Expression Profiling by DNA Microarrays: Development of Amplification Methods for the Analysis of Minimal Tumor Samples

DISSERTATION

submitted to the
Combined Faculties for the Natural Sciences and Mathematics
of the Ruperto-Carola-University of Heidelberg
for the degree of

Doctor of Natural Sciences

presented by

DIPL.-BIOL. JÖRG SCHLINGEMANN

born in Hagen in 1975

2005

DISSERTATION

submitted to the
Combined Faculties for the Natural Sciences and Mathematics
of the Ruperto-Carola-University of Heidelberg
for the degree of

Doctor of Natural Sciences

presented by

DIPL.-BIOL. JÖRG SCHLINGEMANN

born in Hagen in 1975

oral examination: September 30th, 2005

Expression Profiling by DNA Microarrays: Development of Amplification Methods for the Analysis of Minimal Tumor Samples

Referees:

PD Dr. Karsten Rippe
Prof. Dr. Peter Lichter

I. TABLE OF CONTENTS PAGE

II. PUBLICATIONS, PEER-REVIEWED

Schlingemann, J., Habtemichael, N., Ittrich, C., Toedt, G., Kramer, H., Hambek, M., Knecht, R., Lichter, P., Stauber, R. & Hahn, M. Patient-based cross-platform comparison of oligonucleotide microarray expression profiles. *Lab. Invest.* **85**, 1024-1039 (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 Res.* **33**, e29 (2005).

Schlingemann, J., Hess, J., Wrobel, G., Breitenbach, U., Gebhardt, C., Steinlein, P., Kramer, H., Furstenberger, G., Hahn, M., Angel, P. & Lichter, P. Profile of gene expression induced by the tumour promotor TPA in murine epithelial cells. *Int. J. Cancer* **104**, 699-708 (2003).

Wrobel, G., **Schlingemann, J.**, Hummerich, L., Kramer, H., Lichter, P. & Hahn, M. Optimization of high-density cDNA-microarray protocols by 'design of experiments'. *Nucleic Acids Res.* **31**, e67 (2003).

III. PUBLICATIONS, NON PEER-REVIEWED

Kramer, H. & **Schlingemann, J.** RNA-Amplifikation für Oligonukleotid-Microarrays und deren praktische Anwendung in der Krebsforschung. *MTA Dialog* **6**, 350-353 & 410-413 (2005).

Fritz, B., Kokocinski, F., **Schlingemann, J.** & Hahn, M. Anwendungen der DNA-Chiptechnologie in der Krebsforschung. *Biospektrum* **9**, 78-83 (2003).

IV. POSTERS

Hummerich, L., **Schlingemann, J.**, Hess, J., Mueller, R., Breitenbach, U., Wrobel, G., Kokocinski, F., Fuerstenberger, G., Hahn, M., Angel, P. & Lichter, P. Identification of tumour associated genes in the process of skin carcinogenesis. *Genomics and Cancer Conference*, Heidelberg (Mai 2003).

Schlingemann, J., Wrobel, G., Hahn, M., Breitenbach, U., Hess, J., Angel, P. & Lichter, P. A Comprehensive Murine cDNA Microarray Applied for Expression Profiling of TPA Induced Epithelial Cells of the Skin. *Dechema Meeting 'Status of Chip Technologies'*, Frankfurt am Main (Februar 2003).

Schlingemann, J., Hess, J., Wrobel, G., Breitenbach, U., Gebhardt, C., Steinlein, P., Hummerich, L., Kokocinski, F., Kramer, H., Fuerstenberger, G., Hahn, M. & Lichter, P. Profile of Gene Expression Induced by TPA in Murine Epithelial Cells. *NGFN/DHGP Syposium*, Berlin (November 2002).

Hummerich, L., **Schlingemann, J.**, Hess, J., Breitenbach, U., Wrobel, G., Kokocinski, F., Fuerstenberger, G., Hahn, M., Angel, P. & Lichter, P. Expression profiling of epithelial cells in skin cancerogenesis. *NGFN/DHGP Symposium*, Berlin (November 2002).

Thuerigen, O., Hummerich, L., **Schlingemann, J.**, Wrobel, G., Lichter, P. & Hahn, M. Evaluation of Printed Oligonucleotide Microarrays for Large-scale Expression Profiling in Human Applications. *NGFN/DHGP Symposium*, Berlin (November 2002).

Wrobel, G., **Schlingemann, J.**, Hummerich, L., Kramer, H., Lichter, P. & Hahn, M. Refinement of Spotting Conditions for High-Density cDNA-Microarrays. *NGFN/DHGP Symposium*, Berlin (November 2002).

Schlingemann, J., Hummerich, L., Hahn, M. & Lichter, P. Auf der Suche nach den Krebsgenen - Roboter, Chips und das Genom. *DKFZ 'Tag der offenen Tür'*, Heidelberg (Oktober 2002).

V. ACKNOWLEDGEMENTS

I wish to thank Prof. Lichter for providing the research subject and for his thoughtful supervision. PD Dr. Karsten Rippe is acknowledged for acting as referee for this thesis.

I sincerely thank Dr. Meinhard Hahn for his kind and qualified support, for valuable comments as well as frequent proofreading of both this dissertation and the scientific publications that originated from this work.

Drs. Roland Stauber and Negusse Habtemichael are gratefully acknowledged for providing tissue samples as well as RQ-PCR data in the course of a very productive collaboration.

Many thanks go to Olaf Thuerigen for his both practical and theoretical contributions to the development and evaluation of the TAcKLE protocol.

Daniel Haag deserves my gratitude for his dedication and commitment towards his work on RCA-based microarray expression analysis.

Prof. Ulf Landegren, Jonas Jarvius and Olle Ericsson are acknowledged for their kind cooperation and continuous contributions concerning the RCA protocol and for enabling a working visit in Uppsala.

Special thanks go to Axel Benner and particularly to Dr. Carina Ittrich, who reliably and proficiently supported me in the analysis of microarray experiments. Their knowledge and helpfulness made a substantial contribution to the success of this thesis.

I would like to express my heartfelt thanks to Heidi Kramer for her qualified and reliable assistance in conducting many of the experiments shown below and for very private matters...

Finally, I wish to express my profound gratitude to my parents for their almost endless support of my education and for being at my side in any circumstance of life.

VI. ABBREVIATIONS

aRNA	antisense RNA
A	Ampère
AMV	avian myeloblastosis virus
b	base
bp	basepair
BSA	bovine serum albumin
CAS #	chemical abstracts service registry number
Cat #	catalogue number
cDNA	copy DNA
cds.	coding sequence
Cy3	cyanine 3-dUTP
Cy5	cyanine 5-dUTP
ddH$_2$O	double distilled water
DKFZ	Deutsches Krebsforschungszentrum
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
dNTP	2’-deoxynucleoside-5’-triphosphate
dATP	2’-deoxyadenosine-5’-triphosphate
dCTP	2’-deoxycytidine-5’-triphosphate
dGTP	2’-deoxyguanosine-5’-triphosphate
dTTP	2’-deoxythymidine-5’-triphosphate
DTT	dithiothreitol
ds	double-stranded
DWD	distance weighted discrimination
E. coli	*Escherichia coli*
EDTA	ethylenediaminetetraacetic acid, disodium salt
EST	expressed sequence tag
EtBr	ethidium bromide
EtOH	ethanol
Exp.	experiment
Fig.	figure

GO	gene ontology
HIV	human immunodeficiency virus
HNSCC	head and neck squamous cell carcinoma
IVT	*in vitro* transcription
min	minute
MMLV	Moloney mouse leukemia virus
mRNA	messenger RNA
N	A / C / G / T (degenerated DNA code)
NaAc	sodium acetate
NCBI	National Center for Biotechnology Information
OD	optical density
PCR	polymerase chain reaction
Pfu	*Pyrococcus furiosus*
PMT	photomultiplier tubes
RNA	ribonucleic acid
RNase	ribonuclease
rpm	revolutions per minute
RQ-PCR	reverse transcription quantitative real-time PCR
RT	room temperature, or reverse transcription
RZPD	Resource Center (for Genome Research) and Primary Database
rxn	reaction
SDS	sodium dodecyl sulfate
SPA	single primer amplification
SPF	specified pathogen free
ss	single stranded
SSC	saline sodium citrate
SSII	SuperScript II
Tab.	table
TAcKLE	T7 Amplification and Klenow Labeling for Expression Analysis
Taq	*Thermus aquaticus*
TBE	Tris/HCl-borate-EDTA-buffer
TE	Tris/HCl-EDTA-buffer
Tris	tris(hydroxymethyl)aminomethane
TS	template switch

UV	ultraviolet
V	A / C / G (degenerated DNA code)
vol.	volume
v/v	volume/volume-ratio
w/v	weight/volume-ratio

VII. ABSTRACT

Recently, microarrays of synthetic long sense-oriented oligonucleotides were introduced as an alternative expression profiling platform with distinct advantages to both cDNA arrays and commercial arrays produced by *in situ* synthesis of multiple short oligonucleotides per gene. However, gene expression analysis using microarrays of long oligonucleotides is limited in that it requires substantial amounts of RNA. The objective of this thesis was to develop protocols that allow for the analysis of gene expression even in minimal samples. Two different approaches were taken, one that amplifies the RNA target material before hybridization and another that amplifies the signal generated on the array. Most existing target amplification protocols linearly amplify mRNA by cDNA synthesis and *in vitro* transcription. Since orientation of the product is antisense (aRNA), it is inapplicable for dye-labeling by reverse transcription and hybridization to sense-oriented oligonucleotide arrays. Here, a novel protocol (TAcKLE) is introduced in which a combination of two reverse and one forward transcription reactions followed by dye-incorporation using the Klenow fragment of *E. coli* DNA polymerase I generates fluorescent antisense cDNA. This protocol provides high fidelity and up to 10^5-fold amplification, starting from 2 ng total RNA. The generated data are highly reproducible and maintain relative gene expression levels between samples.

Signal amplification is another option if only minimal amounts of sample material are available. Therefore, a method was evaluated that uses on-chip rolling circle replication of circularized oligonucleotides for the amplified detection of gene expression profiles. This principle should allow for a faster and cheaper experimental procedure, circumventing sequence-dependent amplification bias. The preliminary results provide evidence for the method’s applicability, but further experiments are required to reduce the required amount of starting material and to define a stable protocol.

As the TAcKLE protocol performed particularly well, it was subsequently applied to evaluate the utility of spotted oligonucleotide microarrays compared to a widely-used and accepted commercial reference platform. There are numerous ways to perform global transcriptional profiling, among which microarray technology has certainly gained a premier position. The comparison of gene expression measurements obtained with different array-based approaches is therefore of substantial interest in order to clarify whether inter-platform differences may conceal biologically significant information. To address this concern, global gene expression was analyzed in a set of clinical head and neck squamous cell carcinoma samples, using both spotted oligonucleotide microarrays made from a large collection of 70-mer probes and commercial arrays produced by *in situ* synthesis of sets of multiple 25-mer oligonucleotides per gene. Expression measurements were compared for 4,425 genes represented on both platforms, which revealed strong correlations between the corresponding data sets and similar profiles of relative gene expression.

In conclusion, combining the TAcKLE protocol with spotted oligonucleotide arrays is an attractive alternative for transcriptional profiling of limited source material, offering a high potential for gene expression analysis in a multitude of disease situations.

VIII. ZUSAMMENFASSUNG

Microarrays bestehend aus langen, *sense*-orientierten Oligunukleotiden sind seit kurzem als eine Alternative zu cDNA-*Arrays* und *Arrays* kurzer, *in-situ*-hergestellter Oligonukleotide erhältlich, welche deutliche Vorteile zu beiden anderen Systemen aufweisen. Das Spektrum wissenschaftlicher Fragestellungen, das mit Hilfe dieser neuartigen *Microarray*-Technologie bearbeitet werden kann, ist allerdings limitiert durch die großen RNA-Mengen, die für die Experimente benötigt werden. Ziel dieser Arbeit war die Entwicklung von Methoden, die das Erstellen von Genexpressionsprofilen auch bei Fragestellungen mit limitierter RNA-Menge erlauben. Es wurden hierzu zweierlei Ansätze gewählt, nämlich einerseits die Amplifikation des Ausgangsmaterials, und andererseits die Verstärkung des Signals direkt auf dem *Array*.

Etablierte Amplifikationsprotokolle schreiben die zu analysierende mRNA zunächst in cDNA um und amplifizieren sie anschließend durch *in-vitro*-Transkription. Sie produzieren so allerdings RNA in *antisense*-Orientierung, die nach reverser Transkription in fluoreszenzmarkierte, *sense*-orientierte cDNA nicht auf *sense*-orientierte Oligonukleotid-Sonden hybridisiert werden kann. Im Rahmen dieser Arbeit wurde daher ein neuartiges Protokoll (*TAcKLE*) entwickelt, in welchem eine Kombination von zwei reversen Transkriptionen, einer nicht-reversen Transkription und einer Markierungsreaktion mittels Klenow-Fragment fluoreszenzmarkierte cDNA in *antisense*-Orientierung erzeugt.

Die mit diesem Protokoll generierten Daten sind reproduzierbar und geben relative Expressionsniveaus wahrheitsgemäß wieder. Der maximale Amplifikationsfaktor liegt bei 10^5, bei einer minimalen Ausgangsmenge von lediglich 2 ng gesamt-RNA.

Eine weitere Option bei limitiertem Untersuchungsmaterial ist die Verstärkung des Signals, welches auf dem Array generiert wird. Im Rahmen dieser Arbeit wurde eine Methode erarbeitet, bei der diese Signalamplifikation durch *in-situ*-Replikation zirkulärer Oligonukleotide (*rolling circle replication*) auf dem Array erzielt wird. Dieses Prinzip ermöglicht eine schnellere und billigere Durchführung der Experimente und vermeidet sequenzbedingte Verzerrungen der Ergebnisse. Weiterführende Experimente sind allerdings notwendig, um den Amplifikationsfaktor zu erhöhen und ein stabiles Protokoll zu etablieren.

Da mit dem *TAcKLE*-Protokoll besonders gute Ergebnisse erzielt worden waren, wurde es im Anschluss verwendet, um die Nützlichkeit und Leistungsfähigkeit selbst hergestellter Oligonukleotid-*Microarrays* im Vergleich zu einem etablierten, kommerziellen System zu untersuchen. Es gibt verschiedenste Methoden für eine globale Expressionsanalyse, unter welchen die *Microarray*-Technologie sicherlich am häufigsten zum Einsatz kommt. Der Vergleich von Expressionsprofilen, die mit unterschiedlichen Varianten dieser Methodik erstellt wurden, sollte klären, ob plattformspezifische Unterschiede biologische Daten überlagern können. Es wurden diesbezüglich die Genexpressionsprofile von sechs *HNSCC*-Tumoren bestimmt, und zwar sowohl mittels selbst produzierter 70-mer Oligonukleotid-*Microarrays*, die pro Gen nur ein Sondenmolekül verwenden, als auch über kommerzielle, durch *in-situ*-Synthese hergestellte *Arrays*, auf denen pro Gen mehrere 25-mer-Oligonukleotide aufgebracht sind. Die Expressionsdaten wurden verglichen für insgesamt 4.425 Gene, die auf beiden Plattformen vertreten waren. Die Korrelationen unter den Datensätzen waren sehr gut, und es wurden mit beiden Ansätzen sehr ähnliche Genexpressionsprofile erhalten.

Die Kombination aus *TAcKLE*-Protokoll und selbst produzierten 70-mer-*Arrays* ist somit eine attraktive Alternative für die Expressionsanalyse von limitiertem Probenmaterial, deren hohes Potential für die wissenschaftliche Untersuchung bei einer Vielzahl von Erkrankungen Verwendung finden kann.

1 Introduction

The beginning is half the whole.

ARISTOTELES

The completion of the Human Genome Project (HGP) in April 2003[1-4] was a landmark event, making the genomic era a reality. Notably, the human genome seems to encode only 20,000 - 25,000 protein-coding genes, but the exact number is still a matter of investigation, as is the function of the majority of genes. Functional analysis of these genes is a scientific and technical challenge which requires the use of novel high-throughput methods.

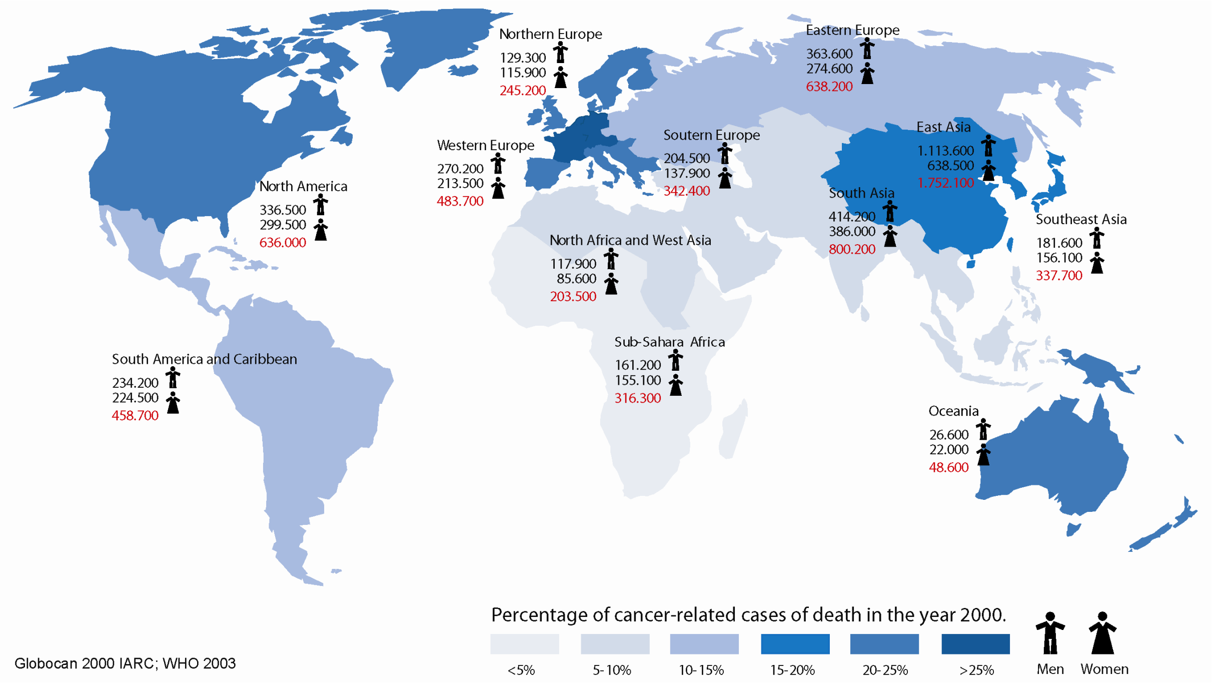
DNA microarray technology is a powerful approach for the parallel expression analysis of thousands of genes. It provides a comprehensive and accurate snapshot of gene expression in the analyzed samples and gave rise to the term and concept of the “transcriptome”, which denotes the entirety of transcripts present in a given sample at a given time. On average, a cell uses only about 5% of its genes at the same time, whereas the remaining 95% are repressed at the transcriptional or, less commonly, at the translational level[5]. Transcriptome analysis by means of expression profiling on DNA microarrays can therefore provide an image of a cell’s differentiation state, its function and its phenotype. This can help to clarify a multitude of biological and medical questions, including cellular functions, biochemical reaction cascades or regulatory mechanisms.

Many human diseases associate with abnormal changes in gene expression. Expression analysis of diseased cells or tissues can reveal pathomechanisms, open up new concepts for therapy, improve diagnosis and substantiate prognosis. Profound genetic reprogramming can, *e.g.*, be found in cancer, which is responsible for one in eight cases of death worldwide[6].

1.1 Cancer Statistics

Cancer is the second most frequent cause of death worldwide (12.6%) and is only exceeded by circulatory diseases [6]. More people dye of cancer than of AIDS, malaria and tuberculosis taken together. On average, worldwide, there is about a 10% chance of getting a cancer before the age of 65 [7]. There were an estimated 10.1 million new cases (incidence), 6.2 million deaths (mortality) (Fig. 1) and 22.4 million persons living with cancer (prevalence) in the year 2000 [7]. No attempt has been made to estimate the incidence or mortality of non-melanoma skin cancer due to the lack of reliable data, and these tumors are therefore excluded from average calculations.

The estimate of the global cancer burden for the year 2000 [7] presents an increase of around 22% in incidence and mortality since the previous comprehensive estimate for 1990; the prospected incidence for the year 2020 is 15.7 million new cases, which would mean an increase of 50% compared to the year 2000. This can in part be explained by an increased overall life expectancy and the constantly growing world population.



Globocan 2000 IARC; WHO 2003

Figure 1. Global cancer mortality 2000. The color coding indicates the respective percentage of cancer-related cases of death in the year 2000. In addition, the absolute numbers of deaths are given as totals for major regions of the world. The data were gathered from the Globocan 2000 database of the International Agency for Research on Cancer (IARC), which is part of the World Health Organization (WHO). The figure was originally taken and modified from the website of the Deutsche Krebshilfe e.V. (http://www.krebshilfe.de).

Lung cancer is the main cancer in the world today, both in terms of numbers of cases (1.2 million) or deaths (1.1 million), because of the high case fatality (ratio of mortality / incidence = 0.9) (Fig. 2). However, breast cancer, although it is the second most common cancer overall (1.05 million new cases) ranks much lower as the 5th most common cause of death by cancer because of the relatively favorable prognosis (ratio of mortality / incidence = 0.4). In terms of prevalence (the proportion of a population that has a disease at a given point in time; for cancer, the most common specification is 5-year prevalence, which denotes those cases neither dead nor considered cured until 5 years after the initial diagnosis), the most frequent cancers are breast (3.9 million cases), colorectal cancers (2.4 million) and prostate (1.6 million)[7].

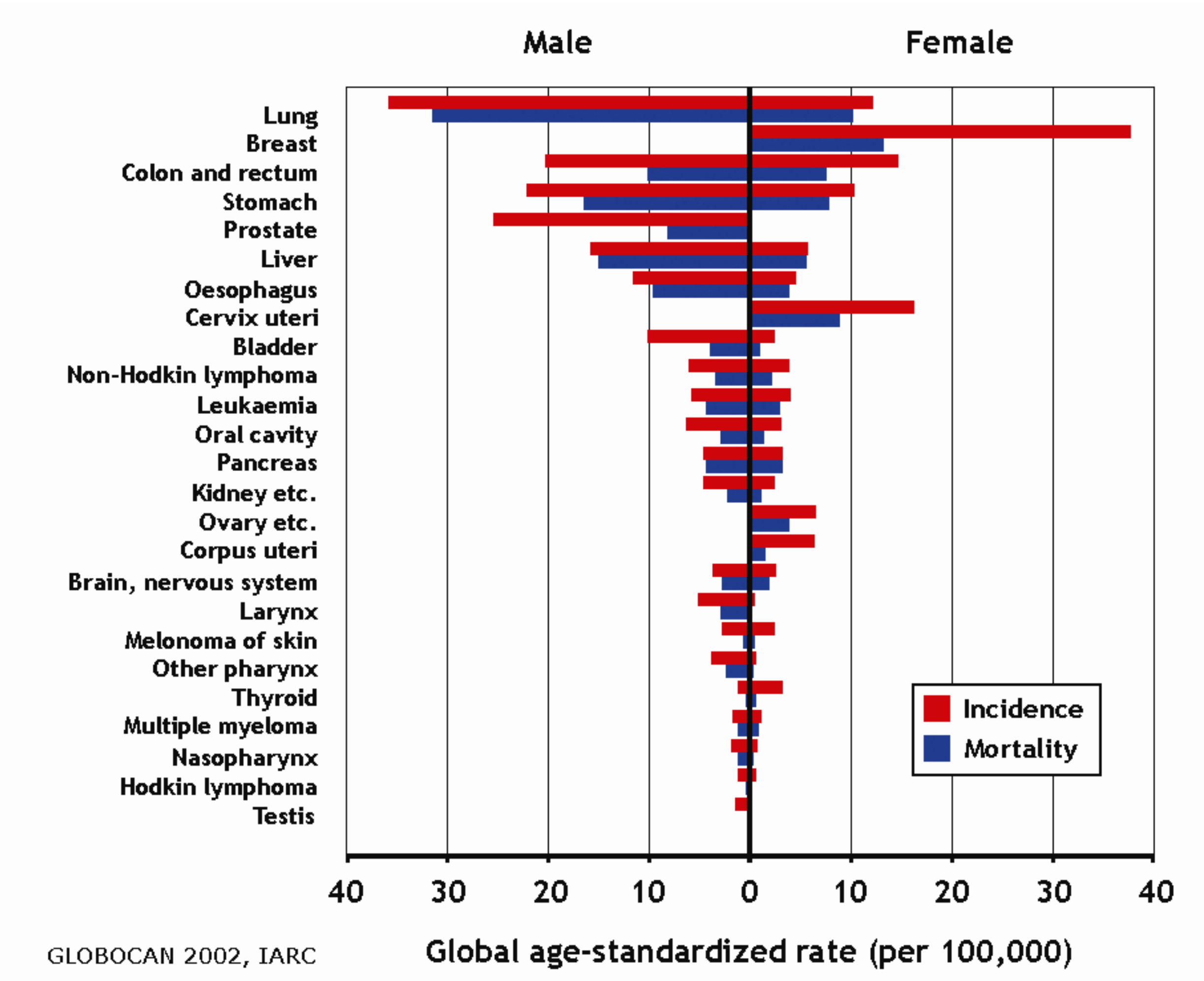


Figure 2. Global age-standardized incidence and mortality rates for the most frequent types of cancer in the year 2002. The data were extracted online from the Globocan 2002 database at the CancerMondial website (http://www-dep.iarc.fr/) of the IARC. Regional statistics may exhibit considerable deviations from this generalized compilation. There is an increased incidence of breast and colorectal cancers in more developed countries as compared to less developed countries, whereas the latter have higher rates for liver, stomach and cervical cancers, partly due to the elevated risk of infections.

Statistics on German cancer mortality in the year 2000 (Fig. 3) confirm that primarily senior citizens are affected (Fig. 3A), the average age at diagnosis being 65-67 years. But whereas the worldwide summarized mortality rate increased considerably during the decade 1990-2000[7], the German rate revealed a downward-trend (Fig. 3B). Looking more closely at various common types of cancer, one can find that the prospects improved clearly for stomach cancer or male lung cancer, remained almost unchanged for malignancies of the female breast or the prostate, and regrettably worsened for female lung cancer (Fig. 3C). The latter can probably be explained by changes in female smoking habits (more and more women are smokers).

Considering all these numbers and the suffering of people that they represent, it is one of the most urgent objectives of biomedical research to gain a detailed understanding of the causes and the processes involved in carcinogenesis and to use these insights to develop new concepts for an effective therapy.

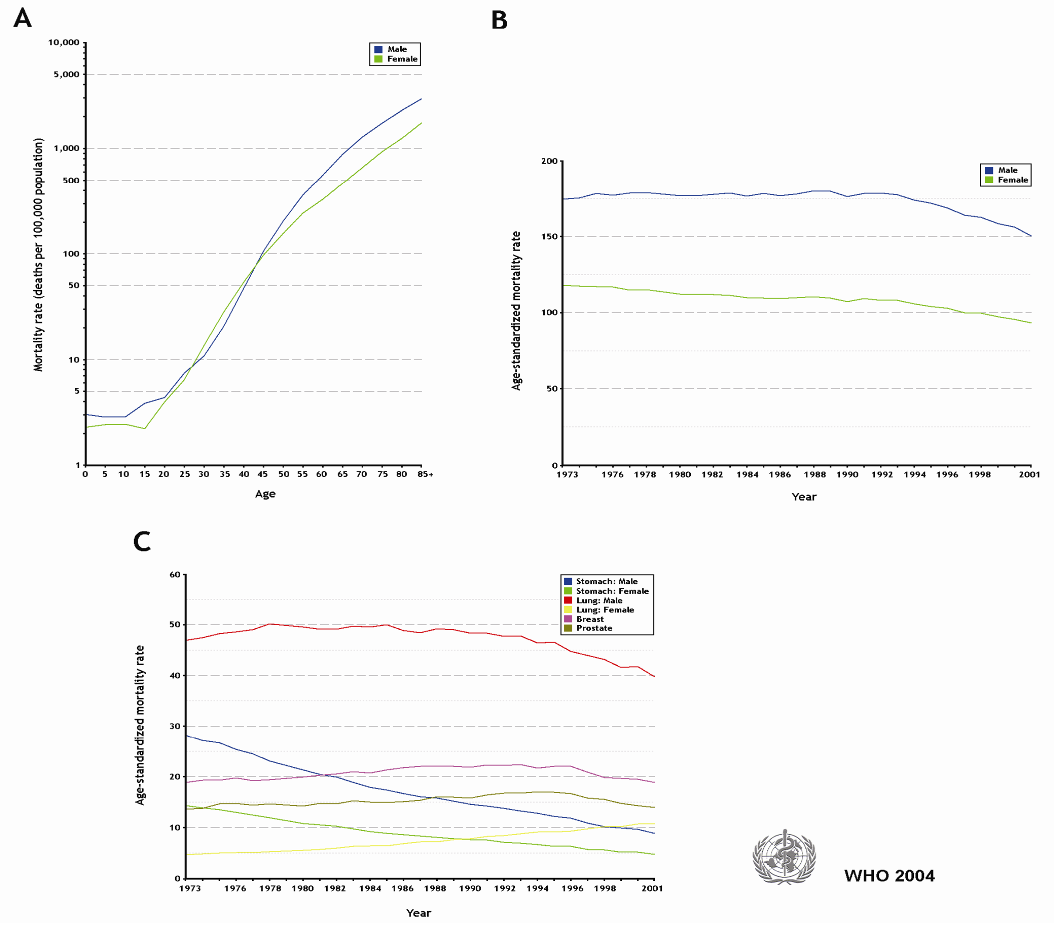


Figure 3. German cancer mortality statistics for the year 2000. (**A**) Age-specific mortality rates for men and women, averaged over "all cancers" excluding non-melanoma skin cancer. (**B**) Trends of age-standardized mortality rates for "all cancers" excluding non-melanoma skin cancer (**C**) Trends of age-standardized mortality rates for various common types of cancer. Source: CancerMondial / IARC (http://www-dep.iarc.fr/).

1.2 The Human Genome

Many human diseases have a causal connection to changes in the genetic constitution and may therefore be seen as diseases of the genome. But whereas classical hereditary diseases trace back to mutational events in germ line cells of the parental generation (or earlier), thereby affecting all somatic cells of their offspring, cancer is caused by the accumulation of genetic changes in a single somatic cell, which collectively promote its clonal expansion.

The human genome (entirety of genetic information) consists of nuclear as well as mitochondrial DNA. The nuclear genome contains more than 99.99% of the total genetic information, most of which specifies protein synthesis on cytoplasmic ribosomes. Nuclear DNA is arranged in 46 chromosomes representing 23 pairs of chromosomes. This condition is called a diploid set of chromosomes and consists of a haploid set from each parent, each comprising 22 autosomes as well as one gonosome. The haploid human genome spans approximately 3.08×10^9 bp, the euchromatic portion being 2.88×10^9 bp in length [8].

It is currently estimated that approximately 20,000 - 25,000 genes exist in the human genome [8,9], but this number has been a source of continued controversy with other estimates reaching as high as 120,000 genes [10], which is almost certainly much too high [11-16]. Of the estimated genes, 42% have an unknown function. The average size of a human gene is around 27,000 bp, with typical ranges between 20,000 and 50,000 bp. However, only about 1,300 bp are required to encode an average-sized human protein of approx. 430 amino acids [17]. The vast majority of protein-coding genes from higher eukaryotes consist of both protein-coding sequences (exons) and sequences that do not code for protein (introns). The transcripts of these genes are called pre-mRNA (precursor-mRNA), from which the intervening intron sequences are removed in a process termed splicing. After some further processing (capping and tailing), mature mRNA exits the nucleus and is translated in the cytoplasm. The average number of exons in human genes ranges between 8 - 9. The average size of an exon is around 150 nucleotides, whereas introns average approx. 3,500 nucleotides and can be as much as 100 times larger [17,18].

Coding regions in the human genome are estimated to account for only around 3% of the total DNA sequence, intronic sequences together with pseudogenes (nonfunctional homologues of functional genes) and gene fragments contribute

~22%, and intergenic regions account for the remaining ~75%[19]. The proportion of non-coding DNA in humans is particularly striking when compared to other metazoan eukaryotes. For example, the human genome is 30 x larger than the genomes of *Caenorhabditis elegans* (nematode, roundworm)[20] and *Drosophila melanogaster* (fruit fly)[21], but has only ~2 - 3x as many genes. Sequence repeats (interspersed repetitive DNA, *i.e.*, SINEs, LINEs; tandemly repeated DNA, *i.e.*, megasatellite DNA, satellite DNA, ministatellite DNA, microsatellite DNA) are another very prominent feature of the human genome[19]. 35% of the entire human genome (including coding regions) is classified as repetitive, compared to only 10% in *Arabidopsis thaliana* (wall cress)[22]. If exclusively non-coding regions are considered, the proportion of repetitive DNA climbs to 46%[19].

Another important feature of the human (and other mammalian) genomes are CpG islands. The cytosines of most CpG dinucleotides in the human genome are methylated. However, methyl-cytosine frequently mutates to thymine via deamination[23] (whereas deamination of cytosine gives rise to uracil, which is easily recognized as foreign within the DNA strand and replaced), and so CpG dinucleotides tend to decay to TpG / CpA. This is why CpG dinucleotides are vastly underrepresented genome-wide compared to what would be expected by chance (five times less frequently). A CpG island is a region of DNA that has a higher relative proportion of CpG dinucleotides when compared to the entire genome, in which the predominant absence of methylation slows CpG decay. About 56% of human genes are associated with CpG islands[24]. CpG island methylation correlates with gene inactivation during gene imprinting[25], X-chromosome inactivation[26] and tissue specific gene expression[27]. The methyl group displaces transcription factors that normally bind to the DNA; and it attracts methyl-binding proteins, which in turn are associated with gene silencing and chromatin compaction, probably through interactions with complexes that modify the tails of histone proteins[28]. Histone proteins form octamers around which the DNA helix loops to form the nucleosome, the individual packaging unit of genomic DNA. The histone tails that extrude from the nucleosomes can be modified by methylation[29], acetylation[30], phosphorylation[31] or ubiquitylation[32] at different sites, creating potential combinations that have been referred to as the 'histone code'[33] in which gene regulatory information is encrypted. Cell-type-specific cytosine methylation and histone-tail modifications could contribute to the differences in gene expression patterns between cell types.

1.3 Genetic Aberrations of Cancer Cells

The formation of cancer is a consequence of genetic changes in somatic cells with a capability for cell division. However, not every mutation results in carcinogenesis, as very special kinds of genes have to be affected. Their gene products are generally involved in the cellular maintenance or repair machinery, or they fulfill key functions in control and/or regulatory mechanisms, which in normal cells modulate the rate of growth, proliferation and apoptosis according to the needs of the organism. The German Biologist Theodor Boveri was among first to realize the connection between cancer and genetic aberrations. Working on the fertilization of sea-urchin eggs by two sperms instead of one, he discovered that distribution of unequal numbers of chromosomes to the daughter cells results in specific characteristics that depend on the random combinations of inherited chromosomes. Whereas some daughter cells survive but develop abnormally, others have a genetic imbalance that is too severe for survival. Boveri concluded that individual chromosomes carry different information[34-36], and he suggested that tumors might likewise arise as a consequence of particular, incorrect chromosome combinations[37]. Boveri even postulated the existence of 'growth stimulatory' as well as 'growth inhibitory chromosomes', and he attributed the unlimited proliferation of tumor cells to an increase in the number of growth-promoting and/or to a physical removal of the growth-inhibiting chromosomes. The concept that cancer originates from defects in particular genes was originally proposed by Fritz Anders in 1967 and deduced from his work with breed hybrids of the fish *Platypoecilus maculatus* (southern platyfish) and *Xiphophorus helleri* (swordtail) from the genus *Xiphophorus*[38]. He showed that the occurrence and the degree of the so-called Gordon-Kosswig melanoma is cooperatively controlled by a set of two genes, one with a tumor-promoting and the other with a tumor-suppressing effect.

1.3.1 Oncogenes

In 1976, Harold Varmus and Mike Bishop discovered that normal cells contain genes that are related to the transforming genes of RNA tumor viruses (retroviruses)[39]. They showed that some of these retroviruses had captured and modified cellular genes that, when expressed at high level or in mutant form in normal cells, confer a tumor phenotype of rapid, uncontrolled growth[40]. The group of Robert Weinberg could demonstrate that these cellular proto-oncogenes are activated in the DNA of chemically transformed cells. They transfected mouse fibroblasts with genomic DNA of these cells and, through this, were able to transfer the altered phenotype[41]. Therefore, oncogenes are defined as genes with products capable of transforming cells in culture[42-44] or inducing cancer in an organism[45]. Most oncogenes actually trace back to genes which, in normal cells, fulfill key functions in the regulatory networks controlling and promoting proliferation. They may also be involved in the repression of apoptosis. In normal cells, proto-oncogenes can become deregulated, amplified or overexpressed and contribute to malignancy. This activation of a proto-oncogene, which is the conversion to its corresponding oncogene, is usually caused by a gain of function mutation triggered by one of several possible events. Point mutations within the proto-oncogene may generate a constitutively active protein product, which is insensitive to cellular control and promotes proliferation even without prior induction. Alternatively, the amplification of a DNA segment harboring a proto-oncogene may result in an over-expression of the corresponding protein and thereby cause an unwanted stimulation of proliferation. Another mechanism of activation is chromosomal translocation that, *e.g.*, puts the proto-oncogene under control of a strong promoter or enhancer, once more resulting in an inadequate level of gene expression. Independent of its underlying mechanism, a gain of function mutation is always dominant, meaning that a mutation in one allele is sufficient to promote carcinogenesis.

Prominent examples for oncogenes are *MYC*, *RAS* and *ABL*[46]. ABL, for instance, is a non-receptor tyrosine kinase that transduces signals from cell-surface growth factors and adhesion receptors. ABL can shuttle between the nucleus and the cytoplasm of cells and interacts with a large variety of proteins, including signaling adaptors, kinases, phosphatases, cell-cycle regulators, transcription factors and even components of the cytoskeleton, affecting cellular processes including the regulation of cell growth and survival, the response to oxidative stress and DNA damage as well as cell migration[47]. In chronic myelogenous leukemia (CML), the hallmark genetic abnormality is the reciprocal translocation t(9;22)(q34;q11)[48], which results in an abnormal, small chromosome, called the 'Philadelphia chromosome'[49], and generates the *BCR-ABL* fusion gene. BCR is also a signaling protein, which contains multiple modular domains. The fusion of *BCR* sequences to *ABL*[50] during the CML-specific translocation increases the tyrosine-kinase activity of ABL and even appends additional regulatory domains, contributing to the neoplastic transformation of cells in the pathogenesis of CML [49,51].

1.3.2 Tumor Suppressor Genes

Tumor suppressor genes encode proteins which, in contrast to oncogenes, play important roles in the inhibition of cellular proliferation, *e.g.*, as regulators of the cell cycle, receptors for anti-proliferatory hormones etc. They may also be involved in apoptosis induction or cellular adhesion. A single copy of the gene is usually sufficient to maintain its proper function. To promote the formation of a tumor, both alleles have to get lost by mutations or become inactivated, *e.g.*, via the methylation of promoter-associated CpG islands [52], eliminating the generation of functional tumor suppressor (loss of function mutation). This principle was originally formulated in Alfred Knudson's 'two hit' hypothesis [53], which implies that cancer predisposition can result from an inherited mutation in a tumor suppressor gene, but that the development of a tumor requires additional somatic alterations that result in the loss of the wild-type allele. Still, only a small fraction of all cancers , 0.1-10%, depending on the cancer type, are estimated to occur in patients with an inherited mutation [54].

The prototype tumor suppressor gene is *RB*, the retinoblastoma-susceptibility gene discovered by Robert Weinberg and co-workers [55]. Germline mutation in the *RB* gene causes the highly penetrant hereditary retinoblastoma, which results from the bi-allelic loss of *RB* in embryonic retinoblasts [56]. Consistent with its tumor-suppressor function, RB inhibits cell growth and proliferation by blocking cell-cycle progression at the G1/S boundary. This is mediated through the interaction of RB with the E2F family of transcription factors [57] as well as the recruitment of chromatin-modifying enzymes [57-59], resulting in the repression of genes that are required for DNA synthesis [57]. Subsequent to the identification of *RB* as the retinoblastoma gene, *FAP* [60] and *BRCA1* [61]were discovered as the primary factors conferring susceptibility to hereditary colon cancer (familial adenomatous polyposis) or ovarian and breast cancer, respectively.

A very special example for a tumor suppressor gene is *TP53*, which is one of the most commonly mutated genes in human cancer. The protein p53 was first identified in a complex with SV40 T antigen, an oncogenic protein produced by a DNA tumor virus (adenovirus SV40) [62]. This is why *TP53* was initially assumed to act as an oncogene. It was then shown that adenoviral oncoproteins act through the binding and concomitant inactivation of cellular tumor suppressor proteins[63], and that *TP53* maps to a chromosomal region that was consistently deleted or inactivated in colorectal carcinomas[64-66]. This clearly demonstrated the role of p53 as a tumor suppressor. p53 acts primarily as a transcription factor and can mediate different downstream functions by activating or repressing a large number of target genes[67,68], and mutations generally affect amino acids within the DNA-binding domain[69,70]. In addition to its role in transcriptional regulation, p53 is even involved in the (transcriptionally independent) regulation of apoptosis, genome integrity, DNA repair and DNA recombination[71].

1.3.3 'Mutator Genes'

The majority of human cancers show signs of a dramatically enhanced mutation rate. The cells are said to be genetically unstable[72]. In 1976, Lawrence Loeb and colleagues postulated the existence of 'mutator genes', which increase the rate of mutation within tumor cells when they themselves are mutated[73]. This concept proved to be true when the mismatch repair gene *MSH2* was identified as a susceptibility gene for hereditary non-polyposis colon cancer (HNPCC)[74]. Loss of function of MSH2 leads to microsatellite instability, notable as frequent point mutations, insertions and deletions affecting mono- and dinucleotides repeats. More recently, genome-wide hypomethylation was discovered as a reason for chromosomal instability[75]. Generally speaking, 'mutator genes' are involved in the repair of local DNA damage, the correction of replication errors or the maintenance of chromosomal and genome integrity. Analogous to tumor suppressor genes, a single copy of a 'mutator gene' is usually sufficient to maintain its proper function. If both alleles are lost, the affected cells adopt a 'mutator phenotype'[72,76] that drives further mutations in oncogenes and tumor suppressor genes and, thereby, provides a selective growth advantage[77].

1.4 Multistage Carcinogenesis

Tumors can be defined as diseases in which a single cell acquires the ability for abnormal proliferation. Cancers are those tumors that have additionally gained the ability to invade through surrounding normal tissues. If the cancer cells can finally break away from their original location, penetrate into lymphatic and blood vessels, circulate through the bloodstream, and grow in a distant focus, the state of metastasis has been reached.

Despite the existence of many forms of cancer and global changes in gene expression observed in many of them, a relatively small number of essential alterations, affecting few essential pathways, seems to be shared by most, if not all tumors [78]. Tumorigenesis is generally thought to be a multi-step process, in which genetic events that activate oncogenes or inactivate tumor suppressor genes are sequentially acquired, and whereby each genetic change confers a proliferative advantage [79]. Each event is thought to contribute specific malignant features, such as cell-autonomous proliferation, cellular immortalization, induction of angiogenesis, blocked differentiation, genomic destabilization or metastasis, and the accumulation of these genetic events is thought to be responsible for the neoplastic phenotype of a tumor. In human solid tumors, at least four to six mutations are required to reach this state [80]. Liquid tumors, *i.e.*, leukemias and lymphomas, can even evolve from a smaller number of mutations [54]. It is statistically most unlikely, though, that a cell acquires even this relatively small number of mutations merely by random (stochastic) genotoxic events (*e.g.*, chemicals, radiation, viruses, DNA replication errors, oxidative DNA damage, deamination). Therefore, it is believed that a ‘mutator phenotype’ (see 1.3.3) has to be induced at an early stage or even the first stage of carcinogenesis (initiation step), which subsequently induces genetic changes at much higher frequencies [72,77].

Today, human colorectal cancer is the most intensively studied example of this principle [80]. Taking “snap-shots” of the genetic aberrations coinciding with its successive pathological stages of tumor progression, researchers found convincing evidence for what is also referred to as ‘multistage carcinogenesis’, whereby malignant neoplasias are a consequence of multiple genetic defects which successively accumulate over time in a self-accelerating process [81].

1.5 Microarray Technology in Cancer Research

In order to further extend our knowledge on the molecular principles of cancer, we have to identify the genes that are frequently affected by mutations (oncogenes, tumor suppressor genes) as well as those that become deregulated as a consequence. This helps to reveal underlying regulatory pathways and, ultimately, provides an informative basis regarding the search for new drug targets. Ideally, this search should be performed using multiplexed assays on the level of proteins, as proteins are both the endpoint of gene expression and the usual target of drugs. However, it is challenging to define conditions that are equally suitable for a large number of different proteins, both in terms of solubility, stability as well as capture molecule (*e.g.*, antibody) cross-reactivity and availability. Consequently, multiplexed proteome analysis is still in its infancy [82]. Nucleic acid molecules, on the other hand, bind tightly and specifically to complementary sequences, they are much more uniform in terms of biophysical properties and can easily be produced in bulk amounts using standard methods. Transcriptome analysis (gene expression analysis on the level of RNA) is therefore much less of a challenge.

Within the past years, many powerful methods for the detection and quantification of gene expression were developed, including Northern Blotting [83], analysis of S1 endonuclease-digested hybrids [84], Differential Display [85], large scale cDNA sequencing [86,87] and Serial Analysis of Gene Expression (SAGE) [88]. Additionally, there are two array-based approaches, namely, cDNA [89-91] and oligonucleotide microarrays [92-94].

The latter methods provide the valuable potential for multiplexing, which means that many thousand genes can be expression-monitored in parallel within a single experiment. This allows for the generation of both static (which tissues express a certain gene, which genes are expressed in a certain tissue; spatial expression patterns) and dynamic (what is the chronological expression pattern of a certain gene compared to other genes) expression profiles. In this thesis, the nucleic acid molecules to be detected and quantified are referred to as ‘targets’, whereas those molecules that contain target-complementary sequences to allow for the identification of specific targets in a large pool via the formation of detectable hybrids are denoted as ‘probes’.

Like many other biomolecular techniques, microarrays use the principle of hybridization, which is based on the mutual affinity of complementary nucleic acid strands. But whereas most of the established hybridization-based techniques make use of a single or few labeled oligonucleotide or polynucleotide probes in solution, in combination with complex mixture of polynucleotide targets immobilized on a solid support, microarrays employ a different strategy to allow for multiplexed target detection. In Northern Blotting, an RNA pool is separated by electrophoresis and transferred to a nitrocellulose membrane. Selected gene transcripts (targets) are subsequently detected by hybridization of radioactively labeled complementary probes. Parallel analysis of different transcripts within the same experiment is difficult, unless the transcript lengths are both known and differ sufficiently to allow for electrophoretic separation. Even so, only small numbers of targets can be analyzed in parallel.

Microarrays, on the other hand, bypass this limitation by using large numbers of gene- or transcript-specific oligonucleotide or polynucleotide probes attached in close proximity and marginal amounts but at defined positions on a solid support (small arrays of probes, hence microarrays), which is either a membrane, coated glass or silicon. Labeled cDNA or cRNA targets corresponding to complete transcriptomes are hybridized to the arrays, allowing for a localized, fluorescence-based detection of gene-specific signals. The ratios of fluorescence intensities from different transcriptomes (extracted from, *e.g.*, cancer tissue and corresponding healthy tissue), which can either be obtained by competitive hybridization using two distinguishable labels or by comparing the results from two individual array hybridizations, represent the relative expression of the assayed genes relating to the analyzed cells or tissues.

Different methods were established for the production of gene- or transcript-specific array probes. PCR-amplification of cDNA libraries can be used to generate gene-specific amplicons usually ranging from several hundred to a few thousand basepairs in size [89]. As an alternative, several commercial suppliers offer large collections of 50-80-mer oligonucleotides synthesized by phosphoramidite chemistry [94]. In both cases, the DNA probes are dissolved in an appropriate buffer, and high precision robotic devices are employed for the localized deposition of the probe solutions on a chemically modified glass surface (Fig. 4). In a different approach, using *in situ* synthesis of oligonucleotides [95,96] via a combination of custom phosphoramidite chemistry and either photolithography [92,97,98] (Fig. 5) or ink-jet technology [93], subsequent deposition of the DNA probes is circumvented. Light-directed synthesis of oligonucleotides using photolithography is a complex procedure commercialized by the company Affymetrix. Their microarrays are available under the trade name ‘GeneChip’ and contain sets of 11-16 perfect-match and single-base mismatch 25-mer probes, as single 25-mers are too short to specifically detect unique transcripts (Fig. 5C).

With regard to these explicit technical differences, it is necessary to determine whether the results from different array-based expression profiling platforms actually are comparable.

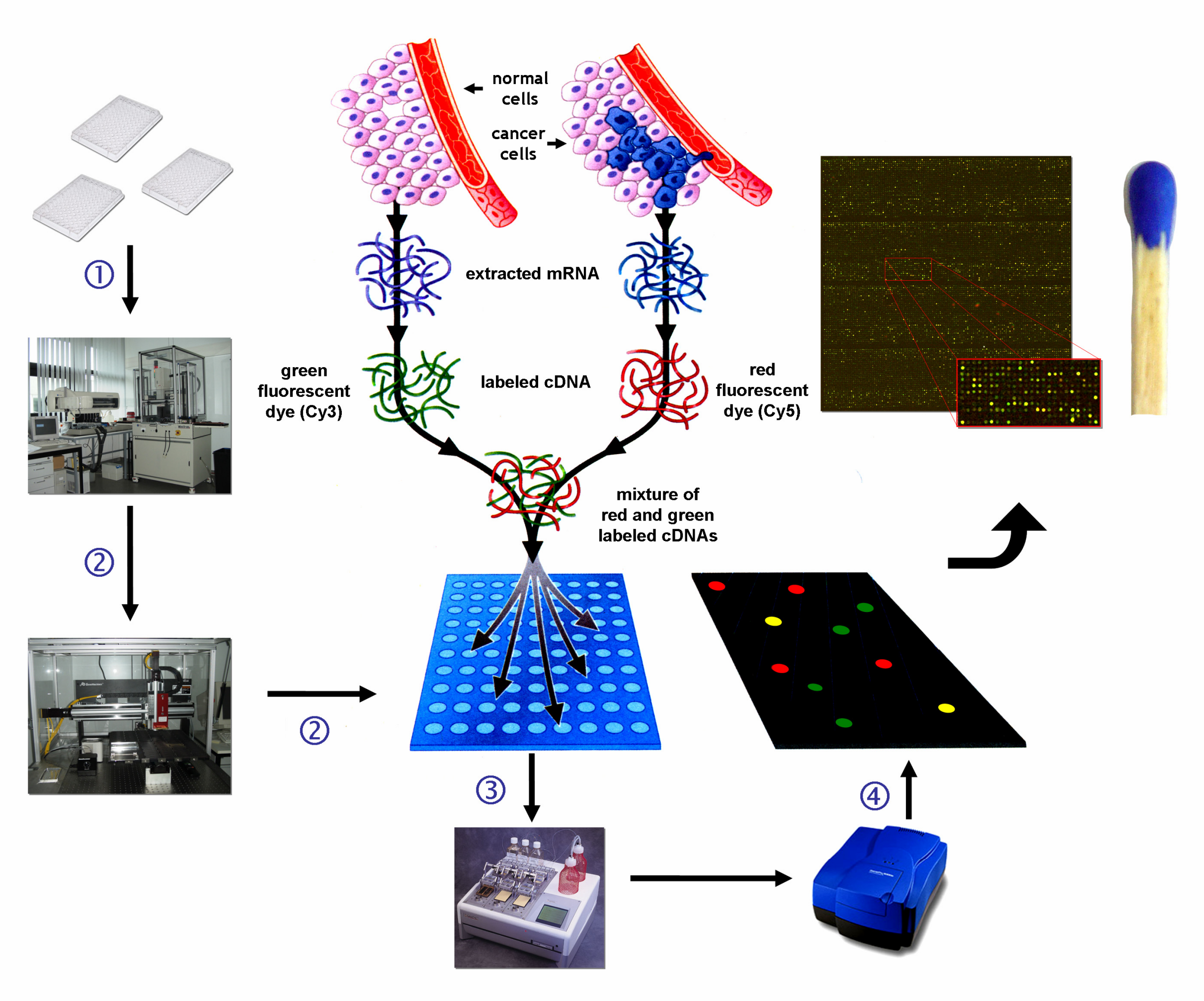

normal cells
cancer cells
extracted mRNA
green fluorescent dye (Cy3)
labeled cDNA
red fluorescent dye (Cy5)
mixture of red and green labeled cDNAs
①
②
②
③
④

Figure 4. Principle of spotted DNA microarrays. The workflow includes (**1**) the preparation and (**2**) robotic deposition of DNA probes onto chemically modified glass slides, (**3**) hybridization of fluorescent DNA targets, usually generated in reverse transcription reactions with Cy3- or Cy5-dUTP, respectively, as well as (**4**) data acquisition and evaluation. Parts of the figure reproduced with kind permission of Aventis Pharma Deutschland GmbH.

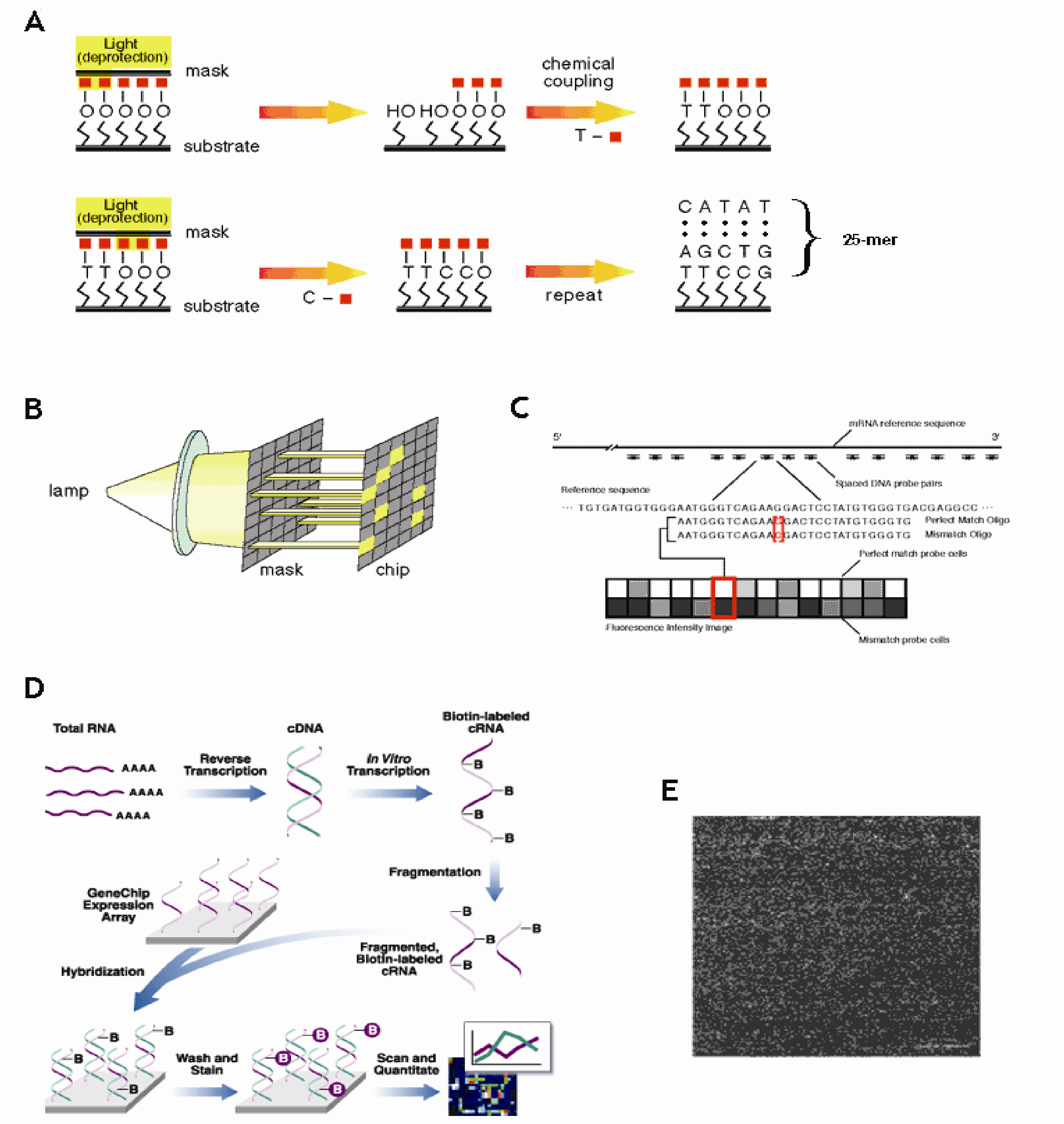


Figure 5. Principle of commercial Affymetrix GeneChip microarrays. (**A**) Light-directed synthesis of oligonucleotides. The surface of a solid support modified with photolabile protecting groups is illuminated through a photolithographic mask, yielding reactive hydroxyl groups in the illuminated regions. A 3'-O-phosphoramidite-activated deoxynucleoside protected at the 5'-hydroxyl with a photolabile group is then presented to the surface and coupling occurs at sites that were exposed to light. Following capping and oxidation, the substrate is rinsed and the surface is illuminated through a second mask, to expose additional hydroxyl groups for coupling. A second 5'-photoprotected, 3'-0-phosphoramidite-activated deoxynucleoside is presented to the surface. The selective photodeprotection and coupling cycles are sequentially repeated until the desired set of products is completed. Most commonly, all synthesis steps are repeated until 25-mer oligonucleotides have been obtained. (**B**) Site-specific photodeprotection occurs via illumination at 365 nm through custom-built sets of photolithographic masks. (**C**) As 25-mer sequences are too short to specifically detect unique transcripts, genes are analyzed by sets of oligonucleotides consisting of 11–16 perfect-match and the same number of single-base mismatch probes, which span representative regions of the selected transcripts. (**D**) For target preparation, total RNA is used in a T7-based linear amplification procedure to produce biotin-labeled cRNA targets. Subsequent to RNA fragmentation, hybridization and washing, target detection occurs via sequential staining with streptavidin-phycoerythrin, biotinylated antistreptavidin and once more streptavidin-phycoerythrin, which allows for fluorescence detection using standard hardware. (**E**) Section of a fluorescence image obtained from a stained GeneChip array. Figure modified from Lipshutz *et al.* [98] and from http://www.affymetrix.com.

1.6 RNA Amplification

Expression profiling by DNA microarrays is based on hybridization methods developed by Ed Southern about 30 years ago [99,100]. Early microarray studies used large amounts of cells (> 10^6), which were obtained either from large tumors or cell lines [101].

Although cancers emerge from single malignant cells, the analysis of tumor sections detected a strong heterogeneity of the cell populations [102]. Metastasizing cells often acquire numerous genotypic and phenotypic alterations, which may also influence their response to anti-tumor therapy [103]. The expression analysis of large and hence potentially heterogeneous tumors may therefore yield inconclusive results.

The large amount of RNA material needed for array-based expression profiling thereby represents a serious obstacle for cancer-related research intending to benefit from the potential of this promising new technology. Using fine-needle aspiration, a method for the collection of small tumor samples (< 100,000 cells) for histological characterization, the amount of RNA which can be obtained (about 2 µg) is hardly sufficient for most of the standard protocols, and the results may still be impaired by a high proportion (20 - 60%) of non-tumor cells [104,105]. To generate a representative expression profile of a cancer, one should ideally use microdissection approaches to obtain single or at least very small numbers of individually selected tumor cells [106,107], but this will yield even smaller amounts of RNA material.

Proper amplification procedures are therefore indispensable if limited source material is to be analyzed by microarray technology (Fig. 6). In principle, this amplification can either be performed on the limited source material itself (target amplification) or on the signal generated by this material (signal amplification). Target amplification can be conducted exponentially using PCR-based approaches [108-110] or linearly via the generation of cDNA and subsequent *in vitro* transcription using T7 RNA polymerase [111,112]. The kinetics of PCR [113,114] implies that sequence dependent and concentration dependent bias accumulates in an exponential fashion as well, and at very low amounts of starting material sampling errors become increasingly problematic [115,116]. Due to these reasons, amplification by PCR is generally considered less appropriate for the parallel, microarray-based analysis of multiple targets.

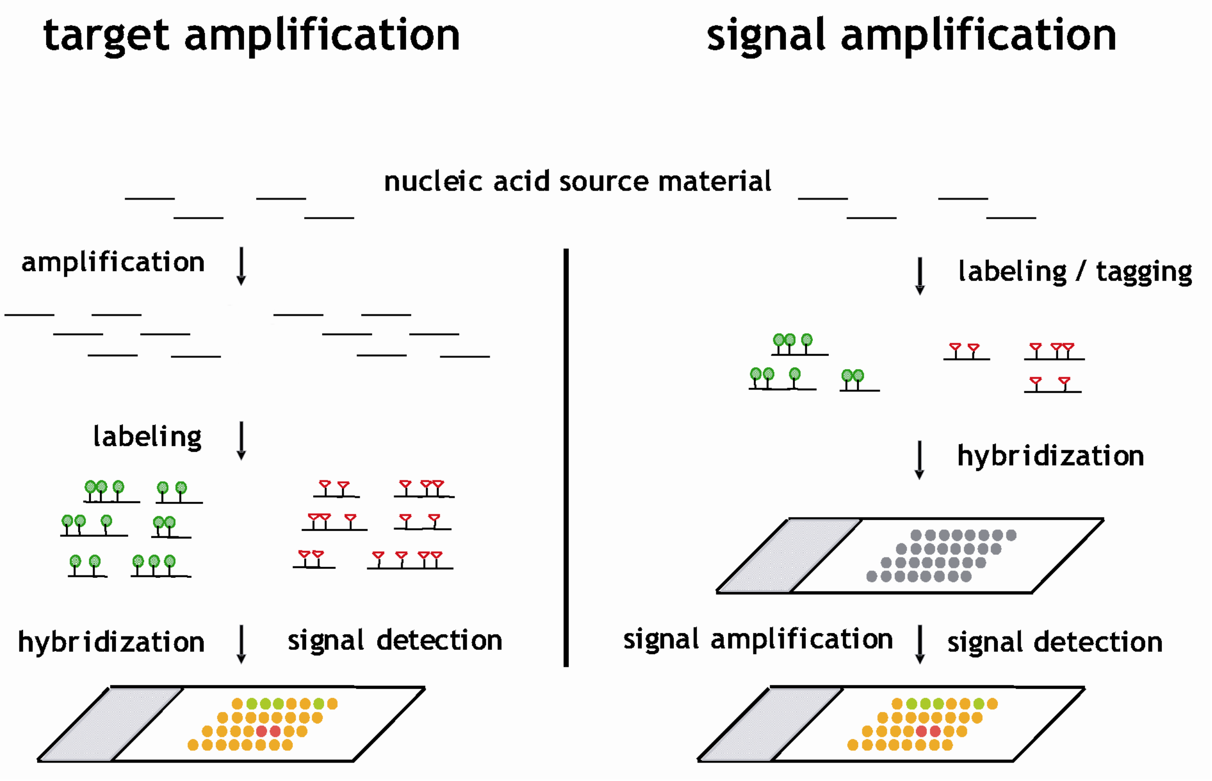


Figure 6. Conceptual differences of target amplification and signal amplification procedures. Methods for target amplification duplicate the source material prior to the hybridization step. The amplified target material is labeled with, *e.g.*, fluorescent dyes and subsequently hybridized to the array, where it can be detected. Signal amplification procedures, on the other hand, employ the limited source material for hybridization and amplify the signal generated on the array. Although different approaches exist for the latter step, they all require some sort of tagging (*e.g.*, via haptens or capture sequences) of the source material that allows for a localized amplification. Modified from a figure kindly provided by Prof. Peter Lichter.

Most of the current target amplification procedures therefore perform a linear amplification by *in vitro* transcription, as initially proposed by van Gelder *et al.* for the cloning and expression analysis of low-abundance transcripts from small populations of neuronal cells [112]. For this purpose, total RNA is reversely transcribed to cDNA, using a modified oligo(dT) primer that contains the promoter sequence of phage T7 RNA polymerase, and an RNase H$^-$ MMLV RNA dependent DNA polymerase (*e.g.*, SuperScript II). This reaction yields an RNA/DNA hybrid. In a subsequent step, partial digestion of the RNA portion of the heteroduplex with RNase H generates small RNA primers that initiate second strand synthesis by *E. coli* DNA polymerase I. Fragments that originated from different RNA primers are joined by enzymatic ligation, similar to the joining of Okazaki fragments in the discontinuous lagging strand synthesis of eukaryotic DNA replication [117-119]. This yields uninterrupted double-stranded cDNA containing the T7 promoter sequence downstream of the transcript sequences. Subsequent *in vitro* transcription with T7 RNA polymerase yields multiple antisense-oriented copies of the initial transcripts, as transcription from each promoter is initiated repeatedly (Fig. 7). Commercial T7-based amplification kits are available ,*e.g.*, from Arcturus (Mountain View, USA; RiboAmp) and Ambion (Austin, USA; MessageAmp) and have been reported to perform well [120,121].

Unfortunately, the original T7-based protocols cannot be used with commercially available oligonucleotide probe libraries when combined with conventional reverse transcription labeling methods. Their sequences are sense-oriented to be compatible with cDNA targets obtained by reverse transcription. T7 amplification yields RNA with antisense orientation, and RT labeling would transform these sequences into sense-oriented cDNA targets, incompatible with sense-oriented oligonucleotide probes.

A T7-based study by t' Hoen et al. [122] used aminoallyl-UTP nucleotides in the *in vitro* transcription reaction, which allowed them to label the aRNA in a subsequent coupling reaction with Cy-NHS-esters. Labeled aRNA would be suitable for hybridization to sense-oriented oligonucleotide arrays (Fig. 8).

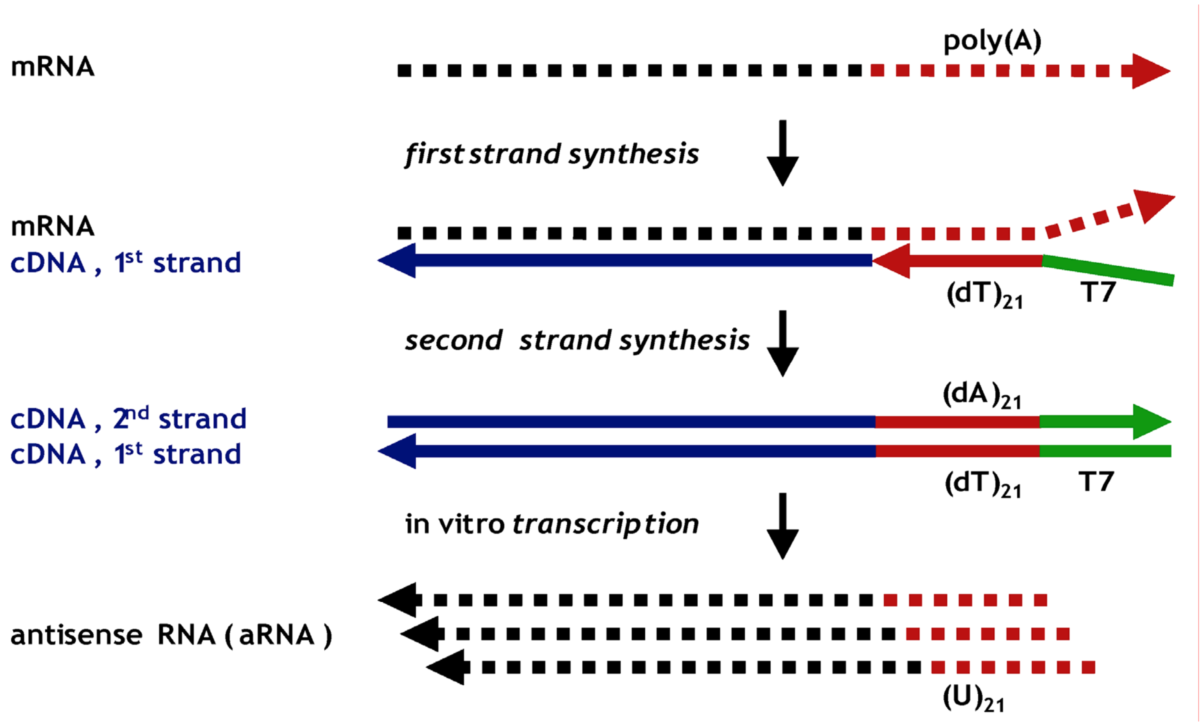


Figure 7. T7 RNA amplification. An RNA polymerase promoter is incorporated into each cDNA molecule by priming cDNA synthesis with a synthetic oligonucleotide containing the phage T7 RNA polymerase promoter. After synthesis of double-stranded cDNA, T7 RNA polymerase is added and antisense RNA is transcribed from the cDNA template. The processive synthesis of multiple RNA molecules from a single cDNA template results in amplified antisense RNA (aRNA).

A recent publication by Smith *et al.*[123] claimed that their amplification procedure, termed "Single Primer Amplification" (SPA), could also be used for microarrays containing oligonucleotide probes in sense orientation. The first steps in this protocol generate double-stranded cDNA, initially primed by a modified oligo(dT) primer. A primer equivalent to the heel of the modified oligo(dT) primer is then used to direct semi-linear *Taq* DNA polymerase amplification of the first strand cDNA. Hence, the protocol is essentially a modified cycle sequencing reaction so far. Then, Klenow fragment is used to perform a randomly primed labeling reaction, initially producing fluorescent cDNA in sense orientation. However, it can be expected that at least a portion of this primary product will be used to template another round of polymerization ("strand switch") that generates antisense-oriented cDNA. This would allow hybridization to both sense and antisense strands and consequently render the method compatible with oligonucleotide microarrays containing probes in sense orientation (Fig. 9).

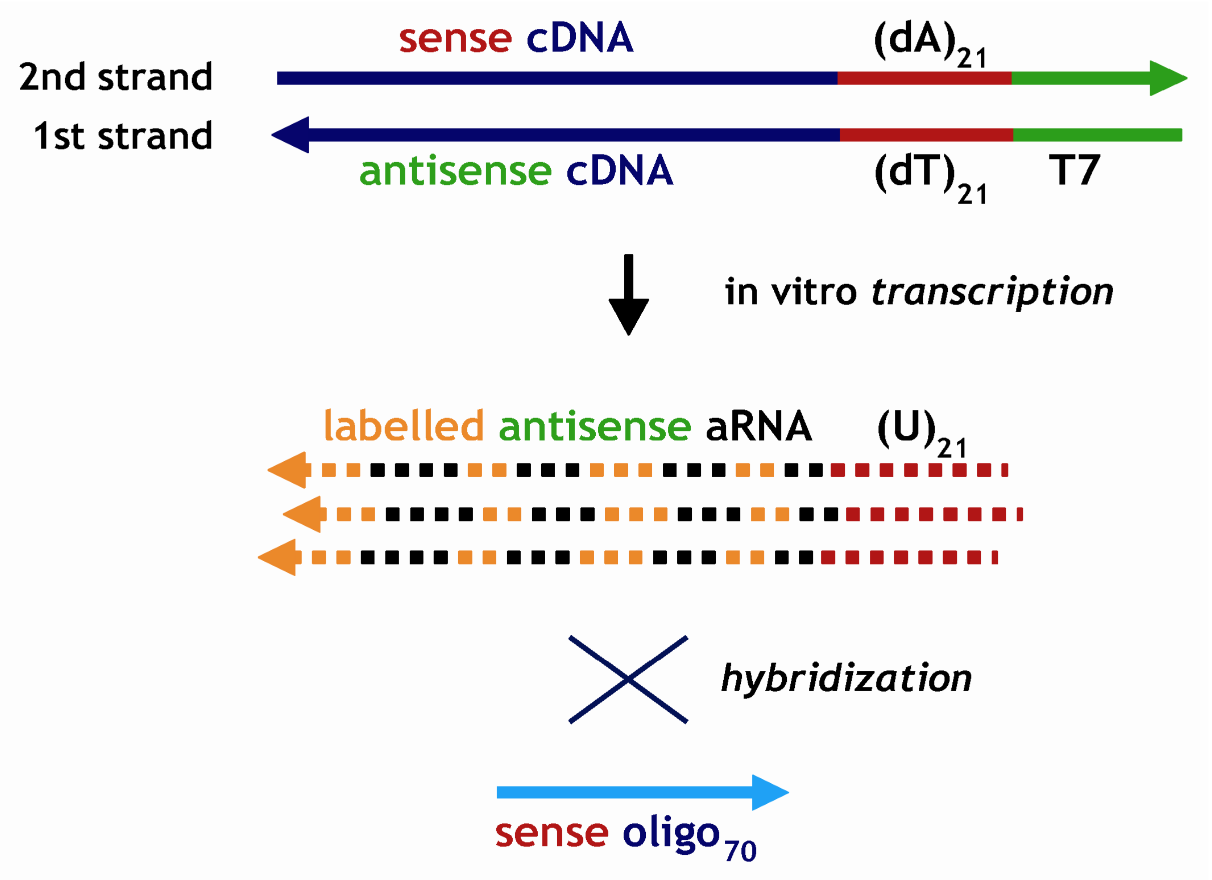


Figure 8. Preparation of fluorescent cRNA targets by *in vitro* transcription. Following the generation of double stranded cDNA containing the T7 promoter downstream of the coding sequence, repeatedly initiated *in vitro* transcription yields multiple copies of fluorescent aRNA. These molecules are antisense-oriented and can therefore be hybridized to sense-oriented oligonucleotides.

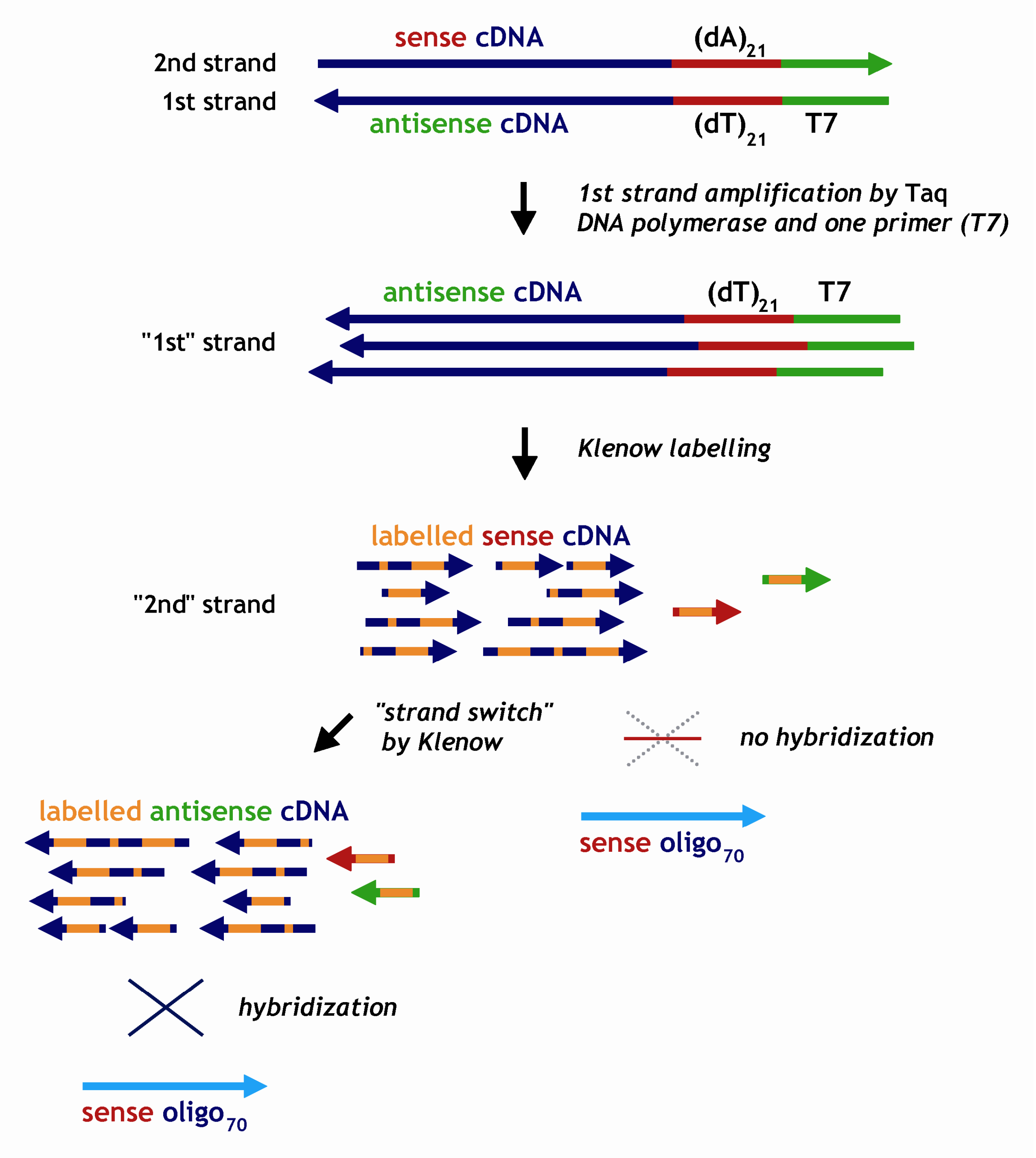


Figure 9. Single Primer Amplification (SPA). In theory, the SPA method can yield both sense- and antisense-oriented fluorescent cDNA targets. The generation of antisense cDNA depends on a "strand switch" of the polymerase, in which the primary, sense-oriented product serves as a template for another, randomly primed round of polymerization.

1.7 Signal Amplification

Target amplification is just one approach to permit DNA microarray experiments with limited amounts of source material. A further possible procedure is to amplify the signal generated on the array subsequent to the hybridization step (Fig. 9). Several methods have been proposed to achieve this amplification.

Tyramide signal amplification (TSA) [124,125] uses fluorescein (as hapten for subsequent labeling with the fluorescent dye Cy3) or biotin (as hapten for subsequent labeling with the fluorescent dye Cy5) modified nucleotides, which are incorporated during first-strand cDNA synthesis. After array hybridization and stringent washing, two antibody-enzyme conjugates recognize the fluorescein or biotin haptens. The enzyme portion of these conjugates is horseradish peroxidase (HRP), which, in subsequent step, catalyzes the local deposition of Cy3/5-labeled tyramide immediately adjacent to the immobilized HRP. In this enzymatic process, HRP-activated highly reactive and short-lived tyramide radicals undergo covalent coupling to nucleophilic residues in the vicinity of the HRP-target interaction site, whereby the amount of tyramide relative to cDNA hapten label is greatly amplified. However, previously published evaluations concerning the TSA method report inconsistent labeling [126], increased background fluorescence and insufficient reproducibility, *i.e.*, poor signal correlations from co-hybridizations of identical cDNA targets [127].

The 3DNA system offered by Genisphere uses fluorescently labeled DNA dendrimers [128]. DNA dendrimers are complex, branched molecules built from interconnected monomeric subunits [129]. The 3DNA dendrimers contain an average of about 850 fluorescent labels and recognize a capture sequence introduced by a modified RT primer during first-strand cDNA synthesis. The results from present studies on the 3DNA system for microarray analysis are inconclusive. Yu et al. [126] report strong signal correlations in self-self hybridization experiments and reasonable consistency of 3DNA expression profiles with those from direct labeling experiments, whereas Richter et al. [127] obtained high signal intensities but no meaningful expression patterns.

Rolling-circle DNA replication is the basic principle for a further signal amplification procedure, termed Rolling Circle Amplification (RCA). In RCA, a circular, single-stranded DNA molecule is duplicated repeatedly by a DNA polymerase, creating long concatameric copies of the circular template (for more details, see 1.7.2). For this purpose, the polymerase has to offer an extremely strong strand displacement activity as well as superior processivity. The monomeric 66.52 kDa DNA polymerase of the lytic bacteriophage Φ29 possesses both of these properties and is therefore of fundamental importance for any RCA protocol. Since RCA for microarray expression analysis is one of the subjects of this thesis, I will provide more detailed information on the Φ29 phage, its DNA polymerase and the rolling-cirlce mechanism of DNA replication.

1.7.1 Phage Φ29

The Φ29-like genus of phages includes, in addition to Φ29, phages PZA, Φ15, BS32, B103, M2Y, Nf, and GA-1 [130]. They are all lytic phages belonging to the *Podoviridae* family. They infect bacteria of the genus *Bacillus*, which incorporates many species of gram-positive, aerobic, endospore-forming bacteria that normally inhabit the soil of decaying plant material. Φ29-like phages are commonly found in *Bacillus subtilis*, but often they may also infect related species such as *Bacillus pumilus*, *Bacillus amyloliquefaciens*, and *Bacillus licheniformis*.

Phage Φ29 has been subject to intensive studies, and the results have contributed to the understanding of several molecular mechanisms of biological processes, such as DNA replication, regulation of transcription, phage morphogenesis, and phage DNA packaging. Electron microscopy analysis revealed that Φ29 comprises a head of 41.5 by 31.5 nm with sixfold radial symmetry and a short noncontractile tail of 32.5 by 6.0 nm [131]. Being the smallest *Bacillus* phages isolated so far, the Φ29-like phages are also among the smallest phages containing dsDNA. The genomes of the Φ29-like phages consist of a linear dsDNA molecule of about 20 kb (19,285 kb for Φ29 [132]), which has a phage-encoded terminal protein (TP) covalently attached at each 5'-end. Genomes consisting of a TP covalently linked to their 5'-ends have also been found for animal viruses (*e.g.*, adenoviruses) and bacteria (*e.g.*, *Streptomyces*), and in all of these cases, initiation of DNA replication occurs via a so-called protein-priming mechanism [133,134]. DNA polymerases are unable to initiate *de novo* DNA synthesis on a DNA template but require the existence of a primer containing a free hydroxyl group. Generally, RNA primers provide the 3'-hydroxyl group needed by the DNA polymerase to elongate the DNA chain. However, in most linear genomes containing a 5’-TP, the 3’-OH of a specific serine, threonine or tyrosine residue of the TP is used for DNA elongation.

Bacteriophage Φ29 DNA polymerase, the product of the viral gene 2, was originally characterized as a protein involved in the initiation of Φ29 DNA replication based on both *in vivo*[135] and *in vitro*[133,136,137] studies. The cloning[138]of gene 2, the overproduction and purification of its product[139], and the development of an *in vitro* system for complete Φ29 DNA replication[134] allowed the characterization of protein p2 as the viral DNA replicase[140]. It belongs to the B-type superfamily of DNA-dependent DNA polymerases, which are also referred to as eukaryotic or α-like polymerases and contain many prokaryotic and eukaryotic enzymes that are sensitive to certain drugs (aphidicolin, phosphonoacetic acid) and nucleotide analogs (butylanilino-dATP, butylphenyl-dGTP). The monomeric Φ29 DNA polymerase has a size of 66.52 kDa and catalyzes both the initiation and elongation stages of Φ29 DNA replication. To accomplish this, it is able to carry out two distinct synthetic reactions: TP deoxynucleotidylation, which consists of the formation of a covalent phosphoester bond between the hydroxyl group of a specific serine residue (Ser^{232}) in Φ29 TP and dAMP, and DNA polymerization[139]. In addition to the synthetic activities, Φ29 DNA polymerase has two degradative activities: pyrophosphorolysis and ssDNA 3’-5’-exonuclease (proofreading), which processively degrades DNA substrates longer than six nucleotides. Moreover, it has the intrinsic properties of high processivity (> 70 kb) and strand displacement activity, even during the polymerization process[141]. Due to these characteristics, Φ29 DNA polymerase is the only enzyme required for efficient *in vitro* replication of the Φ29 genome, with the Φ29 TP, the initiation primer, as the only additional protein requirement[134]. DNA synthesis starts non-simultaneously from either end of the DNA molecule with the covalent linkage of dAMP (from dATP) to a free molecule of the TP. The subsequent elongation of the DNA chain occurs by a strand-displacement mechanism, which allows for efficient *in vitro* synthesis of full length Φ29 DNA.

1.7.2 Rolling Circle Replication

The rolling-circle mechanism of DNA replication (RCR) is used by small prokaryotic genomes, such as single-stranded phages and plasmids. Filamentous single-stranded phages are intermediates between lytic phages (*e.g.*, fX174) and plasmids (*e.g.*, pT181): they do not cause cell lysis but exist intracellularly as stable plasmids and generate infective particles more or less indefinitely. Several groups have demonstrated that even *in vitro*-generated circular ssDNA can support a rolling circle replication [142] to produce concatemeric repeats of monomers as short as 34 nucleotides, using either *E. coli* DNA polymerase I, *E. coli* DNA polymerase large fragment (Klenow fragment) or modified T7 DNA polymerase (Sequenase) [142], and the method has been applied for amplified detection (rolling circle amplification; RCA) of viral RNA from tissue samples [143] and for preparative *in vitro* synthesis of catalytic antisense RNA [144]. It was then realized that Φ29 DNA polymerase, due to its superior processivity and strand displacement activity, is a better choice for this purpose [145,146]. For the same reasons, Φ29 polymerase is also used for the generation of sequencing templates [147,148], for whole genome DNA amplification prior to genotyping [149,150] or array-CGH [151-154] and even for the amplification of CGH array clones [155,156].

In RCA, a circle of DNA, a short DNA primer (complementary to a portion of the circle) and an enzyme catalyst convert dNTPs into a single-stranded concatemeric DNA molecule that is composed of thousands of tandemly repeated copies of the circle. Unlike other amplification procedures, RCA produces a single amplified product that remains linked to the DNA primer. Consequently, RCA is well suited to solid phase formats such as microarrays for generating localized signals at specific microarray locations. PCR or other solution phase amplification procedures, on the other hand, cannot be configured for on-chip amplification due to the lack of accumulation of amplified signal at the site of amplification, *i.e.*, diffusion of products into the solution, and/or deleterious effects of temperature cycling on reaction components, such as analytes, samples or microarray substrates. Applications of RCA signal amplification for the detection of DNA targets immobilized on solid surfaces [157] as well as for ultrasensitive detection of proteins on microarrays [158,159] have been previously demonstrated. RCA signal amplification on microarrays involves a universal amplification circle, regardless of the nature and number of targets being assayed. This approach minimizes bias during amplification. Therefore, and in contrast to target amplification methods, such as PCR or T7-based *in vitro* transcription, universal RCA signal amplification can be expected to introduce significantly less sequence-dependent bias. In addition, the RCA procedure is less costly and time-consuming. The circular oligonucleotides needed for the amplification can be generated from padlock probes, which are linear oligonucleotides with 5' and

3' end regions designed to base pair next to each other on a target strand. If properly hybridized, the ends can be joined by enzymatic ligation, converting the probes to circularly closed molecules that are catenated to their target sequence [160]. So far, RCA on microarrays has not been applied for the amplified detection of expression profiles. In this thesis, I evaluate the applicability of a protocol I conceived to adapt on-chip RCA for expression profiling approaches (Fig. 10).

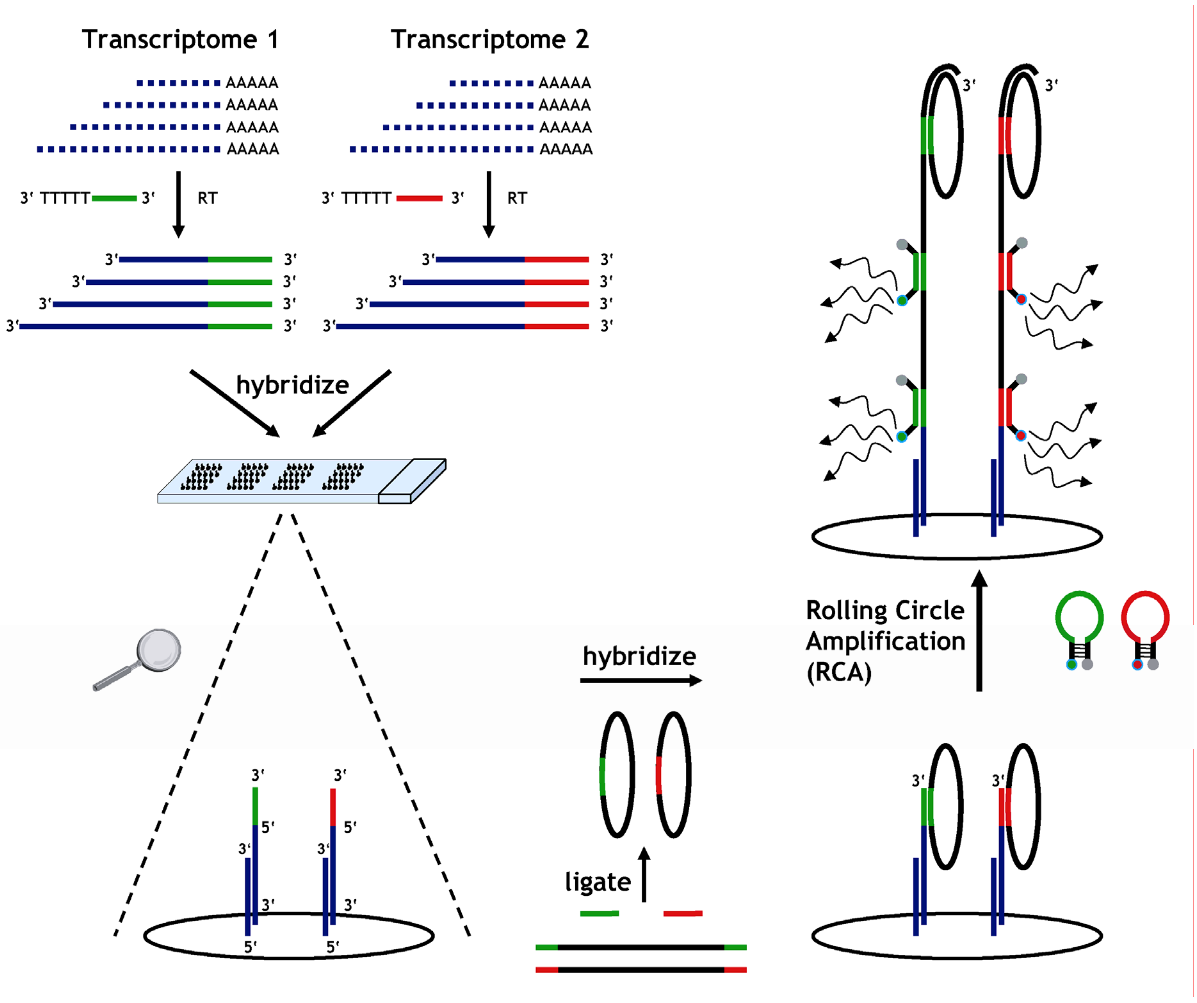


Figure 10. Signal amplification by rolling circle amplification (RCA) on DNA microarrays. Total RNA is reversely transcribed to cDNA, using a special oligo(dT) primer that additionally contains a sequence complementary to a portion of the circular DNA to be amplified. This part of the primer is synthesized in reverse (5' to 3'), providing an additional 3'-end to prime the downstream RCA reaction. Subsequenty, the cDNA is hybridized to standard spotted DNA microarrays, followed by hybridization of the circular oligonucleotides generated in a separate enzymatic ligation reaction. The circle-complementary part of the cDNA contains a free 3'-end to which the circles hybridize, serving as templates for the RCA reaction. The linear concatameric RCA product, comprising repeating units of a sequence complementary to the circle, can be detected by hybridization of molecular beacons or conventional short oligonucleotides containing fluorescent labels.

1.8 Objective

All truths are easy to understand once they are discovered; the point is to discover them.

GALILEO

Initially, there were just two alternatives for microarray-based analysis of gene expression, which were self-spotted cDNA arrays and commercial 25-mer arrays from Affymetrix. More recently, spotted microarrays of long sense-oriented oligonucleotides were introduced as an attractive alternative to the two former methods. Their probes are designed to have similar biophysical properties and avoid secondary structures as well as repetitive sequences, which is a considerable advantage compared to cDNA arrays. Additionally, they are long enough (50 - 80 nt) to allow for a specific detection of target molecules with just one probe, compared to 11 - 16 probes for the Affymetrix platform.

To use spotted oligonucleotide arrays for the analysis of minimal amounts of RNA material, either a target or a signal amplification method must be applied in order to yield adequate results. T7-based target amplification protocols had been reported to perform well with cDNA microarrays. However, reverse transcription labeling transforms the antisense-oriented amplification products to fluorescent cDNA targets with sense orientation, which are obviously inapplicable for hybridization to sense-oriented oligonucleotides. Available methods for signal amplification, on the other hand, which used, *e.g.*, three-dimensional multi-labeled structures (DNA dendrimers), or the enzymatically activated deposition of dye molecules via tyramide (TSA), had been evaluated with rather heterogeneous results. Signal amplification based on rolling circle mechanism of DNA replication (RCA), had very successfully been used for the detection of antigens on protein microarrays, but the method had not yet been implemented for microarray expression analysis. The objective of this thesis was therefore to develop and optimize a target amplification protocol specific to the problems and characteristics of spotted oligonucleotide microarrays. The desired protocol had to generate fluorescent target molecules with antisense orientation, compatible for hybridization to sense-oriented oligonucleotide probes.

Additionally, an RCA-based signal amplification protocol should be developed, as an alternative to the target amplification procedure. Pre-ligated circular oligonucleotides should be used to detect a common sequence introduced to the cDNA targets and serve as templates for a subsequent rolling circle replication reaction. This would yield long concatameric amplification products continuous with the hybridized cDNAs, which could finally be detected by complementary molecular beacons or short labeled oligonucleotides.

Both protocols should finally be used to evaluate the performance of the novel oligonucleotide arrays in comparison to the commercial, well-established Affymetrix GeneChip system. Both platforms should be applied to generate expression profiles from a set of head and neck squamous cell carcinoma patients. These data should be used as a basis to estimate the degree of concordance between the two systems and, thereby, answer the question whether or not spotted oligonucleotide arrays can deliver on their promise.

2 Materials and Methods

Give me a lever long enough and a fulcrum on which to place it, and I shall move the world.

ARCHIMEDES

2.1 Materials

2.1.1 Chemicals and Biochemicals

Table 1. Chemicals and Biochemicals.

Chemical	Supplier (Cat #)
2-Mercaptoethanol (thioethylene glycol)	Sigma-Aldrich, Munich (M3148)
2-Propanol (isopropanol)	Merck, Darmstadt (109634)
Aminoallyl-dUTP	Fermentas, St. Leon-Rot (R0091)
Ammonium acetate solution, 7.5 M	Sigma-Aldrich, Munich (A2706-100ML)
Betaine (trimethylglycine), inner salt, unhydrous	Sigma-Aldrich, Munich (B2629)
Bovine serum albumin (BSA)	New England Biolabs, Frankfurt am Main (B9001S)
C_0t human DNA	Roche Diagnostics, Mannheim (11 581 074 001)
Chloroform, *pro analysis*	Merck, Darmstadt (102445)
Cyanin 3-dUTP	Amersham Biosciences, Freiburg (PA53022)
Cyanin 5-dUTP	Amersham Biosciences, Freiburg (PA55022)
Deoxynucleotide (dNTP) set, 100 mM solutions	Amersham Biosciences, Freiburg (27-2035-02)
DEPC-treated water	Ambion, Austin, USA (9922)
EDTA, 0.5 M, pH 8.0	Ambion, Austin, USA (9260G)
Formamide, *pro analysis*	Merck, Darmstadt (109684)
Hydrochloric acid, 1 M	JT Baker, Pillipsburg, USA (7088)
Linear polyacrylamide (LPA), 5 µg/µl	Ambion, Austin, USA (9520)
NHS-psoralen (SPB)	Pierce Biotechnology, Rockford, USA (23013)
Nitrocellulose (cellulose nitrate)	Sigma-Aldrich, Munich (N7892)
Nuclease-free water (not DEPC-treated)	Ambion, Austin, USA (9937)
Nucleotide (NTP) set, 100 mM solutions	Amersham Biosciences, Freiburg (27-2025-01)
PCR nucleotide mix, 10 mM each dNTP	Amersham Biosciences, Freiburg (US77212-500µl)
Phenol-chloroform-isoamyl alcohol 25:24:1, pH 8.0	Sigma-Aldrich, Munich (P2069-100ML)

Polyadenylic acid (Poly(A)), potassium salt	Sigma-Aldrich, Munich (P9403-500MG)
Primer random (random hexamers)	Roche Diagnostics, Mannheim (1034731)
RNA 6000 ladder	Ambion, Austin, USA (7152)
RNasin ribonuclease inhibitor	Promega, Mannheim (N2111)
SDS solution, 10%	Ambion, Austin, USA (9823)
Second-strand buffer for SSII cDNA synthesis, 5x	Invitrogen, Karlsruhe (10812-014)
Sodium chloride (NaCl) solution, 5 M	Ambion, Austin, USA (9760G)
Sodium hydroxide (NaOH) solution, 10 M	Sigma-Aldrich, Munich (72068-100ML)
SSC, 20X	Ambion, Austin, USA (9763)
T4 gene 32 protein, high concentration	USB, Cleveland, USA (74029Y)
TE, pH 8.0	Ambion, Austin, USA (9858)
Total RNA, human breast, female, Lot #0330574	Stratagene, La Jolla, USA (735044)
Trizol reagent	Invitrogen, Karlsruhe (15596-018)
ULTRAhyb hybridization buffer	Ambion, Austin, USA (8670)
Universal human reference RNA, Lot # 0810006	Stratagene, La Jolla, USA (740000)
Yeast tRNA (lyophilized)	Invitrogen, Karlsruhe (15401-011)

2.1.2 Enzymes

Table 2. Enzymes.

Enzyme	Supplier (Cat #)
Advantage cDNA polymerase mix	BD Biosciences Clontech, Heidelberg (8417-1)
AmpliTaq DNA polymerase	Applied Biosystems, Weiterstadt (N808-0155)
DNA ligase (*E.coli*)	Amersham Biosciences, Freiburg (E70020Z)
DNA polymerase I (*E. coli*)	Promega, Karlsruhe (M2055)
Phi29 DNA polymerase	Fermentas, St. Leon-Rot (EP0092)
PowerScript reverse transcriptase	BD Biosciences Clontech, Heidelberg (8460-1)
Ribonuclease H (*E. coli*)	Epicentre, Madison, USA (R0601K)
SuperScript II reverse transcriptase	Invitrogen, Karlsruhe (18064-071)
T4 DNA ligase	Amersham Biosciences, Freiburg (E70042X)
T4 DNA polymerase	New England Biolabs, Frankfurt am Main (M0203L)
Terminal deoxynucleotidyl transferase	Amersham Biosciences, Freiburg (E2230Y)

2.1.3 Kits

Table 3. Kits.

Kit	Supplier (Cat #)
Buffer kit	Ambion, Austin, USA (9010)
Cooled RNA 6000 nano reagents	Agilent Technologies, Waldbronn (5065-4475)
MEGAscript T7 kit	Ambion, Austin, USA (1334)
MessageAmp aRNA kit	Ambion, Austin, USA (1750)
RiboMAX large scale RNA production system-T7	Promega, Karlsruhe (P1300)
RNA 6000 nano LabChip kit	Agilent Technologies, Waldbronn (5065-4476)
RNeasy mini kit	Qiagen, Hilden (74104)
RNeasy midi kit	Qiagen, Hilden (75144)

2.1.4 Other Materials

Table 4. Other Materials.

Material	Supplier
ART 1000E nuclease free 1000 µl tips	Molecular BioProducts, San Diego, USA (2079E)
ART 100E nuclease free 100 µl tips	Molecular BioProducts, San Diego, USA (2065E)
ART 200 nuclease free 200 µl tips	Molecular BioProducts, San Diego, USA (2169)
ART 20P nuclease free 20 µl tips	Molecular BioProducts, San Diego, USA (2149P)
ART REACH nuclease free 10 µl tips	Molecular BioProducts, San Diego, USA (2140)
AutoSeq G-50 columns	Amersham Biosciences, Freiburg (27-5340-01)
Dynabeads M-280 streptavidin	Dynal Biotech, Hamburg (112.06)
Micro Bio-Spin 6 columns in Tris buffer	Bio-Rad Laboratories, Munich (732-6222)
Microcentrifuge tubes, PCR clean, 0.5 ml	Eppendorf, Hamburg (0030 123.301)
Microcentrifuge tubes, PCR clean, 1.5 ml	Eppendorf, Hamburg (0030 123.328)
Microcentrifuge tubes, PCR clean, 2.0 ml	Eppendorf, Hamburg (0030 123.344)
Microcon YM-100 centrifugal filter devices	Millipore, Schwalbach (42412)
Microcon YM-30 centrifugal filter devices	Millipore, Schwalbach (42411)
mSeries LifterSlips 22 x 60 mm	Erie Scientific, Portsmouth, USA (22x60I-M-5522)
Nexterion slide E	Schott Jenaer Glas, Jena (1066643)
Nexterion slide E with barcode	Schott Jenaer Glas, Jena (1064016)
PCR Tubes, 0.2 ml	Molecular BioProducts, San Diego, USA (3412)

Phase lock gel heavy, 0.5 ml	Eppendorf, Hamburg (0032 005.055)
RNase away	Molecular BioProducts, San Diego, USA (7003)
RNaseZap	Ambion, Austin, USA (9780)
Stealth SMP3 microarray spotting pins	TeleChem / ArrayIt, Sunnyvale, USA (SMP3)
Stealth SPH48 printhead matrix	TeleChem / ArrayIt, Sunnyvale, USA (SPH48)

2.1.5 Instruments

Table 5. Instruments.

Instrument	Manufacturer
Agilent 2100 Bioanalyzer	Agilent Technologies, Waldbronn
Biofuge Fresco refrigerated tabletop centrifuge	Heraeus / Kendro, Hanau
Cary 50 Bio UV-photometer	Varian, Darmstadt
DNA Engine Dyad PCR-cycler	MJ Research
GenePix 4000B array scanner	Molecular Devices / Axon Instruments, Union City, USA
GeneTAC hybridization station	Genomic Solutions (GeneMachines), Ann Arbor, USA
Heating block QBT2	Grant Instruments
Heating cabinet series 6000	Heraeus / Kendro, Hanau
HMT 702 C microwave oven	Bosch, Stuttgart
Microcentrifuge	neoLab Laborbedarf, Heidelberg
MiniTrak liquid handling system	PerkinElmer, Rodgau-Jügesheim
MultiPROBE IIex liquid handling system	PerkinElmer, Rodgau-Jügesheim
OmniGrid microarrayer	Genomic Solutions (GeneMachines), Ann Arbor, USA
Stratalinker 2400 UV-crosslinker	Stratagene, La Jolla, USA
Ultra Turrax T25 tissue homogenizer	Janke & Kunkel, Staufen
Varifuge 3.0/3.0R floor model centrifuge	Heraeus / Kendro, Hanau
VersArray ChipWriter Pro system	Bio-Rad Laboratories, Munich
Water bath SW22	Julabo Labortechnik, Seelbach

2.1.6 Software

Table 6. Software.

Software	Manufacturer
Adobe Creative Suite CS	Adobe Systems, Unterschleissheim
ChemDraw Ultra 9.0	CambridgeSoft, Cambridge, USA
EditPlus 2.11 text editor	ES-Computing, Chinju, South Korea
EndNote 8.0	Thomson ResearchSoft, Carlsbad, USA
Expression Analysis Systematic Explorer 2.0	National Institutes of Health, Rockville, USA
FlashFXP 3.0 FTP client	IniCom Networks, Socorro, USA
GenePix Pro 4.0, 5.0 & 5.1	Molecular Devices / Axon Instruments, Union City, USA
Microsoft Office professional edition 2003	Microsoft, Unterschleissheim
Microsoft Windows 2000 & service pack 4	Microsoft, Unterschleissheim
R 1.9.1 & 2.0.0	R Foundation for Statistical Computing, Vienna, Austria

2.1.7 Standard Solutions

Table 7. Standard Solutions.

Solution	Composition
5 x TBE, pH 8.0	0.445 M Tris-Borat, pH 8.0 10 mM EDTA, pH 8.0
1 x TE, pH 8.0	10 mM Tris-HCl, pH 8.0 1 mM EDTA, pH 8.0
SDS, 20%	20% SDS (w/v)
20 x SSC, pH 7.0	0.3 M NaCl 0.03 M Sodium citrate
DEPC-H_2O	0.1% DEPC (v/v), autoclaved

2.1.8 Human Total RNA

To assess the performance of RNA amplification protocols, high quality total RNA was purchased from Stratagene (La Jolla, USA). Universal Human Reference RNA precipitate in ethanol was pelleted, washed in 70% (v/v) ethanol, air dried and dissolved in RNase-free water at 5 µg/µl, 500 ng/µl, 50 ng/µl, 5 ng/µl and 0.5 ng/µl. Human Adult Breast (female) RNA was precipitated at -80 °C for 30 min with 5 µg linear polyacrylamide (LPA), 2.5 vol 100% (v/v) ethanol and 0.5 vol 7.5 M NH_4OAc and subsequently processed as described for the Reference RNA. Integrity and purity of total RNA were assessed on a Bioanalyzer 2100 using an RNA 6000 Nano LabChip Kit according to the manufacturer's instructions (see 2.3.3).

2.2 Tumor Samples

2.2.1 Human HNSCC Tumor Samples

Tissue samples from six patients were obtained in the years 1998 - 2002 from patients undergoing surgical resection at the department of otolaryngology, J.-W.-Goethe-Universität Frankfurt. All cases were diagnosed histopathologically as HNSCC and staged according to the TNM classification of malignant tumors [161], based on criteria recommended by the 'Union International contre le Cancer (UICC)' (Table 1). The study protocol was approved by the local ethics committee after obtaining the patients' informed consent to participate in the study, and was processed anonymously. Grade 2 HNSCC specimens, corresponding healthy control mucosa surrounding the tumor and lymph node metastases were surgically resected, immediately frozen in liquid nitrogen and stored at -80 °C. The neoplastic specimens contained > 80% tumor tissue and < 10% necrotic debris.

The samples, collected and processed as described above, were kindly provided by Dr. med. Markus Hambek and Prof. Dr. med. Rainald Knecht

Table 8. Patient and disease characteristics.

Patient	Primary Site	Age	Sex	pT	pN	pM	Grading	Samples Analyzed[a]
160	hypopharynx	48	M	3	1	0	2	PT / N
171	hypopharynx	58	M	3	2a	0	2	PT / M
173	oropharynx	56	M	3	2	0	2	PT / N
180	hypopharynx	57	M	2	3	0	2	PT / N
186	hypopharynx	47	F	2	2	0	2	PT / N
205	oropharynx	49	M	3	1	0	2	PT / M

[a] All cases were diagnosed histopathologically as HNSCC and staged according to the TNM classification of malignant tumors. The indicated tissues were used for gene expression profiling. N: normal mucosa, PT: primary HNSCC, M: lymph node metastasis.

2.3 RNA Extraction

2.3.1 RNA Extraction from Tissue Samples

Frozen tissue samples (30-50 mg) were combined with 1 ml Trizol and dispersed using an Ultra-Turrax T25 tissue homogenizer. Total RNA was extracted according to the recommendations given by the Trizol protocol and further purified on RNeasy Mini spin columns. Extracted RNA material was kindly provided by Dr. Negusse Habtemichael (Chemotherapeutisches Forschungszentrum Georg-Speyer-Haus, Frankfurt am Main). Integrity and purity of total RNA were assessed on a Bioanalyzer 2100 using an RNA 6000 Nano LabChip Kit according to the manufacturer's instructions (see 2.3.3).

2.3.2 RNA Quantification by Spectrophotometry

The concentration of RNA and DNA solutions is usually determined by spectrophotometry. This procedure is based on the absorption maximum of nucleic acids at 260 nm, which is caused by the aromatic ring structures of the nucleotide bases. The absorption is measured in quartz cuvettes with a gage of 1 cm, which facilitates the estimation of concentrations via the law of Lambert-Beer:

$$A = \varepsilon \cdot c \cdot d$$

A: absorption (optical density; OD)

ε: molar absorption coefficient [$l \cdot mol^{-1} \cdot cm^{-1}$]

c: concentration [mol/l]

d: gage of the measured solution [cm]

Following calibration with the appropriate solvent, the sample readings should range between 0.05 and 1 OD, which is the linear range of common photometric devices. The concentrations can then be determined using one of these formulas:

$[DNA_{ds}] = OD_{260}$ x 50 µg/ml x dilution factor

$[DNA_{ss}] = OD_{260}$ x 37 µg/ml x dilution factor

$[RNA_{ss}] = OD_{260}$ x 40 µg/ml x dilution factor

The absorption maximum of proteins is at 280 nm, due to the absorbance of aromatic amino acids. The ratio OD_{260}/OD_{280} can be used as an indication for the purity of nucleic acid solutions. Pure DNA solutions have a ratio of about 1.8, whereas the ratio of pure RNA solutions is about 2.0. Contamination by phenol or proteins causes a significant decrease of this value, and solutions with ratios smaller than 1.5 have to be considered inappropriate for further analyses.

2.3.3 Assessment of RNA Integrity and Purity

Slab gel electrophoresis in cross-linked sieving matrices is one of the most powerful tools for nucleic acid analysis, but a major limitation is the low speed of analysis as well as the difficulty of automating the entire process, including sample loading, gel imaging and data analysis. The dimensions of a slab gel limit the electrical field strength that can be applied before severe band broadening occurs due to Joule heating and diffusion. This problem was addressed by introducing capillary electrophoresis (CE) in polymer matrices [162], which allowed nucleic acid separation under much higher electrical field strengths. CE allows for very fast nucleic acid separations when high electrical field strengths are applied. A more recent development involves the separation of DNA or RNA fragments on microfabricated devices (microfluidic chips; lab-on-a-chip technology) [163]. These chips allow the same speed of analysis as do separations on CE equipment and, moreover, due to their planar structure, make it easier to achieve parallel analysis. In addition, microfabricated chips allow the integration of several sequential steps in an automated manner.

The Agilent 2100 Bioanalyzer is the first commercially available system to utilize chip-based nucleic acid separation technology. Chips are fabricated from glass and comprise an interconnected network of fluid reservoirs and microchannels made by semiconductor-like microfabrication techniques, which must be filled with a polymer-dye mixture. Each chip contains a total of 16 wells. Three are used for loading the gel-dye mixture consisting of a linear polymer and a fluorescent, intercalating dye of a proprietary nature. One well is used for a molecular size ladder, and the 12 remaining ones for experimental samples. Both, the ladder and the samples have to be blended with a marker mixture consisting of a buffer along with lower and upper molecular size markers, which the Bioanalyzer uses as references when sizing nucleic acid fragments. The upper marker is also used as a reference for calculating the concentration of DNA or RNA fragments in each sample. The movement of nucleic acids through the microchannels is controlled by a series of electrodes, which create electrokinetic forces capable of driving fluids and DNA molecules. As the electrical voltage is applied, DNA fragments of different sizes intercalating with the fluorescent dye are separated according to their mass and can be quantified by laser-induced fluorescence detection. The Bioanalyzer displays data as both migration-time plots and as computer-generated virtual gels (Fig. 11).

Briefly, RNA 6000 nano chips were filled by pipetting 9 µl of polymer-dye mix into the appropriate well and then forcing the mix into the microchannels by applying pressure to the well via a 1 ml syringe. 5 µl of marker mix were loaded into each sample well, followed by 1 µl of molecular weight ladder into the ladder well and 1 µl aliquots of samples into the sample wells. The contents of each well on the chip were mixed by vortexing for 1 min at setting 4 using a Fisherbrand Vortex Genie-2. Subsequently, chips were immediately inserted into the Bioanalyzer and processed using the Bioanalyzer software settings “Eucaryotic total RNA nano” or “Eucaryotic mRNA nano”.

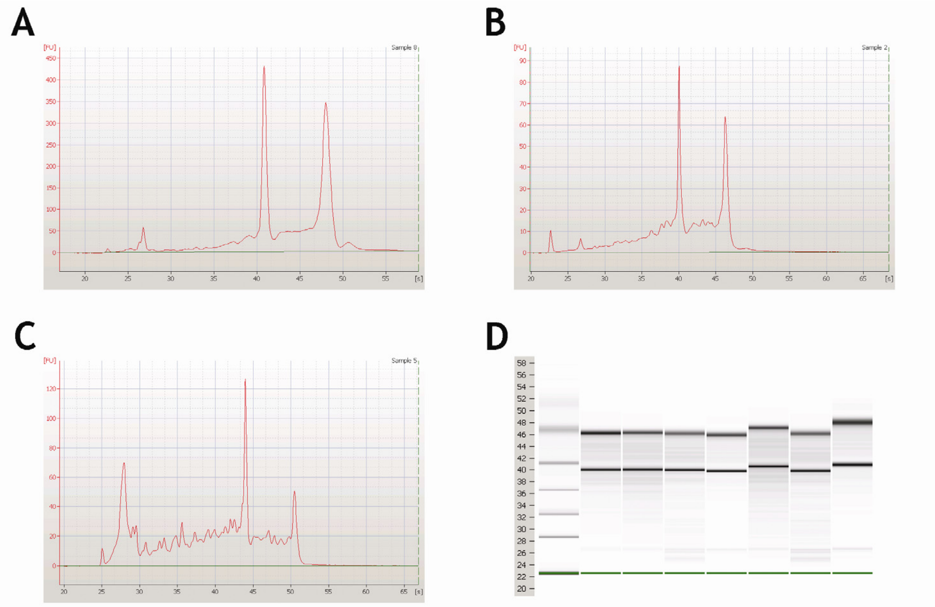


Figure 11. Assessment of total RNA (Stratagene Universal Human Reference RNA) integrity and purity with the Agilent 2100 Bioanalyzer. The Agilent 2100 Bioanalyzer allows for a fast and accurate estimation of the quality of RNA samples, both in an automated fashion and by visual inspection. Migration-time plots are shown for (**A**) high quality total RNA, (**B**) slightly degraded total RNA and (**C**) heavily degraded total RNA (inappropriate for further analyses). Computer-generated virtual gels (**D**) are useful for direct comparison of various samples.

2.4 Oligonucleotides

2.4.1 Probe Sequences for High-Density Oligonucleotide Microarrays

The Human Genome Oligo Set 2.1 and Human Genome Oligo Set 2.1 Upgrade, containing 21,329 and 5,462 70-mer probes, respectively, were purchased from Operon Technologies (Cologne). For enhanced coupling to the microarray surface, an amino linker is attached to the 5’ end of each oligo. The design of the original oligo set 2.1 is based on representative sequences from NCBI *Homo sapiens* UniGene build #147 (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene) and the NCBI RefSeq database, whereas the upgrade uses sequence and mapping information provided by the human Ensembl database build #13.31 (http://www.ensembl.org). All oligos are within a T_M range of 78 °C ± 5 °C and within 1,000 (original) or 2,000 nucleotides (upgrade) from the 3’ end of the available gene sequence. Single nucleotide repeats and stem-loop structures are avoided. Each oligo has ≤ 70% sequence identity and ≤ 20 contiguous bases common to all other genes represented in the collection.

2.4.2 Primer Sequences for RQ-PCR Analysis

The following primer sequences were used to verify gene expression ratios in the context of comparing results between the Operon long oligonucleotide and the Affymetrix GeneChip microarray platform. The primers were designed by Dr. Negusse Habtemichael (Chemotherapeutisches Forschungszentrum Georg-Speyer-Haus, Frankfurt am Main) an purchased from Biospring (Frankfurt am Main).

Table 9. Primer sequences used for RQ-PCR.

Gene Symbol	Forward Primer (5'→3')	Reverse Primer (5'→3')
OSF2	ATTAGGCTTGGCATCTGCTC	CTCGCGGAATATGTGAATCG
GMDS	GCGCTCATCACCGGTATCAC	CTCTGGGCTCCAAGGTTGTAG
TMPRSS2	TCCTGACGCAGGCTTCCAAC	CGAACACACCGATTCTCGTCC
BGN	TGGTTCAGTGCTCCGACCTG	GGATCTCCACCAGGTGGTTC

2.4.3 Padlock probes and associated sequences for RCA on microarrays

The following oligonucleotides (Biospring) were used for on-chip signal amplification via RCA.

Table 10. Padlock probes and associated sequences.

Oligo Name	Function	Modification	Sequence (5'→3')
PL-1216g	Padlock oligo, green system	5'-phosphate	GGGATTATAAAGAACTGTTGCCTCGACCG TTAGCAGCATGATTCCGAGATGTACCGCT ATCGTGTTGATGTCATGTGTCGCACTTCT TCTGGGCTAATTACAGC
PL-6121g	Padlock oligo, red system	5'-phosphate	CCCTAATATTTCTTGACAACGGAGCTGGC AATCGTCGTACTAAGGCTCTACATGGCGA TAGCACAACTACAGTACACAGCGTGAAGA AGACCCGATTAATGTCG
T-1216g	Ligation template, green system	5'-biotin	*GCAACAGTTCTTTATAATCCCGCTGTAAT TAGCCCAGAAGAA*
T-6121g	Ligation template, red system	5'-biotin	*CGTTGTCAAGAAATATTAGGGCGACATTA ATCGGGTCTTCTT*
1216g-double-3'	RT primer, green system	Inverse synthesis; 3'-thioat	3'-*dAsdCTACAGTACACAGCGTG*-5'-TTTTTTTTTTTTTTTTTTTTTVN-3'
6121g-double-3'	RT primer, red system	Inverse synthesis; 3'-thioat	3'-*dTsdGATGTCATGTGTCGCAC* -5'-TTTTTTTTTTTTTTTTTTTTTVN-3'
Detect1-1216g	Detection probe 1, green system	5'-Cy3; 3'-thioat	GGGATTATAAAGAACTGTdTsdG
Detect2-1216g	Detection probe 2, green system	5'-Cy3; 3'-thioat	CTCGACCGTTAGCAGCATGsdA
Detect3-1216g	Detection probe 3, green system	5'-Cy3; 3'-thioat	TCCGAGATGTACCGCTATCsdG
Detect4-1216g	Detection probe 4, green system	5'-Cy3; 3'-thioat	TCTTCTGGGCTAATTACAGsdC
Detect1-6121g	Detection probe 1, red system	5'-Cy5; 3'-thioat	CCCTAATATTTCTTGACAAsdC
Detect2-6121g	Detection probe 2, red system	5'-Cy5; 3'-thioat	GAGCTGGCAATCGTCGTACsdT
Detect3-6121g	Detection probe 3, red system	5'-Cy5; 3'-thioat	AGGCTCTACATGGCGATAGsdC
Detect4-6121g	Detection probe 4, red system	5'-Cy5; 3'-thioat	AGAAGACCCGATTAATGTCsdG
Inv1216Arl	RCA primer for spotting, green system	5'-Amino-C6	TCTCTCTCTCTCTCTCTCTCTCTCTCTCT CTCTCTCTCTCTCTCTCTC*GTGCGACACA TGACATCAAC*

2.5 Oligonucleotide Microarrays

2.5.1 Preparation of Fluorescent Targets

2.5.1.1 Reverse Transcriptase (RT) Labeling

For preparation of unamplified cDNA target, 40 µg of total RNA were heated for 4 min at 70° C in the presence of 2 µg oligo(dT_{21})VN in a total volume of 13.9 µl and chilled on ice. Labeling-mix was added, yielding final concentrations of 1x First-Strand Buffer, 10 mM dithiothreitol, 500 µM each of dATP, dGTP and dCTP, 200 µM dTTP, 100 µM Cy3- or Cy5-dUTP, 2 U/µl RNasin ribonuclease inhibitor as well as 13.33 U/µl Superscript II reverse transcriptase in a total volume of 30 µl. Samples were incubated first at 25° C for 3 min and, thereafter, at 42°C for 2 h, with further 200 U Superscript II (200 U/µl) added after one hour. 15 µl 0.1 M NaOH, containing 2 mM EDTA, were added to stop the reaction. RNA was hydrolyzed at 70 °C for 20 min. Finally, the pH was neutralized by addition of 15 µl 0.1 M HCl.

2.5.1.2 TAcKLE

For amplification and labeling using the TAcKLE protocol, 2000, 200, 20 or 2 ng total RNA were employed in first- and second-strand cDNA synthesis as previously described[164], with minor modifications. Briefly, RNA was mixed with 100 ng (dT)-T7 primer (5'-GCATTAGCGGCCGCGAAATTAATACGACTCACTATAGGGAGA$(T)_{21}$VN-3') to a final volume of 5 µl, denatured 4 min at 70 °C and chilled on ice. 5 µl ice-cold RT-mix were added to the samples, yielding final concentrations of 1x First-Strand Buffer, 10 mM DTT, 500 µM of each dNTP, 400 ng/µl T4gp32, 2 U/µl RNasin ribonuclease inhibitor as well as 10 U/µl Superscript II reverse transcriptase. Reverse transcription was performed for 1 h at 50 °C and reactions were stopped by heating to 65 °C for 15 min. Following the addition of 65 µl ice-cold reaction-mix, second-strand synthesis was performed for 2 h at 15 °C in 1x Second-Strand buffer, 200 µM of each dNTP, 0.27 U/µl DNA polymerase I, 1 U RNase H and 5 U *E. coli* DNA ligase. Then, 10 U T4 DNA polymerase (3.33 U/µl) were added to the samples and cDNA ends were polished for 15 min at 15 °C. Enzymes were heat inactivated by 10 min incubation at 70 °C. To extract double-stranded cDNA, samples were mixed with 75 µl phenol/chloroform/isoamylalcohol (pH 8.0) and transferred to prespun

0.5 ml PLG heavy tubes. After 5 min centrifugation at 13,000 rpm, the aqueous phase was further purified on a P-6 Micro BioSpin column according to the manufacturer's instructions, followed by ethanol precipitation. The cDNA was dissolved in 10 µl nuclease-free water and employed in an *in vitro* transcription reaction using a RiboMAX Large Scale RNA Production System T7 according to the manufacturer's recommendations, but in 40 µl reaction volume and regularly mixing the samples every 30 min for 6 h. Following purification on RNeasy Mini filters and ethanol precipitation, aRNA was dissolved in nuclease-free water, preferentially at 0.25 µg/µl.

Second round RT was performed on 1 µg aRNA (where available) as described above, but with the following modifications: 0.5 µg random hexamer primer was used instead of (dT)-T7 primer. Samples were incubated 5 min at room temperature before addition of RT-mix to allow for annealing of N_6-primer. The following temperature profile was employed for reverse transcription: 20 min at 37 °C, 20 min at 42 °C, 10 min at 50 °C, 10 min at 55 °C, 15 min at 65 °C. RNase H digestion (1 U per reaction) was carried out for 30 min at 37 °C, followed by 2 min at 95 °C to degrade enzymes.

When starting with 20 ng total RNA or less, two rounds of amplification were performed. For this purpose, purified aRNA samples were precipitated, dissolved in 4 µl nuclease-free water and subjected to second round reverse transcription as described above. First-strand cDNA was mixed with 100 ng (dT)-T7 primer in a final volume of 11 µl, incubated 10 min at 42 °C and chilled on ice. Thereafter, second-strand synthesis, cDNA purification, *in vitro* transcription, aRNA clean-up and third round reverse transcription (primed with random hexamers) were performed as described above.

cDNA labeling by Klenow fragment was performed using the Bioprime Kit, but with a modified protocol. Briefly, 10 µl cDNA sample were mixed with 90 µl Klenow-mix to yield a reaction mixture that contained 1x random primer solution, 200 µM each of dATP, dCTP and dGTP, 50 µM dTTP, 30 µM Cy3- or Cy5-dUTP and 0.8 - 1 U/µl Klenow fragment. DNA polymerization was carried out at 37 °C for 16 h.

2.5.1.3 Sample Purification

To remove unincorporated nucleotides and nucleotide-dye conjugates, cDNA samples were purified on Microcon YM-30 filter columns. This separation is based on size exclusion centrifugation (ultrafiltration) through a custom cellulose membrane. The membrane pores are permeable for salts, nucleotides and small macromolecules of up to 30 kDa (50 bp, 60 b), whereas the cDNA product is retained on the membrane.

Corresponding cDNA samples were combined and purified on Microcon YM-30 filter columns, as previously described[90]. For blocking of repetitive sequence elements, 25 µg C_0t-1 DNA, 25 µg poly-A RNA and 75 µg yeast tRNA were added before the final washing step.

2.5.2 Preparation and Post-Processing of Microarrays

2.5.2.1 Microarray Spotting

Synthetic 70mer oligonucleotides ("Human Genome Oligo Set Version 2.1"; consisting of 21,329 oligonucleotides representing human genes and transcripts plus 24 controls, as well as “Human Genome Oligo Set Version 2.1 Upgrade”, consisting of 5,462 human 70mer probes) were purchased from Operon Technologies (Cologne) and dissolved in FBNC spotting buffer[165] (25% (v/v) **f**ormamide, 0.5 M **b**etaine, 0.5 µg/µl = 0.05% (w/v) **n**itro**c**ellulose, 2.5% (v/v) DMSO) at 40 µM, using a MiniTrak robotic liquid handling system. DNA spotting was performed in duplicates on QMT epoxysilane coated slides using an OmniGrid Microarrayer equipped with Stealth SMP3 Micro Spotting Pins. Spot centers were 129 µm apart.

2.5.2.2 DNA Immobilization

DNA adhesion to the glass surface was accomplished by 1 h incubation at 60 °C, followed by UV irradiation (2x 120 mJ/cm^2 at 254 nm) in a Stratalinker Model 2400 UV illuminator.

2.5.3 Microarray Hybridization

Just prior to hybridization, slides were washed for 2 min in 0.2% SDS (w/v), 2 min in ddH_2O at room temperature and 2 min in boiling ddH_2O (95 °C), followed by 3 min centrifugation at 2,000 rpm.

Purified, dye-labeled cDNA was mixed with 120 µl UltraHyb hybridization buffer (Ambion), agitated for 30-60 min at 60 °C, then for 10 min at 70 °C on a thermo mixer and subsequently applied to pre-heated (60 °C) microarrays mounted in a GeneTAC Hybridization Station. Hybridizations were performed for 16 h at 42 °C with gentle agitation. Thereafter, the arrays were automatically washed at 36 °C with (i) 0.5x SSC, 0.1% (w/v) SDS for 5 min; (ii) 0.05x SSC, 0.1% (w/v) SDS for 3 min; (iii) 0.05x SSC for 2 min. Flow time was set to 40 sec, respectively. Immediately after completion of the final washing step, the arrays were unmounted, immersed in 0.05x SSC, 0.1% (w/v) Tween 20 and dried by centrifugation in 50 ml Falcon tubes (30 sec at 500, 1000 and 1500 rpm, respectively, followed by a final step of 90 sec at 2000 rpm).

2.5.4 Data Acquisition

Hybridized microarrays were scanned at 5 µm resolution and variable PMT voltage to obtain maximal signal intensities with <0.1% probe saturation, a count ratio of 0.8-1.2 (Cy5 / Cy3) and maximal congruence of histogram curves, using a GenePix 4000B microarray scanner. Subsequent image analysis was performed with the corresponding software GenePix Pro 5.0. Spots not recognized by the software were excluded from further considerations.

2.5.5 Affymetrix GeneChip Arrays

5 µg of total RNA were used to prepare fragmented, biotinylated cRNAs for hybridization, following the guidelines given in the Affymetrix GeneChip Expression Analysis Technical Manual [166]. cRNA clean-up was performed on RNeasy Mini filters. 10 µg of fragmented, labeled cRNA were hybridized to Affymetrix HG U133A arrays (Affymetrix, Santa Clara, CA) using standard conditions (16 h, 45 °C). Arrays were washed and stained in a Fluidics Station 400 (Affymetrix) and scanned on a Gene Array Scanner 2500 (Agilent), as recommended by Affymetrix. This work was performed by Dr. Negusse Habtemichael (Chemotherapeutisches Forschungszentrum Georg-Speyer-Haus, Frankfurt am Main).

2.5.6 Data Processing and Analysis

If debugging is the process of removing bugs, then programming must be the process of putting them in.

EDSGER DIJKSTRA

Most of the data processing and analysis was carried out in close collaboration with Dr. Carina Ittrich (Central Unit Biostatistics, German Cancer Research Center). Dipl.-Biol. Grischa Toedt (Division of Molecular Genetics, German Cancer Research Center) provided valuable support, especially concerning data normalization, by means of the 'ChipYard' framework for microarray data analysis.

2.5.6.1 Data Normalization

Result files containing all relevant scan data were further processed using the open source statistical software environment **R** (http://www.r-project.org) [167] together with libraries (packages) of the Bioconductor project (http://www.bioconductor.org) [168]. For each hybridization, raw fluorescence intensities were normalized applying variance stabilization [169]. For GeneChip arrays, raw fluorescence intensities from all hybridizations were normalized applying variance stabilization [169] with additional scaling. Additionally, *MAS5* [170] as well as *gcRMA* [171] expression values were calculated.

2.5.6.2 Data Filtering

To eliminate low quality data from spotted microarrays, the data points were ranked according to spot homogeneity, as assayed by the ratio of median to mean fluorescence intensity, the ratio of spot to local background intensity and the standard deviation of the logarithmic ratios ($\log_2$ Cy5 / Cy3) between spot replicates. Those data points ranked among the lower 20% were removed from the data set. Genes that could not be quantified in more than 33% of all experiments after filtering were excluded as well. To combine the data of dye swap experiments, the $\log_2$-transformed intensity ratios of one array were inverted and averaged with the corresponding values of the other array.

Concerning the comparison of the array platforms, an optional filtering procedure additionally excluded those data points considered unreliable [172,173] as they correspond to probes associated with signal intensities less than two standard deviations above local background for at least one channel of the pair of Operon chips or to probe sets with mean $\log_2$ expression values below the median for all probe sets of the pair of GeneChips [172,173]. This strategy was chosen in order to extract high quality data from both array platforms as a sound foundation for quantitative comparisons. More sophisticated filtering based on variance rather than absolute expression levels was not applied due to the deliberate shortage of replicates. Expression ratios of genes with a signal close to the background (low abundance) in only one of the two investigated conditions are clearly significant in a biological context. They were, however, considered less appropriate to this comparative study, as their results were expected to carry an increased and mathematically inevitable degree of variation not caused by characteristics of the investigated platforms.

2.5.6.3 Orthogonal Regression and Pearson Correlation

To investigate the linear relationship between data points, regression lines were determined by minimizing the sum of squares of the Euclidean distance of points to the fitted line (“orthogonal regression”), as there is no clear assignment of dependent and independent variables. Correlations were estimated using the Pearson correlation coefficient together with its 95% confidence interval.

2.5.6.4 Linear Modeling

To compare $\log_2$ ratios obtained by TAcKLE amplifications of 2,000, 200, 20 and 2 ng starting material to those obtained by RT labeling, a linear model with RT labeling as reference was fitted separately for each gene. *p*-values were calculated using Wald statistics. This analysis was performed for all spots with quantified $\log_2$ ratios in at least 9 of the 10 arrays remaining after exclusion of self-self and dye swap hybridizations (see Table 11), hence the Wald statistics were checked for significance using a *t*-distribution with 4 or 5 degrees of freedom, respectively. The magnitude of the effects as well as the corresponding *p*-values are illustrated as volcano plots [174].

2.5.6.5 Identification of Differentially Expressed Genes

Identification of differentially expressed genes was performed by empirical Bayes inference for paired data[175]. Moderated *t*-statistics, based on shrinkage of the estimated sample variance towards a pooled estimate and corresponding *p*-values, were calculated using the Bioconductor **R** package *limma*[176]. *P*-values were adjusted according to the method proposed by Benjamini and Hochberg[177] to control the false discovery rate at a level of 10%. The magnitude of the effects as well as the corresponding *p*-values are illustrated as volcano plots[174].

2.5.6.6 Distance Weighted Discrimination

To remove systematic variation resulting from the different technical approaches of the investigated array platforms or differences in sample handling procedures between the two labs participating in this study, 'Distance Weighted Discrimination (DWD)'[178] was performed on normalized $\log_2$-ratios from both array platforms, using Matlab software freely available at https://genome.unc.edu/pubsup/dwd/. Further details about cross platform adjustment of microarray data can be obtained at http://genome.med.unc.edu:8080/caBIG/DWDNCI60.htm and http://genome.med.unc.edu:8080/caBIG/paper1.pdf. Identification of differentially expressed genes and DWD were only performed for spots with quantified $\log_2$ ratios in all four primary HNSCC *versus* normal mucosa experiments.

2.5.6.7 GO Data Mining and EASE Overrepresentation Analysis

GO (Gene Ontology)[179] data mining was performed using the GOCharts functionality of the 'Database for Annotation, Visualization and Integrated Discovery (DAVID)'[180], which is available at http://david.niaid.nih.gov/david/. Overrepresentation analysis was carried out with the software application 'Expression Analysis Systematic Explorer (EASE)'[181], downloaded from http://david.niaid.nih.gov/david/ease.htm.

2.5.6.8 Matching of Oligonucleotide Probe Sequences

The Bioconductor **R** package *AnnBuilder*[182] and GenBank accession numbers, provided by Affymetrix and Operon, were used to map probe sequences to corresponding UniGene clusters (build #175). Microarray data were only used if the Affymetrix probe set and the Operon probe corresponded to the same UniGene cluster from the intersection of both platforms (n = 4,425). For simplicity, if probe sets

(Affymetrix) mapped to multiple UniGene clusters or if several probes (Operon) or probe sets (Affymetrix) mapped to the same UniGene, they were excluded from further analyses.

2.5.7 Accession Numbers

Science is organized knowledge. Wisdom is organized life.

IMMANUEL KANT

All relevant data from the TAcKLE study are available from GEO (http://www.ncbi.nlm.nih.gov/geo) under the accession numbers GPL1384 (for the array platform), GSM27816-GSM27819, GSM27835, GSM27836 and GSM27915-GSM27928 (for expression data from individual arrays) as well as GSE1645 (for the experimental series).

Data from the cross-platform comparison are available under the accession numbers GPL96 and GPL1384 (for the array platforms), GSM29702-GSM29705, GSM29747-GSM29758, GSM29808-GSM29813, GSM29818 and GSM29820 (for expression data from individual arrays) as well as GSE1722 (for the experimental series).

2.6 RQ-PCR Analysis

For selected genes, changes in mRNA levels detected in microarray experiments were evaluated by reverse transcription (RT) and quantitative real-time PCR analysis, using the iCycler (BioRad, Munich, Germany). This work was carried out by Dr. Negusse Habtemichael (Chemotherapeutisches Forschungszentrum Georg-Speyer-Haus, Frankfurt am Main). One microgram of total RNA was converted to cDNA using Superscript II reverse transcriptase (Invitrogen) and oligo(dT) primer, according to the manufacturer's specifications. PCR reaction mixtures consisted of 12.5 µl of 2x iQ™ SYBR® Green Supermix (Abgene, Hamburg, Germany), 0.5 µl of each 10 µM target primer and 1 µl diluted cDNA template (1:10) in a reaction volume of 25 µl. Thermal cycling conditions comprised an initial denaturation step of 15 min at 95 °C, 40 cycles of 30 s at 95 °C and 30 s variable annealing/elongation temperature, depending on the respective set of target primers. dsDNA-specific fluorescence was measured at the end of each extension phase. Product-specific amplification was confirmed by a melting curve analysis. The relative expression ratio (*R*) of a target gene was calculated using the equation:

$$R = \frac{\left(E_{\text{target}}\right)^{\Delta CP_{\text{target (control−sample)}}}}{\left(E_{\text{ref}}\right)^{\Delta CP_{\text{ref (control−sample)}}}}$$

based on its real-time PCR efficiencies (*E*) and the crossing point (*CP*) differences of sample *versus* a control, and expressed in comparison to a reference gene [183]. The target gene expression was normalized to glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*).

2.7 Rolling Circle Amplification

2.7.1 Padlock Probe Ligation

To prepare the circular oligonucleotides used as templates for the rolling circle amplification, DNA ligation was performed for 1.5 h at 37 °C in a final volume of 1 ml containing 1 µM linear padlock oligo (Table 10), 3 µM biotinylated ligation template (Table 10), 1x T4 DNA ligase buffer, 0.5 µg/µl BSA, 1 mM ATP as well as 0.05 U/µl T4 DNA ligase. The reaction was stopped by addition of EDTA *ad* 5 mM.

2.7.2 Purification of Circularized Padlock Probes

To separate the ligated oligonucleotides from the biotinylated ligation templates and, thus, obtain pure circular probes, the ligation product was purified using streptavidin-coated magnetic beads. Briefly, 1500 µl suspension of beads was washed with an equal amount of 1x B&W buffer (binding and washing buffer; 5 mM Tris-HCl, 0.5 mM EDTA, 1 M NaCl, pH 7.5) - as a magnetic particle concentrator was unavailable, the beads were separated from the supernatant by 1 min centrifugation at 2000 rpm - and resuspended in 1500 µl 2x B&W buffer. Subsequently, 1000 µl ligation product as well as 500 µl ddH_2O was added and the suspension was incubated 30 min at RT, applying gentle rotation or frequent mixing to keep the beads in suspension and thus allow for efficient binding of the ligation templates bound to the circularized probes. Then, the beads were again washed with 1500 µl 1x B&W buffer, separated by centrifugation and resuspended in 500 µl ddH_2O. Finally, the sample was denatured for 2 min at 95 °C and the beads were spun down. The supernatant containing the circularized probes was collected and transferred to a fresh tube. The DNA concentration was measured by spetrophotometry and, if applicable, adjusted in a vacuum concentrator to obtain an appropriate working concentration of 0.1 - 1 µM.

2.7.3 Reverse Transcription with Aminoallyl-dUTP

For preparation of aminoallyl-tagged cDNA suitable for subsequent rolling circle amplification, 50 µg total RNA were heated for 4 min at 70 °C in the presence of 2 µg

1216g-double-3’ or 6121g-double-3’ primer (Table 10) in a total volume of 13.9 µl and chilled on ice. Labeling-mix was added, yielding final concentrations of 1x First-Strand Buffer, 10 mM DTT, 500 µM each of dATP, dGTP and dCTP, 320 µM dTTP, 166.66 µM aminoallyl-dUTP, 2 U/µl RNasin ribonuclease inhibitor as well as 13.33 U/µl Superscript II reverse transcriptase in a total volume of 30 µl. Samples were incubated first at 25 °C for 3 min and, thereafter, at 42 °C for 1 h. 15 µl 0.1 M NaOH, containing 2 mM EDTA, were added to stop the reaction. RNA was hydrolyzed at 70 °C for 20 min. Finally, the pH was neutralized by addition of 15 µl 0.1 M HCl, and the samples were purified using Microcon YM-30 centrifugal filter devices and PBS.

2.7.4 Psoralen Conjugation (Preparation of cDNA-psoralen conjugates)

Psoralen ($C_{11}H_6O_3$; CAS # 66-97-7) is the basic material of the furocoumarin family of drugs. The substance can be found in essential oils from plants of the *Psoralea* genus of *Leguminosae*. If activated by UV-A light, psoralen can introduce inter-strand cross-links between nucleotides in complementary DNA strands. For this reason, it is used together with UV light to treat skin diseases such as psoriasis, vitiligo and even skin nodules of cutaneous T-cell lymphoma. We intended to use the cross-linking potential of psoralen as a means of covalently coupling hybridized cDNA strands to their complementary oligonucleotides attached to the microarray surface.

To derivatize the aminoallyl-tagged cDNA with psoralen, 20 µl purified cDNA sample were mixed with 15 µl conjugation buffer (1 M sodium bicarbonate $NaHCO_3$, 150 mM NaCl) and 40 nmole SPB (succinimidyl-[4-(psoralen-8-yloxy)]butyrate; $C_{19}H_{15}NO_8$; NHS-psoralen; Pierce Biotechnology, Rockford, USA) in 1.2 µl DMSO. SPB is an amine-reactive NHS ester coupled to the DNA-intercalating, photoreactive psoralen-group. The NHS ester will cross-link to primary amines in target molecules (aminoallyl-tagged cDNA) at pH 7-9 to form stable amide bonds (Fig. 12). For this purpose, the sample was protected from light and incubated for 30 min at RT. Subsequently, it was purified by 2 PBS washes using a Microcon YM-30 centrifugal filter device.

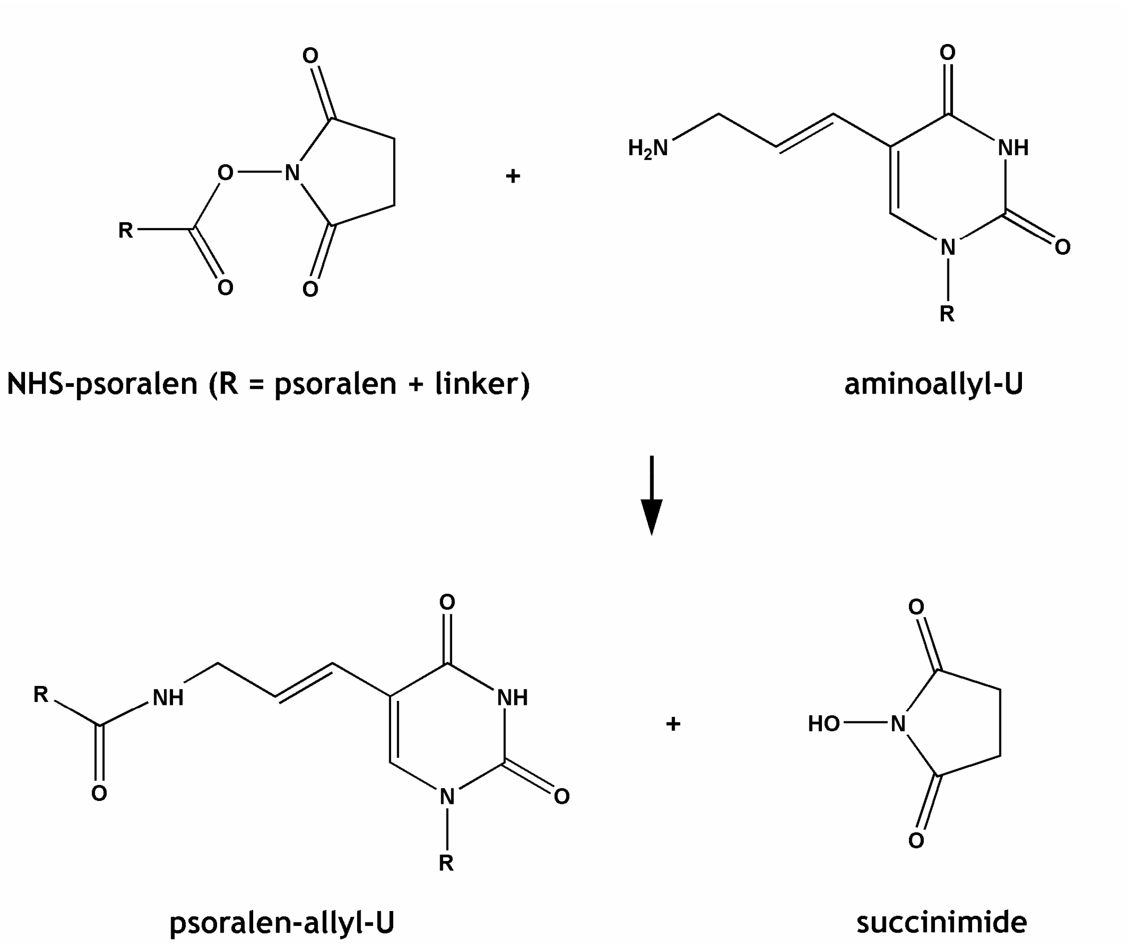


Figure 12. Coupling of NHS-psoralen to aminoallyl residues.

2.7.5 Photoreactive Coupling

For photoreactive coupling of psoralen-derivatized cDNA molecules to their complementary array probes, human 27K oligonucleotide microarrays were prepared as described in 2.5.2.1 and hybridized with psoralen-cDNA as described in 2.5.3. Subsequently, the arrays were exposed for 15 min to 366 nm UV-A light (8 W) from a distance of 3 cm, yielding inter-strand cross-links to the 5,6 double bonds in thymine residues (Fig. 13).

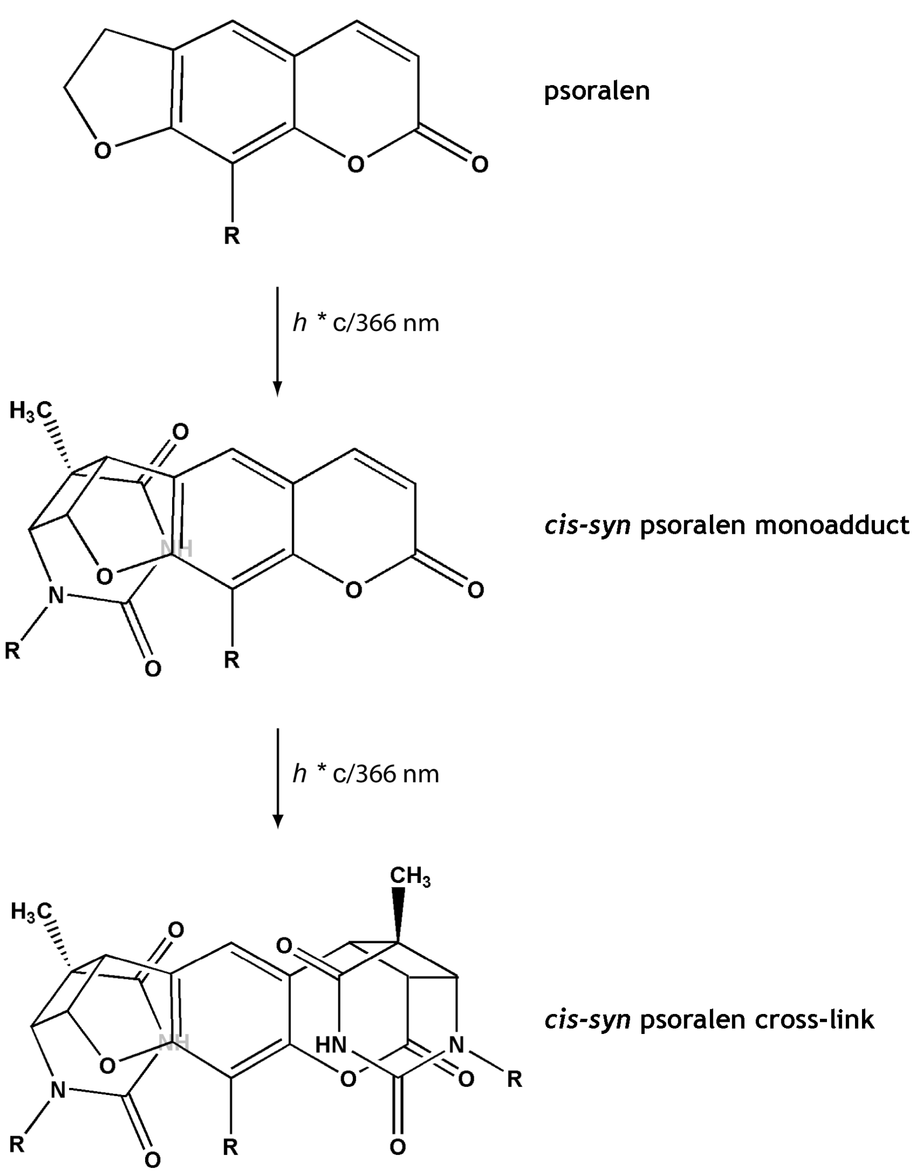


Figure 13. Photoreactive inter-strand cross-linking of dsDNA via psoralen. The figure was modified from an illustration kindly provided by cand. biol. Daniel Haag.

2.7.6 Rolling Circle Amplification on Oligonucleotide Microarrays

The reaction mixture for on-chip RCA contained 1x Φ29 buffer, 0.5x SSC, 200 µM of each dNTP, 200 nM circularized PL-1216g and/or circularized PL-6121g, 20 nM of each Detect1-1216g - Detect4-1216g and/or 20 nM of each Detect1-6121g - Detect4-6121g as well as 350 ng (35 U) Φ29 DNA polymerase in a total volume of 120 µl. To perform the reaction, human 27K oligonucleotide microarrays containing complementary, psoralen-coupled cDNA were mounted in a GeneTAC Hybridization Station and overlaid with 120 µl reaction mixture. Subsequently, the following temperature profile was applied: 15 min at 25 °C to allow for hybridization of the circular oligonucleotide templates to their complementary sequences in the cDNA, 30 - 90 min at 37 °C for rolling circle replication and another 15 min at 25 °C for efficient hybridization of the dye-labeled detection probes. Finally, the arrays were automatically washed at 25 °C with (i) 0.75x TNT buffer (0.1 M Tris-HCl, pH 7.5) for 3 min, (ii) 0.1x SSC, 0.1% (w/v) SDS for 2 min and (iii) 0.05x SSC for 1 min, followed by 5 min of automatic draining. The arrays were scanned as decribed in 2.5.4.

3 Results

3.1 Target Amplification for Oligonucleotide Microarrays

3.1.1 Motivation

The large amount of required RNA material is a restricting aspect of any array-based expression profiling approach. cDNA arrays usually require at least 15 µg total RNA, and the preferred amount for spotted oligonucleotide arrays is increased to about 50 µg, due to the decrease in possible base pairings. Hence, reliable transcriptome amplification is essential for many quantitative analytical approaches, such as RNA expression analysis of tumor biopsies [107], sorted cell populations [184], laser capture microdissected cells and tissues [185] or any other study based on small tissue samples or minute numbers of cells. Methods were developed that amplify initial poly(A) RNA and, thereby, increase detection sensitivity by orders of magnitude.

Amplification can, in principle, either be performed exponentially using PCR-based approaches [108-110], or in a linear fashion, mostly by generation of cDNA followed by *in vitro* transcription with T7 RNA polymerase [111,112,164,186]. However, the kinetics of PCR-based methods implies that both sequence-dependent and copy-number dependent bias will be amplified exponentially as well and accumulate. Another important issue is the influence of sampling errors when handling very limited amounts of RNA [115,187]. For these reasons, exponential amplification protocols are generally considered less applicative for quantitative transcriptome analyses.

T7-based methods, on the other hand, are routinely used for expression profiling studies in combination with cDNA microarrays, and several studies have demonstrated their reliability [164,186]. Recently, large collections of long oligonucleotides (50-80 bases) have become increasingly popular as probes for spotted DNA arrays. Technical advantages of oligonucleotide arrays include a constant DNA concentration across all spots and biophysically optimized sequences, reducing secondary structures, avoiding repetitive sequences and providing a fixed range for both T_m and length. This accounts for more uniform, stable and predictable hybridization conditions.

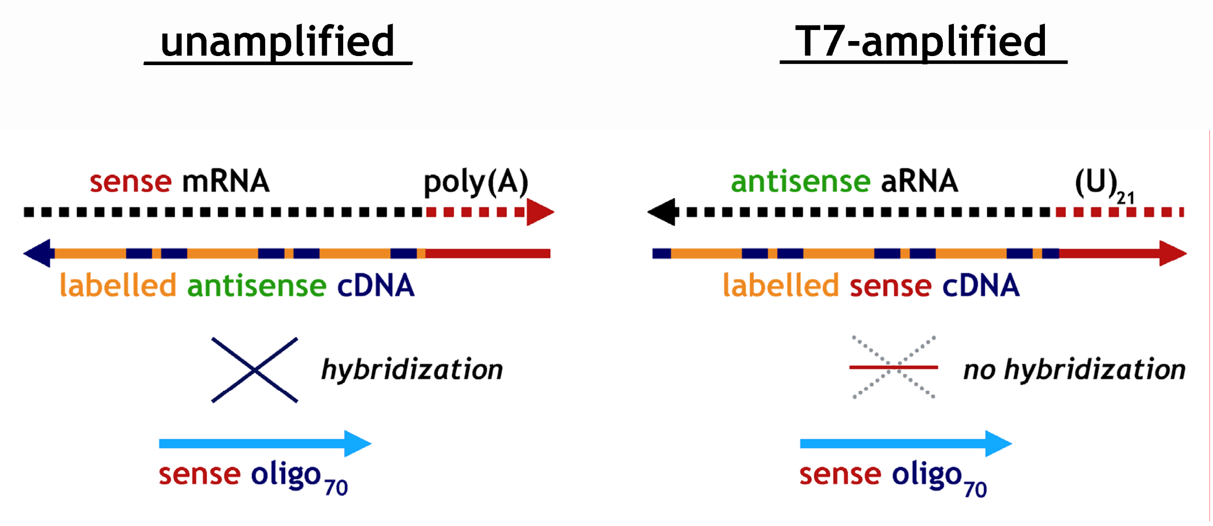


Figure 14. Compatibility of unamplified and T7-amplified target molecules with sense-oriented oligonucleotide arrays. Sense-oriened mRNA can be transformed to fluorescent antisense-oriented cDNA by means of a simple reverse transcription reaction. The resulting first-strand cDNA can be hybridized to oligonucleotide arrays with probes in sense orientation. Fluorescent cDNA targets obtained by reverse transcription labeling of antisense-oriented amplified RNA, on the other hand, are incompatible for hybridization to sense-oriented oligonucleotide libraries.

However, starting from cellular, sense-oriented mRNA, the orientation of T7-amplified RNA will be antisense (aRNA). Therefore, it cannot be used for reverse transcription labeling and hybridization to sense-oriented, gene specific oligonucleotide libraries. Oligonucleotides of commercial libraries are sense-oriented to complement antisense targets produced by reverse transcription of unamplified RNA. Sense cDNA derived from aRNA is incompatible for hybridization to these sequences (Fig. 14).

3.1.2 Protocol Variants

The best way to have a good idea is to have lots of ideas.

LINUS PAULING

Both the conception and the execution of the initial experiments presented in 3.1.2 must be ascribed in part to Dipl.-Biol. Dirk Olaf Thürigen.

Some approaches try to overcome the strand orientation problem by producing labeled aRNA during *in vitro* transcription [122] (Fig. 8), but in our hands the yield of this procedure was insufficient. Therefore, the method was not further investigated.

Another option for the generation of antisense-oriented target molecules would be *in vitro* transcription from a promoter positioned upstream of the cDNA sequence, *e.g.*, by exploiting the template-switching effect[188] of Moloney murine leukemia virus (MMLV) reverse transcriptase ("inverse IVT", Fig. 15).

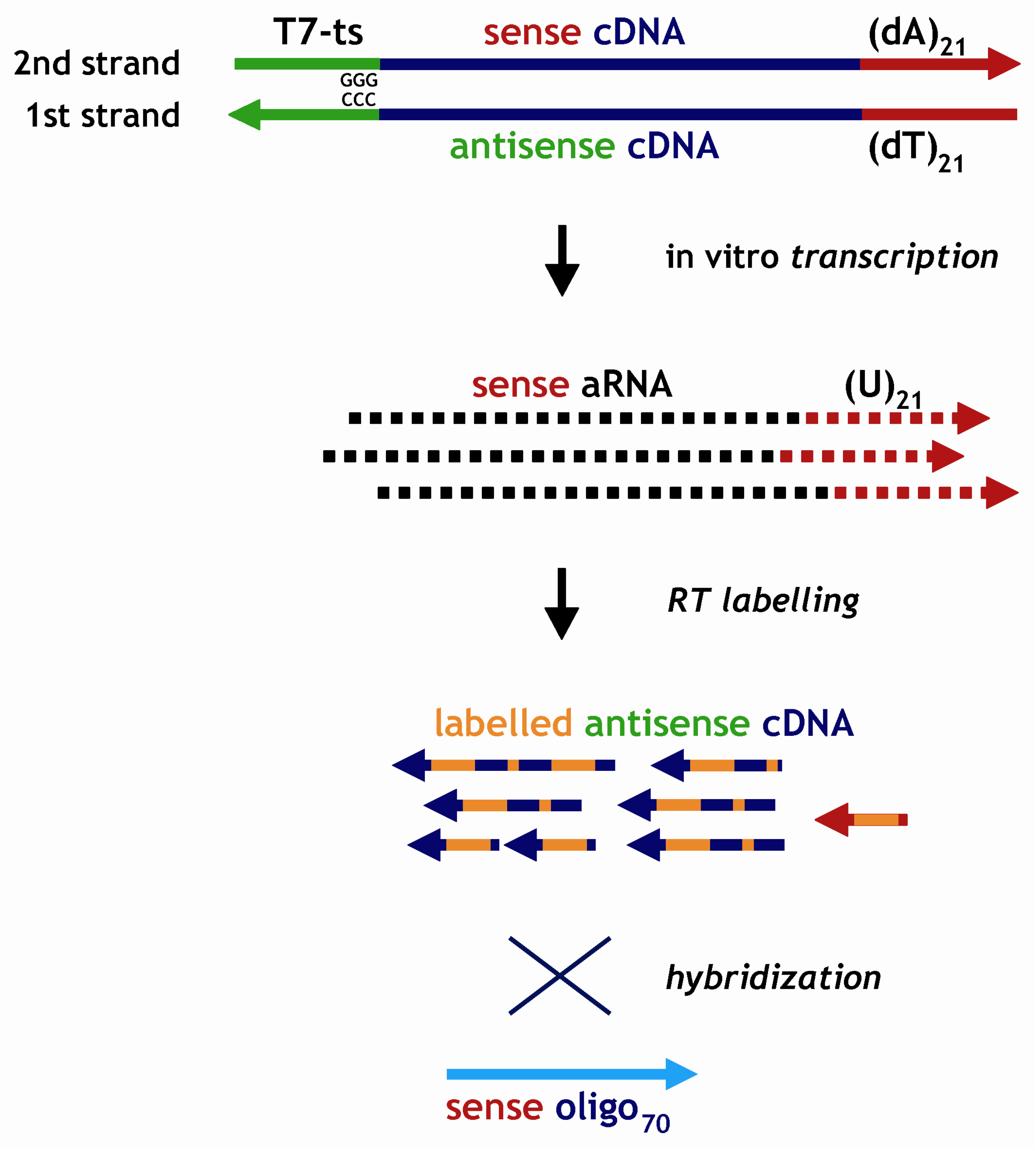


Figure 15. Inverse *in vitro* transcription. The inverse IVT method exploits the template-switching effect of MMLV reverse transcriptase to permit *in vitro* transcription from a promoter positioned upstream of the cDNA sequence. See text for details.

It had been observed that, upon reaching the 5'-end of the RNA template, MMLV reverse transcriptases add a few non-template nucleotides, primarily deoxycytidine, to the 3'-end of a newly synthesized cDNA strand. Oligonucleotides with an oligo(dG) sequence at the 3'-end (so-called template-switch or TS-oligos) can basepair with the deoxycytidine stretch at the 3'-end of the cDNA, creating an extended template that causes the enzyme to continue replicating to the end of the oligonucleotide [188]. Thus, a sequence complementary to the TS-oligo is attached to the cDNA 3'-end. Following second strand cDNA synthesis, repeated transcription from the T7-promoter sequence incorporated by the TS-oligo yields multiple copies of sense-oriented RNA molecules. These can be labeled by reverse transcription and hybridized to microarrays containing oligonucleotide probes in sense orientation (Fig. 15).For unknown reasons, however, this method did not generate sufficient amounts of RNA material, and no analyzable results could be obtained.

In a recent publication, Smith *et al.* [123] claimed that their amplification procedure, termed "Single Primer Amplification" (SPA), could also be used for microarrays containing oligonucleotide probes in sense orientation, enabled by a "strand switch" of the polymerase used for target labeling (Fig. 9). Unfortunately, these considerations could not be confirmed in practice. Arrays hybridized with cDNA targets prepared by the SPA method displayed hardly any perceivable signals, and no interpretable data were obtained.

A protocol termed "Template Switch Single Primer Amplification" (ts-SPA) was devised as a combination of the basic concepts of the inverse IVT approach and SPA. Once again the template-switching effect [188] of Moloney murine leukemia virus (MMLV) reverse transcriptase is exploited, but this time the specific sequence attached upstream of the cDNA is used as a priming site for *Taq* DNA polymerase rather than being a promoter for an RNA polymerase. A primer complementary to the sequence introduced by the TS-oligo binds to the first-strand cDNA and drives amplification by *Taq* DNA polymerase cycling extensions. The resulting sense-oriented second strand molecules can be used as templates for dye labeling by Klenow fragment (Fig. 16).

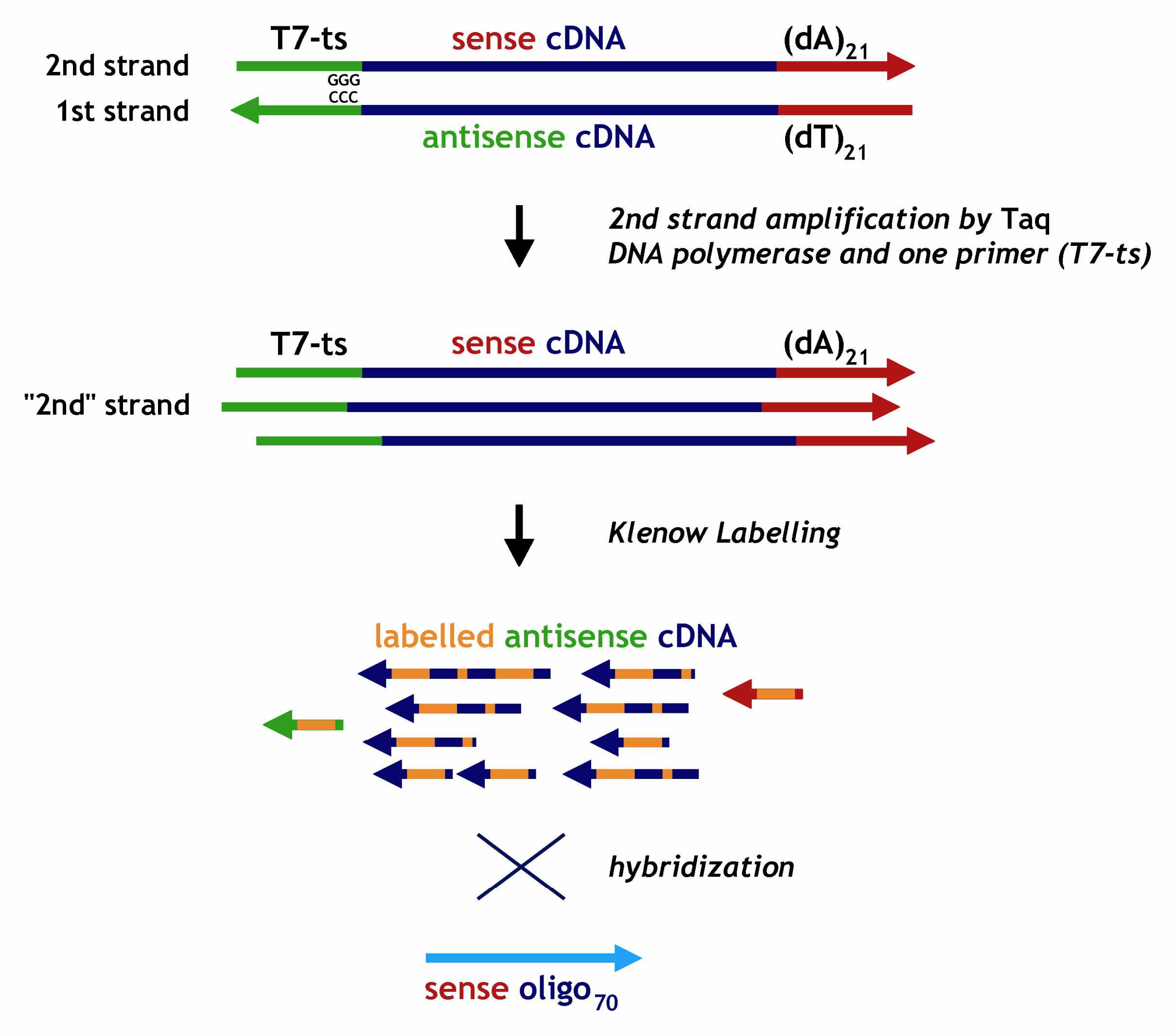


Figure 16. Template Switch Single Primer Amplification (TS-SPA). The protocol exploits the template-switching effect of MMLV reverse transcriptase, thermal cycling with *Taq* DNA polymerase and dye labeling by Klenow fragment. Details are given in the text.

Two-color co-hybridizations of fluorescent target molecules prepared from 2 µg total RNA of the cell line HL-60 yielded reasonable correlations of fluorescence intensities (Fig. 17A). However, hybridizing targets prepared from 2 µg HL-60 total RNA *versus* such made from 2 µg total RNA of DLHL cells (RNA from both cell lines was kindly provided by Dipl.-Biol. Olaf Thuerigen) did not significantly reduce the degree of correlation (Fig. 17B). Consequently, using the TS-SPA method would drastically decrease our ability to identify differentially expressed genes, and it was decided not to conduct further investigations concerning this protocol.

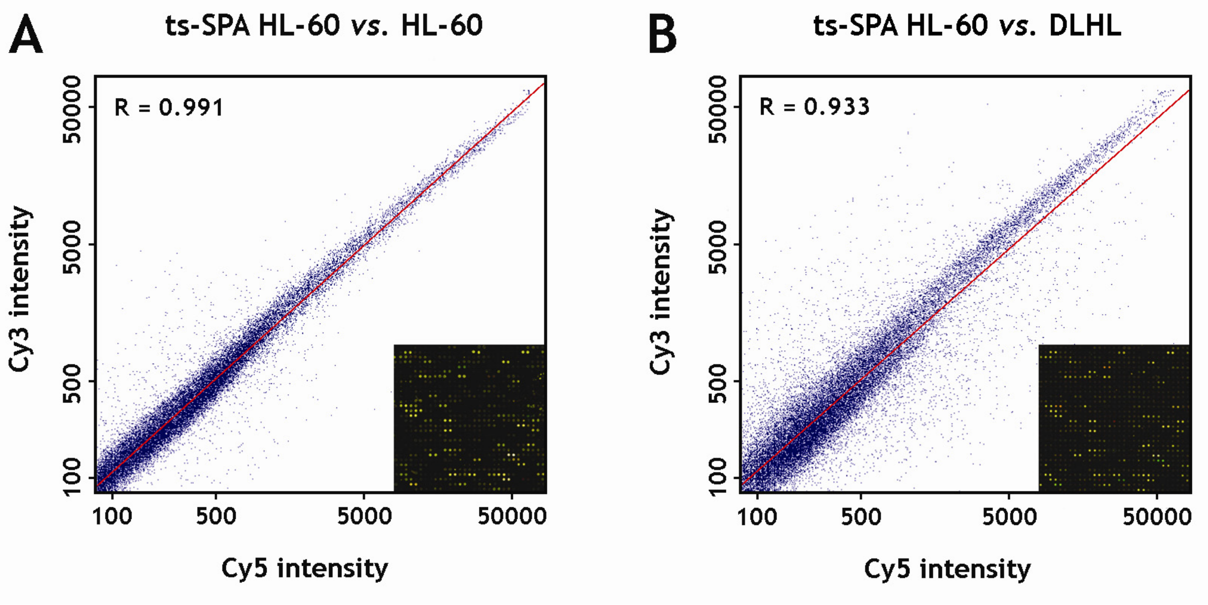


Figure 17. Scatter plots of fluorescence intensities from amplification and labeling reactions using the TS-SPA method. (**A**) Co-hybridizations of independently amplified HL-60 RNA were used to assess the reproducibility of amplification. Fluorescent targets prepared from HL-60 RNA were hybridized *versus* targets prepared from DLHL RNA to estimate the method's potential to identify differentially expressed genes.

3.1.3 T7 Amplification and cDNA Klenow Labeling for Expression Analysis

One more protocol for the generation of labeled antisense cDNA was devised. The method utilizes mRNA amplification by *in vitro* transcription (IVT) of cDNA, as first described by van Gelder *et al.* [112], and fluorescent labeling by Klenow fragment.

Initial mRNA is copied by an RNase H^- MMLV reverse transcriptase, using a modified oligo(dT)-primer to incorporate the promoter sequence of phage T7 RNA polymerase. An RNase H^- polymerase is used to ensure that an RNA/DNA hybrid is produced. RNase H treatment of this heteroduplex creates RNA fragments that prime second strand synthesis by *E. coli* DNA polymerase I. Repeated transcription from the T7 promoter on the cDNA template results in multiple copies of antisense RNA (aRNA), which may be reamplified as previously described [111].

Briefly, random hexamers prime another round of reverse transcription, which generates sense-oriented cDNA with 5'-oligo(dA). An oligo(dT)-T7 primer recognizes this sequence and initiates second strand synthesis, once more generating double-stranded cDNA with a T7-promoter downstream of the transcript sequence. Then, another round of *in vitro* transcription yields additional antisense-oriented copies (aRNA) of the initial mRNA.

Finally, and regardless of the number of transcription reactions, random hexamers are used to reversely transcribe the aRNA into sense cDNA. This serves as template for randomly primed Klenow labeling, yielding mainly fluorescent antisense cDNA as a suitable target for oligonucleotide libraries in sense orientation (Fig. 18).

As initial experiments with this protocol were very promising (data not shown), it was decided to evaluate the method, termed **T**arget **A**mplification and **c**DNA **K**lenow **L**abeling for **E**xpression analysis (**TAcKLE**), in more detail.

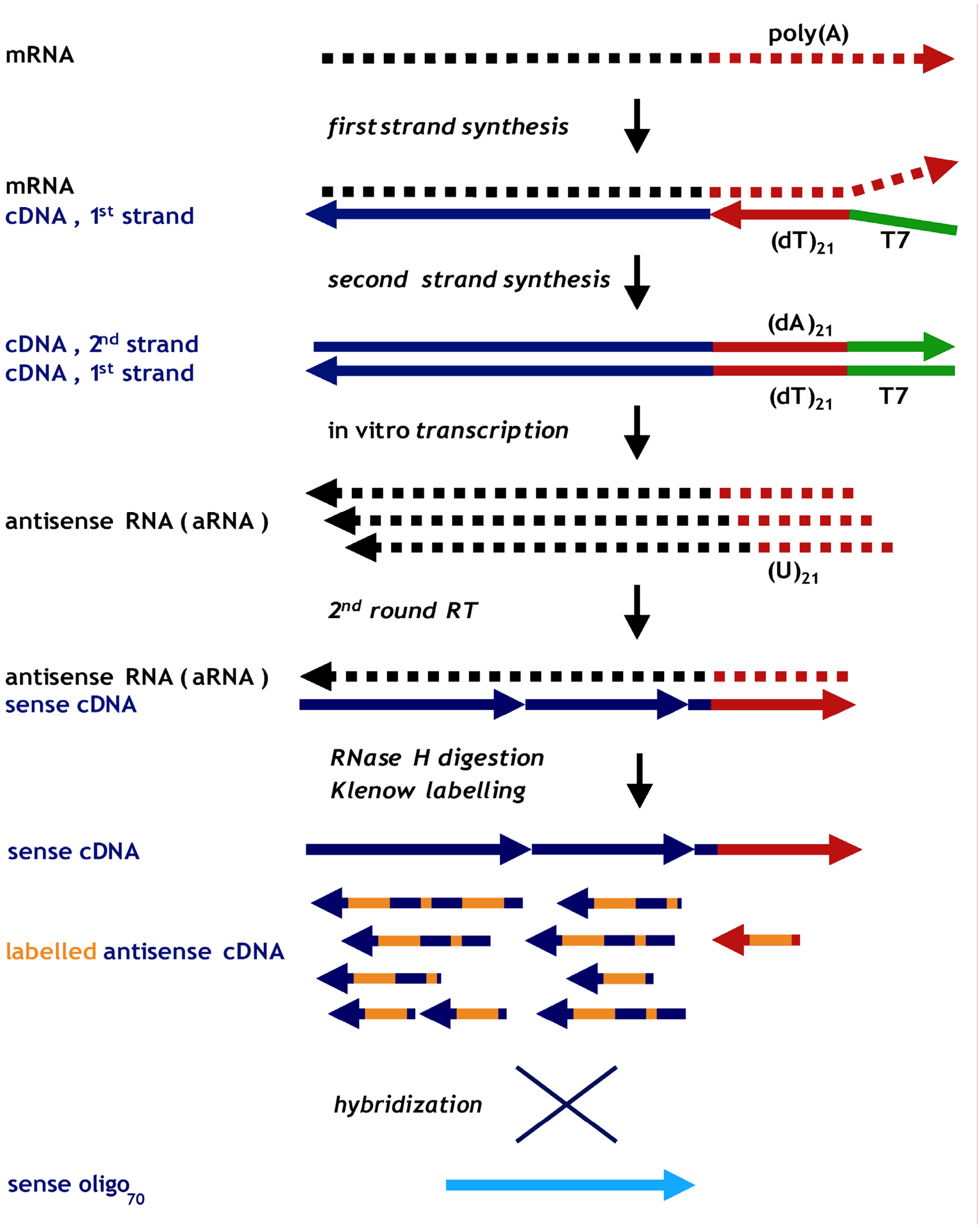


Figure 18. Schematic overview of the TAcKLE protocol. mRNA is linearly amplified by *in vitro* transcription ("T7 amplification"). The resulting aRNA is subsequently converted to cDNA and labeled by dye-dUTP incorporation using Klenow fragment. The resulting antisense-oriented cDNA is suitable for hybridization to sense-oriented oligonucleotide microarrays.

3.1.4 Experimental Design

Things should be made as simple as possible, but not any simpler.

ALBERT EINSTEIN

A single source of reference (pooled from ten human cell lines representing distinct tissues) and breast total RNA was used for all experiments to avoid variations in transcript abundance imposed by the RNA preparation. Each RNA pool was serially diluted to provide four distinct starting quantities equivalent to 2, 20, 200 and 2,000 ng. In total, 20 two-color hybridizations were performed, comprising one co-hybridization of reference RNA, two hybridizations of breast RNA *versus* reference RNA (Cy5 / Cy3) and one hybridization of reference RNA versus breast RNA (dye swap), both for TAcKLE amplifications of all four amounts of input material and for reverse transcription labeling (Table 11). All dye labeling reactions using Klenow fragment were made from separately amplified RNA aliquots. One round of linear RNA amplification resulted in approximately 10^3-fold amplification of starting mRNA and two rounds yielded up to 10^5-fold the starting amount, as determined by spetrophotometry and based on an estimated initial mRNA content of 2%. Labeled cDNAs were hybridized to microarrays containing 26,791 gene specific 70mer oligonucleotide probes, each spotted in duplicate.

Table 11. Experimental design (*n* denotes the number of arrays).

Group	Labeling Method	Input / Channel	Breast *vs*. Ref. (*n*)	Ref. *vs*. Breast (*n*)	Ref. *vs*. Ref. (*n*)
1	TAcKLE	2 ng	2	1	1
2	TAcKLE	20 ng	2	1	1
3	TAcKLE	200 ng	2	1	1
4	TAcKLE	2000 ng	2	1	1
5	RT	40 µg	2	1	1

3.1.5 Reproducibility of Amplification

Hybridizations of differentially labeled targets, independently prepared from the same dilutions of reference RNA, were performed as a first assessment of random bias introduced by the amplification and labeling procedure. The Pearson correlation coefficient of fluorescence intensities (Fig. 19) was high for all tested amounts of input RNA (r = 0.9945, r = 0.9900, r = 0.9905 and r = 0.9657 for 2,000 ng, 200 ng, 20 ng and 2 ng starting material, respectively) and in good agreement with previous reported values for T7-based amplification protocols [164,186]. This reflects a reliable amplification and consistent labeling with both Cy5- and Cy3-dUTPs. There is an increased scattering of low intensity data points for 2 ng of starting material, which might be attributed to sampling errors [115,187] (*i.e.*, errors resulting from the stochastic distribution of low-copy-number templates) and represents a restricting aspect when depending on very strong amplifications. Still, the reproducibility of the amplification is equivalent or even superior when compared to target preparation by reverse transcription (r = 0.987; Fig. 20A).

Of note, the correlation drops considerably if targets prepared from different amounts of starting material are compared (Fig. 22A).

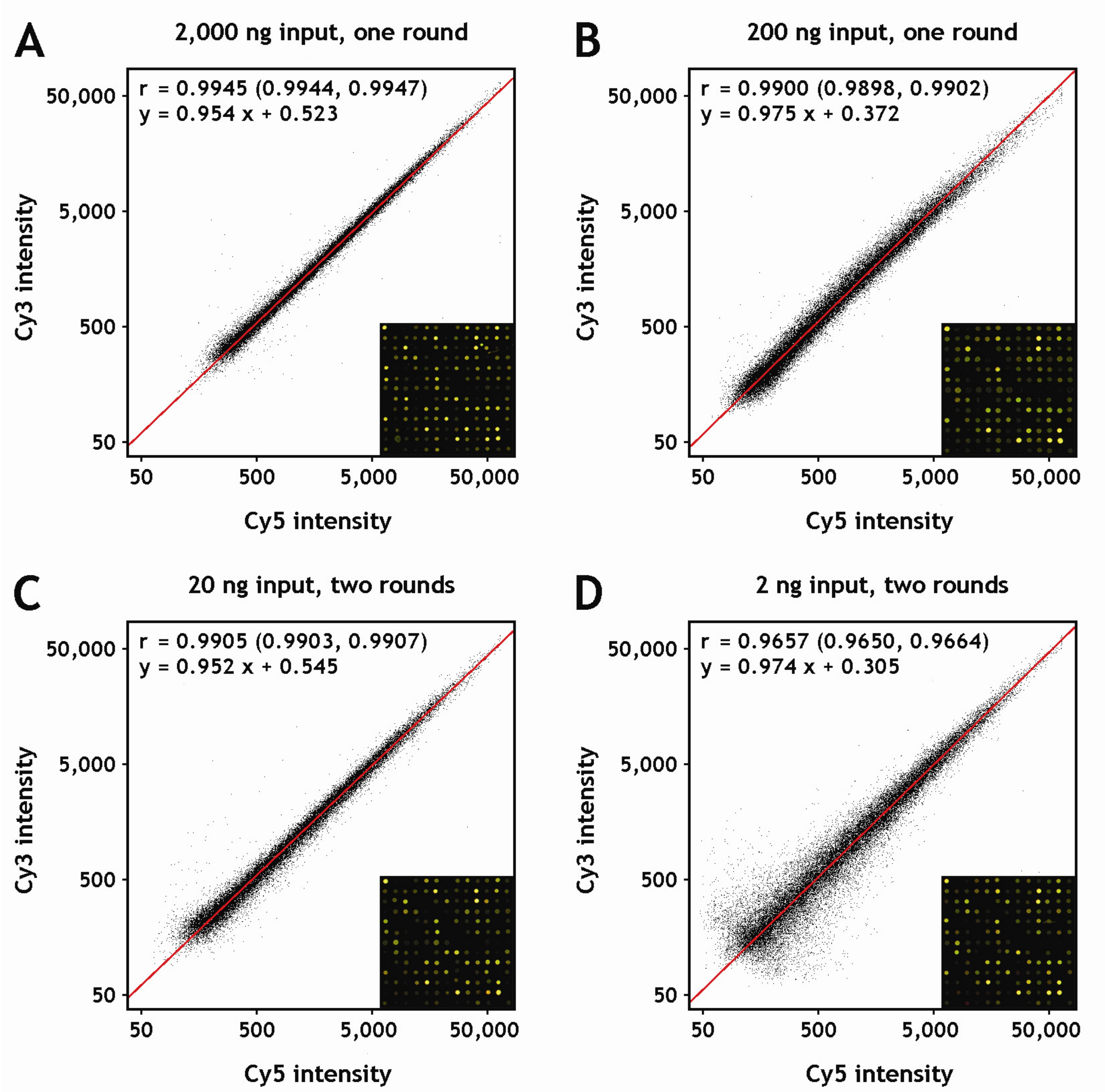


Figure 19. Scatter plots of fluorescence intensities from replicate amplification and labeling reactions. Co-hybridizations of independently amplified reference RNA were used to assess the reproducibility of amplification under diverse conditions. (**A**) 2,000 ng, (**B**) 200 ng, (**C**) 20 ng, (**D**) 2 ng starting material. Orthogonal regression lines are shown in red; the corresponding linear equations are given together with Pearson correlation coefficients and their 95% confidence intervals. A defined section of the respective microarray image is displayed in the lower right corner of each plot.

3.1.6 Reproducibility of Expression Ratios with and without Dye Swap

Hybridizations of targets derived from human reference RNA and RNA extracted from normal human breast tissue were compared to determine the effect of the amplification procedure on the reproducibility of expression ratios. The Pearson correlations of $\log_2$-transformed normalized expression ratios were r = 0.9948, r = 0.9889, r = 0.9780 and r = 0.9938 for identically repeated hybridizations as well as r = -0.9803, r = -0.9496, r = -0.9424 and r = -0.9017 for hybridizations repeated with inverse assignment of fluorophores (dye swap), starting from 2,000 ng, 200 ng, 20 ng and 2 ng RNA material, respectively (Fig. 21). Apparently, the concordance of expression ratios is stable and independent of the amount of input RNA for identically repeated experiments, but decreases considerably in case of dye swap repeats as the amount of starting material is reduced. This might reflect differences in dye incorporation between cyanine-3 and cyanine-5 labeled dUTP, a known bias previously reported for fluorescent cDNA prepared by reverse transcription labeling [189]. The respective correlations for these unamplified targets were r = 0.986 and r = -0.872 (Fig. 20B-C).

Once again, using unequal amounts of starting material for the hybridization experiments to be compared resulted in considerably decreased correlations (Fig. 22B-D).

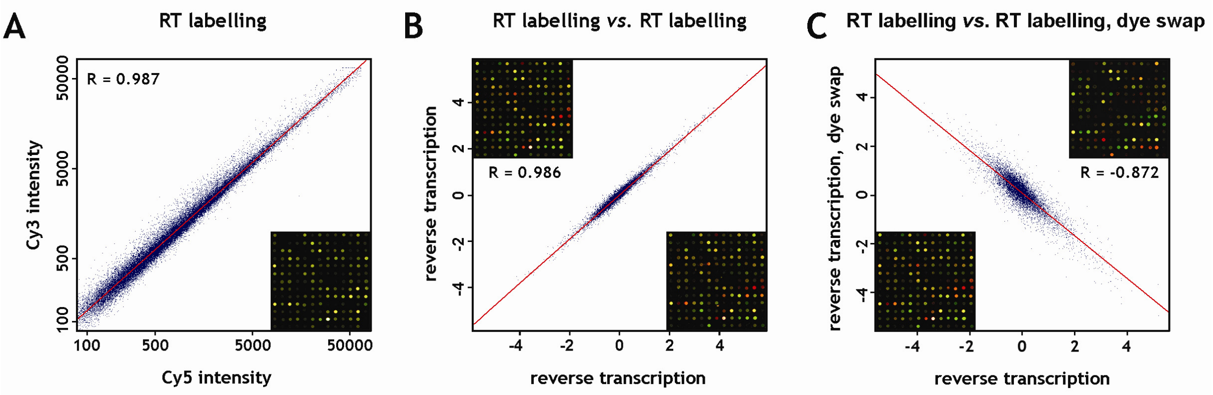


Figure 20. Characteristics of reverse transcription labeling reactions. (**A**) Scatter plot of fluorescence intensities from duplicate reverse transcription labeling reactions using universal reference RNA. (**B**-**C**) Scatter plots of $\log_2$-transformed expression ratios ($\log_2$ Cy5 / Cy3) from duplicate hybridizations of RT-labeled breast and reference RNA, with and without reversed assignment of fluorophores (dye swap); replicate spots were averaged before plotting. Defined sections of the respective microarray images are displayed in (**A**) the lower right or (**B**-**C**) the lower right (adscissa) and upper left (ordinate) corner of each plot. Orthogonal regression lines are shown in red.

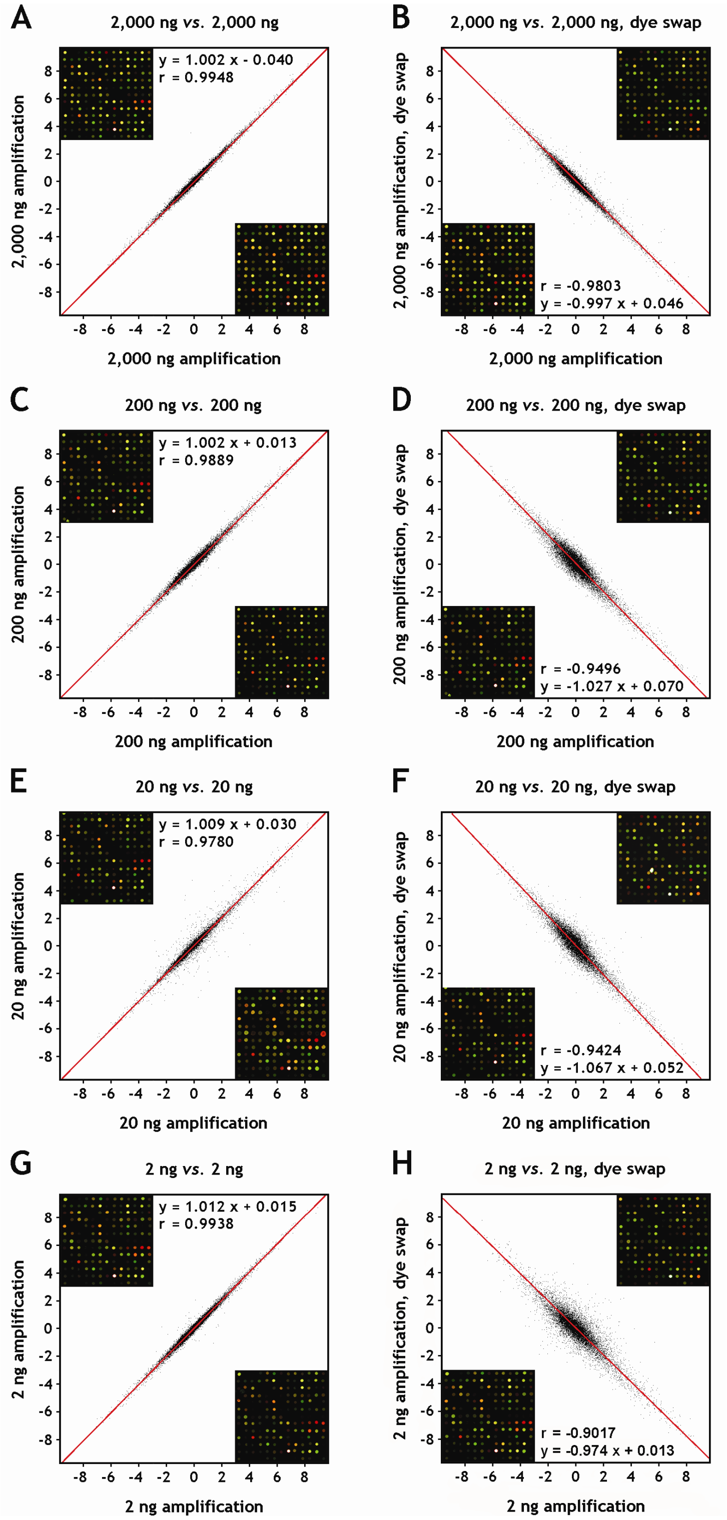


Figure 21. Scatter plots of $\log_2$-transformed expression ratios ($\log_2$ Cy5 / Cy3) from duplicate hybridizations. Amplified breast and reference RNA, with and without reversed assignment of fluorophores (dye swap) was employed to evaluate the accuracy and reproducibility of the experiment. Replicate spots were averaged. (**A**) 2,000 ng; (**B**) 2,000 ng, dye swap; (**C**) 200 ng; (**D**) 200 ng, dye swap; (**E**) 20 ng; (**F**) 20 ng, dye swap; (**G**) 2 ng; (**H**) 2 ng, dye swap. The data were subjected to orthogonal regression analysis (red lines), associated linear equations are listed along with Pearson correlation coefficients. The 95% confidence intervals of the correlation coefficients are (0.9946, 0.9950), (-0.9809, -0.9796), (0.9885, 0.9892), (-0.9513, -0.9478), (0.9772, 0.9788), (-0.9444, -0.9404), (0.9936, 0.9940) and (-0.9050, -0.8983) for panels (**A**) through (**H**). Underlying microarray images are shown as fixed sections in an upper (ordinate) and lower (abscissa) corner of each plot.

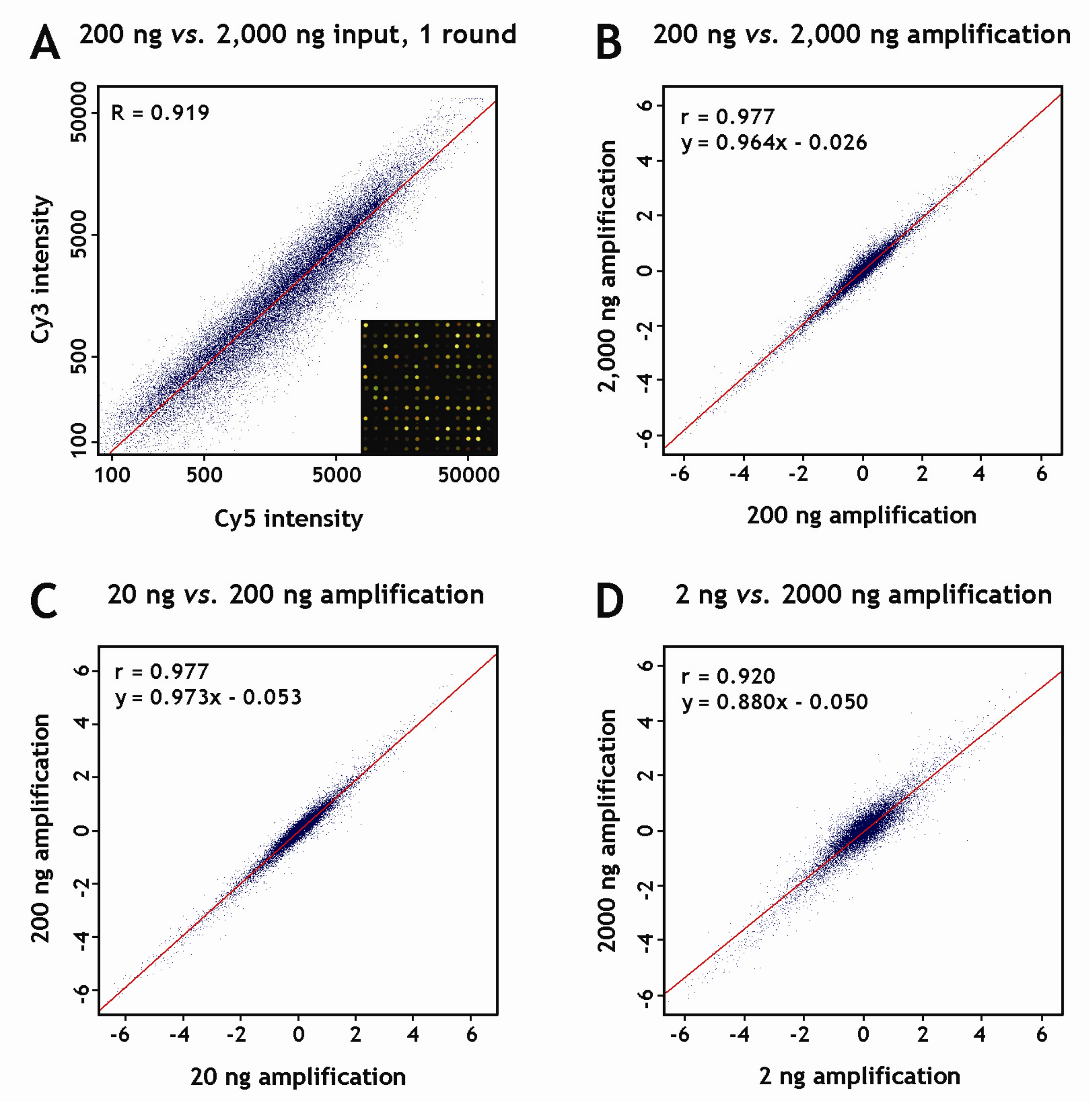


Figure 22. Adverse effect of using unequal amounts of starting material. (**A**) 200 ng and 2,000 ng universal reference RNA were used in discrete amplification and labeling reactions with the TAcKLE protocol. The resulting cDNA targets were co-hybridized to the same array, and their fluorescence intensities were used to calculate the Pearson correlation. A defined section of the array is shown in the lower right corner of the plot. (**B-D**) To assess the influence of variable amounts of starting material on the agreement of expression ratios from replicate chip experiments, 2,000 ng, 200 ng, 20 ng and 2 ng breast and reference RNA were independently amplified by the TAcKLE protocol and used for duplicate microarray hybridizations with reversed assignment of fluorophores (dye swap). The $\log_2$-transformed expression ratios were averaged and used to calculate Pearson correlations of (**B**) 200 ng *vs.* 2,000 ng, (**C**) 20 ng *vs.* 200 ng and (**D**) 2 ng *vs.* 2,000 ng starting material.

3.1.7 Comparison of Amplified and Unamplified Targets

The main practical application of microarray analysis is the identification of transcripts whose abundance differs between samples. To test the fidelity of target amplification, we determined the ratios of amplified breast cDNA *versus* amplified universal reference cDNA hybridizations, and examined how these correlated with the corresponding ratios obtained with unamplified targets. This analysis was used to test whether amplified targets would identify the same set of differentially expressed transcripts recognizable with unamplified targets. Not unexpectedly, Pearson correlations of the corresponding $\log_2$ ratios ($r = 0.8727$, $r = 0.8713$, $r = 0.8565$ and $r = 0.8441$ for the comparison of RT labeling to amplifications of 2000 ng, 200 ng, 20 ng and 2 ng starting material) were not as high as for the comparison of repeated experiments (Fig. 23). The scattering of corresponding values increases towards higher absolute $\log_2$ ratios. Additionally, we observed an increase in the slope of the regression lines ($m = 1.325$, $m = 1.338$, $m = 1.355$ and $m = 1.379$; same order as above), demonstrating a common deviance in the absolute $\log_2$ ratios. On average, absolute ratios obtained with amplified targets were higher than those corresponding to the unamplified samples, prepared by reverse transcription labeling.

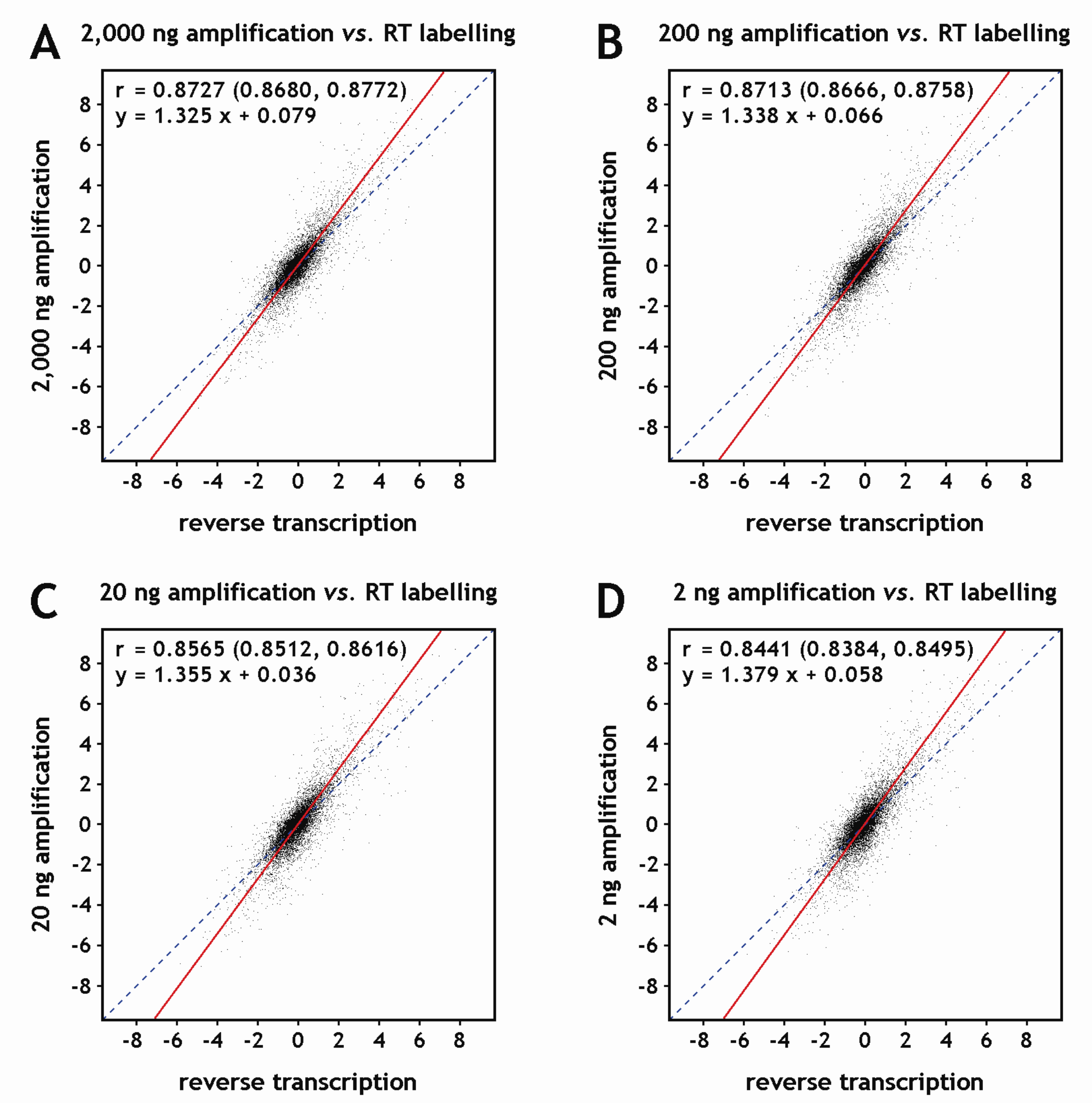


Figure 23. Scatter plots comparing $\log_2$-transformed expression ratios of amplified targets to ratios obtained with unamplified targets. Breast and reference RNA was used as starting material. Dye swap experiments were combined before plotting. Target amplified (TAcKLE) from (**A**) 2,000 ng, (**B**) 200 ng, (**C**) 20 ng and (**D**) 2 ng starting material was compared to unamplified target prepared by reverse transcription labeling. Orthogonal regression analysis was performed to derive the regression lines shown in red and their respective linear equations. Dashed lines through origin with slope 1 are displayed to accentuate the elevated slope. Pearson correlation coefficients and their associated 95% confidence intervals are listed as well.

3.1.8 Linear Modelling and Statistical Analysis

To determine whether target amplification affected our ability to reliably profile gene transcription in the breast tissue, the relationship of the observed differences of $\log_2$ ratios between amplified *versus* unamplified targets and the degree of differential expression was analyzed. We found 1479, 1483, 1444 and 1667 genes to be up-regulated as well as 1237, 1291, 1376 and 1598 genes to be down-regulated in samples TAcKLE-amplified from 2000 ng, 200 ng, 20 ng and 2 ng RNA of healthy human breast tissue when compared to universal human reference RNA. 1171 and 993 genes were identified as up- or down-regulated by reverse transcription labeling, respectively. Apparently, and in agreement with previous reports, target amplification yielded a slightly larger number of differentially expressed genes [190,191]. The distribution of $\log_2$ ratios for the genes detected as differentially expressed in amplified and/or unamplified targets is depicted in Figure 24, which shows that a substantial number of those genes found by merely one method were close to reaching the threshold for differential expression (2-fold difference) with the other method as well. This observation is strengthened in Figure 25, where of the genes common to the data sets under comparison, only very few displayed a deviation of $\log_2$ ratios greater than 1 or smaller than -1 (44 and 47, 72 and 57, 45 and 66 as well as 85 and 115 genes, respectively, for the comparison of dye labeling by reverse transcription to TAcKLE amplifications using 2000 ng, 200 ng, 20 ng and 2 ng starting material; Fig. 25). Additionally, we applied a linear model to assign *p*-values to these differences. The results are displayed as volcano plots [174,192] (see 2.5.6.4) of *p*-value against $\log_2$ ratio difference (Fig. 26). These show the statistical significance of the observed differences in relation to their magnitude. Supporting the findings of Figure 25, similarly small numbers of genes (26 and 33, 59 and 43, 34 and 52, 68 and 100) showed a significant ($p < 0.001$) difference of $\log_2$-transformed ratios when comparing across the target preparation techniques. In Figure 25, the intersection of the "outliers" from all amounts of starting material contains 275 genes for the unfiltered data sets and is empty for the filtered data sets. For Figure 26, the respective numbers of genes are 246 for the unfiltered data sets and 18 for the filtered data sets. No more than 1-4% of the considered probes were affected by a ≥ 2-fold difference. Accordingly, there is strong concordance between expression ratios obtained with amplified and unamplified targets.

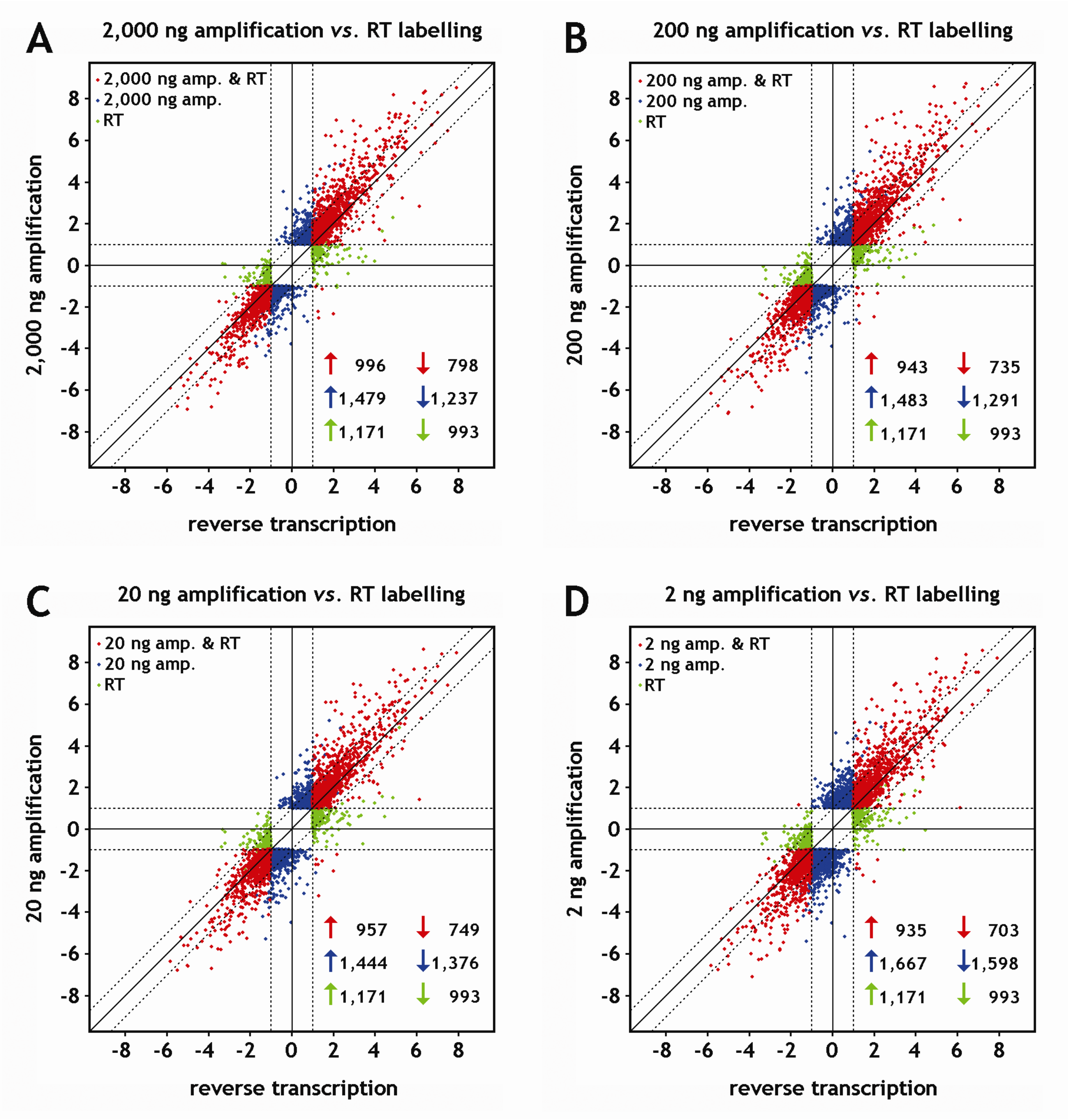


Figure 24. Scatter plots showing $\log_2$ ratios of the genes detected as differentially expressed between breast and reference RNA by either one or both target preparation techniques (reverse transcription labeling and amplification *via* the TAcKLE protocol). Data are shown for the comparisons of RT labeling *versus* targets prepared from (**A**) 2,000 ng, (**B**) 200 ng, (**C**) 20 ng and (**D**) 2 ng starting material. Genes showing differential expression with both methods are shown as red dots, while blue and green dots denote genes only found by either amplification or RT labeling, respectively. The numbers of genes found up- or down-regulated with either one or both methods are given in the lower right corners of the plots.

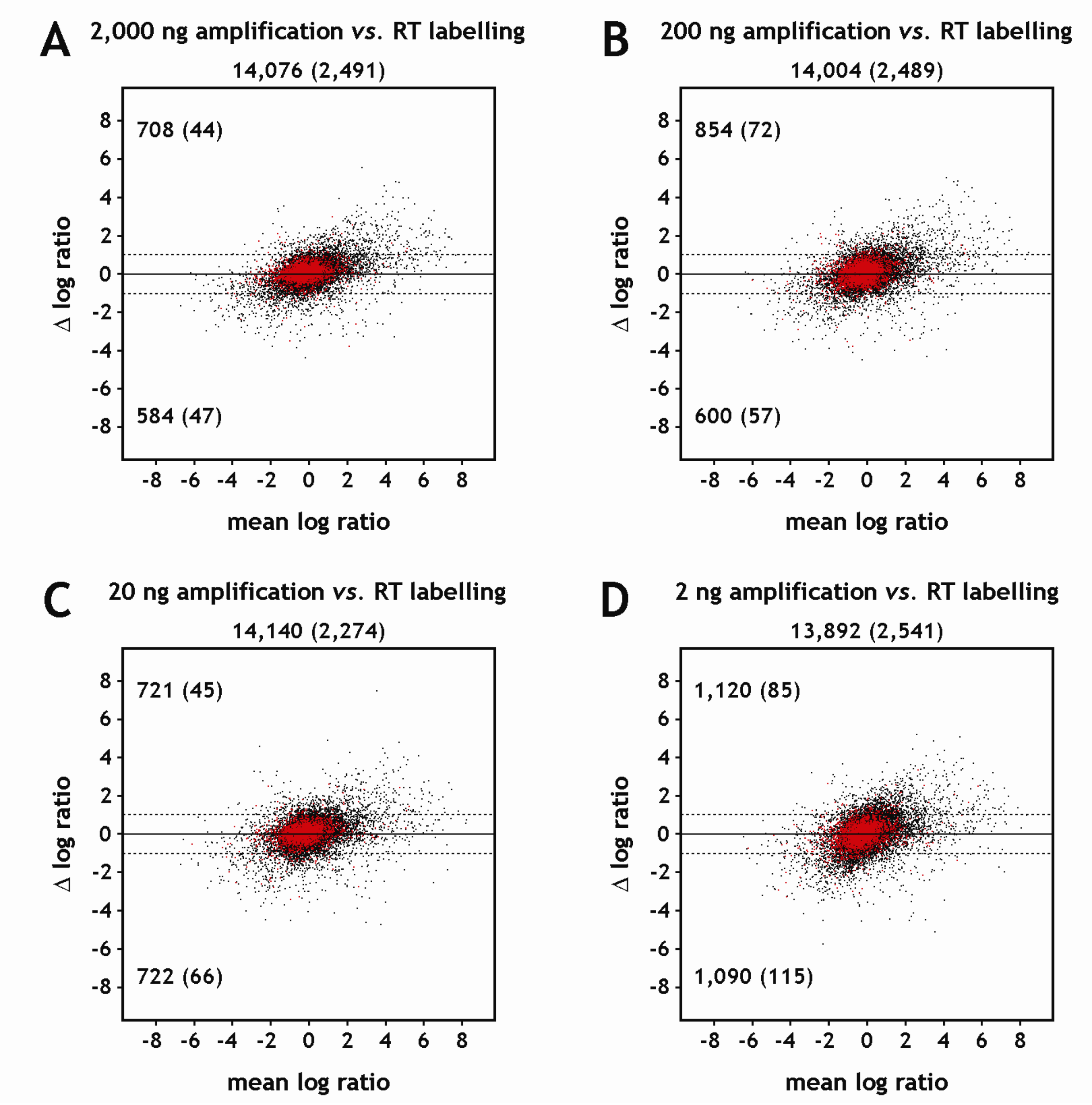


Figure 25. Mean difference (*MA*) plots displaying the difference of $\log_2$ ratios against the mean of $\log_2$ ratios. *M* is a measure for the difference of $\log_2$ ratios observed between amplified and unamplified targets, prepared from breast and universal human reference RNA ($\log_2$ [breast / reference]$_{\text{TAcKLE}}$ - $\log_2$ [breast / reference]$_{\text{RT}}$). *A* is a measure for the average differential expression (½ [$\log_2$ [breast / reference]$_{\text{TAcKLE}}$ + $\log_2$ [breast / reference]$_{\text{RT}}$]). Ratios of targets amplified from (**A**) 2,000 ng, (**B**) 200 ng, (**C**) 20 ng and (**D**) 2 ng starting material were compared to ratios of unamplified targets. Replicated experiments were averaged before calculating the differences and means of $\log_2$ ratios. Black dots correspond to probes detected on at least one array of each considered target preparation approach, probes shown as red dots additionally reached fluorescence intensities at least two standard deviations above local background. The respective quantities are specified underneath the panel headings, values for red dots given in parentheses. Values in the upper and lower left corners of each plot indicate genes that show at least a twofold change of expression ratios to either direction, as illustrated by horizontal dashed lines.

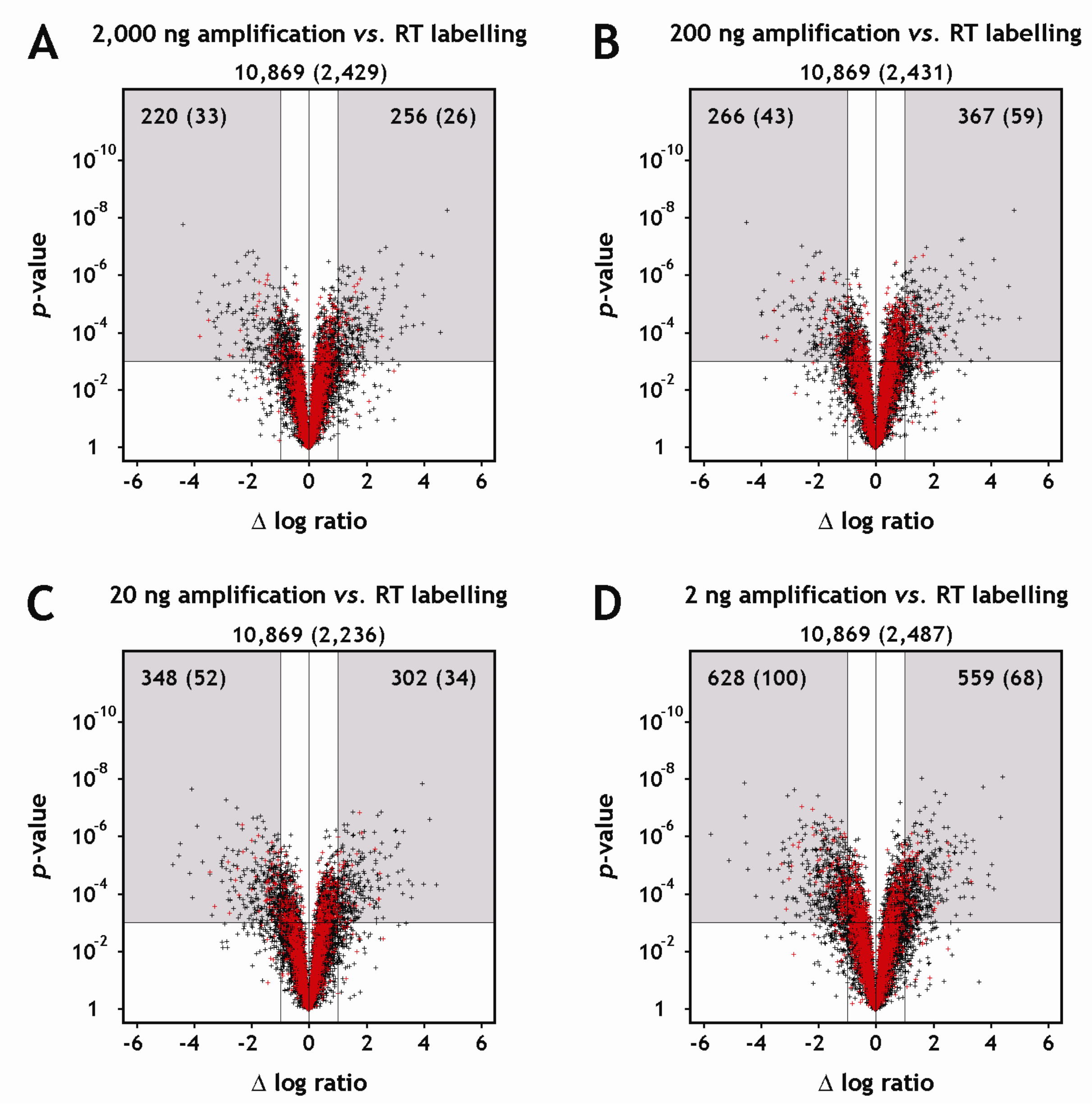


Figure 26. Volcano plots of *p*-values against the difference of $\log_2$-transformed expression ratios. The difference of $\log_2$ ratios observed between amplified and unamplified targets ($\log_2$ [breast / reference]$_{\mathrm{TAcKLE}}$ – $\log_2$ [breast / reference]$_{\mathrm{RT}}$) is shown on the x-axis. The corresponding *p*-value of significance, derived by linear modeling, is shown on the y-axis. Ratios of targets amplified from (**A**) 2,000 ng, (**B**) 200 ng, (**C**) 20 ng and (**D**) 2 ng starting material were compared to ratios of unamplified targets. Black dots correspond to probes detected on all or all but one arrays of all target preparation approaches, red dots indicate probes which additionally reached fluorescence intensities at least two standard deviations above local background on the arrays under consideration. The associated numbers of genes are given underneath the panel headings, values for red dots printed in parentheses. The plots were segmented to illustrate the relation of statistical significance ($p < 0.001$) to significance based on a twofold change criterion. Only genes indicated by spots in the upper left and right segments of the plots satisfy both criteria, their numbers explicitly shown. Genes located in the lower left and right segments display a large fold-change difference between amplified and unamplified targets but fail to achieve statistical significance. Genes found in the middle segments show no relevant difference of expression ratios, with (upper segments) or without (lower segments) additional statistical significance associated with this observation.

3.2 Signal Amplification by Rolling Circle Replication

3.2.1 Motivation

Proper amplification procedures are required to analyze limited source material by microarray technology. The amplification can either be performed on the material itself (target amplification) or on the signal it generates on the array (signal amplification). For target amplification, T7-based protocols had most commonly been used. However, spotted oligonucleotide microarrays contain sense-strand probes, so traditional T7 amplification schemes producing anti-sense RNA are not appropriate when combined with conventional reverse transcription labeling methods (see 1.6). As shown in 3.1, target amplification via the TAcKLE protocol is a good solution to this problem.

In terms of signal amplification, both tyramide signal amplification (TSA) and three-dimensional multi-labeled structures (DNA dendrimers, 3DNA) had previously been used [126,127], but the results of these studies showed little promise. Furthermore, it had been shown that rolling circle replication could be adapted to an array-based format (RCA), and the methodology had successfully been used to analyze minimal amounts of cytokines on protein microarrays [158]. It had also been demonstrated that spotted oligonucleotides could serve as primers for on-chip RCA [157]. What was missing, although it had been suggested [157], was the implementation of the method for microarray-based expression profiling. There were, however, several both practical and theoretical considerations in favor of adapting on-chip RCA for a use in expression profiling. As shown above for the TAcKLE protocol, target amplification by *in vitro* transcription can yield excellent results. It is, however, both time-consuming, error-prone (*e.g.*, for RNA degradation) and expensive due to the numerous enzymes and reagents necessary to complete the protocol. In addition, off-chip target amplification strategies can lead to sequence-dependent quantitation bias, because different transcripts with different biophysical properties have to be processed. In RCA, identical circles of DNA hybridize to short DNA primers (complementary to a portion of the circle) attached to the target molecules, and Φ29 polymerase synthesizes single-stranded concatameric DNA molecules composed of thousands of tandemly repeated copies of the circle. A more detailed explanation has already been given in the introduction. Since RCA, unlike other amplification procedures, produces single amplified products that remain linked to the DNA primers, it is well suited to solid phase formats such as microarrays for generating localized signals at specific array locations.

For these reasons, it was decided to evaluate the applicability of on-chip RCA for microarray expression analysis as a possible alternative to the TAcKLE protocol.

3.2.2 Preparatory Experiments

To choose a substrate with properties favorable for on-chip RCA, several commercially available microarray slides with different surface coatings (A: Slide E, epoxysilane-coated, Schott Nexterion, Jena; B: Slide AL, coated with aldehyde-derivatized aminosilane, Schott Nexterion; C: GAPSII, aminosilane-coated, Corning Life Sciences, Acton, USA; D: CodeLink, coated with a three-dimensional amine-reactive polymer, Amersham Biosciences) were selected and used for spotting of the RCA primer Inv1216Arl (Table 10). This oligonucleotide allows for a simplified experimental setup, in which circularized PL-1216g can directly be hybridized to DNA immobilized on the array surface without the need for prior generation and hybridization of compatible cDNA. Both RCA and detection were performed as described in 2.7.4, but without any oligonucleotides from the 6121g collection. These experiments were performed in the Research Group on Molecular Medicine (Department for Genetics and Pathology, Rudbeck Laboratory, University Uppsala, Sweden) in collaboration with Jonas Jarvius, MD, MSc.

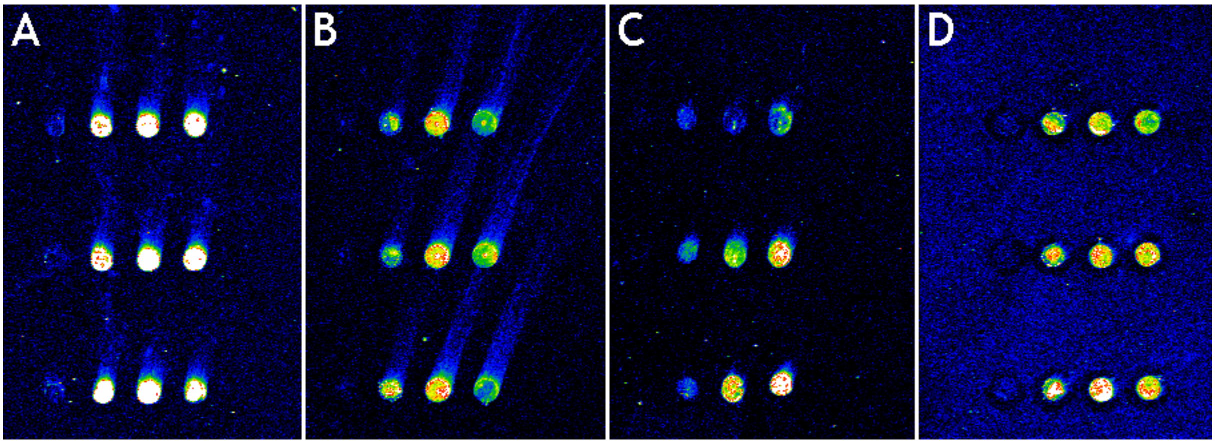


Figure 27. Evaluation of RCA performance with regard to the microarray surface coating. The oligonucleotide Inv1216Arl (Table 11), containing 20 nucleotides complementary to a portion of PL-1216g (Table 11) as well as an oligo(dTdC) sequence as spacer, was spotted on substrates coated with either (**A**) epoxysilane, (**B**) aldehyde-derivatized aminosilane, (**C**) aminosilane or (**D**) a three-dimensional long-chain, hydrophilic polymer containing amine-reactive groups. RCA and detection were performed as described in 2.7.4. All arrays were scanned with the same PMT voltage (700 V) and are displayed with identical settings for brightness and contrast.

The strongest signals were clearly obtained with epoxysilane-coated slides (Fig. 27A). Both the aldehyde coating (Fig. 27B) and the three-dimensional polymer (Fig. 27D) yielded approximately comparable and second strongest signal intensities, but with increased background fluorescence detected on the 3D-coated substrates. The lowest intensities were obtained on aminosilane-coated substrates (Fig. 27C). Accordingly, Slide E epoxysilane-coated substrates were chosen for all consecutive experiments.

3.2.3 The Challenge Caused by Long Amplification Products

All of the subsequent experiments were conceptuated and carried out in collaboration with cand. biol. Daniel Haag.

Depending on the assay conditions [141,145,146], Φ29 polymerase possesses an exponentially decaying polymerization rate of initially 1000 - 3180 nt/min and a half life of 11 h. This means that Φ29-catalyzed RCA yields approximately 600 - 2000-fold amplification of a circular 104-mer within 60 min, corresponding to 60 - 190 kb or 30 - 95 µm of concatameric ssDNA.

However, only 70 nucleotides of each cDNA carrying an attached amplification product are available for base-pairing to complementary array probes. Furthermore, these probes are contained in spots with an average diameter of 60 µm, which is in the range of the length of the amplification product.

It is therefore of fundamental importance to keep the forces that act on the hybridized cDNA as week as possible. Otherwise, the amplification products might become stretched and dispersed or even displaced from the spots. Drying the microarrays by centrifugation is obviously not compatible with this situation (Fig. 28A).

To approach this problem, the centrifugation step was omitted and replaced by the automatic draining functionality of the GeneTAC hybridization station (Fig. 28B). This procedure yielded clearly improved results, but a tendency towards stretching and possibly even dislocation remained.

We therefore devised a procedure that covalently links the hybridized cDNA molecules to their complementary array probes. To accomplish this, the cDNA is initially tagged with a primary amine by addition of aminoallyl-dUTP to the reverse transcription reaction. Then, amine reactive NHS-psoralen (Succinimidyl-[4-(psoralen-8-yloxy)]-butyrate; SPB) is coupled to the aminoallyl residues of the cDNA (Fig. 12), forming stable amide bonds.

Upon hybridization, the psoralen tricyclic planar ring system intercalates into the DNA duplex formed by the cDNA molecules and their complementary oligonucleotide probes on the array. Photoreactive coupling of psoralen to thymine residues is achieved by brief exposure to long UV light (366 nm), creating covalent inter-strand cross-links (Fig. 13) that firmly attach the cDNA to the array via their complementary oligonucleotides that are covalently linked to the array's surface coating.

This approach yielded a vast improvement in array performance with virtually no dispersion of the amplification products (Fig. 28C). As the basic protocols now seemed to function properly, the next goal was to implement a second oligonucleotide system to allow for competitive hybridization of two differently tagged cDNA populations, system-specific RCA and subsequent detection by distinguishable fluorophores.

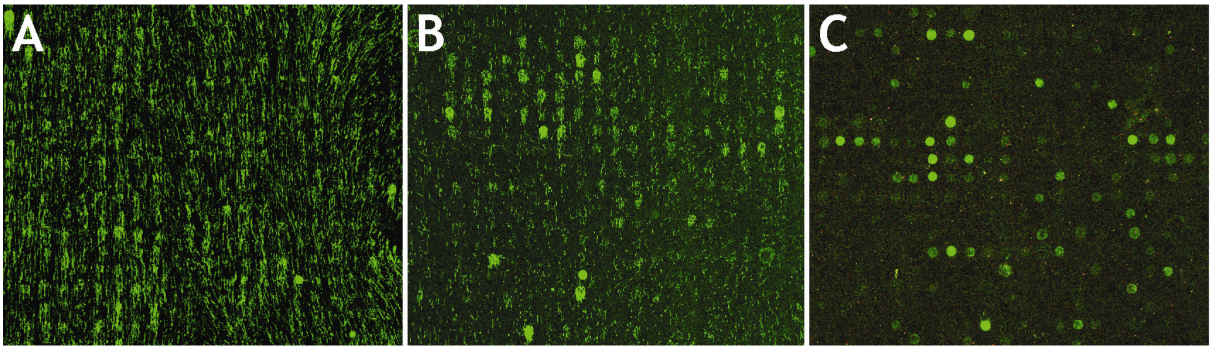


Figure 28. Successive improvements of the RCA protocol. (**A**) Section of a microarray after hybridization with tagged HL-60 cDNA and subsequent RCA using the system 1216g. The array was dried by centrifugation, as described in 2.5.3. A large portion of the concatemeric RCA product, visible as green threads, was stretched out by centripetal forces and detached from the site of amplification. (**B**) Section of a microarray hybridized with tagged HL-60 cDNA, RCA-amplified using the system 1216g and dried by automated draining. Compared to slides dried by centrifugation, considerably less RCA product was dispersed from the array spots. (**C**) Section of a microarray containing cDNA cross-linked to its oligonucleotide probes. HL-60 cDNA containing the 1216g-tag as well as aminoallyl residues was reacted with NHS-psoralen as described in 2.7.4 and hybridized to the array. Inter-strand DNA cross-linking was achieved by 15 min irradiation with UV-A light (366 nm), as described in 2.7.5. Finally, RCA was performed using the system 1216g, and the array was dried by automated draining.

3.2.4 Two-Color Hybridizations

To obtain a second set of oligonucleotides for the various steps of the RCA protocol (reverse transcription, oligonucleotide circularization, replication and detection), all sequences from the set 1216g were modified by exchanging every base for its complementary base, creating the set 6121g (Table 10). In this way, both sets were as different as possible regarding the respective sequences but still had similar biophysical properties, *i.e.*, same length, same GC content *etc*. The oligonucleotides used for the detection of 6121g RCA product were labeled with Cy5 instead of Cy3. Now, both 1216g- and 6121g-tagged cDNA was prepared independently and applied on the same array for on-chip RCA (Fig. 29).

Unfortunately, not only yellow spots of variable intensities were detected, as one would expect for a two-color co-hybridization of cDNAs prepared from the same RNA. There was also a considerable number of spots that were primarily if not exclusively green or red. Possible reasons for this are discussed in 4.2. Further improvements can be expected from a stepwise refinement of the protocols, and some conceivable points are mentioned in the discussion part of this thesis. The current procedure for on-chip RCA is summarized in Figure 30.

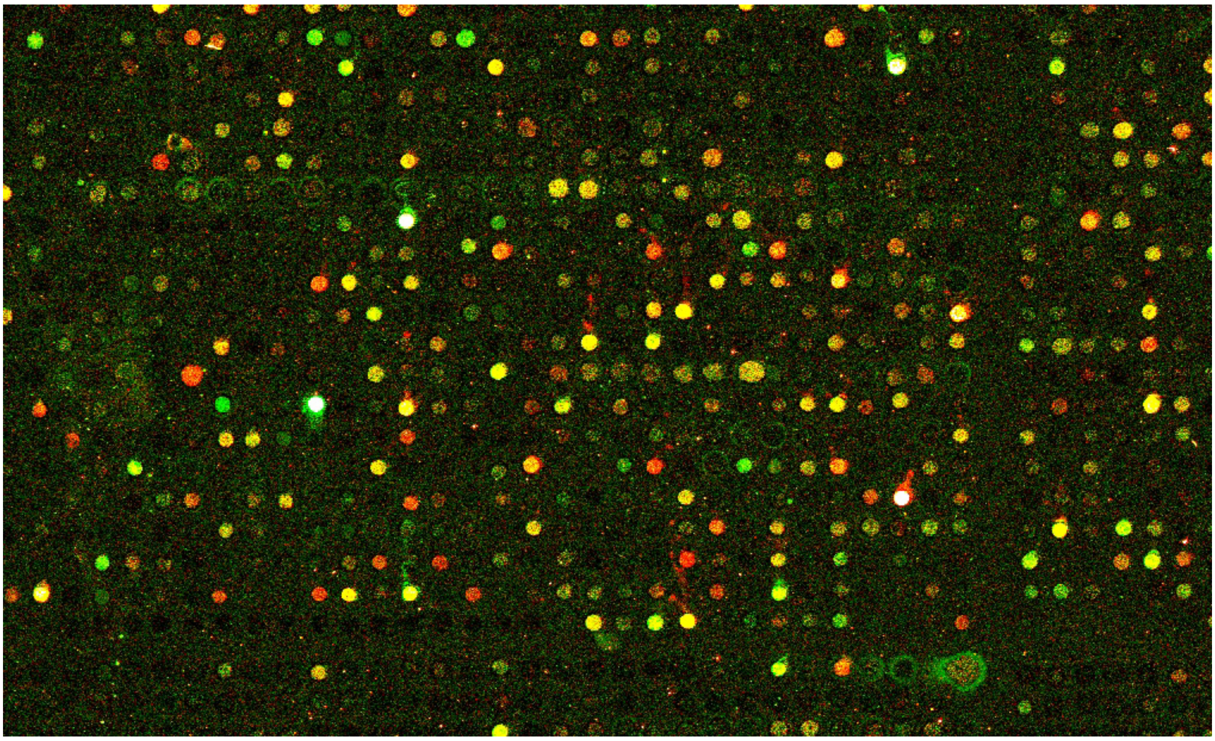


Figure 29. Section of a microarray containing both 1216g-tagged and 6121g-tagged cDNA. Following hybridization and cross-linking, on-chip RCA products were detected by dye-labeled detection probes, using Cy3 for 1216g and Cy5 for 6121g.

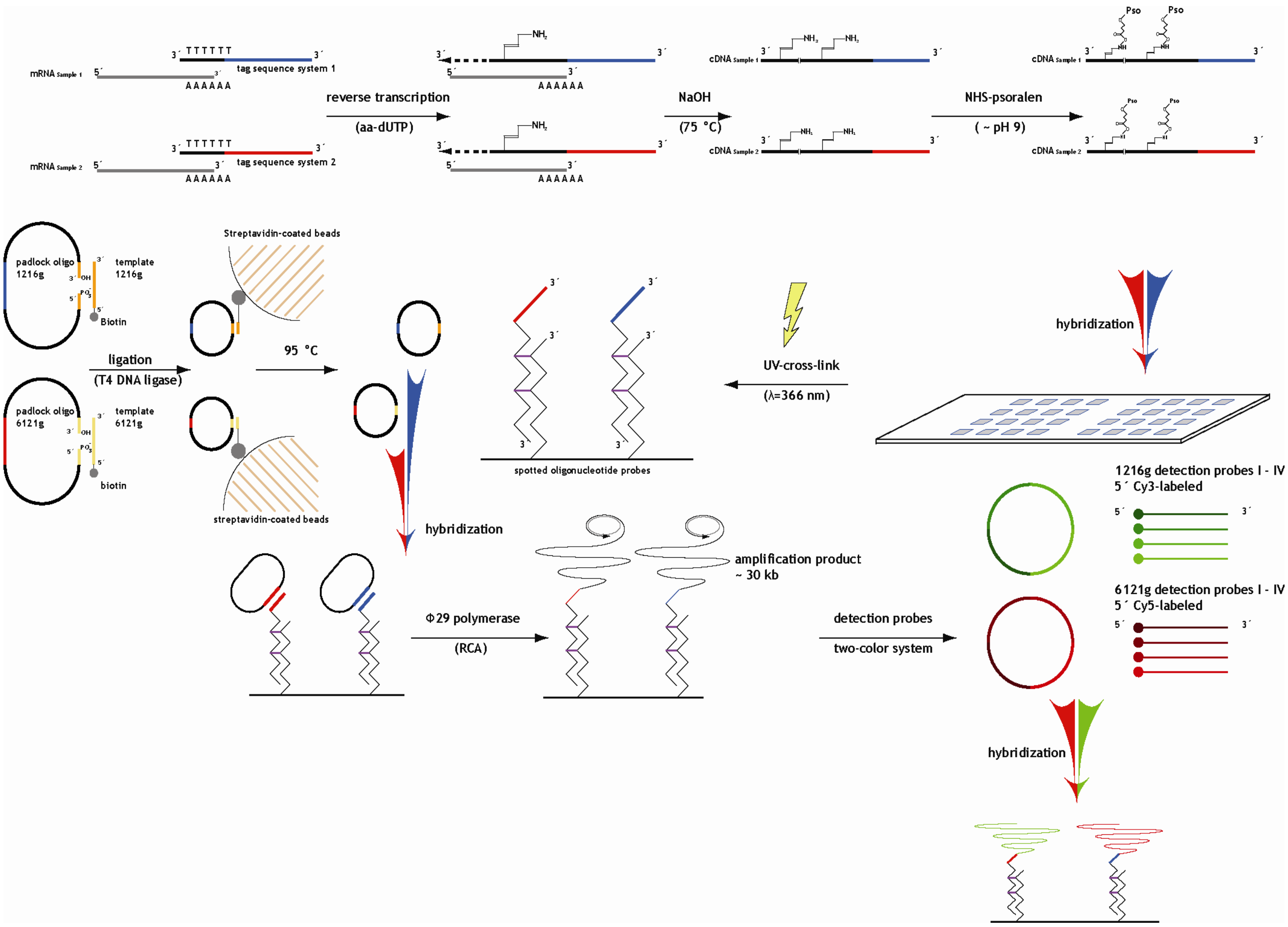


Figure 30. On-chip RCA for microarray expression analysis. To allow for a larger illustration and, thus, achieve better readability, the figure was rotated 90° counterclockwise. Details of the successive protocol steps are given in the text. Modified from a figure kindly provided by cand. biol. Daniel Haag.

3.3 Cross-Platform Reproducibility of Oligonucleotide Microarray Expression Profiles

One who asks a question is a fool for five minutes; one who does not ask a question remains a fool forever.

CHINESE PROVERB

3.3.1 Motivation

In the scope of this thesis, the protocol for on-chip RCA for microarray expression analysis could unfortunately not be sufficiently advanced to allow for a reasonable analysis of biological samples. Target amplification via the TAcKLE protocol, however, yielded excellent results, at least in terms of reproducibility and comparability of expression profiles to those generated on the same array platform with unamplified targets (see 3.1). Since it was planned to use this new method for a large-scale expression profiling study of primary breast cancer samples, it was of outstanding interest to find out about the comparability of expression profiles generated using the TAcKLE protocol and our platform of spotted 70-mer oligonucleotide arrays to profiles from other platforms.

Today, researchers can choose from a broad variety of methods for global transcriptional profiling. Among the different technical approaches, microarray technology has gained a premier position. In principle, microarrays can be produced either by robotic printing (“spotting”) of DNA on a chemically modified glass surface [89], or by *in situ* synthesis of oligonucleotides via custom phosporamidite chemistry using either photolithography on a silane-reacted quartz substrate [92] or ink-jet technology on a hydrophobic glass support [93].

Spotted arrays usually contain cDNA-specific PCR amplicons (cDNA arrays), ranging from several hundred to a few thousand basepairs in size. Generally, no more than one amplicon is used to probe a given gene. Although they are technically challenging and require both optimized protocols [165] and workflow [193], cDNA arrays are typically produced by individual research groups or core facilities. Alternatively, they can be purchased from several commercial suppliers. But after the discovery of frequent discrepancies in the annotation of cDNA clones [194], investigators began to realize potential drawbacks of this highly advocated technology. *In situ* synthesis of oligonucleotide probes requires sophisticated equipment for photolithography and solid phase chemistry, which is usually too complex and elaborate for an academic environment. A widespread commercial implementation of this technology is the Affymetrix GeneChip platform [92], which currently uses 11-16 pairs (11 for the arrays used in this study) of perfect-match and single-base-mismatch 25-mer oligonucleotides for each gene. Recently, large collections of longer oligonucleotides (50-80 bases), produced by established suppliers using conventional phosphoramidite chemistry, have become increasingly popular as probes for spotted DNA arrays. Technical advantages of oligonucleotide arrays include a constant DNA concentration across all spots and biophysically optimized sequences, reducing secondary structures, avoiding repetitive sequence motives and providing a fixed range for both T_m and length. All this accounts for more uniform, stable and predictable hybridization conditions. The overall costs for long oligonucleotide arrays will often be lower when labor and other costs associated with cDNA libraries, such as replication, amplification or sequence verification, are regarded.

Considering this diversity of approaches and the resulting technical differences, researchers are highly interested in the general accuracy and reliability of microarray data and the cross-platform comparability. Several independent methods like Northern blotting or real-time quantitative reverse transcription PCR (RQ-PCR) have been used to validate microarray results for a small number of transcripts. Generally, there was good agreement between the corresponding values, affirming the ability to accurately profile gene expression with array-based approaches.

Former studies also compared global expression measurements between cDNA arrays and short oligonucleotide arrays [195,196] or SSH [197]. Recently, Barczak *et al.* [173] compared results between spotted arrays of 70-mer oligonucleotides and *in situ* synthesized Affymetrix GeneChip arrays. Using RNA of a cell line and a commercial reference RNA, they found strong correlations of the corresponding data sets. Despite these studies clarifying some fundamental questions, there still remains considerable uncertainty regarding the comparability of data from clinical specimens.

This lack of understanding constitutes a barrier, which keeps researchers from an immense amount of potentially valuable information (via efficient integration of microarray data generated on different array platforms). In this thesis, a comparison with a small set of HNSCC tumor samples from clinical practice evaluates the cross-platform reproducibility between spotted 70-mer oligonucleotide arrays and the well-established commercial Affymetrix GeneChip platform in a practical setting.

3.3.2 Experimental Design

To assess the degree of concordance between expression profiles obtained with either spotted oligonucleotide microarrays made from a large collection of 70-mer probes or commercial arrays produced by *in situ* synthesis of sets of multiple 25-mer oligonucleotides per gene, we analyzed relative gene expression in a set of six human head and neck squamous cell carcinoma (HNSCC) samples *versus* either healthy control mucosa (n = 4) or lymph node metastases (n = 2) of the respective patients as the reference (Table 12). For the spotted 70-mer arrays, relative expression levels were calculated by averaging the normalized $\log_2$-ratios of two replicate two-color hybridizations per patient, one performed with inverse assignment of fluorophores (dye swap). This procedure was used to eliminate dye-related signal correlation bias [198,199]. For the commercial 25-mer arrays, relative expression levels were derived by subtracting normalized $\log_2$-transformed probe-level data (fluorescence intensities) of two single-color hybridizations per patient, corresponding to the respective tumor and reference tissue.

Table 12. Patient and disease characteristics.

Patient	Primary Site	Age	Sex	pT	pN	pM	Grading	Samples Analyzed[a]
160	hypopharynx	48	M	3	1	0	2	PT / N
171	hypopharynx	58	M	3	2a	0	2	PT / M
173	oropharynx	56	M	3	2	0	2	PT / N
180	hypopharynx	57	M	2	3	0	2	PT / N
186	hypopharynx	47	F	2	2	0	2	PT / N
205	oropharynx	49	M	3	1	0	2	PT / M

[a] All cases were diagnosed histopathologically as HNSCC and staged according to the TNM classification of malignant tumors. The indicated tissues were used for gene expression profiling. N: normal mucosa, PT: primary HNSCC, M: lymph node metastasis.

3.3.3 Probe Matching

In this study, the gene expression profiles of 12 specimens obtained from six head and neck cancer patients were analyzed (Table 12). Four primary HNSCC were assayed *versus* corresponding healthy mucosa and another two primary HNSCC *versus* corresponding lymph node metastases of the respective patients.

This analysis was performed both on *in situ*-synthesized Affymetrix HG-U133A arrays, containing 22,283 sets of 25-mer probes, and on spotted long oligonucleotide arrays containing 26,791 70-mer probes of the Operon Human Genome Oligo Set Version 2.1 and Version 2.1 Upgrade. A total of 9,867 UniGene clusters were found for the probe sets of the HG-U133A arrays, while 13,604 were retrieved for the Operon arrays, using GenBank accession numbers provided by the manufacturers. 4,425 genes were represented on both platforms, as identified by consistent assignment of UniGene clusters to the corresponding probes or probe sets (Fig. 31). This large set of genes was used as a basis for comparing expression data from the two array systems.

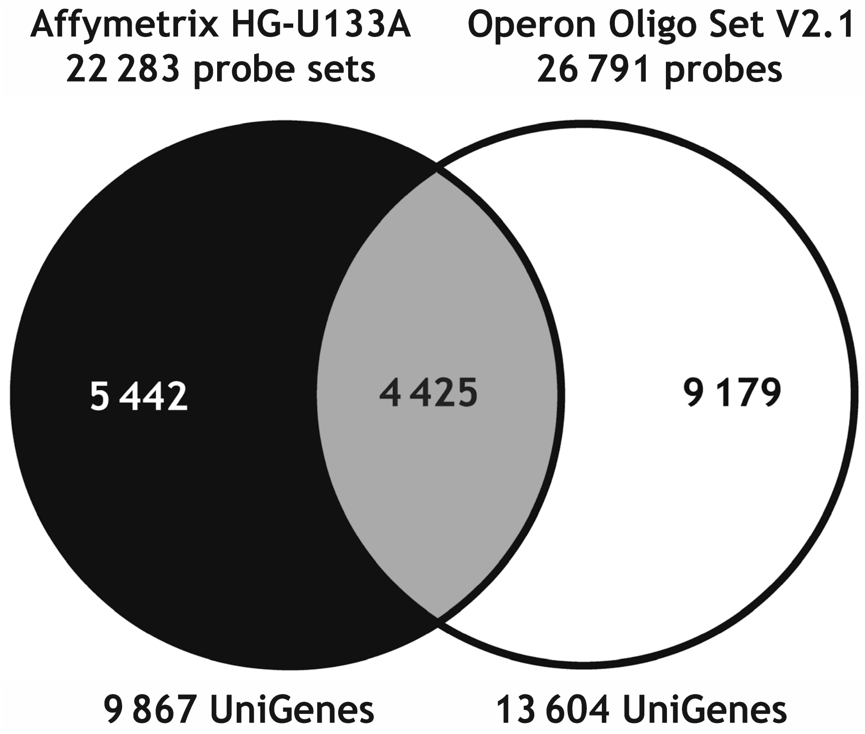


Figure 31. Intersection of probes or probe sets from the different array platforms. Probe sequences were mapped to UniGene clusters (build #175), based on GenBank accession numbers provided by the manufacturers. 9,867 UniGene clusters were found for the probe sets of the HG-U133A arrays, while 13,604 were retrieved for the Operon arrays. A total of 4,425 genes were represented on both array types.

3.3.4 Intra-Platform Reproducibility of Expression Ratios

For the platform of spotted long oligonucleotide arrays, correlations of expression ratios measured on individual arrays were r = 0.99 for identically repeated hybridizations and r = -0.98 for dye swap hybridizations repeated with inverse assignment of fluorophores (Fig. 21). Similar correlations had been reported for the Affymetrix system [164]. Hence, both array platforms provide highly reproducible measurements of gene expression profiles, which is an essential pre-requisite for the success of a cross-platform comparison.

3.3.5 Cross-Platform Reproducibility of Expression Ratios

Normalized $\log_2$-transformed absolute signal intensities were calculated for the arrays from both platforms using variance stabilization by *vsn* [169]. For GeneChip arrays, $\log_2$ expression ratios were obtained by subtracting $\log_2$-transformed absolute signal intensities of the two respective arrays from each patient. For the spotted long oligonucleotide array, $\log_2$-ratios from two-color dye swap hybridizations were inverted and averaged. To ensure that the observed effects were not due to characteristics of the data processing algorithm, the analyses of GeneChip arrays were repeated using background correction and normalization by *gcRMA* [171] as well as the *MAS5* algorithm [170]. For all patients, there was a clear correlation between differential expression measurements made with either array type (r = 0.56 - 0.76), and the correlation improved substantially (r = 0.61 - 0.85) when measurements from probes with low intensity signals were excluded (Fig. 32). Except for patients 160 and 186, the respective regression lines all showed a slope clearly smaller than 1, indicating that, on average, absolute log ratios obtained on the Operon long oligonucleotide platform were lower than the corresponding values measured with Affymetrix arrays. The changes in correlation were marginal when *gcRMA* was used to normalize the GeneChip results. *MAS5* yielded lower correlations with unfiltered data, but the results were similar to those of *vsn* or *gcRMA* when filtered data were used (Table 13).

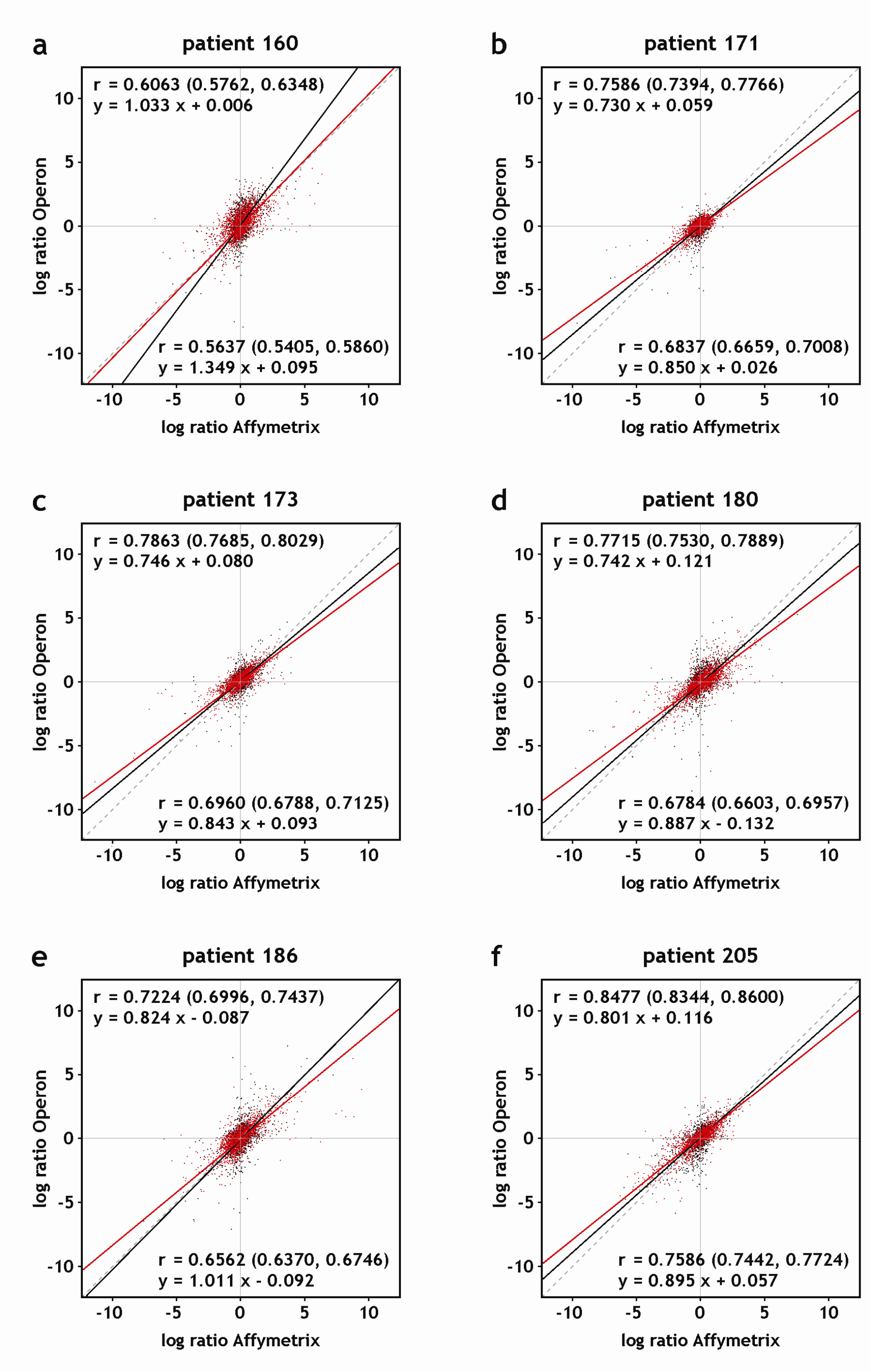


Figure 32. Scatter plots comparing normalized, $\log_2$-transformed expression ratios of spotted long oligonucleotide arrays to ratios obtained with Affymetrix GeneChip short oligonucleotide arrays. For the spotted arrays, normalized ratio data from dye swap experiments were combined. For Affymetrix arrays, the ratios of normalized intensity values from corresponding arrays were used. Hybridized targets were derived from (**a**) patient 160 (**b**) patient 171 (**c**) patient 173 (**d**) patient 180 (**e**) patient 186 and (**f**) patient 205. Orthogonal regression analysis was performed to derive the regression lines shown in black (unfiltered data) and red (filtered data) as well as their respective linear equations (shown in the lower part of the plots for unfiltered data and in the upper part for filtered data). Pearson correlation coefficients and their associated 95% confidence intervals are listed as well. Dashed lines through origin with slope 1 are displayed to accentuate the reduced slope. For panels (**a**) - (**f**), calculations were based on 3,472, 3,595, 3,600, 3,569, 3,522 and 3,474 data points for the unfiltered data sets as well as 1,796, 2,011, 1,889, 1,954, 1,816 and 1,866 data points for the filtered data sets.

Table 13. Correlation of gene expression ratios obtained with either Affymetrix GeneChip arrays or Operon long oligonucleotide arrays.

Patient	Affy vsn *vs.* Operon vsn[a]	Affy MAS5 *vs.* Operon vsn[b]	Affy gcRMA *vs.* Operon vsn[c]
160	0.564 (0.606)	0.456 (0.571)	0.544 (0.585)
171	0.684 (0.759)	0.453 (0.707)	0.671 (0.749)
173	0.696 (0.786)	0.553 (0.783)	0.710 (0.795)
180	0.678 (0.772)	0.577 (0.763)	0.681 (0.777)
186	0.656 (0.722)	0.561 (0.712)	0.664 (0.734)
205	0.759 (0.848)	0.636 (0.837)	0.762 (0.841)

[a]GeneChip results were normalized by variance stabilization (*vsn*), [b]*MAS5* and [c]*gcRMA*. Values obtained upon removal of low-intensity-signals are given in parentheses.

3.3.6 Systematic Bias Correction by 'Distance Weighted Discrimination (DWD)'

As the samples were processed at different institutions (the 70-mer arrays were processed at the DKFZ, whereas all Affymetrix experiments were performed at the Georg-Speyer-Haus in Frankfurt a.M.) and assayed using different array platforms and protocols, considerable systematic biases were expected to be manifested in the data sets as differences in gene expression patterns. In order to identify and adjust systematic biases imposed by characteristics of the different array platforms, we used the method of 'Distance Weighted Discrimination (DWD)' [178]. Following this procedure, there was a clear improvement in the correlations of relative expression measurements (Table 14). As before, correlations obtained after normalization by *MAS5* were lower than the respective values generated with *vsn* or *gcRMA*, unless low-intensity signals were excluded from the analyses. The respective orthogonal regression lines showed little if any change in slope when the data from both platforms had been normalized by the *vsn* algorithm. Moderate changes were detected upon DWD in case *gcRMA* had been used to normalize Affymetrix data, whereas normalization by MAS5 tended to cause more severe variation. On average, the slopes were closest to 1 when *vsn* was used to normalize the Affymetrix data and deviated the most from 1 upon normalization by *MAS5*. As expected, systematic bias correction by DWD shifted the slopes towards one in almost all cases (Table 14).

Table 14. Correlation of gene expression ratios before and after systematic bias correction by DWD.

Patient	Affy vsn *vs.* Operon vsn[a]	Affy MAS5 *vs.* Operon vsn[b]	Affy gcRMA *vs.* Operon vsn[c]
160	r: 0.550 (0.704) [0.735]	r: 0.431 (0.577) [0.693]	r: 0.525 (0.678) [0.703]
	s: 1.343 (1.289) [1.126]	s: 0.755 (0.899) [1.280]	s: 1.137 (1.237) [1.083]
173	r: 0.686 (0.848) [0.893]	r: 0.544 (0.701) [0.886]	r: 0.707 (0.850) [0.896]
	s: 0.865 (0.906) [0.867]	s: 0.539 (0.677) [0.911]	s: 0.745 (0.853) [0.811]
180	r: 0.697 (0.785) [0.816]	r: 0.577 (0.680) [0.806]	r: 0.696 (0.778) [0.821]
	s: 0.852 (0.878) [0.828]	s: 0.644 (0.747) [0.940]	s: 0.767 (0.838) [0.805]
186	r: 0.667 (0.771) [0.830]	r: 0.568 (0.680) [0.819]	r: 0.674 (0.767) [0.824]
	s: 0.937 (0.956) [0.874]	s: 0.647 (0.744) [0.895]	s: 0.739 (0.809) [0.726]

[a]GeneChip results were normalized by variance stabilization (*vsn*), [b]*MAS5* and [c]*gcRMA*. Values obtained upon bias correction by DWD are given in parentheses. Values in brackets were derived after additionally removing low-intensity-signals. r: Pearson correlation. s: slope of the respective orthogonal regression lines.

3.3.7 Significant Differences and Similarities

The comparison of the sets of genes identified as differentially expressed was used as a further approach to detect differences between the two array systems. In Figure 34, volcano plots [174] show the $\log_2$-ratios of those 2,861 genes consistently detected in the 4 primary HNSCC *versus* normal mucosa experiments and their respective *p*-values. The two platforms identified similar numbers of differentially expressed genes (≥ 2-fold difference), both regarding raw *p*-values ($p \leq 0.001$) or FDR-adjusted *p*-values (adj. $p \leq 0.1$) [177]. There were 45 genes identified as differentially expressed in all tumor samples on the Affymetrix platform, 53 were scored on Operon arrays, and the intersection contained 21 genes discovered on both systems (Table 15, Fig. 33a). Plotting of the corresponding mean log ratios (Fig. 33b) revealed that even genes scored by only one of the systems generally showed the same direction, but not the same degree of differential expression on the other. GO data mining [180] for ‘biological process’ (at level 3) assigned the majority of annotated genes from each platform to cell growth and/or maintenance as well as various metabolic pathways (Fig. 35). However, using the software EASE [181], which performs a statistical analysis of the GO categories assigned to the differentially expressed genes, accounting for the distribution of GO categories in the list of all analyzed genes to find those categories that are the most overrepresented (and can therefore be described as 'themes'), revealed a trend towards components of the extracellular matrix for both of the platforms. Furthermore, genes involved in lipid metabolism were significantly overrepresented only among the differentially expressed genes identified on the Affymetrix system, whereas the Operon platform additionally detected genes engaged in ion binding (Table 16).

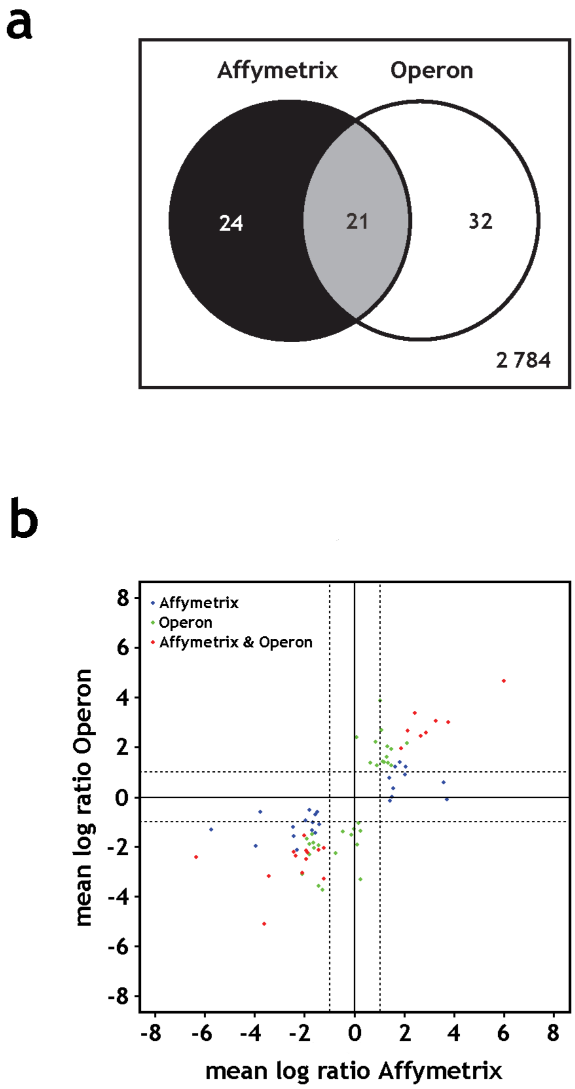


Figure 33. Summary of genes scored as differentially expressed with either one or both evaluated platforms. (**a**) Venn diagram showing subsets of genes that exhibit a significant differential expression with either technology, taken from a pool that contained only those genes which could repeatedly be quantified in all Operon hybridizations. (**b**) log-log-plot illustrating the relationship of the log-ratios for the 77 genes shown in (**a**).

Table 15. Genes scored as differentially expressed with either one or both evaluated platforms.

UniGene	Gene Name	Gene Symbol	OMIM	$\log_2$ Affy[a]	*p*-value Affy[b]	$\log_2$ Operon[a]	*p*-value Operon[b]	Δ exp Affy[c]	Δ exp Operon[c]
Hs.136348	osteoblast specific factor 2 (fasciclin I-like)	*OSF-2*		6.01	0.07	4.67	0.01	+	+
Hs.443625	collagen, type III, alpha 1	*COL3A1*	120180	1.05	0.19	3.87	0.01	-	+
Hs.75823	ALL1-fused gene from chromosome 1q	*AF1Q*	604684	3.78	0.10	2.99	0.04	+	+
Hs.28792	inhibin, beta A (activin A, activin AB alpha polypeptide)	*INHBA*	147290	3.72	0.05	-0.11	0.85	+	-
Hs.232115	collagen, type I, alpha 2	*COL1A2*	120160	3.60	0.06	0.58	0.39	+	-
Hs.528321	collagen, type V, alpha 1	*COL5A1*	120215	2.44	0.09	3.38	0.01	+	+
Hs.437173	collagen, type IV, alpha 1	*COL4A1*	120130	3.28	0.06	3.05	0.04	+	+
Hs.409602	sulfatase 1	*SULF1*		2.88	0.10	2.59	0.03	+	+
Hs.372679	Fc fragment of IgG, low affinity IIIa, receptor for (CD16)	*FCGR3A*	146740	1.10	0.40	2.67	0.07	-	+
Hs.434488	chondroitin sulfate proteoglycan 2 (versican)	*CSPG2*	118661	2.66	0.08	2.43	0.04	+	+
Hs.821	biglycan	*BGN*	301870	2.15	0.09	2.65	0.07	+	+
Hs.83354	lysyl oxidase-like 2	*LOXL2*	606663	0.10	0.86	2.39	0.05	-	+
Hs.435795	insulin-like growth factor binding protein 7	*IGFBP7*	602867	0.87	0.23	2.20	0.07	-	+
Hs.118893	Melanoma associated gene	*D2S448*	600134	2.11	0.23	2.15	0.07	-	+
Hs.408096	fragile X mental retardation, autosomal homolog 1	*FXR1*	600819	2.06	0.07	1.21	0.14	+	-
Hs.102308	potassium inwardly-rectifying channel, subfamily J, member 8	*KCNJ8*	600935	1.33	0.13	2.04	0.08	-	+
Hs.15099	Rho-related BTB domain containing 1	*RHOBTB1*	607351	2.03	0.06	0.88	0.17	+	-
Hs.122645	laminin, beta 1	*LAMB1*	150240	1.89	0.07	1.96	0.02	+	+
Hs.81988	disabled homolog 2, mitogen-responsive phosphoprotein (*Drosophila*)	*DAB2*	601236	1.48	0.26	1.93	0.07	-	+
Hs.235935	nephroblastoma overexpressed gene	*NOV*	164958	1.82	0.10	1.40	0.13	+	-
Hs.85195	myeloid leukemia factor 1	*MLF1*	601402	1.64	0.10	1.20	0.16	+	-
Hs.246875	DRE1 protein	*DRE1*		1.31	0.23	1.61	0.07	-	+
Hs.436708	Kruppel-like factor 7 (ubiquitous)	*KLF7*	604865	1.55	0.10	0.35	0.49	+	-
Hs.278469	taste receptor, type 2, member 14	*TAS2R14*		1.52	0.10	0.01	1.00	+	-
Hs.7753	calumenin	*CALU*	603420	1.48	0.15	1.27	0.07	-	+
Hs.528298	Sec23 homolog A (*S. cerevisiae*)	*SEC23A*		1.43	0.10	-0.16	0.78	+	-
Hs.433452	HEG homolog	*HEG*		1.42	0.07	0.77	0.23	+	-
Hs.179657	plasminogen activator, urokinase receptor	*PLAUR*	173391	1.15	0.23	1.41	0.08	-	+
Hs.16530	chemokine (C-C motif) ligand 18 (pulmonary and activation-regulated)	*CCL18*	603757	1.20	0.28	1.40	0.08	-	+
Hs.370774	ankyrin repeat and BTB (POZ) domain containing 2	*ABTB2*		0.65	0.36	1.38	0.07	-	+
Hs.312419	origin recognition complex, subunit 3-like (yeast)	*ORC3L*	604972	1.36	0.11	1.37	0.08	-	+
Hs.462693	zinc finger protein 22 (KOX 15)	*ZNF22*	194529	0.90	0.35	1.27	0.09	-	+
Hs.130958	ribonuclease/angiogenin inhibitor	*RNH*		0.18	0.78	-1.06	0.10	-	+
Hs.1321	coagulation factor XII (Hageman factor)	*F12*	234000	-0.02	0.98	-1.28	0.07	-	+
Hs.434933	regulator of G-protein signalling 12	*RGS12*	602512	0.26	0.67	-1.37	0.08	-	+
Hs.112028	misshapen/NIK-related kinase	*MINK*		-0.46	0.51	-1.39	0.08	-	+
Hs.5215	integrin beta 4 binding protein	*ITGB4BP*	602912	-1.41	0.10	-1.11	0.18	+	-
Hs.211556	ELOVL family member 6, elongation of long chain fatty acids	*ELOVL6*		-1.48	0.08	-0.60	0.32	+	-
Hs.40968	heparan sulfate (glucosamine) 3-O-sulfotransferase 1	*HS3ST1*	603244	-0.11	0.85	-1.53	0.08	-	+
Hs.132853	enthoprotin	*ENTH*	607265	-1.56	0.10	-0.70	0.55	+	-
Hs.528666	RAR-related orphan receptor A	*RORA*	600825	-1.56	0.10	-1.45	0.14	+	-
Hs.15519	oxysterol binding protein-like 2	*OSBPL2*	606731	-1.66	0.06	-1.03	0.17	+	-
Hs.91139	solute carrier family 1, member 1	*SLC1A1*	133550	-1.68	0.08	-1.34	0.14	+	-
Hs.378738	AHNAK nucleoprotein (desmoyokin)	*AHNAK*		-1.71	0.20	-1.51	0.09	-	+
Hs.393239	sterol-C4-methyl oxidase-like	*SC4MOL*	607545	-1.81	0.10	-0.52	0.61	+	-
Hs.77870	hypothetical protein FLJ12750	*FLJ12750*		-1.66	0.17	-1.83	0.08	-	+
Hs.212787	Microtubule associated serine/threonine kinase family member 4	*MAST4*		-1.80	0.12	-1.89	0.07	-	+
Hs.105435	GDP-mannose 4,6-dehydratase	*GMDS*	602884	-1.91	0.13	-1.68	0.09	-	+
Hs.82237	tripartite motif-containing 29	*TRIM29*		0.13	0.83	-1.93	0.05	-	+
Hs.166311	SAM and SH3 domain containing 1	*SASH1*	607955	-1.42	0.11	-1.96	0.07	-	+
Hs.1588	4-aminobutyrate aminotransferase	*ABAT*		-1.97	0.08	-0.94	0.18	+	-
Hs.424551	integral type I protein	*P24B*		-2.02	0.08	-1.56	0.08	+	+
Hs.434243	KIBRA protein	*KIBRA*		-1.21	0.10	-2.04	0.06	+	+
Hs.90797	O-acyltransferase (membrane bound) domain containing 2	*OACT2*		-1.61	0.16	-2.06	0.08	-	+
Hs.437043	KIAA0540 protein	*KIAA0540*		-1.43	0.10	-2.13	0.06	+	+
Hs.446429	prostaglandin D2 synthase 21kDa (brain)	*PTGDS*	176803	-1.93	0.10	-2.15	0.05	+	+
Hs.5541	ATPase, Ca++ transporting, ubiquitous	*ATP2A3*	601929	-1.89	0.09	-2.23	0.02	+	+
Hs.356726	scinderin	*SCIN*		-0.76	0.32	-2.26	0.02	-	+
Hs.118747	solute carrier family 15 (H+/peptide transporter), member 2	*SLC15A2*	602339	-2.30	0.07	-2.13	0.11	+	-
Hs.430324	annexin A9	*ANXA9*	603319	-1.80	0.34	-2.30	0.07	-	+
Hs.206501	hypothetical protein from clone 643	*LOC57228*		-2.35	0.07	-2.36	0.02	+	+
Hs.169238	fucosyltransferase 3 (galactoside 3(4)-L-fucosyltransferase)	*FUT3*	111100	-2.43	0.08	-1.59	0.10	+	-
Hs.348350	dehydrogenase/reductase (SDR family) member 1	*DHRS1*		-2.43	0.10	-2.20	0.03	+	+
Hs.257697	programmed cell death 4 (neoplastic transformation inhibitor)	*PDCD4*		-2.45	0.06	-1.21	0.27	+	-
Hs.282975	carboxylesterase 2 (intestine, liver)	*CES2*	605278	-1.92	0.10	-2.49	0.02	+	+
Hs.31130	transmembrane 7 superfamily member 2	*TM7SF2*	603414	-2.08	0.07	-3.05	0.01	+	+
Hs.436657	clusterin (complement lysis inhibitor, apolipoprotein J)	*CLU*	185430	-2.10	0.28	-3.11	0.04	-	+
Hs.167218	BarH-like homeobox 2	*BARX2*	604823	-1.23	0.09	-3.28	0.01	+	+
Hs.134478	RecQ protein-like 5	*RECQL5*	603781	0.26	0.65	-3.31	0.04	-	+
Hs.439309	transmembrane protease, serine 2	*TMPRSS2*	602060	-3.44	0.06	-3.19	0.02	+	+
Hs.298023	aquaporin 5	*AQP5*	600442	-1.44	0.36	-3.58	0.06	-	+
Hs.272813	dual oxidase 1	*DUOX1*	606758	-1.28	0.28	-3.72	0.08	-	+
Hs.103944	mucin 7, salivary	*MUC7*	158375	-3.77	0.10	-0.62	0.32	+	-
Hs.116651	epithelial V-like antigen 1	*EVA1*	604873	-3.96	0.07	-1.97	0.22	+	-
Hs.438862	EPS8-like 1	*EPS8L1*		-3.63	0.07	-5.11	0.02	+	+
Hs.226391	anterior gradient 2 homolog (*Xenopus laevis*)	*AGR2*	606358	-5.75	0.05	-1.33	0.14	+	-
Hs.13775	homeodomain-only protein	*HOP*	607275	-6.35	0.06	-2.42	0.07	+	+

Differentially expressed genes were selected from a subset (n = 2 861) consistently detected in the 4 primary HNSCC *versus* normal mucosa experiments. [a]$\log_2$-transformed expression ratios were averaged within each platform, [b]corresponding *p*-values of significance derived by empirical Bayes inference and subsequent adjustment to control the FDR. [c]To be scored as differentially expressed (Δ exp), genes had to satisfy both statistical significance ($p \leq 0.1$) and significance based on a twofold change criterion.

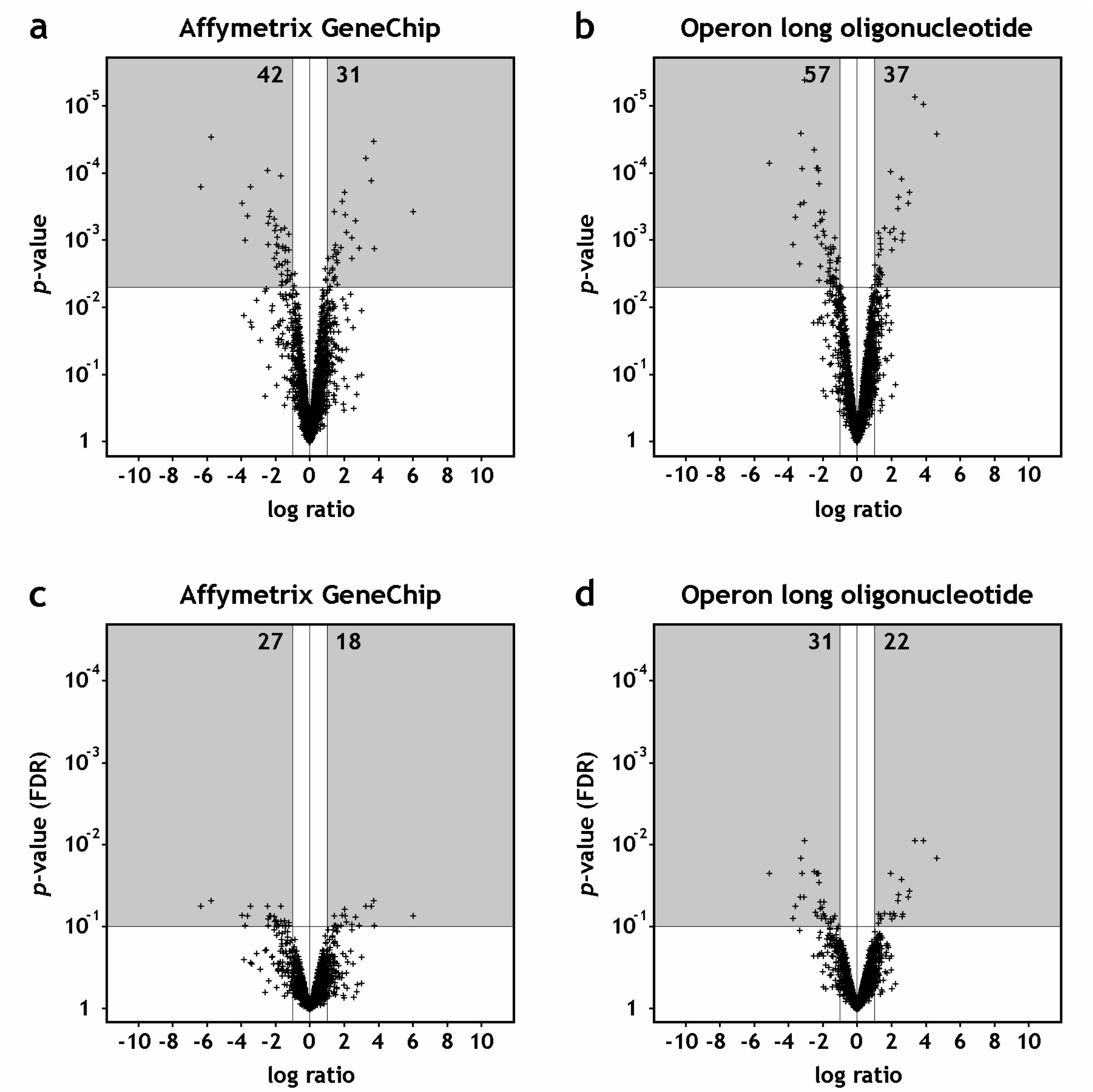


Figure 34. Volcano plots of *p*-values against $\log_2$-transformed expression ratios. Mean $\log_2$ ratios of tumor *versus* reference samples are shown on the x-axis. The corresponding *p*-values of significance, derived by empirical Bayes inference (**a**-**b**) or empirical Bayes inference and subsequent adjustment to control the FDR (**c**-**d**), are displayed on the y-axis. Results are shown for those 2,861 genes consistently detected in the 4 primary HNSCC *versus* normal mucosa experiments. The plots were segmented to illustrate the relation of statistical significance ($p \leq 0.005$, adj. $p \leq 0.1$) to significance based on a twofold change criterion. Only genes indicated by spots in the upper left and right segments of the plots satisfy both criteria, their numbers explicitly shown. Genes located in the lower left and right segments display a large fold-change but fail to achieve statistical significance. Genes found in the middle segments show no relevant difference of expression, with (upper segments) or without (lower segments) additional statistical significance associated with this observation.

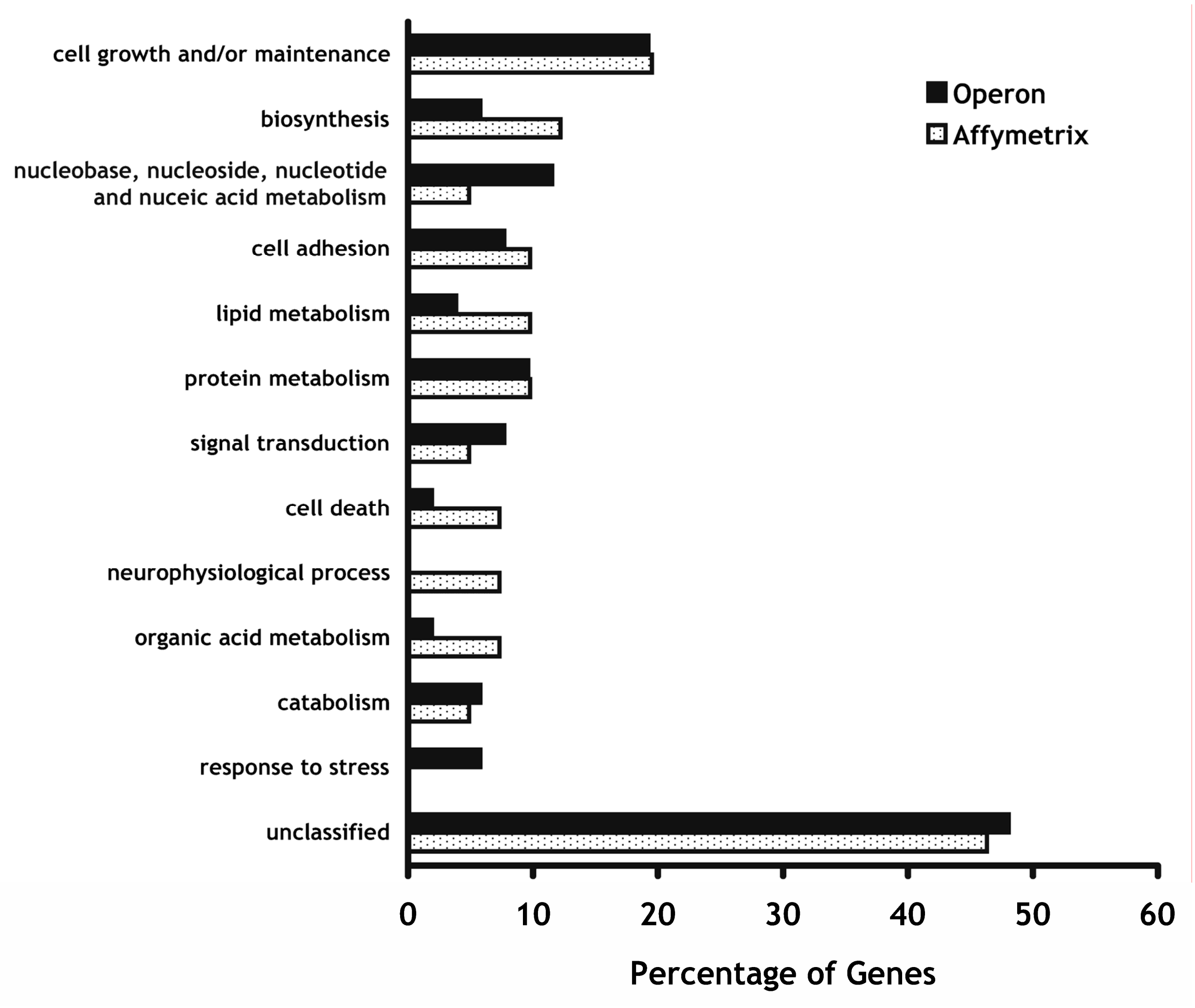


Figure 35. GO data mining. The 45 regulated genes detected with the Affymetrix system as well as the 52 regulated genes found with Operon arrays were characterized according to their biological process classification in the GO database (at level 3). Roughly half of the genes did not have a GO classification at this level. The majority of the remaining genes were involved with cell growth and/or maintenance as well as various metabolic pathways.

Table 16. EASE overrepresentation analysis of the genes listed in Table 15.

System[a]	Gene Category[b]	List Hits[c]	List Total[d]	Population Hits[e]	Population Total[f]	EASE Score[g]	LH in PH (%)[h]	LH in Δ exp (%)[i]
***Affymetrix*[j]**								
GO Cellular Component	extracellular matrix	6	34	55	2348	0.00084	10.91	14.63
GO Molecular Function	extracellular matrix structural constituent	4	37	14	2410	0.0010	28.57	9.76
GO Biological Process	cell adhesion	6	37	99	2386	0.015	6.06	14.63
GO Biological Process	lipid metabolism	6	37	112	2386	0.025	5.36	14.63
GO Biological Process	lipid biosynthesis	4	37	44	2386	0.027	9.09	9.76
GO Biological Process	cellular lipid metabolism	5	37	78	2386	0.028	6.41	12.20
GO Cellular Component	endoplasmic reticulum	5	34	99	2348	0.048	5.05	12.20
GO Cellular Component	extracellular matrix (sensu Metazoa)	3	34	26	2348	0.050	11.54	7.32
GO Biological Process	fatty acid metabolism	3	37	26	2386	0.057	11.54	7.32
GO Biological Process	steroid metabolism	3	37	27	2386	0.061	11.11	7.32
GO Molecular Function	extracellular matrix structural constituent c. t. s.[k]	2	37	5	2410	0.073	40.00	4.88
GO Cellular Component	extracellular region	4	34	75	2348	0.086	5.33	9.76
GO Cellular Component	collagen	2	34	7	2348	0.094	28.57	4.88
***Operon*[j]**								
GO Cellular Component	extracellular matrix	7	45	55	2348	0.00045	12.73	13.46
GO Molecular Function	extracellular matrix structural constituent	4	48	14	2410	0.0022	28.57	7.69
GO Molecular Function	cation binding	10	48	192	2410	0.010	5.21	19.23
GO Molecular Function	metal ion binding	10	48	225	2410	0.027	4.44	19.23
GO Molecular Function	ion binding	10	48	225	2410	0.027	4.44	19.23
GO Biological Process	cell adhesion	6	44	99	2386	0.031	6.06	11.54
GO Molecular Function	calcium ion binding	6	48	104	2410	0.049	5.77	11.54
GO Cellular Component	extracellular region	5	45	75	2348	0.050	6.67	9.62
GO Molecular Function	scavenger receptor activity	2	48	3	2410	0.057	66.67	3.85
GO Cellular Component	extracellular	5	45	79	2348	0.058	6.33	9.62
GO Cellular Component	extracellular matrix (sensu Metazoa)	3	45	26	2348	0.084	11.54	5.77
GO Biological Process	organismal physiological process	8	44	214	2386	0.085	3.74	15.38
GO Cellular Component	Golgi apparatus	5	45	93	2348	0.094	5.38	9.62
GO Molecular Function	extracellular matrix structural constituent c. t. s.[k]	2	48	5	2410	0.094	40.00	3.85

[a]System: the system of categorizing genes, in this case the GO classification type. [b]Gene Category: the specific category of genes within the classification system, in this case the GO category of the superordinate GO classification type (different levels are possible). [c]List Hits (LH): number of genes in the list of differentially expressed genes that belong to the respective GO category. [d]List Total: number of differentially expressed genes that could be annotated within the respective GO classification system. [e]Population Hits (PH): number of genes in the list of all analyzed genes (n = 2,861) belonging to the respective GO category. [f]Population Total: number of analyzed genes with annotation data in the respective GO classification system. [g]EASE Score: The upper bound of the distribution of Jackknife Fisher exact probabilities given the List Hits, List Total, Population Hits and Population Total. Categories with the lowest EASE score are significantly overrepresented in the list of differentially expressed genes. [h]LH in PH: percentage of differentially expressed genes belonging to the respective category in the group of all analyzed genes in this category. [i]LH in Δ exp: percentage of differentially expressed genes belonging to the respective category in the group of all differentially expressed genes. [j]Differentially expressed genes from each platform were analyzed separately. [k]c.t.s.: conferring tensile strength.

3.3.8 RQ-PCR Analysis

For a small subset of genes, the differential expression measurements were verified by RQ-PCR analysis (Fig. 36). There was good qualitative agreement between the values determined by either GeneChip arrays, Operon arrays or RQ-PCR. All platforms showed the same direction of regulated gene expression. However, the magnitude of differential expression differed considerably depending on both the experimental approach and the algorithm applied for normalization. Firstly, GeneChip intensity measurements were transformed by variance stabilization (*vsn*), which was also used for the long oligonucleotide arrays and derives an approximately constant variance along the complete intensity range[169]. Normalization was additionally accomplished employing the *MAS5* algorithm from the current version of the Affymetrix Microarray Suite software package[170]. At least for the small number of genes and patients shown here, there is a tendency for higher ratios with *vsn* normalization.

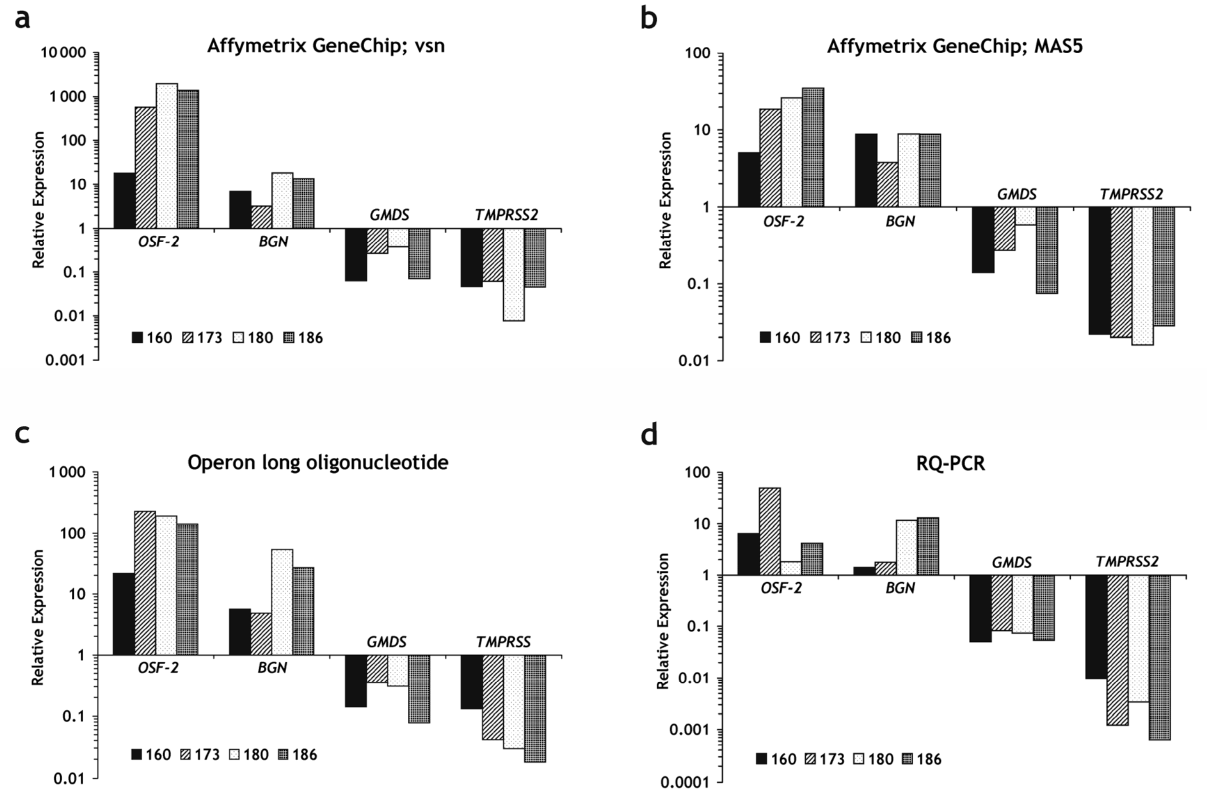


Figure 36. Comparison of relative gene expression for the genes *OSF2*, *GMDS*, *TMPRSS2* and *BGN*. Expression ratios were determined for tumor *versus* control tissue of the indicated patients, using either Affymetrix GeneChip arrays (**a-b**), Operon long oligonucleotide arrays (**c**) or real-time quantitative PCR analysis (**d**). Affymetrix ratios were either normalized by variance stabilization (**a**) or the *MAS5* algorithm (**b**).

4 Discussion

There is no statement so absurd that no philosopher will make it.

CICERO

The objective of this thesis was to develop protocols that allow for the analysis of gene expression in minimal samples by means of spotted long oligonucleotide microarrays. Two different approaches were taken, one that amplifies the target material before hybridization (TAcKLE protocol) and another that amplifies the signal generated on the array (on-chip RCA). As the TAcKLE protocol performed particularly well, it was subsequently applied to evaluate the utility of spotted oligonucleotide microarrays compared to an accepted commercial reference platform.

4.1 The TAcKLE Protocol

RNA amplification by *in vitro* transcription yields up to 10^5-fold linear amplification of high quality aRNA starting from nanogram quantities of total RNA[164] and has been applied for microarray studies of differential gene expression for several years. In this thesis, a newly developed protocol broadens the utility of this approach to the application with spotted oligonucleotide microarrays and, thus, expands the utilization of these microarrays to the analysis of rare cell populations (Fig. 37). These could be derived by fine-needle aspiration or microdissection of clinical specimens, by cell sorting or micromanipulation of single cells. Utilizing elements of the approved Eberwine procedure[111,112], the TAcKLE protocol can easily be implemented, and even aRNA, produced for other applications, can be made accessible for oligonucleotide arrays by adding another reverse transcription and labeling step.

The amplification itself does not increase the overall variability above that encountered during cDNA synthesis. This is clearly demonstrated by co-hybridization of material independently amplified from the same source. The reproducibility of a single round and even two rounds of amplification, estimated by the correlation coefficient, is comparable or even superior to that obtained with unamplified targets and possibly more biased by the variability of the chip hybridization and readout procedure than by the enzymatic manipulations.

A further level of amplification is added by the strong strand displacement activity of Klenow fragment, combined with random priming of DNA polymerisation [200,201], which adds a further level of amplification and, thereby, decreases the amount of RNA necessary for labeling. The amplification was estimated to be about 5-fold by spectrophotometrically measuring the amount of cDNA subsequent to the labeling reaction. This effect facilitates the conduction of additional experiments even with marginal amounts of starting material. This value seems reasonable since as little as 1 µg Klenow-labeled material (500 ng still work fine) can be used for hybridization, whereas protocols using labeled aRNA or RT-labeled cDNA require as much as 3-6 µg. Additionally, Klenow fragment is known to have a superior efficiency with modified nucleotides compared to any known reverse transcriptase.

The generated data demonstrate that the ability to reproducibly identify differentially expressed genes after amplification is retained compared to conventional labeling by reverse transcription. This is true even when using as little starting material as 2 ng total RNA.

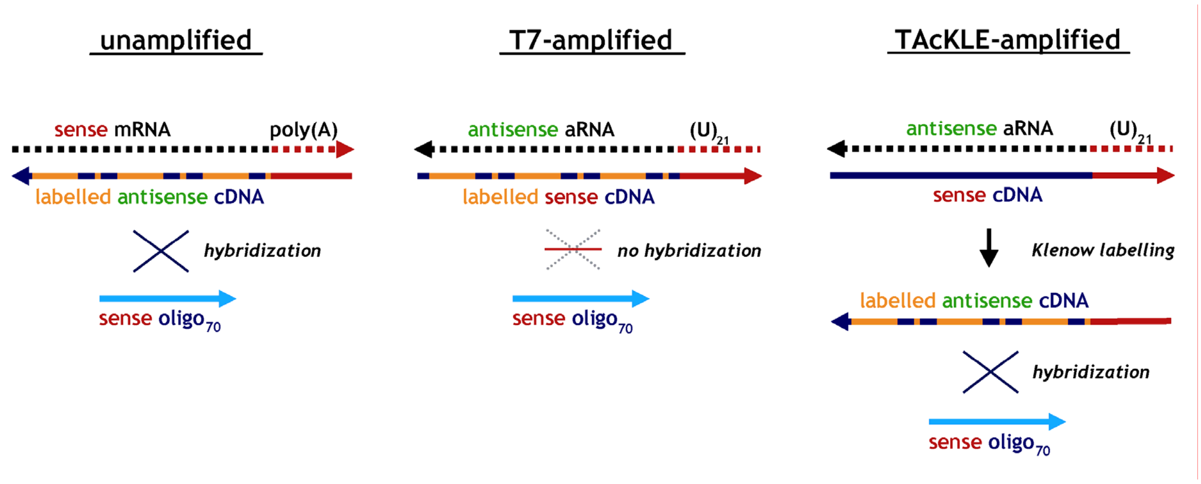


Figure 37. Special demand on target amplification protocols for oligonucleotide microarrays containing sense-oriented probe molecules. Fluorescent cDNA targets prepared by reverse transcription labeling of mRNA or rather the mRNA content of total RNA are antisense-oriented and thus compatible for hybridization to sense-oriented arrays. However, antisense RNA (aRNA) generated by T7 amplification procedures cannot be used in this way, as the resulting cDNA targets would have sense orientation and hence could not basepair with the sense-oriented array probes. Instead, a modification of the original procedure must be used, in which aRNA labeling is achieved by reverse transcription and subsequent dye incorporation using Klenow fragment.

Some minor differences between transcription profiles generated from 2000 ng and 2 ng of total RNA can still be detected, probably due to additional bias introduced by a second round of amplification, which includes a randomly primed RT reaction. But even after two rounds of amplification, reproducibility is sufficiently high for reliable quantification of differences between samples. Furthermore, and equally important, there is no compression of differences between RNA samples with either one or two rounds of amplification. In contrast, there is a systematic and reproducible expansion of expression ratios in amplified targets. A possible explanation can be differences in RT efficiency, depending on the template concentration.

The presented analyses also indicate that reverse transcription labeling represents a significant source of variation between identical RNA samples and reaffirm the need for dye swap replicates. A part of the deviating ratios detected when comparing amplified and unamplified targets can probably be attributed rather to the inaccuracy of reverse transcription labeling than to systematic bias or random errors of the amplification procedure.

A different approach to overcome the problem of strand orientation is the addition of fluorescent nucleotide derivatives to the *in vitro* transcription reaction. Barczak *et al.*[173] reported decreased signal intensities of fluorescent aRNA targets, compared to cDNA prepared by reverse transcription labeling. This could be confirmed in initial experiments that were performed with this method (data not shown). Apparently, RNA polymerase is not a favorable enzyme for the incorporation of dye-labeled nucleotides. As it clearly discriminates bulky nucleotide modifications, ratios of labeled to unlabeled nucleotides have to be optimized. It has been reported that the addition of DMSO during *in vitro* transcription can improve incorporation rates[122], and that utilization of aminoallyl-dUTP, followed by chemical coupling of reactive dye derivatives, may overcome some of the problems connected to the bulky nature of dye-labeled nucleotides. Still, there is no additional amplification by the labeling procedure. Smith et al.[123] claimed that their method termed ‘Single Primer Amplification’ would generate both sense-oriented and antisense-oriented fluorescent cDNA targets, the latter via a ‘strand switch’ during the Klenow labeling reaction. This effect, however, did not generate sufficient amounts of the antisense-oriented cDNA required for the spotted oligonucleotide arrays used in this thesis. A recent study by Rajeevan *et al.*[202] exploits the template-switching effect[188] of Moloney murine leukemia virus (MMLV) reverse transcriptase to incorporate an RNA polymerase promoter sequence upstream of the generated cDNA, producing sense-oriented RNA (sRNA) by subsequent *in vitro* transcription. In a similar approach, the method of terminal continuation has been used to generate amplified transcripts with either sense or antisense orientation[203]. However these reports had not yet been published at the time when the experiments described in 3.1.2 were conceived and performed. Similar to the report by Rajeevan *et al.*[202], it was tried to use the template switching effect of MMLV to incorporate a T7-promoter upstream of the cDNA sequence (inverse IVT, Fig. 15). However, the amount of sense-oriented RNA that could be generated was insufficient. One can only speculate about the reasons, which may be handling errors or, more likely, unfavourable conditions during second-strand cDNA synthesis that did not allow for the generation of a functional, double-stranded T7-promoter. The template switching itself was probably successful, since the TS-SPA method (Fig. 16), which is identical to inverse IVT in terms of first-strand cDNA synthesis, produced large amounts of sense-oriented second-strand cDNA via thermal cycling with *Taq* DNA polymerase and T7-ts primer. This requires a functional priming site in the first-strand cDNA, which can only be generated by template switching. It was impossible, however, to use TS-SPA for the generation of reasonable expression profiles, since differentially expressed genes could hardly be detected (Fig. 17B). So far, it has been impossible to explain this observation.

Commercial solutions utilize novel signal amplification and/or detection procedures, as in the QIAGEN HiLight Platform (http://www1.qiagen.com/Products/MicroArrayAnalysis/MicroArrayAnalysisSystems.aspx), which uses resonance light scattering (RLS), a technology based on the optical light scattering properties of nano-sized metal colloidal particles[204]. The system requires 1-2 µg total RNA and generates biotinylated and/or fluorescein-labeled target cDNA, which can be hybridized to commercial or custom made arrays. Gold particles, coated with anti-biotin antibodies, and/or silver particles, coated with anti-fluorescein antibodies, are used to stain the targets after hybridization. Detection is performed on a specialized reader. The SensiChip System developed by QIAGEN and Zeptosens AG (http://www.zeptosens.com) uses planar waveguide (PWG) technology[205,206] and requires a minimum of 1 µg total RNA. Hybridizations are carried out on 70mer oligonucleotide arrays of a special format using the SensiChip HybStation.

The SenseAmp RNA amplification kit offered by Genisphere, which was introduced after the completion of this study, generates amplified RNA in sense orientation (senseRNA). This is achieved by poly(dT)-tailing of first strand cDNA and subsequent hybridization of a modified oligo(dA), which carries the T7 promoter sequence attached to its 5’-end and a 3’-ddA to block second strand synthesis. In this way, nucleotides can only be attached to the 3’-end of the first strand cDNA, which creates a double-stranded T7 promoter sequence upstream of the coding sequence. Although no double-stranded cDNA is created, the double-stranded promoter sequence is sufficient to initiate *in vitro* transcription, which yields multiple copies of sense-oriented RNA molecules, which can optionally be tailed with poly(A). A RQ-PCR-based comparison of 192 transcripts before and after amplification yielded strong correlations [121]. For a microarray-based analysis, however, the amplified RNA would have to be labeled by reverse transcription, primed either by oligo(dT) (in case of poly(A)-tailing) or by random hexamers, accepting dye-related incorporation bias that the TAcKLE protocol avoids by means of Klenow labeling.

The Ovation RNA amplification system from NuGEN uses a combination of a proprietary primer, RNase H and a DNA polymerase. First strand cDNA is produced using a DNA/RNA chimeric primer. The DNA portion of the primer contains an oligo(dT)-sequence, whereas the RNA portion, attached at the 5’-end, contains a unique sequence that is used to incorporate a priming site for the subsequent linear amplification step. Second strand cDNA is produced by a combination of RNase H, a DNA polymerase and a DNA ligase, similar to the original T7 protocol. The amplification is initiated by the action of RNase H, which also degrades the RNA portion of the chimeric primer. A new primer molecule hybridizes to its released complementary sequence in the second strand cDNA. DNA polymerase binds to the annealed 3'-end of the primer and initiates primer extension. Due to its strand displacement activity, the forward anti-sense strand of the cDNA duplex is displaced from the template strand as cDNA elongation takes place. Simultaneously, RNase H degrades the RNA of the primer that is being elongated, exposing the binding site for another primer molecule and starting another round of primer extension. This cyclic process continues in a linear fashion until sufficient amounts of antisense cDNA have been obtained. For microarray applications, the cDNA can be labeled by addition of aminoally-dUTP to the amplification reaction, which allows for subsequent labeling with fluorescent dyes. Technical reports published on NuGEN’s website (http://www.nugentechnologies.com) look promising, but independent validations of the method have to be awaited before reliable predictions on its performance can be made.

In conclusion, it was shown that TAcKLE can faithfully amplify and label as little as 2 ng of total RNA, an amount which can be obtained from a few hundred cells. It represents a robust method for the sensitive detection of expression profiles, which is particularly suited for the use with microarrays consisting of long sense-oriented oligonucleotides, which are currently gaining popularity. Meanwhile, the TAcKLE protocol has been used for a large-scale expression profiling study of more than 100 primary mammary carcinoma samples, supported by a large pharmaceutical company. The project's objective was to identify a gene expression signature that predicts the patient benefit from a novel neoadjuvant chemotherapy combining the drugs Gemcitabine, Epirubicin and Docetaxel [207]. This goal has been achieved, and the results are currently being prepared for publication.

4.2 RCA for Microarray Expression Analysis

Rolling circle amplification holds some compelling theoretical advantages compared to PCR-based or T7 RNA polymerase-based target amplification procedures. A distinctive property of RCA is that the amplified product remains linked to the DNA primer attached to the target molecule. As another unique feature, RCA uses identical circular oligonucleotides for the amplified detection of all target molecules, circumventing sequence-dependent amplification bias. In addition, on-chip RCA can be expected to be less time-consuming, laborious and expensive. Schweitzer *et al.* [158] used RCA for the amplified detection of cytokines on protein microarrays. For this application, specific monoclonal antibodies targeting various cytokines were attached to a solid support and hybridized with supernatant samples from cultured Langerhans cells. Biotinylated, polyclonal secondary antibodies were added to detect the cytokines bound to the arrayed primary antibodies. Subsequently, a universal anti-biotin antibody attached to the 5'-end of a DNA primer could bind to the biotinylated secondary antibodies and initiate rolling circle amplification of a circular oligonucleotide. The amplified product was finally detected by hybridization of multiple fluorescent, complementary oligonucleotide probes. In terms of DNA microarrays, Nallur *et al.* [157] used spotted oligonucleotides as primers for on-chip RCA. In a more advanced genotyping application, Lizardi *et al.* [145] performed allele discrimination by on-chip ligation and subsequent RCA. Briefly, a gene-specific oligonucleotide probe was covalently attached to a solid support via a reactive 3'-amino group, ensuring that the 5'-phosphate was available for ligation. This orientation was preferred because it eliminates the possibility of non-specific priming by the 3'-end, which could otherwise interact with the circular oligonucleotide templates used for RCA. Two additional oligonucleotides were added in solution, capable of discriminating the allelic variants of the target gene. For this purpose, an allele-specific base was located at the 3'-end of a 20-mer target-complementary portion of these probes. The opposite end of these probes comprised a specific RCA primer sequence with an additional free 3'-OH terminus, obtained by reversal of backbone polarity during chemical synthesis. Guided by the target sequence hybridized to the immobilized oligonucleotide, allele-specific ligation of the cognate probe generated a surface-bound oligonucleotide with a free 3´ terminus capable of priming rolling circle amplification of the circular template complementary to the associated primer sequence. Finally, specific detection of the amplified DNA was achieved by hybridization of differentially labeled oligonucleotides complementary to either the mutant or the wild type RCA product.

Several problems had to be addressed in order to adapt RCA for microarray expression analysis. One challenge was the inevitable requirement to generate cDNA that contained a primer sequence with a free 3'-OH terminus to initiate the RCA reaction. This was impossible by means of ordinarily coupling the sequence to an oligo(dT) primer for reverse transcription. In this case, the RCA primer sequence had either been located at the 5'-end, since the 3'-end of the molecule was necessary to prime the reverse transcription, and it had therefore been unavailable to prime the RCA reaction. *Vice versa*, locating the RCA primer at the 3'-end had created another completely useless molecule, as it had been unable to prime the RT. To solve this problem, oligonucleotides were used that were similar to the ones Lizardi *et al.* applied for genotyping by on-chip ligation and subsequent RCA [145]. One part of the molecules was synthesized with reversed backbone polarity, assuring that both the oligo(dT) portion and the RCA primer portion had an opposite orientation, thereby creating two distinct free 3'-ends.

Another issue was the single-strand specific 3' to 5' exonuclease activity of Φ29 DNA polymerase. The problem was actually twofold, since this activity would degrade any single-stranded molecule with an unprotected 3'-terminus and simultaneously occupy a substantial amount of polymerase molecules, thereby keeping them from synthesizing DNA. To minimize these implications, the 3'-ends of the detection probes and the RCA primers were protected by the introduction of single phosphorothioate bonds [208]. The phosphorothioate bond was described to be a much less favoured substrate to nuclease activity than the naturally occuring phosphodiester bond [209]. It might also be beneficial to protect the cDNA analytes from nucleolytic degradation, which could be achieved by addition of either 2'-O-methyl-3'-deoxy-NTP or 2',3'-dideoxynucleoside-5'-O-(1-thiotriphosphate) (Fig. 38) to the reverse transcription reaction. The lack of a 3'-hydroxyl would terminate the polymerization and leave a nuclease-resistant 3'-end protected either by a 2'-O-methyl residue [210] or a thiophosphate (phosphorothioate) bond [208,209]. These reagents are commercially available, and respective experiments will be part of the diploma thesis of cand. biol. Daniel Haag.

An additional complication resulted from the considerable length of the RCA-amplified DNA product. Being several kb in length, corresponding to scores of nanometers, there was a risk that it would be dispersed or even flushed away from the site of amplification. Initial results showed that drying RCA-amplified slides by centrifugation was completely inapplicable. Using the automated draining

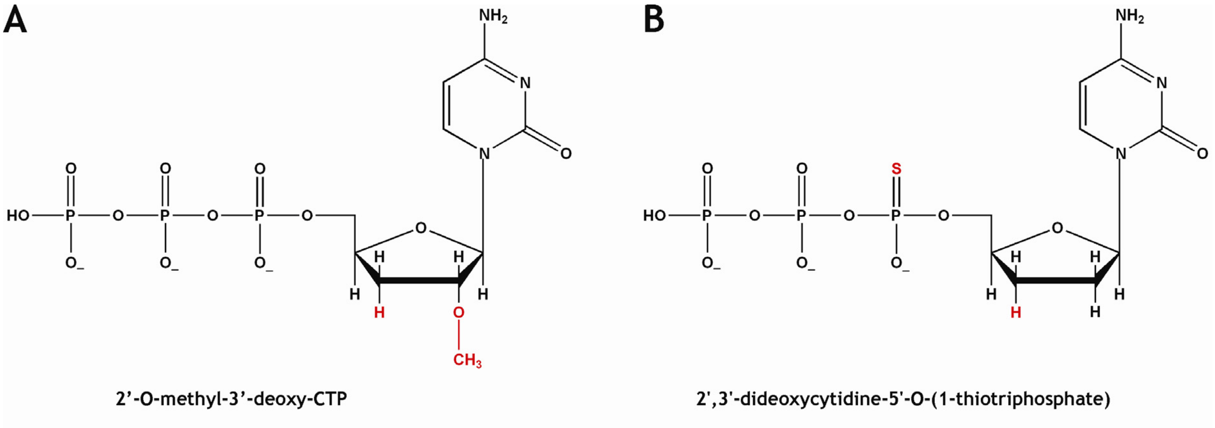


Figure 38. Nucleoside triphosphates with modifications conferring nuclease protection and causing chain termination. (**A**) 2'-O-methyl-3'-deoxy-CTP incorporation during reverse transcription causes chain termination due to the lack of a 3'-hydroxyl and renders the cDNA resistant to 3'-exonucleolytic attacks by means of the 2'-O-methyl residue. (**B**) 2',3'-dideoxycytidine-5'-O-(1-thiotriphosphate) features the same properties, but here nuclease resistance is obtained via a non-hydrolyzable thiophosphate.

functionality of the GeneTAC hybridization station improved the results to a certain extent, but the basic problem remained and interfered with any reasonable quantitative analysis. For that reason, a method was conceived that covalently links the hybridized cDNA molecules, to which the RCA reaction will append the amplified concatemeric copies of the circular DNA template, to their complementary array probes. Initially, the cDNA was tagged with aminoallyl residues, derivatized with the amine-reactive, DNA intercalating reagent NHS-psoralen and hybridized to the array. Irradiation with long UV light (366 nm) created inter-strand cross-links between thymine residues in the cDNA and the arrayed oligonucleotides. This procedure was able to efficiently prevent the spreading and release of hybridized cDNA. For substantially longer RCA incubations, yielding considerably longer amplification products, it might become necessary to condense the DNA via cross-linking with multivalent anti-hapten IgM, using hapten-tags in the detection probes, as suggested by Lizardi *et al.* [145].

To move towards a comparative analysis of different transcriptomes on the same array, a second set of oligonucleotides for RCA was devised, and both systems were used in parallel for the RCA-amplified detection of HL-60 cDNA.

Unfortunately, not only yellow spots of variable intensities were detected, as one would expect for a two-color co-hybridization of cDNAs prepared from the same RNA, but also a considerable number of spots that were primarily if not exclusively green or red. The most likely explanation is an overlap of specific RCA signals and unspecific background due to insufficient washing, and more stringent protocols are currently being tested by cand. biol. Daniel Haag.

As anticipated, it turned out that efficient purification of the circularized padlock probes was essential for successful rolling circle amplification. Remaining ligation template (connector oligonucleotide) could prime an unlocalized, target-independent amplification, whereas unligated probes would arrest the reaction once it reaches the end of the linear template. As described in the “Materials and Methods” section, streptavidin-coated magnetic beads were used to obtain circularized probes for subsequent amplification. Although this procedure was fast and straightforward, it did not always efficiently remove the ligation template (data not shown). Furthermore, this approach is incapable of removing unligated probes. Separation by electrophoresis and subsequent gel extraction yielded extremely pure products (data not shown), but the procedure was both time consuming and difficult to scale up.

HPLC certainly holds the highest potential for a large-scale preparation of circularized probes, even circumventing the need for subsequent exonucleolytic treatment, but this method has not yet been implemented. We currently favor a combination of streptavidin-sepharose columns and subsequent exonuclease treatment to remove unligated probes, but this remains to be evaluated in more detail, despite of promising preliminary results (data not shown).

4.3 Comparison of Oligonucleotide Microarray Platforms

The comparison of relative gene expression measurements obtained with different technical approaches or different implementations of a proven technology is of considerable interest to researchers from all fields of the biological and biomedical sciences. Several studies have addressed this topic, with rather heterogeneous results.

Mah *et al.*[196] compared absolute expression levels quantified on Affymetrix short oligonucleotide and radioactively labeled cDNA-based filter-arrays. The expression values from the two technologies showed merely poor correlations. Tan *et al.*[211] evaluated the performance of three commercial microarray platforms and found merely modest correlations when comparing both absolute and relative gene expression measurements. Strikingly, $\log_2$-ratios from the two platforms using short oligonucleotide probes and biotinylated cRNA targets (Affymetrix and Amersham; r = 0.52) did not correlate better with each other than with those of cDNA arrays (Agilent; r = 0.53 or r = 0.59). In a comparison of Affymetrix GeneChip arrays and two different collections of 70-mer oligonucleotides, Barczak *et al.*[173] found moderate correlations of corresponding signal intensities (r = 0.56 - 0.60), but strong correlations of respective relative expression values (r = 0.80 without filtering, r = 0.83 - 0.89 after exclusion of probes or probe sets with low signal intensities). Similarly, Shippy *et al.*[212] described improved correlations between expression measurements from Affymetrix GeneChip and Amersham CodeLink arrays upon removal of genes within platform noise (r = 0.62 *versus* r = 0.79). Measuring relative gene expression values on Affymetrix short oligonucleotide arrays, commercial (Agilent) and custom-made, sequence-validated cDNA arrays, Järvinen *et al.*[213]

observed reasonable correlations of $\log_2$-ratios. Interestingly, the correlation between the two different cDNA platforms (r =0.73) was weaker than the correlations between the commercial or custom-made cDNA arrays and the Affymetrix system (r = 0.84 and r = 0.76, respectively). A recent study by Tan *et al.*[211], showing very little correlation between Affymetrix, Amersham and Agilent arrays, came to broad public attention[214] and raised general concerns regarding the comparability of expression data across labs and platforms. Shortly after the completion of this project, the prestigious journal 'Nature Methods' dedicated its May 2005 issue to the comparison of array platforms and published several new reports on this topic, emphasizing its relevance and the enduring interest of the scientific community. Larkin *et al.*[215] compared gene expression between Affymetrix GeneChips and spotted cDNA arrays in a mouse model of angiotensin II-induced hypertension and found that biological treatment had a greater impact on the measured expression than the platform for more than 90% of the analyzed genes. In a multiple-laboratory comparison of Affymetrix GeneChips, spotted cDNA arrays and spotted oligonucleotide arrays, Irizarry et al.[216] showed that the results were more affected by the labs than by the different platforms, *i.e.*, it was more important where an experiment was done than on which platform it was done. Bammler *et al.*[217] compared expression profiles between seven labs and across twelve array platforms. Initially, they found poor reproducibility both between labs and across platforms. Not unexpectedly, the reproducibility between labs could clearly be improved by the implementation of standardized protocols for target labeling, hybridization, microarray processing, data acquisition and data normalization.

A different approach to review the possibility for meaningful translation of microarray data is meta-analysis of extensive datasets of similar type (generated from the same type of samples, but not from identical samples), produced in different labs and on different platforms[218-220]. Since many additional parameters such as classification of the samples or individual laboratory practices (see Irizarry *et al.*[216] and Bammler *et al.*[217]) influence the outcome of these studies, the results are rather inconclusive concerning comparability on the technological level. Generally, at least common patterns and/or groups of genes could be confirmed.

The above mentioned studies were, except for meta-analyses, usually based on data generated with homogeneous cell lines and by averaging over several technical replicates. In this study, it was intended to increase the practical significance by the use of clinical samples in combination with modest technical replication (two single arrays per patient for the Affymetrix platform and two dye swap replicates per patient for the spotted oligonucleotide arrays). Additionally, the protocols for target preparation were kept as comparable as possible. Since the Affymetrix platform utilizes biotinylated cRNA generated by *in vitro* transcription (IVT), the linear, IVT-based TAcKLE protocol, which was used for the spotted oligonucleotide arrays, was much more similar compared to conventional dye-labeling by reverse transcription. It could be shown that TAcKLE generates highly reproducible expression profiles with down to 2 ng of starting material[199]. Additionally, it was demonstrated that the correlation of expression ratios obtained with spotted oligonucleotide arrays is higher between replicate amplified sample pairs than between amplified and RT-labeled sample pairs or replicate RT-labeled sample pairs. Accordingly, one can expect that consistent target amplification would also be beneficial, if expression ratios are to be compared across platforms. Comparative studies that do not account for this consideration might introduce additional systematic bias, resulting in reduced agreement between platforms.

To match the probes from the two platforms, accession numbers provided by Affymetrix and Operon were mapped to the current version of the UniGene database. Transcript identifiers from the RefSeq collection[221,222] were not chosen for matching the platforms, since reference sequences can change through consolidation of the database. Recently, Mecham *et al.*[223] showed that up to 50% of Affymetrix probes do not have a matching sequence in the current version of RefSeq. Despite these considerations, platform matching by RefSeq identifiers yielded approximately similar and partly even improved results in terms of cross platform correlation (9,922 genes could be assigned as represented on both platforms, correlations of unfiltered $\log_2$-ratios were r = 0.66 - r = 0.81; data not shown). Evolution of the UniGene database (accession numbers that were removed due to misalignment or retraction by their submitters; UniGene clusters that were retired as they could be joined or split) and the associated loss of cross-references may also explain why we identified less genes common to both array types than previously reported by Barczak *et al.*[173]. It was also decided against matching by GenBank accession numbers, since corresponding probes and probe sets can be annotated by different accession numbers of the same UniGene, causing this procedure to exclude large amounts of potentially useful information.

Considerable variations in the degree of correlation were detected (Fig. 32) when comparing unfiltered, $\log_2$-transformed expression ratios of individual patients, obtained with either GeneChip short oligonucleotide arrays or spotted long oligonucleotide arrays,. As reported previously [173,224], these correlations improved after the exclusion of probes and probe sets associated with low signal intensities. This observation might, at least in part, be attributed to variations in the performance of individual array experiments.

The correlations between expression ratios could further be improved (Table 14) by the application of systematic bias adjustment via 'Distance Weighted Discrimination (DWD)'. DWD is an advanced method for the adjustment of various systematic differences across microarray experiment subpopulations, including sample source, batch and platform effects [178], which facilitates the merging of different data sets. DWD uses an approach similar to that of support vector machines (SVM) [225], but delivers improved performance in the context of high-dimensional, low sample size (HDLSS) data such as those obtained by microarray analyses. Both methods aim at finding a hyperplane in high-dimensional space, which separates defined subpopulations of data as completely as possible. The essential difference is that, while SVM tries to maximize the minimum distance (margin) of all the data to the separating plane, DWD works by maximizing the sum of the inverse distances. In this way, all data points have an influence on the result (optimized position of the hyperplane), and data piling at the margins is avoided, a problem associated with the minimum distance criterion of SVM. After determination of the DWD direction vector, all data points of each subpopulation are projected onto the direction given by this vector. Finally, data points from each subpopulation are shifted in the DWD direction by subtracting the DWD direction vector multiplied by their projected means, thereby effectively removing systematic variation while preserving any variation in the DWD direction not caused by systematic effects. Applied to the data generated in the course of this thesis, the DWD approach clearly and consistently improved cross platform correlations while shifting the slopes of corresponding regression lines towards one (Table 14). The latter effect was minimal in case both data sets had been normalized by the same algorithm (*vsn*), as this procedure not unexpectedly yielded slopes closest to one even before DWD. A slope close to one implies that genes are more likely to yield similar results (regardless of differential expression) on both of the investigated platforms. Further improvements of DWD performance can be anticipated for more extensive data sets, and the method might greatly enhance agreement in future comparative studies.

A goal of this project was to compare reliable measurements from both systems, both of which can be regarded as detecting overlapping but different subsets of the actual set of differentially expressed genes. This was confirmed by EASE overrepresentation analysis [181], which revealed that some of the differentially expressed genes could be assigned to the same 'theme' on both platforms, whereas others were exclusive to one of the platforms (Table 16). On each array system, approximately 50 genes were consistently and repeatedly scored as differentially expressed, and the intersection of these groups contained 21 common genes (Fig. 33a). The majority of genes restricted to one of the platforms showed no sufficient degree and/or significance, but at least the same direction of regulated expression on the other platform (Fig. 33b). Therefore, it doesn't matter if a clinical study uses Affymetrix or Operon long oligonucleotide arrays, as long as these are used consistently and combined with high quality control standards throughout the whole investigation.

For a subset of genes, microarray-derived expression ratios were verified by RQ-PCR, finding good qualitative agreement between the two array platforms and the PCR-based method (Fig. 36).

It was shown that, overall, expression profiles obtained with either long (Operon) or multiple short (Affymetrix) oligonucleotide microarrays display a reasonable correlation, with variable concordance of individual genes. Based on patient samples, results were obtained that are in good agreement with previous studies utilizing cell line-derived RNA. Projecting these findings to a larger series of array experiments, one could expect to obtain similar albeit not identical results, concerning, *e.g.*, a hierarchical clustering or a gene expression signature, with either of the two investigated platforms. On the level of individual genes and quantitative precision, however, the results of this study reaffirm that microarrays have to be considered a screening technology and that their data should be regarded with caution. This should be kept in mind particularly when comparing data from different array platforms. Recently, important progress has been made to facilitate this transfer of information. Guidelines provided by the 'Microarray Gene Expression Data Society (MGED)' (http://www.mged.org), which developed the 'Minimum Information About a Microarray Experiment (MIAME)' specifications [226], assist researchers in the annotation of their microarray experiments. Further improvement is provided by public microarray repositories, which facilitate the publication and sharing of properly annotated gene expression data. Statistical methods like Distance Weighted Discrimination [178] can further improve the comparability of microarray data sets, since systematic biases arising from platform-specific parameters, such as measurement precision (reproducibility), accuracy (regarding the "true" values), specificity and sensitivity or differences in protocol performance, can be properly weighted and adjusted accordingly. The utility of future array studies could further improve if the 'External RNA Control Consortium (ERCC)' is successful in its effort to standardize controls for the calibration of microarray experiments. But ultimately, meaningful comparison, translation and integration of expression data will be impaired as long as industrial standards are missing for all stages of the experiment, *i.e.*, the design of array probes, the production of the arrays, sample handling, RNA extraction, target amplification and target labeling, hybridization, washing, data acquisition, data filtering and data normalization.

4.4 Conclusions

Prediction is very difficult, especially of the future.

NIELS BOHR

In this thesis, a novel procedure (TAcKLE) for the generation of fluorescent, antisense-oriented target molecules was successfully conceived, implemented, evaluated and employed. It is particularly valuable for applications of spotted oligonucleotide microarrays where only limited amounts of RNA source material are available. It could be shown that the novel method generates expression profiles of exceptional quality and reproducibility. As spotted oligonucleotide arrays are becoming increasingly popular and no equally qualified protocols are currently available, it is to be expected that the TAcKLE will be widely used in the fields of basic biological and biomedical research as well as for biotechnology applications.

The protocol has already been used for a large-scale expression profiling study of more than 100 primary mamma carcinoma samples obtained by core needle biopsies, which could identify a gene expression signature that predicts the patient benefit concerning a novel neoadjuvant chemotherapy combining the drugs gemcitabine, epirubicine and docetaxel [207].

As an alternative to target amplification via the TAcKLE protocol, a signal amplification method was devised. It uses the principle of rolling circle replication, known from the duplication of circular phage genomes or plasmids, to achieve a localized amplification of a circular oligonucleotide to enhance the detection of hybridization signals. The method is straightforward, cost-efficient and offers the compelling theoretical advantage of eliminating sequence-dependent amplification bias. Further experimentation will be needed to establish a sufficiently robust protocol, but ultimately this new procedure might greatly simplify the way in which microarrays are used to assay limited source material.

Finally, gene expression profiles of HNSCC patients were generated both on spotted 70-mer oligonucleotide microarrays using the TAcKLE protocol, and on commercial microarrays produced by photolithographic *in situ* synthesis of 25-mer probes (Affymetrix). This comparison was particularly important, since spotted oligonucleotide arrays had just recently been introduced as an alternative to cDNA arrays and the Affymetrix system, which combines attractive advantages from both platforms. Like cDNA arrays, spotted oligonucleotide arrays are produced by individual research groups, providing a considerable cost advantage. As for the Affymetrix platform, their probes are designed to have similar biophysical properties and avoid secondary structures as well as repetitive sequences. Additionally, they are long enough to allow for a specific analysis via just one probe for each target, whereas Affymetrix, due to the reduced specificity of 25-mer sequences, requires 11 - 16 probes per target. These analyses revealed strong correlations between the data sets generated on the platform of spotted 70-mer oligonucleotides and the Affymetrix system, and RQ-PCR analysis confirmed the concordance for selected genes. There were reproducible differences between the two platforms, though, and meaningful comparison and translation of gene expression data will be impaired as long as industrial standards are missing for the production and handling of arrays as well as for the design of array probes. Both array systems detect overlapping, equally large but non-identical subsets of the actual set of differentially expressed genes.

Consequently, they are equally qualified for any expression profiling study, as long as they are used consistently.

In conclusion, combining the TAcKLE protocol with spotted oligonucleotide arrays is an attractive alterative for transcriptional profiling of limited source material, which easily bears comparison with an accepted commercial reference platform.

5 References

1. Collins, F. S., Morgan, M. & Patrinos, A. The Human Genome Project: lessons from large-scale biology. *Science* **300**, 286-90 (2003).

2. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

3. Collins, F. S., Green, E. D., Guttmacher, A. E. & Guyer, M. S. A vision for the future of genomics research. *Nature* **422**, 835-47 (2003).

4. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J. *et al.* The sequence of the human genome. *Science* **291**, 1304-51 (2001).

5. Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. USA* **99**, 4465-70 (2002).

6. World Health Organization. World Health Statistics 2005. *WHO Press*, Geneva, Switzerland (2005).

7. Parkin, D. M., Bray, F., Ferlay, J. & Pisani, P. Estimating the world cancer burden: Globocan 2000. *Int. J. Cancer* **94**, 153-6 (2001).

8. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-45 (2004).

9. Stein, L. D. Human genome: end of the beginning. *Nature* **431**, 915-6 (2004).

10. Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L. & Quackenbush, J. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* **25**, 239-40 (2000).

11. Pennisi, E. Bioinformatics. Gene counters struggle to get the right answer. *Science* **301**, 1040-1 (2003).

12. Pennisi, E. Human genome. A low number wins the GeneSweep Pool. *Science* **300**, 1484 (2003).

13. Wright, F. A., Lemon, W. J., Zhao, W. D., Sears, R., Zhuo, D. *et al.* A draft annotation and overview of the human genome. *Genome Biol.* **2**, RESEARCH0025 (2001).

14. Claverie, J. M. Gene number. What if there are only 30,000 human genes? *Science* **291**, 1255-7 (2001).

15. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **25**, 232-4 (2000).

16. Aparicio, S. A. How to count ... human genes. *Nat. Genet.* **25**, 129-30 (2000).

17. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. Molecular Biology of the Cell. *Garland Publishing*, New York (2002).

18. Gottlieb, S. The splice of life. *Understanding the RNAissance*, NPG Horizon Symposia, Prouts Neck, USA (2003).

19. Strachan, T. & Read, A. P. Human Molecular Genetics 2. *BIOS Scientific Publishers, Ltd*, Oxford, UK (1999).

20. C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-8 (1998).

21. Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D. *et al.* The genome sequence of Drosophila melanogaster. *Science* **287**, 2185-95 (2000).

22. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815 (2000).

23. Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* **287**, 560-1 (1980).

24. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. USA* **90**, 11995-9 (1993).

25. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366**, 362-5 (1993).

26. Pfeifer, G. P., Tanguay, R. L., Steigerwald, S. D. & Riggs, A. D. In vivo footprint and methylation analysis by PCR-aided genomic sequencing: comparison of active and inactive X chromosomal DNA at the CpG island and promoter of human PGK-1. *Genes Dev.* **4**, 1277-87 (1990).

27. Fazzari, M. J. & Greally, J. M. Epigenomics: beyond CpG islands. *Nat. Rev. Genet.* **5**, 446-55 (2004).

28. Bird, A. P. & Wolffe, A. P. Methylation-induced repression--belts, braces, and chromatin. *Cell* **99**, 451-4 (1999).

29. Sims, R. J., 3rd, Nishioka, K. & Reinberg, D. Histone lysine methylation: a signature for chromatin function. *Trends Genet.* **19**, 629-39 (2003).

30. Roth, S. Y., Denu, J. M. & Allis, C. D. Histone acetyltransferases. *Annu. Rev. Biochem.* **70**, 81-120 (2001).

31. Thomson, S., Clayton, A. L. & Mahadevan, L. C. Independent dynamic regulation of histone phosphorylation and acetylation during immediate-early gene induction. *Mol. Cell* **8**, 1231-41 (2001).

32. Zhang, Y. Transcriptional regulation by histone ubiquitination and deubiquitination. *Genes Dev.* **17**, 2733-40 (2003).

33. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074-80 (2001).

34. Boveri, T. Über mehrpolige Mitosen als Mittel zur Analyse des Zellkerns. *Verh. Phys. Med. Ges. Würzburg N. F.* **35**, 67-90 (1902).

35. Boveri, T. Ueber die Konstitution der chromatischen Kernsubstanz. *Verh. Zool. Ges.* **13** (1903).

36. Boveri, T. Ergebnisse ueber die Konstitution der chromatischen Substanz des Zellkerns. *Gustav Fischer*, Jena (1904).

37. Boveri, T. in *Zur Frage der Entstehung maligner Tumoren* 1-64 (Gustav Fischer, Jena, 1914).

38. Anders, F. Tumour formation in platyfish-swordtail hybrids as a problem of gene regulation. *Experientia* **23**, 1-10 (1967).

39. Stehelin, D., Varmus, H. E., Bishop, J. M. & Vogt, P. K. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260**, 170-3 (1976).

40. Bishop, J. M. Enemies within: the genesis of retrovirus oncogenes. *Cell* **23**, 5-6 (1981).

41. Shih, C., Shilo, B. Z., Goldfarb, M. P., Dannenberg, A. & Weinberg, R. A. Passage of phenotypes of chemically transformed cells via transfection of DNA and chromatin. *Proc. Natl. Acad. Sci. USA* **76**, 5714-8 (1979).

42. Shilo, B. Z. & Weinberg, R. A. Unique transforming gene in carcinogen-transformed mouse cells. *Nature* **289**, 607-9 (1981).

43. Murray, M. J., Shilo, B. Z., Shih, C., Cowing, D., Hsu, H. W. & Weinberg, R. A. Three different human tumor cell lines contain different oncogenes. *Cell* **25**, 355-61 (1981).

44. Shih, C., Padhy, L. C., Murray, M. & Weinberg, R. A. Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. *Nature* **290**, 261-4 (1981).

45. Sukumar, S., Notario, V., Martin-Zanca, D. & Barbacid, M. Induction of mammary carcinomas in rats by nitroso-methylurea involves malignant activation of H-ras-1 locus by single point mutations. *Nature* **306**, 658-61 (1983).

46. Felsher, D. W. Cancer revoked: oncogenes as therapeutic targets. *Nat. Rev. Cancer* **3**, 375-80 (2003).

47. Hantschel, O. & Superti-Furga, G. Regulation of the c-Abl and Bcr-Abl tyrosine kinases. *Nat. Rev. Mol. Cell Biol.* **5**, 33-44 (2004).

48. Rowley, J. D. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243**, 290-3 (1973).

49. Nowell, P. C. & Hungerford, D. A. Chromosome studies on normal and leukemic human leukocytes. *J. Natl. Cancer Inst.* **25**, 85-109 (1960).

50. Heisterkamp, N., Stephenson, J. R., Groffen, J., Hansen, P. F., de Klein, A., Bartram, C. R. & Grosveld, G. Localization of the c-ab1 oncogene adjacent to a translocation break point in chronic myelocytic leukaemia. *Nature* **306**, 239-42 (1983).

51. Ren, R. Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia. *Nat. Rev. Cancer* **5**, 172-83 (2005).

52. Baylin, S. B., Esteller, M., Rountree, M. R., Bachman, K. E., Schuebel, K. & Herman, J. G. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Hum. Mol. Genet.* **10**, 687-92 (2001).

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

54. Vogelstein, B. & Kinzler, K. W. The Genetic Basis of Human Cancer. *McGraw-Hill*, New York (2002).

55. Friend, S. H., Bernards, R., Rogelj, S., Weinberg, R. A., Rapaport, J. M., Albert, D. M. & Dryja, T. P. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323**, 643-6 (1986).

56. Chau, B. N. & Wang, J. Y. Coordinated regulation of life and death by RB. *Nat. Rev. Cancer* **3**, 130-8 (2003).

57. Nevins, J. R. The Rb/E2F pathway and cancer. *Hum. Mol. Genet.* **10**, 699-703 (2001).

58. Harbour, J. W. & Dean, D. C. Rb function in cell-cycle regulation and apoptosis. *Nat. Cell Biol.* **2**, E65-7 (2000).

59. Nielsen, S. J., Schneider, R., Bauer, U. M., Bannister, A. J., Morrison, A. *et al.* Rb targets histone H3 methylation and HP1 to promoters. *Nature* **412**, 561-5 (2001).

60. Groden, J., Thliveris, A., Samowitz, W., Carlson, M., Gelbert, L. *et al.* Identification and characterization of the familial adenomatous polyposis coli gene. *Cell* **66**, 589-600 (1991).

61. Futreal, P. A., Liu, Q., Shattuck-Eidens, D., Cochran, C., Harshman, K. *et al.* BRCA1 mutations in primary breast and ovarian carcinomas. *Science* **266**, 120-2 (1994).

62. Lane, D. P. & Crawford, L. V. T antigen is bound to a host protein in SV40-transformed cells. *Nature* **278**, 261-3 (1979).

63. Whyte, P., Buchkovich, K. J., Horowitz, J. M., Friend, S. H., Raybuck, M., Weinberg, R. A. & Harlow, E. Association between an oncogene and an anti-oncogene: the adenovirus E1A proteins bind to the retinoblastoma gene product. *Nature* **334**, 124-9 (1988).

64. Baker, S. J., Fearon, E. R., Nigro, J. M., Hamilton, S. R., Preisinger, A. C. *et al.* Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science* **244**, 217-21 (1989).

65. Baker, S. J., Markowitz, S., Fearon, E. R., Willson, J. K. & Vogelstein, B. Suppression of human colorectal carcinoma cell growth by wild-type p53. *Science* **249**, 912-5 (1990).

66. Fearon, E. R., Hamilton, S. R. & Vogelstein, B. Clonal analysis of human colorectal tumors. *Science* **238**, 193-7 (1987).

67. Polyak, K., Xia, Y., Zweier, J. L., Kinzler, K. W. & Vogelstein, B. A model for p53-induced apoptosis. *Nature* **389**, 300-5 (1997).

68. Vogelstein, B., Lane, D. & Levine, A. J. Surfing the p53 network. *Nature* **408**, 307-10 (2000).

69. Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C. C. p53 mutations in human cancers. *Science* **253**, 49-53 (1991).

70. Levine, A. J., Momand, J. & Finlay, C. A. The p53 tumour suppressor gene. *Nature* **351**, 453-6 (1991).

71. Sengupta, S. & Harris, C. C. p53: traffic cop at the crossroads of DNA repair and recombination. *Nat. Rev. Mol. Cell Biol.* **6**, 44-55 (2005).

72. Loeb, L. A. Mutator phenotype may be required for multistage carcinogenesis. *Cancer Res.* **51**, 3075-9 (1991).

73. Loeb, L. A., Springgate, C. F. & Battula, N. Errors in DNA replication as a basis of malignant changes. *Cancer Res.* **34**, 2311-21 (1974).

74. Parsons, R., Li, G. M., Longley, M. J., Fang, W. H., Papadopoulos, N., Jen, J., de la Chapelle, A., Kinzler, K. W., Vogelstein, B. & Modrich, P. Hypermutability and mismatch repair deficiency in RER+ tumor cells. *Cell* **75**, 1227-36 (1993).

75. Eden, A., Gaudet, F., Waghmare, A. & Jaenisch, R. Chromosomal instability and tumors promoted by DNA hypomethylation. *Science* **300**, 455 (2003).

76. Loeb, L. A. Cancer cells exhibit a mutator phenotype. *Adv. Cancer Res.* **72**, 25-56 (1998).

77. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396**, 643-9 (1998).

78. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).

79. Hahn, W. C. & Weinberg, R. A. Modelling the molecular circuitry of cancer. *Nat. Rev. Cancer* **2**, 331-41 (2002).

80. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759-67 (1990).

81. Marks, F. & Furstenberger, G. Cancer chemoprevention through interruption of multistage carcinogenesis. The lessons learnt by comparing mouse skin carcinogenesis and human large bowel cancer. *Eur. J. Cancer.* **36**, 314-29 (2000).

82. Abbott, A. Betting on tomorrow's chips. *Nature* **415**, 112-4 (2002).

83. Alwine, J. C., Kemp, D. J. & Stark, G. R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci. USA* **74**, 5350-4 (1977).

84. Berk, A. J. & Sharp, P. A. Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell* **12**, 721-32 (1977).

85. Liang, P. & Pardee, A. B. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257**, 967-71 (1992).

86. Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651-6 (1991).

87. Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y. & Matsubara, K. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat. Genet.* **2**, 173-9 (1992).

88. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484-7 (1995).

89. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-70 (1995).

90. Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O. & Davis, R. W. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA* **93**, 10614-9 (1996).

91. DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A. & Trent, J. M. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **14**, 457-60 (1996).

92. Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14**, 1675-80 (1996).

93. Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J. *et al.* Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* **19**, 342-7 (2001).

94. Kane, M. D., Jatkoe, T. A., Stumpf, C. R., Lu, J., Thomas, J. D. & Madore, S. J. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.* **28**, 4552-7 (2000).

95. Maskos, U. & Southern, E. M. Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. *Nucleic Acids Res.* **20**, 1679-84 (1992).

96. Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T. & Solas, D. Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**, 767-73 (1991).

97. Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P. & Fodor, S. P. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. USA* **91**, 5022-6 (1994).

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

99. Southern, E. M. An improved method for transferring nucleotides from electrophoresis strips to thin layers of ion-exchange cellulose. *Anal. Biochem.* **62**, 317-8 (1974).

100. Southern, E. M. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98**, 503-17 (1975).

101. Forozan, F., Mahlamaki, E. H., Monni, O., Chen, Y., Veldman, R., Jiang, Y., Gooden, G. C., Ethier, S. P., Kallioniemi, A. & Kallioniemi, O. P. Comparative genomic hybridization analysis of 38 breast cancer cell lines: a basis for interpreting complementary DNA microarray data. *Cancer Res.* **60**, 4519-25 (2000).

102. Wild, P., Knuechel, R., Dietmaier, W., Hofstaedter, F. & Hartmann, A. Laser microdissection and microsatellite analyses of breast cancer reveal a high degree of tumor heterogeneity. *Pathobiology* **68**, 180-90 (2000).

103. Nishizuka, I., Ishikawa, T., Hamaguchi, Y., Kamiyama, M., Ichikawa, Y. *et al.* Analysis of gene expression involved in brain metastasis from breast cancer using cDNA microarray. *Breast Cancer* **9**, 26-32 (2002).

104. Assersohn, L., Gangi, L., Zhao, Y., Dowsett, M., Simon, R., Powles, T. J. & Liu, E. T. The feasibility of using fine needle aspiration from primary breast cancers for cDNA microarray analyses. *Clin. Cancer Res.* **8**, 794-801 (2002).

105. Symmans, W. F., Ayers, M., Clark, E. A., Stec, J., Hess, K. R. *et al.* Total RNA yield and microarray gene expression profiles from fine-needle aspiration biopsy and core-needle biopsy samples of breast carcinoma. *Cancer* **97**, 2960-71 (2003).

106. Zhu, G., Reynolds, L., Crnogorac-Jurcevic, T., Gillett, C. E., Dublin, E. A. *et al.* Combination of microdissection and microarray analysis to identify gene expression changes between differentially located tumour cells in breast cancer. *Oncogene* **22**, 3742-8 (2003).

107. Ellis, M., Davis, N., Coop, A., Liu, M., Schumaker, L. *et al.* Development and validation of a method for using breast core needle biopsies for gene expression microarray analyses. *Clin. Cancer Res.* **8**, 1155-66 (2002).

108. Iscove, N. N., Barbara, M., Gu, M., Gibson, M., Modi, C. & Winegarden, N. Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. *Nat. Biotechnol.* **20**, 940-3 (2002).

109. Klein, C. A., Seidl, S., Petat-Dutter, K., Offner, S., Geigl, J. B. *et al.* Combined transcriptome and genome analysis of single micrometastatic cells. *Nat. Biotechnol.* **20**, 387-92 (2002).

110. Makrigiorgos, G. M., Chakrabarti, S., Zhang, Y., Kaur, M. & Price, B. D. A PCR-based amplification method retaining the quantitative difference between two complex genomes. *Nat. Biotechnol.* **20**, 936-9 (2002).

111. Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettel, M. & Coleman, P. Analysis of gene expression in single live neurons. *Proc. Natl. Acad. Sci. USA* **89**, 3010-4 (1992).

112. Van Gelder, R. N., von Zastrow, M. E., Yool, A., Dement, W. C., Barchas, J. D. & Eberwine, J. H. Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc. Natl. Acad. Sci. USA* **87**, 1663-7 (1990).

113. Mullis, K. B. & Faloona, F. A. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol.* **155**, 335-50 (1987).

114. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487-91 (1988).

115. Stenman, J. & Orpana, A. Accuracy in amplification. *Nat. Biotechnol.* **19**, 1011-2 (2001).

116. Kenzelmann, M. & Muhlemann, K. Transcriptome analysis of fibroblast cells immediate-early after human cytomegalovirus infection. *J. Mol. Biol.* **304**, 741-51 (2000).

117. Okazaki, R., Okazaki, T., Sakabe, K. & Sugimoto, K. Mechanism of DNA replication possible discontinuity of DNA chain growth. *Jpn. J. Med. Sci. Biol.* **20**, 255-60 (1967).

118. Okazaki, R., Okazaki, T., Sakabe, K., Sugimoto, K. & Sugino, A. Mechanism of DNA chain growth. I. Possible discontinuity and unusual secondary structure of newly synthesized chains. *Proc. Natl. Acad. Sci. USA* **59**, 598-605 (1968).

119. Sugimoto, K., Okazaki, T. & Okazaki, R. Mechanism of DNA chain growth, II. Accumulation of newly synthesized short chains in E. coli infected with ligase-defective T4 phages. *Proc. Natl. Acad. Sci. USA* **60**, 1356-62 (1968).

120. Xiang, C. C., Chen, M., Ma, L., Phan, Q. N., Inman, J. M., Kozhich, O. A. & Brownstein, M. J. A new strategy to amplify degraded RNA from small tissue samples for microarray studies. *Nucleic Acids Res.* **31**, e53 (2003).

121. Goff, L. A., Bowers, J., Schwalm, J., Howerton, K., Getts, R. C. & Hart, R. P. Evaluation of sense-strand mRNA amplification by comparative quantitative PCR. *BMC Genomics* **5**, 76 (2004).

122. 't Hoen, P. A., de Kort, F., van Ommen, G. J. & den Dunnen, J. T. Fluorescent labelling of cRNA for microarray applications. *Nucleic Acids Res.* **31**, e20 (2003).

123. Smith, L., Underhill, P., Pritchard, C., Tymowska-Lalanne, Z., Abdul-Hussein, S. *et al.* Single primer amplification (SPA) of cDNA for microarray expression analysis. *Nucleic Acids Res.* **31**, e9 (2003).

124. Bobrow, M. N., Harris, T. D., Shaughnessy, K. J. & Litt, G. J. Catalyzed reporter deposition, a novel method of signal amplification. Application to immunoassays. *J. Immunol. Methods* **125**, 279-85 (1989).

125. Bobrow, M. N., Litt, G. J., Shaughnessy, K. J., Mayer, P. C. & Conlon, J. The use of catalyzed reporter deposition as a means of signal amplification in a variety of formats. *J. Immunol. Methods* **150**, 145-9 (1992).

126. Yu, J., Othman, M. I., Farjo, R., Zareparsi, S., MacNee, S. P., Yoshida, S. & Swaroop, A. Evaluation and optimization of procedures for target labeling and hybridization of cDNA microarrays. *Mol Vis* **8**, 130-7 (2002).

127. Richter, A., Schwager, C., Hentze, S., Ansorge, W., Hentze, M. W. & Muckenthaler, M. Comparison of fluorescent tag DNA labeling methods used for expression analysis by DNA microarrays. *Biotechniques* **33**, 620-8, 630 (2002).

128. Stears, R. L., Getts, R. C. & Gullans, S. R. A novel, sensitive detection system for high-density microarrays using dendrimer technology. *Physiol. Genomics* **3**, 93-9 (2000).

129. Nilsen, T. W., Grayzel, J. & Prensky, W. Dendritic nucleic acid structures. *J. Theor. Biol.* **187**, 273-84 (1997).

130. Meijer, W. J., Horcajadas, J. A. & Salas, M. Phi29 family of phages. *Microbiol. Mol. Biol. Rev.* **65**, 261-87 (2001).

131. Anderson, D. L., Hickman, D. D. & Reilly, B. E. Structure of Bacillus subtilis bacteriophage phi 29 and the length of phi 29 deoxyribonucleic acid. *J. Bacteriol.* **91**, 2081-9 (1966).

132. Vlcek, C. & Paces, V. Nucleotide sequence of the late region of Bacillus phage phi 29 completes the 19,285-bp sequence of phi 29 genome. Comparison with the homologous sequence of phage PZA. *Gene* **46**, 215-25 (1986).

133. Watabe, K., Shin, M. & Ito, J. Protein-primed initiation of phage phi 29 DNA replication. *Proc. Natl. Acad. Sci. USA* **80**, 4248-52 (1983).

134. Salas, M. Protein-priming of DNA replication. *Annu. Rev. Biochem.* **60**, 39-71 (1991).

135. Mellado, R. P., Penalva, M. A., Inciarte, M. R. & Salas, M. The protein covalently linked to the 5' termini of the DNA of Bacillus subtilis phage phi 29 is involved in the initiation of DNA replication. *Virology* **104**, 84-96 (1980).

136. Blanco, L., Garcia, J. A., Penalva, M. A. & Salas, M. Factors involved in the initiation of phage phi 29 DNA replication in vitro: requirement of the gene 2 product for the formation of the protein p3-dAMP complex. *Nucleic Acids Res.* **11**, 1309-23 (1983).

137. Matsumoto, K., Saito, T. & Hirokawa, H. In vitro initiation of bacteriophage phi 29 and M2 DNA replication: genes required for formation of a complex between the terminal protein and 5'dAMP. *Mol. Gen. Genet.* **191**, 26-30 (1983).

138. Blanco, L., Garcia, J. A. & Salas, M. Cloning and expression of gene 2, required for the protein-primed initiation of the Bacillus subtilis phage phi 29 DNA replication. *Gene* **29**, 33-40 (1984).

139. Blanco, L. & Salas, M. Characterization and purification of a phage phi 29-encoded DNA polymerase required for the initiation of replication. *Proc. Natl. Acad. Sci. USA* **81**, 5325-9 (1984).

140. Blanco, L. & Salas, M. Replication of phage phi 29 DNA with purified terminal protein and DNA polymerase: synthesis of full-length phi 29 DNA. *Proc. Natl. Acad. Sci. USA* **82**, 6404-8 (1985).

141. Blanco, L., Bernad, A., Lazaro, J. M., Martin, G., Garmendia, C. & Salas, M. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J. Biol. Chem.* **264**, 8935-40 (1989).

142. Fire, A. & Xu, S. Q. Rolling replication of short DNA circles. *Proc. Natl. Acad. Sci. USA* **92**, 4641-5 (1995).

143. Zhang, D. Y., Brandwein, M., Hsuih, T. C. & Li, H. Amplification of target-specific, ligation-dependent circular probe. *Gene* **211**, 277-85 (1998).

144. Daubendiek, S. L. & Kool, E. T. Generation of catalytic RNAs by rolling transcription of synthetic DNA nanocircles. *Nat. Biotechnol.* **15**, 273-7 (1997).

145. Lizardi, P. M., Huang, X., Zhu, Z., Bray-Ward, P., Thomas, D. C. & Ward, D. C. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat. Genet.* **19**, 225-32 (1998).

146. Baner, J., Nilsson, M., Mendel-Hartvig, M. & Landegren, U. Signal amplification of padlock probes by rolling circle replication. *Nucleic Acids Res.* **26**, 5073-8 (1998).

147. Nelson, J. R., Cai, Y. C., Giesler, T. L., Farchaus, J. W., Sundaram, S. T. *et al.* TempliPhi, phi29 DNA polymerase based rolling circle amplification of templates for DNA sequencing. *Biotechniques* **Suppl**, 44-7 (2002).

148. Predki, P. F., Elkin, C., Kapur, H., Jett, J., Lucas, S., Glavina, T. & Hawkins, T. Rolling circle amplification for sequencing templates. *Methods Mol. Biol.* **255**, 189-96 (2004).

149. Alsmadi, O. A., Bornarth, C. J., Song, W., Wisniewski, M., Du, J. *et al.* High accuracy genotyping directly from genomic DNA using a rolling circle amplification based assay. *BMC Genomics* **4**, 21 (2003).

150. Hosono, S., Faruqi, A. F., Dean, F. B., Du, Y., Sun, Z., Wu, X., Du, J., Kingsmore, S. F., Egholm, M. & Lasken, R. S. Unbiased whole-genome amplification directly from clinical samples. *Genome Res.* **13**, 954-64 (2003).

151. Paez, J. G., Lin, M., Beroukhim, R., Lee, J. C., Zhao, X. *et al.* Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res.* **32**, e71 (2004).

152. Wang, G., Maher, E., Brennan, C., Chin, L., Leo, C., Kaur, M., Zhu, P., Rook, M., Wolfe, J. L. & Makrigiorgos, G. M. DNA amplification method tolerant to sample degradation. *Genome Res.* **14**, 2357-66 (2004).

153. Lage, J. M., Leamon, J. H., Pejovic, T., Hamann, S., Lacey, M. *et al.* Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH. *Genome Res.* **13**, 294-307 (2003).

154. Cardoso, J., Molenaar, L., de Menezes, R. X., Rosenberg, C., Morreau, H., Moslein, G., Fodde, R. & Boer, J. M. Genomic profiling by DNA amplification of laser capture microdissected tissues and array CGH. *Nucleic Acids Res.* **32**, e146 (2004).

155. Buckley, P. G., Mantripragada, K. K., Benetkiewicz, M., Tapia-Paez, I., Diaz De Stahl, T. *et al.* A full-coverage, high-resolution human chromosome 22 genomic microarray for clinical and research applications. *Hum. Mol. Genet.* **11**, 3221-9 (2002).

156. Albertson, D. G. & Pinkel, D. Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet.* **12 Spec No 2**, R145-52 (2003).

157. Nallur, G., Luo, C., Fang, L., Cooley, S., Dave, V., Lambert, J., Kukanskis, K., Kingsmore, S., Lasken, R. & Schweitzer, B. Signal amplification by rolling circle amplification on DNA microarrays. *Nucleic Acids Res.* **29**, E118 (2001).

158. Schweitzer, B., Roberts, S., Grimwade, B., Shao, W., Wang, M. *et al.* Multiplexed protein profiling on microarrays by rolling-circle amplification. *Nat. Biotechnol.* **20**, 359-65 (2002).

159. Schweitzer, B., Wiltshire, S., Lambert, J., O'Malley, S., Kukanskis, K., Zhu, Z., Kingsmore, S. F., Lizardi, P. M. & Ward, D. C. Inaugural article: immunoassays with rolling circle DNA amplification: a versatile platform for ultrasensitive antigen detection. *Proc. Natl. Acad. Sci. USA* **97**, 10113-9 (2000).

160. Nilsson, M., Malmgren, H., Samiotaki, M., Kwiatkowski, M., Chowdhary, B. P. & Landegren, U. Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* **265**, 2085-8 (1994).

161. UICC. TNM Classification of Malignant Tumours. (eds. Sobin, L. H. & Wittekind, C.) *John Wiley & Sons*, New York (2002).

162. Heiger, D. N., Cohen, A. S. & Karger, B. L. Separation of DNA restriction fragments by high performance capillary electrophoresis with low and zero crosslinked polyacrylamide using continuous and pulsed electric fields. *J. Chromatogr.* **516**, 33-48 (1990).

163. Woolley, A. T. & Mathies, R. A. Ultra-high-speed DNA fragment separations using microfabricated capillary array electrophoresis chips. *Proc. Natl. Acad. Sci. USA* **91**, 11348-52 (1994).

164. 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).

165. Wrobel, G., Schlingemann, J., Hummerich, L., Kramer, H., Lichter, P. & Hahn, M. Optimization of high-density cDNA-microarray protocols by 'design of experiments'. *Nucleic Acids Res.* **31**, e67 (2003).

166. Affymetrix. GeneChip Expression Analysis Technical Manual. *Affymetrix*, Santa Clara (2004).

167. R Development Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna (2004).

168. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).

169. 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-S104 (2002).

170. Affymetrix. Microarray Suite User Guide. *Affymetrix*, Santa Clara (2001).

171. Wu, Z., Irizarry, R., Gentleman, R., Martinez-Murillo, F. & Spencer, F. A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.* **99**, 909-917 (2004).

172. Yang, M. C., Ruan, Q. G., Yang, J. J., Eckenrode, S., Wu, S., McIndoe, R. A. & She, J. X. A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. *Physiol. Genomics* **7**, 45-53 (2001).

173. Barczak, A., Rodriguez, M. W., Hanspers, K., Koth, L. L., Tai, Y. C., Bolstad, B. M., Speed, T. P. & Erle, D. J. Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Res.* **13**, 1775-85 (2003).

174. Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. & Paules, R. S. Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* **8**, 625-37 (2001).

175. Smyth, G. K., Yang, Y. H. & Speed, T. Statistical issues in cDNA microarray data analysis. *Methods Mol. Biol.* **224**, 111-36 (2003).

176. Smyth, G. K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article 3 (2004).

177. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B.* **57**, 289-300 (1995).

178. Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M. & Marron, J. S. Adjustment of systematic microarray data biases. *Bioinformatics* **20**, 105-14 (2004).

179. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25-9 (2000).

180. Dennis, G., Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C. & Lempicki, R. A. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4**, P3 (2003).

181. Hosack, D. A., Dennis, G., Jr., Sherman, B. T., Lane, H. C. & Lempicki, R. A. Identifying biological themes within lists of genes with EASE. *Genome Biol.* **4**, R70 (2003).

182. Zhang, J., Carey, V. & Gentleman, R. An extensible application for assembling annotation for genomic data. *Bioinformatics* **19**, 155-6 (2003).

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

184. St Croix, B., Rago, C., Velculescu, V., Traverso, G., Romans, K. E. *et al.* Genes expressed in human tumor endothelium. *Science* **289**, 1197-202 (2000).

185. Luzzi, V., Mahadevappa, M., Raja, R., Warrington, J. A. & Watson, M. A. Accurate and reproducible gene expression profiles from laser capture microdissection, transcript amplification, and high density oligonucleotide microarray analysis. *J. Mol. Diagn.* **5**, 9-14 (2003).

186. 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).

187. Kenzelmann, M., Klaren, R., Hergenhahn, M., Bonrouhi, M., Grone, H. J., Schmid, W. & Schutz, G. High-accuracy amplification of nanogram total RNA amounts for gene profiling. *Genomics* **83**, 550-8 (2004).

188. Matz, M., Shagin, D., Bogdanova, E., Britanova, O., Lukyanov, S., Diatchenko, L. & Chenchik, A. Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic Acids Res.* **27**, 1558-60 (1999).

189. Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. & Speed, T. P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15 (2002).

190. Polacek, D. C., Passerini, A. G., Shi, C., Francesco, N. M., Manduchi, E. *et al.* Fidelity and enhanced sensitivity of differential transcription profiles following linear amplification of nanogram amounts of endothelial mRNA. *Physiol. Genomics* **13**, 147-56 (2003).

191. Feldman, A. L., Costouros, N. G., Wang, E., Qian, M., Marincola, F. M., Alexander, H. R. & Libutti, S. K. Advantages of mRNA amplification for microarray analysis. *Biotechniques* **33**, 906-12, 914 (2002).

192. Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G. & Gibson, G. The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nat. Genet.* **29**, 389-95 (2001).

193. Kokocinski, F., Wrobel, G., Hahn, M. & Lichter, P. QuickLIMS: facilitating the data management for DNA-microarray fabrication. *Bioinformatics* **19**, 283-4 (2003).

194. Knight, J. When the chips are down. *Nature* **410**, 860-1 (2001).

195. Yuen, T., Wurmbach, E., Pfeffer, R. L., Ebersole, B. J. & Sealfon, S. C. Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.* **30**, e48 (2002).

196. Mah, N., Thelin, A., Lu, T., Nikolaus, S., Kuhbacher, T. *et al.* A comparison of oligonucleotide and cDNA-based microarray systems. *Physiol. Genomics* **16**, 361-70 (2004).

197. Schlingemann, J., Hess, J., Wrobel, G., Breitenbach, U., Gebhardt, C. *et al.* Profile of gene expression induced by the tumour promotor TPA in murine epithelial cells. *Int. J. Cancer* **104**, 699-708 (2003).

198. Cox, W. G., Beaudet, M. P., Agnew, J. Y. & Ruth, J. L. Possible sources of dye-related signal correlation bias in two-color DNA microarray assays. *Anal. Biochem.* **331**, 243-54 (2004).

199. 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 Res.* **33**, e29 (2005).

200. Walker, G. T., Fraiser, M. S., Schram, J. L., Little, M. C., Nadeau, J. G. & Malinowski, D. P. Strand displacement amplification--an isothermal, in vitro DNA amplification technique. *Nucleic Acids Res.* **20**, 1691-6 (1992).

201. Walker, G. T., Little, M. C., Nadeau, J. G. & Shank, D. D. Isothermal in vitro amplification of DNA by a restriction enzyme/DNA polymerase system. *Proc. Natl. Acad. Sci. USA* **89**, 392-6 (1992).

202. Rajeevan, M. S., Dimulescu, I. M., Vernon, S. D., Verma, M. & Unger, E. R. Global amplification of sense RNA: a novel method to replicate and archive mRNA for gene expression analysis. *Genomics* **82**, 491-7 (2003).

203. Che, S. & Ginsberg, S. D. Amplification of RNA transcripts using terminal continuation. *Lab. Invest.* **84**, 131-7 (2004).

204. Bao, P., Frutos, A. G., Greef, C., Lahiri, J., Muller, U., Peterson, T. C., Warden, L. & Xie, X. High-sensitivity detection of DNA hybridization on microarrays using resonance light scattering. *Anal. Chem.* **74**, 1792-7 (2002).

205. Duveneck, G. L., Abel, A. P., Bopp, M. A., Kresbach, G. M. & Ehrat, M. Planar waveguides for ultra-high sensitivity of the analysis of nucleic acids. *Analytica Chimica Acta* **469**, 49-61 (2002).

206. Voeroes, J., de Paul, S. M., Textor, M., Abel, A. P., Kaufmann, E. & Ehrat, M. Polymer cushions to analyze genes and proteins. *BioWorld* **4**, 16-17 (2003).

207. Schneeweiss, A., Thuerigen, O., Toedt, G., Warnat, P., Hahn, M., Rudlowski, C., Benner, A., Brors, B., Sohn, C. & Lichter, P. Gene expression profiles predict pathologic complete response to preoperative chemotherapy with gemcitabine, epirubicin and docetaxel in primary breast cancer. *J Clin Oncol (Meeting Abstracts)* **23**, 2001- (2005).

208. Skerra, A. Phosphorothioate primers improve the amplification of DNA sequences by DNA polymerases with proofreading activity. *Nucleic Acids Res.* **20**, 3551-4 (1992).

209. Eckstein, F. Nucleoside phosphorothioates. *Annu. Rev. Biochem.* **54**, 367-402 (1985).

210. Sproat, B. S., Lamond, A. I., Beijer, B., Neuner, P. & Ryder, U. Highly efficient chemical synthesis of 2'-O-methyloligoribonucleotides and tetrabiotinylated derivatives; novel probes that are resistant to degradation by RNA or DNA specific nucleases. *Nucleic Acids Res.* **17**, 3373-86 (1989).

211. Tan, P. K., Downey, T. J., Spitznagel, E. L., Jr., Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M. & Cam, M. C. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* **31**, 5676-84 (2003).

212. Shippy, R., Sendera, T. J., Lockner, R., Palaniappan, C., Kaysser-Kranich, T., Watts, G. & Alsobrook, J. Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations. *BMC Genomics* **5**, 61 (2004).

213. Järvinen, A. K., Hautaniemi, S., Edgren, H., Auvinen, P., Saarela, J., Kallioniemi, O. P. & Monni, O. Are data from different gene expression microarray platforms comparable? *Genomics* **83**, 1164-8 (2004).

214. Marshall, E. Getting the noise out of gene arrays. *Science* **306**, 630-1 (2004).

215. Larkin, J. E., Frank, B. C., Gavras, H., Sultana, R. & Quackenbush, J. Independence and reproducibility across microarray platforms. *Nat. Methods* **2**, 337-44 (2005).

216. Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S. *et al.* Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2**, 345-50 (2005).

217. Bammler, T., Beyer, R. P., Bhattacharya, S., Boorman, G. A., Boyles, A. *et al.* Standardizing global gene expression analysis between laboratories and across platforms. *Nat. Methods* **2**, 351-6 (2005).

218. Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D. & Chinnaiyan, A. M. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.* **62**, 4427-33 (2002).

219. Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA* **100**, 8418-23 (2003).

220. Mitchell, S. A., Brown, K. M., Henry, M. M., Mintz, M., Catchpoole, D., LaFleur, B. & Stephan, D. A. Inter-platform comparability of microarrays in acute lymphoblastic leukemia. *BMC Genomics* **5**, 71 (2004).

221. Pruitt, K. D., Katz, K. S., Sicotte, H. & Maglott, D. R. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.* **16**, 44-7 (2000).

222. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137-40 (2001).

223. Mecham, B. H., Wetmore, D. Z., Szallasi, Z., Sadovsky, Y., Kohane, I. & Mariani, T. J. Increased measurement accuracy for sequence-verified microarray probes. *Physiol. Genomics* **18**, 308-15 (2004).

224. Baum, M., Bielau, S., Rittner, N., Schmid, K., Eggelbusch, K. *et al.* Validation of a novel, fully integrated and flexible microarray benchtop facility for gene expression profiling. *Nucleic Acids Res.* **31**, e151 (2003).

225. Boser, B. E., Guyon, I. M. & Vapnik, V. N. A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, Pittsburgh (1992).

226. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P. *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365-71 (2001).

6 Appendix

6.1 R Scripts for the Comparison of Microarray Platforms

6.1.1 Normalization of Affymetrix GeneChip Data (affynorm.R)

```
# introduction

cat("\n")
cat("affynorm.R © 2004 J. Schlingemann \n")
cat("\n")
cat("This script loads *.CEL files and performs MAS 5.0 normalization and \n")
cat("vsn variance stabilization. You need the R-packages affy and vsn \n")
cat("(1.2.0 ++) installed to run this script. You also need to install the \n")
cat("CDF-package corresponding to the type of array You used. \n")
cat("All required packages are available at http://www.bioconductor.org/ \n")
cat("\n")


# load necessary packages

library(affy)
library(vsn)


# define functions

choose.option = function (choices, title = "") {
   nc <- length(choices)
   cat("\n")
   cat(title, "\n")
   for (i in seq(length = nc)) cat(i, ":", choices[i], " \n", sep = "")
   repeat {
      option <- .Internal(menu(as.character(choices)))
      if (option <= nc && option !=0) {
         return(option)
      } else {
         cat("Please enter an item from the menu !! \n")
      }
   }
}


# define variables

PSEP = .Platform$file.sep


# check if you can go on or load new data

go.on = "2"
if (exists("res.dir")){
   if (file.access(res.dir,0)!=-1) {
      go.on = choose.option(c("yes","no"),
              title=paste("Is '",res.dir,"' the correct data path ? "))
   } else {
      cat("Data path no longer valid !! \n")
   }
}

while (go.on != "1") {
   res.dir <- "xxx"
   while (res.dir == "xxx") {
      cat("\n")
      res.dir = readline(prompt="Specify the path to the data files (*.CEL): ")
      setwd(res.dir)
      cat("\n")
      if (file.access(res.dir,0)==-1) {
         res.dir <- "xxx"
         cat("This is not a valid path !! \n")
      }
   }
  go.on = choose.option(c("yes","no"),
          title=paste("Is '",res.dir,"' the correct data path ? "))
  if (exists("cel")) {
      rm(joinedReplicas)
  }
 }


# load *.CEL files

```

```
cel <- ReadAffy()
cdf <- paste(annotation(cel),"cdf", sep="")
library(cdf, character.only=TRUE)

cat("\n")
cat("The following *.CEL files will be processed: \n")
cat("\n")
cels <- list.celfiles()
print(cels)
cat("\n")
dims <- dim(pm(cel))
cat("These are", dims[2], "chips of the type", cdfName(cel), "with",
    dims[1], "perfect match probes each... \n")
cat("\n")


# get MAS 5.0 expression, calls and p-values

try(memory.limit(size = 2000))
MAS5.data <- mas5(cel)
MAS5.es <- exprs(MAS5.data)
colnames(MAS5.es)<-cels
write.table(MAS5.es, file = "MAS5.es.csv", sep = ",", col.names = NA)
MAS5.calls <- mas5calls(cel)
calls <- exprs(MAS5.calls)
pvalues <- se.exprs(MAS5.calls)
colnames(calls)<-cels
colnames(pvalues)<-cels
write.table(calls, file = "calls.csv", sep = ",", col.names = NA)
write.table(pvalues, file = "pvalues.csv", sep = ",", col.names = NA)
cat("Extracted MAS 5.0 normalized expression, calls and p-values \n")
cat("\n")


# start vsn

normalize.AffyBatch.methods <- c(normalize.AffyBatch.methods, "vsn")

data.vsn <-expresso(cel,
        pmcorrect.method = "pmonly",
        bgcorrect.method = "none",
        normalize.method = "vsn",
        summary.method   = "medianpolish", verbose = TRUE)

log2.vsn.es <-exprs(data.vsn)
log2.vsn.scaled.es <- 1.5*exprs(data.vsn)-8
colnames(log2.vsn.scaled.es) <-cels
lin.vsn.scaled.es <- 2^log2.vsn.scaled.es

write.table(log2.vsn.scaled.es, file = "log2.vsn.scaled.es.csv", sep = ",",
            col.names = NA)
write.table(lin.vsn.scaled.es, file = "lin.vsn.scaled.es.csv", sep = ",",
            col.names = NA)

vsndims <- dim(exprs(data.vsn))
cat("vsn normalization finished !  \n")
cat("Expression values were extracted both lin- and log2-scaled \n")
cat("Your data contains", vsndims[1], "normalized probe sets for",
    vsndims[2], "chips of the type", cdfName(cel), "! \n")
cat("\n")

for (i in 1:dims[2]) {
    combi <-cbind(lin.vsn.scaled.es[,i],log2.vsn.scaled.es[,i],MAS5.es[,i],
            calls[,i],pvalues[,i])
    colnames(combi) = c("VALUE", "VSN_LOG2", "MAS5_VALUE",
                        "ABS_CALL", "DETECTION_P_VALUE")
    write.table(combi, file = paste(cels[i], "_summary.csv"), sep = ",",
                col.names = NA)
}

cat("Summary tables with all relevant data for have been created !!\n")
cat("All tasks successfully completed !!\n")


# the end
```

6.1.2 Matching of Oligonucleotide Probe Sequences (alignment.R)

```
# introduction

cat("\n")
cat("alignment.R © 2004 J. Schlingemann \n")
cat("\n")
cat("This script will query various databases to collect available information \n")
cat("about two chip platforms to be compared, based on Genebank \n")
cat("accession numbers as primary identifiers.\n")
cat("Probes or probe sets common to both platforms are determined \n")
cat("via affiliation to UniGene clusters. \n")
cat("You need the R packages annotate and Annbuilder. \n")
cat("All required packages are available at http://www.bioconductor.org/ \n")
cat("\n")


# increase memory size

try(memory.limit(size = 2000))
gc(verbose=FALSE)
cat("\n")
try(cat("Memory limit:     ",memory.limit(size = NA)/1048576,"MB \n"))
try(cat("Currently in use:",round(memory.size(max = FALSE)/1048576,2),"MB \n"))
try(cat("Available:       ",round((memory.limit(size = NA)/1048576)-
                            (memory.size(max = FALSE)/1048576),2),"MB \n"))
cat("\n")
cat("\n")


# load packages and Affy data

library(annotate)
library(AnnBuilder)


# define variables

URLs <- getSrcUrl(src = "ALL", organism = "human")
URLs[2] <- "http://hgdownload.cse.ucsc.edu/goldenPath/hg17/database/"
baseType <- "gb"
setwd("E:\\Stauber all gpr\\Affy-Daten\\Affy metadata\\new\\")
resDir <- getwd()


# wait for KEGG...

japs <- "1"
attpt <- 0
while (japs == "1") {
   attpt <- attpt + 1
   cat("\n")
   cat("Looking for KEGG server (ping)... Try", attpt, "... \n")
   japs <- try(shell("ping www.genome.ad.jp", intern=FALSE, wait=TRUE,
               translate=FALSE, mustWork=FALSE, invisible=TRUE), silent=TRUE)
}
cat("\n")
cat("KEGG server finally online... !!! \n")


# get UNIGENE IDs for Operon

cat("Collecting data for Operon platform...\n")
cat("\n")
ABPkgBuilder(baseName = "Operon3.txt", srcUrls = URLs, baseMapType = baseType,
             pkgName = "Operon_human_2.1_21k", pkgPath = resDir,
             organism = "human", version = "04.08.19", makeXML = FALSE,
             author = list(author = "Joerg Schlingemann",
             maintainer = "j.schlingemann@dkfz.de"), fromWeb = TRUE)


# wait for KEGG...

japs <- "1"
attpt <- 0
while (japs == "1") {
   attpt <- attpt + 1
   cat("\n")
   cat("Looking for KEGG server (ping)... Try", attpt, "... \n")
   japs <- try(shell("ping www.genome.ad.jp", intern=FALSE, wait=TRUE,
               translate=FALSE, mustWork=FALSE, invisible = TRUE), silent=TRUE)
}
cat("\n")
cat("KEGG server finally online... !!! \n")


# get UNIGENE IDs for Affy

cat("Collecting data for Affy platform...\n")
cat("\n")
ABPkgBuilder(baseName = "Affy1.txt", srcUrls = URLs, baseMapType = baseType,
             otherSrc = "Affy3.txt", pkgName = "Affy_HG_U133A",
             pkgPath = resDir, organism = "human", version = "04.08.19",
             makeXML = FALSE, author = list(author = "Joerg Schlingemann",
             maintainer = "j.schlingemann@dkfz.de"), fromWeb = TRUE)


# data processing

cat("\n")
cat("Data collection finished! Processing comparison... \n")

setwd("E:\\Stauber all gpr\\Affy-Daten\\Affy metadata\\new\\Operon_human_2.1_21k\\")
temp <- data()
data(Operon_human_2.1_21kUNIGENE, package = "Operon_human_2.1_21k",
     lib.loc = resDir)
res.Operon <- as.list(Operon_human_2.1_21kUNIGENE)
res.Operon <- as.matrix(unlist(res.Operon))
res.Operon <- as.matrix(res.Operon[!is.na(res.Operon)])
res.Operon <- cbind(rownames(res.Operon),res.Operon[,1])
rownames(res.Operon) <- res.Operon[,2]
colnames(res.Operon) <- c("Operon_ID", "UniGene_cluster")

setwd("E:\\Stauber all gpr\\Affy-Daten\\Affy metadata\\new\\Affy_HG_U133A\\")
temp <- data()
data(Affy_HG_U133AUNIGENE, package = "Affy_HG_U133A", lib.loc = resDir)
res.Affy <- as.list(Affy_HG_U133AUNIGENE)
res.Affy <- as.matrix(unlist(res.Affy))
res.Affy <- as.matrix(res.Affy[!is.na(res.Affy)])
res.Affy <- cbind(rownames(res.Affy),res.Affy[,1])
rownames(res.Affy) <- res.Affy[,2]
colnames(res.Affy) <- c("Affy_ID", "UniGene_cluster")

consensus <- intersect(res.Affy[,2],res.Operon[,2])
final.table <- matrix(NA,length(consensus),1)
final.table[,1] <- consensus
final.table <- cbind(final.table,res.Operon[consensus,1,drop=FALSE],
                     res.Affy[consensus,1,drop=FALSE])
final.table.csv <- cbind(final.table[,2], final.table[,3])
rownames(final.table) <- NULL
colnames(final.table) <- c("UniGene_cluster","Operon_ID","Affy_ID")
colnames(final.table.csv) <- c("Operon_ID","Affy_ID")
setwd("E:\\Stauber all gpr\\Affy-Daten\\Affy metadata\\new\\")
write.table(consensus, file = "Unigene.txt", sep ="\t", quote = FALSE,
            row.names = FALSE, col.names = "Unigene ID")
write.table(final.table, file = "FINAL.txt", sep ="\t", quote = FALSE,
            row.names = FALSE)
write.table(final.table.csv, file = "FINAL.csv", sep = ",",
            col.names = NA, quote = FALSE)

aL <- length(res.Affy[,1])
oL <- length(res.Operon[,1])
cL <- length(consensus)

cat("Found", aL, "Unigene IDs for Affy_HG-U133A and",
    oL, "Unigene IDs for Operon_human_2.1_21k \n")
cat("The intersection contains", cL, "Unigene IDs !!")


# the end
```

6.1.3 Comparison of Operon and Affymetrix GeneChip Data (compare.R)

```
# introduction

cat("\n")
cat("\n")
cat("compare.R © 2004 J. Schlingemann \n")
cat("This script calculates correlations of Affy and Operon datasets \n")
cat("\n")


# variables

PSEP = .Platform$file.sep
plotRange = c(-6,6)


# needed functions

dir.check <- function (path, create = "TRUE", message = "") {
   if (file.access(path,0)==-1) {
      if (create == "TRUE") {
         dir.create(path)
         cat("\n")
         cat("Created directory '", path, "' \n")
         cat(message, "\n")
      } else {
         cat("\n")
         cat("'",path,"' not found !! \n")
      }
   } else {
      cat("\n")
      cat("Found directory '", path, "' \n")
      cat(message, "\n")
   }
}

file.check <- function (path, file, vari = "x", read = "TRUE",
                        error.message = "", sep = "\t",
                        header = TRUE, row.namez = 1) {
   if (!exists(vari)) {
      if (file.access(paste(path,PSEP,file,sep=""),0)!=-1) {
         if (read == "TRUE") {
            vari =  read.delim(paste(path,PSEP,file,sep=""),sep=sep,
                               header=header, row.names=row.namez)
            cat("\n")
            cat(file,"has been loaded! \n")
            return(vari)
         } else {
            cat("\n")
            cat(file,"was found! \n")
         }
      } else {
         cat("\n")
         cat(path,PSEP,file," not found! \n")
         cat(error.message, "\n")
      }
   } else {
      cat("\n")
      cat(file,"was still in memory \n")
      vari <- get(vari)
      return(vari)
   }
}

choose.option = function (choices, title = "") {
   nc <- length(choices)
   cat("\n")
   cat(title, "\n")
   for (i in seq(length = nc)) cat(i, ":", choices[i], " \n", sep = "")
   repeat {
      option <- .Internal(menu(as.character(choices)))
      if (option <= nc && option !=0) {
         return(option)
      } else {
         cat("Please enter an item from the menu !! \n")
      }
   }
```

```
}


regLineLin <- function(x,y) {
   fitout <- lsfit(x, y, intercept = TRUE)
   lines(x, fitout$coefficients[1] + fitout$coefficients[2] * x, lty="solid" ,
         lwd=1, col="red")
}


# check if you can go on or load new data

go.on = "2"
if (exists("res.dir")){
   if (file.access(res.dir,0)!=-1) {
      go.on = choose.option(c("yes","no"), title=paste("Is '",res.dir,
                            "' the correct data path ? "))
   } else {
      cat("Data path no longer valid !! \n")
   }
}

while (go.on != "1") {
   res.dir <- "xxx"
   while (res.dir == "xxx") {
      res.dir = readline(prompt="Specify the path to the data files: ")
      cat("\n")
      if (file.access(res.dir,0)==-1) {
         res.dir <- "xxx"
         cat("This is not a valid path !! \n")
      }
   }
  go.on = choose.option(c("yes","no"), title=paste("Is '",res.dir,
                        "' the correct data path ? "))
  if (exists("affyRaw")) {
      rm(joinedReplicas)
  }
  if (exists("operonRatios")) {
      rm(colorswitchData)
  }
}


## load data if necessary

setwd(res.dir)
affyRaw <- file.check(res.dir, "log2.vsn.scaled.es.csv", "affyRaw",
                      error.message = "Affy data missing !!", header = TRUE,
                      sep = ",", row.namez=1)
operonRatios <- file.check(res.dir, "ColorswitchData.txt", "operonRatios",
                           error.message = "Operon data missing !!")


# calculate Affy ratios

tumorChips <- list()
controlChips<- list()
colnames.tumor=list()
colnames.control=list()


if(!exists("affyRaw") || !exists("operonRatios")) {
   cat("you can't perform analyses as necessary data files are missing !! \n")
   start.analysis <- FALSE
   go.on <- FALSE
}

if(exists("affyRaw") && exists("operonRatios")) {
   start.analysis <- TRUE
   cat("Please assign corresponding tumor- and control-chips (Affy) \n")
   i <- 0
   while (i < 1/2 * length(colnames(affyRaw))) {
      i <- i+1
      tumorChips[i] <- choose.option(colnames(affyRaw),
                                     paste("Please give # of tumor chip ",i,
                                     " !!", sep=""))
      cat("You chose",colnames(affyRaw)[as.integer(tumorChips[i])], "\n")
      colnames.tumor[i] <- colnames(affyRaw)[as.integer(tumorChips[i])]
      controlChips[i] <- choose.option(colnames(affyRaw),
                                       paste("Please give # of control chip ",i,
                                       " !!", sep=""))
      cat("You chose",colnames(affyRaw)[as.integer(controlChips[i])], "\n")
      colnames.control[i] <- colnames(affyRaw)[as.integer(controlChips[i])]
   }
}
colnames.final <- paste(colnames.tumor," vs ", colnames.control,sep="")

if (start.analysis == "TRUE") {
   affyRatios = matrix(0,dim(affyRaw)[1],length(colnames.final))
   colnames(affyRatios) <- colnames.final
   rownames(affyRatios) <- rownames(affyRaw)

   for (j in 1:length(tumorChips)) {
      tumorChip <- as.integer(tumorChips[j])
      controlChip <- as.integer(controlChips[j])
      affyRatios[,j] <- affyRaw[,tumorChip] - affyRaw[,controlChip]
   }
}


# assign Operon-Experiments

cat("Please assign corresponding Affy- and Operon-experiments \n")

affyExp <- list()
operonExp <- list()

i <- 0
while (i < length(colnames(affyRatios))) {
   i <- i+1
   affyExp[i] <- choose.option(colnames(affyRatios),
                               paste("Please give # of Affy-experiment ",i,
                               " !!", sep=""))
   cat("You chose",colnames(affyRatios)[as.integer(affyExp[i])], "\n")
   operonExp[i] <- choose.option(colnames(operonRatios),
                                 paste("Please give # of Operon-experiment ",i,
                                 " !!", sep=""))
   cat("You chose",colnames(operonRatios)[as.integer(operonExp[i])], "\n")
}

# replace Operon_IDs with ChipYard_IDs

chipyardID <- as.matrix(file.check(res.dir, "ChipYard_IDs.txt", "chipyardID",
                        error.message = "ChipYard data missing !!",
                        row.namez=NULL))
rownames(chipyardID) <- chipyardID[,2]

replaceID <- as.matrix(file.check(res.dir, "Operon_IDs.txt", "replaceID",
                       error.message = "Operon data missing !!",
                       row.namez=NULL))
rownames(replaceID) <- replaceID[,1]

combi <- as.matrix(intersect(replaceID[,1],chipyardID[,2]))
combi <- cbind(replaceID[combi,2,drop=FALSE],combi,chipyardID[combi,1,drop=FALSE])
rownames(combi) <- combi[,1]
colnames(combi) <- c("Operon_ID","CloneY_ID","ChipYard_ID")

refseq.data <- file.check(res.dir, "FINAL_refseq.csv", "refseq.data",
                          error.message = "RefSeq data missing !!", header = TRUE,
                          sep = ",", row.namez=1)
refseq.data[,3] <- rownames(refseq.data)
rownames(refseq.data) <- refseq.data[,1]
colnames(refseq.data)[3] <- "RefSeq"
refseq.data <- as.matrix(refseq.data)

unigene.data <- file.check(res.dir, "FINAL_unigene.csv", "unigene.data",
                           error.message = "Unigene data missing !!",
                           header = TRUE,
                           sep = ",", row.namez=1)
unigene.data[,3] <- rownames(unigene.data)
rownames(unigene.data) <- unigene.data[,1]
colnames(unigene.data)[3] <- "UniGene"
```

```
unigene.data <- as.matrix(unigene.data)

refseq.rep <- as.matrix(intersect(combi[,1],refseq.data[,1]))
refseq.rep <- cbind(refseq.rep,combi[refseq.rep,2,drop=FALSE],combi[refseq.rep,3,
                    drop=FALSE],refseq.data[refseq.rep,2,drop=FALSE],
                    refseq.data[refseq.rep,3,drop=FALSE])
colnames(refseq.rep) <- c("Operon_ID", "CloneY_ID", "ChipYard_ID","Affy_ID",
                          "RefSeq")

unigene.rep <- as.matrix(intersect(combi[,1],unigene.data[,1]))
unigene.rep <- cbind(unigene.rep,combi[unigene.rep,1,drop=FALSE],
                     combi[unigene.rep,3,drop=FALSE],unigene.data[unigene.rep,2,
                     drop=FALSE],unigene.data[unigene.rep,3,drop=FALSE])
colnames(unigene.rep) <- c("Operon_ID", "CloneY_ID", "ChipYard_ID","Affy_ID",
                           "UniGene")


# filter Affy-data for probe sets common to both platforms

rownames(refseq.rep) <- refseq.rep[,3]
rownames(unigene.rep) <- unigene.rep[,3]
operon.ratios.refseq.filtered <- as.matrix(cbind(operonRatios
                                           [rownames(refseq.rep),],
                                           refseq.rep[,]))
operon.ratios.unigene.filtered <- as.matrix(cbind(operonRatios
                                            [rownames(unigene.rep),],
                                            unigene.rep[,]))


# filter Affy-data for probe sets common to both platforms

rownames(refseq.rep) <- refseq.rep[,4]
rownames(unigene.rep) <- unigene.rep[,4]
affy.ratios.refseq.filtered <- as.matrix(cbind(affyRatios[rownames(refseq.rep),],
                                         refseq.rep[,]))
affy.ratios.unigene.filtered <- as.matrix(cbind(affyRatios[rownames(unigene.rep),],
                                          unigene.rep[,]))


# correlate Affy- and Operon-data

refseq.dir <- paste(res.dir,PSEP,"refseq_data",sep="")
dir.check(refseq.dir, message = "All RefSeq results will be saved to that path!")

unigene.dir <- paste(res.dir,PSEP,"unigene_data",sep="")
dir.check(unigene.dir, message="All UniGene data will be saved to that path!")

correl.refseq = matrix(0,4,length(colnames(affyRatios)))
colnames(correl.refseq) = colnames(affyRatios)
rownames(correl.refseq) = c("correlation","slope","intercept","equation")
correl.unigene = matrix(0,4,length(colnames(affyRatios)))
colnames(correl.unigene) = colnames(affyRatios)
rownames(correl.unigene) = c("correlation","slope","intercept","equation")

for (j in 1:length(affyExp)) {

   affy.cor <- as.integer(affyExp[j])
   operon.cor <- as.integer(operonExp[j])

   x1 = affy.ratios.refseq.filtered[,affy.cor]
   x2 = affy.ratios.unigene.filtered[,affy.cor]
   y1 = operon.ratios.refseq.filtered[,operon.cor]
   y2 = operon.ratios.unigene.filtered[,operon.cor]

   correl.refseq[1,j] = cor(x1,y1,use="pairwise.complete.obs")
   correl.unigene[1,j] = cor(x2,y2,use="pairwise.complete.obs")
   R1 <- round(as.double(correl.refseq[1,j]),3)
   R2 <- round(as.double(correl.unigene[1,j]),3)
   fit1 <- lsfit(x1, y1, intercept = TRUE)
   fit2 <- lsfit(x2, y2, intercept = TRUE)
   slope1 <- round(fit1$coefficients[2],3)
   intercept1 <- round(fit1$coefficients[1],3)
   fit2 <- lsfit(x2, y2, intercept = TRUE)
   slope2 <- round(fit2$coefficients[2],3)
   intercept2 <- round(fit2$coefficients[1],3)
   correl.refseq[2,j] <- slope1
   correl.unigene[2,j] <- slope2
   correl.refseq[3,j] <- intercept1
   correl.unigene[3,j] <- intercept2
   if(intercept1 < 0) operator = "-"  else operator = "+"
   if(intercept1 < 0) intercept1 <- abs(intercept1)
   if(intercept2 < 0) operator = "-"  else operator = "+"
   if(intercept2 < 0) intercept2 <- abs(intercept2)
   equation1 <- paste("Y=",slope1,"*x", operator ,intercept1 ,sep="")
   equation2 <- paste("Y=",slope2,"*x", operator ,intercept2 ,sep="")
   correl.refseq[4,j] <- equation1
   correl.unigene[4,j] <- equation2

   cat("\n")
   cat("Analyzing ",colnames(affy.ratios.refseq.filtered)[affy.cor]," vs ",
       colnames(operon.ratios.refseq.filtered)[operon.cor],"\n",sep="")
   print(paste("R (refseq):  ",R1,sep=""))
   print(paste("R (unigene): ",R2,sep=""))
   print(paste("Equation (refseq):  ",equation1,sep=""))
   print(paste("Equation (unigene): ",equation2,sep=""))

   bmp(file=paste(refseq.dir,PSEP,colnames(correl.refseq)[j],".bmp", sep=""),
       width = 688, height = 688)
   plot(x1, y1, , xlim=range(plotRange), ylim=range(plotRange), pch="°",
        col="darkblue", xlab=colnames(affy.ratios.refseq.filtered)[affy.cor],
        ylab=colnames(operon.ratios.refseq.filtered)[operon.cor],
        main=paste(colnames(correl.refseq)[j],"  R = ",R1, "  ", equation1,
        sep=""))
   regLineLin(x1[order(x1)], y1[order(x1)])
   dev.off()

   bmp(file=paste(unigene.dir,PSEP,colnames(correl.unigene)[j],".bmp", sep=""),
       width = 688, height = 688)
   plot(x2, y2, , xlim=range(plotRange), ylim=range(plotRange), pch="°",
        col="darkblue", xlab=colnames(affy.ratios.unigene.filtered)[affy.cor],
        ylab=colnames(operon.ratios.unigene.filtered)[operon.cor],
        main=paste(colnames(correl.unigene)[j],"  R = ",R2, "  ",
             equation2, sep=""))
   regLineLin(x1[order(x1)], y1[order(x1)])
   dev.off()

}
   write.table(correl.refseq, paste(refseq.dir,PSEP,
               "correlation_refseq.txt",sep=""),col.names=NA)
   write.table(correl.unigene, paste(unigene.dir,PSEP,
               "correlation_unigene.txt",sep=""),col.names=NA)
   cat("\n")
   cat("All tasks successfully completed !!!!!!! \n",sep="")
   cat("\n")


# the end
```

Science is a wonderful thing if one does not have to earn one's living at it.

ALBERT EINSTEIN