rpm, remove s/n completely			
allow pellet to dry 2-3' @ 20-40 °C			
resuspend in 10 µl water			
PCR mix, *on ice!*	**vol**	**Conc.**	**OK**
10x PCR buffer	5.00		
25 mM $MgCl_2$	3.00		
10 mM dATP	1.00		
10 mM dCTP	1.00		
10 mM dGTP	1.00		
10 mM dTTP	0.88		
1 mM Cy-dUTP	1.25		
BSA (10 µg/µl)	0.50	0.1 µg/µl	
T7 PCR primer [100 µM]	25.00	50 µM	
2nd strand cDNA	10.00		
Taq Pol [5 U/µl]	1.00	0.02 U/µl	
water	0.38		
total	**50.00**		
initital denaturing: 1' @ 95 °C			
2x 50 cycles of: 25" @ 95 °C (denature) 45" @ 65 °C (anneal) 60" @ 72 °C (elongate)			
final elongation: 5' @ 72 °C			
4 °C forever			
after first 50 cycles: (take 1 µl, measure DNA content) add 1µl Taq Pol [5 U/µl] run 2nd time 50 cycles			
clean with Microcon YM-30			
wash with 450 µl TE			
wash with 500-50 µl TE[§] / blocking mix			
concentrate to ~ 10 µl and "elute"			

3. PCR Labeling *Pfu* Polymerase			
SSS-eluate @ RT	**vol**	**m [µg]**	**OK**
transfer eluate to 0.6 ml PCR tube			
add LPA		5	
add 5 M NaCl	3.5		
add 100% EtOH	220.0		
mix well			
precipitate 60' @ -70 °C or 2h - o/n @ -20 °C			
spin tube 30' @ 13,000 rpm, remove s/n			
wash pellet with 500 µl 70% Et-OH			
spin tube 5' @ 13,000 rpm, remove s/n			
spin tube 2' @ 13,000 rpm, remove s/n completely			
allow pellet to dry 2-3' @ 20-40 °C			
resuspend in 10 µl water			
PCR mix, *on ice!*	**vol**	**Conc.**	**OK**
10x PCR buffer	5.00		
25 mM $MgCl_2$	3.00		
10 mM dATP	1.00		
10 mM dCTP	1.00		
10 mM dGTP	1.00		
10 mM dTTP	0.88		
1 mM Cy-dUTP	1.25		
BSA (10 µg/µl)	0.50	0.1 µg/µl	
T7 PCR primer [100 µM]	25.00	50 µM	
2nd strand cDNA	10.00		
Pfu Pol exo$^-$ [2.5 U/µl]	1.00	0.01 U/µl	
water	0.38		
total	**50.00**		
initital denaturing: 1' @ 95 °C			
2x 50 cycles of: 45" @ 95 °C (denature) 45" @ 65 °C (anneal) 120" @ 72 °C (elongate)			
final elongation: 10' @ 72 °C			
4 °C forever			
after first 50 cycles: (take 1 µl, measure DNA content) add 1µl Pfu Pol exo$^-$ [2.5 U/µl] run 2nd time 50 cycles			
clean with Microcon YM-30			
wash with 450 µl TE			
wash with 500-50 µl TE[§] / blocking mix			
concentrate to ~ 10 µl and "elute"			

[§] Optionally measure incorporation rate before adding mix (use 50 µl of 500 µl TE resuspension)

To circumvent this restraint, a similar reaction which comprises only one primer was used, excluding the limiting factor of transcript length and thereby ensuring that the amplification occurs linearly. Analogous to the IVT labeling, the procedure was set up to incorporate the nucleotides during the polymerization reaction.

Reverse transcription and second strand synthesis as well as extraction of the double-stranded DNA were performed as described before. Then, a PCR-like reaction with a single primer, containing a part of the T7 promotor sequence (5'-GCG-GCC-GCG-AAA-TTA-ATA-CGA-CTC-ACT-ATA-GGG-3'), was performed. PCR conditions were optimized for this primer. As the *Taq* DNA polymerase (from *Thermus aquaticus*, Amersham) has a limited tolerance for the bulky dye-associated nucleotides, single primer PCR reactions were performed for 2 x 50 cycles and the *Taq* DNA polymerase was renewed after the first 50 cycles. The general feasibility of incorporation of fluorescently labeled nucleotides during the PCR reaction had been assured by the manufacturer. In addition, the same protocol was also tested with another thermally stable polymerase, *Pfu* exo$^-$ (from *Pyrococcus furiosus*, Stratagene), which had been modified by the manufacturer for elimination of exonuclease activity. This modification was similar to a modification that the M-MLV reverse transcriptase had been subjected to in order to allow for incorporation of bulky nucleotide modifications. The washing and blocking steps were performed as described above.

Single Primer Amplification (SPA)

Another protocol based on amplification with *Taq* DNA polymerase was tested, which had been published just before the comparison was started.[128] To be able to compare this published protocol with the ones introduced here, the reactions were modified to be used with the same enzymes and reagents already made use of (Table 10). In brief, mRNA from total RNA was reverse transcribed using oligo-d(T) - T7 primer and cDNA was complemented to double-stranded DNA as described before. A single primer PCR was performed as above, but using only unlabeled dNTPs (10 mM nucleotide mix). Amplified antisense DNA was then labeled using the Klenow fragment as described.

Table 10 **SPA Protocol, using *Taq* Polymerase**

3. SPA			
SSS-eluate @ RT	**vol [µl]**	**m [µg]**	**OK**
transfer eluate to 0.6 ml PCR tube			
add LPA		5	
add 1/25 vol 5 M NaCl	3.5		
add 2.5 vol 100% EtOH	220.0		
mix well			
precipitate 60' @ -70 °C or 2 h - o/n @ -20 °C			
spin tube 30' @ 13,000 rpm, remove s/n			
wash pellet with 500 µl 70% Et-OH			
spin tube 5' @ 13,000 rpm, remove s/n			
spin tube 2' @ 13,000 rpm, remove s/n completely			
allow pellet to dry 2-3' @ 20-40 °C			
resuspend in 10 µl water			

PCR mix, *on ice!*	**vol [µl]**	**Conc.**	**OK**
10x PCR buffer	10.0		
10 mM dNTP	2.0		
BSA (10 µg/µl)	1.0	0.1 µg/µl	
T7 PCR primer [100 µM]	2.0	2 µM	
2nd strand cDNA	5.0	0.01 µg/µl	
Taq DNA Pol. [5 U/µl]	4.0	0.2 U/µl	
water	76.0		
total	**100.0**		
initital denaturing: 3' @ 94 °C			
50 cycles of: 1' @ 94 °C (denature) 1' @ 62 °C (anneal) 2' @ 72 °C (elongate)			
final elongation: 5' @ 72 °C			
4 °C forever			
clean with Microcon YM-30			
wash with 450 µl TE			
wash with 450 µl TE			
concentrate to ~ 50 µl and "elute"			

PCR-eluate	**vol [µl]**	**m [µg]**	**OK**
transfer eluate to 0.6 ml PCR tube			
add LPA		5	
add 1/25 vol 5 M NaCl	3.5		
add 2.5 vol 100% EtOH	220.0		
mix well			
precipitate 60' @ -70 °C or 2h - o/n @ -20 °C			
spin tube 30' @ 13,000 rpm, remove s/n			
wash pellet with 500 µl 70% Et-OH			
spin tube 5' @ 13,000 rpm, remove s/n			
spin tube 2' @ 13,000 rpm, remove s/n completely			
allow pellet to dry 2-3' @ 20-40 °C			
resuspend in 10 µl water			

4. Klenow Labeling			
Klenow mix, *on ice!*	**vol [µl]**	**m [µg]**	**OK**
eluted cDNA	10.0	1 µg	
2.5x Random Primer	40.0		
10x dNTP (low dTTP)	10.0		
Cy-dUTP	3.0		
water ad 98 µl	35.0		
mix briefly			
add Klenow fragment	2.0		
total	**100.0**		
mix gently but thoroughly			
centrifuge 15-30"			
incubate o/n @ 37 °C (~16 h)			
clean with Microcon YM-30			
wash with 450 µl TE			
wash with 500-25 µl TE§ / blocking mix			
concentrate to ~ 10 µl and "elute"			

§ Optionally measure incorporation rate before adding mix (use 25 µl of 500 µl TE resuspension)

As this published protocol was originally created and optimized for microarrays generated from PCR-amplified cDNA libraries, strand specificity had not been a consideration of the authors. The primary product of the labeling step here, using Klenow fragment and the antisense SPA product as a template, was sense DNA. As described, antisense DNA was expected to be generated by Klenow fragment reactions with the labeled sense product as a new template. Therefore, it was tested whether this protocol produced enough labeled antisense DNA to be hybridized against the sense-orientated DNA on the microarrays, and whether the signal was sufficient despite the additionally created labeled sense DNA.

Template-Switch Single Primer Amplification (ts-SPA)

A special feature of the M-MLV reverse transcriptase as used in the initial step of all protocols described here is the terminal addition of a few nucleotides to the cDNA transcript, mostly three cytosines. Though this represents only a short template, it had been shown to be sufficient for priming a 3'-extension of the first strand cDNA.[127]

This effect had been used for switching the template from sense (second) to antisense (first) strand of the double-stranded DNA.[129] Therefore, the single primer PCR could be used to amplify DNA using the elongated first strand as template, yielding multiple copies of the sense strand. These could then be replicated again, using the Klenow fragment, into labeled antisense DNA.

The only difference in the ts-SPA reactions to ones described for SPA was the addition of a TS primer to the oligo-d(T) - T7 primer, during the RT reaction (TS primer sequence: 5'-CG-GCC-AGT-GAA-TTG-TAA-TAC-GAC-TCA-CTA-TAG-GCG-3']. The TS primer then remained present during the second strand synthesis; consequently the first strand could be extended in the course of the reaction. For the single-primer amplification PCR, the same TS primer was used again to create the sense DNA. Klenow labeling reaction was performed as described before.

PCR

To have a benchmark for analysis using abovementioned different PCR amplification variants, a normal or standard PCR amplification with two primers and fluorescently labeled nucleotides was also performed. Since these bulky nucleotide derivatives were known to have a low incorporation rate, the protocol was modified according to the above described PALDA protocols. In brief, the number of cycles was also increased to 100 with intermediate refreshment of *Taq* DNA polymerase after 50 cycles, but the usage of the modified dCTP nucleotide derivatives instead of dUTP. The primers used for the reverse transcription using SuperScript II were oligo-d$(T)_{21}$VN primer and the TS-primer (see above). Subsequently, cDNA was cleaned using Microcons columns as described before. Then the PCR reaction was performed on 1 µg cDNA with internal primers, to suppress amplifying false products, and in

presence of fluorescently labeled dCTP nucleotides. Amplified and labeled DNA was again cleaned up with Microcon columns as described.

Blocking Mix

The products of all abovementioned protocols were hybridized to microarrays in presence of a mix consisting of Cot-1 DNA (Roche), tRNA and poly(A)-RNA (both Sigma-Aldrich). These nucleotide sequences were used to inhibit binding of free repetitive RNA sequences or non-messenger RNA molecules, mostly in an unspecific manner, to probes on the microarray. Although this procedure is not necessary in each of the above introduced protocols, it was nevertheless performed in all of them for comparative reasons.

Table 11 **Blocking Mix**

Blocking Mix	vol [µl]	m [µg]	OK
Cot-1 DNA (1 µg/µl)	25.0	25	
Poly(A) RNA (5 µg/µl)	5.0	25	
tRNA (10 µg/µl)	7.5	75	
total	**37.5**		

3.2.3. Analysis of Comparative Amplifications

DNA and RNA Purity

Nucleic acid purity and concentration were measured to ensure the optimum prerequisites for a successful hybridization event. This was performed during the last washing step at the end of each protocol, before the blocking mix was added and the volume was reduced on the Microcon columns. A small volume (5-10%) was taken and measured on the UV-spectrometer Cary 50 Bio (Varian Inc.), taking the absorption value at the peak of 260 nm (A_{260}) for DNA or RNA content and applying multiplication factors (37 µg/ml, single stranded DNA; 40 µg/ml, single stranded RNA) to estimate the nucleic acid concentration. Calculating the ratio of A_{260} over A_{280} was used for purity measurements, anticipating values for DNA of 1.8 to 2.0 and for RNA of 1.9 to 2.1 for good results.

Incorporation Rates of the Fluorescently Labeled Nucleotides

The performance of each of the amplifying and labeling methods tested was evaluated by measuring and calculating the incorporation rate of labeled nucleotides. This was carried out together with the DNA or RNA concentration and purity estimations on the UV-spectrometer.

Incorporation of fluorescently labeled nucleotides was estimated from the characteristic absorption maxima for Cy3 and Cy5 dyes, at 550 nm and 650 nm, respectively. These peak values were set into relation with the nucleic acid concentration to determine the incorporation rate. The rates for each channel were compared with each other, to identify imbalances in the tolerance of the different enzymatic approaches to the two types of bulky nucleotides. To estimate the incorporation yield of the corresponding reaction of each labeling procedure, an approximation to the average incorporation efficiency, given as

$$r_I = \frac{A_{dye}}{A_{260}} \times \frac{\varepsilon_{Nucl.\ Acid}}{\varepsilon_{dye}} \qquad [1]$$

was used. Here, r_I is the incorporation ratio, A is the absorbance at a specific wavelength in nm (550 nm for Cy3 dye and 650 nm for Cy5 dye, respectively), and ε is the extinction coefficient in $cm^{-1}M^{-1}$ at the absorption maximum for either nucleic acid or the respective dyes. Values for DNA (ε = 10,162.5 $cm^{-1}M^{-1}$) and RNA (ε = 10,418.75 $cm^{-1}M^{-1}$) were estimated from averages per nucleotide of measured and published values for any possible dinucleotide.[134] Values for the two dyes Cy3 (ε = 150,000 $cm^{-1}M^{-1}$) and Cy5 (ε = 250,000 $cm^{-1}M^{-1}$) were provided by the manufacturer. The r_I was additionally multiplied with 1,000 to calculate the average number of incorporated fluorescently labeled nucleotides per 1,000 nucleotides.

Outlier Features

After scanning the microarray images, the grid was placed in GenePix Pro Software onto the image and feature alignment as well as detection of false positive spots, which were flagged as outliers, were performed as described. Only features of good quality were used for further analyses, and the rate of outlier features was considered as a quality estimate.

Spot Homogeneity

Each feature spot consists of some 50 to 500 pixels, which are taken as foreground measurement. To receive a single intensity value for the feature, either the median or mean of the intensities of these pixels can be used. To see whether the intensity distribution within the spot of a feature is homogenous, the ratio of mean over median was taken. Ideally, this ratio should have the value of one for a homogenous spot. A deviation of up to 0.2 below or 0.25 above this quotient was considered acceptable, features with a mean to median ratio outside the interval were considered inhomogeneous.

Feature and Background Signal Comparison

A general assessment of the amplification and hybridization success was achieved when comparing the median signal intensities of features *versus* each feature's local background (the surrounding area of the DNA spot). These were set into relation to each other and averaged across all features of each chip to estimate a compatible value for the different protocols.

Scatter Plots

Scatter plots represent the intensities of the features in one dye channel *versus* the other. For this scheme, median intensity values from the raw data tables for each feature were plotted on a two-dimensional scale, each feature being represented by its corresponding logarithmic values for Cy3 and Cy5 intensity. To integrate dye-swap experiments, the ratios between the two channel values were taken for each DNA spot and transformed by natural logarithm. These ratios were then plotted against each other for both experiments.

This method allowed for a more elaborate examination of the distribution for the following reasons: A bias, e.g. towards smaller intensities or loss of dynamic range, could be detected much better when looking at the corresponding plots as compared to looking at the images themselves or an average of signal to background. In this respect, the different amplification and labeling procedures were much easier to compare on the basis of their scatter plots. In addition, differential and same-*versus*-same hybridization results could be compared with each other to identify effects derived from the

amplification procedure but independent from the distribution of the different mRNA copy numbers between the two cell lines.

Linear Trend of Intensity Scatter Plots

Taking the median feature intensity values from the raw data table for each channel, the differences between procedures to amplify and label were also accountable for in measurable parameters. The distribution of these intensities, as seen on the scatter plots, was described by calculating the slope and intercept of the trend line derived from the feature intensity spots scattered on the plot. A deviation of the slope from the value one, which represents the bisecting line of the plot, could be explained by different input amounts. The intercept or offset on the ordinate, however, expresses a bias towards one of the channels, e.g. by incorporation incompatibilities of the dyes.

Correlation of Gene Expression Patterns

To account for similarity between the distributions of intensities across all features of a chip for both channels, the Pearson's coefficient was estimated. On a scale from -1 to 1, it provided an assessment of similarity between two data sets, or matrices representing them, independent of the slope of the trend line of their distribution. The grades of similarity between same-*versus*-same hybridization intensities or between repeat experiments, such as the dye-swaps, were of particular interest in this comparison. In the case of repeat experiments, this value signified the reproducibility of the amplification procedure.

The calculations described in this chapter were performed with the statistical software "R" [www.r-project.org], a script-based programming environment.[135]

3.3. Gene Expression Signature Predictive for Chemotherapy in Primary Breast Cancer

This study was conducted to investigate whether a gene expression signature could be identified in tissue specimens taken from primary breast tumors of patients that allows for predicting the patient's outcome, or response, to a chemotherapy applied after taking the specimen.

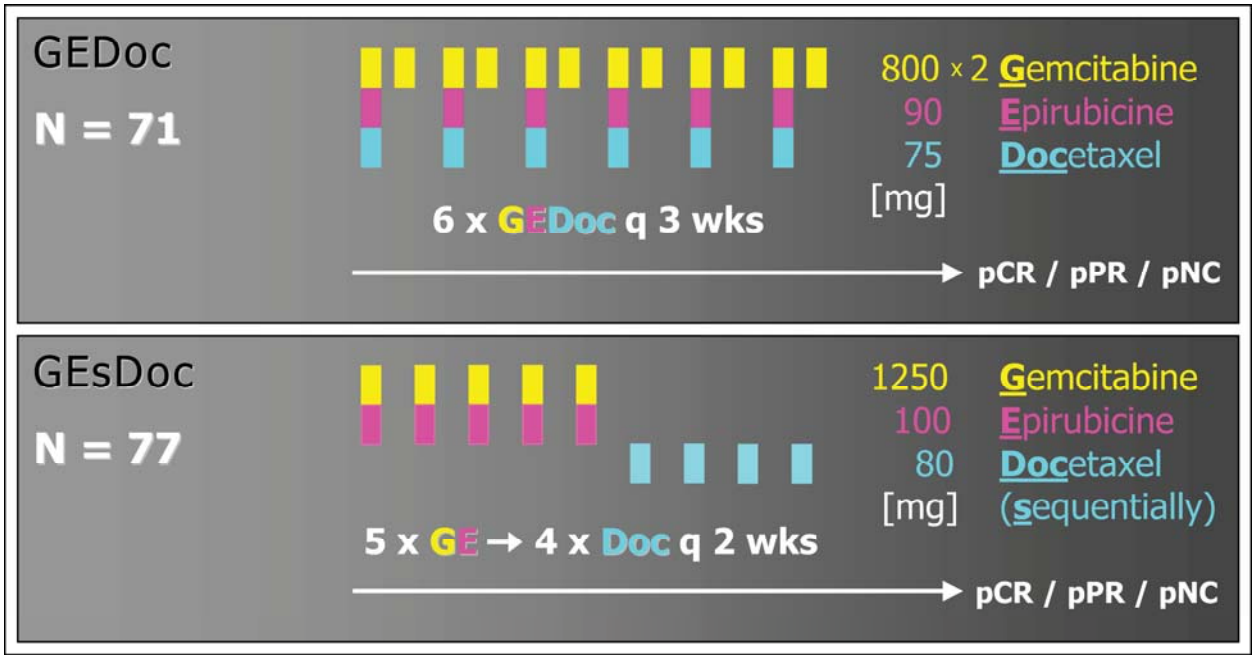
To generate this gene expression signature, the principles of the techniques described above were applied, extracting the RNA from the tumor samples, amplifying the mRNA with the chosen procedure (see chapter 4.1), labeling them with fluorescent dyes in the process, and hybridizing them to microarrays generated from the Human Oligo Set 2.0 or 2.1.1.

3.3.1. Patients and Chemotherapy Protocol

The primary breast cancer specimen used here were provided by the Department of Gynecology and Obstetrics of the University Hospital Heidelberg (Universitäts-Frauenklinik) from female patients (n=148) who participated in two similar studies evaluating new chemotherapy protocols, combining gemcitabine, epirubicin and docetaxel as anti-cancerous agents. Patients were recruited with their voluntary commitment, if they had no prior chemotherapy treatment, their tumor had a diameter of at least 2 cm, they had a maximum of nine metastatic local (internal) or axillary lymph nodes and no distal metastases (WHO classification: T2-4 N0-2 M0), among other criteria.

The evaluation of the chemotherapy protocols concerned the treatment dosages and schedule. In the course of this assessment, these parameters were modified, resulting in two different cohorts, treated with either the "GEDoc" or "GEsDoc" protocols (Figure 8). The major difference between the two schedules was comprised of the sequential application of docetaxel.

Figure 8



Patients enrolled in clinical study evaluating GEDoc or GEsDoc chemotherapies. Patients participated in one of either studies of neoadjuvant chemotherapy protocols administering gemcitabine, epirubicine and docetaxel in the Clinic for Gynecology and Obstetrics, University of Heidelberg. After therapy, surgery of the residual tumor was performed and the response to chemotherapy was estimated by pathology (pCR, pathological complete remission; pPR, pathological partial remission; pNC, pathological no change). Modified after A. Schneeweiß, University of Heidelberg.

3.3.2. Tumor Samples and Reference

Tumors were sampled by taking core cut biopsies with a 14-gauge needle under surveillance of sonographic life-imaging. Up to five of these biopsies, each yielding a tissue sample with a maximal size of 20 × 2 × 2 mm, were taken from each tumor at the time of diagnosis, before chemotherapy treatment of the patient. One or two of these samples per patient were available for gene expression measurements with the microarray technique. These tissue samples were locked in cryo tubes and shock frozen in liquid nitrogen (-196 °C) within 10' after being taken from the tumor and were kept at -80 °C until processing. The total number of investigated tumor samples was 174, taken from 148 patients.

Extraction of RNA

Deep frozen tumor tissue samples were cooled in liquid nitrogen to -196 °C and quickly transferred from the cryo tubes to polytetrafluoroethylene (PTFE or Teflon®) containers (NeoLab) suitable for ball milling, which had been pre-cooled in liquid nitrogen as well. The containers containing the tissue sample

were then equipped with the appropriatly pre-cooled tungsten balls (5 mm caliber, NeoLab), locked and again cooled in liquid nitrogen to ensure that the sample remained deeply frozen. Then the sample was milled in the container with a dismembrator (B.Braun) at 3,000 rpm for 10", if necessary for several times with intermediate cooling of sample, ball and container in liquid nitrogen, until the tissue sample was completely ground to powder. After another cooling cycle, the tissue powder was collected into a 15 ml-Falcon tube which had been prepared to contain 5 ml of TRIzol solution at 4 °C. The mix was vigorously shaken and left for 5' to equilibrate at room temperature. Then, 1 ml of chloroform was added and the suspension was again vigorously shaken and mixed on a vortex. Centrifugation and obtaining of the aqueous phase containing the RNA was carried out as described before (Chapter 3.2.1). After mixing of the aqueous phase 1:2 with ethanol (p.A.), it was applied to RNeasy mini columns (Qiagen) and centrifuged at room temperature for 4' at 8,000 rpm (6,000 × g) in a bench-top Biofuge fresco (Kendro). The columns were washed as recommended by the manufacturer with the buffers provided, and the RNA was eluted twice with 30 µl RNase-free water. After taking 5 µl of the elution for RNA yield and quality assessments, the remaining total RNA dilution was stored at -80 °C until used for amplification and labeling. The yield and purity were measured in 1:25 dilution of the RNA (3 µl in 75 µl total volume) as described on the UV-spectrometer and the quality was assessed by application of the "Lab-on-a-chip" system (Agilent) as described before (Chapter 3.2.1).

Reference RNA

For clinical and ethical reasons, only a restricted amount of non-cancerous tissue of the mammary was available and could not be used as source for reference RNA. The RNA chosen was Human Universal Reference RNA (Stratagene), consisting of total RNA from a mixture of ten cell lines of different cancer origin. Since expression levels of different tumor classes between each other were compared here, namely those from patients with complete remission *versus* those with partial remission or no change, the origin of the reference RNA could be neglected.

3.3.3. Amplification of mRNA from Samples and Reference RNA

The Baugh + Klenow protocol, also named TAcKLE (see chapters 3.2, 4.1 and 5.1) was chosen for amplification and labeling of the tumor mRNA. To avoid introducing a bias between RNA extracted from the tumor specimens and the reference RNA, the latter was also amplified using this procedure. Since tumor samples were recruited consecutively, reference RNA was amplified in bulk and the c*DNA was pooled before labeling the aliquots as needed.

The RNA extracted from tumor samples was precipitated using LPA, ammonium acetate and ethanol as before for aRNA cleanup during the TAcKLE procedure, and suspended to a concentration 0.5 µg/µl. Such prepared RNA was amplified in two different aliquots with a maximum of 2 µg total RNA input each, and the sense-orientated c*DNA was copied, including labeling using Cy3- and Cy5-modified dUTPs, respectively, to perform color switch repeat experiments. Each labeled tumor sample was then mixed with adversely labeled reference RNA sample, washed on Microcon columns and complemented with blocking mix containing 25 µg Cot-1 DNA (Roche), 25 µg poly(A) RNA and 75 µg tRNA (both Sigma-Aldrich). These combined sample-reference mixes, ready for hybridization, were stored at -20 °C for a maximum time of two weeks, if they were not applied to the microarrays on the same day.

3.3.4. Hybridization to Microarrays and Data Pre-Processing

Labeled tumor and reference sample mixes were hybridized to the oligonucleotide microarrays generated on the HybStation as described. The hybridized microarray slides were scanned using the Axon 4000B scanner at 5 µm resolution with adaptation of the optimal settings for the PMT voltages to correct for different incorporation rates and bleaching characteristics of the different Cy-dyes attached to the dUTP nucleotides.

Raw data tables were generated by imposing the appropriate grid on the microarray images, inspecting and flagging degenerated resulting spots and saving the data set to the computer. Since all tumor RNA samples had been amplified and labeled with both Cy3- and Cy5-modified nucleotides, the entire data set for each patient contained two of these raw data tables. Together with the clinical data relevant for analysis of the gene expression signatures, these

raw data were uploaded into a database (ChipYard), developed and maintained by Grischa Tödt from the Division of Molecular Genetics.

3.3.5. Data Analysis

The pre-processing and data analysis described in this chapter were performed in collaboration with Grischa Tödt from the Division Molecular Genetics.
In a first step, the raw data tables from the microarrays were inspected for aspects of quality concerning the hybridization experiment, the amplification and the integrity of the underlying RNA input. These quality assessments were achieved by (i) plotting the median intensity values for both channels in scatter plots, (ii) box plots showing average, upper and lower 25% percentiles for each microarray and channel, (iii) screening of the averages for red and green intensities per spot versus their log ratio (M/A plots) and (iv) plotting the distribution of the spot background intensities according to spot localization. In the course of this quality examination, repeat hybridizations with switched dye assignments were compared with each other for estimating reproducibility. Based on these analyses, experiments showing low quality of RNA, improper amplification or hybridization results were identified and excluded from further processing and analysis. Data sets for patients with excess RNA, whose low quality outcome could not be explained by low RNA quality as evaluated with the BioAnalyzer, where highlighted. If possible, for these both dye swap experiments were repeated from the beginning of the amplification procedure, even if only one of the experiments showed low quality in the assessment of the raw data. This guaranteed prevention from a bias introduced by different experiments within dye-swap pairs.

Individual spots were scored for homogeneity by calculating mean over median of the raw intensity values, and for dynamic range (signal-to-background intensity ratio) by calculating corresponding median fore- over background intensity ratios. As a third component, the standard deviation of intensity ratios (before normalization) between replicate spots within one microarray was also calculated. The scores of these three feature quality measurements were combined by multiplication, features were ranked and the lower 30% of them were eliminated.

After choosing the data sets for experiments with sufficient reliability, the values for median spot intensities were normalized to balance out the different dynamic ranges between individual microarrays. For this purpose, the variance stabilizing normalization (vsn) method was applied.[136] In the course of this calibration, the median intensities (x) were transformed by

$$h\,(x) = \operatorname{arsinh}\,(a + bx) \qquad [2]$$

The parameters a and b describe the onset and magnitude of contraction of small intensities (near the detection limit) towards zero in comparison with the logarithmic conversion and are estimated iteratively by the vsn function. Large intensities, on the other hand, are affected by this contraction to a much lesser extent and therefore coincide with a natural logarithmic transformation. For this reason, the variance stabilizing normalized values h could be used for building ratios between intensities in the same manner as logarithmically transformed values, meaning that while

$$\log_n\,(x_{i,red} \,/\, x_{i,green}) = \log_n\,(x_{i,red}) - \log_n\,(x_{i,green}) \qquad [3],$$

the logarithm was in this case substituted with the vsn transformation $h\,(x)$ to

$$\log_n\,(x_{i,red} \,/\, x_{i,green}) \approx h\,(x_{i,red}) - h\,(x_{i,green}) \qquad [4].$$

This derivation of the logarithmic transformation, however, required caution when considering ratios resulting from low intensity values.

After normalization, repeat spots within one microarray and corresponding values for color-switch repeat experiments were averaged. If, according to the filtering that had been performed before, individual features had been removed from one array, the matching dye swap partner had to be eliminated as well.

3.3.6. Identification of the Gene Expression Signature

The data analysis described in this chapter was performed in collaboration with Grischa Tödt from the Division Molecular Genetics and Dr. Patrick Warnat from the Division Theoretical Bioinformatics.

To detect the most relevant genes to distinguish between patients with a complete response to chemotherapy (pCR) and patients with a partial response or no change in tumor growth (pPR or pNC, respectively), the genes had to be ranked according to their predictive power. The ranking of genes was performed in the process of learning to classify the patients by usage of the Support Vector Machines (SVM) algorithm on the training subset (GEsDoc patients) with five times repeated five-fold cross-validation.[137] Afterwards, the most predictive genes were chosen and the predictive power of the set of genes was estimated. To test independently whether the prediction based on the ranked genes is accurate, the expression data were divided into two sets, one to identify the predictive gene subset (training set), and the other data set to test the predictive power (test set).

To minimize the number of genes, the Recursive Feature Elimination (RFE) algorithm was applied.[138] The RFE approach recursively reduces the number of genes used in the predictor function by removal of those genes with lowest weights and re-fitting of the SVM algorithm using the remaining genes. In the first step, the number of genes used is reduced to the highest power of two that is smaller than the total number of genes. In each following step of the RFE procedure, half of the genes are eliminated from the predictor model until only one gene is left. The minimal number of used genes with a predictive value of at least 0.8, which was set as a threshold, was the constraint of this selection. Finally, the RFE approach was applied once on the training set to generate a final predictive model with the optimal number of genes as determined in cross-validation. Microarray data for patients receiving GEDoc therapy (48 patients) were used as a test set for independent validation of this gene expression based predictor of pCR. The final predictive model generated a predictor score for every patient in the test set. Sensitivity and specificity, resulting from different cut-off values, were visualized by a Receiver Operating Characteristic (ROC) graph. A ROC graph shows how sensitivity and specificity vary together as the cut-off value (that determines the class prediction for a

given sample) on the output of a prediction function for a given test set of samples is varied between the extremes of the prediction function output. The cut-off value yielding a maximal Youden's Index (sensitivity + specificity - 1) was used to determine the binary classification of the test set into pCR and non-pCR cases.

3.4. Real-time Quantitative PCR

To validate the results from the microarray data for gene expression, a real-time quantitative PCR (RQ-PCR) was performed on reverse transcribed mRNA from all patients with sufficient RNA after performing the microarray analysis, for selected genes. The selection was comprised of genes belonging to the signature predicting the patient's outcome, plus two genes (*ESR1*, *HER2*) which belong to markers classically used in immuno-histochemical pathology and another two genes (DCTN2, GALNAC4S-6ST) which were used as reference genes. The references were chosen from the genes on the microarray by analyzing the results of these for minimal differential expression between the two patient groups (responder *versus* non-reponders), minimal standard deviation across patients within these groups, and functional ontology in the sense of metabolic and/or structural activity within living cells. This was necessary to ensure that the results were unbiased for tumor cell activity, since these standard genes were used to normalize between individual patients and across all genes of interest (target genes).

3.4.1. Reverse Transcription

The Reverse Transcription (RT) reaction to create cDNA, and the following RQ-PCR reaction mix to measure the content of each specific gene transcript were carried out as given in Table 12. For estimation of amplification efficiency, the cDNA generated from of Universal Reference RNA (Stratagene) was diluted in seven serial 1:4 dilution steps, starting with cDNA corresponding to 512 ng total RNA.

Table 12 **RQ-PCR Protocol**

Reverse Transcription for RQ-PCR			
RNA/primer mix	**amount [ng]**	**vol [µl]**	**OK**
$d(T)_{21}VN$ primer, 1 µg/µl	300	0.30	
total RNA, 2 µg/µl	3,000	1.50	
dNTP, 10 mM		0.60	
ddH_2O		5.40	
mix, 5' @ 65 °C, snap cool on ice, spin down			
add 5x 1st strand buffer		2.40	
add DTT, 0.1 mM		1.20	
mix and incubate 2' @ 42 °C			
add SuperScript II, 200 U/µl		0.60	
	total volume:	12.00	
incubate 50' @ 42 °C			
incubate 15' @ 70 °C			
dilute by 1:1.25 to 0.01 µg/µl cDNA (≈ 5% of total RNA), add 3 µl H_2O			

RQ-PCR			
2x SYBR Mix (Thermo Scientific)		10.00	
upper primer, 5 mM		0.20	
lower primer, 5 mM		0.20	
template cDNA, 0.01 µg/µl	32	3.20	
ddH_2O		6.40	
	total volume:	20.00	

3.4.2. Primer Design

For each of the measured genes, at least two primer pairs were designed. For exclusion of quantitative bias resulting from amplification of genomic DNA during the PCR reaction, the two primers of each pair were located within different exons of the gene with the largest possible intron, or more than one intron embedded between them, to create a minimum genomic distance of 2kb. The maximal distance on the mature mRNA between the two primers, however, was limited to a size of 100 to 400 bases, so the amplicon could be optimally duplicated during each round of PCR, as given by the distinct elongation time. To ensure a uniform annealing performance, all primers were designed to have a T_m of 60 °C, and the resulting amplicon was required to have a minimum T_m of 78 °C (both estimated with standard PCR conditions of 50 mM Na^+ and 250 pM primer concentration).

Each primer pair was tested for optimal results with templates of 25, 50 and 100 ng of total RNA from the reference RNA to estimate correct efficiency, 50 ng of genomic DNA to exclude amplification of genomic templates and "no template control" to exclude primer dimerization products. Primer pairs with efficiency outside 1.7 to 2.05 or products with a crossing point (CP) larger than 40 cycles were neglected, and a new primer pair was generated for the respective gene. Primer pairs were also neglected if their product's melting curve, as measured with the RQ-PCR thermocycler by fluorescence, did not match with the estimated melting temperature characteristics. For resulting primer sequences, see Appendix C.

3.4.3. RQ-PCR Measurements and Analysis

Measurements of real-time quantitative PCR were performed on an ABI PRISM 7700 Sequence Detection System (Applied Biosystems) in triplicates; using 10 µl of 2x SYBR Mix (Thermo Scientific), cDNA derived from 32 ng of total RNA and a concentration of 50 nM of each primer in a total volume of 20 µl per well (Table 12). Only one 384-well plate per gene was used, and all control RNA dilutions and samples were measured on the same plate to eliminate plate-to-plate variance bias within genes.

Analysis was performed using algorithms based on the efficiency of the PCR reaction for each tested gene. This efficiency estimation [5] uses the slope of the CP values *versus* the logarithmic input for the dilution series of the reversely transcribed control RNA.[139]

$$E = 10^{[-1/\text{slope}]} \qquad [5]$$

The difference of expression levels of each gene in question (target genes) between cDNA from the control RNA and sample RNA were then normalized to the differences between controls and samples for the reference genes [6], resulting in a normalized ratio of expression in samples *versus* control.

$$\text{ratio} = \frac{(E_{\text{target}})^{\Delta \text{CP}_{\text{target}}(control - sample)}}{(E_{\text{ref}})^{\Delta \text{CP}_{\text{ref}}(control - sample)}} \quad [6]$$

These normalized expression ratios for each gene of interest (target gene) against the two different reference genes were averaged to further lower any bias produced by different expression levels of these reference genes. This average was used as the final expression ratio and compared between the groups of patients either completely responding to the chemotherapy (pCR) or not completely responding (pPR and pNC combined) to validate the microarray expression data.

3.5. Antibody Generation

To create a monoclonal antibody, it was necessary to produce recombinant protein of interest and isolate it to a very high grade. This was performed for the BAMBI protein, a TGF-β receptor - like protein residing in the plasma membrane of certain epithelial cells. Due to its nature as a plasma membrane protein and its considerable amino acid sequence similarity to the TGF-β receptor family protein BMP-receptor type 1b, only the cytosolic parts of the protein not homologous to BMPR1B were expressed for generating an antibody against (BAMBI cytosolic fragment). For later testing of the generated antibody serum candidates, BAMBI full length and BMPR1B proteins were expressed both in eukaryotic and prokaryotic settings.

3.5.1. Preparation of cDNA and Cloning into Expression Vectors

To obtain vectors for prokaryotic and eukaryotic expression of BAMBI protein, the mRNA was first reversely transcribed into cDNA, as described above and in Table 12. Then the BAMBI coding sequence (CDS) was isolated using a PCR with gene-specific primers and a low cycle number (n=15). Using agarose gel-electrophoresis to identify and excise the correct DNA strands, and Rapid Gel Extraction Protocol (Marligen Bioscience) to clean them up, excess cDNA was discarded. In a second step PCR reaction, primers were used which additionally contain the necessary sequences for enzymatic restriction and re-ligation into the expression vectors pBCHGs and pQC-His, using the sites for *Bam* HI and *Hin*d III.

As a negative control for later experiments, the same procedure was also conducted for BMPR1B protein. However, for the successful isolation of the *BMPR1B* coding sequence, the PCR had to be performed using nesting primers, containing locus-specific sequences outside of the actual coding sequence. After agarose gel-electrophoresis and clean-up of this longer PCR product, the second PCR was performed again using primers containing the gene-specific sequences and the additional restriction sites. Primer sequences are given in Table 13.

Table 13 **Primers used for Cloning from Reversely Transcribed mRNA**

Gene	Position	Use	Sequence (5'- ... -3')
BAMBI	upper	cDNA	ATG-GAT-CGC-CAC-TCC-AGC-TAC-ATC
	lower	cDNA	TCA-TAC-GAA-TTC-CAG-CTT-CCC-GTG
	upper	cloning	GAG-AGA-GGA-TCC-ATG-GAT-CGC-CAC-TCC-AGC-TAC-ATC
	lower	cloning	GAG-AGA-AAG-CTT-TCA-TAC-GAA-TTC-CAG-CTT-CCC-GTG-C
	upper	sequencing	GGC-GGA-TCC-ATG-GAT-CGC-CAC-TCC
	lower	sequencing	GGC-AAG-CTT-TCA-TAC-GAA-TTC-CAG-CTT-CCC
BMPR1B	upper	cDNA, nesting	CAG-CCG-CGG-GGT-GGA-GTT
	lower	cDNA, nesting	TGA-TGT-CTT-TTG-CTC-TGC-CCA-CAA
	upper	cloning	GAG-AGA-GGA-TCC-ATG-CTT-TTG-CGA-AGT-GCA-GGA-AA
	lower	cloning	GAG-AGA-AAG-CTT-TCA-GAG-TTT-AAT-GTC-CTG-GGA-CTC-TGA-C
	upper	sequencing	TCA-GAG-TTT-AAT-GTC-CTG-GGA-CTC-TGA-C
	middle 1	sequencing	GAA-GTG-GAT-CAG-GCC-TCC-CTC-TG
	middle 2	sequencing	CGA-GTT-GGC-ACC-AAA-CGC-TAT-ATG
	lower	sequencing	ATG-CTT-TTG-CGA-AGT-GCA-GGA-AAA
pQC-His plasmid, 40bp upstream of *Bam* HI site		sequencing	CGG-ATA-ACA-ATT-TCA-CAC-AG
pBCHGs plasmid, 60bp upstream of *Bam* HI site		sequencing	GGT-CCT-TCT-TGA-GTT-TGT-AAC-AG

Afterwards, gel-electrophoresis and clean-up was performed again for both constructs to isolate the correct length CDS / restriction site product and discard excess primers, nucleotides and polymerase. DNA concentration was determined using spectrometric measurements and the DNA was stored at -20 °C. To ensure that no mutations had been introduced during the previous PCR reactions, a small aliquot of 500 ng was used for sequencing.

Vector DNA for pBCHGs and pQC-His was provided by the laboratory of Prof. Dr. Hanswalter Zentgraf from the DKFZ. 100 µl of *E.coli* bacterial cells of strain XL1-Blue (Stratagene) were transformed with 1 - 10 ng of the vector DNA [retransformation], harboring an ampicillin resistance gene (*β*-lactamase), and the competent bacterial cells were plated in 1/5 and 4/5 aliquots onto two LB-agar plates each, containing 100 µg/ml ampicillin. Cells were grown overnight (12-16 h) at 37 °C, then stored at 4 °C.

For each preparation, up to six isolated bacteria clones from these plates were picked and used to inoculate LB-Amp mini-cultures (5 ml LB containing 100 µg/ml ampicillin). Again, these cultures were grown overnight (12-16 h)

and plasmid DNA was harvested with Plasmid Mini kits (Qiagen) according to the manufacturer's protocol. The DNA was checked for concentration by photometric measurement on a NanoDrop spectrometer and for plasmid purity by electrophoresis on 1.2% agarose gels.

The plasmids and cDNA inserts were then each digested using both restriction enzymes *Bam* HI and *Hin*d III (Roche) in a combined restriction reaction, using SureCut buffer B (Roche) according to the manufacturer's protocol. Linearized plasmid DNA was then separated by agarose gel electrophoresis, excision, clean-up as above and concentration measurement.

Afterwards, the linearized vector strands were processed with shrimp alkaline phosphatase (Roche) according to the manufacturer's protocol to dephosphorylate the 5'-phosphate from the DNA. This optional step was performed to limit the number of unspecifically re-ligating vectors.

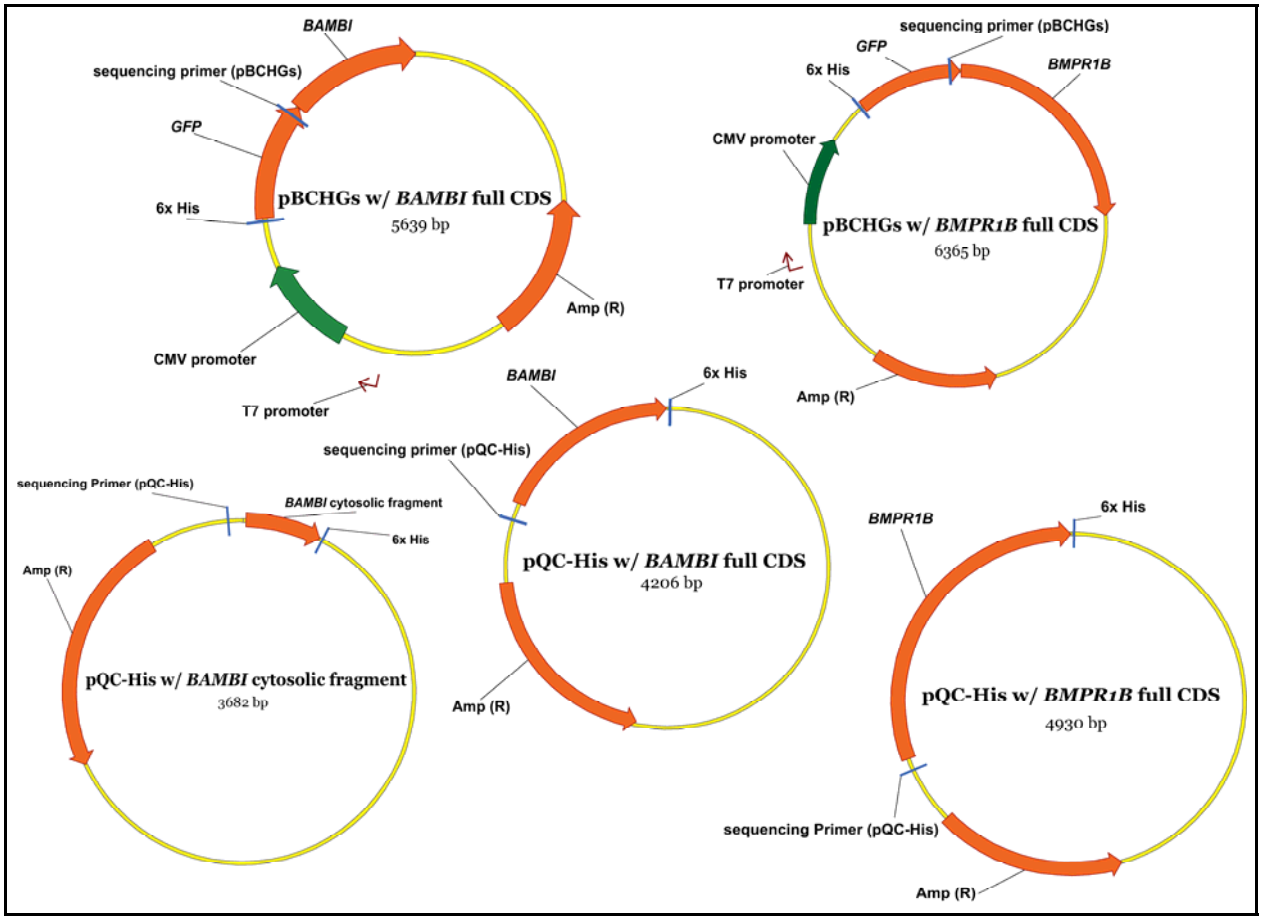
Ligations of plasmids with insert constructs were performed using T4 DNA ligase (Roche) according to the manufacturer's protocol, but with 4 U of enzyme in 40 µl reaction volume. Input DNA used were 60 ng for pQC-His and 100 ng for pBCHGs, and exactly three times more molecules, as calculated by the number of basepairs, for each PCR product (*BAMBI* cytosolic fragment, 15 ng / --; *BAMBI* full length CDS, 42 ng / 50 ng; *BMPR1B*, 80 ng / 94 ng; for ligations with pQC-His / pBCHGs, respectively). Ligation reactions were performed for 10 h at 12 °C followed by 4 h at 8 °C. The vector charts for the ligated products are given in Figure 9.

After ligation, 20 µl of the reaction volume were directly used for transformation of *E.coli* XL-10 gold strains and seeded on LB-Amp agar plates to select for re-ligated colonies, as described above. Of each of the ligation constructs, clones were picked to conduct a colony PCR as well as grow a mini culture (10 ml LB-Amp medium). The remainders of the ligation reactions were stored at 4 °C for a few days or at -20°C for long term, while the LB-Amp plates were stored at 4 °C.

If the colony PCR proved the plasmids positive for inserts, 2 ml of the corresponding mini cultures were harvested and the remainder was used to seed maxi cultures (400 ml). The harvested bacteria were used to isolate the DNA, again using Plasmid Mini kits (Qiagen) and the DNA was sequenced to check for mutations. After confirming the correct sequence, the plasmid DNA

was harvested from the maxi cultures with Plasmid Maxi kits (Qiagen), using the manufacturer's recommendations but with 30 ml of each buffer P1 - P3 and adjusting the protocol accordingly.

Figure 9



Vector charts. Eukaryotic pBCHGs and prokaryotic pQC-His vectors were subcloned to contain the respective coding sequences (CDS) of BAMBI cytosolic fragment, BAMBI full length CDS or BMPR1B full length CDS. 6x His, tag consisting of 6 subsequent histidine amino acids for purification purposes; Amp (R), ampicillin resistance gene β-lactamase. Charts were generated using Vector NTI software (Invitrogen).

3.5.2. Protein Expression in the Prokaryotic System and Isolation

For the expression of proteins in *E.coli*, BL21 CodonPlus (DE3)-RIL Competent Cells (Stratagene) were transformed with pQC-His construct plasmids bearing coding sequences of BAMBI cytosolic fragment, BAMBI full length protein or BMPR1B full length protein. These cells had been engineered to allow for a higher protein yield than normal XL-Blue or XL-Gold strains, and easy induction of the T7 RNA polymerase-driven expression by the manufacturer. Transformation was carried out with 50 µl of cells and 100 ng of plasmid DNA

each. Cells were kept 15' on ice, then heat-shocked for 3' at 37 °C and cooled again 3' on ice before adding 400 µl of LB medium and incubating at 37 °C for 45' in a shaker at 600 rpm. Cells were then plated on LB-Amp agar, incubated overnight, and clones were picked for each construct and transferred to mini-cultures (5 ml) as described above.

E.coli mini cultures were diluted in a rich medium (TB-Amp) to a final volume of 30 ml and grown under periodical surveillance of growth at OD_{600} on the spectrometer until an optical density of 0.6 was reached. Then, isopropyl *β*-D-1-thiogalactopyranoside (IPTG, dioxane-free, Fermentas) was added to a final concentration of 1 mM for induction of desired gene expression. Cells were continuously grown at 37 °C and 200 rpm. After 6 h of incubation, cells were harvested by centrifugation at 2,000 rpm in a Heraeus Varifuge 3.0 and the medium supernatant was discarded.

Depending on the purification of protein with native or denaturing protocol, cells were resuspended in 10 ml of the appropriate lysis buffer, as recommended by the Ni-Agarose manufacturer, sonified three times for 30' with intermediate cooling on ice, and the lysed cells were incubated overnight at 4 °C.

After pelletting the cell debris for 90' at 4 °C and 4,300 rpm (4,000 x g) in the Heraeus Varifuge, the supernatant containing the protein was decanted into a new tube. 5 ml of the solution were mixed with 4 ml of Ni-Agarose slurry (Qiagen) and the protein purification protocol was continued according to the manufacturer's protocol, but with 3 x 2 ml aliquots per washing step. Cleared protein fractions were stored at 4 °C.

To check for the protein content, 20 µl of the mini culture, lysis supernatant as well as each of the fractions were mixed with 10 µl Laemmli gel loading buffer, incubated for 15' at 96 °C and loaded onto a 15% polyacrylamide gel (containing SDS, Bio-Rad). Size marker used for direct staining was Unstained Protein Molecular Weight Marker and for subsequent blotting PageRuler Prestained Protein Ladder (both Fermentas). Electrophoresis was performed for approximately 2 h with 20W constant electrical current per mini gel, until the front of the loading buffer reached the end of the gel.

For immediate results, gels were stained with Coomassie Blue for at least 60' or overnight, destained with 20% isopropanol / 7% acetic acid (v/v) until the staining of background had diminished, and fixed in 7% acetic acid (v/v) for at least 30' or overnight. For long term storage, stained gels were spread on Whatman paper and dried under vacuum at 80 °C for 105'.

3.5.3. Immunization of Mice and Generation of Hybridoma Cells

The generation of antibodies from mice hybridoma cells was carried out in cooperation with the laboratory of Prof. Dr. Hanswalter Zentgraf, DKFZ Heidelberg.

In brief, mice of strain BALB/c, 8-12 weeks old, were injected subcutaneously with 20 µg of soluble protein as immunogen. The preparation of the immunogen included a nonspecific immunogenic stimulator (Freund's Adjuvant) containing mineral oils. This procedure is administrated for the primary immunization to enhance the immune response of the animals and protect the immunogen from rapid catabolism. Primary immunization of the mice was performed with a conjugate made of BAMBI cytosolic fragment with keyhole limpet hemocyanin (KLH) by thiol-coupling according to Sawin and co-workers.[140] Second and third immunizations, in two weeks intervals, were performed with 20 µg of the antigen alone to boost the specific immune response against the BAMBI cytosolic fragment. Three days later, spleen cells from the immunized mice were fused with cells of the myeloma line P3x63Ag8.65 3 using polyethylene glycol as described.[141] Cell culture supernatants were screened for antibodies by ELISA and immunoblotting. Positive cell lines were subcloned by limited dilution.

3.5.4. Validation by Western Blotting

For testing of antibodies, prokaryotic expression of proteins and PAGE were performed as described above, but with BAMBI full length protein as well as BMPR1B protein, on 15% polyacrylamide gels.

Gels were blotted using polyvinylidene fluoride (PVDF) membrane (Millipore) in a standard mini-gel tank-blotting apparatus (Bio-Rad) according to manufacturer's instructions. The buffer used for transfer was Tris/Glycine based but without SDS and contained 20% methanol (v/v). Blots were then

washed, blocked and incubated with antibodies. Transfer and washing buffers were prepared and used according to protocols given in the "QIAexpress Detection and Assay Handbook for Anti·His Antibodies" (Qiagen). The only exception to these recommendations was the blocking mix, for which 5% milk powder (w/v) and 3% BSA (w/v) in TBS-Tween buffer containing 0.05% Tween-20 (v/v) were used.

Hybridoma cell supernatants were used undiluted to test antibodies. For positive controls, anti-P53 antibody was used on the blots against P53 protein (both kindly provided by Hanswalter Zentgraf), which was additionally loaded onto the acrylamide gels. As negative controls, antibodies against either P53 or BRWD3 protein (the latter provided by Magdalena Schlotter) were used against whole cell lysate. All controls were performed at concentrations recommended by the provider.

3.6. Pathway Analysis

For the identification of cellular signaling pathways involved in the course of the disease, as characterized by the classification between responders and non-responders with the gene expression signature, the contained genes were analyzed using designated software and tools.

The most comprehensive and best known database is Gene Ontology (GO). Since GO and its use are public, other tools use it to integrate the information for providing statistical analyses (AmiGO, FatiGO)[142,143] or graphical illustrations of the interplay (KEGG)[144] between proteins or lists of genes and proteins. To identify pathways deregulated or altered between the responder and non-responder groups of breast cancer patients, the FatiGO and KEGG analysis tools and databases were used, as well as the raw information published and stored in the GO database and information collected in the Gene and Pubmed databases of the National Center for Biotechnology Information (NCBI), which belongs to the United States National Institutes of Health (NIH).

3.7. Immuno-Histochemistry

To validate results from the gene expression data acquired by microarray and RQ-PCR measurements, the translation of the deregulated genes into proteins was assessed by immuno-histochemical measurement in sections of breast tumor samples from the same patients. These formalin-fixed and paraffin-embedded tissue samples as well as the sections thereof were provided by Prof. Hans-Peter Sinn from the Department of Pathology at the Hospital for Gynecology and Obstetrics of the University of Heidelberg. In total, 80 patients from both GEDoc and GEsDoc studies were available for immuno-histochemical experiments, with four consecutive sections per patient, cut at 5 µm thickness.

Deparaffination and Antigen Retrieval

Embedded tissue sections were deparaffinated and antigen retrieval was carried out by incubation of the pre-processed tissue section slide by boiling for 25' in either citrate buffer (10 mM citric acid, 0.05% Tween-20, pH 6.0) or EDTA buffer (10 mM Tris, 1 mM EDTA, 0.05% Tween-20, pH 9.0) and left to cool down at room temperature for another 25' (Table 14).[145-147] For each antibody, the optimal retrieval method was tested on excess sections and the optimal protocol was then applied to the sections from tumor specimen of the breast cancer patients from the GEDoc and GEsDoc study cohorts.

Table 14 **Preprocessing of Paraffin Sections**

Incubate in xylol, 3x 5'
Incubate in 100% EtOH, 2x 5'
Incubate in 95% EtOH, 2x 5'
Incubate in 80% EtOH, 2x 2'
Incubate in aqua dest., 1'
Incubate in Tris-EDTA (pH 9.0) / citrate (pH 6.0) buffer, 25' @ 97 °C
25' cool down at room temperature

Immuno-Staining Reactions

Antibodies against BAMBI, BMP4, BRCA1, LMO4, SMAD3 and SRC proteins were purchased and used as given in Table 15. Washing of the sections, incubation with the antibody dilutions and staining with chromogens was carried out on a TechMate Horizon (Dako) using protocol MSIP and the

solutions provided, as recommended by the manufacturers. Double-stains (BAMBI/SRC and BMP4/SMAD3) were performed sequentially using NovaRed chromogen first and SG chromogen (both Vector) second, and in accordance with deparaffination method and subcellular location compliance of the two antigens. Counterstaining with hematoxylin & eosin solution was performed during the first staining run and substituted with Washing Buffer 4 during the second staining run.

Table 15 **Antibodies and Dilutions**

Target Protein	Antibody No.	Manufacturer	Clonality	Origin Species	Dilution	Chromogen
BAMBI [a]	H00025805-M01	Abnova	monoclonal	Mouse	1:150	NovaRed (red, Vector)
BMP4 [b]	NCL-BMP4	NovoCastra	monoclonal	Mouse	1:50	NovaRed (red, Vector)
BRCA1	ab16780	Abcam	monoclonal	Mouse	1:100	NovaRed (red, Vector)
LMO4	sc-11120	Santa Cruz	polyclonal	Goat	1:100	DAB (red, Dako)
SMAD3 [b]	ab28379	Abcam	polyclonal	Rabbit	1:100	SG (grey, Vector)
SRC [a]	ab32102	Abcam	monoclonal	Rabbit	1:150	SG (grey, Vector)

[a,b] Double stains performed on the same sections.

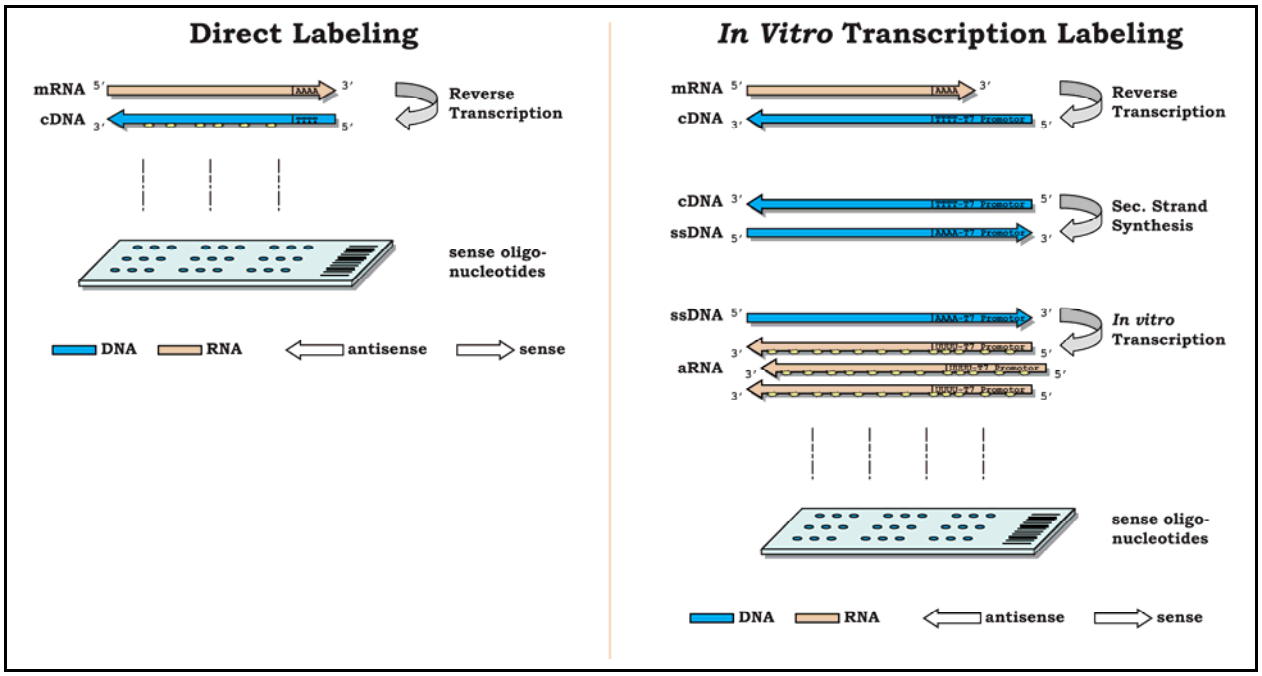
4. Results

In the presented study, the discovery of a gene expression signature based on RNA samples from small tissue biopsies, obtained from primary tumors of the breast, was elucidated. As the RNA yield was very small, the appropriate method for amplification of mRNA to be used with long gene-specific oligonucleotide microarrays had to be developed. Then, the amplification procedure was applied to a large set of biopsies from breast cancer patients, and whole-genome gene expression experiments were performed to identify a gene signature predicting response of the patients to chemotherapy. The prediction performance of the obtained gene expression signature was then tested for significance in an independent set of patients, who received chemotherapy with the same drugs. Genes contained in the predictive gene signature demonstrating biological relevance of the corresponding pathways were selected to confirm the microarray expression results using RQ-PCR and to further investigate these pathways by immuno-histochemical staining of tumor biopsy sections from the same patients.

4.1. Messenger RNA Amplification Protocols

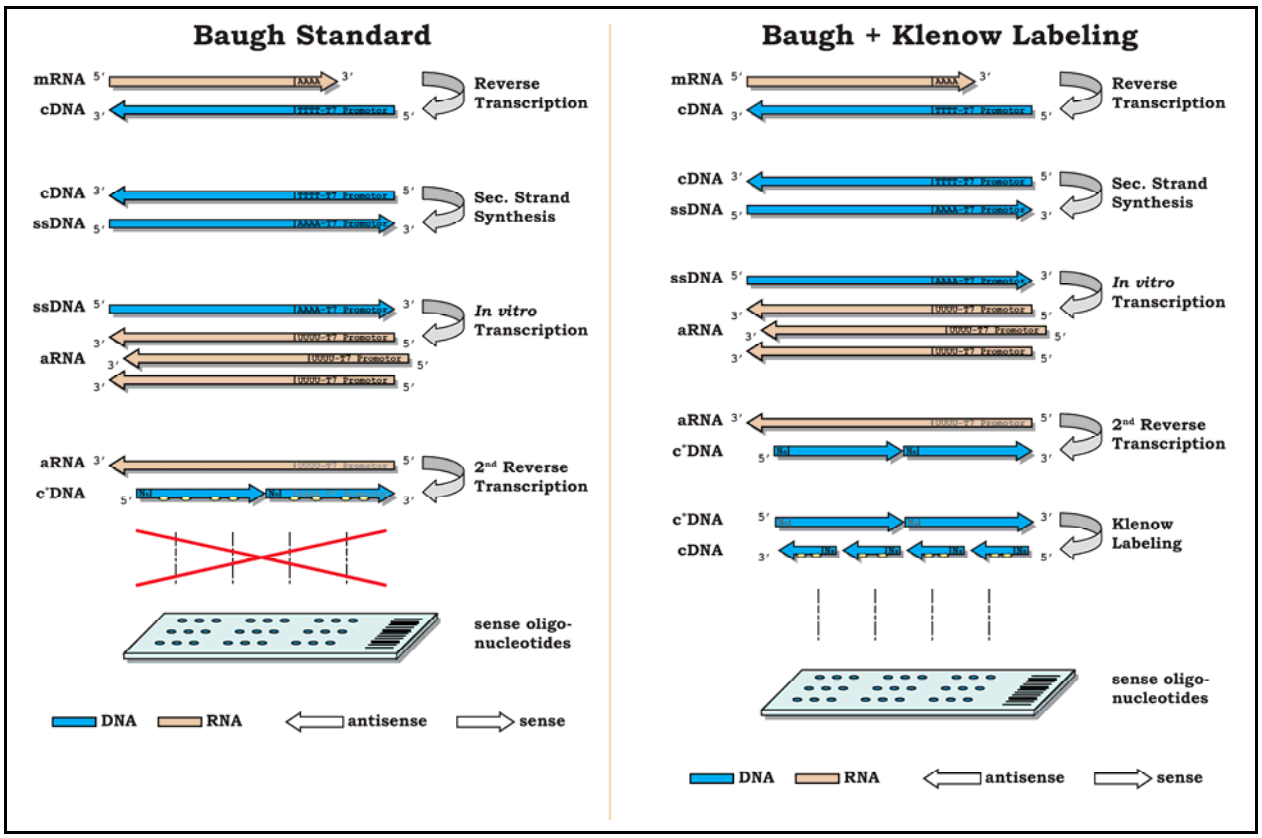
To overcome restrictions imposed by the strand incompatibility of the sense-orientated oligonucleotide DNA probes of the microarray with established mRNA amplification protocols yielding sense orientated labeled DNA samples, six different procedures and additional variations thereof to amplify nucleic acids were developed or introduced and tested in collaboration with Dr. Jörg Schlingemann. In order to have a direct comparison and minimize bias, the protocol steps that were shared between the individual methods were performed correspondingly, e.g. using the same enzymes and concentrations of reagents. Then, various measurements were made to estimate the performance of these methods like fluorescent dye incorporation, microarray signal intensity, reproducibility, and others. The source material used for testing was generated from two different cell lines, HL-60 and NU-DHL-1, which both are from myeloid origin but represent different genetic alterations (see Appendix A). For a short summary of the different amplification methods, see Figures 10-12.

Figure 10



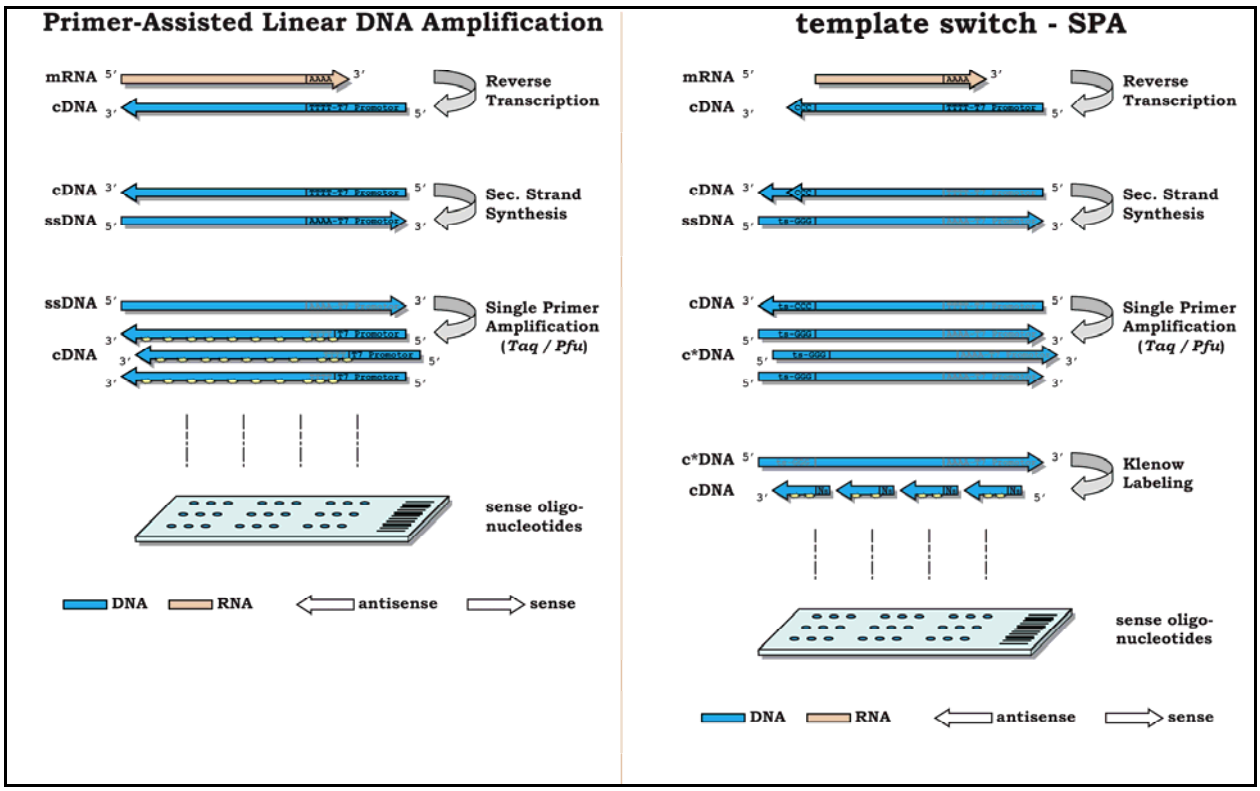
Schematic view of Direct Labeling and *In Vitro* Labeling protocols. ssDNA, second strand DNA, complementary strand of cDNA; aRNA, antisense RNA. Sense and antisense refer to the orientation of mRNA.

Figure 11



Schematic view of Baugh Standard and Baugh + Klenow protocols. ssDNA, second strand DNA, complementary strand of cDNA; aRNA, antisense RNA; c*DNA complimentary sense DNA. Sense and antisense refer to the orientation of mRNA.

Figure 12



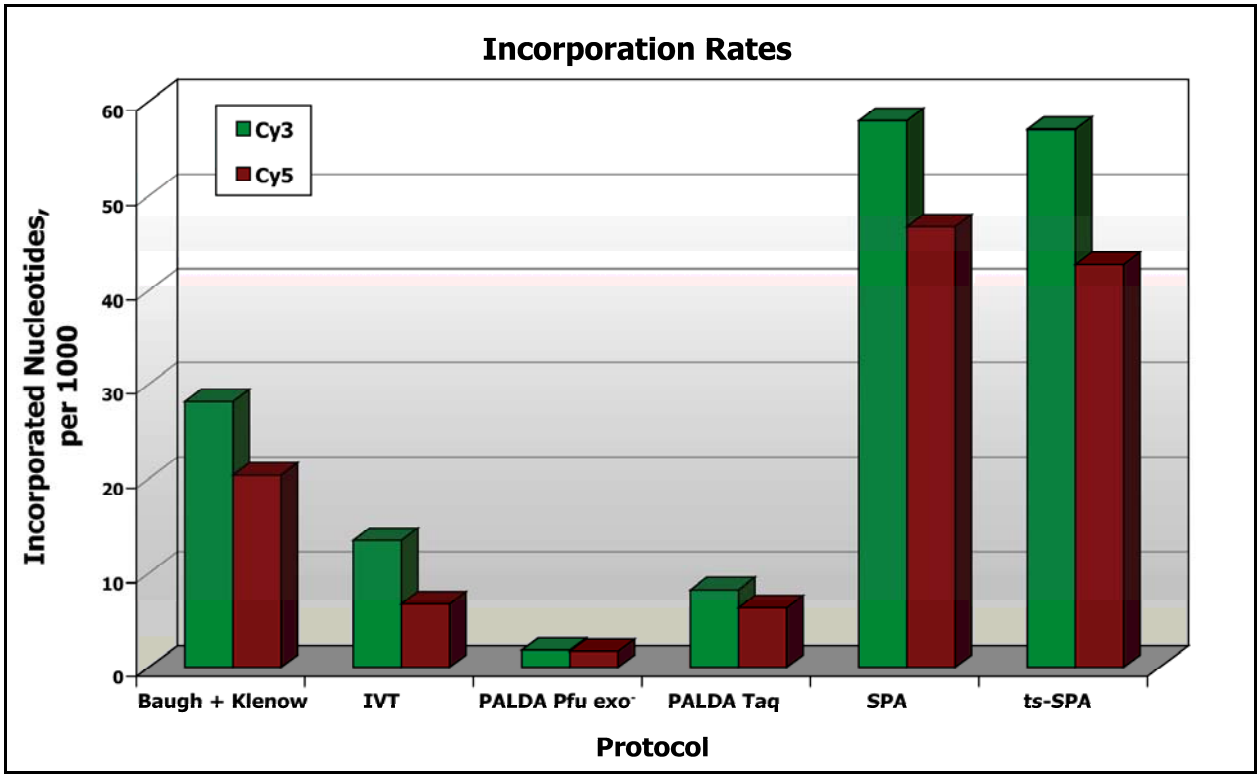
Schematic view of PALDA and ts-SPA protocols. ssDNA, second strand DNA, complementary strand of cDNA; c*DNA complimentary sense DNA. Sense and antisense refer to the orientation of mRNA.

4.1.1 Incorporation of Fluorescently Labeled Dyes

The first step in evaluating the performance of the different protocols for amplifying nucleic acid material was comparing the rate of incorporated fluorescent dyes per 1000 nucleotides. Since the Cy dyes are covalently bound to the nucleotides, resulting in a much higher molecular weight and surface area, this rate of incorporation depends on the different enzymes that are used to polymerize the nucleic acids. Some of the enzymes had been modified accordingly by their manufacturers, e.g. by reducing the proof-reading capacity or modifying the size of the grooves for nucleotide entry and polynucleotide exit. For unmodified enzymes, the tested protocols had been adapted to increase the ratio of labeled to normal nucleotides, thereby increasing the probability to incorporate the labeled ones (PALDA *Taq* / *Pfu* and IVT Labeling). The Primer-Assisted Linear DNA Amplification (PALDA) protocols did not incorporate these bulky nucleotides well (Figure 13), although the used *Pfu* DNA polymerase with the lowest incorporation rate (below 2 per 1000 nucleotides) has a decreased exonucleoase activity (exo$^-$). PALDA using the *Taq*

DNA polymerase as well as *In Vitro* Transcription (IVT) labeling performed significantly better (7.3 and 10.1 per 1000 nucleotides, respectively), but their incorporation rates were considerably low when compared to the other protocols. Using the Klenow fragment of DNA polymerase I, as in Baugh + Klenow and the two Single Primer Amplifications (SPA) protocols, performed substantially better (24.3, 49.9 and 52.3 per 1000 nucleotides for Baugh + Klenow, ts-SPA and SPA, respectively).

Figure 13

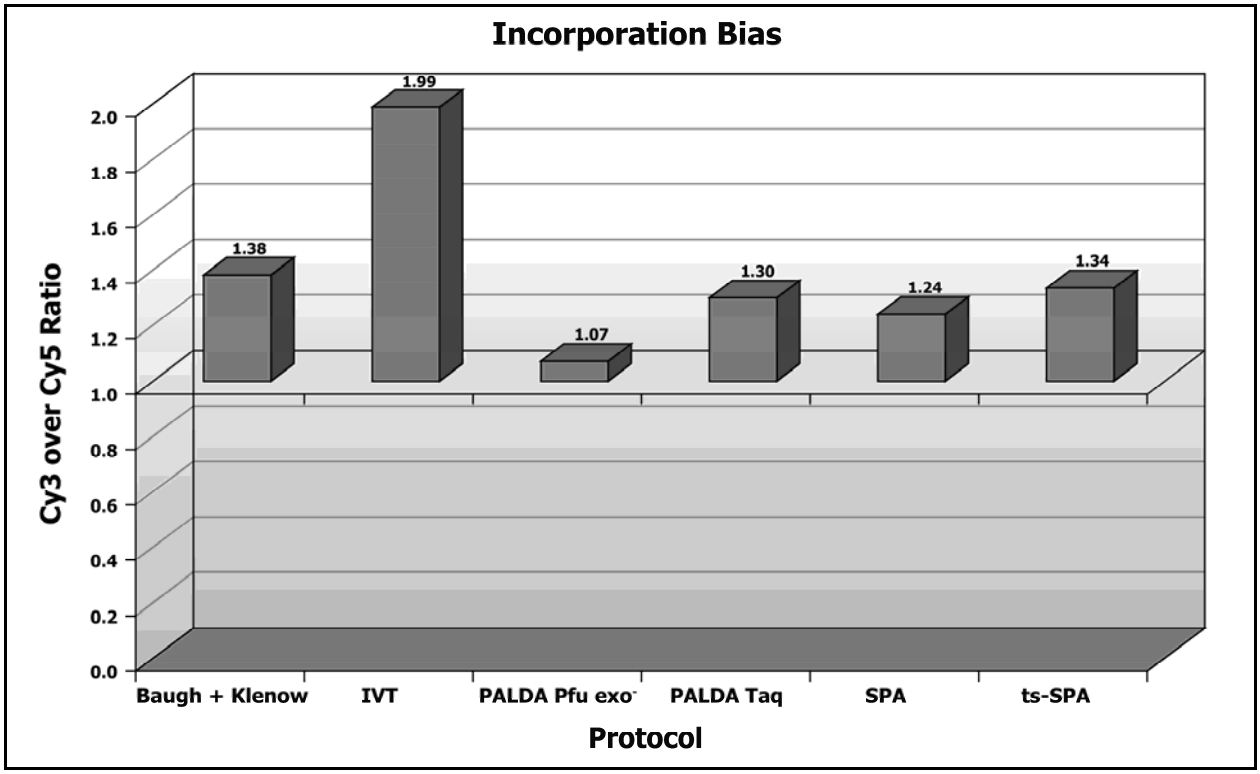


Incorporation rates for fluorescently labeled nucleotides. Bars represent labeled nucleotides per 1000 nucleotides as a result of the different protocols for amplification, estimated by photometric measurements after amplification.

Another very important aspect in evaluating the incorporation rates of the labeled nucleotides is the ratio between the two different fluorescent dyes. As these have different molecular weights and surface areas, a difference in their incorporation into the nucleic acids was expected, especially for the incorporating enzymes with low processivity for these bulky nucleotides. However, the only protocol showing a dramatic bias towards one of the dyes was the IVT Labeling protocol (1.99 for Cy3 over Cy5, Figure 13 & 14). The RNA polymerase II, which was used in this protocol with the labeled nucleotides, showed a strong discrimination in its processivity between the dyes, as the Cy5 dye has the bulkier fluorescent molecule group and therefore requires more

space to be integrated. The enzyme with the lowest dye bias was the *Pfu* exo$^-$ polymerase, although it had a very limited total incorporation. The protocols using the Klenow fragment, as well as the PALDA *Taq* method, had a moderate bias for the benefit of Cy3 dye (1.24 to 1.38).

Figure 14



Biased incorporation of fluorescently labeled nucleotides between dye channels. For different amplification protocols, the ratio of Cy3 over Cy5 incorporation rates after amplification was calculated.

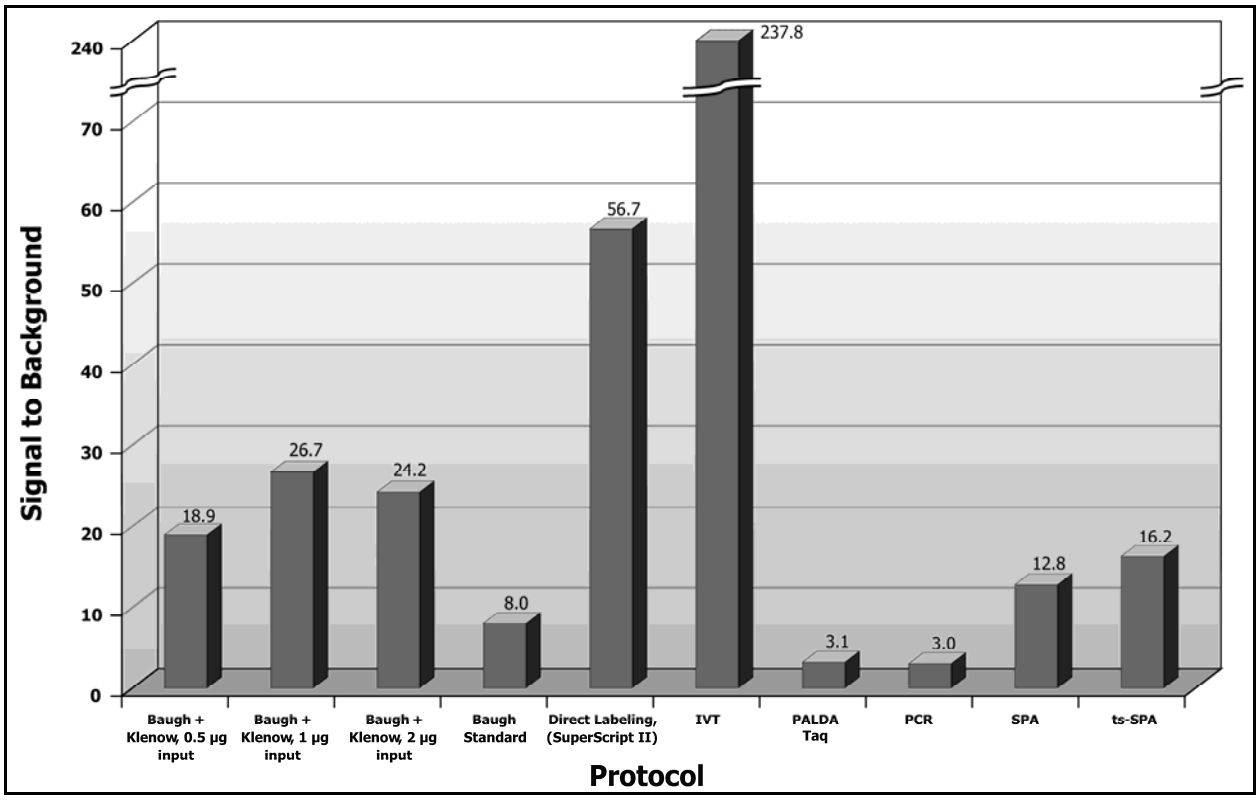
4.1.2 Performance of Amplified Messenger RNA on Oligonucleotide Microarrays

After evaluation of the dye incorporation, the second but even more critical aspect of the amplification protocols was to investigate their performance on the oligonucleotide microarrays. In order to obtain valuable data, the experiments were performed at least in duplicates. For comparison, directly labeled cDNA as well as DNA amplified and labeled with the Klenow standard protocol and by PCR were also included. Since the PALDA protocol with *Pfu* exo$^-$ enzyme did not yield sufficient fluorescent labeling, this protocol was not pursued anymore.

Signal Intensity

Beyond the general incorporation as estimated above, the amount of fluorescently labeled nucleic acid actually available for hybridization to the DNA probes on the microarray was analyzed. To obtain this measure, the total intensity of each feature, representing hybridized sample molecules, was calculated *versus* its local background of the surrounding area and averaged for all valid features. These ratios were compared for the different protocols (Figure 15).

Figure 15



Signal to background ratio. Intensity values from feature foreground (signal) to local area surrounding feature (background) were averaged for both dye channels of microarray experiments, and the ratio was calculated.

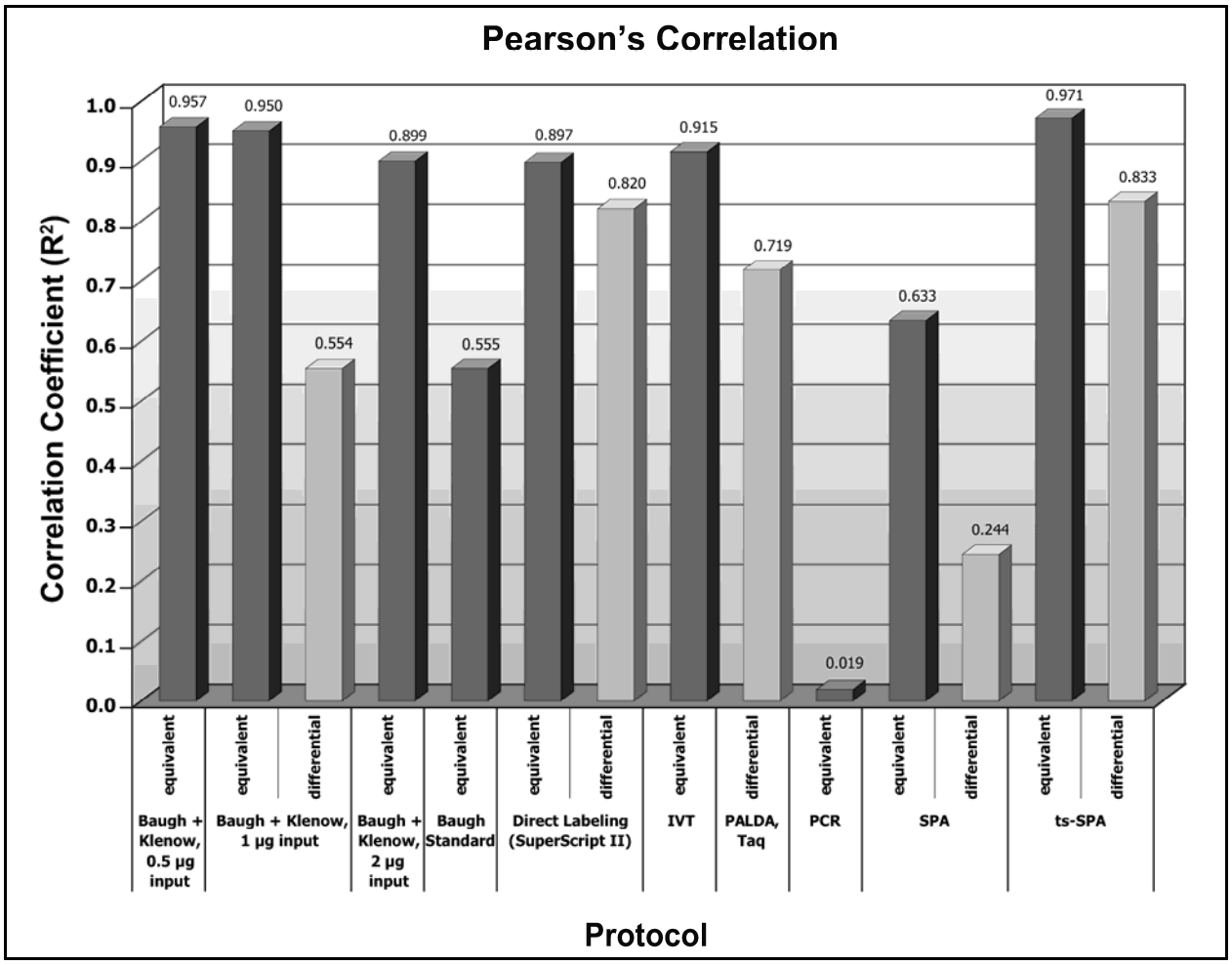
Compared to the direct labeling procedure used as benchmark (signal to background ratio of 56.7), only one method showed superior signal to background ratio, the IVT Labeling (237.8). Next closest to the direct labeling, but lower, is the Baugh + Klenow method. Even though the input amount of total RNA was modified (0.5, 1.0 and 2.0 µg), the protocol yielded higher signal to background ratios than the next best method (18.9, 26.7 and 24.2, respectively). Both SPA and template-switch SPA yielded relatively low ratios when compared with the direct labeling (12.8 and 16.2, respectively), while the

other protocols (Baugh Standard, 8.0; PALDA *Taq*, 3.1; PCR, 3.0) did not yield sufficient signal to background ratio.

Same-versus-same (Equivalent) and Differential Expression Correlation

To analyze whether the signals received from the microarray hybridizations were gene-specific and reproducible, expression ratios from hybridizations of same-*versus*-same (referred to as equivalent) experiments as well as from experiments with the two different cell lines (differential) were compared. Unfortunately, the PALDA *Taq* protocol did not yield enough labeled product to perform same-*versus*-same hybridizations, so only differential experiments could be carried out. The Pearson's Correlation for all valid features was calculated both in equivalent and differential experiments, to evaluate reproducibility or difference (Figure 16). While most protocols show good reproducibility, as shown by high values for the equivalent comparisons, the Baugh Standard, SPA and PCR methods have low or no correlation. The PALDA *Taq* protocol could not be evaluated in this aspect. In the differential experiments, a medium (0.5) to slightly elevated (0.65) correlation was expected, as the two cell lines have some, but not high similarities. However, the Direct Labeling, PALDA *Taq* and ts-SPA methods have a higher than expected correlation, while the SPA protocol shows very low correlation.

Figure 16

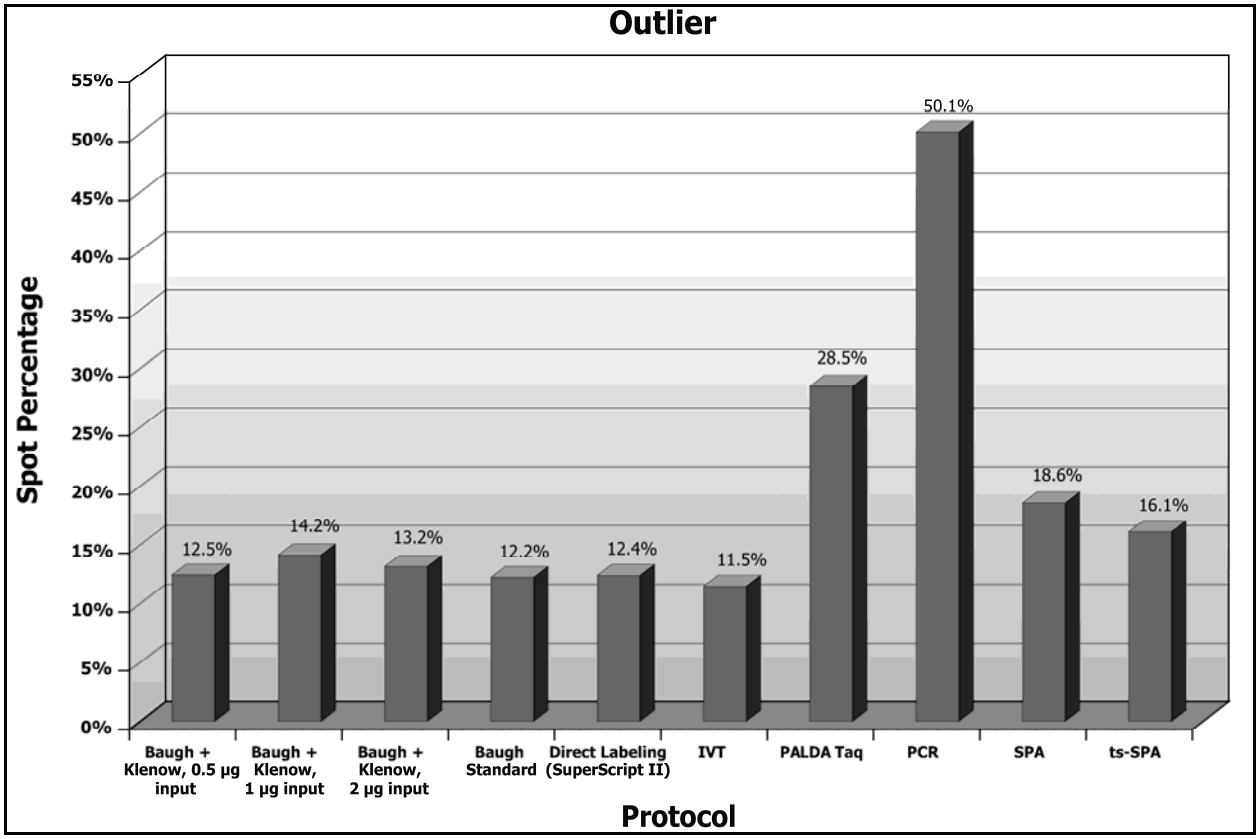


Squared Pearson's correlation coefficient. Repeat hybridizations in same-*versus*-same (equivalent) or differential hybridizations between cell lines HL-60 or NU-DHL-1 were compared across all valid features of the respective microarrays.

Outlier Features

The GenePix software detecting the hybridized microarray features, as guided by the manual inspection of the user, was used to identify features without hybridization as well as spots representing artificial or otherwise false positive signals. The proportion of these so-called outlier features could be determined for each protocol (Figure 17). A certain extent of missing features was expected, since not all genes are expressed in one or both of the cell lines. A very high percentage of outlier features was seen in amplifications with the PCR (50.1%) and the PALDA *Taq* (28.5%) methods, while the SPA and ts-SPA protocols showed an elevated percentage (18.6% and 16.1%, respectively), when compared to the remaining protocols (11.5% - 14.2%).

Figure 17

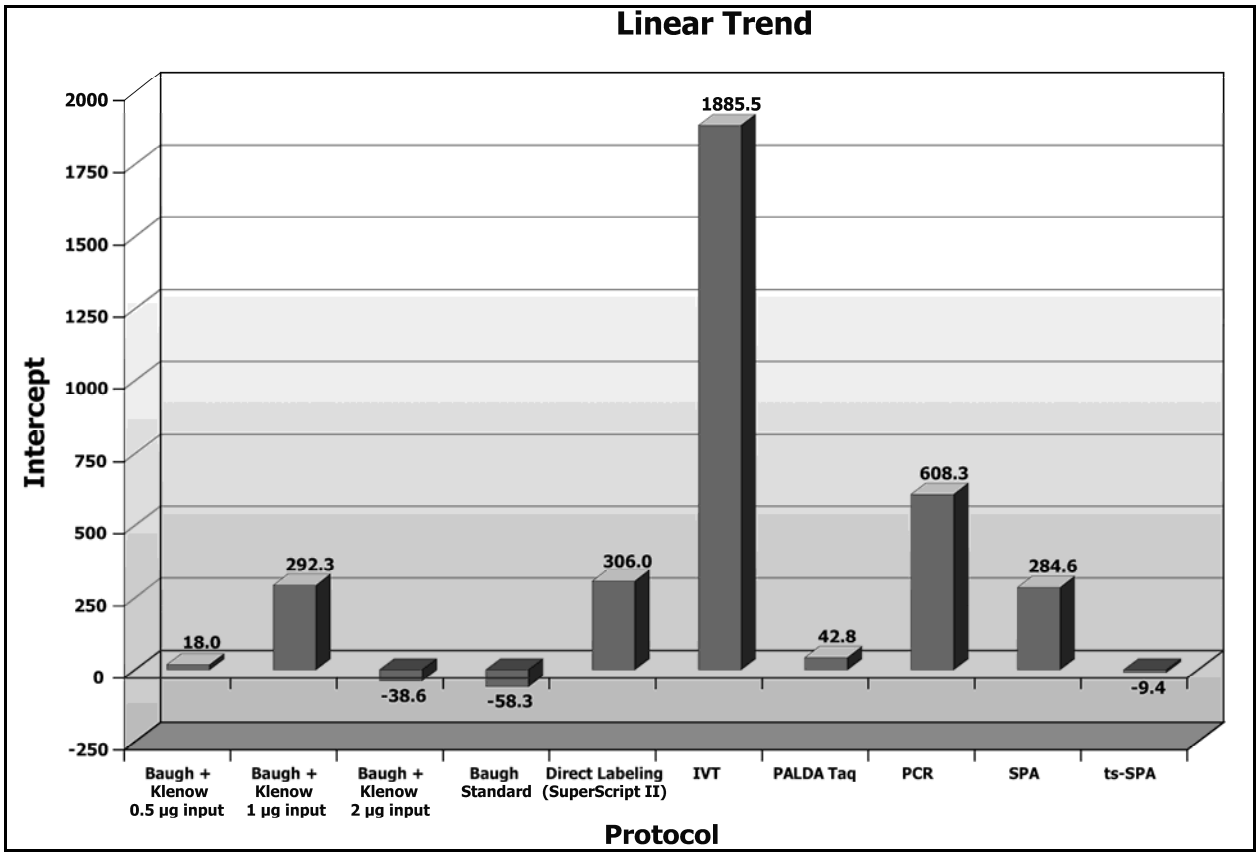


Percentage of non-valid features (outliers). Outlier features were identified by microarray scanning software or manually flagged by user inspection, and set in relation to total number of spots per microarray.

Linear Trend of Intensity Scatter Plots

By spreading the median intensities of all spots on the microarrays on scatter plots, either in same-*versus*-same (equivalent) or differential hybridizations, and calculating the linear trend of the resulting data points, the slope and intercept of the trend line could be used to describe key features of the amplification performance and hybridization to the probes on the microarrays. The intercept of this line expresses a bias of the corresponding dye intensity, as the scanner was set to a higher PMT voltage in the corresponding channel to compensate for the intensity loss. Such a bias is seen as a very high intercept of the trend in IVT labeling (1885.5), and as an elevated intercept in the PCR protocol (608.3).

Figure 18

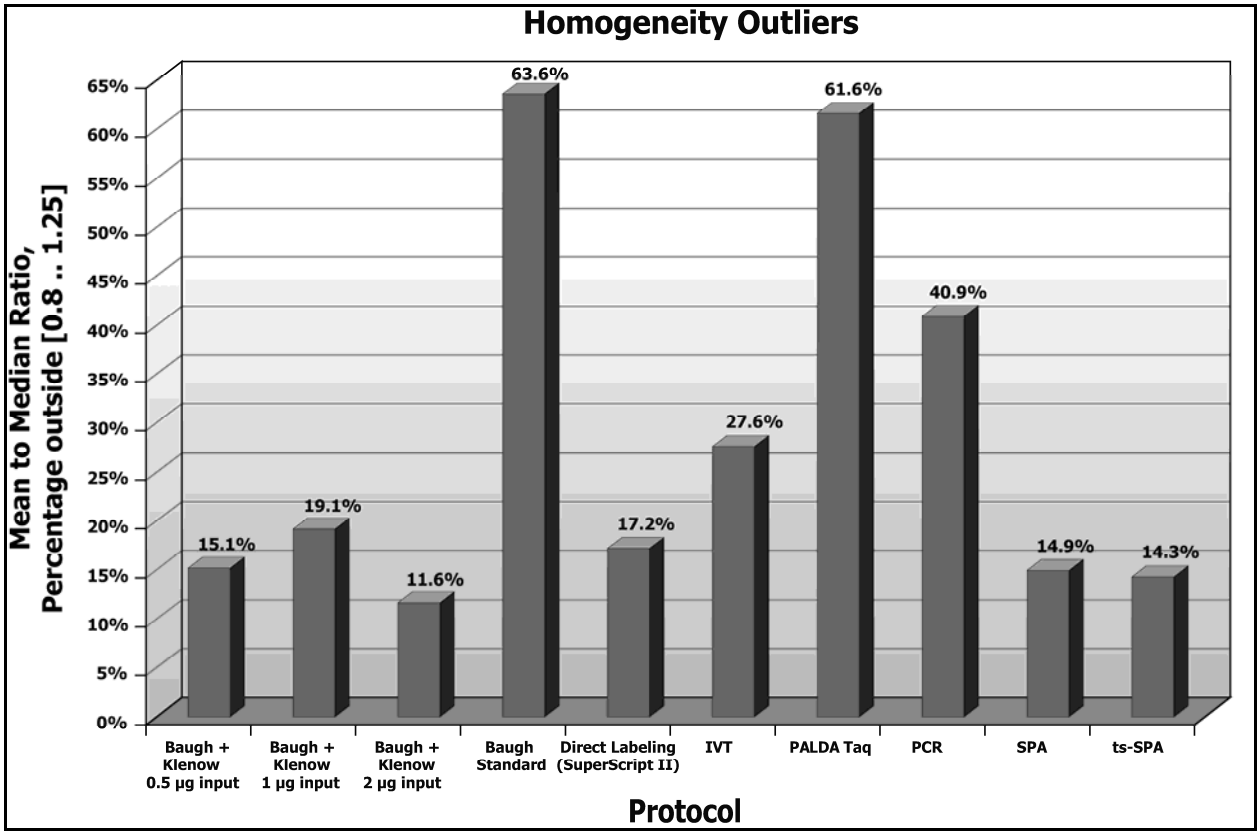


Ordinate intercept points of the linear trend. Linear models were fitted to the scatter plots of feature intensities between both dye channels, Cy3 and Cy5, and the ordinate intercept was calculated.

Spot Homogeneity

The homogeneity of the spots was another factor used for evaluation of the amplification protocols. Spots with inhomogeneous pixel intensities have to be dismissed, as they do not represent a substantial intensity value. Therefore, the percentage of homogeneity outlier features was used as a measure of intensity validity, accepting features for data analyses if the ratio of mean to median was between 0.8 and 1.25. The percentage of the features with a ratio outside this interval was estimated for each protocol (Figure 19). With a homogeneity outlier percentage above that of the PCR amplification (40.9%), the Baugh Standard (63.6%) and PALDA Taq protocols (61.6%) showed very high values. IVT labeling showed an elevated homogeneity outlier percentage of 27.6%, while the Baugh + Klenow protocols (11.6% - 19.1%) and the SPA methods (ts-SPA, 14.3%; SPA, 14.9%) were within acceptable range of the Direct Labeling protocol (17.2%).

Figure 19



Homogeneity outlier percentage. Valid features with a mean to median ratio of feature intensity pixels outside the acceptable interval of 0.8 to 1.25 were calculated in respect to total number of valid features.

Based on the results shown, it was evident that the protocol for amplification of messenger RNA to be used for small amounts such as those obtained from clinical biopsies to be used together with the oligonucleotide microarray technology was best met with the Baugh + Klenow method. In all aspects measured and displayed, it performed best or second to best and showed good reliability in laboratory practice. Since its denomination was derived from the original author and the enzyme nickname that was used for its additional extension, it was finally entitled "**T**7-based **A**mplification of **c**DNA and **K**lenow **L**abeling for **E**xpression Analysis", abbreviated *TAcKLE* Analysis.

4.2 Gene Expression Signature Predictive for Chemotherapy in Primary Breast Cancer

A total of 148 patients had been enrolled in two combined clinical phase II studies to determine the efficacy and dosage compatibility of a neoadjuvant chemotherapeutic regimen of **g**emcitabine, **e**pirubicin and (or **s**equentially followed by) **doc**etaxel, termed GEDoc and GEsDoc, respectively, for therapy of patients with primary breast cancer at the Clinic for Gynecology and Obstetrics of the University of Heidelberg. This protocol includes the addition of the drug gemcitabine, a nucleoside analog like fluorouracil, to the combination therapy of an anthracycline (epirubicine) and a taxane (docetaxel). The addition of gemcitabine promised an improvement of the response of the tumors to established chemotherapy regimens, resulting in an increase in both the number of patients with a complete response (no residual tumor) and the number of patients with a partial response (decreased tumor size).[47,104]

Table 16 **Patient Characteristics**

	GEsDoc		GEDoc	
	n	**%**	**n**	**%**
No. of patients	52	100	48	100
Age [median (range)], years	45 (29 to 65)	-	49 (30 to 65)	-
Tumor size by US [median (range)], cm	3.5 (2.1 to 8.0)	-	3.5 (2.1 to 10.0)	-
Histology, ductal / lobular / other	45 / 4 / 3	87 / 8 / 6	37 / 7 / 4	77 / 15 / 8
Histological grade, 1 / 2 / 3 / n.a.	2 / 22 / 25 / 3	4 / 42 / 48 / 6	2 / 19 / 23 / 4	4 / 40 / 48 / 8
Clinical nodal status, N0 / N+	31 / 21	60 / 40	20 / 28	42 / 58
Hormone receptor expression, ER or PGR score ≥ 1 / ER & PGR score 0	37 / 15	71 / 29	35 / 13	73 / 27
HER2 expression, 0 / 1+ / 2+ / 3+	40 / 2 / 1 / 9	77 / 4 / 2 / 17	33 / 3 / 1 / 11	69 / 6 / 2 / 23
KI67 expression, ≤ 50% / > 50% / n.a.	34 / 18 / 0	65 / 35 / 0	37 / 8 / 3	77 / 17 / 6
P53 expression, ≤ 20% / > 20%	39 / 13	75 / 25	38 / 10	79 / 21
BCL2 expression, 0 / 1+ / 2+ / 3+	21 / 10 / 12 / 9	40 / 19 / 23 / 17	29 / 4 / 10 / 5	60 / 8 / 21 / 10
Clinical response after 6 weeks of PST, CR / PR / NC / PD / n.a.	1 / 28 / 21 / 0 / 2	2 / 54 / 40 / 0 / 4	0 / 30 / 18 / 0 / 0	0 / 63 / 38 / 0 / 0
Pathologic response at surgery				
pT0 / pTis / pT1-4	12 / 3 / 37	23 / 6 / 71	5 / 4 / 39	10 / 8 / 81
pN0 / pN+ / n.a.	19 / 32 / 1	37 / 62 / 2	34 / 14 / 0	71 / 29 / 0
pCR breast (pT0 and pTis)	15	29	9	19
pCR breast+axilla [(pT0 or pTis) & pN0]	13	25	9	19

CR, complete remission; ER, estrogen receptor; n.a., not available; N, nodal status; NC, no change; PD, progressive disease; pN, pathologically determined nodal status; PGR, progesterone receptor; PR, partial remission; PST, primary systemic chemotherapy; pT, pathologically determined tumor status (is, tumor *in situ*); US, ultrasound

To supplement these studies, an accompanying trial was conducted at the German Cancer Research Center (Deutsches Krebsforschungszentrum) in

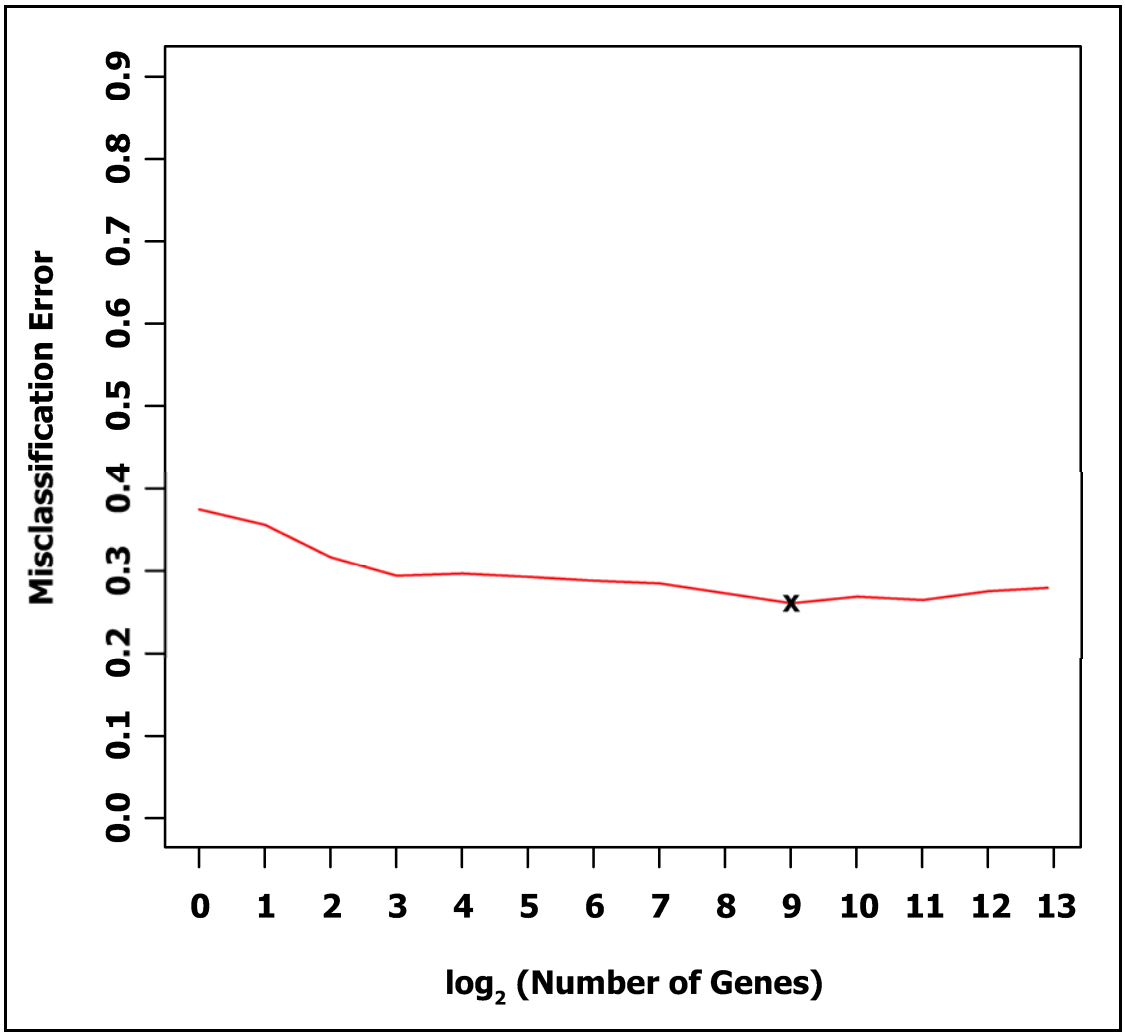
Heidelberg to establish and test a gene expression signature that allows the prediction of the response of patients to this chemotherapy.
To conduct the necessary experiments, core needle biopsies of the primary tumor were taken from the patients under sonographic surveillance at the clinic and one to two of these tumor specimens were subjected to microarray analysis in our laboratory as described. For the number of 100 patients, the mRNA could be successfully extracted, amplified and analyzed on the oligonucleotide microarrays. A detailed overview of the characteristics of this cohort of patients is given in Table 16.

4.2.1. Identificaton of the Gene Expression Signature
In order to obtain statistical significant genes predicting the response of the patients to the chemotherapy, two preconditions had to be set. First, the patients needed to be classified as responders or non-responders. According to the clinical behavior of the patients, only those who achieved a pathologically confirmed complete remission of their tumor, defined as the disappearance of all viable tumor cells in the specimen at surgery after chemotherapy, termed pathological complete remission (pCR), were classified as responders. All other patients were classified as non-responders. Secondly, to minimize overfitting bias of the algorithms identifying the classifier, the total number of 100 patients had to be divided into two cohorts of similar size. One of these sets was used as a training set to discover the gene signature, while the other completely independent set was exclusively used to test the gene signature and estimate its predictive power. As the patients had enrolled in two slightly different studies, of which almost identical numbers of patients were successfully analyzed, these were used as training and test sets. For the training set, the GEsDoc study patients were chosen, as they had the larger proportion of responders (29%) and thus the greater probability to establish a significantly predicting gene signature. The GEDoc study patients were then used to test the gene signature.

Of the initially 21,329 gene-specific oligonucleotides contained on the microarray, 15,355 were expressed and passed quality checks in at least 80% of the patients. Therefore, only these genes were considered for the

establishment of the gene signature. Applying the Support Vector Machines algorithm, the training set was used to discriminate the predictive power of these genes. Subsequently, the number of genes was halved stepwise by Recursive Feature Elimination, and each time the predictive power of the subset of genes, given as misclassification error, was estimated by cross-validation (Figure 20). On the basis of the minimal misclassification error, the selected gene signature contained the 512 (2^9) most predictive genes. The list of the genes contained in the signature is given in Appendix D.

Figure 20

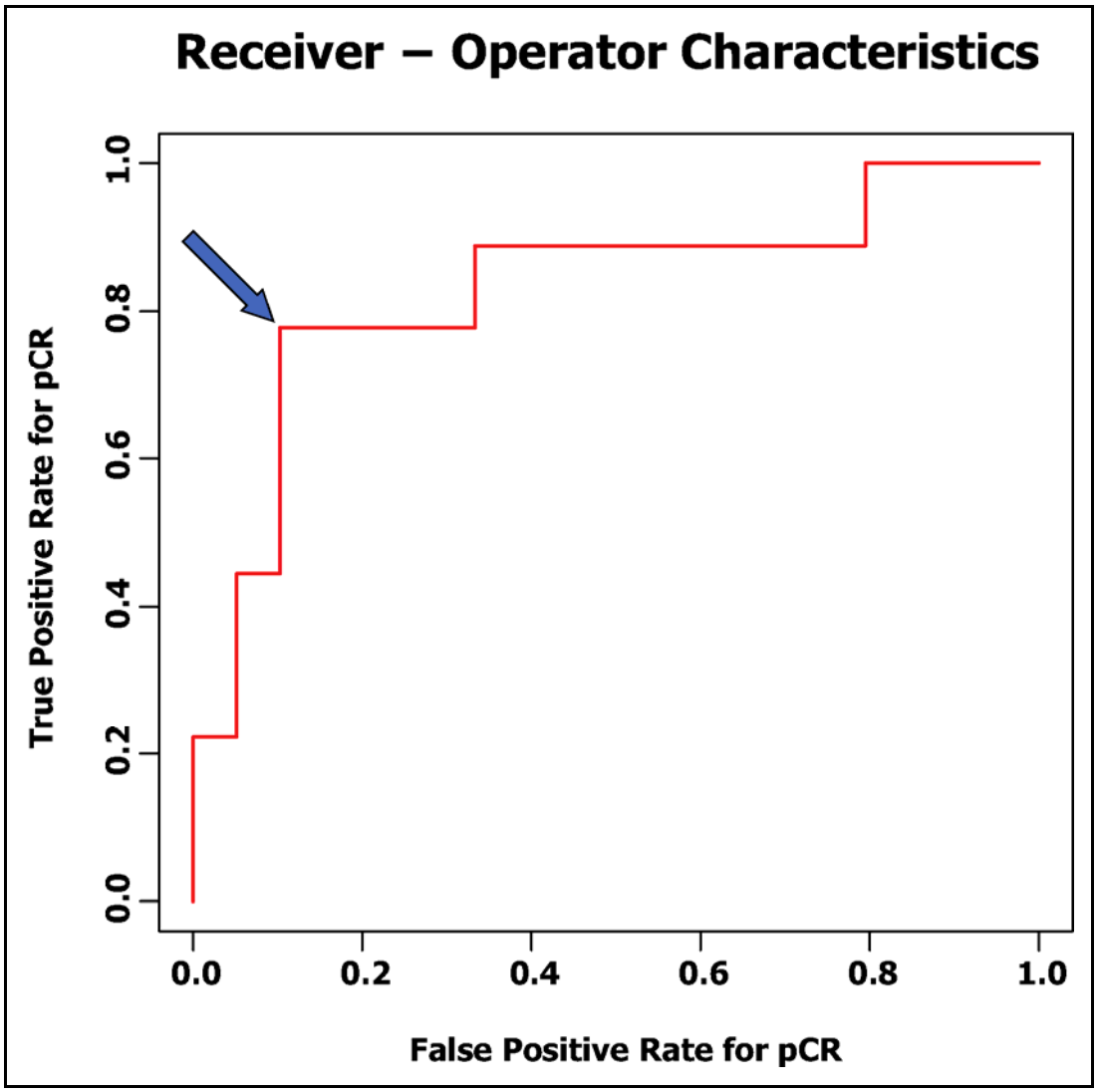


Misclassification error. The sum of false positives and negatives, estimated by cross-validation depending on the number of genes of the model within the training set of patients. In collaboration with P. Warnat.

Using the Receiver-Operator Characteristic graph, in combination with the Youden's Index (Sensitivity + Specificity - 1), the predictive characteristics of the chosen gene signature was calculated on the test set (Figure 21). For the optimal Youden's Index of 0.68, the parameters sensitivity (true positive rate,

estimated within observed pCR cases), specificity (true negative rate, estimated within observed non-pCR cases), positive predictive index (PPV, observed within estimated pCR cases), negative predictive index (NPV, observed within estimated non-pCR cases) and overall accuracy (PPV + NPV) were calculated (Table 17). With a total of seven correctly predicted pCR cases, sensitivity was 78% and the positive predictive value was 64%. Non-pCR patients were correctly classified by the predictor in 35 cases, yielding a specificity of 90% and a negative predictive value of 95%. In total, the accuracy was 88%, with 42 cases correctly classified. Due to the small number of pCR cases, the confidence interval of sensitvity and PPV are high.

Figure 21



Receiver-Operator Characteristics graph. Displayed are the dependencies of the selected classification model between true positive rate (sensitivity) and false positive rate (1 - specificity), as estimated by cross-validation of the training set of patients. The optimal balance between low false positive rate and high true positive rate is marked by the blue arrow. In collaboration with P. Warnat.

Table 17 **Patient Prediction Characteristics (test set)**

		predicted		
		pCR	**non-pCR**	**total**
observed	**pCR**	7	2	9
	non-pCR	4	35	39
	total	11	37	48

	cases	**percentage**	**95% C.I.**
sensitivity	7 of 9	78%	40% to 97%
specificity	35 of 39	90%	76% to 97%
PPV	7 of 11	64%	31% to 89%
NPV	35 of 37	95%	82% to 99%
accuracy	42 of 48	88%	75% to 95%

C.I., confidence interval; NPV, negative predictive value;
pCR, pathologic complete remission; PPV, positive predictive value

When comparing the predictive power of the gene signature with the best clinical factors in multivariate analysis, it shows superior predictive power, as calculated by the Odds Ratio (Table 18). The Odds Ratio describes the relative risks of patients in the respective classes for the predicted negative outcome of not reaching a pathological complete remission. Of the clinical factors, only HER2 (Score 0-2 *versus* Score 3) has a significant predictive power in the tested patient cohort. With p-values lower than 0.25 demonstrating factors to be not statistically relevant trends, as the low grading of the tumors for positive outcome as well as smaller tumors for negative outcome, the other factors, estrogen and progesterone negativity as well as the clinical tumor response after six weeks of therapy, are statistically irrelevant.

Table 18 **Penalized Logistic Regression of Signature and Clinical Factors**

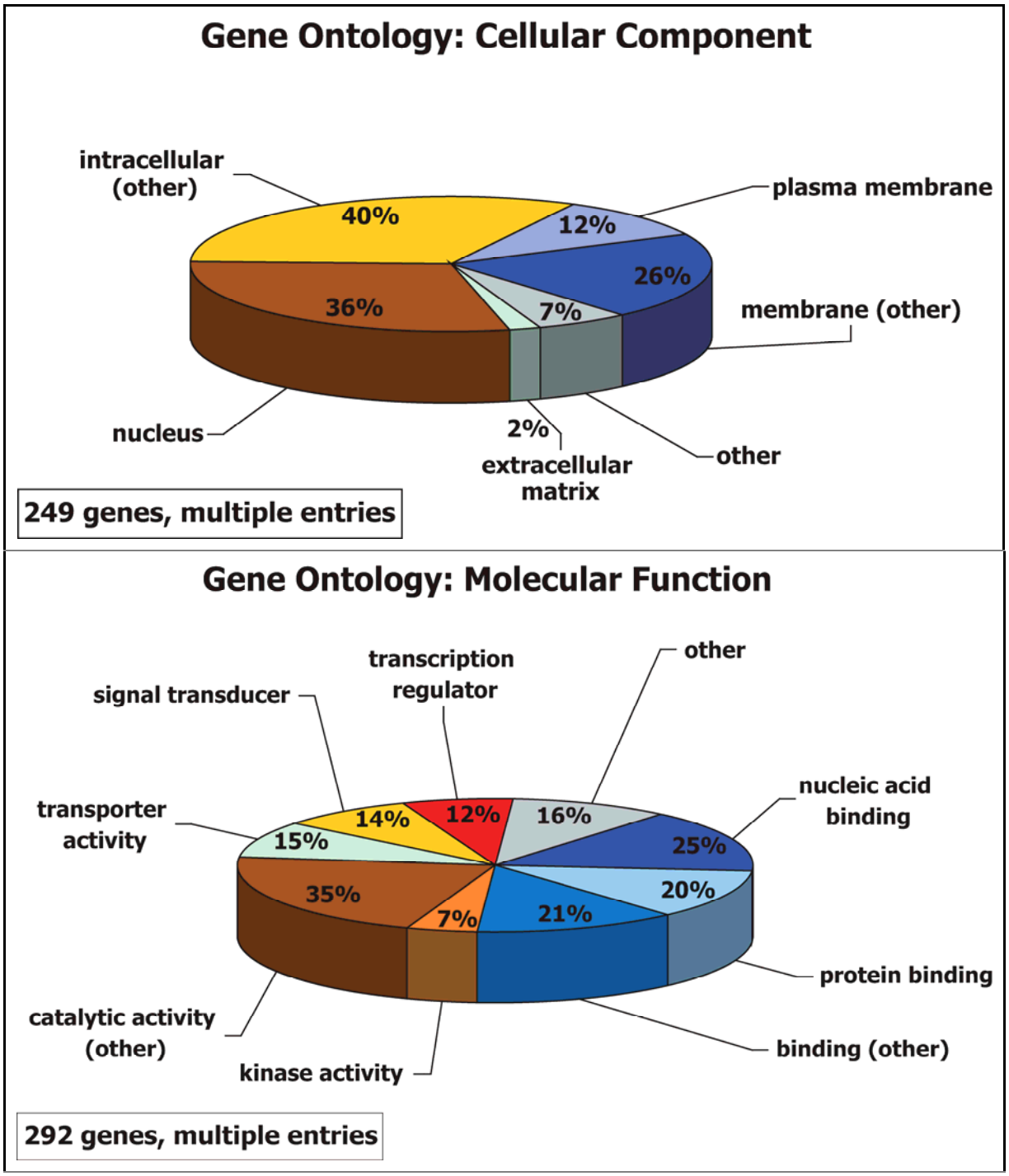
Factor	**Odds Ratio**	**95% C.I.**	***p***
Signature (Predicted negative *vs.* positive)	38.3	2.43 - 6560	0.01
Grading (G1, G2 *vs.* G3)	0.2	0.00 - 2.75	0.23
HER2 (0-2 *vs.* 3)	10.5	1.26 - 151	0.03
ER/PgR (ER 0 & PgR 0 *vs.* ER >0 or PgR > 0)	0.5	0.03 - 8.19	0.64
Response after 6 weeks PST (PR, CR vs. NC, PD)	0.6	0.02 - 10	0.70
cT max (cT 2-5 cm vs. cT >5 cm)	6.9	0.33 - 1128	0.22

C.I., confidence interval; CR, complete response; cT, clinical tumor size; ER, estrogen receptor;
NC, no change; PgR, progesterone receptor; PR, partial response; PST, primary systemic chemotherapy

4.2.2. Genes and Pathways of the Predictive Signature

In order to understand the biological implications of response to chemotherapy in breast cancer patients, it was of biological interest to investigate the genes contained in the gene expression signature in further detail. For this purpose, the Gene Ontology entries of the genes in the signature, which had an annotation, were studied in respect to cellular localization and molecular function (249 and 292 genes, respectively). As these numbers of genes were large, they were first depicted in their corresponding groups (Figure 22). Only a very small percentage of genes codes for proteins located in the extracellular matrix (2%), while the proportion of proteins in the nucleus is relatively large (36%), as the cellular components graph (upper panel) illustrates. The molecular function graphic (lower panel) shows a large proportion of genes coding for proteins involved in catalytic activity (35%) and nucleic acid binding (25%), binding of other molecules (21%) and proteins (20%) as well as signal transducing (14%) and transcriptional regulation (12%) activities.

Figure 22



Analysis of genes contained in the predictive expression signature. Genes are grouped according to their annotation in Gene Ontology for cellular localization and molecular function. Due to missing annotations, only 249 and 292 genes could be categorized, respectively. Pies amount to more than 100% as genes may have entries in multiple categories.

In a second step, the GO annotation terms of the genes from the signature were analyzed for statistically significant enrichment when compared with all genes represented on the microarray using Fisher test (Table 19). Genes significantly associated with the metabolic pathways directly targeted by the

chemotherapeutic agents gemcitabine and epirubicin were grouped as nucleotide metabolisms (yellow), and genes targeted by docetaxel belong to the functional groups of microtubular depolymerization and regulation of spindle apparatus during the mitotic phase (pink). Protein farnesylation/ prenylation, associated with Ras proteins, showed a very high significance (orange), as did the significant genes from the TGF-β pathway subfamily of bone remodeling proteins (blue) and DNA damage response genes (green).

Table 19 **Significantly enriched GO Terms**

GO ID	GO Term	p value
GO:0018343	protein farnesylation	0.0015
GO:0018347	protein amino acid farnesylation	0.0015
GO:0018342	protein prenylation	0.0141
GO:0018346	protein amino acid prenylation	0.0141
GO:0007265	Ras protein signal transduction	0.0347
GO:0045669	positive regulation of osteoblast differentiation	0.0391
GO:0030501	positive regulation of bone mineralization	0.0391
GO:0046852	positive regulation of bone remodeling	0.0391
GO:0045778	positive regulation of ossification	0.0391
GO:0009117	nucleotide metabolism	0.0209
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	0.0330
GO:0009134	nucleoside diphosphate catabolism	0.0391
GO:0046939	nucleotide phosphorylation	0.0391
GO:0009132	nucleoside diphosphate metabolism	0.0087
GO:0006014	D-ribose metabolism	0.0391
GO:0009191	ribonucleoside diphosphate catabolism	0.0391
GO:0046785	microtubule polymerization	0.0391
GO:0045842	positive regulation of mitotic metaphase/anaphase transition	0.0391
GO:0000718	nucleotide-excision repair, DNA damage removal	0.0391
GO:0042769	DNA damage response, perception of DNA damage	0.0391

A more detailed inspection of the signature genes was performed using the KEGG database of genes or proteins, which illustrates the grouping of the proteins by their participation in signaling or metabolic cellular pathways. Moreover, using the NCBI database of genes, detailed annotations of the genes and the function of the encoded proteins were allocated and functional or signaling connections between them were identified, as described in chapter 5.3.

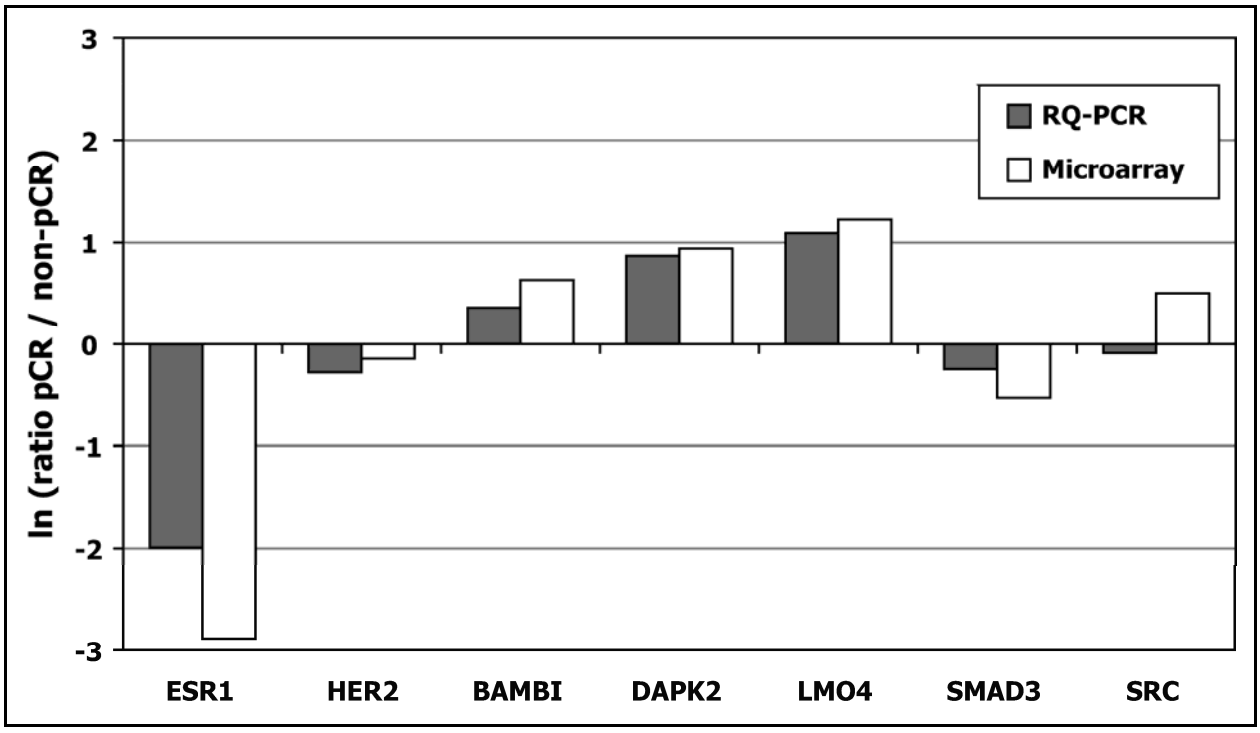
4.3 Validation of Microarray Results by Real-time Quantitative PCR

In order to validate the results yielded by the TAcKLE amplification method in combination with the oligonucleotide gene expression microarrays, and to gain more detailed insights into the molecular pathways involved with the prediction of chemotherapy response, four genes from the signature were chosen to be analyzed for their expression by real-time quantitative PCR (RQ-PCR).

As discussed in chapter 5.3, the genes selected from the signature were *DAPK2*, *BAMBI*, *LMO4* and *SRC* for their associations with the DNA damage response or TGF-*β* pathway. Additionally chosen genes for RQ-PCR were *SMAD3*, coding for a transcription factor interconnecting signaling between BAMBI and LMO4, as well as *ESR1* and *HER2*, two genes for which the corresponding protein levels had been measured for all patients in the clinic.

The results were averaged for all patients in the corresponding responder class (pCR/non-pCR) and the averages set into relation, both in the microarray and RQ-PCR data sets (Figure 23).

Figure 23

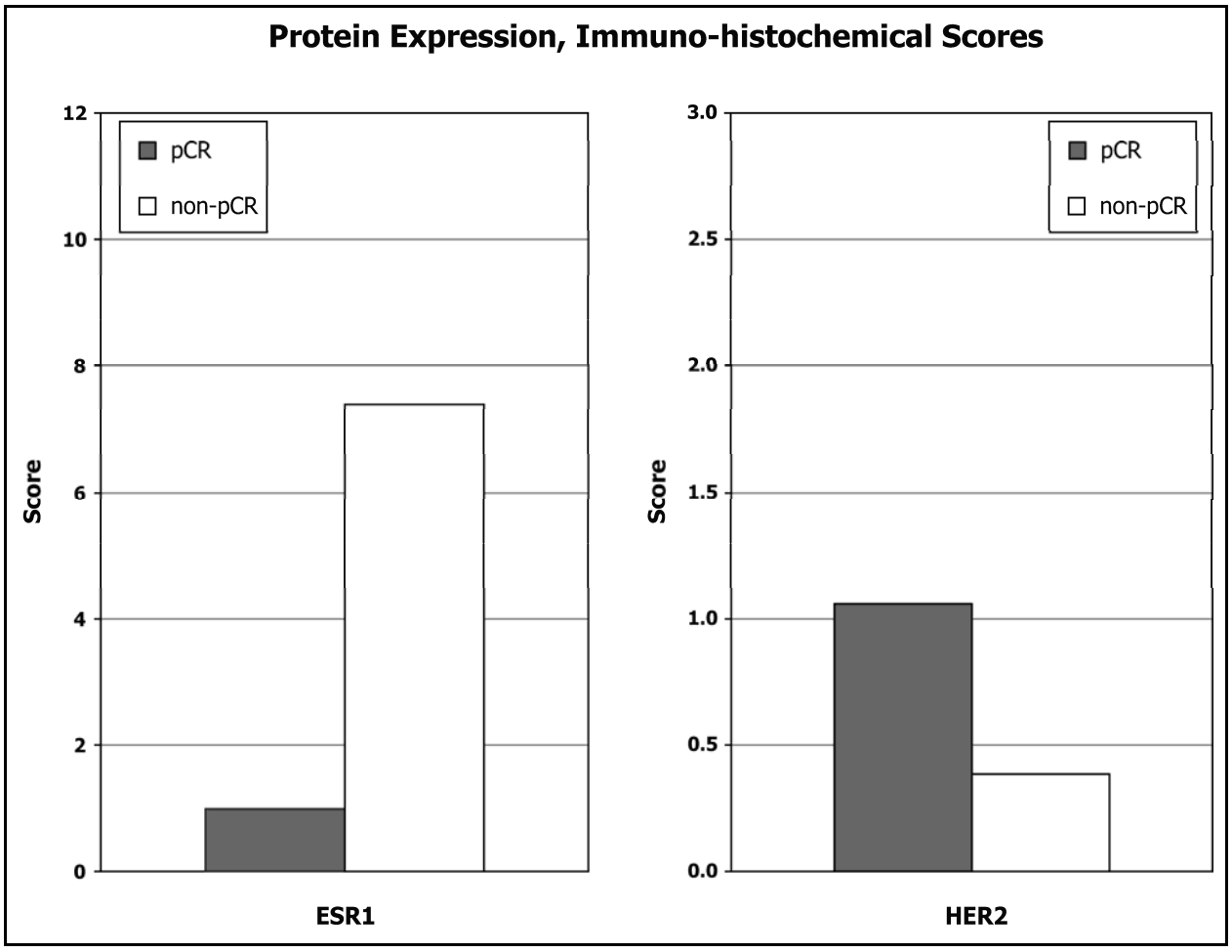


Comparison of gene expression results as measured by Microarray and RQ-PCR techniques. Displayed here are the ratios between expression values averaged for patients with pCR or non-pCR to chemotherapy.

When comparing these data, it became evident that for all genes except *SRC*, the expression differences between classes relate well between RQ-PCR and microarray results. However, the ratios between the patient classes had in these cases higher absolute values for the microarrays as compared to the RQ-PCR, except for *HER2*. Genes downregulated in pCR patients as compared to non-pCR patients were estrogen receptor (*ESR1*) and the TGF-*β*/bone morphogenic pathway (BMP) signal transducer *SMAD3*. *BAMBI*, the negatively regulating pseudoreceptor of the TGF-*β*/BMP signaling family, as well as *LMO4*, a transcription regulator associated with suppression of TGF-*β* target genes, were both upregulated in tumors pCR patients as compared to the tumors of non-pCR patients. Another gene upregulated in tumors of pCR patients is the death-associated protein kinase *DAPK2*, which is thought to induce apoptosis. The epithelial growth factor receptor gene *HER2* did not show significant regulation differences between tumors of pCR and non-pCR patients on the transcription level by both methods.

For the proteins estrogen receptor 1 (ESR1) and HER2, the averages of immuno-histochemical staining scores of the tumor biopsies (Figure 24) were included into the comparison by taking the natural logarithms of the ratios between patients from the different response classes (pCR over non-pCR). These log-ratios amount to -2.0 for ESR1, given as pCR (average score 1.0) *versus* non-PCR (7.4), and 1.0 for HER2 (pCR, 1.06 *versus* non-pCR, 0.38), respectively. While the protein expression value corresponds very well in case of ESR1 to the RQ-PCR and microarray values, it does not reflect transcript data for *HER2*. It should be noted, however, that the HER2 protein expression levels were scored on a scale of 0 to 3, and while their means result 1.06 and 0.38 for pCR and non-pCR, repectively, the medians of the scores have the value of 0 in both classes.

Figure 24

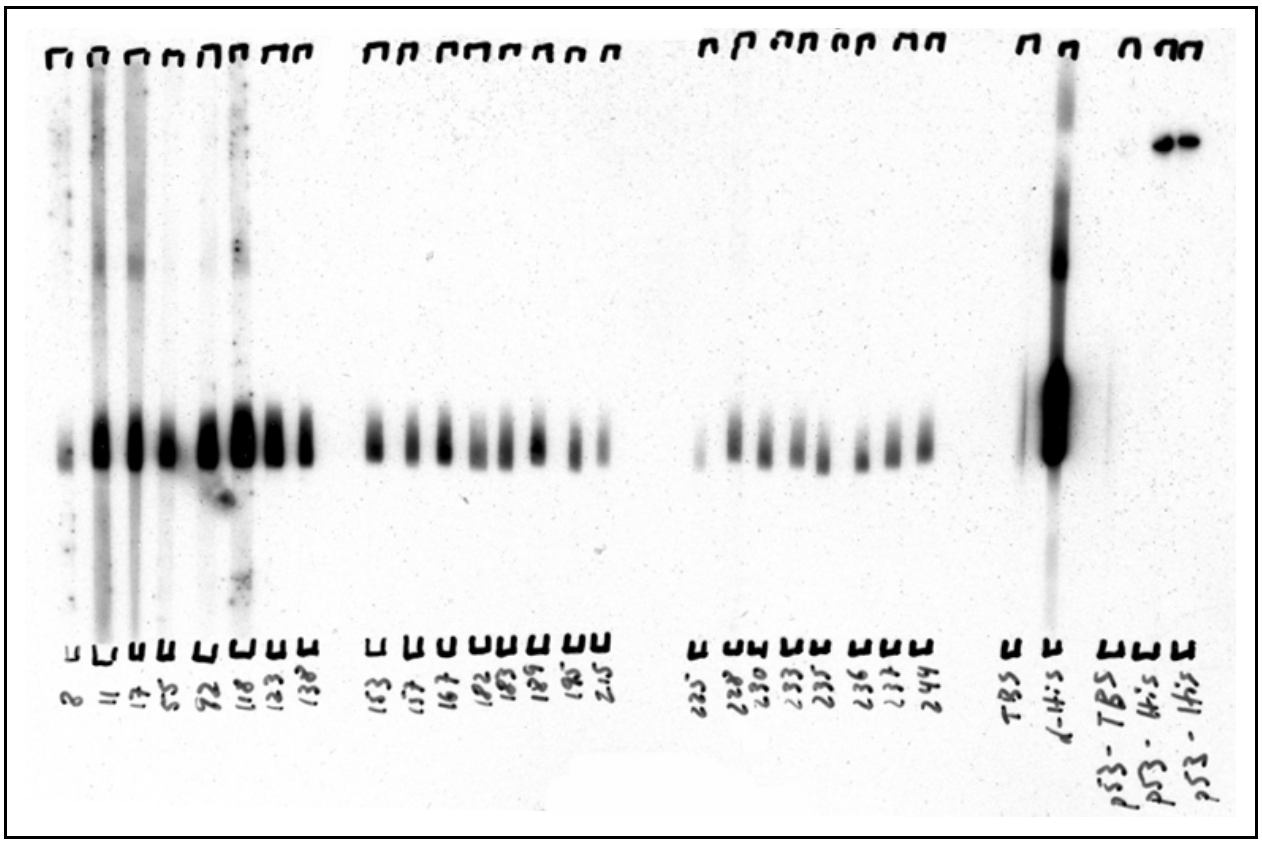


Immuno-histochemical evaluation of tumor biopsies. Tumor sections from patients participating in GEDoc and GEsDoc studies were scored for protein expression of estrogen receptor (ESR1) and HER2/NEU, respectively. Maximum score is 12 for ER expression and three for HER2 expression. Scores were averaged for pCR and non-pCR patient classes. Data was provided by the Clinic for Gynecology and Obstetrics, University of Heidelberg.

4.4 Generation of Antibody using Mouse Hybridoma Cells

One of the genes contained in the signature discriminating between patients classified as responders or non-responders in respect to the chemotherapy was the *BAMBI* gene. It was considered an important protein, as it negatively regulates the TGF-*β* signaling cascade, which was significantly enriched in the gene signature. For this reason, it was decided to measure the protein expression of BAMBI to reveal insight into its role in breast cancer chemotherapy response. Since there was no antibody available at the time the signature was discovered, it was necessary to generate one against BAMBI protein that would be suited for Western blot analysis and immuno-histochemistry. In collaboration with the laboratory of Prof. Hanswalter Zentgraf at the DKFZ, such antibodies were aimed to be developed.

Figure 25



Antibody sensitivity test. Western blot strips containing BAMBI full length protein (1-1 to 4-2), isolated from whole cell lysates of *E.coli* cells overexpressing the protein, or TP53 protein (4-3 and 4-4) with His-tag attached were stained by immune reaction with antibody containing conditioned cell media from different hybridoma cell clones, *α*-His (Qiagen) antibody for positive or TBS buffer for negative control. TP53-His was kindly provided by Prof. Hanswalter Zentgraf.

Conditioned media of the generated hybridoma cell clones were used for testing of the antibodies on Western blots (Figure 25). The sensitivity of the antibodies contained in the conditioned media was very high in samples 11, 17, 55, 92, 118 and 123, but very low in samples 215 and 225. Specificity is low in samples 11, 17 and 118, as seen by their detection of other bands. TBS control was negative, showing only overshadowing of very strong signal from *α*-His control, and TP53 positive controls are also positive. Samples with the best balance of high sensitivity and specificity were 123, 138, 153, 167 and 189.

In order to validate the specificity of the antibodies produced by the hybridoma cells, the antibody containing media needed to be tested against whole cell lysates containing either BAMBI or BMPR1B proteins in their native form. To do such a test, it was therefore necessary to express these proteins in mammalian cell lines. However, it proved impossible to generate such cells: While transfection of cells was successful with the "empty" pBCHGs vectors, as seen by positive fluorescence of GFP protein, the cells transfected with either pBCHGs containing GFP-BAMBI or GFP-BMPR1B fusion genes showed no significant fluorescence (data not shown). To exclude any cell-type specific effects of GFP-BAMBI or GFP-BMPR1B overexpression, the transfection method was evaluated with cell lines from different tissue origins, as e.g. epithelial cells (MCF-7, HeLa), osteosarcoma cells (U20S), hepatocellular carcinoma cells (HEPG2) or embryonic cells (HEK-293). However, even though all of these cells incorporated the DNA vectors, none of them produced sufficient amounts of fluorescent fusion proteins of GFP with BAMBI or BMPR1B to be used for testing with immuno-histochemical methods.
Some of the tested cell lines did produce very rarely a faint signal of GFP fusion proteins, but the signal was too weak. The transfection with GFP-only vectors had an efficiency of 60-70%, depending on cell type (data not shown). At this point, a commercial monoclonal antibody against BAMBI protein had become available, and the development of an antibody was not pursued.

4.5 Immuno-Histochemical Analysis of Patients

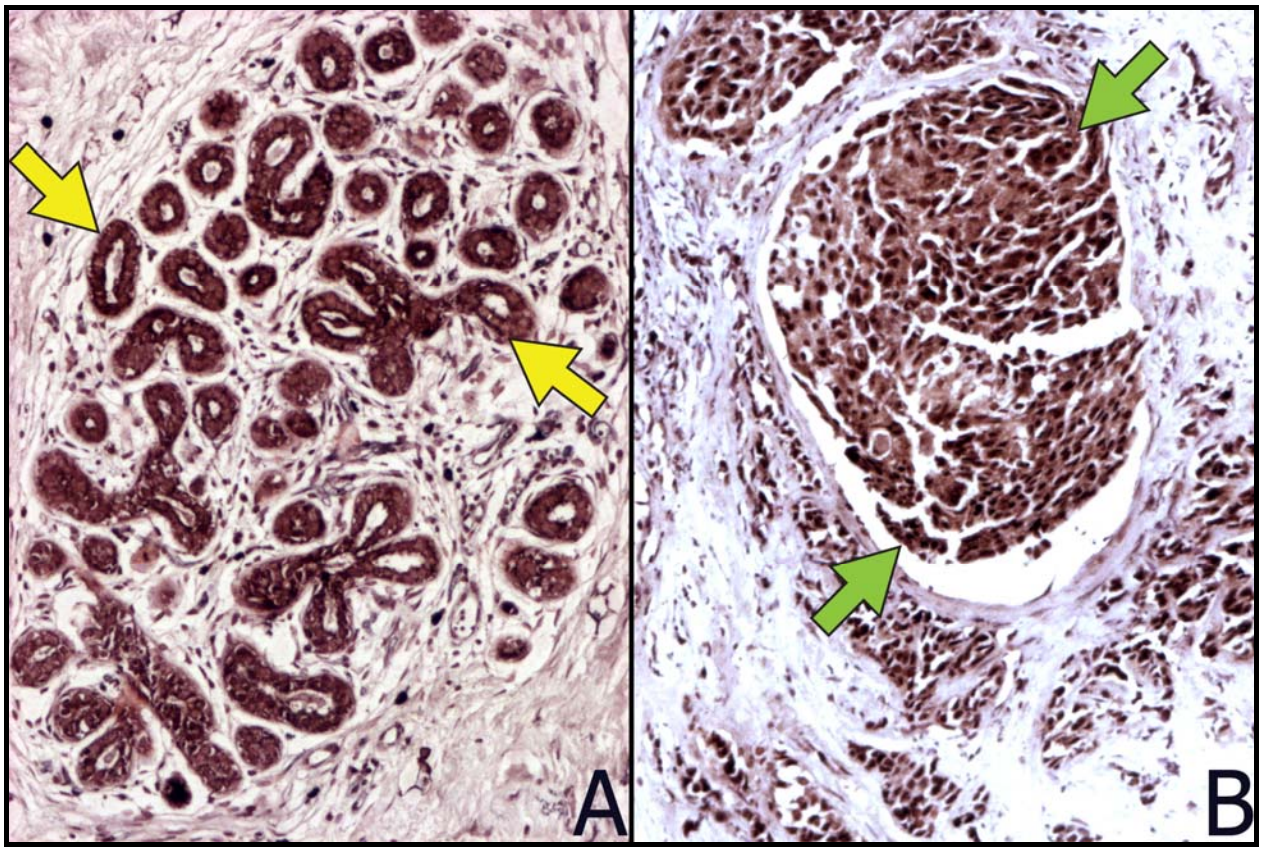
With the identification of genes, of which the expression in the tumors of primary breast cancer patients allows classification of these patients into responders or non-responders to the tested chemotherapy, candidates for corresponding marker proteins were identified.

To test some of these markers for clinical applicability, sections of paraffin-embedded tumor biopsies were obtained from the Clinic for Gynecology and Obstetrics of the University Heidelberg, from 80 of the 100 patients that were included in the final microarray analysis. Unfortunately, only four sections per patient were available, limiting the number of possible tests.

Antibodies against six different proteins were selected for their involvement in pathways enriched in the gene expression signature (BAMBI, BMP4, LMO4, SMAD3, SRC) or for the known responsibility for hereditary predisposition to develop breast cancer (BRCA1). Four of these proteins were used in double-stains, and two solitary in single stains. These proteins were chosen Following the recommendations of the manufacturer, the double stains were performed with the use of NovaRed and SG (dark grey) chromogens. However, the applicability of the technique was limited, as strong and therefore dark red stains obscured weaker grey staining (Figure 26). For that reason, the results of the staining against SMAD3 and BRCA1 were evaluated with caution.

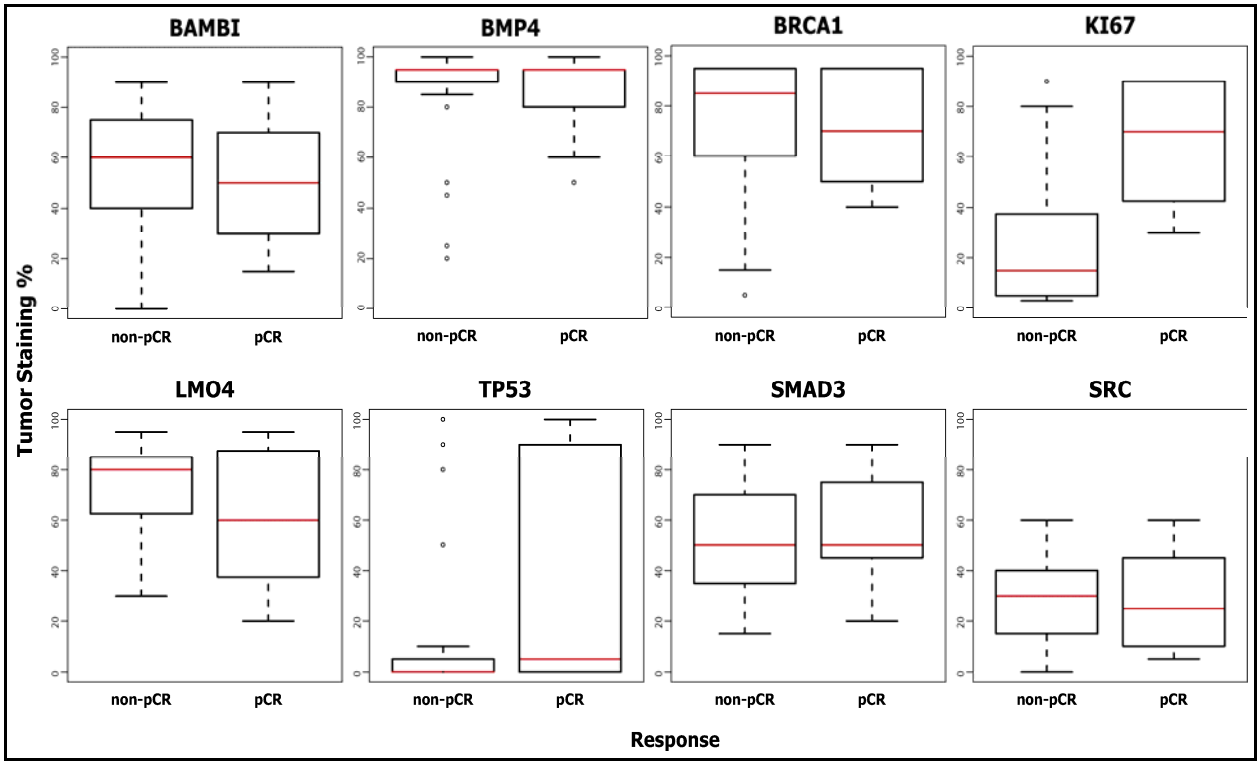
The staining for the different proteins was assessed by scoring of the tumor cells in the tissue biopsies only. Scoring was performed both in respect to staining intensity (scores 0-3) as well as percentage of stained cells. Localization of the cell staining was not accounted for.

Figure 26



Immuno-histochemical double-staining. Example of breast cancer tissue samples stained with both BMP4 (red) and SMAD3 (grey) chromogens. **A.** Image representing a tissue sample with good discrimination between both stainings (yellow arrows). **B.** Image showing a tissue sample with dark red and/or grey stainings that were difficult to discriminate between (green arrows).

Figure 27



Box plots showing protein expression of markers. Marker expression was measured by immuno-histochemical staining, in tumors from patients of the different response groups. Tumor cells in three different fields of sight were evaluated at 10x magnification. Boxes represent values within 1st to 3rd quartile of patients, red lines indicate medians of patients and whiskers represent minimum and maximum value within statistical significance. Outlier values, representing not statistically significant tumor staining percentages, are represented by circles. KI67 and TP53 measurements were performed in the Clinic for Gynecology and Obstetrics of the University of Heidelberg.

For statistical analysis, box-plots of the percentages were performed with regard to the classification of the patients into responder and non-responder patients (Figure 27). Additionally, all different combinations of linear dependencies between the staining patterns as well as between them and important clinical markers were analyzed using linear models for the estimation of probabilities (Table 20). Probability values lower than 0.01 are marked by green fields and bold italic writing, while p-values between 0.01 and 0.05 are marked by yellow fields and bold writing. The upper panel A shows the dependencies of the markers as measured by immuno-histochemistry, among each other. The lower panel B shows also dependencies of these with the clinical markers, as well as therapy (GEDoc *versus* GEsDoc), patient response (pCR *versus* non-pCR) and pathological status of their tumors (pCR, pPR and pNC). The input denotes the percentage value that was used for the classification of the scored other variables (output).

Table 20 **Probability Values for Dependencies Based on Linear Models** *

A

Output (Staining Intensity Score) \ Input (Percentage of Stained Tumor Cells)	BAMBI %	BMP4 %	BRCA1 %	LMO4 %	SMAD3 %	SRC %	TP53 %	KI67 %
BAMBI		0.119	***0.004***	0.353	0.064	0.412	0.466	0.903
BMP4	0.145		0.514	0.106	0.058	0.449	0.662	0.720
BRCA1	***0.004***	0.398		0.441	0.517	0.121	0.423	0.394
LMO4	0.182	**0.026**	0.075		0.693	0.157	0.615	0.248
SMAD3	0.134	0.078	0.255	0.167		0.068	0.527	0.308
SRC	0.113	**0.023**	0.366	0.155	0.386		0.536	**0.047**
TP53	0.972	0.720	0.664	0.117	0.472	0.631		0.174
KI67	0.674	0.574	0.340	0.319	0.378	0.964	***0.003***	

B

Output (Staining Intensity Score / Factor) \ Input (Staining Intensity Score / Factor)	Re-sponse	The-rapy	Gra-ding	ER	PgR	HER2	BCL2	BAM-BI	BMP4	BRCA 1	LMO4	SMAD 3	SRC
Response		0.385	**0.040**	***0.000***	***0.000***	**0.012**	0.052	0.557	0.472	0.944	0.101	0.332	0.548
Therapy	0.385		0.610	0.901	0.745	0.701	0.311	***0.001***	0.600	0.560	0.920	0.654	***0.009***
Grading	0.123	0.563		**0.025**	**0.037**	**0.050**	***0.000***	0.374	0.441	0.507	0.490	0.331	0.908
ER	***0.000***	0.823	***0.008***		***0.000***	0.099	***0.005***	0.507	0.644	0.688	0.307	0.679	0.405
PgR	***0.000***	0.409	0.050	***0.000***		**0.048**	0.054	0.312	0.811	0.399	0.743	0.468	0.786
HER2	**0.029**	0.637	0.056	0.058	0.162		0.491	0.585	0.922	0.358	0.591	0.217	0.711
BCL2	**0.040**	0.436	***0.001***	**0.010**	**0.023**	0.621		0.151	0.352	**0.020**	0.866	0.482	0.397
BAMBI	0.781	***0.005***	0.133	0.560	0.969	0.498	0.768		0.076	***0.000***	0.711	0.222	**0.016**
BMP4	0.593	0.557	0.495	**0.013**	**0.035**	0.548	0.356	0.157		0.818	**0.041**	**0.026**	0.139
BRCA1	0.116	0.443	0.838	0.466	0.060	0.653	**0.025**	***0.000***	0.803		0.716	0.140	**0.028**
LMO4	0.262	0.862	0.453	0.426	0.192	0.595	0.804	0.835	***0.010***	0.876		0.956	0.236
SMAD3	0.563	0.901	0.712	0.598	0.355	0.290	0.978	0.808	0.110	0.223	0.509		0.534
SRC	0.708	**0.033**	0.830	0.101	0.175	0.374	0.411	***0.005***	0.220	**0.024**	0.085	0.214	
Pathol. Status	***0.000***	0.377	0.122	***0.000***	***0.000***	**0.042**	0.153	**0.030**	0.628	0.189	0.153	0.428	0.394

Response (pCR/non-pCR); Therapy (GEDoc/GEsDoc); Grading (G1/G2/G3); ER, estrogen receptor score (0-12); PgR, progesterone receptor score (0-12); Pathol. Status, pathological status at surgery (pCR/pPR/pNC); all other markers scored 0-3. Data for ER, PgR, HER2, BCL2, TP53, KI67 as well as clinical factors provided by the Clinic for Gynecology and Obstetrics, University of Heidelberg.

* P-values below 0.01, green; p-values between 0.01 and 0.05, yellow.

In some cases, a strong reciprocal dependency was detected, e.g. for BAMBI with BRCA1 or ER with PgR. However, there are cases that show unidirectional dependencies, as for example SRC depending on BMP4 percentage but not *vice versa*.

According to these linear models, the pathological status (pCR/pPR/pNC) was highly associated with ER and PgR status, and HER2 and BAMBI protein expression showed good association. Response to chemotherapy (pCR/non-pCR), a slightly different classification, was highly associated with ER and PgR status. HER2 protein expression showed good association, but there was no significance in expression of BAMBI. Notably, the type of therapy given was

associated with BAMBI protein expression, although the therapies were administered after aquisition of the samples.

Between the different marker proteins, the linear models showed strong associations of BAMBI and BRCA1, KI67 and TP53, and good associations between both LMO4 or SRC and BMP4, respectively.

One clinical marker, the grade of the tumors, was strongly associated with BCL2 score and showed significant association with ER, PgR and HER2 status.

5. Discussion

5.1. Messenger RNA Amplification and Labeling Protocol

The use of oligonucleotide microarrays for expression analysis of small biopsies from tumor material from female patients set the task of developing an appropriate protocol for amplification of the messenger ribonucleic acids. Since the established protocols could not be applied *per se* to be used with sense-orientated oligonucleotides spotted on arrays for complementarity reasons, the known protocols needed to be either adapted or entirely substituted.
Several different methodologies were tested for this purpose, and evaluated with respect to amplification rate, dye incorporation efficiency in total and in comparison of both dyes used, linearity of the amplification across different mRNA molecule sizes and applicability to the oligonucleotide microarray technology. Additionally, the usefulness from the economic and laboratory handling standpoints were also taken into consideration for the decision of the most appropriate protocol.

The first analysis, the incorporation rate in total and in comparison between both fluorescent dye nucleotide types (Figures 13 and 14), shows a strong disadvantage of the PALDA and IVT methods. As the Primer-Assisted Linear DNA Amplification (PALDA) uses *Taq* or *Pfu* exo$^-$ DNA polymerases for integration of fluorescently labeled dyes, a low incorporation rate was initially expected. To overcome this restraint, a high concentration of labeled dyes was used initially, and the amplification procedure was performed for 2x 50 cycles to reach the necessary amplification efficiency. Nevertheless, the yield of fluorescently labeled DNA was very low in comparison with the other protocols. The IVT labeling protocol, designed to integrate fluorescently labeled RNA molecules during *in vitro* transcription, was also expected to have a low incorporation rate and thus was started with a high concentration of the labeled nucleotides. Although the efficiency of fluorescent nucleotide incorporation is not comparable with the best methods in this respect, it is significantly better compared to the PALDA protocols and sufficient for hybridization to the microarrays. However, the incorporation was shown to have a strong bias, supposedly from a preference of the RNA polymerase used

in this protocol towards Cy3-labeled nucleotides. This effect, seen as a Cy3:Cy5 ratio of approximately two, is significantly higher in IVT labeling than all other tested protocols, which have a maximum ratio of 1.4.

While the yield and incorporation ratio were measured on the amplified nucleic acids directly and in total, the following analyses were performed with amplified and labeled nucleic acids actually hybridized to the oligonucleotides on the microarrays. As a benchmark for the comparison of the protocols, the direct labeling protocol was performed and evaluated along with the amplification procedures.
Consecutive to the yield measurements was the issue of what proportion of amplified molecules effectively participated in the hybridization, or to which extent the amplified material was an interfering side-product. To answer this question of amplification specificity, analyses of signal-to-background intensity ratios of the array features were performed (Figure 15). The IVT labeling method results in a very high signal-to-background ratio of more than 200, approximately 4-fold higher than the direct labeling method with a ratio of approximately 57. Not surprisingly, and probably resulting from both the fractionation and the strand non-specificity of the Klenow enzyme used in these protocols, the Baugh + Klenow (TAcKLE) as well as Single Primer Amplification (SPA) methods yielded approximately 50% or 30% of the ratio of direct labeling, respectively. However, these results were acceptable, whereas the results of the PALDA and Baugh Standard protocol were not.

Continuing with specificity, the results of repeat experiments were analyzed both in the same-*versus*-same (equivalent) and differential hybridization setting by estimating correlation coefficients across all valid features of the microarrays (Figure 16). Unfortunately, the correlation could not be calculated between two differentially hybridized samples for all protocols. In general, the correlation allows elucidating, whether the measured signal intensities are really specific for the genes. Sufficient reliability was seen for Baugh + Klenow (TAcKLE), IVT, ts-SPA and direct labeling, with correlation coefficients of 0.9 or higher for R^2. Baugh Standard and SPA performed poorly, while for PALDA this analysis was not possible due to the very limited yield. On the other hand,

results from differential hybridizations showed a limited decrease of correlation in case of the direct labeling protocol, as compared to the value for the equivalent hybridizations. This is also the case in ts-SPA, while the Baugh+Klenow (TAcKLE) shows a strong decrease to approximately 60% of the R^2 of equivalent hybridizations.

While in the microarray analyses for the performance of amplification methods above only valid spots were taken into consideration, the percentage of spots not valid for analysis was of great interest too. Figure 17 depicts the ratio of these so-called outlier spots, averaged for the arrays of each protocol. Outliers are those spots that either have no intensity and were therefore flagged as such by the software or those that were manually flagged as false positives by the user. The PCR protocol, which was used for negative control, had an extremely high percentage of such outliers (50%), and the PALDA protocol (28.5%) also had a significantly high proportion. The other PCR-based methods had tolerable, but higher percentages than direct labeling (18.6% and 16.1% for SPA and ts-SPA, respectively), while the Baugh Standard and Baugh + Klenow methods performed similar to direct labeling (12.2%-14.2%). The IVT labeling had the lowest outlier percentage (11.5%).

Another view taken on the performance of the amplification and nucleic acid labeling was the analysis of linear trend lines from scatter plots of feature intensities. This trend should ideally be close to the bisecting line of the plot, as the total raw intensity distributions from the two dye channels should be similar. Irregularities were displayed by the slope and intersection point on the ordinate of this trend. Significant deviations of the slope between the intensities for both channels from one or deviations of the intersection on the ordinate from the origin were considered hazardous. However, as the slope also varies with the amount of input mRNA, slight variations seen with all tested methods were considered acceptable. Apart from the PCR method, used as a negative control, the slopes were all within 1 ± 0.25 (data not shown). The interception on the ordinate showed a very strong variation for the IVT labeling, reflecting the differences in the incorporation between Cy3- and Cy5-modified

nucleotides (Figure 18). All other protocols considered for amplification were well within the acceptable range of 0 ± 500.

Finally, the quality of the valid features was assessed by estimating the homogeneity, averaging all ratios of mean to median for each spot per array. This parameter is generally used for filtering of features in microarray analysis and was therefore an important determinant of data validity. To normalize for the different numbers of valid spots, the percentage of valid features outside the accepted homogeneity interval was calculated (Figure 19). As the threshold for filtering microarray features with this parameter varies between 20% and 30% in final data analysis, the IVT labeling protocol and PALDA method were considered as too erratic.

In summary, the amplification procedure best applicable and most stable in the comparison of the protocols was the Baugh + Klenow method, which had been based on studies by Eberwine *et al.*, Baugh *et al.* and Kenzelmann and co-workers. It was later named "**T**7-based **A**mplification of **c**DNA and **K**lenow **L**abeling for **E**xpression Analysis", or *TAcKLE* analysis.
The second most appropriate procedure, which had been considered for multiple reasons, was the template-switch Single Primer Amplification, or ts-SPA. This method was based on works of Ena Wang *et al.* and Matz and co-workers.[127,129] The advantages mainly comprised laboratory handling and economic rationales, as the fewer reaction steps are also very commonly used and cost-effective. However, the lack of discrimination between differentially hybridized samples was the major objective to disregard the method, along with its weaker overall performance, e.g. signal-to-background ratio and outlier feature percentage.

The TAcKLE analysis procedure was therefore chosen as the method of choice to amplify and label mRNA from core needle biopsy samples for subsequent hybridization to oligonucleotide microarrays.

5.2. Gene Expression Signature Predictive for Chemotherapy in Primary Breast Cancer

The study investigated and presented here aimed at the identification of a gene expression signature, which allows for the classification of patients according to their response to GE(s)Doc chemotherapy. For female primary non-metastatic breast cancer patients receiving a neoadjuvant triple chemotherapy consisting of the regimen gemcitabine, epirubicin and docetaxel, this classifier should allow prediction of reaching the pathologically proven complete remission of their tumor at time of surgery with high sensitivity and accuracy.

The patients were enrolled in two slightly different clinical studies, designated GEDoc and GEsDoc, which differ in their administration schedule and dosage (Figure 8). While the GEDoc cohort received all three therapeutics in parallel, the GEsDoc cohort received the third therapeutic, docetaxel, sequentially after gemcitabine and epirubicin treatment. As both clinical studies yielded the same percentage of pathological complete remission (pCR) of 26%, and all other parameters remained similar, they were considered to be comparable also on the molecular biology level.[47,104] The GEsDoc cohort of patients was used as training set to identify the gene expression signature, while the GEDoc cohort was used as independent test set to estimate predicive power of this signature.
The patients agreed to contribute to the microarray study with a core biopsy taken from their tumor, from which RNA was extracted, amplified, labeled and hybridized to the microarrays using the TAcKLE analysis procedure.

As the threshold for the sensitivity of the prediction classifier was set to be at least 80% or 12 of 15 patients in the training group to prove its clinical applicability, the number of genes necessary for prediction was 512. Although it is possible to deduce a ranking of these genes in respect to their discriminating power, it is important to point out that in the theory of the used algorithm, the Support Vector Machines (SVM), none of the genes has a higher importance than the others. Each of the genes used for the classification has the same weight, and is therefore as necessary to make the prediction as the other genes.

When comparing the predictive power of the gene expression signature with the clinical predictive factors for chemotherapy to date, e.g. tumor grading, HER2 expression, hormone receptor status or clinical response after six weeks of therapy, the signature proved to be of superior predictive value (Table 18). In the test patients group, only the HER2 score (0-2 *versus* 3) shows a significant independent predictive value, but with a lower predictive power than the signature. Other proposed candidates, e.g. clinical response after 6 weeks of treatment, do not have any statistically significant predictive values, while tumor size (smaller than 5 cm) and grading (G1/2 *versus* G3) only show a predictive trend.

For comparison of the gene expression signature with other clinically relevant expression data published in respect to breast cancer and the usefulness of microarrays, these have to be divided into three groups:[95,105,148,149]

(a) Prognostic molecular profiles, using unsupervised clustering of all or only pre-selected subsets of genes, were aimed at molecular classification or the prospective classification of disease-free and overall survival or metastasis. These data sets, as those of Sorlie *et al.*, van 't Veer *et al.*, Perou *et al.* and others following since, provide valuable information about the patient's tumors genetic setup, and have helped to understand and interpret the heterogeneity of the clinical course of patients.[93,105,106,150] However, these studies are not related with the type of treatment, and thus cannot be compared with this study.

(b) Predictive molecular profiles that are aimed at long-term effects of (adjuvant) treatments, e.g. tamoxifen, trastuzumab and others, or at the prediction of chemo-resistance had been done in retrospective manner. These studies need long follow-up monitoring of the patients, as the effects of treatment or resistance can only be seen after five or more years. Therefore, results are not available yet for comparison.

(c) Predictive molecular profiles that identify gene expression signatures for chemotherapy, ideally in the neoadjuvant setting, are the only studies that the gene expression signature presented here can currently be compared to, if the therapy settings are relatively comparable.

Table 21 **Studies of Gene Signatures Predicting Chemotherapy Response**

Authors	Tumors (n)	Prediction Endpoint	Molecular Tool	Genes (n)	Publication
Chang JC *et al.*	**24**	Response* in (A)	Affymetrix HgU95-Av2 (12k)	**92**	08/2003, [110] Lancet
Ayers M *et al.*	**42**	pCR in (T+FAC)	custom cDNA array (31k)	**74**	06/2004,[111] J Clin Oncol
Iawo-Koizumi K *et al.*	**70**[§]	CR in (D)	ATAC-PCR (2,453)	**85**	01/2005,[151] J Clin Oncol
Hannemann J *et al.*	**48**	"near" pCR in (AD vs. AC)	custom cDNA (18k)	**--**	05/2005, [114] J Clin Oncol
Rouzier R *et al.*	**82 (22**[b]**)**	pCR in (T+FAC)	Affymetrix U133A	**-- (61**[b]**)**	08/2005,[152] Clin Cancer Res
Gianni L *et al.*	**89; 82**[¶]	pCR in (AT+T); (T+FAC)	RQ-PCR (384); Affymetrix U133A (14k)	**86 (79)**	10/2005,[153] J Clin Oncol
Dressman HK *et al.*	**37**	Clinical Response[#] in (AT)	Affymetrix U133 Plus 2.0 (38.5k)	**38**	02/2006,[154] Clin Cancer Res
Paik S *et al.*	**424**	Response[†] in (Tam+CMF/MF)	RQ-PCR (21)	**21**	08/2006,[115] J Clin Oncol
Sørlie T *et al.*	**81**	PR in (A vs. FMi)	custom cDNA (8k A / 30k FMi)	**--**	11/2006,[155] Mol Cancer Ther
Bonnefoi H *et al.*	**66; 59**[¶]	pCR in (FEC); (D+ED)	Affymetrix X3P (38.5k)	**NA**	12/2007,[156] Lancet Oncol

A, doxorubicin; C, cyclophosphamide; E, epirubicin; F, 5-fluorouracil; D, docetaxel; M, methotrexate; Mi, mitomycin; T, paclitaxel; Tam, tamoxifen; CR, complete response; pCR, pathological complete response; PR, partial response

* defined as ≥75% regression of tumor; [§] primary or locally recurring breast cancers; [b] basal-like patient subgroup; [¶] two differently treated patient cohorts; [#] defined as absence of pos. lymph nodes; [†] defined as freedom of distant recurrence

The latter group of studies is aimed at identifying gene expression patterns or signatures predicting response to chemotherapy. This topic is an intensely investigated research field, and a number of studies have already been published since the start of this thesis (Table 21). It is necessary to define the clinical setting that can be effectively compared with the study introduced here. Two of these recent trials, by Hannemann *et al.* and Sorlie *et al.*, completely failed to find a predictive gene signature, possibly due to their definition of response. Although other studies (Chang JC *et al.*, Iawo-Koizumi K *et al.*, Dressman HK *et al.*, Paik S *et al.*) with a similar definition of response succeeded in making predictions based on gene expression, endpoints other than pathological complete remission have been shown to be only weakly associated with patient overall or disease-free survival.[98,101] The two most comparable of those (Chang, Dressman) also do not show any gene overlap in

their predictive signatures. This could be due to the facts that both studies used a very limited number of patients, and without validating their signatures in an additional patient set. The studies by Iawo-Koizumi as well as Paik and respective co-workers were performed on pre-selected gene sets, and are thus not comparable to any of the other studies, including the one performed here.

A limited number of patients (n=42) was also the basis for the analysis Ayers *et al.* performed on chemotherapy comprised of sequentially administering T (paclitaxel) and triple therapy with FAC (5-fluorouracil, doxorubicin, cyclophosphamide). Nevertheless, the authors split the patients into two groups to independently build and then test their signature classifier. While all other parameters differed only insignificantly between their training and validation cases, the percentage of patients reaching pathologic complete remission was significantly higher in the validation group (39%) *versus* the training group (25%). The overall accuracy of their 74-gene classifier in the test group was 78%, with a sensitivity of 43% and a specificity of 100%. When compared to the present study, the focus of Ayers *et al.* on the high specificity becomes evident, making sure to select only patients that would benefit from the therapy. In contrast, the focus of the study discussed here was put on the highest overall accuracy (88%), resulting in a much higher sensitivity (78%) but accepting a comparably lower specificity (90%).
Patients receiving the same chemotherapy regimen as in Ayers and co-workers study were analyzed in a trial performed by Rouzier *et al.* In their investigation, the authors decided to differentiate the patients first by usage of the "breast cancer intrinsic gene set", previously published by Sorlie *et al.* in 2001.[106] Then, using the four different molecular subtypes of patients they received, these were subjected separately to the identification of gene expression signatures. As the resulting patient subgroups were again very small, and two of them had a very low to no percentage of pCR patients, only the HER2+ and basal-like subsets could be used. As a benefit, these two subsets had a higher pCR patient percentage (45% in both) than the entire collective. However, only for the basal-like subset a gene signature predictive for pCR could be identified, containing 61 genes. Due to the pre-clustering of patients, the results of the

Rouzier study cannot be compared with the outcome of the study presented here.

Gianni *et al.* also included patients receiving T/FAC as chemotherapy treatment. However, these authors first performed the gene expression signature identification on a different set of patients, who received doxorubicin and paclitaxel followed by another paclitaxel regimen (AT+T; INT-Milan cohort). After identification of the gene signature, it was then tested on patients treated with T/FAC (MDACC-Houston cohort). Adding to this complication, the INT-Milan group was assessed using RQ-PCR to identify a predictive gene subset out of 384 pre-selected genes, yielding an 86-gene signature classifier. This signature was then tested on the MDACC-Houston cohort, which had been profiled using gene expression microarrays, to validate its performance within that dataset. Only 79 of the 86 genes were represented in the microarray dataset, and 24 of these showed an association with pCR in the MDACC-Houston dataset of $p \leq 0.05$. Again this study cannot be directly compared to the investigation here, due to Gianni *et al.* limiting the number of investigated genes for identification of the signature.

In the most recent study in the field, performed by Bonnefoi and co-workers, gene signatures identified by cell culture experiments to be predictive for resistance against single chemotherapy agents were used in a combinatorial approach. Both investigated patient cohorts, treated either with FEC (5-fluorouracil, epirubicine and cyclophosphamide) or D+ED (docetaxel sequentially followed by epirubicine and docetaxel, published "TET"), were used to validate gene expression signatures deduced from cell line experiments for the single chemotherapy agents in an earlier study.[157] The investigated patients of this study, however, were pre-selected to be ER-negative. Although it is also not comparable with the study performed in this thesis, the work of Bonnefoi *et al.* is very interesting the way it may lead into the future, as it successfully integrated separate gene signatures for each chemotherapeutic drug on a bioinformatic level. If this procedure of identifying signatures for single drugs and applying them in a combinatorial fashion to patients receiving multi-drug therapies proves successful in general, it could facilitate the urgent solution to tailor the chemotherapy to each patient's best benefit.

In summary, most of these trials, which were published in the four years since the presented study was started, are not comparable with it. Often the investigators lacked the necessary accuracy in defining the clinically relevant study endpoint, a pathologically assured complete remission (response) of the tumor after chemotherapy. Additionally, the datasets are often strongly biased, either by a pre-selection of genes based on literature research and historical presumptions, by the pre-selection of patients to increase the response rates of the investigated cohorts artificially, or both.
The study that can be best compared with the one presented in this thesis is the one performed by Ayers *et al.* Unfortunately, the authors took a different perspective on the focus of the statistical analysis, in terms of sensitivity *versus* specificity, than this study. The overlap of genes from their 74-gene signature classifier and the 512-gene signature identified here amounts to three genes only (*APOE*, *NME2*, *SCARA3*). Taken into consideration that the gene expression methods, statistical approaches and most importantly the chemotherapeutical treatment for the patients used differ largely, this is not surprising.

Whether the signatures derived from all these studies, including the one presented here, are specific to the chemotherapeutic regimens used to treat the patients, provide a general applicability with any chemotherapy, or a mixture thereof remains to be determined. But as the two studies that can be best compared directly show very little overlap, a general applicability seems to be more unlikely. A more standardized approach to the clinical endpoints as well as the molecular methods would be necessary to address this question for different studies. However, some of the research groups that investigated gene expression signatures for prediction in breast cancer mixed different methodologies even within the same study. A truly comparative clinical trial, evaluating patients receiving different treatments with the same biological methods and statistical approach, is therefore essential to find a definitive answer.

5.3. Genes and Pathways involved in Prediction of Chemotherapy Response

The genes of which the expression levels enable to discriminate between patients fully responding to the chemotherapy and those not fully responding were identified and ranked according to their discriminating power by statistical analysis. Of these, the top 512 genes had been determined to be necessary for a predictive classification into responders and non-responders with a prediction sensitivity of ~80% and thus ensure its significance, as proven by multivariate logistic regression testing.

Most generally, bioinformatic tools to identify gene or protein associations rely on the correct annotation of genetic information with the function of their corresponding proteins and the interplay they have in a body or cell. This information, which is gathered by the scientific community and made known through their publications, is collected and stored in databases, which are maintained by bodies of scientific consortia. The 512 genes of the signature presented here were analyzed using databases incorporating Gene Ontology annotation data (GO and FatiGO), revealing functional and signaling interconnections of encoded proteins (KEGG) and harboring published functional and protein interaction information (NCBI Gene and Pubmed databases).[142-144]

A Fisher test was used to identify statistically significant enrichments of genes within the gene signature as compared to all genes represented on the microarrays, determined by Gene Ontology (GO) terms (Table 19). According to these, the gene signature represents a number of genes that could be associated with the chemotherapeutic action of the regimen the patients received, consisting of gemcitabine (a cytidine nucleoside analogue to which no other nucleoside can be attached), epirubicine (a DNA-intercalating anthracycline additionally producing free radicals in the cells) and docetaxel (stabilizing GDP-bound *β*-tubilin and thus preventing depolymerization of microtubules). However, groups of genes could be shown to be active in other pathways that were not directly associated with the action of the chemotherapeutics (Table 22). These include genes associated with RAS signaling and the related protein farnesylation metabolic pathway, the TGF-*β* signaling and associated bone remodeling pathways as well as gene involved in

DNA damage perception or response and genes playing a role in apoptotic pathways.

Table 22 **Signature Genes Related to Pathways or Cellular Processes**

rank	gene	gene description
Regulation of TGF-β / EP300 pathway		
19	*LMO4*	LIM domain transcription factor LMO4 (LIM-only protein 4) (LMO-4) (Breast tumor autoantigen).
23	*BAMBI*	BMP and activin membrane-bound inhibitor homolog precursor (Putative transmembrane protein NMA) (Non-metastatic gene A protein).
24	*EP300*	E1A-associated protein p300 (EC 2.3.1.48).
97	*BMP4*	Bone morphogenetic protein 4 precursor (BMP-4) (BMP-2B).
98	*CREB3*	Cyclic AMP-responsive element-binding protein 3 (Luman protein) (Transcription factor LZIP-alpha).
107	*SMURF2*	Smad ubiquitination regulatory factor 2 (EC 6.3.2.-) (Ubiquitin-protein ligase SMURF2) (Smad-specific E3 ubiquitin ligase 2) (hSMURF2).
171	*SRC*	Proto-oncogene tyrosine-protein kinase Src (EC 2.7.1.112) (p60-Src) (c-Src).
222	*TRIP6*	Thyroid receptor interacting protein 6 (TRIP6) (OPA-interacting protein 1) (Zyxin related protein 1) (ZRP-1).
325	*TGIF2*	Homeobox protein TGIF2 (TGFB-induced factor 2) (5'-TG-3' interacting factor 2) (TGF(beta)-induced transcription factor 2).
RAS pathway		
2	*RASAL1*	RasGAP-activating-like protein 1.
30	*A-RAF*	A-Raf proto-oncogene serine/threonine-protein kinase (EC 2.7.1.37) (A-raf-1) (Proto-oncogene Pks).
48	*DAB2IP*	DAB2 interacting protein isoform 2.
133	*RAB32*	Ras-related protein Rab-32.
142	*RRAGC*	Ras-related GTP binding C.
227	*RAB5A*	Ras-related protein Rab-5A.
298	*RASL11B*	RAS-like family 11 member B.
307	*RASA3*	Ras GTPase-activating protein 3 (GAP1(IP4BP)) (Ins P4-binding protein).
396	*RASSF1*	Ras association domain family 1 (Ras association, RalGDS/AF-6, domain family 1).
412	*RHEB*	GTP-binding protein Rheb (Ras homolog enriched in brain).
456	*MRAS*	Ras-related protein M-Ras (Ras-related protein R-Ras3).
484	*RSU1*	Ras suppressor protein 1 (Rsu-1) (RSP-1).
Regulation of apoptosis		
1	*DAPK2*	Death-associated protein kinase 2 (EC 2.7.1.37) (DAP kinase 2) (DAP- kinase related protein 1) (DRP-1).
58	*DIP*	death-inducing-protein
99	*KIAA1303*	Regulatory-associated protein of mTOR (Raptor) (P150 target of rapamycin (TOR)-scaffold protein).
168	*BAK1*	Bcl-2 homologous antagonist/killer (Apoptosis regulator BAK) (BCL2- like 7 protein).
181	*MRPS30*	Mitochondrial 28S ribosomal protein S30 (S30mt) (MRP-S30) (Programmed cell death protein 9) (BM-047).
208	*MRPL37*	Mitochondrial ribosomal protein L37
274	*MRPL30*	Mitochondrial ribosomal protein L30 isoform a.
360	*FRAP1*	FKBP-rapamycin associated protein (FRAP, mTOR) (Rapamycin target protein).
DNA damage response		
88	*BRAP*	BRCA1-associated protein (EC 6.3.2.-) (BRAP2) (Impedes mitogenic signal propagation) (IMP) (RING finger protein 52).
346	*TP53BP1*	Tumor suppressor p53-binding protein 1 (p53-binding protein 1) (53BP1).
355	*CHEK2*	Serine/threonine-protein kinase Chk2 (EC 2.7.1.37) (Cds1).
446	*TP53RK*	TP53 regulating kinase (EC 2.7.1.37) (p53-related protein kinase) (Nori-2).
487	*RAD51C*	DNA repair protein RAD51 homolog 3.

Genes contained in the signature from these functional groups cannot be considered to act independently, as the analysis of protein function of the respective relevant genes from these pathways reveals. Signature genes are highlighted by blue letters in the following.

RAS Signaling Pathway

Analysis of genes in the signature using Fisher's test identified protein farnesylation as a highly significant metabolic pathway in the Gene Ontology terms. The corresponding genes in the signature are members of or closely related to the Ras superfamily of proteins, which is well known as potential target for oncogenic transformation of cells. Many of the small GTPases, including the Ras, Rho and Arf subfamilies of these proteins are post-translationally modified by covalent addition of a farnesyl group, an isoprenoid, which anchors these proteins in the plasma membrane. This modification step is administered by an enzyme called farnesyl-transferase. Due to the oncogenic potential of Ras, the application of farnesyl-transferase inhibitors (FTI) for therapeutical use is currently under investigation in clinical trials.[158]

The genes belonging to the Ras signaling pathway or related to it, that were contained in the gene expression signature predicting the response, were *MRAS*, a homologue to the main signaling kinase HRAS. *MRAS* was initially found in muscle cells but is now known to be expressed also e.g. in epithelial cells, *A-RAF*, coding for a downstream effector kinase of Ras proteins as well as genes coding for other Ras-associated proteins like e.g. *RHEB*, *RRAGC* or *RASAL1*, that are held responsible for GTP recruitment or GDP/GTP exchange. The latter are necessary for RAS and RAF proteins to perform their phosphorylating enzymatic function or for recycling the GTP molecules.

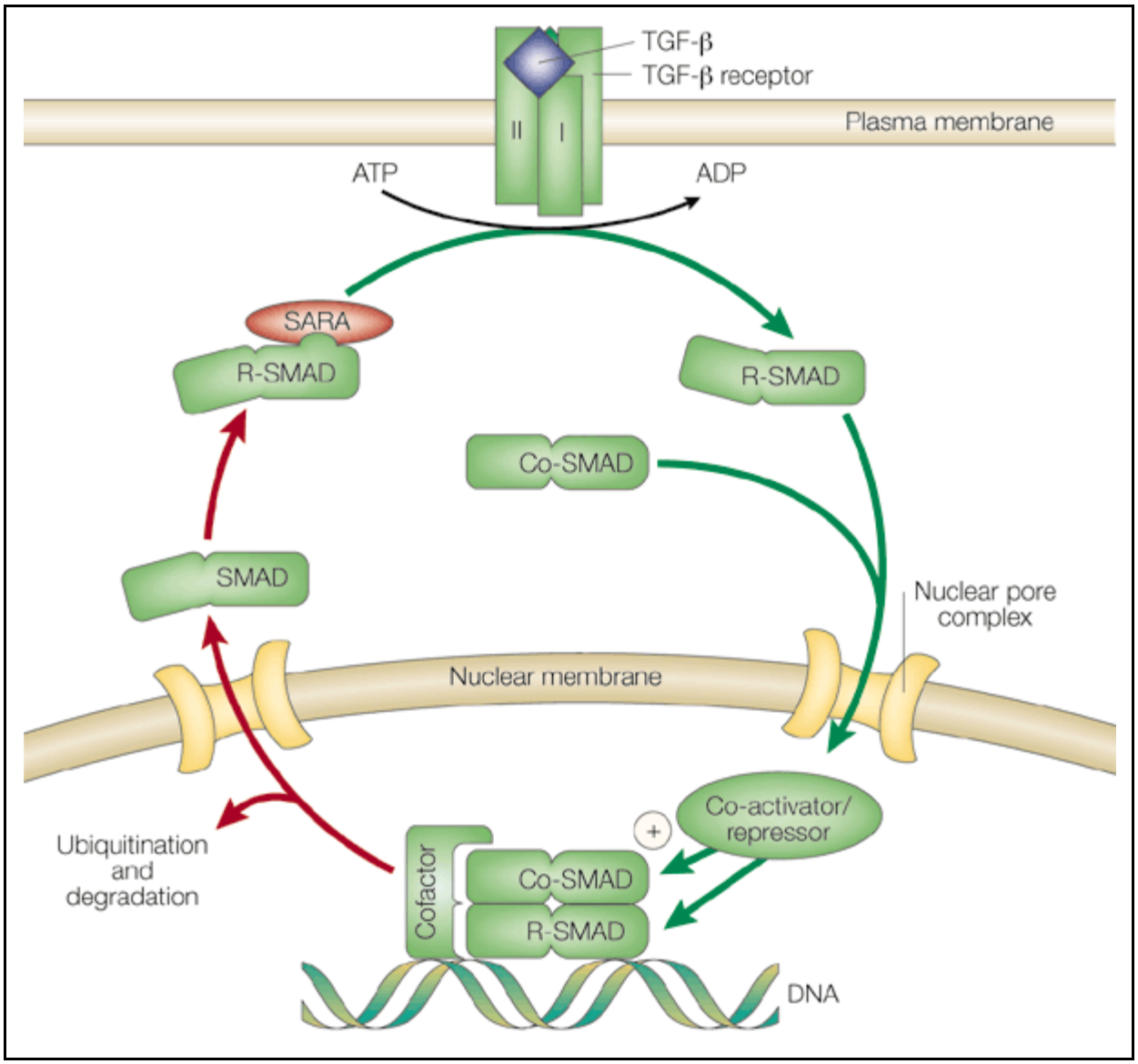
Muscle RAS oncogene homolog (*MRAS*), one of the genes of the Ras family, was reported to be engaged in regulating cell-cell adhesion via intracellular adhesion molecules (ICAMs).[159] It was reported to bind to different downstream effectors of the RAF subfamily, leading to the inhibition of activation of the transcription factor FOS in a competitive manner.[160] Mutant *MRAS* was more recently reported to induce epithelial-mesenchymal transition (EMT) and tumorigenesis.[161]

The downstream effector of RAS proteins, v-raf murine sarcoma viral oncogene homolog (A-RAF) activates the MAP kinase kinase MEK1 in epidermal growth factor-stimulated HeLa cells in the classical RAS signaling cascade.[162] Additionally, A-RAF was shown to bind directly to phosphatidylinositol 3-kinase (PI3K), thus the A-RAF kinase also interacts with the G-protein coupled receptor signaling pathway.[163] As a third role, A-RAF was also reported to interact specifically with two novel human proteins, referred to as hTOM and hTIM, which are similar to components of mitochondrial outer and inner membrane protein-import receptors from lower organisms.[164] A-RAF was detected in purified mitochondrial fractions of cells, suggesting that a proportion of A-RAF is present in the inner matrix compartment of mitochondria. While a current hypothesis exists that the major effector kinase RAF1 is involved in apoptotic signaling through its association with BCL2 and mitochondrial outer membrane, a similar mode of action for A-RAF could be implied by the finding.

TGF-β Signaling Pathway

The general TGF-*β* signaling pathway includes TGF-*β* receptor proteins, which tetramerize upon binding of their ligands (BMP or TGF-*β*) and activate R-SMAD transcription factors by phosphorylation. Activated R-SMADs dimerize with Co-SMAD proteins to form a transcription factor that is able to enter the nucleus and bind DNA (Figure 28).[165] The transcription is further regulated by site-specific co-factors as well as co-activator and co-repressor proteins that bind to the SMAD-DNA complex and mediate or inhibit the transcription of target genes. Members of TGF-*β* signaling family protein genes were identified to be significantly over-represented in the signature, e.g. the bone morphogenic ligand *BMP4*, the pseudo-receptor "BMP and activin membrane-bound inhibitor homolog precursor" (*BAMBI*), the inhibitory downstream regulator "Smad ubiquitination regulatory factor 2" (*SMURF2*) as well as the TGF-*β* -induced transcription factor 2 (*TGIF2*), the co-factor *LMO4*, and the co-activators of transcription *EP300*, *CREB* and the proto-oncogene *SRC*.

Figure 28



Basic TGF-*β* pathway. Receptor-regulated SMAD transcription factors (R-SMADs) require transforming growth factor-beta (TGF-*β*) -induced phosphorylation to assemble transcription regulatory complexes with partner SMADs (co-SMADs). R-SMADs can move into the nucleus on their own but, to be accessible to membrane receptors, R-SMADs are tethered in the cytoplasm by proteins such as SARA (SMAD anchor for receptor activation). Receptor activation occurs when TGF-*β* induces the association of two type I and two type II receptors. Both receptor components have a serine/threonine protein kinase domain in the cytoplasmic region. In the basal state, the type I receptor is kept inactive by a wedge-shaped GS region, which presses against the kinase domain, dislocating its catalytic centre. In the ligand-induced complex, the type II receptor phosphorylates the GS domain and this activates the type I receptor, which catalyses R-SMAD phosphorylation. Phosphorylation decreases the affinity of R-SMADs for SARA and increases their affinity for co-SMADs. The resulting SMAD complex is free to move into the nucleus and competent to associate with transcriptional co-activators or co-repressors. SMADs can contact DNA, but effective binding to particular gene regulatory sites is enabled by specific DNA-binding co-factors. R-SMADs that move into the nucleus may return to the cytoplasm, but their ubiquitylation- and proteasome-dependent degradation in the nucleus provide a way to terminate TGF-*β* responses. From J. Massague, 2000.[165]

The transforming growth factor TGF-*β* plays a dual role in its mode of action, as it can both act as mediator of transformation as well as an inhibitor of proliferation. Its dual role is mainly cell-type dependent, and it has been shown to have tumor suppressor activity in early stages of tumorigenesis, while operating as a promotor of tumor cell invasiveness and metastasis in advanced

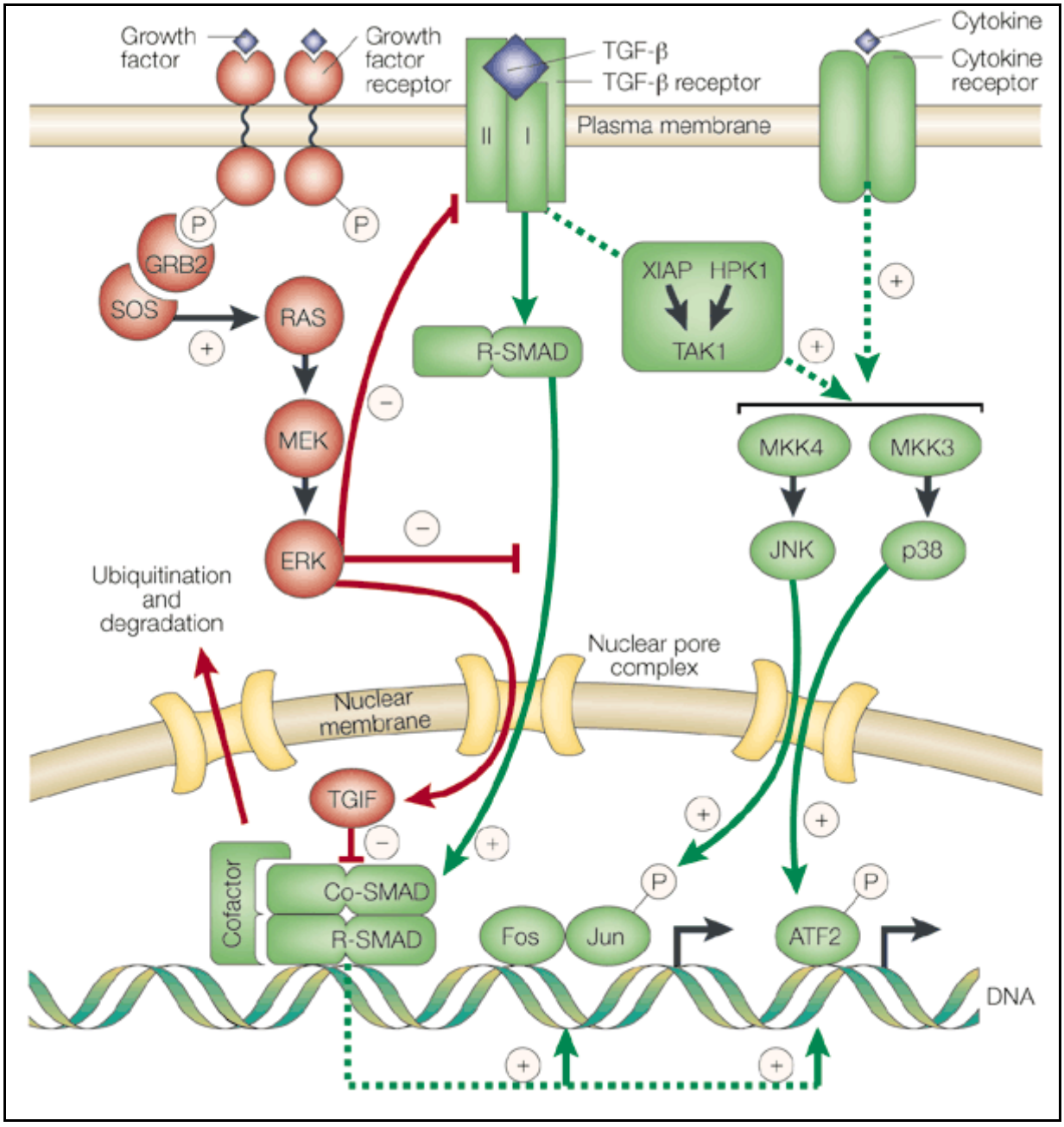
tumors.[165-167] The tumor suppressor activity includes the arrest of the cell cycle in epithelial, endothelial and hematopoetic cells at the early G1 phase via SMAD protein-mediated transcriptional regulation of critical regulators of the cell cycle, e.g. by transcriptionally inhibitory promotor elements leading to repression of c-Myc and CDK4, maintenence of Rb in hypo-P state as well as control of cell-cycle inhibitors (CKIs) such as p15 (Ink4) and p21/p27/p57 (Cip/Kip family) proteins.

TGF-β is also known to induce the epithelial-mesenchymal transition (EMT) in an oncogenic manner, thus enhancing proliferative, migratory, invasive and metastatic potential of the cells. TGF-β thereby acts in an autokrine loop, sustaining its acivity on the invasive cells. On the contrary, closely related BMP proteins fail to elicit EMT, and higher levels of BMP proteins inhibit TGF-β from inducing EMT. BMPs have been shown to be able to reverse EMT and lead to MET. Therefore, the ratio between BMP and TGF-β in tumor cells may be of importance in the decision of migration potential and invasiveness. However, in a cooperative manner, active RAS/RAF signaling further enhances the establishment of EMT, as do PI3K and Rho GTPase signaling.

The transcriptional co-activators or -repressors present in the nucleus facilitate and determine the mode of action of the activated SMAD dimers. Transcriptional activators include CREB binding proteins, EP300 and repressors include TGIFs like TGIF2, among others. The repressors bind histone deacetylases (HDACs), while activators generally act as histone acetyltransferases (HATs). In general, TGIFs bind to SMAD2 and SMAD3 in a competitive manner to EP300, so the relative levels of these transcriptional regulator proteins determine the activating versus repressing mode of action of TGF-β. Thus, TGF-β signaling leads to inactivation of gene expression in many cases.[168] Additionally, there is evidence for cross-talk between RAS/MAP-kinase and the TGF-β pathway in that TGIF2 was shown to be phosphorylated in response to EGF signaling. Another known debranching from the classical TGF-β pathway leads to a cross-talk with the MAP kinase pathways on the level of Jun-amino-terminal kinase (JNK, MAPK8) associated transcription factors, also known as AP-1 family.[169] It was shown that TGF-β harbors the ability to increase the activity of AP-1 (JUN-FOS) complexes through phosporylation by

JNK or the activity of ATF2 transcription factor (which also contains HAT activity) binding to CREB complexes, either resulting in an activation of AP-1 or CREB target genes (Figure 29).

Figure 29



Crosstalk between the SMAD and mitogen-activated protein kinase pathways. The three principal MAPK pathways in mammalian cells may affect the SMAD pathway through various mechanisms. The Ras–MEK–ERK pathway can decrease TGF-*β* receptor levels by controlling expression, attenuate SMAD accumulation in the nucleus by phosphorylating SMADs in the linker region and increase the level of the SMAD corepressor TGIF by stabilizing this protein. The MKK4/JNK and MKK3/p38 pathways, which can be activated by various cytokines, enhance the activity of Jun and ATF2 transcription factors that may cooperate with SMADs through direct physical contacts. In certain cell types and conditions, the MKK4/JNK and MKK3/p38 pathways are reportedly activated by TGF-*β* itself, and the proteins XIAP, HPK1 and TAK1 might be involved in this link. The direct nature and physiological relevance of these interactions remain to be established. (ATF2, activating transcription factor 2; ERK, extracellular-signal-regulated kinase; GRB2, growth factor receptor-binding proteins 2; JNK, Jun amino-terminal kinase; XIAP, Xenopus inhibitor of apoptosis; HPK1, haematopoietic progenitor kinase 1; TAK1, TGF-*β* -activated kinase; MAPK, mitogen-activated protein kinase; MKK, MAPK kinase; R-SMAD, Receptor-regulated SMAD transcription factors; sos, son of sevenless; TGF-*β*, transforming growth factor-beta.) From J. Massague, 2000.[165]

The receptor protein BAMBI acts as a negative regulator in the TGF-β signaling pathway. It is located in the plasma-membrane of cells, has high homology to the BMP receptor BMPR1B, but lacks an intracellular serine/threonine kinase domain required for signaling. As the BMP receptor proteins are required to tetramerize upon BMP or TGF-β signaling to perform their activating phosphorylation function, the BAMBI protein acts as a so-called pseudo-receptor, inhibiting downstream signaling. Expression of *BAMBI* was described by Sekiya *et al.* to be upregulated by TGF-β/BMP signaling-mediated activation of the transcription factor SMAD3/4 dimer, thus acting in a negative feedback loop.[170] *BAMBI* expression was also found by the authors to be elevated in colorectal and hepatocellular cancers.

In 2001, Visvader *et al.* explored a role for *LMO4*, initially described as a human breast tumor autoantigen, in developing mammary epithelium and breast oncogenesis.[171] The gene was expressed predominantly in the lobuloalveoli of the mammary gland during pregnancy. Consistent with its role in proliferation, forced expression of this gene inhibited differentiation of mammary epithelial cells. Overexpression of *LMO4* mRNA was observed in 5 of 10 human breast cancer cell lines. Moreover, in situ hybridization analysis of 177 primary invasive breast carcinomas revealed overexpression of *LMO4* in 56% of specimens. Immuno-histochemistry confirmed overexpression in a high percentage (62%) of tumors. These studies implied a role for *LMO4* in maintaining proliferation of mammary epithelium and suggested that deregulation of this gene may contribute to breast tumorigenesis.

LMO proteins act as transcription factor regulators, do not bind to DNA directly but associate with other transcription factors (CLIM1 and especially CLIM2 for *LMO4*). LMO4 expression was also reported to be associated with a worse prognosis and overexpression in mammary glands of mice led to inhibition of mammary gland development, hyperplasia and intraepithelial neoplasia.[172,173] Identifying LMO4 as a transcription regulation factor that does not bind directly to DNA but other transcription factors, Ning Wang *et al.* recently identified BMP7 protein (a BMP4 homologue) as one of the highly significant target genes of *LMO4* regulation through recruitment of the histone deacetylase HDAC2 to the binding site.[172,174] Other upregulated genes include e.g. AKT1, RHOB, SMAD5 and TGFBRAP1 while among the downregulated genes were e.g.

MBD1 (methyl-CpG binding domain protein 1), RERG (RAS-like estrogen-regulated growth inhibitor) and RRAS. Markedly, the only statistically significant Gene Ontology process enriched in the set of deregulated genes was apoptosis. In the study presented by Ning Wang *et al.*, BMP7 decreased proliferation and induced apoptosis. In 2003, a different role for LMO4 was suggested by Sutherland and co-workers as a BRCA1-interacting protein, repressing its transcriptional activity.[175] The authors concluded that the high expression of LMO4 in sporadic breast cancers may alter the stoichiometry of BRCA1 expression, leading to an inhibition of its tumor-suppressing function.

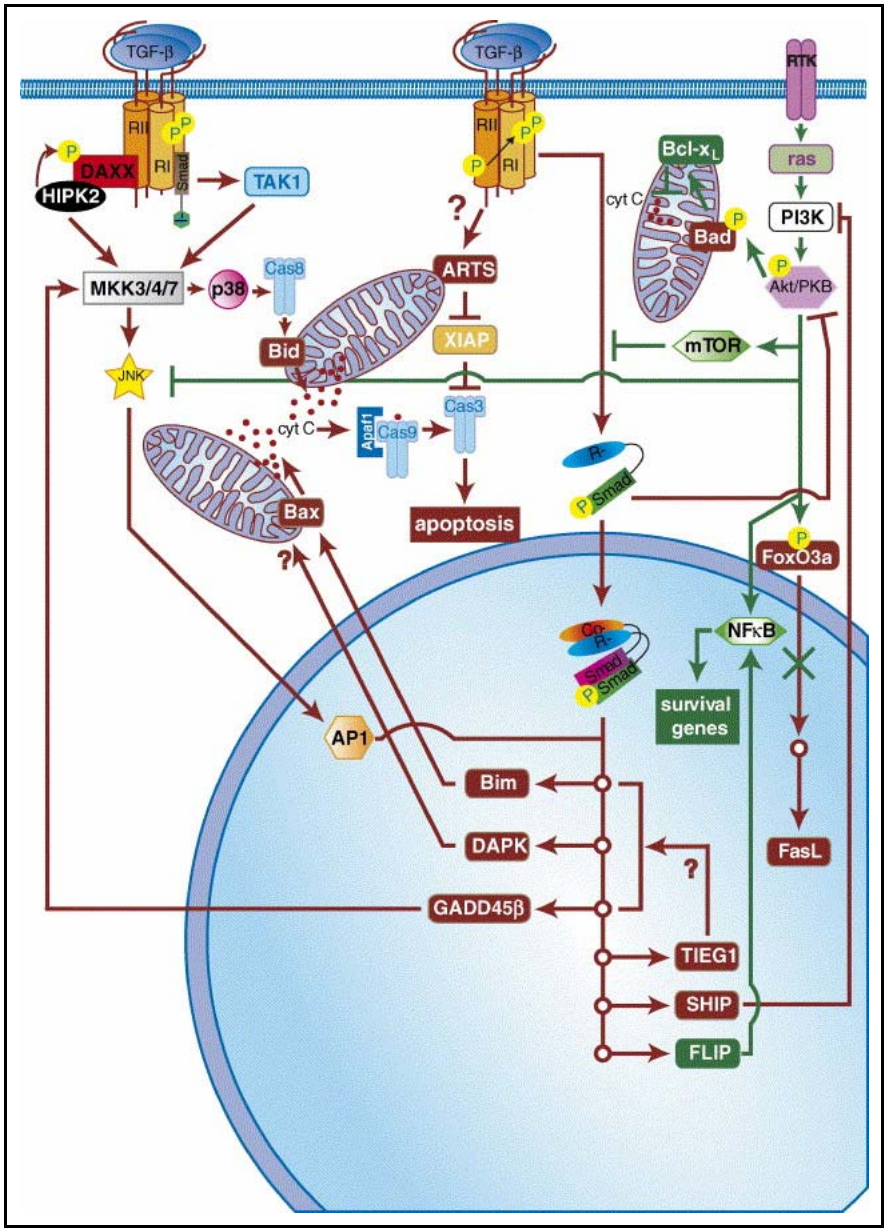
Regulation of Apoptosis

Among the genes contained in the signature, apoptosis regulation seemed to play a particular role in classifying patient response groups. While cells that undergo apoptosis can be triggered for the programmed cell death by either of two pathways, it seemed striking that only those of the mitochondria-related apoptosis mechanism were contained, but not genes from the mechanism regulated by the caspase protein signaling cascade.

The signature genes involved in apoptosis regulation included the death-associated protein kinase 2 (*DAPK2*), death-induced protein (*DIP*), BCL2-antagonist/killer 1 protein *BAK1* and other mitochondrial proteins as well as the rapamycin-associated protein genes *KIAA1303* (*RPTOR*) and *FRAP1*.

Whether DAPK proteins regulate the activation of mitochondrial apoptosis program via BAX and BAK1 upon TGF-β signaling, as proposed by Pardali and Moustakas (Figure 30), remains unclear.[167] However, a direct interaction of the TGF-β pathway with the proteins of the AKT/mTOR pathway was demonstrated by association of SMAD3 with FRAP1 (also called mammalian target of rapamycin, mTOR), in which FRAP1 suppresses the phospho-activation of SMAD3.[176] The authors proposed a model of an AKT kinase-dependent inhibition of SMAD3 through FRAP1, and a resulting loss of tumor suppression by TGF-β in cancer. In 2007, Creighton discovered a link between gene expression patterns derived from overexpression of AKT in mouse-models and human breast cancer gene expression studies by meta-analysis.[177] Genes upregulated by AKT and dependent on FRAP1 activity were associated with poor prognosis in these studies.

Figure 30



The apoptotic response program. TβRII in the TGF-β receptor complex directly binds DAXX, which recruits HIPK2 and becomes phosphorylated by HIPK2, leading to activation of MKK3/4/7. The same kinases can be activated by TAK1 which is activated by Smad7 (I-Smad) bound to the receptor complex. MKKs then phosphorylate and activate JNK or p38 MAPKs. JNK activates the AP-1 transcriptional complex, leading to induction of pro-apoptotic genes (red circular nodes) in cooperation with Smads. p38 activates caspase-8 (Cas8), which activates the pro-apoptotic factor Bid, leading to cytochrome C (cyt C) release and activation of the apoptosome (cyt C/Apaf1/Caspase-9 (Cas9) complex), which activates caspase-3 (Cas3) and executes apoptosis. The TGF-β receptor complex signals by unknown mechanism (?) to mitochondrial ARTS, which inhibits XIAP, the inhibitor of caspase 3, leading to apoptosis. The activated nuclear Smad complex induces transcription of pro-apoptotic genes such as Bim, DAPK, GADD45β, TIEG1, and SHIP. Bim activates the pro-apoptotic Bax, which leads to cytochrome C release and caspase activation. DAPK modulates the action potential of the mitochondrial membrane and induces apoptosis via yet unknown molecular mechanisms (?). GADD45β interacts with and activates MKK4, thus activating the p38 pro-apoptotic pathway. TIEG1 is a transcription factor that regulates additional pro-apoptotic genes, but it is not clear whether these include those listed in the figure (?). SHIP inhibits PI3K. Smad3 can directly interact and inhibit the activity of Akt/PKB in addition to transactivating pro-apoptotic target genes. Activated Smads also induce expression of the pro-survival factor FLIP, which exits the nucleus and activates the transcriptional activity of NF-κB, thus inducing the expression of other anti-apoptotic factors. Growth factors signaling via receptor tyrosine kinases (RTK) activate the Ras/PI3K/Akt/PKB pathway. Akt phosphorylates the pro-apoptotic protein Bad, thus activating the anti-apoptotic protein Bcl-xL, which blocks cytochorome C release. Akt also activates FRAP1 (mTOR), which inhibits R-Smad phosphorylation by the TGF-β receptor complex, and directly inhibits the pro-apoptotic JNK, while activating the pro-survival NF-κB pathway. In addition to NFκB Akt phosphorylates the pro-apoptotic transcription factor FoxO3a, leading to its cytoplasmic retention and transcriptional inactivation of its target pro-apoptotic genes such as Fas ligand (FasL). All pro-apoptotic events are shown in dark red and all pro-survival events are shown in green. From Pardali & Moustakas, 2006.[167]

DNA Damage Response

Although representing a smaller group of genes in the signature as the others described before, the gene activities involved in perception and regulation of DNA damage appeared to be significantly altered (Table 19). Signature genes belonging to these pathways were *BRAP* and *RAD51C*, two genes associated with the BRCA1 protein, *TP53BP1* and *TP53RK*, which are both associated with the tumor suppressor TP53, and *CHEK2*, a cell cycle regulator important for checking DNA damage before entering replication.

The BRCA1-associated protein BRAP was identified by its ability to bind to the nuclear localization signal of BRCA1 and to regulate nuclear targeting by retaining proteins with a nuclear localization signal in the cytoplasm. In 2004, a direct association of BRAP (also named IMP) with the RAS pathway was reported and its function as an ubiquitin E3 ligase for a RAF/MEK1 complex inhibitor was discovered.[178] Whether the interaction of BRCA1 with BRAP has to do with its transcription factor or the DNA-damage dependent function of BRCA1 is not known.

RAD51C is known to be involved in the homologous recombinational repair pathway of damaged DNA and in meiotic recombination. However, the gene coding for this protein is located on a region of chromosome 17q23 where amplification occurs frequently in breast tumors. It is therefore unclear, which effect is causal for the finding of highly expressed RAD51C transcript.

Tumor protein p53 binding protein 1, TP53BP1, is clearly associated with activation of ATM in response to DNA double strand breaks.[179] It binds to ATM as well as to TP53 protein. TP53RK, or TP53 regulating kinase, is known to activate TP53, but apart from one report claiming a binding of the protein to HER2, very little is known about its upstream signaling.[180]

Expression of the Signature Genes in Perspective of Pathways

In order to evaluate the measured expression in the female breast cancer studies in respect to the response of the patients to the chemotherapy given, it was necessary to define the view point of the gene expression regulation between the two response classes. Here, the expression ratios between pCR and non-responder patients were estimated, and should be taken as up- or downregulated in pCR patients *versus* non-pCR patients (Table 23).

As the non-responder patients were comprised of patients with pathologically no change as well as partial remission, and the latter in a much larger number, these ratios should not be taken as absolute truth. However, it was expected that the genes could provide valuable information about the differences and maybe even the cause of response to chemotherapy.

Table 23 **RNA Expression of Signature Genes and BRCA1**

Gene Symbol	Oligo ID	$\log_n$ (pCR)	$\log_n$ (non-pCR)	$\log_n$ (pCR/non-pCR)
A-RAF	OL006360	-0.56	-0.18	-0.378
MRAS	OL020413	1.81	1.39	0.416
RASAL1	OL014803	0.67	0.10	0.577
RHEB	OL017532	-0.71	-1.04	0.325
BAMBI	OL006551	-0.93	-1.59	0.657
BMP4	OL005655	-1.49	-0.94	-0.550
SMURF2	OL014622	0.48	-0.07	0.554
CREB3	OL009199	0.20	0.24	-0.035
EP300	OL003498	0.36	0.55	-0.194
SRC	OL014799	0.84	0.58	0.265
LMO4	OL000848	1.32	0.39	0.932
DAPK2	OL012804	2.32	1.88	0.441
FRAP1	OL019848	-1.04	-0.83	-0.212
RPTOR	OL002920	-0.33	-0.16	-0.173
BAK1	OL010567	0.19	-0.10	0.292
BRAP	OL012396	-0.73	-0.79	0.053
RAD51C	OL001998	-1.16	-1.50	0.341
BRCA1	OL014601	-1.94	-2.03	0.086
TP53BP1	OL007925	-0.45	-0.34	-0.112
TP53RK	OL017620	-0.38	-0.26	-0.120
CHEK2	OL013553	-0.71	-1.07	0.359

As the literature proposed, it may be suggested that the major difference between patients not responding to the chemotherapy to the responders is given by the fact that activation of RAS was shown to induce EMT. In the patients investigated here, the effector kinase MRAS shows a higher expression in the responders, leading to the conclusion that responder patients might have a higher induction of tumor cell transformation. However, the signal transducing protein A-RAF shows the direct opposite expression ratio, with a higher expression in the non-responder patients. It is therefore unclear,

whether a stronger activation of the Ras pathway and consequently a higher induction of EMT in responder patients could be reasoned by these results.

The transcription regulator LMO4, acting most probably as inhibitor of differentiation, is strongly upregulated in responder patients, which would fit very well to the fact that undifferentiated cells are more likely to respond to chemotherapy due to their proliferative turnover, which is targeted by the chemotherapeutic regimen. As the interactions of the TGF-β signaling and Ras signaling were described, the combination of the findings here would lead to the conclusion that the responder patients had tumors cells which were proliferating more and had a lower grade of differentiation. This is very well in accordance with the KI67 measurements, as seen by IHC (Figure 27), and with the findings in the literature that LMO4 maintains cells in a proliferative state, and tumors with high LMO4 expression have a worse prognosis.
While the stronger activation of BMPs was seen as generating the opposite effect and rather maintain cells in the differentiated state according to the literature, the finding that BMP4 was also upregulated in responder patients must not be in disagreement. It is possible to postulate that the highly proliferative cells try to find the balance and thus counteract by expressing BMPs. The results of a higher expression of *BAMBI* and *SMURF2* in the responder tumors, both representing inactivating feedback loops in the TGF-β signaling, supports this idea.

As Pao *et al.* reported, the activating co-regulators of TGF-β -induced transcription EP300 and CREB interact with BRCA1 and activate its transcription.[181] Here, a clear difference in the expression of these coactivators could not be seen between response classes, and there were also no differences in the expression of BRCA1. However, the highly probable dependencies between BAMBI and BRCA1, as well as between SRC and BRCA1, as seen in the linear model analysis, imply a strong positive association of the TGF-β pathway signaling with the mechanism of DNA damage response.

The upregulation of the DNA damage-related genes *RAD51C* and *CHEK2* measured in responder patients could not be explained by a more proliferative

tumor tissue. It may be possible that due to high proliferation or other mechanisms, the DNA of these tumor cells is unstable. However, for TP53BP1 and TP53RK, the proteins associated with the tumor suppressor TP53, the gene expression were measured as slightly downregulated in responder patients, again supporting the idea of a more proliferative tumor tissue in these patients.

Concerning the genes differentially regulated between patient classes in the apoptosis pathways, it is difficult to decide, whether the DNA and microtubule damages triggered by the chemotherapy drugs have an additive effect to the pre-therapeutic results presented here or not. However, as the transcription of *DAPK2* and its possible downstream pathway molecule *BAK1* showed a very clear differentiation between both patient groups, with an upregulation in responder tumors, it is explainable that these cells can be killed more effectively by chemotherapeutic intervention.

The results presented here for FRAP1 (mTOR), and the associated scaffolding protein KIAA1303 (RPTOR) do not conform well to the published data. As FRAP1 activation was reported to repress TGF-β-dependent tumor suppression, and a dependence of genes predicting poor prognosis on the activation of FRAP1 was shown, it is unclear how the downregulation of both proteins in responder patients could be matched with these reports.

5.4. Antibody Generation using Mouse Hybridoma Cells

As there was no antibody available against the BAMBI protein at the time when the gene expression signature was compiled, a validation of the genetic results on protein expression was impossible. However, as the protein seemed to be an interesting candidate for a more detailed analysis, a good antibody was required. In order to generate it, a truncated form of the BAMBI protein was expressed in *E.coli* cells, containing only the unique cytoplasmatic and transmembrane protein domains, and isolated using a His-tag.

Hybridoma cells, generated using plasma cells from mice vaccinated with this truncated BAMBI, were grown in cell culture, and conditioned media of different cell clones were tested. These antibody containing media were shown to have a sufficient sensitivity for detection of BAMBI, as tested by staining of Western blots containing the full length protein, again generated using *E.coli* cells (Figure 26). For a thorough test of the specificity of these antibodies and evaluating them in the native state, an expression of the BAMBI protein was necessary in eukaryotic cell cultures.

However, it was not possible to express this gene in such cells in culture, although they showed sufficient transfection efficiencies with the vectors containing GFP control protein. When human culture cells from different tissue origins were transfected with either GFP-BAMBI or GFP-BMPR1B (negative control) fusion proteins, they showed no fluorescence and an elevated apoptosis rate, so the expression could not be observed (data not shown).

Therefore, it was impossible to test for specificity of the antibodies from the conditioned media, and an appropriate antibody could not be obtained. At this point, a commercially produced monoclonal antibody had become available, and the generation of an antibody against BAMBI protein was abandoned.

5.5. Immuno-histochemical Analysis of Tumors

In order to validate the results that were generated with gene expression profiling on microarrays and with RQ-PCR, a suitable number of genes was selected to be analyzed with immuno-histochemical staining on tissue sections from the same patient tumor samples. For this purpose, four consecutive tumor sections of 80 patients could be obtained. As the number of selected proteins was six, however, two of these sections were required to be stained with two antibodies each.

For data analysis, sections for each patient and antibody were interpreted in staining intensity (score 0-3) and the percentage of tumor cells, for three representative areas, if available. Data for ER and PgR scores, TP53, KI67, HER2 and BCL2 were provided by the pathology department of the university clinic. The generated data were then analyzed for the patient classes. Using a linear model prediction algorithm, the immuno-histochemical stainings were analyzed for associations among each other and in comparison with clinical data of the patients: tumor grading, pathological outcome and response of the patients, as well as the therapy administered to the patients.

Due to the small sample size, these associations were considered statistically significant only if the p-value was below 0.01, even though a p-value below 0.05 was considered an association. As expected, significant associations were seen between ER and PgR scores. Very good associations were also seen between response (pCR *versus* non-pCR) and both ER and PgR, BCL2 and tumor grading (G1-2/G3), BRCA1 and BAMBI as well as BAMBI and therapy (GEDoc/GEsDoc). The latter seemed not explainable, as the therapy was administered after the acquisition of the tumor biopsies analyzed here and as both therapies were considered equally effective. However, due to the small sample size especially in the case of pCR patients in both therapy studies, such a finding might be explained by minor accidental differences between patients, or the ability of the linear model algorithms to detect small but consistent differences between the groups. However, as the majority of other parameters and stainings were far from such a highly significant p-value, a connection between differences in the patients achieving pCR concerning the BAMBI expression levels of their tumors and the therapy schedule administered cannot be ruled out.

Statistically not as significant associations were shown between LMO4 and each of BMP4, SMAD3 or SRC proteins, but as these belong to the same signaling pathway, this was considered a probable result. More astonishing in this respect seemed the fact that LMO4 was exclusively associated with BMP4, but none of the other TGF-β proteins.
The protein markers best associated with the response of patients were ER, PgR, HER2, BCL2 and LMO4, in decreasing order.

To illustrate the extent of relative differences in the protein expression, according to their IHC staining percentages, these were depicted as box-plots (Figure 27). In case of TP53, the box plot seems misleading: Due to the fact that patients either showed a very high or low (or no) expression of TP53 protein, the statistical analysis here was restricted by the different sizes of the response groups. As the non-pCR group was large (n=52), the relatively few highly TP53-positive patients appear as outliers, while the much smaller pCR group (n=17) appears to have a wider dynamic range, thus the high percentage tumors are represented as the 3rd quartile. In fact, only one patient shows intermediate expression of TP53 (40%), while the others show values of either ≥90% or ≤20%. This circumstance is represented by the medians in both groups, which show similar values.
In general, for most protein stainings, the box plots showed only lesser differences between the patient classes, except for KI67 and LMO4. This reflects the results from the microarray gene expression study only in part, as some genes show a markedly higher difference in their RNA expression levels, as in the cases of BAMBI and BMP4 (Table 23). A significant difference between IHC staining and gene expression data, as measured both by microarrays and RQ-PCR (Figure 23), is seen in the lower median expression of LMO4 protein in the pCR group. However, it should be taken into consideration that only samples from 80 patients were available for IHC staining as compared to the microarrays (n=100), that the number of pCR patients was small in respect to the non-pCR group, that the staining of SMAD3 and BRCA1 were preliminary due to double staining evaluation difficulties and that the protein expression could be regulated post-transcriptionally as well as post-translationally.

The deviations seen for some factors between the protein expression, as measured by IHC, and the gene activity, as measured by microarray and RQ-PCR, seem to make the interpretation of the results uncertain. However, given the fact that for all but one of the gene expression values, these correspond very well between RQ-PCR and the microarray results, these can be considered absolutely valid. Therefore, the RNA expression experiments used for the prediction of therapy response and pathway analysis are validated and undisputed.

The differences between the protein expression and the RNA expression, as seen in this study, could be explained: Firstly, the mechanisms of post-transcriptional regulation, like e.g. RNA transport from the nucleus, modifications to or even degradation of the mRNA (editing, silencing and interference mechanisms) as well as the translation into proteins have long been known.[2,182] Secondly, such differences between the expression levels of RNA and protein, especially in cancer, have been reported, for example by the oncogenic deregulation of RNA translation into protein by phosphatidylinositol-3-kinase (PI3K), which has been identified as required for the transformation cells to and has been reported to be dependent on mTOR and RHEB.[183]

Whether the proteins of the genes, for which a deregulation between RNA expression and protein expression in this study is seen, are activated or deactivated by such a mechanism, could not be elucidated here. A much more detailed protein analysis, which also includes analysis of post-translational modifications such as phosphorylation, would be needed for this question to be answered.

6. Outlook

In the thesis presented here, the method of microarray expression profiling analysis using long oligonucleotide DNA probes was applied for the identification of a gene signature predicting response to chemotherapy in breast cancer, starting from small tumor biopsies taken at time of diagnosis. For this aim, a novel procedure to amplify and label mRNA from small RNA sources for hybridization to oligonucleotide microarrays had to be developed and introduced, and had also proven applicable. However, the input minimum of 2 µg of total RNA limits its usefulness, and thus has been improved to 500 ng in the meantime. Further improvements to the minimum input amount are currently only possible through a repetition of the amplification steps, resulting in a further fragmentation and shortening of the labeled nucleic acid chains yielded and therefore a loss of dynamic range of the expression profiles. Protocols using two-round amplification are currently being investigated for the exploitation of microdissected cells retrieved from freshly frozen or paraffin-embedded tissue sections to be used for oligonucleotide microarrays. A further reduction of total RNA input amounts in a single round amplification, below 500 ng, could improve the quality of the results obtained in such experiments and at the same time make more patient samples available for investigation using whole-genome expression analysis with long oligonucleotide microarrays.

The signature identified for prediction of the pathologic complete remission of primary breast tumors after neoadjuvant application of gemcitabine, epirubicin and docetaxel in a tri-fold chemotherapy regimen is comprised of 512 genes, and displays a higher sensitivity and specificity than the classical markers currently used in the clinic.

In a new study pursuing the one presented here, other chemotherapy regimens for the neoadjuvant therapy of primary female breast cancers are tested in the same manner to identify such predictive gene expression signatures. This study is comprised of two different treatment arms, is performed in a double-blind setup, and contains a similar number of patients in each arm as in the study presented here. In addition to the task of providing predictive gene signatures for each of the treatment arms, the new study will also allow for an evaluation of the gene signature presented here in its general applicability of

predicting response of female breast cancer to chemotherapy treatment. In this way, the genes responsible or predictive for a general response to chemotherapy may possibly be identified as well as the genes predictive or responsible for the individual chemotherapeutic drugs. The recent finding of Bonnefoi *et al.*, who successfully combined genes into a predictive signature based on individual gene signatures derived from cell culture experiments with single drugs, encourages this point of view.[156] Such cell line experiments could also be performed for comparison with the gene signatures presented here and identified in the pursuing study.

Some of the genes within the predictive signature, that classifies patients according to their therapy response as discussed in this thesis, were identified to be functionally related. Upon analyzing these genes for possible pathway interactions, several genes from the TGF-β and Ras signaling pathways as well as genes involved in DNA damage response and apoptosis were identified. These genes could possibly represent pathways that are not only functionally related to the response of the patients, but also to the development of breast cancers and its different tumor types. A closer investigation of the relationships between these pathways and the development of breast tumors might be very helpful in understanding the heterogeneity of breast cancer patients. Additionally, they could provide new drug targets and be good candidates for further improvements of existing targeted therapies. As some of these had already been identified as genes or proteins related to breast cancer, as e.g. BAMBI and LMO4, they should be investigated in greater detail.

"What we observe is not nature itself,
but nature exposed to our method of questioning."

Werner Heisenberg

7. References

1. Smigal, C. Cancer Statistics 2006, A Presentation From the American Cancer Society. in *American Cancer Society* (2006).

2. Alberts, B. et al. (eds.). *Molecular Biology of the Cell*, (Garland Science, New York and London, 2002).

3. Knudson, A.G., Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820-3 (1971).

4. Knudson, A.G. Hereditary cancer: two hits revisited. *J Cancer Res Clin Oncol* **122**, 135-40 (1996).

5. Vogelstein, B. & Kinzler, K.W. The multistep nature of cancer. *Trends Genet* **9**, 138-41 (1993).

6. Al-Hajj, M., Wicha, M.S., Benito-Hernandez, A., Morrison, S.J. & Clarke, M.F. Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci U S A* **100**, 3983-8 (2003).

7. Pardal, R., Clarke, M.F. & Morrison, S.J. Applying the principles of stem-cell biology to cancer. *Nat Rev Cancer* **3**, 895-902 (2003).

8. Al-Hajj, M. & Clarke, M.F. Self-renewal and solid tumor stem cells. *Oncogene* **23**, 7274-82 (2004).

9. Stingl, J. & Caldas, C. Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. *Nat Rev Cancer* **7**, 791-9 (2007).

10. Aguirre-Ghiso, J.A. Models, mechanisms and clinical evidence for cancer dormancy. *Nat Rev Cancer* **7**, 834-46 (2007).

11. Brackstone, M., Townson, J.L. & Chambers, A.F. Tumour dormancy in breast cancer: an update. *Breast Cancer Res* **9**, 208 (2007).

12. Demicheli, R. Tumour dormancy: findings and hypotheses from clinical research on breast cancer. *Semin Cancer Biol* **11**, 297-306 (2001).

13. Demicheli, R. et al. Breast cancer recurrence dynamics following adjuvant CMF is consistent with tumor dormancy and mastectomy-driven acceleration of the metastatic process. *Ann Oncol* **16**, 1449-57 (2005).

14. WHO. Fact sheet N°310, The Top 10 Causes of Death. in *WHO Fact sheets* (2007).

15. Ries LAG, M.D., Krapcho M, Mariotto A, Miller BA, Feuer EJ, Clegg L, Horner MJ, Howlader N, Eisner MP, Reichman M, Edwards BK (eds). SEER Cancer Statistics Review, 1975-2004, based on November 2006 SEER data submission. (National Cancer Institute. Bethesda, 2007).

16. Urruticoechea, A., Smith, I.E. & Dowsett, M. Proliferation marker Ki-67 in early breast cancer. *J Clin Oncol* **23**, 7212-20 (2005).

17. Fisher, B. et al. Prevention of invasive breast cancer in women with ductal carcinoma in situ: an update of the national surgical adjuvant breast and bowel project experience. *Semin Oncol* **28**, 400-18 (2001).

18. Schwartz, G.F., Solin, L.J., Olivotto, I.A., Ernster, V.L. & Pressman, P.I. Consensus Conference on the Treatment of In Situ Ductal Carcinoma of the Breast, April 22-25, 1999. *Cancer* **88**, 946-54 (2000).

19. Kuschel, B., Aigner, M. & Kiechle, M. Das hereditäre Mammakarzinom. *Der Onkologe* **8**, 790-6 (2002).

20. Hemminki, K., Ji, J. & Forsti, A. Risks for familial and contralateral breast cancer interact multiplicatively and cause a high risk. *Cancer Res* **67**, 868-70 (2007).

21. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet* **358**, 1389-99 (2001).

22. Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group. *Br J Cancer* **83**, 1301-8 (2000).

23. Domchek, S.M. & Weber, B.L. Clinical management of BRCA1 and BRCA2 mutation carriers. *Oncogene* **25**, 5825-31 (2006).

24. Antoniou, A.C. & Easton, D.F. Models of genetic susceptibility to breast cancer. *Oncogene* **25**, 5898-905 (2006).

25. Oldenburg, R.A., Meijers-Heijboer, H., Cornelisse, C.J. & Devilee, P. Genetic susceptibility for breast cancer: how many more genes to be found? *Crit Rev Oncol Hematol* **63**, 125-49 (2007).

26. Antoniou, A.C. et al. A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. *Br J Cancer* **86**, 76-83 (2002).

27. Rahman, N. et al. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet* **39**, 165-7 (2007).

28. Renwick, A. et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat Genet* **38**, 873-5 (2006).

29. Curtin, N.J. PARP inhibitors for cancer therapy. *Expert Rev Mol Med* **7**, 1-20 (2005).

30. Zaremba, T. & Curtin, N.J. PARP inhibitor development for systemic cancer targeting. *Anticancer Agents Med Chem* **7**, 515-23 (2007).

31. Kühn, T., Helms, G. & Kreienberg, R. Operative Behandlung des Mammakarzinoms. *Der Onkologe* **8**, 837-46 (2002).

32. von Minckwitz, G. & Kaufmann, M. Primäre und adjuvante systemische Therapie beim operablen Mammakarzinom. *Der Onkologe* **8**, 847-52 (2002).

33. Polychemotherapy for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group. *Lancet* **352**, 930-42 (1998).

34. Fifth main meeting of the Early Breast Cancer Trialist's Collaborative Group. in *EBCTCG* (Oxford, 2000).

35. Tamoxifen for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group. *Lancet* **351**, 1451-67 (1998).

36. Tamoxifen for early breast cancer. *Cochrane Database Syst Rev*, CD000486 (2001).

37. Honkoop, A.H., Wagstaff, J. & Pinedo, H.M. Management of stage III breast cancer. *Oncology* **55**, 218-27 (1998).

38. Fisher, B. et al. Effect of preoperative chemotherapy on local-regional disease in women with operable breast cancer: findings from National Surgical Adjuvant Breast and Bowel Project B-18. *J Clin Oncol* **15**, 2483-93 (1997).

39. Kaufmann, M. et al. International expert panel on the use of primary (preoperative) systemic treatment of operable breast cancer: review and recommendations. *J Clin Oncol* **21**, 2600-8 (2003).

40. Howell, A. et al. Results of the ATAC (Arimidex, Tamoxifen, Alone or in Combination) trial after completion of 5 years' adjuvant treatment for breast cancer. *Lancet* **365**, 60-2 (2005).

41. Ellis, M.J. et al. Letrozole is more effective neoadjuvant endocrine therapy than tamoxifen for ErbB-1- and/or ErbB-2-positive, estrogen receptor-positive primary breast cancer: evidence from a phase III randomized trial. *J Clin Oncol* **19**, 3808-16 (2001).

42. Kaufmann, M. et al. Improved overall survival in postmenopausal women with early breast cancer after anastrozole initiated after treatment with tamoxifen compared with continued tamoxifen: the ARNO 95 Study. *J Clin Oncol* **25**, 2664-70 (2007).

43. Land, S.R. et al. Patient-reported symptoms and quality of life during treatment with tamoxifen or raloxifene for breast cancer prevention: the NSABP Study of Tamoxifen and Raloxifene (STAR) P-2 trial. *Jama* **295**, 2742-51 (2006).

44. Vogel, V.G. et al. Effects of tamoxifen vs raloxifene on the risk of developing invasive breast cancer and other disease outcomes: the NSABP Study of Tamoxifen and Raloxifene (STAR) P-2 trial. *Jama* **295**, 2727-41 (2006).

45. Bear, H.D. et al. The effect on tumor response of adding sequential preoperative docetaxel to preoperative doxorubicin and cyclophosphamide: preliminary results from National Surgical Adjuvant Breast and Bowel Project Protocol B-27. *J Clin Oncol* **21**, 4165-74 (2003).

46. Bear, H.D. et al. Sequential preoperative or postoperative docetaxel added to preoperative doxorubicin plus cyclophosphamide for operable breast cancer:National Surgical Adjuvant Breast and Bowel Project Protocol B-27. *J Clin Oncol* **24**, 2019-27 (2006).

47. Schneeweiss, A. et al. Gemcitabine, epirubicin and docetaxel as primary systemic therapy in patients with early breast cancer: results of a multicentre phase I/II study. *Eur J Cancer* **40**, 2432-8 (2004).

48. Slamon, D.J. et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N Engl J Med* **344**, 783-92 (2001).

49. Pegram, M. et al. Inhibitory effects of combinations of HER-2/neu antibody and chemotherapeutic agents used for treatment of human breast cancers. *Oncogene* **18**, 2241-51 (1999).

50. von Minckwitz, G. et al. [Evidence-based recommendations on primary treatment of carcinomas of the breast]. *Zentralbl Gynakol* **124**, 293-303 (2002).

51. Telli, M.L., Hunt, S.A., Carlson, R.W. & Guardino, A.E. Trastuzumab-related cardiotoxicity: calling into question the concept of reversibility. *J Clin Oncol* **25**, 3525-33 (2007).

52. O'Donovan, N. & Crown, J. EGFR and HER-2 antagonists in breast cancer. *Anticancer Res* **27**, 1285-94 (2007).

53. Agus, D.B. et al. Phase I clinical study of pertuzumab, a novel HER dimerization inhibitor, in patients with advanced cancer. *J Clin Oncol* **23**, 2534-43 (2005).

54. de Bono, J.S. et al. Open-label phase II study evaluating the efficacy and safety of two doses of pertuzumab in castrate chemotherapy-naive patients with hormone-refractory prostate cancer. *J Clin Oncol* **25**, 257-62 (2007).

55. Engel, R.H. & Kaklamani, V.G. HER2-positive breast cancer: current and future treatment strategies. *Drugs* **67**, 1329-41 (2007).

56. Gralow, J.R. The role of bisphosphonates as adjuvant therapy for breast cancer. *Curr Oncol Rep* **3**, 506-15 (2001).

57. Diel, I.J. et al. Reduction in new metastases in breast cancer with adjuvant clodronate treatment. *N Engl J Med* **339**, 357-63 (1998).

58. Hillner, B.E. et al. American Society of Clinical Oncology guideline on the role of bisphosphonates in breast cancer. American Society of Clinical Oncology Bisphosphonates Expert Panel. *J Clin Oncol* **18**, 1378-91 (2000).

59. Coleman, R. Potential use of bisphosphonates in the prevention of metastases in early-stage breast cancer. *Clin Breast Cancer* **7 Suppl 1**, S29-35 (2007).

60. Saarto, T., Blomqvist, C., Virkkunen, P. & Elomaa, I. Adjuvant clodronate treatment does not reduce the frequency of skeletal metastases in node-positive breast cancer patients: 5-year results of a randomized controlled trial. *J Clin Oncol* **19**, 10-7 (2001).

61. Pfisterer, J., Hilpert, F., Budach, W. & Jonat, W. Therapie des nichtinvasiven Mammakarzinoms. *Der Onkologe* **8**, 826-31 (2002).

62. Boyages, J., Delaney, G. & Taylor, R. Predictors of local recurrence after treatment of ductal carcinoma in situ: a meta-analysis. *Cancer* **85**, 616-28 (1999).

63. Silverstein, M.J. et al. A prognostic index for ductal carcinoma in situ of the breast. *Cancer* **77**, 2267-74 (1996).

64. Asjoe, F.T. et al. The value of the Van Nuys Prognostic Index in ductal carcinoma in situ of the breast: a retrospective analysis. *Breast J* **13**, 359-67 (2007).

65. Fisher, E.R. et al. Pathologic findings from the National Surgical Adjuvant Breast Project (NSABP) eight-year update of Protocol B-17: intraductal carcinoma. *Cancer* **86**, 429-38 (1999).

66. Fisher, B. et al. Tamoxifen in treatment of intraductal breast cancer: National Surgical Adjuvant Breast and Bowel Project B-24 randomised controlled trial. *Lancet* **353**, 1993-2000 (1999).

67. Bodian, C.A., Perzin, K.H. & Lattes, R. Lobular neoplasia. Long term risk of breast cancer and relation to other factors. *Cancer* **78**, 1024-34 (1996).

68. Bembenek, A., Stroszczynski, C., Hünerbein, M., Markwardt, J. & Schlag, P.M. Neue Diagnoseverfahren beim Mammakarzinom. *Der Onkologe* **8**, 817-25 (2002).

69. Ikeda, D.M., Birdwell, R.L. & Daniel, B.L. Potential role of magnetic resonance imaging and other modalities in ductal carcinoma in situ detection. *Magn Reson Imaging Clin N Am* **9**, 345-56, vii (2001).

70. Goldhirsch, A., Glick, J.H., Gelber, R.D., Coates, A.S. & Senn, H.J. Meeting highlights: International Consensus Panel on the Treatment of Primary Breast Cancer. Seventh International Conference on Adjuvant Therapy of Primary Breast Cancer. *J Clin Oncol* **19**, 3817-27 (2001).

71. Harbeck, N., Aigner, M., Kuschel, B. & Kiechle, M. Mammakarzinom - prognostische und prädiktive Faktoren. *Der Onkologe* **8**, 808-16 (2002).

72. Hayes, D.F. et al. Tumor marker utility grading system: a framework to evaluate clinical utility of tumor markers. *J Natl Cancer Inst* **88**, 1456-66 (1996).

73. Fitzgibbons, P.L. et al. Prognostic factors in breast cancer. College of American Pathologists Consensus Statement 1999. *Arch Pathol Lab Med* **124**, 966-78 (2000).

74. Goldhirsch, A., Glick, J.H., Gelber, R.D. & Senn, H.J. Meeting highlights: International Consensus Panel on the Treatment of Primary Breast Cancer. *J Natl Cancer Inst* **90**, 1601-8 (1998).

75. Carter, C.L., Allen, C. & Henson, D.E. Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer* **63**, 181-7 (1989).

76. Allred, D.C., Harvey, J.M., Berardo, M. & Clark, G.M. Prognostic and predictive factors in breast cancer by immunohistochemical analysis. *Mod Pathol* **11**, 155-68 (1998).

77. Clark, G.M. Prognostic and Predictive Factors for Breast Cancer. *Breast Cancer* **2**, 79-89 (1995).

78. Ellis, I.O. et al. Pathological prognostic factors in breast cancer. II. Histological type. Relationship with survival in a large study with long-term follow-up. *Histopathology* **20**, 479-89 (1992).

79. Elston, C.W. & Ellis, I.O. 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* **19**, 403-10 (1991).

80. Nixon, A.J. et al. Relationship of patient age to pathologic features of the tumor and prognosis for patients with stage I or II breast cancer. *J Clin Oncol* **12**, 888-94 (1994).

81. Janicke, F. et al. Randomized adjuvant chemotherapy trial in high-risk, lymph node-negative breast cancer patients identified by urokinase-type plasminogen activator and plasminogen activator inhibitor type 1. *J Natl Cancer Inst* **93**, 913-20 (2001).

82. Look, M.P. et al. Pooled analysis of prognostic impact of urokinase-type plasminogen activator and its inhibitor PAI-1 in 8377 breast cancer patients. *J Natl Cancer Inst* **94**, 116-28 (2002).

83. Harbeck, N. et al. HER-2/neu gene amplification by fluorescence in situ hybridization allows risk-group assessment in node-negative breast cancer. *Int J Oncol* **14**, 663-71 (1999).

84. Press, M.F. et al. HER-2/neu gene amplification characterized by fluorescence in situ hybridization: poor prognosis in node-negative breast carcinomas. *J Clin Oncol* **15**, 2894-904 (1997).

85. Andrulis, I.L. et al. neu/erbB-2 amplification identifies a poor-prognosis group of women with node-negative breast cancer. Toronto Breast Cancer Study Group. *J Clin Oncol* **16**, 1340-9 (1998).

86. Mass, R.D. et al. Evaluation of clinical outcomes according to HER2 detection by fluorescence in situ hybridization in women with metastatic breast cancer treated with trastuzumab. *Clin Breast Cancer* **6**, 240-6 (2005).

87. Konigshoff, M., Wilhelm, J., Bohle, R.M., Pingoud, A. & Hahn, M. HER-2/neu gene copy number quantified by real-time PCR: comparison of gene amplification, heterozygosity, and immunohistochemical status in breast cancer tissue. *Clin Chem* **49**, 219-29 (2003).

88. Esteva, F.J. et al. Phase II study of weekly docetaxel and trastuzumab for patients with HER-2-overexpressing metastatic breast cancer. *J Clin Oncol* **20**, 1800-8 (2002).

89. Allred, D.C. et al. HER-2/neu in node-negative breast cancer: prognostic significance of overexpression influenced by the presence of in situ carcinoma. *J Clin Oncol* **10**, 599-605 (1992).

90. Muss, H.B. et al. c-erbB-2 expression and response to adjuvant therapy in women with node-positive early breast cancer. *N Engl J Med* **330**, 1260-6 (1994).

91. Paik, S. et al. HER2 and choice of adjuvant chemotherapy for invasive breast cancer: National Surgical Adjuvant Breast and Bowel Project Protocol B-15. *J Natl Cancer Inst* **92**, 1991-8 (2000).

92. Petricoin, E.F. et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572-7 (2002).

93. van 't Veer, L.J. et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-6 (2002).

94. Bast, R.C., Jr. et al. 2000 update of recommendations for the use of tumor markers in breast and colorectal cancer: clinical practice guidelines of the American Society of Clinical Oncology. *J Clin Oncol* **19**, 1865-78 (2001).

95. Hortobagyi, G.N., Hayes, D.F. & Pusztai, L. Integrating newer science into breast cancer prognosis and treatment: a review of current molecular predictors and profiles. in *American Society of Clinical Oncology, Annual Meeting* Vol. 2002 Summaries 191-202 (2002).

96. Ozols, R.F. et al. Clinical cancer advances 2006: major research advances in cancer treatment, prevention, and screening--a report from the American Society of Clinical Oncology. *J Clin Oncol* **25**, 146-62 (2007).

97. von Minckwitz, G. Evidence-based treatment of metastatic breast cancer - 2006 recommendations by the AGO Breast Commission. *Eur J Cancer* **42**, 2897-908 (2006).

98. Bear, H.D. et al. A randomized trial comparing preoperative (preop) doxorubicin/cyclophosphamide (AC) to preop AC followed by preop docetaxel (T) and to preop AC followed by postoperative (postop) T in patients (pts) with operable carcinoma of the breast: results of NSABP B-27. *Breast Cancer Res Treat* **88**, S16 (abstract 26) (2004).

99. Fisher, B. et al. Effect of preoperative chemotherapy on the outcome of women with operable breast cancer. *J Clin Oncol* **16**, 2672-85 (1998).

100. Kuerer, H.M. et al. Clinical course of breast cancer patients with complete pathologic primary tumor and axillary lymph node response to doxorubicin-based neoadjuvant chemotherapy. *J Clin Oncol* **17**, 460-9 (1999).

101. Wolmark, N., Wang, J., Mamounas, E., Bryant, J. & Fisher, B. Preoperative chemotherapy in patients with operable breast cancer: nine-year results from National Surgical Adjuvant Breast and Bowel Project B-18. *J Natl Cancer Inst Monogr*, 96-102 (2001).

102. Smith, I.C. et al. Neoadjuvant chemotherapy in breast cancer: significantly enhanced response with docetaxel. *J Clin Oncol* **20**, 1456-66 (2002).

103. von Minckwitz, G. et al. Doxorubicin with cyclophosphamide followed by docetaxel every 21 days compared with doxorubicin and docetaxel every 14 days as preoperative treatment in operable breast cancer: the GEPARDUO study of the German Breast Group. *J Clin Oncol* **23**, 2676-85 (2005).

104. Schneeweiss, A. et al. Dose-dense primary systemic chemotherapy with gemcitabine plus epirubicin sequentially followed by docetaxel for early breast cancer: final results of a phase I/II trial. *Anticancer Drugs* **16**, 1023-8 (2005).

105. Brenton, J.D., Carey, L.A., Ahmed, A.A. & Caldas, C. Molecular classification and molecular forecasting of breast cancer: ready for clinical application? *J Clin Oncol* **23**, 7350-60 (2005).

106. Sorlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**, 10869-74 (2001).

107. Sorlie, T. et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* **100**, 8418-23 (2003).

108. Sotiriou, C. et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A* **100**, 10393-8 (2003).

109. van de Vijver, M.J. et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999-2009 (2002).

110. Chang, J.C. et al. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet* **362**, 362-9 (2003).

111. Ayers, M. et al. Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer. *J Clin Oncol* **22**, 2284-93 (2004).

112. Paik, S. et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* **351**, 2817-26 (2004).

113. Wang, Y. et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671-9 (2005).

114. Hannemann, J. et al. Changes in gene expression associated with response to neoadjuvant chemotherapy in breast cancer. *J Clin Oncol* **23**, 3331-42 (2005).

115. Paik, S. et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J Clin Oncol* **24**, 3726-34 (2006).

116. Bogaerts, J. et al. Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. *Nat Clin Pract Oncol* **3**, 540-51 (2006).

117. Mook, S., Van't Veer, L.J., Rutgers, E.J., Piccart-Gebhart, M.J. & Cardoso, F. Individualization of therapy using Mammaprint: from development to the MINDACT Trial. *Cancer Genomics Proteomics* **4**, 147-55 (2007).

118. Eisen, M.B. & Brown, P.O. DNA arrays for analysis of gene expression. *Methods Enzymol* **303**, 179-205 (1999).

119. Solinas-Toldo, S. et al. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer* **20**, 399-407 (1997).

120. Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. & Lockhart, D.J. High density synthetic oligonucleotide arrays. *Nat Genet* **21**, 20-4 (1999).

121. Schlingemann, J. et al. Patient-based cross-platform comparison of oligonucleotide microarray expression profiles. *Lab Invest* **85**, 1024-39 (2005).

122. *Two-Color Microarray-Based Gene Expression Analysis,* , 66 (Agilent Technologies, Inc., 5301 Stevens Creek Rd, Santa Clara, CA 95051, USA, 2007).

123. Baugh, L.R., Hill, A.A., Brown, E.L. & Hunter, C.P. Quantitative analysis of mRNA amplification by in vitro transcription. *Nucleic Acids Res* **29**, E29 (2001).

124. Eberwine, J. et al. Analysis of gene expression in single live neurons. *Proc Natl Acad Sci U S A* **89**, 3010-4 (1992).

125. Kenzelmann, M. et al. High-accuracy amplification of nanogram total RNA amounts for gene profiling. *Genomics* **83**, 550-8 (2004).

126. Lockhart, D.J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**, 1675-80 (1996).

127. Matz, M. et al. Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic Acids Res* **27**, 1558-60 (1999).

128. Smith, L. et al. Single primer amplification (SPA) of cDNA for microarray expression analysis. *Nucleic Acids Res* **31**, e9 (2003).

129. Wang, E., Miller, L.D., Ohnmacht, G.A., Liu, E.T. & Marincola, F.M. High-fidelity mRNA amplification for gene profiling. *Nat Biotechnol* **18**, 457-9 (2000).

130. Wrobel, G. et al. Optimization of high-density cDNA-microarray protocols by 'design of experiments'. *Nucleic Acids Res* **31**, e67 (2003).

131. Thuerigen, O. Vergleichende Untersuchung der Hybridisierungseigenschaften Oligonukleotid- und cDNA-basierender Microarrays für die Expressionsanalyse humaner Transkriptome. 102 (Ruprecht-Karls-Universität Heidelberg, Heidelberg, 2002).

132. Agilent Technologies, I. *Two-Color Microarray-Based Gene Expression Analysis,* , 66 (Agilent Technologies, Inc.
5301 Stevens Creek Rd
Santa Clara, CA 95051 USA, 2007).

133. Wessendorf, S. et al. Automated screening for genomic imbalances using matrix-based comparative genomic hybridization. *Lab Invest* **82**, 47-60 (2002).

134. Nucleic Acids. in *Handbook of Biochemistry and Molecular Biology*, Vol. 1 (ed. G.D., F.) 589 (CRC Press, 1975).

135. R Development Core Team, R.F.f.S.C., Vienna, Austria. R: A language and environment for statistical computing. (2004).

136. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1**, S96-104 (2002).

137. Vapnik, V. The support vector method of function estimation. in *Nonlinear Modeling: Advanced Black-box Techniques.* (eds. Suykens, J.A.K. & Vandewalle, J.) 55-85 (Kluwer Academic Publishers, Boston, 1998).

138. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine Learning* **46**, 389-422 (2002).

139. Pfaffl, M.W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* **29**, e45 (2001).

140. Sawin, K.E., Mitchison, T.J. & Wordeman, L.G. Evidence for kinesin-related proteins in the mitotic apparatus using peptide antibodies. *J Cell Sci* **101 (Pt 2)**, 303-13 (1992).

141. Kohler, G. & Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* **256**, 495-7 (1975).

142. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).

143. Al-Shahrour, F., Diaz-Uriarte, R. & Dopazo, J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20**, 578-80 (2004).

144. Kanehisa, M. et al. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* **34**, D354-7 (2006).

145. Suurmeijer, A.J. Optimizing immunohistochemistry in diagnostic tumor pathology with antigen retrieval. *Eur J Morphol* **32**, 325-30 (1994).

146. Cattoretti, G. Standardization and reproducibility in diagnostic immunohistochemistry. *Hum Pathol* **25**, 1107-9 (1994).

147. Iczkowski, K.A., Cheng, L., Crawford, B.G. & Bostwick, D.G. Steam heat with an EDTA buffer and protease digestion optimizes immunohistochemical expression of basal cell-specific antikeratin 34betaE12 to discriminate cancer in prostatic epithelium. *Mod Pathol* **12**, 1-4 (1999).

148. Weigelt, B., Peterse, J.L. & van 't Veer, L.J. Breast cancer metastasis: markers and models. *Nat Rev Cancer* **5**, 591-602 (2005).

149. Lonning, P.E., Knappskog, S., Staalesen, V., Chrisanthar, R. & Lillehaug, J.R. Breast cancer prognostication and prediction in the postgenomic era. *Ann Oncol* **18**, 1293-306 (2007).

150. Perou, C.M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747-52 (2000).

151. Iwao-Koizumi, K. et al. Prediction of docetaxel response in human breast cancer by gene expression profiling. *J Clin Oncol* **23**, 422-31 (2005).

152. Rouzier, R. et al. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin Cancer Res* **11**, 5678-85 (2005).

153. Gianni, L. et al. Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer. *J Clin Oncol* **23**, 7265-77 (2005).

154. Dressman, H.K. et al. Gene expression profiles of multiple breast cancer phenotypes and response to neoadjuvant chemotherapy. *Clin Cancer Res* **12**, 819-26 (2006).

155. Sorlie, T. et al. Gene expression profiles do not consistently predict the clinical treatment response in locally advanced breast cancer. *Mol Cancer Ther* **5**, 2914-8 (2006).

156. Bonnefoi, H. et al. Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial. *Lancet Oncol* **8**, 1071-8 (2007).

157. Potti, A. et al. Genomic signatures to guide the use of chemotherapeutics. *Nat Med* **12**, 1294-300 (2006).

158. Mesa, R.A. Tipifarnib: farnesyl transferase inhibition at a crossroads. *Expert Rev Anticancer Ther* **6**, 313-9 (2006).

159. Yoshikawa, Y. et al. The M-Ras-RA-GEF-2-Rap1 pathway mediates tumor necrosis factor-alpha dependent regulation of integrin activation in splenocytes. *Mol Biol Cell* **18**, 2949-59 (2007).

160. Ehrhardt, G.R., Leslie, K.B., Lee, F., Wieler, J.S. & Schrader, J.W. M-Ras, a widely expressed 29-kD homologue of p21 Ras: expression of a constitutively active mutant results in factor-independent growth of an interleukin-3-dependent cell line. *Blood* **94**, 2433-44 (1999).

161. Ward, K.R., Zhang, K.X., Somasiri, A.M., Roskelley, C.D. & Schrader, J.W. Expression of activated M-Ras in a murine mammary epithelial cell line induces epithelial-mesenchymal transition and tumorigenesis. *Oncogene* **23**, 1187-96 (2004).

162. Wu, X., Noh, S.J., Zhou, G., Dixon, J.E. & Guan, K.L. Selective activation of MEK1 but not MEK2 by A-Raf from epidermal growth factor-stimulated Hela cells. *J Biol Chem* **271**, 3265-71 (1996).

163. Fang, Y., Johnson, L.M., Mahon, E.S. & Anderson, D.H. Two phosphorylation-independent sites on the p85 SH2 domains bind A-Raf kinase. *Biochem Biophys Res Commun* **290**, 1267-74 (2002).

164. Yuryev, A., Ono, M., Goff, S.A., Macaluso, F. & Wennogle, L.P. Isoform-specific localization of A-RAF in mitochondria. *Mol Cell Biol* **20**, 4870-8 (2000).

165. Massague, J. How cells read TGF-beta signals. *Nat Rev Mol Cell Biol* **1**, 169-78 (2000).

166. Elliott, R.L. & Blobe, G.C. Role of transforming growth factor Beta in human cancer. *J Clin Oncol* **23**, 2078-93 (2005).

167. Pardali, K. & Moustakas, A. Actions of TGF-beta as tumor suppressor and pro-metastatic factor in human cancer. *Biochim Biophys Acta* **1775**, 21-62 (2007).

168. Melhuish, T.A., Gallo, C.M. & Wotton, D. TGIF2 interacts with histone deacetylase 1 and represses transcription. *J Biol Chem* **276**, 32109-14 (2001).

169. Engel, M.E., McDonnell, M.A., Law, B.K. & Moses, H.L. Interdependent SMAD and JNK signaling in transforming growth factor-beta-mediated transcription. *J Biol Chem* **274**, 37413-20 (1999).

170. Sekiya, T. et al. Identification of BMP and activin membrane-bound inhibitor (BAMBI), an inhibitor of transforming growth factor-beta signaling, as a target of the beta-catenin pathway in colorectal tumor cells. *J Biol Chem* **279**, 6840-6 (2004).

171. Visvader, J.E. et al. The LIM domain gene LMO4 inhibits differentiation of mammary epithelial cells in vitro and is overexpressed in breast cancer. *Proc Natl Acad Sci U S A* **98**, 14452-7 (2001).

172. Wang, N. et al. Expression of an engrailed-LMO4 fusion protein in mammary epithelial cells inhibits mammary gland development in mice. *Oncogene* **23**, 1507-13 (2004).

173. Sum, E.Y. et al. Overexpression of LMO4 induces mammary hyperplasia, promotes cell invasion, and is a predictor of poor outcome in breast cancer. *Proc Natl Acad Sci U S A* **102**, 7659-64 (2005).

174. Wang, N. et al. The LIM-only factor LMO4 regulates expression of the BMP7 gene through an HDAC2-dependent mechanism, and controls cell proliferation and apoptosis of mammary epithelial cells. *Oncogene* **26**, 6431-41 (2007).

175. Sutherland, K.D. et al. Mutational analysis of the LMO4 gene, encoding a BRCA1-interacting protein, in breast carcinomas. *Int J Cancer* **107**, 155-8 (2003).

176. Song, K., Wang, H., Krebs, T.L. & Danielpour, D. Novel roles of Akt and mTOR in suppressing TGF-beta/ALK5-mediated Smad3 activation. *Embo J* **25**, 58-69 (2006).

177. Creighton, C.J. A gene transcription signature of the Akt/mTOR pathway in clinical breast tumors. *Oncogene* **26**, 4648-55 (2007).

178. Matheny, S.A. et al. Ras regulates assembly of mitogenic signalling complexes through the effector protein IMP. *Nature* **427**, 256-60 (2004).

179. Mochan, T.A., Venere, M., DiTullio, R.A., Jr. & Halazonetis, T.D. 53BP1, an activator of ATM in response to DNA damage. *DNA Repair (Amst)* **3**, 945-52 (2004).

180. Wang, S.C. et al. Binding at and transactivation of the COX-2 promoter by nuclear tyrosine kinase receptor ErbB-2. *Cancer Cell* **6**, 251-61 (2004).

181. Pao, G.M., Janknecht, R., Ruffner, H., Hunter, T. & Verma, I.M. CBP/p300 interact with and function as transcriptional coactivators of BRCA1. *Proc Natl Acad Sci U S A* **97**, 1020-5 (2000).

182. Lodish, H. et al. *Molecular Cell Biology*, (W. H. FREEMAN, 41 Madison Avenue, 10010 New York, NY, USA 2000).

183. Bader, A.G., Kang, S., Zhao, L. & Vogt, P.K. Oncogenic PI3K deregulates transcription and translation. *Nat Rev Cancer* **5**, 921-929 (2005).

184. Epstein, A.L., Variakojis, D., Berger, C. & Hecht, B.K. Use of novel chemical supplements in the establishment of three human malignant lymphoma cell lines (NU-DHL-1, NU-DUL-1, and NU-AMB-1) with chromosome 14 translocations. *Int J Cancer* **35**, 619-27 (1985).

185. Winter, J.N., Variakojis, D. & Epstein, A.L. Phenotypic analysis of established diffuse histiocytic lymphoma cell lines utilizing monoclonal antibodies and cytochemical techniques. *Blood* **63**, 140-6 (1984).

186. Collins, S.J., Gallo, R.C. & Gallagher, R.E. Continuous growth and differentiation of human myeloid leukaemic cells in suspension culture. *Nature* **270**, 347-9 (1977).

187. Gallagher, R. et al. Characterization of the continuous, differentiating myeloid cell line (HL-60) from a patient with acute promyelocytic leukemia. *Blood* **54**, 713-33 (1979).

188. Dalton, W.T., Jr. et al. HL-60 cell line was derived from a patient with FAB-M2 and not FAB-M3. *Blood* **71**, 242-7 (1988).

189. Collins, S.J. The HL-60 promyelocytic leukemia cell line: proliferation, differentiation, and cellular oncogene expression. *Blood* **70**, 1233-44 (1987).

8. Appendix

A. Cell Lines HL-60 and NU-DHL-1

	NU-DHL-1	HL-60
Cell type	human B cell lymphoma	human acute myeloid leukemia
DSMZ N°	ACC 583	ACC 3
Origin	established from the left inguinal lymph node of a 73-year-old Caucasian man with B-cell Non-Hodgkin lymphoma (B-NHL, diffuse large cell lymphoma, non-cleaved cell type) in 1982	established from the peripheral blood of a 35-year-old woman with acute myeloid leukemia (AML FAB M2) in 1976; cells can be used for induction of differentiation studies; described to be responsive to DMSO, phorbol ester TPA and other reagents and to carry amplified MYC gene; present cells are apparently tetraploid derivates of hypodiploid original where MYC was amplified in dmin (instead of hsr)
References	Epstein et al., Int J Cancer 35:619-627 (1985);[184] Winter et al., Blood 63:140-146 (1984).[185]	Collins et al., Nature 270:347-349 (1977);[186] Gallagher et al., Blood 54:713-733 (1979);[187] Dalton et al., Blood 71:242-247 (1988);[188] Collins, Blood 70:1233-1244 (1987, review).[189]
Depositor	Dr. A. L. Epstein, USC, Los Angeles, CA, USA	Dr. E. Porfiri, The Royal Free Hospital, Department of Haematology, London, UK
DSMZ Cell Culture Data		
Morphology	single, round to polymorph cells growing in suspension	round, single cells in suspension
Medium	80-90% RPMI 1640 + 10-20% FBS	90% RPMI 1640 + 10% FBS
Subculture	split saturated culture 1:2 to 1:4 every 2-3 days; seed out at ca. 1.0 x 106 cells/ml; maintain at ca. 0.5-1.0 x 106 cells/ml; recommended to start culture in a 24-well-plate and with 20% FBS	maintain at 0.5-1.0 x 10^6 cells/ml, split 1:2 to 1:5 every 1-2 days; seed out at about 1 x 10^6 cells/ml
Incubation	at 37 °C with 5% CO_2	at 37 °C with 5% CO_2
Doubling time	ca. 50 hours	ca. 25 hours
Harvest	cell harvest of ca. 1.5 x 10^6 cells/ml	maximal density of 1.5-2.0 x 10^6 cells/ml
Storage	frozen with 70% medium, 20% FBS, 10% DMSO at about 5 x 10^6 cells/ampoule	frozen with 70% medium, 20% FBS, 10% DMSO at about 4-5 x 10^6 cells/ampoule
DSMZ Scientific Data		
Mycoplasma	negative in microbiological culture, PCR assays	negative in DAPI, microbiological culture, RNA hybridization, PCR assays
Immunology	CD3 -, CD10 -, CD13 -, CD19 +, CD20 +, CD34 -, CD37 +, CD38 -, CD79a +, cyCD79a +, CD80 -, CD138 -, HLA-DR +, sm/cyIgG -, sm/cyIgM +, sm/cykappa -, sm/cylambda +	CD3 -, CD4 +, CD13 +, CD14 -, CD15 +, CD19 -, CD33 +, CD34 -, HLA-DR -
Fingerprint	fluorescent nonaplex PCR of short tandem repeat markers revealed a unique DNA profile	multiplex PCR of minisatellite markers revealed a unique DNA profile
Species	confirmed as human by cytogenetics and species PCR	confirmed as human with IEF of MDH, NP
Cytogenetics	human hyperdiploid karyotype with 4% polyploidy 51(46-53)<2n>XY, +5, +8, +9, +12, +12, t(3;8)(p25;q24), i(5p), dup(7)(q21.3q31.1), der(8)t(3;8)(p25;q24), i(9p), del(12)(q11), i(12p), der(14)t(14;18)(q32;q21)x1-3 sideline with dup(2)(p2?1p2?4) carries t(3;8) and t(14;18) effecting respective rearrangements of MYC and IGH-BCL2 resembles published karyotype	human flat-moded hypotetraploid karyotype with hypodiploid sideline and 1.5% polyploidy 82-88<4n>XX, -X, -X, -8, -8, -16, -17, -17, +18, +22, +2mar, ins(1;8)(p?31;q24hsr)x2, der(5)t(5;17)(q11;q11)x2, add(6)(q27)x2, der(9)del(9)(p13)t(9;14)(q?22;q?22)x2, der(14)t(9;14)(q?22;q?22)x2, der(16)t(16;17)(q22;q22)x1-2, add(18)(q21) - sideline with: -2, -5, -15, del(11)(q23.1q23.2) - c-myc amplicons present in der(1) and in both markers
Viruses	EBV -, HBV -, HCV -, HIV -, HTLV-I/II -	ELISA: reverse transcriptase negative; PCR: EBV -, HBV -, HCV -, HHV-8 -, HIV -, HTLV-I/II -

B. Manufacturers of Chemicals and Laboratory Material

Agilent	Agilent Technologies, Inc. 5301 Stevens Creek Blvd Santa Clara , CA 95051 USA	Millipore	Millipore 290 Concord Rd. Billerica, MA 01821 USA
Ambion	Applied Biosystems 850 Lincoln Centre Drive Foster City, CA 94404 USA	Molecular Devices	Molecular Devices 1311 Orleans Drive Sunnyvale, CA 94089-1136 USA
Amersham	Amersham Place Little Chalfont Buckinghamshire HP7 9NA United Kingdom	NanoDrop Technologies	NanoDrop Technologies 3411 Silverside Rd Bancroft Building Wilmington, DE 19810 USA
Applied Biosystems	Applied Biosystems 850 Lincoln Centre Drive Foster City, CA 94404 USA	NeoLab	neoLab Migge GmbH Rischerstr. 7-9 69123 Heidelberg Germany
B.Braun	B. Braun Melsungen AG Carl-Braun-Straße 1 34212 Melsungen Germany	New England Biolabs	New England Biolabs 240 County Road Ipswich, MA 01938-2723 USA
Bio-Rad	Bio-Rad Laboratories 2000 Alfred Nobel Drive Hercules, CA 94547 USA	Operon	Operon Biotechnologies, Inc. 2211 Seminole Drive Huntsville, AL 35805 USA
Biospring	BioSpring GmbH Alt Fechenheim 34 60386 Frankfurt am Main Germany	Promega	Promega Corporation 2800 Woods Hollow Road Madison, WI 53711 USA
Dako	Dako Denmark A/S Produktionsvej 42 DK-2600 Glostrup Denmark	Qiagen	QIAGEN GmbH QIAGEN Strasse 1 40724 Hilden Germany
Epicentre	EPICENTRE Biotechnologies 726 Post Road Madison, WI 53713 USA	Roche	F. Hoffmann-La Roche Ltd Grenzacherstrasse 124 4070 Basel Switzerland
Eppendorf	Eppendorf AG Barkhausenweg 1 22339 Hamburg Germany	Schott Nexterion	SCHOTT Jenaer Glas GmbH Otto-Schott-Strasse 13 07745 Jena Germany
Fermentas	Fermentas, Inc. 798 Cromwell Park Drive Glen Burnie, MD 21061 USA	Sigma-Aldrich	Sigma-Aldrich Co. 3050 Spruce Street St. Louis, MO 63103 USA
Genomic Solutions	Genomic Solutions Inc. 4355 Varsity Drive Ann Arbor, MI 48108 USA	Stratagene	11011 N. Torrey Pines Road La Jolla, CA 92037 USA
GibCo	Invitrogen Corporation 1600 Faraday Avenue Carlsbad, California 92008 USA	TeleChem	TeleChem International, Inc. 524 East Weddell Drive Sunnyvale, CA 94089 USA
Invitrogen	Invitrogen Corporation 1600 Faraday Avenue Carlsbad, California 92008 USA	Thermo Scientific	Thermo Fisher Scientific, Inc. 81 Wyman Street Waltham, MA 02454 USA
Kendro	Thermo Fisher Scientific, Inc. 81 Wyman Street Waltham, MA 02454 USA	Varian Inc.	Varian, Inc. 3120 Hansen Way Palo Alto, CA 94304-1030 USA
Marligen Bioscience	Marligen Biosciences, Inc. 2502 Urbana PikeIjamsville, MD 21754USA	Vector	Vector Laboratories 30 Ingold Road Burlingame, CA 94010USA
Merck	Merck KGaA Frankfurter Str. 250 64293 Darmstadt Germany		

C. Primers for Real-time Quantitative PCR

Gene	Primer (Exon:Exon)[§]	Sequence, 5' -> 3'	T_m [°C]
DCTN2	DCTN2 (2:3) upper	CGCCATGGCGGACCCTAAAT	60.5
	DCTN2 (2:3) lower	TTGTCAGCTCCTCCGCATCGAA	61.5
GALNAC4S-6ST	GALNAC4S-6ST (5:6,7) upper	ATCCACGCCTTTCAGCCAAATG	59.8
	GALNAC4S-6ST (5:6,7) lower	AGCCCAACCTGGAGCCTCACA	60.4
HER2	ERBB2 (16:17) upper	CATCAACTGCACCCACTCCTGTGT	59.8
	ERBB2 (16:17) lower	CTCCACCAGCTCCGTTTCCTG	58.3
ESR1	ESR1 (6:7) upper	CTCTTGGACAGGAACCAGGGAAAAT	59.6
	ESR1 (6:7) lower	CAGGGTGCTGGACAGAAATGTGTAC	58.6
BAMBI	BAMBI (1:2) upper	CGTGCTGCTCACCAAAGGTGAAAT	61.1
	BAMBI (1:2) lower	CATGGGTGAGTGGGGAATTTGAG	59.1
DAPK2	DAPK2 (7,8:9,10) upper	GGCCAAGGACTTTATTCGGAAGC	59.1
	DAPK2 (7,8:9,10) lower	CACAGGGACACGATGCTGAAGGA	60.8
LMO4	LMO4 (4:5) upper	GTCCCGGGAGATCGGTTTCACT	60.0
	LMO4 (4:5) lower	ATGGGATCCACCTGTGATGAACAAA	60.2
SMAD3	SMAD3 (3,4:5,6) upper	GAGCCCCAGAGCAATATTCCAGA	58.3
	SMAD3 (3,4:5,6) lower	GGCCGGCTCGCAGTAGGTAACT	60.4
SRC	SRC (3:4) upper	CTGGCCGGTGGAGTGACCAC	59.8
	SRC (3:4) lower	CAAAATACCACTCCTCAGCCTGGAT	58.6

[§] primers spanning exon borders are denoted by kommata; PCR products spanning exon borders are denoted by colons

D. Genes Contained in the Predictive Gene Expression Signature

Rank	Symbol	Description	Mapping	Ensembl ID	Operon ID
1	*DAPK2*	Death-associated protein kinase 2 (EC 2.7.1.37) (DAP kinase 2) (DAP- kinase related protein 1) (DRP-	15q22.31	ENSG00000035664	H200012808
2	*RASAL1*	RasGAP-activating-like protein 1.	12q24.13	ENSG00000111344	H200014809
3	*THAP8*	THAP domain protein 8.	19q13.12	ENSG00000161277	H200020498
4					H200015020
5	*OAT*	Ornithine aminotransferase, mitochondrial precursor (EC 2.6.1.13) (Ornithineoxo-acid aminotransfer	10q26.13	ENSG00000065154	H200006115
6	*CPM*	Carboxypeptidase M precursor (EC 3.4.17.12).	12q15	ENSG00000135678	H200019714
7	*PITPNM2*	phosphatidylinositol transfer protein, membrane-associated 2, PYK2 N-terminal domain-interacting rec	12q24.31	ENSG00000090975	H200016945
8	*SLC35B2*	solute carrier family 35, member B2, 3'-phosphoadenosine 5'-phosphosulfate transporter [Homo sapiens	6p21.1	ENSG00000157593	H200008629
9	*CRYBB2*	Beta crystallin B2 (BP).	22q11.23	ENSG00000100058	H200007805
10	*TCF8*	Transcription factor 8 (NIL-2-A zinc finger protein) (Negative regulator of IL2).	10p11.22	ENSG00000148516	H200015445
11					H200001488
12					H200019153
13	*TOR1B*	Torsin B precursor (Torsin family 1 member B) (FKSG18 protein).	9q34.11	ENSG00000136816	H200016328
14	*SMU1*	smu-1 suppressor of mec-8 and unc-52 homolog, ortholog of rat brain-enriched WD-repeat protein, homo	9p21.1	ENSG00000122692	H200014545
15	*SMYD3*	SET and MYND domain containing protein 3 (Zinc finger MYND domain containing protein 1).	1q44	ENSG00000185420	H200001594
16	*ARMC8*	armadillo repeat containing 8, HSPC056 protein [Homo sapiens].	3q22.3	ENSG00000114098	H200011194
17	*C18orf1*		18p11.21	ENSG00000168675	H200013878
18	*PRDX1*	Peroxiredoxin 1 (EC 1.11.1.-) (Thioredoxin peroxidase 2) (Thioredoxin- dependent peroxide reductase	1p34.1	ENSG00000117450	H200008482
19	*LMO4*	LIM domain transcription factor LMO4 (LIM-only protein 4) (LMO-4) (Breast tumor autoantigen).	1p22.3	ENSG00000143013	H200000848
20	*CSNK2A2*	Casein kinase II, alpha' chain (CK II) (EC 2.7.1.37).	16q21	ENSG00000070770	H200006904
21	*PTPN13*	Protein tyrosine phosphatase, non-receptor type 13 (EC 3.1.3.48) (Protein-tyrosine phosphatase 1E) (	4q21.3	ENSG00000163629	H200015130
22					H200011178
23	*BAMBI*	BMP and activin membrane-bound inhibitor homolog precursor (Putative transmembrane protein NMA) (Non	10p12.1	ENSG00000095739	H200006552
24	*EP300*	E1A-associated protein p300 (EC 2.3.1.48).	22q13.2	ENSG00000100393	H200003499
25		GT198, complete ORF, TBP-1 interacting protein [Homo sapiens].	17q21.2	ENSG00000131470	H200017422
26	*SRF*	Serum response factor (SRF).	6p21.1	ENSG00000112658	H200014058
27	*STK32B*	serine/threonine kinase 32B, gene for serine/threonine protein kinase [Homo sapiens].	4p16.2	ENSG00000152953	H200005311
28	*NICAL*	NEDD9 interacting protein with calponin homology and LIM domains (Molecule interacting with CasL pro	6q21	ENSG00000135596	H200010147
29	*TTC14*	Tetratricopeptide repeat protein 14 (TPR repeat protein 14).	3q26.33	ENSG00000163728	H200004622
30	*ARAF1*	A-Raf proto-oncogene serine/threonine-protein kinase (EC 2.7.1.37) (A- raf-1) (Proto-oncogene Pks).	Xp11.3	ENSG00000078061	H200006361
31	*SLC6A8*	Sodium- and chloride-dependent creatine transporter 1 (CT1).	Xq28	ENSG00000130821	H200014389
32	*EVL*	Ena/vasodilator stimulated phosphoprotein-like protein (Ena/VASP-like protein).	14q32.2	ENSG00000196405	H200015749
33	*POLR1D*	DNA-directed RNA polymerase I 16 kDa polypeptide (EC 2.7.7.6) (RPA16).	13q12.2	ENSG00000186184	H200017450
34			16p11.2	ENSG00000047578	H200004007
35		ezrin-binding partner PACE-1 isoform 1 [Homo sapiens].	1q24.2	ENSG00000000457	H200003382
36	*WDR5B*	WD repeat domain 5B [Homo sapiens].	3q21.1	ENSG00000196981	H200013357

37	*PAFAH1B3*	Platelet-activating factor acetylhydrolase IB gamma subunit (EC 3.1.1.47) (PAF acetylhydrolase 29 kD	19q13.2	ENSG00000079462	H200001306
38			22q11.21	ENSG00000183597	H200020502
39	*WDR8*	WD-repeat protein 8.	1p36.32	ENSG00000116213	H200004114
40	*PTPNS1*	Protein-tyrosine phosphatase non-receptor type substrate 1 precursor (SHP substrate-1) (SHPS-1) (Inh	20p13	ENSG00000198053	H200014140
41	*AGTRAP*	angiotensin II receptor-associated protein, angiotensin II, type I receptor-associated protein [Homo	1p36.22	ENSG00000177674	H200002197
42	*PIK3C2B*	Phosphatidylinositol-4-phosphate 3-kinase C2 domain-containing beta polypeptide (EC 2.7.1.154) (Phos	1q32.1	ENSG00000133056	H200013011
43			9q12, 9p11.2	ENSG00000196635, ENSG00000197068, ENSG00000198119, ENSG00000196164, ENSG00000198052	H200018900
44	*SULT1A3*	Monoamine-sulfating phenol sulfotransferase (EC 2.8.2.1) (Sulfotransferase, monoamine-preferring),	16p11.2	ENSG00000132207, ENSG00000181625	H200017186
45	*PAFAH2*	Platelet-activating factor acetylhydrolase 2, cytoplasmic (EC 3.1.1.47) (Serine dependent phospholip	1p36.11	ENSG00000158006	H200015501
46	*C22orf8*		22q13.31	ENSG00000100376	H200016550
47	*EPS8L2*	epidermal growth factor receptor pathway substrate 8-like protein 2, EPS8-related protein 2, epiderm	11p15.5	ENSG00000177106	H200005141
48	*DAB2IP*	DAB2 interacting protein, nGAP-like protein, DOC-2/DAB2 interactive protein [Homo sapiens].	9q33.2	ENSG00000136848	H200015622
49	*PCTP*	Phosphatidylcholine transfer protein (PC-TP) (StAR-related lipid transfer protein 2) (StARD2) (START	17q23.1	ENSG00000141179	H200008819
50	*TIMM17B*	Mitochondrial import inner membrane translocase subunit Tim17 B (JM3).	Xp11.23	ENSG00000126768	H200002793
51	*C14orf122*	UPF0172 protein C14orf122 (CGI-112).	14q11.2	ENSG00000100908	H200016750
52	*CKAP1*	Tubulin-specific chaperone B (Tubulin folding cofactor B) (Cytoskeleton-associated protein CKAPI).	19q13.12	ENSG00000105254	H200004062
53		Pygopus homolog 2.	1q22	ENSG00000163348	H200008025
54	*XRN2*	5'-3' exoribonuclease 2 (EC 3.1.11.-).	20p11.22	ENSG00000088930	H200016651
55		ubiquitin-conjugating enzyme HBUCE1 [Homo sapiens].	7p13	ENSG00000078967	H200002803
56	*SEC14L2*	SEC14-like protein 2 (Alpha-tocopherol associated protein) (TAP) (hTAP) (Supernatant protein factor)	22q12.2	ENSG00000100003	H200017232
57			8q11.21	ENSG00000164808	H200017329
58			22q13.31	ENSG00000075240	H200010685
59					H200008383
60					H200010303
61			2q35	ENSG00000124006	H200013996
62					H200011048
63	*GALK2*	N-acetylgalactosamine kinase (EC 2.7.1.-) (GalNAc kinase) (Galactokinase 2).	15q21.1	ENSG00000156958	H200012811
64	*PKN3*	protein kinase PKNbeta [Homo sapiens].	9q34.11	ENSG00000160447	H200004669
65	*GBF1*	Golgi-specific brefeldin A-resistance guanine nucleotide exchange factor 1 (BFA-resistant GEF 1).	10q24.32	ENSG00000107862	H200014080
66			7p14.3	ENSG00000105778	H200013693
67	*EXT2*	Exostosin-2 (EC 2.4.1.224) (EC 2.4.1.225) (Glucuronosyl-N- acetylglucosaminyl-proteoglycan/N-acetylg	11p11.2	ENSG00000151348	H200006075
68	*MKRN2*	Makorin 2 (HSPC070).	3p25.2	ENSG00000075975	H200017431
69	*ADCK1*	aarF domain containing kinase 1 [Homo sapiens].	14q24.3	ENSG00000063761	H200002462
70					H200017729
71	*C9orf25*		9p13.3	ENSG00000164970	H200001541
72	*C6orf199*		6q21	ENSG00000155085	H200012502
73	*PEPD*	Xaa-Pro dipeptidase (EC 3.4.13.9) (X-Pro dipeptidase) (Proline dipeptidase) (Prolidase) (Imidodipept	19q13.11	ENSG00000124299	H200005894

74	*BMS1L*	Ribosome biogenesis protein BMS1 homolog.	10q11.21	ENSG00000165733	H200001947
75			2p23.3	ENSG00000163026	H200003419
76	*ZNF335*	Zinc finger protein 335.	20q13.12	ENSG00000198026	H200007640
77	*AP1G2*	Adapter-related protein complex 1 gamma 2 subunit (Gamma2-adaptin) (Adaptor protein complex AP-1 gam	14q11.2	ENSG00000092051	H200019914
78	*HPS4*	Hermansky-Pudlak syndrome 4 protein (Light-ear protein homolog).	22q12.1	ENSG00000100099	H200001277
79					H200004418
80	*CACNG4*	Voltage-dependent calcium channel gamma-4 subunit (Neuronal voltage- gated calcium channel gamma-4 s	17q24.2	ENSG00000075461	H200010418
81					H200002313
82	*KDELR3*	ER lumen protein retaining receptor 3 (KDEL receptor 3).	22q13.1	ENSG00000100196	H200016239
83	*MYST1*	MYST histone acetyltransferase 1, histone acetyltransferase MYST1 [Homo sapiens].	16p11.2	ENSG00000103510	H200004571
84	*OPTN*	optineurin, glaucoma 1, open angle, E (adult-onset), tumor necrosis factor alpha-inducible cellular	10p13	ENSG00000123240	H200017355
85	*PROCR*	Endothelial protein C receptor precursor (Endothelial cell protein C receptor) (Activated protein C	20q11.22	ENSG00000101000	H200006932
86	*ACTG1*	Actin, cytoplasmic 1 (Beta-actin)., Actin, cytoplasmic 2 (Gamma-actin).	7p22.1, 17q25.3	ENSG00000075624, ENSG00000184009	H200002375
87	*SLC1A3*	Excitatory amino acid transporter 1 (Sodium-dependent glutamate/aspartate transporter 1) (Glial glut	5p13.2	ENSG00000079215	H200006090
88	*BRAP*	BRCA1-associated protein (EC 6.3.2.-) (BRAP2) (Impedes mitogenic signal propagation) (IMP).	12q24.12	ENSG00000089234	H200012400
89	*INPP4B*	inositol polyphosphate-4-phosphatase, type II, 105kD, inositol polyphosphate 4-phosphatase II, 4-pho	4q31.21	ENSG00000109452	H200013899
90					H200011830
91					H200010509
92					H200019925
93	*COL12A1*	Collagen alpha 1(XII) chain precursor.	6q13	ENSG00000111799	H200011114
94	*HESX1*	Homeobox expressed in ES cells 1 (Homeobox protein ANF) (hAnf).	3p14.3	ENSG00000163666	H200008014
95	*NMNAT1*	Nicotinamide mononucleotide adenylyltransferase 1 (EC 2.7.7.1) (NMN adenylyltransferase 1).	1p36.22	ENSG00000173614	H200007228
96	*NPAS2*	Neuronal PAS domain protein 2 (Neuronal PAS2) (Member of PAS protein 4) (MOP4).	2q11.2	ENSG00000170485	H200018956
97	*BMP4*	Bone morphogenetic protein 4 precursor (BMP-4) (BMP-2B).	14q22.2	ENSG00000125378	H200005656
98	*CREB3*	cAMP responsive element binding protein 3, cAMP responsive element binding protein 3 (luman), cyclic	9p13.3	ENSG00000107175	H200009202
99		Regulatory associated protein of mTOR (Raptor) (P150 target of rapamycin (TOR)-scaffold protein).	17q25.3	ENSG00000141564	H200002921
100	*ATP6V1C 2*	ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C isoform 2, V-ATPase C2 subunit, ATPase, H+ tr	2p25.1	ENSG00000143882	H200008589
101	*C20orf14*	U5 snRNP-associated 102 kDa protein (U5-102 kDa protein).	20q13.33	ENSG00000101161	H200004082
102	*AURKB*	Serine/threonine-protein kinase 12 (EC 2.7.1.37) (Aurora- and Ipl1- like midbody-associated protein	17p13.1	ENSG00000178999	H200008454
103	*FBXO18*	F-box only protein 18 (EC 3.6.1.-) (F-box DNA helicase 1).	10p15.1	ENSG00000134452	H200001177
104					H200009676
105	*RHOBTB1*	Rho-related BTB domain-containing protein 1.	10q21.2	ENSG00000072422	H200002445
106	*PTP4A1*	protein tyrosine phosphatase type IVA, member 1, Protein tyrosine phosphatase IVA1 [Homo sapiens].	6q12	ENSG00000112245	H200015401
107	*SMURF2*	Smad ubiquitination regulatory factor 2 (EC 6.3.2.-) (Ubiquitin protein ligase SMURF2) (Smad-speci	17q24.1	ENSG00000108854	H200014627
108	*UBTF*	Nucleolar transcription factor 1 (Upstream binding factor 1) (UBF-1) (Autoantigen NOR-90).	17q21.31	ENSG00000108312	H200010392
109					H200002419
110		Autoantigen NGP-1.	1p34.3	ENSG00000134697	H200006126

111	*SUPT4H1*	Transcription initiation protein SPT4 homolog 1.	17q23.2	ENSG00000108372	H200006614
112			17q21.2	ENSG00000141698	H200015595
113			2p11.2	ENSG00000144115	H200001790
114					H200008842
115			12q13.2	ENSG00000135482	H200005456
116					H200017299
117	*CCNL1*	cyclin L1, cyclin L ania-6a [Homo sapiens].	3q25.31	ENSG00000163660	H200000937
118	*RNU3IP2*	U3 small nucleolar RNA interacting protein 2 (U3 small nucleolar ribonucleoproptein-associated 55-kD	3p21.2	ENSG00000114767	H200013906
119	*ASF1B*	ASF1 anti-silencing function 1 homolog B, anti-silencing function 1B, CCG1-interacting factor A-II [	19p13.12	ENSG00000105011	H200003654
120	*C9orf100*		9p13.3	ENSG00000137135	H200015825
121	*SUMO4*	Ubiquitin-like protein SMT3B (Sentrin 2) (HSMT3).	17q25.1, Xq23	ENSG00000180283, ENSG00000188612	H200008406
122					H200007932
123	*WFDC2*	WAP four-disulfide core domain protein 2 precursor (Major epididymis- specific protein E4) (Epididym	20q13.12	ENSG00000101443	H200000668
124	*PABPC4*	Polyadenylate-binding protein 4 (Poly(A)-binding protein 4) (PABP 4) (Inducible poly(A)-binding prot	1p34.3	ENSG00000090621	H200007875
125	*C10orf7*	D123 gene product [Homo sapiens].	10p13	ENSG00000151465	H200006879
126	*GRM4*	Metabotropic glutamate receptor 4 precursor (mGluR4).	6p21.31	ENSG00000124493	H200008305
127	*NDUFS1*	NADH-ubiquinone oxidoreductase 75 kDa subunit, mitochondrial precursor (EC 1.6.5.3) (EC 1.6.99.3) (C	2q33.3	ENSG00000023228	H200001624
128		phosphatidylinositol-specific phospholipase C, X domain containing 1 [Homo sapiens].	Xp22.33	ENSG00000182378	H200003545
129	*TPCN1*	two pore segment channel 1, two-pore channel 1, two-pore segment channel 1 [Homo sapiens].	12q24.13	ENSG00000186815	H200003641
130	*TNFRSF7*	Tumor necrosis factor receptor superfamily member 7 precursor (CD27L receptor) (T-cell activation an	12p13.31	ENSG00000139193	H200021166
131	*RKHD2*	ring finger and KH domain containing 2 [Homo sapiens].	18q21.1	ENSG00000176624	H200002191
132	*DTYMK*	Thymidylate kinase (EC 2.7.4.9) (dTMP kinase).,Thymidylate kinase (EC 2.7.4.9) (dTMP kinase).	2q37.3	ENSG00000168393, ENSG00000188547	H200006601
133	*RAB32*	Ras-related protein Rab-32.	6q24.3	ENSG00000118508	H200004149
134	*WBSCR18*	Williams-Beuren syndrome chromosome region 18 protein.	7q11.23	ENSG00000176410	H200002733
135	*AKR1C1*	Aldo-keto reductase family 1 member C1 (EC 1.1.1.-) (Trans-1,2- dihydrobenzene-1,2-diol dehydrogenas	10p15.1	ENSG00000187134	H200018207
136	*GLTSCR1*	Glioma tumor suppressor candidate region gene 1 protein.	19q13.32	ENSG00000063169	H200010786
137					H200016743
138					H200000805
139	*SOX9*	Transcription factor SOX-9.	17q24.3	ENSG00000125398	H200000590
140	*ST7*	suppression of tumorigenicity 7 isoform a, family with sequence similarity 4, subfamily A, member 1,	7q31.2	ENSG00000004866	H200001109
141	*SLC39A14*	solute carrier family 39 (zinc transporter), member 14, solute carrier family 39 (metal ion transpor	8p21.3	ENSG00000104635	H200010399
142	*RRAGC*	Ras-related GTP binding C, Rag C protein [Homo sapiens].	1p34.3	ENSG00000116954	H200011726
143					H200016422
144	*FAM20A*	Protein FAM20A precursor (UNQ9388/PRO34279).	17q24.2	ENSG00000108950	H200013465
145					H200002577
146	*SALL2*	Sal-like protein 2 (Zinc finger protein SALL2) (HSal2).	14q11.2	ENSG00000165821	H200006732
147	*SLC39A11*	solute carrier family 39 (metal ion transporter), member 11, chromosome 17 open reading frame 26 [Ho	17q24.3	ENSG00000133195	H200000782
148	*GFPT2*	Glucosaminefructose-6-phosphate aminotransferase [isomerizing] 2 (EC 2.6.1.16) (Hexosephosphate am	5q35.3	ENSG00000131459	H200003997
149	*MRS2L*	MRS2-like, magnesium homeostasis factor, MRS2 (S. cerevisiae)-like, magnesium homeostasis factor [Ho	6p22.2	ENSG00000124532	H200003978
150					H200009425

151	*GBA*	Glucosylceramidase precursor (EC 3.2.1.45) (Beta-glucocerebrosidase) (Acid beta-glucosidase) (D-gluc	1q22	ENSG00000177628	H200020059
152	*MYO9A*	myosin IXA [Homo sapiens].	15q23	ENSG00000066933	H200003274
153		HTPAP protein [Homo sapiens].	8p12	ENSG00000147535	H200007814
154	*SETBP1*	SET-binding protein (SEB).	18q12.3	ENSG00000152217	H200013788
155	*SRI*	Sorcin (22 kDa protein) (CP-22) (V19).	7q21.12	ENSG00000075142	H200009922
156	*ABCB8*	ATP-binding cassette, sub-family B, member 8, mitochondrial precursor (Mitochondrial ATP-binding cas	7q36.1	ENSG00000197150	H200012134
157			15q15.1	ENSG00000137824	H200001578
158	*HOXB6*	Homeobox protein Hox-B6 (Hox-2B) (Hox-2.2) (HU-2).	17q21.32	ENSG00000108511	H200010880
159	*ATP11A*	Potential phospholipid-transporting ATPase IH (EC 3.6.3.1) (ATPase class I type 11A) (ATPase IS).	13q34	ENSG00000068650	H200003896
160	*TFCP2L2*	leader-binding protein 32 isoform 1, LBP protein 32, leader-binding protein 32, mammalian grainyhead	2p25.1	ENSG00000134317	H200018964
161	*KIAA1117*		6q14.1	ENSG00000083097	H200017268
162	*ABCD1*	Adrenoleukodystrophy protein (ALDP).	Xq28	ENSG00000101986	H200007466
163	*CCM2*	cerebral cavernous malformation 2, chromosome 7 open reading frame 22 [Homo sapiens].	7p13	ENSG00000136280	H200001861
164		NY-REN-58 antigen [Homo sapiens].	12q22	ENSG00000173588	H200005210
165			17q21.2	ENSG00000131475	H200011600
166		UPF0120 protein DKFZp564C186.	1p36.33	ENSG00000188976	H200013109
167		CGI-29 protein, APAF1-interacting protein [Homo sapiens].	11p13	ENSG00000149089	H200011285
168	*BAK1*	Bcl-2 homologous antagonist/killer (Apoptosis regulator BAK) (BCL2- like 7 protein).	6p21.31	ENSG00000030110	H200010571
169					H200003750
170	*CACNA1H*	Voltage-dependent T-type calcium channel alpha-1H subunit (Voltage- gated calcium channel alpha subu	16p13.3	ENSG00000196557	H200012371
171	*SRC*	Proto-oncogene tyrosine-protein kinase Src (EC 2.7.1.112) (p60-Src) (c-Src).	20q11.23	ENSG00000197122	H200014805
172					H200013303
173	*CXorf37*		Xp11.23	ENSG00000101997	H200003626
174	*GPNMB*	Putative transmembrane protein NMB precursor (Transmembrane glycoprotein HGFIN).	7p15.3	ENSG00000136235	H200006911
175	*RBM10*	RNA-binding protein 10 (RNA binding motif protein 10) (DXS8237E).	Xp11.3	ENSG00000182872	H200013980
176			1p34.1	ENSG00000187147	H200010430
177					H200012709
178			12q24.13	ENSG00000139405	H200016259
179	*APG10L*	APG10 autophagy 10-like [Homo sapiens].	5q14.1	ENSG00000152348	H200009074
180	*UBE2A*	Ubiquitin-conjugating enzyme E2 A (EC 6.3.2.19) (Ubiquitin-protein ligase A) (Ubiquitin carrier prot	Xq24	ENSG00000077721	H200006776
181	*MRPS30*	Mitochondrial 28S ribosomal protein S30 (S30mt) (MRP-S30) (Programmed cell death protein 9) (BM-047)	5p12	ENSG00000112996	H200003850
182					H200013007
183	*C10orf33*		10q24.2	ENSG00000119943	H200010766
184	*TRIP10*	Cdc42-interacting protein 4 (Thyroid receptor interacting protein 10) (TRIP-10).	19p13.3	ENSG00000125733	H200005905
185					H200012966
186			3p21.31	ENSG00000198530	H200020692
187			1p34.3	ENSG00000163875	H200002622
188		XTP3-transactivated protein A [Homo sapiens].	16p11.2	ENSG00000179958	H200015610
189	*NFIB*	Nuclear factor 1 B-type (Nuclear factor 1/B) (NF1-B) (NFI-B) (NF-I/B) (CCAAT-box binding transcripti	9p22.3	ENSG00000147862	H200004230
190	*PRKCH*	Protein kinase C, eta type (EC 2.7.1.-) (nPKC-eta) (PKC-L).	14q23.1	ENSG00000027075	H200018874
191					H200019805
192		CGI-111 protein [Homo sapiens].	5q23.3	ENSG00000066583	H200001967

193			12q21.31	ENSG00000111058	H200002237
194	*TRIM41*	Tripartite motif protein 41.	5q35.3	ENSG00000146063	H200020110
195	*DHRS4*	Dehydrogenase/reductase SDR family member 4 (EC 1.1.1.184) (NADPH- dependent carbonyl reductase/NADP	14q11.2	ENSG00000157326	H200001202
196	*AMACR*	Alpha-methylacyl-CoA racemase (EC 5.1.99.4) (2-methylacyl-CoA racemase).	5p13.3	ENSG00000082196	H200012768
197					H200020351
198	*MINA*	MYC induced nuclear antigen isoform 2, myc-induced nuclear antigen, 53 kDa, mineral dust induced gen	3q11.2	ENSG00000170854	H200003264
199					H200010060
200		Protein HSPC020.	11q13.4	ENSG00000110200	H200004421
201	*FGD1*	Putative Rho/Rac guanine nucleotide exchange factor (Rho/Rac GEF) (Faciogenital dysplasia protein) (	Xp11.22	ENSG00000102302	H200000409
202					H200010990
203		Arylsulfatase G [Homo sapiens].	17q24.2	ENSG00000141337	H200017561
204	*RKHD3*	ring finger and KH domain containing 3 [Homo sapiens].	15q25.2	ENSG00000183496	H200011330
205		Guanine nucleotide exchange factor-related protein (Deafness locus associated putative guanine nucle	11p15.1	ENSG00000129158	H200004794
206	*PTPRU*	Receptor-type protein-tyrosine phosphatase U precursor (EC 3.1.3.48) (R-PTP-U) (Protein-tyrosine pho	1p35.3	ENSG00000060656	H200002847
207	*B4GALT2*	Beta-1,4-galactosyltransferase 2 (EC 2.4.1.-) (Beta-1,4-GalTase 2) (Beta4Gal-T2) (b4Gal-T2) (UDP-gal	1p34.1	ENSG00000117411	H200015011
208	*MRPL37*	mitochondrial ribosomal protein L37, ribosomal protein, mitochondrial, L2 [Homo sapiens].	1p32.3	ENSG00000116221	H200000887
209	*ASH2L*	ASH2-like protein.	8p12	ENSG00000129691	H200001319
210	*PIGS*	GPI transamidase component PIG-S (Phosphatidylinositol-glycan biosynthesis, class S protein).	17q11.2	ENSG00000087111	H200020143
211			5q32	ENSG00000145882	H200016427
212			7p22.3	ENSG00000164818, ENSG00000188246	H200016943
213	*PEX6*	Peroxisome assembly factor-2 (PAF-2) (Peroxisomal-type ATPase 1) (Peroxin-6) (Peroxisomal biogenesis	6p21.1	ENSG00000124587	H200010054
214	*ENTPD7*	ectonucleoside triphosphate diphosphohydrolase 7, lysosomal apyrase-like protein 1 [Homo sapiens].	10q24.2	ENSG00000198018	H200003172
215					H200002658
216	*ARHGDIG*	Protein disulfide-isomerase A2 precursor (EC 5.3.4.1) (PDIp).	16p13.3	ENSG00000185615	H200010174
217	*ESPL1*	Separin (EC 3.4.22.49) (Separase) (Caspase-like protein ESPL1) (Extra spindle poles-like 1 protein).	12q13.13	ENSG00000135476	H200013875
218	*PRKWNK2*	Serine/threonine-protein kinase WNK2 (EC 2.7.1.37) (Protein kinase with no lysine 2) (Protein kinase	9q22.31	ENSG00000165238	H200013033
219	*BCKDHA*	2-oxoisovalerate dehydrogenase alpha subunit, mitochondrial precursor (EC 1.2.4.4) (Branched-chain a	19q13.2	ENSG00000142046	H200006590
220			11q23.1	ENSG00000137702	H200010909
221	*BIRC7*	Baculoviral IAP repeat-containing protein 7 (Kidney inhibitor of apoptosis protein) (KIAP) (Melanoma	20q13.33	ENSG00000101197	H200016398
222	*TRIP6*	Thyroid receptor interacting protein 6 (TRIP6) (OPA-interacting protein 1) (Zyxin related protein 1)	7q22.1	ENSG00000087077	H200012209
223					H200018290
224	*NPY1R*	Neuropeptide Y receptor type 1 (NPY1-R).	4q32.2	ENSG00000164128	H200007802
225		pyruvate dehydrogenase phosphatase regulatory subunit [Homo sapiens].	16q22.1	ENSG00000090857	H200016507
226	*C20orf18*	Ubiquitin conjugating enzyme 7 interacting protein 3 (Hepatitis B virus X-associated protein 4) (HBV	20p13	ENSG00000125826	H200015861
227	*RAB5A*	Ras-related protein Rab-5A.	3p24.3	ENSG00000144566	H200005896
228					H200007233
229	*BRD8*	bromodomain containing 8 isoform 1, skeletal muscle abundant protein, thyroid hormone receptor coact	5q31.2	ENSG00000112983	H200001063
230			11p13	ENSG00000176148	H200003139

231	*NIT1*	nitrilase 1 [Homo sapiens].	1q23.3	ENSG00000158793	H200013566
232	*DEPDC6*	DEP domain containing 6 [Homo sapiens].	8q24.12	ENSG00000155792	H200010250
233	*ZDHHC14*	Zinc finger DHHC domain containing protein 14 (NEW1 domain containing protein) (NEW1CP).	6q25.3	ENSG00000175048	H200004435
234	*CPXM*	Potential carboxypeptidase X precursor (EC 3.4.17.-) (Metallocarboxypeptidase CPX-1).	20p13	ENSG00000088882	H200008267
235	*ARNTL*	Aryl hydrocarbon receptor nuclear translocator-like protein 1 (Brain and muscle ARNT-like 1) (Member	11p15.2	ENSG00000133794	H200005950
236	*PRPSAP2*	Phosphoribosyl pyrophosphate synthetase-associated protein 2 (PRPP synthetase-associated protein 2)	17p11.2	ENSG00000141127	H200002257
237		dudulin 2, tumor suppressor pHyde, six transmembrane prostate protein 3 [Homo sapiens].	2q14.2	ENSG00000115107	H200005252
238	*ELOVL5*	homolog of yeast long chain polyunsaturated fatty acid elongatio, homolog of yeast long chain polyun	6p12.1	ENSG00000012660	H200016204
239					H200014395
240	*SOX13*	SOX-13 protein (Type 1 diabetes autoantigen ICA12) (Islet cell antigen 12).	1q32.1	ENSG00000143842	H200014916
241	*CYR61*	CYR61 protein precursor (Cysteine-rich, angiogenic inducer, 61) (Insulin-like growth factor-binding	1p22.3	ENSG00000142871	H200001697
242					H200001611
243					H200009962
244	*NT5M*	5'(3')-deoxyribonucleotidase, cytosolic type (EC 3.1.3.-) (Cytosolic 5',3'-pyrimidine nucleotidase)	17q25.1	ENSG00000125458	H200005632
245					H200001954
246	*LDLR*	Low-density lipoprotein receptor precursor (LDL receptor).	19p13.2	ENSG00000130164	H200015172
247					H200003707
248	*C5orf19*		5q31.2	ENSG00000132563	H200003814
249	*ADCY3*	Adenylate cyclase type III (EC 4.6.1.1) (Adenylate cyclase, olfactive type) (ATP pyrophosphate-lyase	2p23.3	ENSG00000138031	H200001647
250	*C5orf3*		5q33.2	ENSG00000055147	H200007691
251	*NF1*	Neurofibromin (Neurofibromatosis-related protein NF-1) [Contains: Neurofibromin truncated].	17q11.2	ENSG00000196712	H200010569
252			5q31.1	ENSG00000145835	H200009573
253					H200005644
254	*GNG11*	Guanine nucleotide-binding protein G(I)/G(S)/G(O) gamma-11 subunit.	7q21.3	ENSG00000127920	H200007026
255	*ZNF259*	Zinc-finger protein ZPR1 (Zinc finger protein 259).	11q23.3	ENSG00000109917	H200001377
256	*SUSD2*	sushi domain containing 2, Sushi domain (SCR repeat) containing [Homo sapiens].	22q11.23	ENSG00000099994	H200012965
257	*FGA*	Fibrinogen alpha/alpha-E chain precursor [Contains: Fibrinopeptide A].	4q31.3	ENSG00000171560	H200021184
258	*FNTB*	Protein farnesyltransferase beta subunit (EC 2.5.1.58) (CAAX farnesyltransferase beta subunit) (RAS	14q23.3	ENSG00000125954	H200000075
259	*TIMM23*	Mitochondrial import inner membrane translocase subunit Tim23.	10q11.23	ENSG00000138297	H200002039
260					H200002087
261					H200016707
262					H200001644
263	*CLECSF12*	C-type lectin, superfamily member 12 isoform b, beta-glucan receptor, lectin-like receptor 1, transm	12p13.2	ENSG00000172243	H200014156
264					H200000829
265					H200001681
266	*DGKG*	Diacylglycerol kinase, gamma (EC 2.7.1.107) (Diglyceride kinase) (DGK- gamma) (DAG kinase gamma).	3q27.2	ENSG00000058866	H200010327
267			19q13.2	ENSG00000090924	H200011738
268	*MICA*	MHC class I chain-related gene A protein [Homo sapiens].	6p21.33	ENSG00000184444	H200010444
269			11q23.2	ENSG00000180425	H200010517
270	*TMEM25*	transmembrane protein 25, 0610039J01Rik [Homo sapiens].	11q23.3	ENSG00000149582	H200021089

271					H200003472
272					H200004592
273	*KEAP1*	Kelch-like ECH-associated protein 1 (Cytosolic inhibitor of Nrf2).	19p13.2	ENSG00000079999	H200005267
274	*MRPL30*	mitochondrial ribosomal protein L30 isoform a [Homo sapiens].	2q11.2	ENSG00000185414	H200002065
275	*HELLS*	helicase, lymphoid-specific, SWI/SNF2-related, matrix-associated, actin-dependent regulator of chrom	10q23.33	ENSG00000119969	H200014971
276	*ENC1*	Ectoderm-neural cortex-1 protein (ENC-1) (P53-induced protein 10) (Nuclear matrix protein NRP/B).	5q13.3	ENSG00000171617	H200011342
277	*PSME2*	Proteasome activator complex subunit 2 (Proteasome activator 28-beta subunit) (PA28beta) (PA28b) (Ac	14q11.2	ENSG00000100911	H200008379
278	*CENPJ*	centromere protein J, centrosomal P4.1-associated protein, LYST-interacting protein LIP1, LAG-3-asso	13q12.12	ENSG00000151849	H200017688
279	*DSCR1*	Calcipressin 1 (Down syndrome critical region protein 1) (Myocyte- enriched calcineurin interacting	21q22.12	ENSG00000159200	H200014262
280	*WNT5B*	Wnt-5b protein precursor.	12p13.33	ENSG00000111186	H200018203
281		putative breast adenocarcinoma marker [Homo sapiens].	19q13.43	ENSG00000130724	H200002069
282			10q22.2	ENSG00000138286	H200004671
283	*SCAMP3*	Secretory carrier-associated membrane protein 3 (Secretory carrier membrane protein 3).	1q22	ENSG00000116521	H200014894
284					H200003033
285	*RAI3*	Retinoic acid induced 3 protein (G protein-coupled receptor family C group 5 member A) (Retinoic aci	12p13.1	ENSG00000013588	H200014654
286					H200004670
287	*NFX1*	Transcriptional repressor NF-X1 (EC 6.3.2.-) (Nuclear transcription factor, X box-binding, 1).	9p13.3	ENSG00000086102	H200000753
288	*FOS*	Proto-oncogene protein c-fos (Cellular oncogene fos) (G0/G1 switch regulatory protein 7).	14q24.3	ENSG00000170345	H200003548
289		Cappuccino protein homolog.	4p16.1	ENSG00000186222	H200001457
290	*CARM1*	coactivator-associated arginine methyltransferase 1, coactivator-associated arginine methyltransfera	19p13.2	ENSG00000142453	H200013420
291	*FBXO17*	F-box only protein 17 (F-box only protein 26).	19q13.2	ENSG00000161241	H200016187
292	*NANOG*	Nanog homeobox, homeobox transcription factor Nanog [Homo sapiens].	15q14, 12p13.31	ENSG00000179437, ENSG00000111704	H200019169
293	*PPIL1*	Peptidyl-prolyl cis-trans isomerase like 1 (EC 5.2.1.8) (PPIase) (Rotamase) (CGI-124) (UNQ2425/PRO49	6p21.2	ENSG00000137168	H200003788
294	*MYO5A*	Myosin Va (Myosin 5A) (Dilute myosin heavy chain, non-muscle) (Myosin heavy chain 12) (Myoxin).	15q21.2	ENSG00000197535	H200007910
295	*C14orf138*		14q21.3	ENSG00000100483	H200002224
296	*PRKR*	Interferon-induced, double-stranded RNA-activated protein kinase (EC 2.7.1.-) (Interferon-inducible	2p22.2	ENSG00000055332	H200017076
297			2q37.1	ENSG00000185404	H200016714
298	*RASL11B*	RAS-like family 11 member B [Homo sapiens].	4q12	ENSG00000128045	H200001570
299	*C16orf34*	Crm, cramped-like, Crm (Cramped Drosophila)-like [Homo sapiens].	16p13.3	ENSG00000007545	H200002479
300	*EDG4*	Lysophosphatidic acid receptor Edg-4 (LPA receptor 2) (LPA-2).	19p13.11	ENSG00000064547	H200012385
301	*REN, KCTD11*	Renin precursor (EC 3.4.23.15) (Angiotensinogenase).,potassium channel tetramerisation domain containing 11, chromosome 17 open reading frame 36, retinoi	1q32.1,17p13.1	ENSG00000143839, ENSG00000184542	H200015103
302	*UST*	uronyl-2-sulfotransferase, uronyl 2-sulfotransferase, dermatan/chondroitin sulfate 2-sulfotransferas	6q25.1	ENSG00000111962	H200013102
303	*TCFL5*	Transcription factor-like 5 protein (Cha transcription factor) (HPV-16 E2 binding protein 1) (E2BP-1	20q13.33	ENSG00000101190	H200004024
304	*HIVEP1*	Zinc finger protein 40 (Human immunodeficiency virus type I enhancer- binding protein 1) (HIV-EP1) (	6p24.1	ENSG00000095951	H200000081
305	*NUDT5*	ADP-sugar pyrophosphatase (EC 3.6.1.13) (EC 3.6.1.-) (Nucleoside diphosphate-linked moiety X motif 5	10p14	ENSG00000165609	H200017987
306	*LAT*	Linker for activation of T cells (36 kDa phospho-tyrosine adaptor protein) (pp36) (p36-38).	16p11.2	ENSG00000169678	H200007039

307	*RASA3*	Ras GTPase-activating protein 3 (GAP1(IP4BP)) (Ins P4-binding protein).	13q34	ENSG00000185989	H200012198
308					H200001307
309					H200004113
310	*MTX2*	Metaxin 2.	2q31.1	ENSG00000128654	H200004105
311			7p22.3	ENSG00000146540	H200019508
312	*RHOBTB3*	Rho-related BTB domain-containing protein 3.	5q15	ENSG00000164292	H200001910
313	*TAF1*	Transcription initiation factor TFIID subunit 1 (EC 2.7.1.37) (Transcription initiation factor TFIID	Xq13.1	ENSG00000147133	H200000303
314					H200001320
315	*USP12*	Ubiquitin carboxyl-terminal hydrolase 12 (EC 3.1.2.15) (Ubiquitin thiolesterase 12) (Ubiquitin-speci	13q12.13	ENSG00000152484	H200004574
316	*PHC1*	Polyhomeotic-like protein 1 (Early development regulator protein 1) (HPH1).,Polyhomeotic-like protein 1 (Early development regulator protein 1) (HPH1).	12q13.2, 12p13.31	ENSG00000179899, ENSG00000111752	H200018199
317			15q15.3	ENSG00000137770	H200011302
318	*ATP5J2*	ATP synthase f chain, mitochondrial (EC 3.6.3.14).	7q22.1	ENSG00000160916	H200014108
319	*PLXNA1*	plexin A1, plexin 1 [Homo sapiens].	3q21.2	ENSG00000114554	H200019013
320	*PLCB4*	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase beta 4 (EC 3.1.4.11) (Phosphoinositide pho	20p12.2	ENSG00000101333	H200017629
321	*C1orf6*	Ataxin-1 ubiquitin-like interacting protein A1U.	1q22	ENSG00000160803	H200017798
322	*CCND3*	G1/S-specific cyclin D3.	6p21.1	ENSG00000112576	H200007012
323	*C10orf81*		10q25.3	ENSG00000148735	H200015418
324	*ZNF228*	Zinc finger protein 228.	19q13.31	ENSG00000062370	H200004887
325	*TGIF2*	Homeobox protein TGIF2 (TGFB-induced factor 2) (5'-TG-3' interacting factor 2) (TGF(beta)-induced tr	20q11.23	ENSG00000118707	H200010661
326	*RPP38*	Ribonuclease P protein subunit p38 (EC 3.1.26.5) (RNaseP protein p38).	10p13	ENSG00000152464	H200010679
327	*PMS2L11*	HPMSR6.	7q11.23	ENSG00000186704	H200018214
328	*IARS*	Isoleucyl-tRNA synthetase, cytoplasmic (EC 6.1.1.5) (IsoleucinetRNA ligase) (IleRS) (IRS).	9q22.31	ENSG00000196305	H200008082
329	*SUPT16H*	chromatin-specific transcription elongation factor large subunit [Homo sapiens].	14q11.2	ENSG00000092201	H200002431
330			1p34.1	ENSG00000159596	H200014285
331	*PSCD2L*	Cytohesin 2 (ARF nucleotide-binding site opener) (ARNO protein) (ARF exchange factor).	19q13.32	ENSG00000105443	H200018158
332					H200020364
333	*CD1C*	T-cell surface glycoprotein CD1c precursor (CD1c antigen).	1q23.1	ENSG00000158481	H200000342
334	*POFUT2*	GDP-fucose protein O-fucosyltransferase 2 precursor (EC 2.4.1.221) (Peptide O-fucosyltransferase) (O	21q22.3	ENSG00000186866	H200003223
335	*TBXA2R*	NY-REN-24 antigen (Fragment).Thromboxane A2 receptor (TXA2-R) (Prostanoid TP receptor).	19p13.3	ENSG00000105298, ENSG00000179855, ENSG00000006638	H200012751
336					H200019921
337	*UBE1*	Ubiquitin-activating enzyme E1 (A1S9 protein).	Xp11.3	ENSG00000130985	H200000527
338	*CORO1C*	Coronin 1C (Coronin 3) (hCRNN4).	12q24.11	ENSG00000110880	H200002647
339	*ZNF41*	Zinc finger protein 41.	Xp11.3	ENSG00000147124	H200015400
340					H200010439
341	*DPP3*	Bardet-Biedl syndrome 1 protein (BBS2-like protein 2).	11q13.2	ENSG00000174483	H200003203
342	*PRAF2*	JM4 protein [Homo sapiens].	Xp11.23	ENSG00000102050	H200003933
343			1p36.12	ENSG00000090432	H200001884
344			7q22.1	ENSG00000160993	H200016776
345					H200015636
346	*TP53BP1*	Tumor suppressor p53-binding protein 1 (p53-binding protein 1) (53BP1).	15q15.3	ENSG00000067369	H200007926
347	*WBP11*	WW domain binding protein 11, Npw38-binding protein NpwBP, SH3 domain-binding protein SNP70 [Homo sa	12p12.3	ENSG00000084463	H200019689

348					H200010493
349	*CDK2AP1*	Cyclin-dependent kinase 2-associated protein 1 (CDK2-associated protein 1) (Putative oral cancer sup	12q24.31	ENSG00000111328	H200000786
350		nucleostemin isoform 1, putative nucleotide binding protein, estradiol-induced [Homo sapiens].	3p21.1	ENSG00000163938	H200017556
351					H200008222
352	*VAV3*	Vav-3 protein.	1p13.3	ENSG00000134215	H200016605
353	*PHKG2*	Phosphorylase B kinase gamma catalytic chain, testis/liver isoform (EC 2.7.1.38) (PHK-gamma-T) (Phos	16p11.2	ENSG00000156873	H200014730
354	*ZNF346*	zinc finger protein 346, double-stranded RNA-binding zinc finger protein JAZ [Homo sapiens].	5q35.2	ENSG00000113761	H200008493
355	*CHEK2*	Serine/threonine-protein kinase Chk2 (EC 2.7.1.37) (Cds1).	22q12.1	ENSG00000183765	H200013557
356	*NT5C2*	Cytosolic purine 5'-nucleotidase (EC 3.1.3.5) (5'-nucleotidase cytosolic II).	10q24.33	ENSG00000076685	H200013286
357	*OSR1*	oxidative-stress responsive 1 [Homo sapiens].	3p22.2	ENSG00000172939	H200010689
358	*PRKAB1*	5'-AMP-activated protein kinase, beta-1 subunit (AMPK beta-1 chain) (AMPKb).	12q24.23	ENSG00000111725	H200001142
359	*ATP6V1F*	Vacuolar ATP synthase subunit F (EC 3.6.3.14) (V-ATPase F subunit) (Vacuolar proton pump F subunit)	7q32.1	ENSG00000128524	H200006486
360	*FRAP1*	FKBP-rapamycin associated protein (FRAP) (Rapamycin target protein).	1p36.22	ENSG00000198793	H200019855
361	*PHB*	Prohibitin.	17q21.32	ENSG00000167085	H200012361
362	*ZNF502*	Zinc finger protein 502.	3p21.31	ENSG00000196653	H200015276
363					H200002063
364					H200010421
365	*OSBPL3*	Oxysterol binding protein-related protein 3 (OSBP-related protein 3) (ORP-3).	7p15.3	ENSG00000070882	H200014777
366	*ACTG2*	Actin, gamma-enteric smooth muscle (Alpha-actin 3).	2p13.1	ENSG00000163017	H200006478
367					H200012250
368	*EXOC7*	Exocyst complex component 7 (Exocyst complex component Exo70).	17q25.1	ENSG00000182473	H200016731
369	*STARD7*	StAR-related lipid transfer protein 7 (StARD7) (START domain- containing protein 7) (GTT1 protein).	2q11.2	ENSG00000084090	H200017790
370		CGI-62 protein [Homo sapiens].	8q21.12	ENSG00000104427	H200012156
371	*ALG1*	beta-1,4-mannosyltransferase, beta-1,4 mannosyltransferase [Homo sapiens].	16p13.3	ENSG00000033011	H200004717
372					H200004334
373	*DDX56*	Probable ATP-dependent 61 kDa nucleolar RNA helicase (EC 3.6.1.-) (DEAD-box protein 56) (DEAD-box pr	7p13	ENSG00000136271	H200001883
374	*TBC1D1*	TBC1 domain family member 1.	4p14	ENSG00000065882	H200017315
375	*CYP27A1*	Cytochrome P450 27, mitochondrial precursor (EC 1.14.-.-) (Cytochrome P-450C27/25) (Sterol 26-hydrox	2q35	ENSG00000135929	H200006956
376	*HEMK1*	HemK protein homolog (EC 2.1.1.-) (M.HsaHemKP).	3p21.31	ENSG00000114735	H200004813
377			17q25.3	ENSG00000178927	H200001827
378					H200003632
379			17q25.1	ENSG00000177728	H200009966
380	*LENG8*	leukocyte receptor cluster (LRC) member 8 [Homo sapiens].	19q13.42	ENSG00000167615	H200018209
381					H200001270
382	*PSD4*	pleckstrin and Sec7 domain containing 4, ADP-ribosylation factor guanine nucleotide-exchange factor	2q13	ENSG00000125637	H200011683
383					H200014351
384	*APBB2*	Amyloid beta A4 precursor protein-binding family B member 2 (Fe65-like protein) (Fragment).	4p14	ENSG00000163697	H200019090
385	*WISP2*	WNT1 inducible signaling pathway protein 2 precursor (WISP-2) (Connective tissue growth factor-like	20q13.12	ENSG00000064205	H200014646
386	*TCF3*	Transcription factor E2-alpha (Immunoglobulin enhancer binding factor E12/E47) (Transcription factor	19p13.3	ENSG00000071564	H200011103

387	*SCRT2*	UPF0238 protein c20orf139.	20p13	ENSG00000172070	H200013143
388	*TXNDC4*	Thioredoxin domain containing protein 4 precursor (Endoplasmic reticulum protein ERp44).	9q31.1	ENSG00000023318	H200013926
389					H200009570
390	*PLAGL2*	Zinc finger protein PLAGL2 (Pleiomorphic adenoma-like protein 2).	20q11.21	ENSG00000126003	H200013939
391					H200005593
392					H200008282
393	*TUBGCP3*	Gamma-tubulin complex component 3 (GCP-3) (Spindle pole body protein Spc98 homolog) (hSpc98) (hGCP3)	13q34	ENSG00000126216	H200001849
394	*C6orf79*		6p23	ENSG00000050393	H200016505
395					H200000866
396	*RASSF1*	Ras association domain family 1 (Ras association, RalGDS/AF-6, domain family 1).	3p21.31	ENSG00000068028	H200003705
397	*PDE9A*	High-affinity cGMP-specific 3',5'-cyclic phosphodiesterase 9A (EC 3.1.4.17).	21q22.3	ENSG00000160191	H200002780
398	*PECR*	peroxisomal trans-2-enoyl-CoA reductase, peroxisomal trans 2-enoyl CoA reductase, putative short cha	2q35	ENSG00000115425	H200017595
399	*UAP1*	UDP-N-acetylhexosamine pyrophosphorylase (Antigen X) (AGX) (Sperm- associated antigen 2) [Includes:	1q23.3	ENSG00000117143	H200003000
400	*PBP*	Phosphatidylethanolamine-binding protein (PEBP) (Prostatic binding protein) (HCNPpp) (Neuropolypepti	12q24.23	ENSG00000089220	H200006761
401	*LRP5*	Low-density lipoprotein receptor-related protein 5 precursor.	11q13.2	ENSG00000162337	H200001207
402	*CHST3*	carbohydrate (chondroitin 6) sulfotransferase 3, chondroitin 6-sulfotransferase [Homo sapiens].	10q22.1	ENSG00000122863	H200007322
403	*LEPREL1*	leprecan-like 1, myxoid liposarcoma associated protein 4, prolyl 3-hydroxylase 3 [Homo sapiens].	3q28	ENSG00000090530	H200004600
404	*TNRC18*	CAGL79 (Fragment).	7p22.1	ENSG00000182095	H200020510
405	*KBTBD2*	Kelch repeat and BTB domain containing protein 2 (BTB and kelch domain containing protein 1).	7p14.3	ENSG00000170852	H200002891
406	*C20orf121*		20q13.12	ENSG00000124120	H200008681
407	*PNKP*	Bifunctional polynucleotide phosphatase/kinase (Polynucleotide kinase- 3'-phosphatase) (DNA 5'-kinas	19q13.33	ENSG00000039650	H200006474
408					H200004276
409			10p11.23	ENSG00000165757	H200004811
410	*MYBBP1A*	MYB binding protein 1a, p53-activated protein-2 [Homo sapiens].	17p13.2	ENSG00000132382	H200003194
411	*FOXC1*	Forkhead box protein C1 (Forkhead-related protein FKHL7) (Forkhead- related transcription factor 3)	6p25.3	ENSG00000054598	H200020379
412	*RHEB*	GTP-binding protein Rheb (Ras homolog enriched in brain).	7q36.1	ENSG00000106615	H200017539
413	*GPRC5C*	G protein-coupled receptor family C group 5 member C precursor (Retinoic acid induced gene 3 protein	17q25.1	ENSG00000170412	H200005296
414		Tetraspan NET-7.	10q22.1	ENSG00000099282	H200010703
415	*ETV5*	Ets-related protein ERM (ETS translocation variant 5).	3q27.2	ENSG00000171656	H200004645
416	*PGD*	6-phosphogluconate dehydrogenase, decarboxylating (EC 1.1.1.44).	1p36.22	ENSG00000142657	H200006228
417	*RGS2*	Regulator of G-protein signaling 2 (RGS2) (G0/G1 switch regulatory protein 8).	1q31.2	ENSG00000116741	H200006587
418	*BCL6*	B-cell lymphoma 6 protein (BCL-6) (Zinc finger protein 51) (LAZ-3 protein) (BCL-5) (Zinc finger and	3q27.3	ENSG00000113916	H200014019
419	*UBAP2*	ubiquitin associated protein 2 isoform 1, AD-012 protein [Homo sapiens].	9p13.3	ENSG00000137073	H200002430
420	*COPB*	Coatomer beta subunit (Beta-coat protein) (Beta-COP).	11p15.2	ENSG00000129083	H200000730
421	*GNB4*	Guanine nucleotide-binding protein beta subunit 4 (Transducin beta chain 4).	3q26.32	ENSG00000114450	H200008063
422	*MPHOSPH6*	M-phase phosphoprotein 6.	16q23.3	ENSG00000135698	H200013829
423			12p13.33	ENSG00000171792	H200007245
424					H200016752

425	*C13orf12*		13q12.3	ENSG00000132963	H200017498
426	*CYP1B1*	Cytochrome P450 1B1 (EC 1.14.14.1) (CYPIB1).	2p22.2	ENSG00000138061	H200013981
427	*GSDML*	gasdermin-like [Homo sapiens].	17q21.1	ENSG00000073605	H200002789
428					H200018267
429		CDA02 protein [Homo sapiens].	3q25.1	ENSG00000144895	H200019327
430			2q21.1	ENSG00000169606	H200018879
431	*ZNF160*	Zinc finger protein Kr18 (HKr18).	19q13.41	ENSG00000170949	H200015014
432	*IQCB1*	IQ calmodulin-binding motif containing protein 1.	3q13.33	ENSG00000173226	H200007819
433	*IPO4*	Importin 4 (Importin 4b) (Imp4b) (Ran-binding protein 4) (RanBP4).	14q11.2	ENSG00000196497	H200005457
434	*ZA20D2*	Zinc finger A20 domain containing protein 2 (Zinc finger protein 216).	9q21.13	ENSG00000107372	H200000830
435					H200018807
436			20p11.21	ENSG00000101004	H200015397
437	*MTMR2*	Myotubularin-related protein 2 (EC 3.1.3.-).	11q21	ENSG00000087053	H200008529
438	*PRPF18*	Pre-mRNA splicing factor 18 (PRP18 homolog) (hPRP18).	10p13	ENSG00000165630	H200014051
439	*CDC42EP 1*	CDC42 effector protein 1 (Serum protein MSE55).	22q13.1	ENSG00000128283	H200013626
440	*FAM8A1*	Autosomal Highly Conserved Protein [Homo sapiens].	6p22.3	ENSG00000137414	H200010692
441	*NRD1*	Nardilysin precursor (EC 3.4.24.61) (N-arginine dibasic convertase) (NRD convertase) (NRD-C).	1p32.3	ENSG00000078618	H200000873
442	*LASS2*	LAG1 longevity assurance homolog 2 (Tumor metastasis-suppressor gene 1 protein) (SP260).	1q21.2	ENSG00000143418	H200008861
443	*C19orf33*	Immortalization-up-regulated protein (Hepatocyte growth factor activator inhibitor type 2-related sm	19q13.2	ENSG00000167644	H200020296
444	*ASPSCR1*	alveolar soft part sarcoma chromosome region, candidate 1, ASPL protein, renal cell carcinoma gene o	17q25.3	ENSG00000169696	H200009864
445	*ALG2*	Alpha-1,3-mannosyltransferase ALG2 (EC 2.4.1.-) (GDP- Man:Man(1)GlcNAc(2)-PP-dolichol mannosyltransf	9q22.33	ENSG00000119523	H200004499
446	*TP53RK*	TP53 regulating kinase (EC 2.7.1.37) (p53-related protein kinase) (Nori-2).	20q13.12	ENSG00000172315	H200017627
447	*SSR1*	Translocon-associated protein, alpha subunit precursor (TRAP-alpha) (Signal sequence receptor alpha	6p24.3	ENSG00000124783	H200016256
448	*CYLN2*	cytoplasmic linker 2 isoform 1, Williams-Beuren syndrome chromosome region 4 [Homo sapiens].	7q11.23	ENSG00000106665	H200011328
449	*VIP, CPAMD8*	Vasoactive intestinal peptide precursor (VIP).,C3 and PZP-like, alpha-2-macroglobulin domain containing 8, alpha-2 macroglobulin family protein VIP	6q25.2, 19p13.11	ENSG00000146469, ENSG00000160111	H200012267
450			7q34	ENSG00000006530	H200016499
451		HBxAg transactivated protein 2 [Homo sapiens].	1q24.3	ENSG00000117523	H200005686
452	*MRPL2*	mitochondrial ribosomal protein L2 [Homo sapiens].	6p21.1	ENSG00000112651	H200005144
453			22q11.21	ENSG00000099972	H200017971
454	*PACSIN2*	Protein kinase C and casein kinase substrate in neurons protein 2.	22q13.2	ENSG00000100266	H200002767
455	*ZNF553*	zinc finger protein 553 [Homo sapiens].	16p11.2	ENSG00000180035	H200002116
456					H200020420
457					H200009382
458	*C4orf14*		4q12	ENSG00000084092	H200001673
459	*ICAM3*	Intercellular adhesion molecule-3 precursor (ICAM-3) (ICAM-R) (CDw50) (CD50 antigen).	19p13.2	ENSG00000076662	H200011041
460	*SF3A2*	Splicing factor 3A subunit 2 (Spliceosome associated protein 62) (SAP 62) (SF3a66).	19p13.3	ENSG00000104897	H200011943
461	*EPHB3*	Ephrin type-B receptor 3 precursor (EC 2.7.1.112) (Tyrosine-protein kinase receptor HEK-2).	3q27.1	ENSG00000182580	H200000705
462			1p36.13	ENSG00000169991	H200002318
463					H200013074
464		Protein AF1q.	1q21.2	ENSG00000143443	H200006209
465	*RWDD1*	RWD domain containing protein 1 (CGI-24) (PTD013).	6q22.1	ENSG00000111832	H200003183

466	*NDST2*	Heparin sulfate N-deacetylase/N-sulfotransferase (EC 2.8.2.-) (N-HSST) (N-heparin sulfate sulfotrans	10q22.2	ENSG00000166507	H200006518
467	*PRKCZ*	Protein kinase C, zeta type (EC 2.7.1.37) (nPKC-zeta).	1p36.33	ENSG00000067606	H200006556
468	*PAPD1*	PAP associated domain containing 1 [Homo sapiens].	10p11.23	ENSG00000107951	H200008184
469	*MTRF1*	Peptide chain release factor 1, mitochondrial precursor (MRF-1).	13q14.11	ENSG00000120662	H200006786
470			19q13.31	ENSG00000105771	H200001886
471	*KCND1*	Potassium voltage-gated channel subfamily D member 1 (Voltage-gated potassium channel subunit Kv4.1)	Xp11.23	ENSG00000102057	H200005164
472	*UPF2*	UPF2 regulator of nonsense transcripts homolog, regulator of nonsense transcripts 2, yeast Upf2p hom	10p14	ENSG00000151461	H200000853
473			7q34	ENSG00000006459	H200015246
474		down-regulated in metastasis [Homo sapiens].	12q23.2	ENSG00000120800	H200008329
475	*C20orf23*	Kinesin-like motor protein C20orf23 (Sorting nexin 23).	20p12.1	ENSG00000089177	H200011141
476					H200011386
477	*SGSH*	N-sulphoglucosamine sulphohydrolase precursor (EC 3.10.1.1) (Sulfoglucosamine sulfamidase) (Sulphami	17q25.3	ENSG00000181523	H200004063
478	*MPP1*	55 kDa erythrocyte membrane protein (p55) (Membrane protein, palmitoylated 1).	Xq28	ENSG00000130830	H200000477
479	*THUMPD2*	THUMP domain containing protein 2.	2p22.1	ENSG00000138050	H200004437
480	*DGCR8*	DGCR8 protein (DiGeorge syndrome critical region 8).	22q11.21	ENSG00000128191	H200017967
481					H200012395
482					H200008747
483	*B4GALT1*	Beta-1,4-galactosyltransferase 1 (EC 2.4.1.-) (Beta-1,4-GalTase 1) (Beta4Gal-T1) (b4Gal-T1) (UDP-gal	9p21.1	ENSG00000086062	H200014787
484	*RSU1*	Ras suppressor protein 1 (Rsu-1) (RSP-1).	10p13	ENSG00000148484	H200006130
485			7q11.21	ENSG00000198874	H200002559
486					H200009822
487	*RAD51C*	DNA repair protein RAD51 homolog 3.	17q23.2	ENSG00000108384	H200001999
488	*TBC1D16*	TBC1 domain family, member 16 [Homo sapiens].	17q25.3	ENSG00000167291	H200010522
489					H200012718
490	*DRG1*	Developmentally regulated GTP-binding protein 1 (DRG 1).	22q12.2	ENSG00000185721	H200011945
491	*ARID1A*	SWI/SNF-related, matrix-associated, actin-dependent regulator of chromatin subfamily F member 1 (SWI	1p36.11	ENSG00000117713	H200012438
492	*TNFSF5IP1*	tumor necrosis factor superfamily, member 5-induced protein 1, hepatocellular carcinoma susceptibili	18p11.21	ENSG00000128789	H200000820
493	*MRPS22*	Mitochondrial 28S ribosomal protein S22 (S22mt) (MRP-S22) (GK002).	3q23	ENSG00000175110	H200011494
494	*ZDHHC6*	Zinc finger DHHC domain containing protein 6 (Zinc finger protein 376) (Transmembrane protein H4).	10q25.2	ENSG00000023041	H200003136
495					H200009543
496	*CHD8*	Chromodomain-helicase-DNA-binding protein 8 (CHD-8) (Helicase with SNF2 domain 1) (Fragment).	14q11.2	ENSG00000100888	H200008140
497	*KIAA1618*		17q25.3	ENSG00000180843	H200018846
498	*BHLHB2*	Class B basic helix-loop-helix protein 2 (bHLHB2) (Differentially expressed in chondrocytes protein	3p26.1	ENSG00000134107	H200007994
499	*USP5*	Ubiquitin carboxyl-terminal hydrolase 5 (EC 3.1.2.15) (Ubiquitin thiolesterase 5) (Ubiquitin-specifi	12p13.31	ENSG00000111667	H200000826
500	*ESRRAP*	28S ribosomal protein S31, mitochondrial precursor (S31mt) (MRP-S31) (Imogen 38).	13q14.11	ENSG00000102738	H200013982
501	*FNTA*	Protein farnesyltransferase/geranylgeranyltransferase type I alpha subunit (EC 2.5.1.58) (EC 2.5.1.5	8p11.21	ENSG00000168522	H200020470
502			2q31.1	ENSG00000138382	H200016971
503	*ADCK4*	aarF domain containing kinase 4 [Homo sapiens].	19q13.2	ENSG00000123815	H200012914
504	*BCCIP*	BRCA2 and CDKN1A-interacting protein isoform BCCIPalpha, BRCA2 and CDKN1A-interacting protein, cdk i	10q26.2	ENSG00000107949	H200017518
505	*METAP1*	Methionine aminopeptidase 1 (EC 3.4.11.18) (MetAP 1) (MAP 1) (Peptidase M 1).	4q23	ENSG00000164024	H200006871
506	*C10orf97*		10p13	ENSG00000148481	H200008813

507			1p13.2	ENSG00000143079	H200003341
508	*CSPG6*	Structural maintenance of chromosome 3 (Chondroitin sulfate proteoglycan 6) (Chromosome-associated p	10q25.2	ENSG00000108055	H200003409
509	*ORC5L*	Origin recognition complex subunit 5.	7q22.1	ENSG00000164815	H200013856
510	*AKAP1*	A kinase anchor protein 1, mitochondrial precursor (Protein kinase A anchoring protein 1) (PRKA1) (A	17q23.2	ENSG00000121057	H200006583
511			19q13.43	ENSG00000105136	H200018815
512	*CDC16*	Cell division cycle protein 16 homolog (CDC16Hs) (Anaphase promoting complex subunit 6) (APC6) (Cycl	13q34	ENSG00000130177	H200000418

for Operon IDs http://omad.operon.com/human2/index.php
for ensembl IDs http://www.ensembl.org/Homo_sapiens/textview

E. Publications and Patent

01/2008 *Schneeweiss A, ***Thuerigen O**, …, Lichter P, Toedt G.

In preparation

* Authors contributed equally.

11/2007 Pfister S, Rea S, Taipale M, Mendrzyk F, Straub B, Ittrich C, **Thuerigen O**, Sinn HP, Akhtar A, Lichter P.

The histone acetyltransferase hMOF is frequently downregulated in primary breast carcinoma and medulloblastoma and constitutes a biomarker for clinical outcome in medulloblastoma

International Journal of Cancer: [Epub ahead of print]; in press

04/2006 ***Thuerigen O**, *Schneeweiss A, Toedt G, Warnat P, Hahn M, Kramer H, Brors B, Rudlowski C, Benner A, Schuetz F, Tews B, Eils R, Sinn HP, Sohn C, Lichter P.

Gene Expression Signature Predicting Pathologic Complete Response with Gemcitabine, Epirubicin, and Docetaxel in Primary Breast Cancer

Journal of Clinical Oncology, Vol. 24(12):1839-45

* Authors contributed equally.

02/2005 Schlingemann J, **Thuerigen O**, Ittrich C, Toedt G, Kramer H, Hahn M, Lichter P.

Effective Transcriptome Amplification for Expression Profiling on Sense-oriented Oligonucleotide Microarrays

Nucleic Acids Research, Vol. 33(3):e29

01/2004 **Thuerigen O**, Schlingemann J, Hahn M, Lichter P.

T7 amplification and cDNA Klenow labeling for expression (TAcKLE) analysis with oligonucleotide microarrays

European Patent No. 04 000 343.6