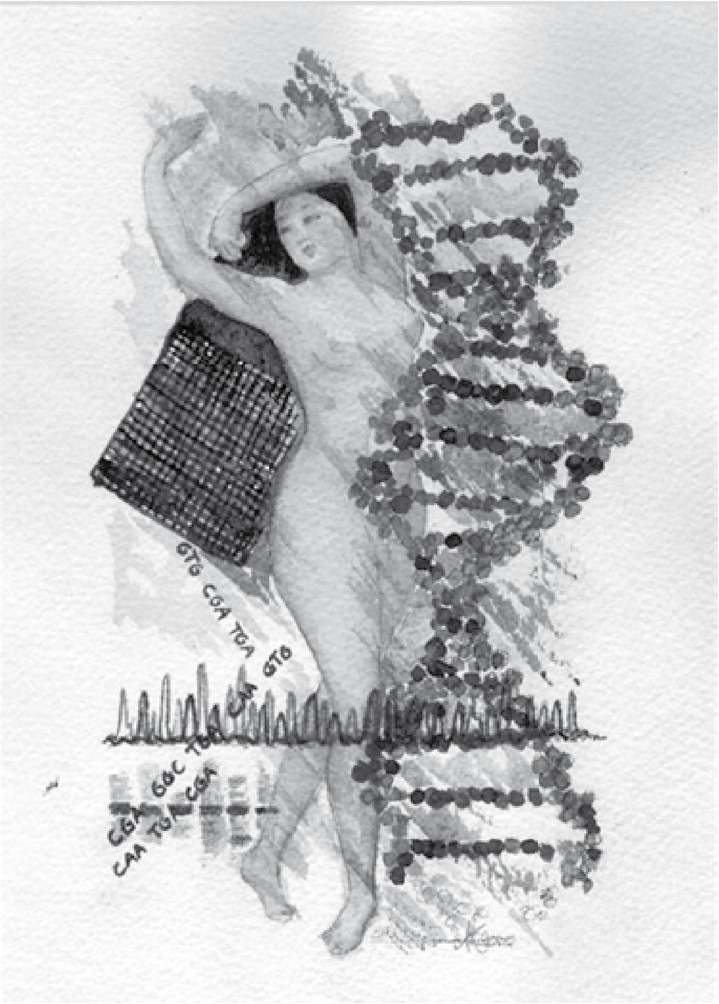
Microarray Bioinformatics and Applications in Oncology



Justine Kate Peeters

The work in this thesis was performed in the Department of Bioinformatics, Erasmus University Medical Center, Rotterdam, The Netherlands.

The printing of this thesis was financially supported by

J.E. Juriaanse Stichting,
Rotterdam, The Netherlands

The Netherlands Bioinformatics Center (NBIC)

Affymetrix Europe

Cover: painted by Anne Karin Pettersen Arvola, with compliments of Therese Sorlie
Print: Printpartners Ipskamp, Enschede www.ppi.nl
Lay-out: Legatron Electronic Publishing, Rotterdam

ISBN: 978-90-9023048-1

Microarray Bioinformatics and Applications in Oncology

Toepassingen van bioinformatica en microarray's in oncologie

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
Prof. dr. S.W.J. Lamberts
en volgens besluit van het college voor Promoties

De openbare verdediging zal plaatsvinden op
Woensdag 11 juni 2008 om 13.45 uur

door

Justine Kate Peeters

geboren te Melbourne, Australia

ERASMUS UNIVERSITEIT ROTTERDAM

Promotiecommissie

Promotor:	Prof.dr. P.J. van der Spek
Co-promotor:	Dr. A.E.M. Schutte
Overige leden:	Prof.dr. F.G. Grosveld
	Prof.dr. P.A.E. Sillevis Smitt
	Prof.dr. L.H.J. Looijenga

This thesis is dedicated to my
Oma 'Wilhelmena Johanna Peeters' (1918-2004).

Table of Contents

Appendices

*For all large supplementary data files, tables and colour versions of figures and PDFs. please refer to the website **http://www-bioinf.erasmusmc.nl/thesis_peeters.**

Chapter 1

Introduction to Microarray Bioinformatics

1.0 Introduction to microarray bioformatics

'Bioinformatics' is one of the newest fields of biological research, and should be viewed broadly as the use of mathematical, statistical, and computational methods for the processing and analysis of biologic data [1]. Over the last decade, the rapid growth of information and technology in both 'genomics[1]' and 'omics[2]' era's has been overwhelming for the laboratory scientists to process experimental results. Traditional gene-by-gene approaches in research are insufficient to meet the growth and demand of biological research in understanding the true biology. The massive amounts of data generated by new technologies as genomic sequencing and microarray chips make the management of data and the integration of multiple platforms of high importance; this is then followed by data analysis and interpretation to achieve biological understanding and therapeutic progress. Global views of analyzing the magnitude of information are necessary and traditional approaches to labwork have steadily been changing towards a bioinformatic era. Research is moving from being restricted to a laboratory environment to working with computers in a 'virtual lab' environment.

1.1 The age of science and the evolution into the 'Omics' era

The invention of the Polymerase Chain Reaction (PCR) technique was a major milestone in molecular research that transformed and revolutionized current research and diagnostics [3]. Since the introduction of PCR, linkage analysis and mutation screening became easier and the rate of identified disease genes increased dramatically. Another milestone that influenced the rate of discovery was the Human Genome project [4,5]. The main aim of this project was to create a detailed physical map of the human genome. Having genomic sequences available made the identification of disease related mutations in Mendelian single gene disorders an easier task; however, the mapping of complex diseases such as diabetes and cancer remains a challenge. The successful completion of the Human Genome Project, as well as the sequencing of the genomes of many other species, has generated a large amount of freely available information, opening the

[1] The study of all of the nucleotide sequences, including structural genes, regulatory sequences, and non-coding DNA segments, in the chromosomes of an organism. Also seen as a branch of biotechnology concerned with applying the techniques of genetics and molecular biology to the genetic mapping and DNA sequencing of sets of genes or the complete genomes of selected organisms using high-speed methods, with organizing the results in databases, and with applications of the data (as in medicine or biology) 2. http://www.dictionary.com. [cited.

[2] 'Omics' refers to a field of study in biology ending in the suffix -omics such as genomics or proteomics. The related neologism 'omes' addresses the objects of study of such fields, such as the genome or proteome respectively 2. Ibid. [cited..

door to a post-genome era where an 'e-science' approach can allow *in silico* research and further mining of available data.

One of the most challenging objectives of the post-genome era is to understand the complex genome and its interacting products. Various techniques to understand gene expression have been developed where it is important to remember that RNA (ribonucleic acid) expression does not directly reflect equal protein levels. Approaches such as subtractive hybridization of cDNA libraries and differential display have been used with some success; however, these techniques are laborious and are not suited for global gene analysis. Techniques such as SAGE (for serial analysis of gene expression) have been used to quantify the expression of multiple genes [6]. Surpassing this technique with its profiling capacity is the microarray technology. Whilst there are multiple platforms of microarray technology used in this thesis, the GeneChip gene expression platform of Affymetrix will be described in detail. All applications of microarray technology presented in the following chapters have been performed using these gene expression GeneChips, and alternate Affymetrix platforms have been further described in chapter 1.1.2.

1.2 Microarray technology

Within the human body, thousands of genes and their products (i.e., RNA and proteins) function in a complicated web and are orchestrated both temporally and spatially. Due to this complexity, the traditional gene-by-gene approach is not powerful enough to define a global view of cellular function. The microarray technology has been designed to measure the activity of gene expression, from the complete genome in a single experiment. Genetic information contained in DNA (deoxyribonucleic acid), is consistent with cells of one individual, and a source of variation within and between species. Gene expression however, varies from tissue to tissue depending on the cell types present in the tissue and its condition (e.g. disease state), giving a source of variation within and between organisms. The ability to measure expression of multiple genes provides the researcher with a method to elucidate the mechanisms behind this process.

Within a couple of years, gene expression microarray technology has developed from profiling a selection of genes on a membrane filter to all mRNA transcripts simultaneously (known as a 'transcriptome') on a solid surface [7,8]. Current microarrays may have up to tens of thousands of unique DNA sequences spotted to it (Figure 1). The underlying principle of the microarray technology is base-pair hybridization. When using a gene expression microarray, one extracts mRNA from the sample of interest, under experimental procedures makes complimentary RNA from this, labels the cRNA with a fluorescent dye and hybridizes it to a glass slide with the spotted DNA sequences. Sequence specific hybridization ensures that the mRNA mostly binds to the DNA from which it is transcribed. Color intensities for each gene can be quantified from a laser scanner using specialized software for scanning microarrays, which can be used for statistical analysis.

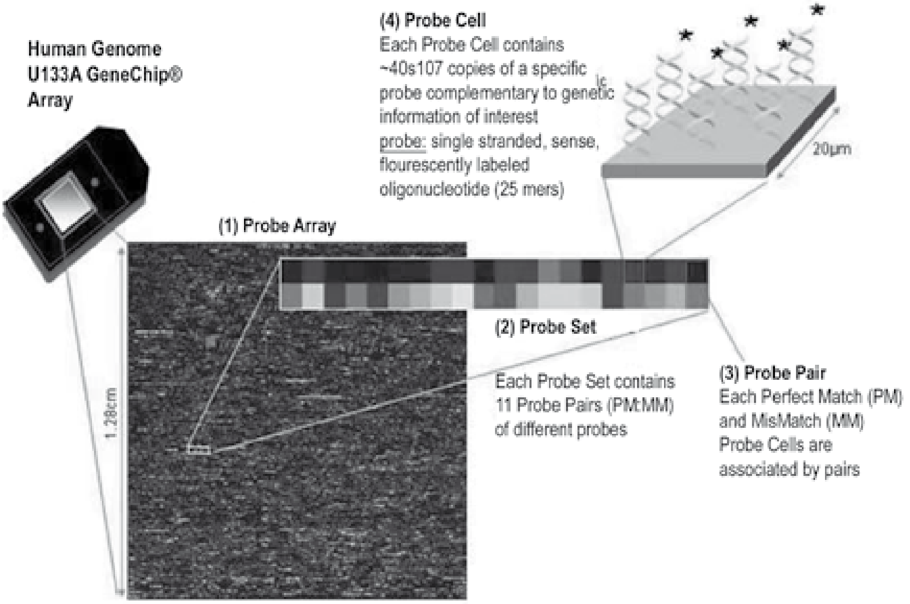


Figure 1: Affymetrix GeneChips Affymetrix gene chips contain more than 400,000 features per chip. Each feature contains millions of identical oligonucleotide probes. These probes are synthesized directly on the chip. http://www.weizmann.ac.il/home/ligivol/research_interests.html

Microarray studies with research questions often aim at increasing the knowledge and understanding of gene functions. This is usually done by investigating genes whose expression levels are correlated with experimental conditions or important phenotypes. This can also involve the identification of biological pathways affected by the expression levels of a particular gene, but also in the aspects of drug targets and drug sensitivity in therapy development. Microarray research can also address questions relating to the phenotype of a particular disease. These studies aim at understanding discovering which biological processes are related to certain aspects or subtypes of disease, or identification of disease–specific molecular markers. Such information can be of great value in unraveling the complex biological mechanisms involved in a disease. A third direction of microarray research is driven by research questions that relate to the patient. Answers to such questions could potentially improve diagnosis and treatment of disease. Microarrays can be especially useful in prognosis, as future events which are not yet clinically detectable may be predicted through measurement of gene expression activity (such as metastasis in cancer).

1.3 Microarray experiment design

To complete a successful microarray experiment, several factors must be addressed, including the experimental design. The design defines the statistical power of the analysis, which is dependent on the number of samples included in an experiment. When biological material is scarce, RNA amplification procedures are available or options to pool RNA samples, but this must be done with caution so as not to introduce bias and variation [9-12]. Identifying sources of variation and taking their relative magnitudes into account is vital in the optimal planning of a microarray and development of quality control procedures [13]. It is very important to consider this not only in the experimental design but also in the statistical analysis and interpretation of controls. Knowing what type of information is attainable from a microarray can make for better and more efficient hypothesis and experimental design.

In general, microarrays provide a large amount of information, and are used to examine a wide range of biological questions. These are:

– Inference of which genes are highly expressed, not expressed, or in general, the relative expression level of all genes. I.e. genes x, y, z are highly expressed under condition W, while genes a, b, c are not expressed [14,15].
– Observing gene expression levels at different time points following a stimulus, allows one to infer how they change with time. I.e. as the disease progresses, Gene x's expression decreases [16,17].
– Comparing the expression level of two genes can provide assumptions about their similar or different expression patterns. I.e. gene's with similar expression patterns might be functionally related or working in the same pathway as co-expressed genes [18-20].
– Comparing gene expression levels under different conditions, allows one to infer which gene's expression levels are affected. I.e. gene x is significantly higher expressed in

disease cells than normal cells; hence gene x could potentially serve as a marker of the disease [21,22].

1.3.1 Biological variation

Determining biological variation in gene expression is the goal of microarray analysis, but unwanted biological variation can sometimes mask the question being asked. Biological variation is known to exceed the technical variation in an experiment. Such biological variation includes collection of samples at different times of the day. Harmer *et al.* 2000 [23] have shown that dramatic transcriptional differences can occur at different times of the day solely due to circadian rhythm. Keeping RNA extraction procedures consistent and making sure that they yield good quality-RNA is of utmost importance. The treatment of tissue before the extraction of RNA is also important. Fresh frozen tissue must have been consistently handled. Tissue that has not been immediately frozen may have degradation of RNA species as will the freeze thawing of samples. Culturing cells under the same conditions is important in an experiment. A change of nutrient in the media can account for transcriptional changes in metabolic pathways. A change in temperature of the incubator can have an effect on the transcription of heat shock genes for example. Biological variation can also be as simple as the differences between man and woman. Moreover, differences in metabolism, can give rise to gene expression changes that are irrelevant to the hypothesis. Considering all the potential variables as part of the experimental design can alleviate biological variation.

1.3.2 Technical and system variation

Technical and system variation can arise at any time when performing the entire experimental procedure, but is easier to control than biological variation. Having two technicians perform the RNA extraction and hybridization assays can create variation in gene expression as the two may have different habits in the lab leading to RNA degradation or decreased labeling in some samples. Changing labeling kits and reagents within an experiment can contribute to variation. One should make sure all reagents are of very high quality. A reduced efficiency of an enzyme can lead to a reduction in labeling efficiency and/or reduced hybridization of the probes to the array. System variation can arise from the equipment used to run a microarray experiment. Using two different scanners to scan the chips within an experiment can contribute to variation as one scanner may have a reduced laser power output and will thus scan the hybridized probes at a reduced intensity. It is important to check all settings of the scanner before beginning.

Some practical questions (and *reasons*) to consider when planning a microarray experiment include:

- How much variation does the system have? *Understanding and minimizing this variation is of utmost importance.*
- What level of significance is needed? *If subtle changes are expected then more replicates may be needed.*
- How many treatments and controls? *Is the experiment a comparative analysis (two experimental conditions such as normal and disease tissue) or a serial analysis (multiple experimental conditions)?*

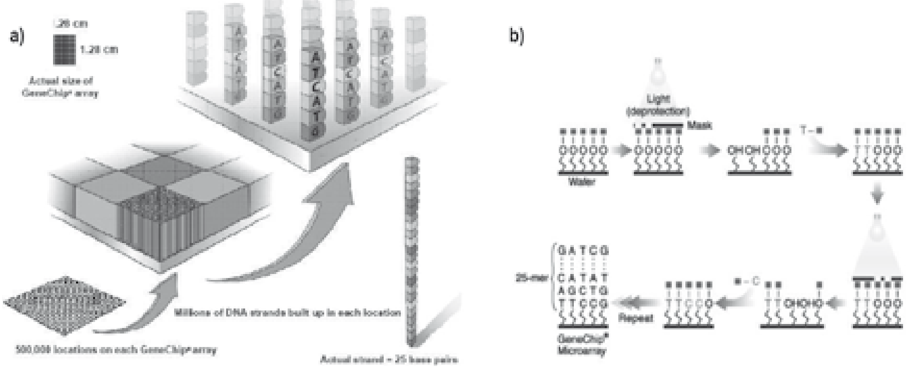

a)
1.28 cm
1.28 cm
Actual size of GeneChip® array
500,000 locations on each GeneChip® array
Millions of DNA strands built up in each location
Actual strand = 25 base pairs
b)
Light (deprotection)
Mask
Wafer
OH OH
T –
– C
25-mer
G A T C G
C A T A T
A G C T G
T T C C G
Repeat
GeneChip® Microarray

Figure 2: Production of Affymetrix GeneChips a) The probes on an Affymetrix GeneChip are synthesized directly on the chip by a lithographic masking method. b) The photolithographic process of synthesizing oligos begins by coating a 5" x 5" quartz wafer with a light-sensitive chemical compound that prevents coupling between the wafer and the first nucleotide of the DNA probe being created. Lithographic masks are used to either block or transmit light onto specific locations of the wafer surface. The surface is then flooded with a solution containing either adenine, thymine, cytosine, or guanine, and coupling occurs only in those regions on the glass that have been de-protected through illumination. The coupled nucleotide also bears a light-sensitive protecting group, so the cycle can be repeated. In this way, the microarray is built as the probes are synthesized through repeated cycles of de-protection and coupling. The process is repeated until the probes reach their full length, usually 25 nucleotides. Commercially available arrays are typically manufactured at a density of over 1.3 million unique features per array. Depending on the demands of the experiment and the number of probes required per array, each quartz wafer can be diced into tens or hundreds of individual arrays.

Figures adapted from http://awww.affymetrix.com/technology/manufacturing/index.affx, http://keck.med.yale.edu/affymetrix/technology.htm

1.4 Gene Chip technology

This thesis will review microarray technology using various GeneChips technologies developed by Affymetrix. All basic explanations of the technology and further analysis will refer to the 3' gene expression arrays (U133A & B and U133-2plus). Exon arrays, an alternative to gene-expression arrays, which can also measure gene expression, are outlined in paragraph 1.12.1.

Affymetrix GeneChips refer to the high-density oligonucleotide-based arrays, which consist of small DNA oligonucleotides; referred to as probes. These DNA probes are chemically synthesized at specific locations on a coated quartz surface. The exact location where a probe is located is called a feature and on a single array there can be millions of features [8]. DNA probes are synthesized *in situ* within a feature on silicon wafers using a photolithographic process (Figure 2). The 11-µm DNA probes on the array are 25 nucleotides long and a probe set consists of 11 different probe pairs (22 probes). The 54,000 different probe sets on the current U133-2plus GeneChip microarray represent approximately 30,000 known genes and EST sequences.

Every probe on an Affymetrix GeneChip is designed to determine whether the complementary sequence of RNA or DNA is present in the sample. At the molecular level, the probe must be specific enough to distinguish a sequence from similar sequences in order to get an accurate assessment of whether the complementary molecule is present in the interrogated sample. With the possibility of having millions of features on a single array, multiple probes are present for each sequence expressed. These multiple measurements provide high sensitivity and reproducibility, just as the 25-mer oligonucleotide probe length confers high specificity. This allows for consistent discrimination between signal and background noise.

On U133-2plus expression GeneChips, 22 probes are used for each expression measurement. For each probe on the array that perfectly matches (PM) its target sequence, Affymetrix also created a paired "mismatch" probe (MM). The mismatch probe contains a single mismatch located directly at the 13th position in the 25-mer probe sequence [24] (Figure 3). This mismatch probe is used as a background control and also to overcome the low specificity of the short oligonucleotide used [25]. While the perfect match probe provides measurable fluorescence when the sample binds to it, the paired mismatch probe is used to detect and eliminate any false or contaminating fluorescence within that measurement [26]. The mismatch probe serves as an internal control for its perfect match partner because it hybridizes to nonspecific sequences about as effectively as its counterpart, allowing misleading signals, from cross hybridization for example, to be efficiently quantified and subtracted from a gene expression measurement or genotype call [24, 25, 27]. Ideally, all 11 PM probes should have the same signal intensity, having measured the same gene; however, this usually is not the case.

There may be enormous differences between individual probes in the probe set, which can be adjusted and taken into account in further analysis (Refer to paragraph 1.6.2).

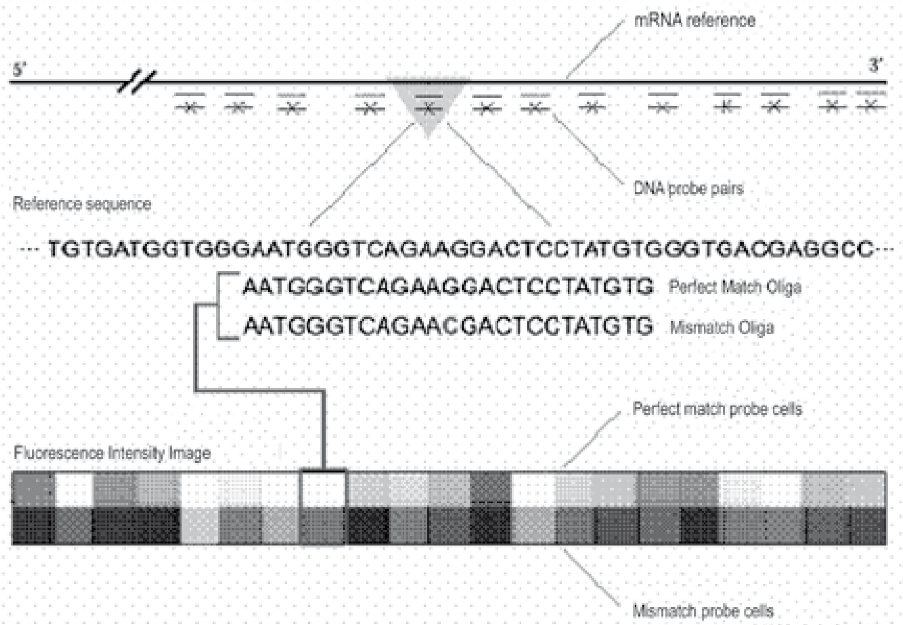

mRNA reference
5'
3'
DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC···
AATGGGTCAGAAGGACTCCTATGTG Perfect Match Oliga
AATGGGTCAGAACGACTCCTATGTG Mismatch Oliga
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

Figure 3: Affymetrix has employed the unique PM-MM probe pair probe design approach

On Affymetrix GeneChips, each probe pair consists of a perfect match oligonucleotide and a mismatch oligonucleotide. The perfect match probe has a sequence exactly complimentary to the particular gene and thus measures the expression of the gene. The mismatch probe differs from the perfect match probe by a single base substitution at the center base position, disturbing the binding of the target gene transcript. These probes are designed to obtain the optimal balance of highest sensitivity and specificity in the presence of a complex sample background. In addition, redundant sampling of each sequence with multiple probe pairs in a probe set provides robustness and reliability in the data obtained. From this picture, the differing intensities represented in the PM probes, at least one of these probe pairs will have less weight in the analysis based on the signal intensity in the mismatch, indicating non-specific hybridization.

Figure taken from http://keck.med.yale.edu/affymetrix/technology.htm

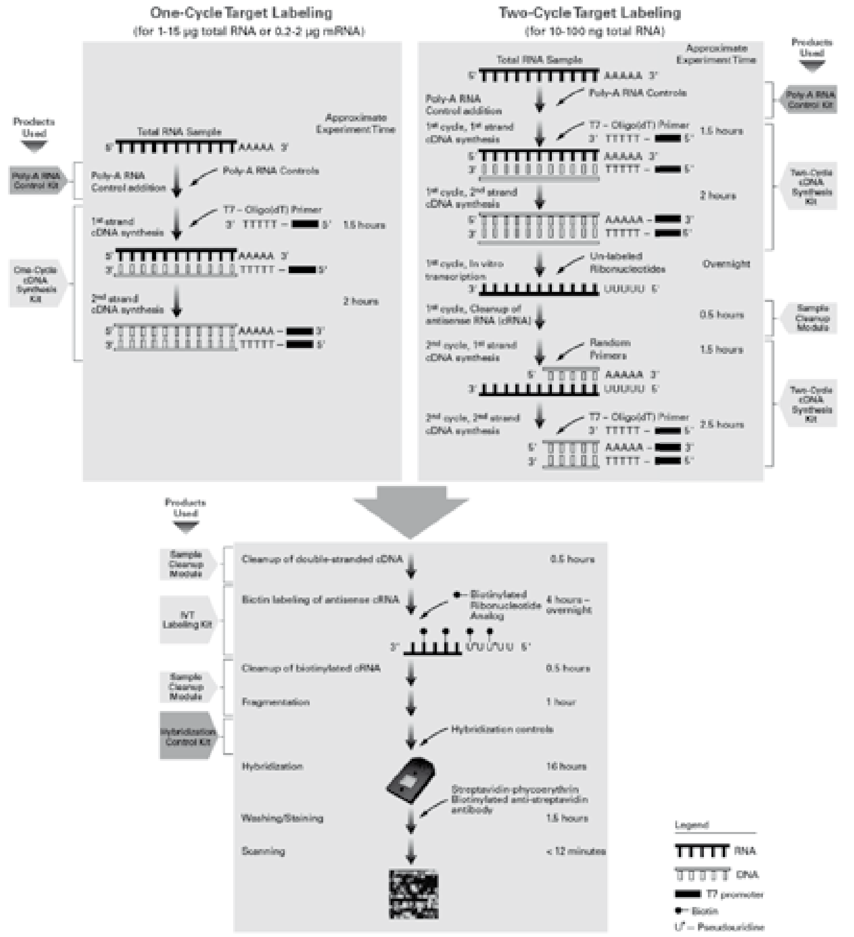

One-Cycle Target Labeling
(for 1-15 μg total RNA or 0.2-2 μg mRNA)
Products Used
Poly-A RNA Control Kit
One-Cycle cDNA Synthesis Kit
Total RNA Sample
Approximate Experiment Time
5' AAAAA 3'
Poly-A RNA Control addition
Poly-A RNA Controls
1st strand cDNA synthesis
T7 – Oligo(dT) Primer
3' TTTTT – 5'
1.5 hours
5' AAAAA 3'
3' TTTTT – 5'
2nd strand cDNA synthesis
2 hours
5' AAAAA – 3'
3' TTTTT – 5'
Two-Cycle Target Labeling
(for 10-100 ng total RNA)
Total RNA Sample
Approximate Experiment Time
Products Used
5' AAAAA 3'
Poly-A RNA Control addition
Poly-A RNA Controls
Poly-A RNA Control Kit
1st cycle, 1st strand cDNA synthesis
T7 – Oligo(dT) Primer
3' TTTTT – 5'
1.5 hours
5' AAAAA 3'
3' TTTTT – 5'
Two-Cycle cDNA Synthesis Kit
1st cycle, 2nd strand cDNA synthesis
2 hours
5' AAAAA – 3'
3' TTTTT – 5'
1st cycle, In vitro transcription
Un-labeled Ribonucleotides
Overnight
3' UUUUU 5'
1st cycle, Cleanup of antisense RNA (cRNA)
0.5 hours
Sample Cleanup Module
2nd cycle, 1st strand cDNA synthesis
Random Primers
1.5 hours
5' AAAAA 3'
3' UUUUU 5'
Two-Cycle cDNA Synthesis Kit
2nd cycle, 2nd strand cDNA synthesis
T7 – Oligo(dT) Primer
3' TTTTT – 5'
2.5 hours
5' AAAAA – 3'
3' TTTTT – 5'
Products Used
Sample Cleanup Module
Cleanup of double-stranded cDNA
0.5 hours
IVT Labeling Kit
Biotin labeling of antisense cRNA
Biotinylated Ribonucleotide Analog
4 hours – overnight
3' U*U U*U U 5'
Sample Cleanup Module
Cleanup of biotinylated cRNA
0.5 hours
Fragmentation
1 hour
Hybridization Control Kit
Hybridization controls
Hybridization
16 hours
Streptavidin-phycoerythrin
Biotinylated anti-streptavidin antibody
Washing/Staining
1.5 hours
Scanning
< 12 minutes
Legend
RNA
DNA
T7 promoter
Biotin
U* – Pseudouridine

Figure 4: GeneChip eukaryotic target labeling assays for gene expression analysis The Affymetrix gene expression assay for labeling samples. Both one-cycle (for small amounts of RNA) and two-cycle assays are available.

Figure taken from http://awww.affymetrix.com/technology/manufacturing/index.affx

1.5 Labeling and hybridization procedure

Affymetrix gene expression arrays use a standardized biotin labeling protocol, which utilizes an Oligo(dT)-primed, *in vitro* transcription based linear amplification strategy (Figure 4). Following this, strict protocols are utilized by the standard Affymetrix fluidics and scanning station. The advantages of GeneChip technologies include the specificity and reproducibility of experiments due to the multiple probes that interrogate a single gene and the automated control of the experimental process from hybridization to quantification [28].

1.6 Scanning expression microarrays: converting probe sets to signal intensity

1.6.1 Data output from the scanner

Each probe pair consists of a perfect match and a mismatch probe. The signal intensity emitted from the probes within a probe set are read by the scanner and then summarized or condensed into an expression index that represents the transcript level of the corresponding gene by the Affymetrix software [29]. The text box below describes the stages of output files in the scanning and processing of the GeneChip.

4 staged Affymetrix scanning output files

- **Experiment File *.EXP:** This file contains the parameters of the experiment such as Probe Array Type, Experiment Name, Equipment parameters, Sample Description, and others. This file is not used for analysis, but is required to open other GCOS* files for the designated chip experiment.
- **Image Data File *.DAT:** This file is the image file generated by the scanner from the Probe Array after processing on the Fluidics Station. This file can be viewed in GCOS or exported as a *.TIFF image. This file is used in GCOS to generate the *.CEL file.
- **Cell Intensity File *.CEL:** The cell file contains the processed cell intensities from the primary image in the *.DAT file. The cell file is used by GCOS to generate the *.CHP file, which contains the numerical data from the *.DAT, and *.CEL files.
- **Probe Array Results File *.CHP:** The chip file is the output file from the GCOS expression analysis of the Probe Array. The chip file contains the data that will be used for statistical analysis and data mining analysis.

* GCOS is the Affymetrix software suite, which controls the hybridization and fluidics station as well as the scanner. GCOS regulates the final laboratory processing producing the specified files as well as having the option to do statistical pre-processing within its environment. Alternatives to pre-processing in this environment are described below.

Pre-processing of microarray data is a very important step in the analysis of GeneChips. Many variables can contribute to variation within and between microarrays, and these can effect the interpretation of signal intensity, leading to the possibility of false positives. A .CEL file has all the signal information associated with each probe feature. Microarray pre-processing of Affymetrix GeneChips traditionally consists of four steps beginning from the .CEL file level. These steps can be continued in the GCOS software after scanning and calculating a .CEL file, or can be used in external processing software such as the R library files from within the

Bioconductor package. The R/Bioconductor project is primarily an academic based project in which usable statistical R libraries have been deposited in the database [30, 31]. The libraries contain both data-preprocessing and normalization methods as well as advanced downstream statistical algorithms that are not often incorporated into commercial software.

Many commercial software packages with GUI's (Graphical user interface) can pre-process and analyze microarray expression data. Examples of this include Spotfire ® DecisionSite (Spotfire, MA, USA) [32], Omniviz (Omniviz, MI, USA) [33] and Rosetta Resolver (Rosetta Biosoftware, WA, USA) [34].

4 GeneChip pre-processing steps

1. Background correction (B) of the probes, which can either be omitted or with such methods as 'Mas', 'RMA' or 'RMA2' can be used.
2. Normalization (N) within the chip to correct technical variation (see also paragraph 1.6.1), or to facilitate between-array comparison. This can be chosen from methods such as 'quantiles', 'invariant set of genes' or 'loess'.
3. PM correction methods such as 'Mas', using 'PM only' and 'subtracting MM from PM'.
4. Summary (S) methods such as; 'Mas', 'average difference' and 'median polish' which converts the 11-22 probe pairs (PM/MM) intensities into one probe set value.

[expression value = S(N(B(X))), where X is .CEL file]

There are also many freely available alternatives including dChip [35-37] and for the more advanced users, libraries within the R statistical analysis platform [31] are available within the Bioconductor project [30] (see figure 5 for an example of R GUI environment). The vignettes found on the Bioconductor sites have thorough descriptions and codes to use within the R environment to process the Affymetrix chips into expression values. During the summarization steps, the researcher can choose various algorithms for each step or just choose a single step algorithm. For example, it is possible to define whether the MM values will be utilized in the calculation of the probe signal. Normalization (Paragraph 1.6.2) can also be directly included within this process or the probe sets can be summarized without normalization into expression values for each transcript. These raw summarized signal intensities can be used in alternative normalization and analysis software. The advantage of the R environment is that the pre-processing can be tailored to one's need. A disadvantage of having multiple pre-processing options is that it can be difficult to compare data in public databases as expression measurements may vary more on account of the pre-processing method than the biology. Other open source tools using the R modules are available as add-ins for popular programs such as Excel. One such tool is the BRB-Array package (Biometric Research Branch, National Cancer Institute, MD, USA) [38], which calls R functions from within the Excel environment to process data, including the initial steps of normalization. Refer to figure 6 for the effects of pre-processing on signal intensity distribution in .CEL files.

1.6.2 Normalization and summarization

To compare different microarrays, intensity values measured by Affymetrix GeneChips must be normalized and summarized, generating the final probe set expression values [39]. Normalization is a type of 'calibration' that serves to remove non-biological or systematic variation among samples such as total brightness, differences in the background and noise levels, measurement, hybridization conditions, handling and instrument inconsistencies [29,40]. Such systematic differences were outlined in chapter 1.3. The normalization may be contained in a complete algorithm of the three steps of data preprocessing, such as Mas or RMA (see also paragraphs 1.6.2.1 and 1.6.2.3, respectively). If this is not the case, a normalization step should be performed before comparing two or more microarrays.

An early approach to normalization was to use one gene or preferably several genes whose expression is invariant across all samples. The standards typically were housekeeping genes. This is based on the assumption that these genes would be expressed at nearly equal levels in all cells. This, however caused problems as the expression of housekeeping genes were found to vary substantially among cell lines and certainly among cell types, therefore the stable expression of the household genes needs to be proven [41,42].

Currently, there are several types of normalization frequently used:

- **Global:** based on all probes on the array assuming that expression of the genes does not change (such as quantile normalization (Paragraph 1.6.2.2), or VSN (Paragraph 1.6.2.4).
- **Internal reference:** based on invariant gene sets or 'household genes' provided by Affymetrix on their GeneChip These include known housekeeping genes including *GAPDH* and *ACTIN,* but there are also probe sets spanning the 3' middle and 5' ends of the *GAPDH* gene for example, to control efficiency of RNA labeling procedures. Affymetrix also provides buffer controls to control for nonspecific hybridization and auto fluorescence. There are also statistical methods to reliably choose the housekeeping genes for normalization [41,42].
- **External reference:** using spiked-in RNA. This requires accurate quantification of amounts of spiked-in RNA, and probes complementary to the spikes on Affymetrix GeneChips [43].

1.6.2.1 Mas normalization

The 'Mas' algorithm is a summary method in the R Bioconductor 'affy' library. This is routinely embedded in Microarray Suite software from Affymetrix, now known as GCOS. The global normalization method implemented in the Mas algorithm (Paragraph 1.6.1) is based on the assumption that the total amount of labeled mRNA in all samples is roughly the same [29]. Mas uses a robust estimator, Tukeys biweight, based on a weighted mean to estimate the variance among probe pairs [24]. Following this, the algorithm applies the Wilcoxon-Signed Rank statistical test to make the confidence calls, which indicate the reliability of each call. The Mas algorithm also uses a global normalization method that applies a scaling factor to bring the signal intensity of the trimmed mean intensity to a user determined target value (after excluding the top and bottom 2%).

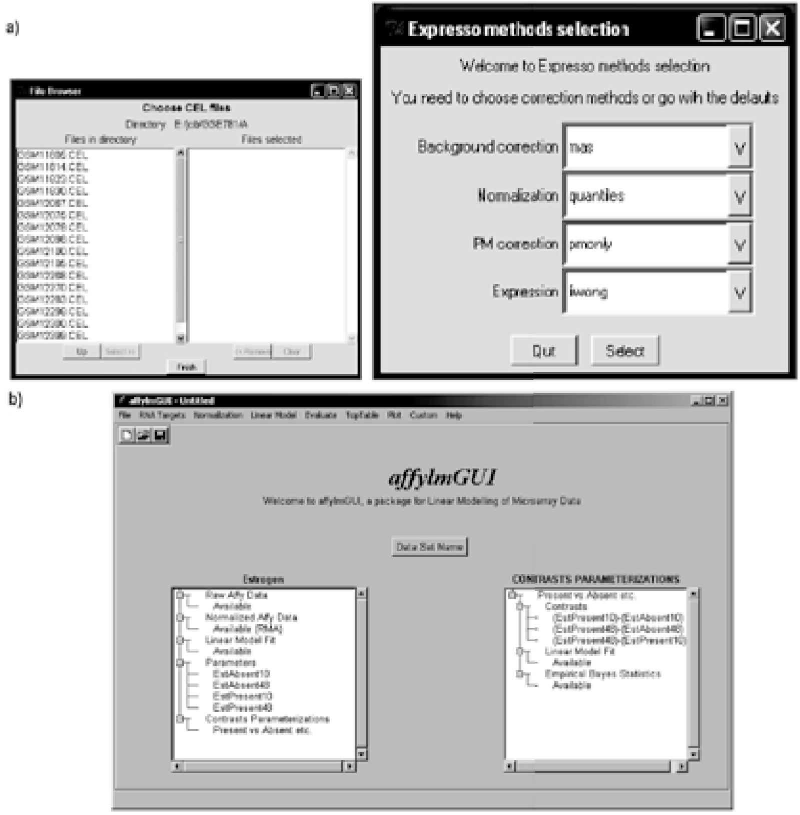
1.6.2.2 Quantile normalization

By definition, a quantile is the sorted percentage of a distribution into four equal ordered subgroups [2]. Quantile normalization assumes that all the genes on the array are expressed at relatively unaltered levels across varying conditions. Quantile normalization can be carried out at the probe level by imposing that all the probe level distributions are equal. By mapping each quantile to the average quantile across the arrays and then summarizing the gene indices from the normalized arrays [44]. Each gene has 11 perfect match probes and the quantile algorithm gives each array the same distribution by calculating the mean of each quantile and substituting it as the data value in the original data set. Quantile normalization is utilized within RMA normalization (1.6.2.3) and can be accessed through the Bioconductor "affy" package.

1.6.2.3 RMA/RMAexpress

Another widely used normalization method is the RMA (Robust Multi-Array analysis) [45, 46]. This algorithm is implemented as a module within Bioconductor and as a stand-alone version with a GUI (RMAexpress) (Figure 5a) [47]. The RMA algorithm adjusts the background to create an ideal match (IM), ignoring MM and removing global background [48]. It utilizes quantile normalization in which the intensities are adjusted, ignoring outliers to produce identical distributions and median polish to estimate the log expression robustly. This median polish repeatedly subtracts row medians and column medians from a matrix of probe intensity values until it stabilizes. A modified version of RMA is GC-RMA, which models the intensity of the probe level data taking into account the stronger binding of G/C pairs presumably resulting in higher intensity values for GC-rich probes [39,49]. GC-RMA uses background estimates based on GC content, by using a subset of probes with the same GC content as a population of MM probes that can be considered as pseudo MM for all PM with the same GC content.

This background estimate is made using maximum likelihood estimate or an Empirical Bayes estimate.



File Browser

Choose CEL files

Directory E:\cbi\GSE781\A

Files in directory

GSM11096.CEL
GSM11014.CEL
GSM11023.CEL
GSM11030.CEL
GSM12087.CEL
GSM12075.CEL
GSM12078.CEL
GSM12098.CEL
GSM12100.CEL
GSM12105.CEL
GSM12268.CEL
GSM12270.CEL
GSM12280.CEL
GSM12296.CEL
GSM12290.CEL
GSM12299.CEL

Files selected

Expresso methods selection

Welcome to Expresso methods selection

You need to choose correction methods or go with the delauts

Background correction: mas

Normalization: quantiles

PM correction: pmonly

Expression: liwong

affylmGUI - Untitled

affylmGUI

Welcome to affylmGUI, a package for Linear Modelling of Microarray Data

Estrogen

```
Raw Affy Data
  Available
Normalized Affy Data
  Available (RMA)
Linear Model Fit
  Available
Parameters
  EstAbsent10
  EstAbsent48
  EstPresent10
  EstPresent48
Contrasts Parameterizations
  Present vs Absent etc.
```

CONTRASTS PARAMETERIZATIONS

```
Present vs Absent etc.
  Contrasts
    (EstPresent10)-(EstAbsent10)
    (EstPresent48)-(EstAbsent48)
    (EstPresent48)-(EstPresent10)
  Linear Model Fit
    Available
  Empirical Bayes Statistics
    Available
```

Figure 5: pre-processing GUI within R environment. By calling the functions as written below within the 'affy' library environment, one can use the GUI for the R environment that will pre-process Affymetrix microarray chips.

a) **> eset ← expresso(data object, widget=true)**

> data ← ReadAffy(widget=true)

By using this GUI, one can specify the .CEL files that they want processed as well as the algorithms to background correct, normalize, PM correct, and expression summarize. The GUI will then process the files within an R environment with this easy step through process.

b) By using the AffylmGUI, it utilizes the Limma package within R for linear modeling of microarray data and finding differentially expressed genes; controlling the problem of multiple testing with FDR. AffylmGUI can be accessed by directly loading the package within the R environment.

1.6.2.4 Normalization by VSN

VSN (Varying stabilizing normalization) is a method for normalization found in the Bioconductor R package “vsn”.

VSN addresses the dependence of the variance on the mean intensity, and a variance stabilizing data transformation [30,50]. Differences between transformed intensities are analogous to “normalized log-ratios” and their variance is independent of the mean, usually being more sensitive and specific in detecting differential transcription [30,50].

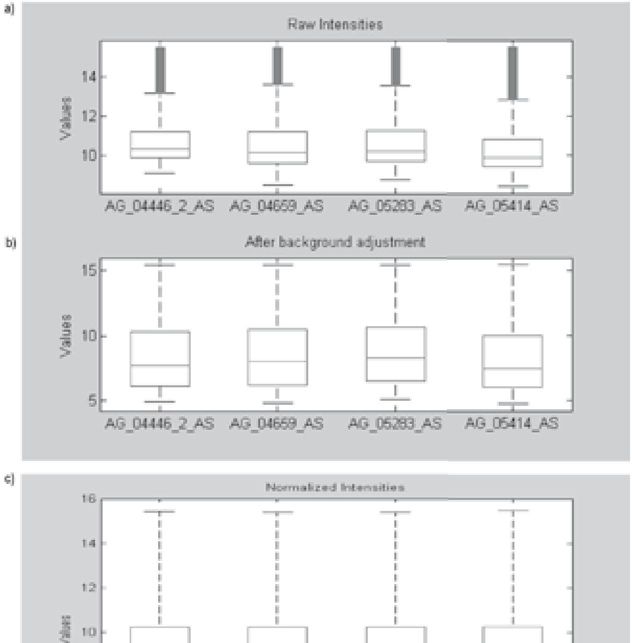


Figure 6: Illustration of the effects of pre-processing to the signal intensity distributions across multiple microarrays. Boxplots illustrate the comparison of intensity distributions across several microarrays a) before and b) after the background adjustment and c) normalization. Several nonlinear normalization methods can be applied to Affymetrix microarray data. In this illustration, GeneChips have been normalized at the probe-level with a quantile normalization method. Expression values intensities were summarized for each probe set across multiple microarrays by the median polish. The expression values are the probe set intensity summaries on log-2 scale. http://www.mathworks.com/products/demos/bioinfo/primategeneexpdemo/primategeneexpdemo.html

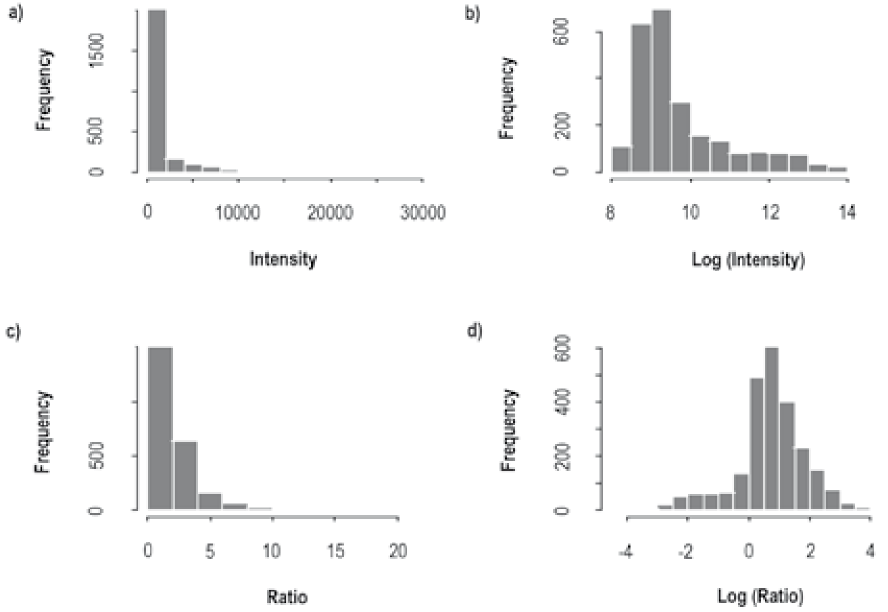


Figure 7: Illustration of the effects of log transformation. The result of log transformation is that positively skewed data (large part of the graph for upregulated data) is transformed into a more symmetrical data distribution around 0 (usually creating a normal distribution). This means that a graph is created where up- and downregulated genes are treated in similar fashion, both using a similar part of the graph (see b and d for the log transformation of raw intensity and ratios). If the data was not log transformed, in comparison of two intensity signals for the same spot, the disadvantage of treating up and down regulated genes differently arises as systematic error. Genes upregulated by a factor of 2 have an expression ratio of 2, whereas those downregulated by the same factor have and expression ratio of 0.5. This will result in a graph where the upregulated genes have a much wider range than the downregulated genes (see a and c). Another result of log transformation is decreasing the influence of very high expression values in the mean or median value, because they will get relatively smaller with log transformation. Small values will be more spread and have more influence. Without log transformation the data in the lower intensity range is harder to detect i.e. in a linear scale the differences between intensities are not as significant as if you were looking at them on a log scale.

Figure taken from http://www.systemsbiology.nl/datgen/transcriptomics/preprocessing/log.html

1.6.3 Other Transformations

Some transformations are also a type of normalization. A simple method of transformation is to divide every intensity value on the array by the arithmetic or geometric mean, or logarithmic median of the entire array. Multiplying all values by a constant will shift the median and change the shape of the distribution. This can effectively establish a common reference for between array comparisons with a linear transformation. This transformation is specific for each array, so the relative expression level differences between genes on the same array do not change [29].

Following normalization, a log transformation of the data should be performed. In log transforming the data (typically base 2 with microarray data), the intensity measurements obtain a distribution that is closer to a normal distribution with the variation being independent of the magnitude (Figure 7). Log transformation gives a more realistic sense of the variation throughout the data and evens out highly-skewed distributions [29,51,52]. This provides values that are easily interpretable and meaningful from a biological point of view, for example the calculation of differentially expressed genes that have a high correlation of expression values [51].

1.6.4 Choice of normalization

It is not trivial to choose which normalization procedure is most appropriate and not a single algorithm has been shown to be superior over others. In chapters 4 and 5, we have applied Mas normalization to the analysis of experiments. In chapters 6 and 7, we have applied the more recent PLIER algorithm as a summary method followed by quantile normalization. Mas and RMA both perform very well with larger data sets. Quantile normalization performs better with smaller data sets and this has been performed with trial and error and known positive controls for the experiments. PLIER may supersede Mas in the newer generation GCOS software. Currently it is implemented in the EXACT software, which is used instead of GCOS for exon arrays. Millenaar *et al.* [24] have compared the calculation of gene expression array by six different algorithms which all resulted in different levels of expression, but all yielded the same list of genes as being differential by expression. In this study it was also found that MM signals do not truly represent non-specific binding for PM signal as in 47% of the cases there was correlation between both PM and MM signals, thus underestimating the true signals [24]. If true, RMA would be more appropriate in calculating normalized expression signal, as it does not take into account the MM signal. Verhaak *et al.* have also evaluated the effects of different pre-processing methods [39]. Using two independent large and small data sets, they found that the choice of pre-processing method is of relatively minor influence on the final analysis outcome of large microarray studies whereas it can have significant effects on the results of a smaller study. They also found that the data source (platform, tissue homogeneity, RNA quality) is potentially of bigger importance than the choice of pre-processing method [39]. Figure 7 illustrates the effects of data preprocessing on the distribution of the signal intensities across the arrays.

1.7 Clustering: Unsupervised analysis

Clustering is a visualization tool that aims at dividing the data in such a way that similar items fall into the same group and dissimilar items fall into different groups [53]. Cluster analysis is considered an unsupervised method when phenotype class information of the samples is not utilized. Cluster analysis is sometimes also used to infer that patients in the same cluster have the same subtype of a disease or to infer that genes in the same cluster have a similar function. One of the first research groups to publish clustered microarray data was Eisen *et al.*, who applied hierarchical clustering to yeast microarray data [54]. Subsequently, there have been many different algorithms developed to cluster microarray data and each method is different in its way to impose biases on the clusters that it constructs. Many good software programs can mathematically and visually cluster microarray data. Commercially, both Omniviz and Spotfire are excellent. BRB-Array tools also offers clustering from the R-package as well as TreeView and Cluster 3.0, which are shareware software programs incorporated into the tool.

One important thing to keep in mind when using clustering methods is that these algorithms will always produce clusters. Even with a data set of completely unrelated data, clustering will always force the data into some sort of groups. Clusters may thus not always be reproducible or biologically meaningful. This is especially a problem when clustering samples based on disease tissue from different patients. The substantial claim that a disease is molecularly heterogeneous requires more evidence than the fact that the clustering algorithm produced clusters. A good proof is to show that patients' tissues are placed in the same cluster when the analysis is repeated using RNA independently extracted from each of the same samples. For most studies, however, independently extracted RNA samples for each sample are not available. When clustering genes this may not be such a severe problem as genes are grouped into biological pathways. The problem then moves more into the biological meaning of the clusters; whether the genes within the same cluster are co-regulated or really are part of the same pathway. One way to assess the validity of the clusters is to resample the genes (with replacement), and repeat the clustering procedure, also known as bootstrapping. BRB-Array tools offers bootstrapping for cluster analysis that will give a probability value of the reproducibility for the clusters on either samples or genes. Below three main types of clustering for microarray expression data are being discussed.

1.7.1 Hierarchical clustering

Hierarchical clustering divides the data set into clusters, which are further subdivided into smaller clusters, forming a dendrogram (Figure 8). There are many different forms of hierarchical clustering with each one differing only by the way in which the clusters are linked together. This is known as the distance (or linkage) measurement. Figure 9 illustrates some of these measurements.

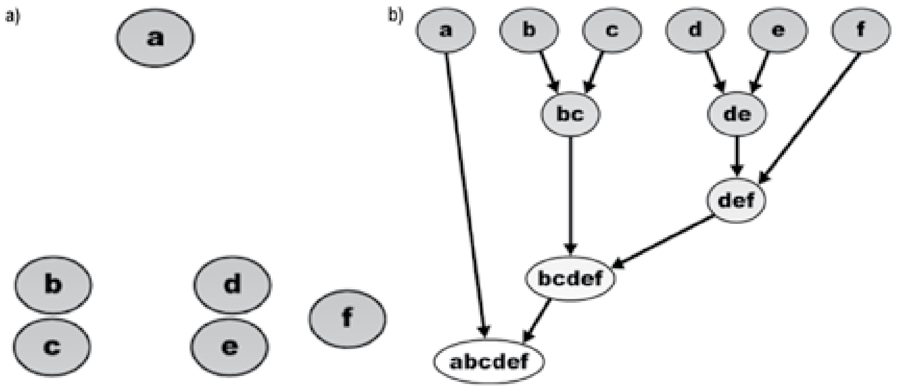
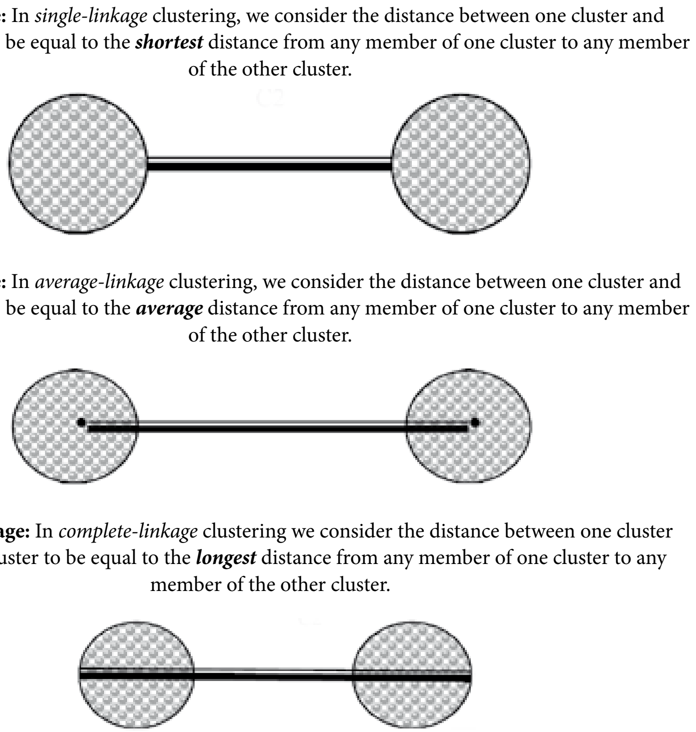


Figure 8: Hierarchical clustering. Given a distance measure such as Euclidean, raw data elements (a) can be combined. Hierarchical clustering builds (agglomerative (as in this example)), or breaks up (divisive), a hierarchy of clusters. The traditional representation of this hierarchy is a tree data structure (called a dendrogram), with individual elements at one end and a single cluster with every element at the other. Agglomerative algorithms begin at the top of the tree, whereas divisive algorithms begin at the bottom (In b, the arrows indicate an agglomerative clustering of the raw data elements). Cutting the tree at a given height will give a clustering at a selected precision. In this example, cutting after the second row will yield clusters {a} {b c} {d e} {f}. Cutting after the third row will yield clusters {a} {b c} {d e f}, which is a coarser clustering, with a fewer number of larger clusters.

1.7.2 Partitioning clustering

Partitioning clustering methods, such as k-means, involve the subdivision of microarray data into a pre-determined number of (k) clusters, without any implied hierarchical relationship between the clusters. K-means clustering begins with k randomly chosen centroids with each gene being assigned to the closest centroid. The centroids are then reset to the average of the genes in each cluster. This process is repeated until there is no more genes that switch clusters. It is important to repeat this clustering method several times using different random seeds, as the initial centroids positions can define different clusters (see figure 10 for an illustration of this method).



Single linkage: In *single-linkage* clustering, we consider the distance between one cluster and another cluster to be equal to the ***shortest*** distance from any member of one cluster to any member of the other cluster.

Average linkage: In *average-linkage* clustering, we consider the distance between one cluster and another cluster to be equal to the ***average*** distance from any member of one cluster to any member of the other cluster.

Complete linkage: In *complete-linkage* clustering we consider the distance between one cluster and another cluster to be equal to the ***longest*** distance from any member of one cluster to any member of the other cluster.

Figure 9: Schematic representation of the different linkage measurements in hierarchical clustering.

Like hierarchical clustering, there are mathematical algorithms that measure the similarity among expression patterns of two genes. There are several types of similarity measures, however the two most commonly used algorithms are Euclidean distance and Pearson correlation coefficient. Euclidean distance measures the absolute distance between two points in space, which in this case are defined by two vectors. Euclidean distance will be affected by both the direction and the amplitude of the vectors, so that two genes that are coordinately expressed may not be seen to be similar if one has a much higher signal than the other [55]. Correlation distance on the other hand, treats the vectors as if they were the same (unit) length, and is thus insensitive to the amplitude of changes that may be seen in expression profiles [55]. Euclidean similarity is sensitive to scaling and differences in average expression level, whereas correlation is not [53].

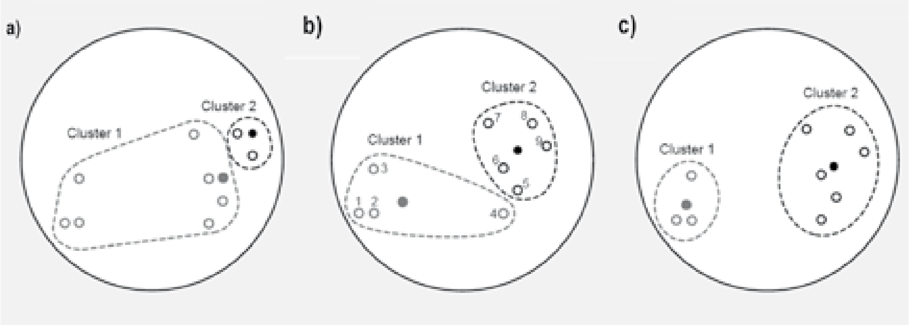


Figure 10: Illustration of the process of calculating k-means clusters

The algorithm for calculating k-means is:

1. Choose k points in m-space at random (or otherwise) as the k means
2. a. Calculate the distance of each gene (its expression pattern) from each of the k means

   b. Assign each gene to the closest k-mean

   c. Calculate the new location of each k-mean as the average of the gene expression patterns assigned to that k-mean
3. Repeat step 2 until the locations of the k-means stabilizes. It may be necessary to experiment with the number of repeats of step 2, observing how the k-means change.

In this example:

(a) Cluster 1 (filled red circle) and cluster 2 (filled black circle) are chosen arbitrarily. All data points (open circles) are then partitioned into two clusters: each data point is assigned to cluster 1 or cluster 2, depending on whether the data point is closer to reference point 1 or 2, respectively.

(b) Results of first iteration: Next, each reference point is moved to the centroid of its cluster. Then each data point is considered in the sequence shown. If the reference point closest to the data point belongs to the other cluster, the data point is reassigned to that other cluster, and both cluster centroids are recomputed.

(c) Results of second iteration: During the second iteration, the process in Figure 3(b) is performed again for every data point. The partition shown above is table; it will not change for any further iteration.

Figure adapted from http://fas.org/sgp/othergov/doe/lanl/pubs/00412967.pdf

1.7.3 Multi-dimensional clustering

Data from microarray experiments is considered high dimensional data. Our visual system is based on four dimensions (3 for space and 1 for time) and although color and shape can be used to visualize more dimensions, we quickly reach the limits of what we can understand and interpret. A natural solution would be to try to reduce the number of dimensions by eliminating those dimensions that are not "important". One way to do this is a multi-dimensional scaling approach, which is similar to cluster analysis in that one is attempting to examine the relations among samples. Principle component analysis (PCA) does exactly that by ignoring the dimensions in which data does not show variation and combining those that are similar. PCA then calculates a new system of coordinates. The objective in this linear scaling is to reduce the dimensionality of the data set while preserving the pair-wise similarities or distances between objects in the low dimensional graphical representation, i.e. without losing any information (Figure 11) [32, 38]. In a PCA representation, each gene or sample is represented by a point and the distances between the points are determined to best preserve the distances between the high dimensional expression profiles [38]. PCA takes a high dimensional data set such as microarray data and produces a new data set that consists of fewer variables. These variables are the linear combinations of the original variables, so it is often possible to deduct the biological meaning of what they represent. This clustering method does not reduce the dimensionality *per se*, but it reveals the true dimensionality of the original data in variables that we can understand. When running a PCA, the analyst can choose the number of dimensions (components) to project the data to, resulting in fewer variables than the original data set, but retaining as much as possible of the information. The majority of variation in the microarray data set should be in the first few dimensions represented by the principle components.

1.7.4 Choice of clustering method

Clustering methods can be used in various areas of biological analysis. To date, there is no single algorithm to determine the correct number of clusters that can best represent data biologically, or even not a consensus of which algorithm to apply to various data sets. However, some general themes are emerging in the literature.

- Complete linkage often outperforms average linkage [56].
- Single linkage works very badly with 'real world' data sets such as microarray data.
- Euclidean distance may be more appropriate for log ratio data, whereas Pearson's correlation seems to work better with absolute values [56].

Clustering algorithms serve the purpose for analysis of microarray data but also many other forms of numerical data.

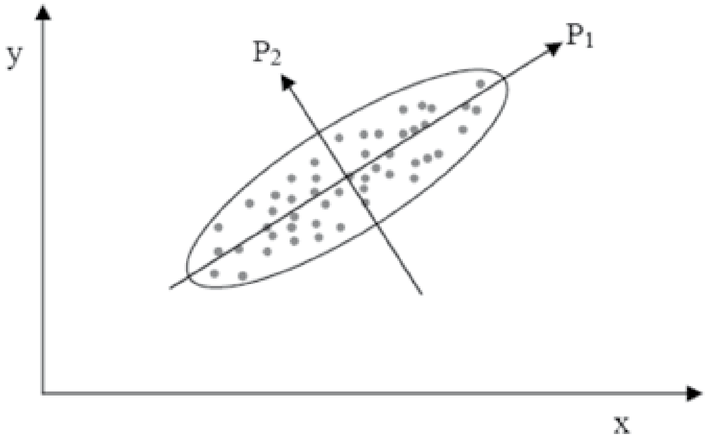


Figure 11: Multi-dimensional analysis: principle components. The objective of the Principal Component Analysis is to reduce the dimensionality of the data set. It involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables that are called principal components. Each data point in this diagram has two coordinates. However, this data set is essentially one dimensional because most of the variance is along the first eigenvector p1. The variance along the second eigenvector p2 is marginal, thus, p2 may be discarded.

1.8 Visualization of gene/sample similarity: Pearson correlation matrix

Visualization by Pearson correlation matrix will aid in discriminating groups of genes/samples that are similar based on pair wise correlations between genes or samples. This can identify clusters of samples that can either be positively or negatively correlated. One hundred percent negative correlation would indicate that genes with a high level of expression in one sample would always have a low level of expression in the other sample and vice versa. An example of this type of visualization is shown in figure 12 [57]. The matrix-ordering algorithm of the Pearson correlation matrix starts with the most highly correlated pair of samples, and through an iterative process, it sorts all samples into correlated blocks. Each sample is joined to a block in an ordered manner so that a correlation trend is formed within a block, with the most correlated samples at the center. The blocks are then positioned along the diagonal of the plot in a similar ordered manner.

1.9 Supervised analysis

Microarray experiments require a lot of planning and a clear experimental design, with regard to not only organizing and preparing the samples to be used in the experiments but also in the choice of analysis methods that will be used to evaluate the hypothesis. Analysis methods are 'supervised' when they include classification information relating to the phenotype of the samples. There are three main forms of supervised analysis:

i. Class comparison identifies differentially expressed genes among pre-defined phenotypic classes.

ii. Class prediction searches for differentially expressed genes with the aim of identifying a set of genes that can accurately predict a biological group, diagnostic stage or the prognosis of a patient.

iii. Class discovery involves grouping of samples that have homogeneous expression profiles that can represent unique subgroups for a particular disease. This can be performed through unsupervised clustering supported with chi-square statistics or similar methods.

1.9.1 Class comparison

Class comparison involves the discovery of differentially expressed genes among different classes or samples. This may be different cell or tissue types or experimental/treatment conditions. For example, when looking at tissues of normal breast and cancerous breast, the genes that are consistently differentially expressed between them, may be involved in the initiation or progression of cancer and these are potential drug targets. When looking for such gene expression differences we use multiple hypotheses testing; testing the association of the gene expression measurements with the phenotype separately for each gene. Multiple hypothesis testing in microarray analysis involves linear statistical models which can be applied to a large number of experimental designs [58]. There are a lot of such linear models including t-test, paired t-test, F-test and ANOVA. By analyzing the genes that are differentially expressed between classes, one can begin to piece together the puzzle of the underlying biology. There are also many algorithms that incorporate the linear models of class comparison such as the SAM tool (***S***ignificant ***A***nalysis of ***M***icroarray data), developed at Stanford University [59]. The SAM algorithm has been implemented in various analysis software packages such as Omniviz and BRB-Array tool. BRB-Array tool also offers a t-test to compare classes. This independent t-test is based on comparing the differences in mean log-ratios (logA-logB) or log-intensities between classes relative to the variation expected in the mean differences. The variation is computed assuming that all the samples are independent, unless otherwise specified in the settings [38]. If two classes are compared and the experiments are paired, then the t-test should be paired. For example, if experiments have been prepared for the primary tumor and metastatic tumor of each patient, then the paired t-test option may improve the statistical power of the analysis [38]. If multiple technical replicates have been performed for some of the samples, then either the analysis must be based on selection of a single replicate for each sample or the samples can be averaged. BRB-Array tools has the option to select univariate/multivariate tests, paired/unpaired tests, control over the number of permutations, randomized variance model for univariate tests and various ways to select significant genes. The test will result in p-values for univariate/multivariate significance, in which an arbitrary cut-off can be selected by either p-value, or the proportion of false discoveries (Paragraph 1.9.2). Tools such as the 'limma' library in R and BRB-Array tools offer such t-tests with the possibility to control or correct for multiple testing error [60]. AffylmGUI is a graphical interface which uses the 'limma' R library [61]. Such a GUI also allows the user to do their pre-processing of the chips in the same environment as class comparison.

1.9.2 Problem of multiple testing: p-values and false discovery rates

In comparing classes, it is statistically easier to reliably determine whether expression profiles for predefined classes are different than to reliably determine exactly which genes are differentially expressed among the classes [38]. In making many independent observations with the same acceptance threshold (t-test criterion) that would be used when considering a single testing event, it can become a problem to control the Type I error (the Type I error rate is the probability of rejecting the null hypothesis when the null hypothesis is in fact true) [62]. For example, each t-test is used to examine the null hypothesis; that there is no difference in gene expression between patients with and without cancer. If we have profiled 5000 genes, 5000 individual t-tests are conducted that will test the hypothesis. If we were to consider each of the tests to be significant with a p-value <0.01, then 50 genes on the list are expected to be false positives, yet the Type I family wise error rate (i.e., over the family of 5000 tests) would be much larger than 0.01. In other words, there would be an excessive number of false rejections and hence the need to account for multiple testing [63].

Most multiple testing correction procedures, such as the classical Bonferroni method [64], the sequential methods of Hochberg and Benjamini [65], and the control of false discover rate (FDR) (see also paragraph 1.8.1.2) are designed to control directly the Type I Family Wise Error Rate [66]. The Bonferroni method of multiple hypothesis testing correction is often considered too conservative for microarray analysis [67]. One reason for this is that it does not take into account the dependency between the test statistics (i.e. for over 50,000 tests, not all tests are completely independent as there are many genes with correlated expression). Another reason is that over-all the type-I error criterion are too strict for microarray research. Microarray research is largely exploratory. It is not intended to give a definitive answer but provides hypotheses that can later be tested using independent biological techniques such as in situ hybridization or immunohistochemical staining. For this reason, it is not so important that each discovery is completely reliable, but that a large proportion of the findings can be trusted as being true. More studies are currently being performed by controlling the FDR rather than the exact type I error. Controlling the FDR was developed in an algorithm by Benjamini and Hochberg in 1995 [60] which can be interpreted as the expected proportion of false positives in the list of declared significant genes. Most packages in R can be adapted to select true genes while controlling multiple testing problems according to the need of the analyst and the biological question. Randomly permutating the classes of samples for each gene multiple times (~1000 times) will give greater power by computing the proportion of random permutations that gave as many genes significant at the significant level as were found in comparing the true classes [38].

1.9.3 Class prediction

Class prediction is the discovery of genes that alone, or in combination can predict which class a sample belongs to. One thing to remember is that the ideal predictors to specify a class of samples may not necessarily be the most differentially expressed genes. An analogy is to think of this like the unique skyline of each city around the world. If you line up all the buildings from a particular city, the total combination of small buildings and large skyscrapers becomes visible. This combination will predict that skyline to be of that particular city. Similarly, gene prediction classifiers consist of a list of genes whose expression pattern is unique to a specific class of samples. This is particularly useful in diagnostics as specific profiles can be drawn that will be able to predict a rare subtype of disease or which drug patients will be effectively treated with. Beer *et al.* [68] for example, have identified a set of genes that can predict survival in early stage lung cancer. From this research, they have also delineated a high-risk group that may benefit from adjuvant therapy. More recently, Valk *et al.* used microarray gene expression profiles from 286 patients with AML to identify and determine the prognosis of AML cases with specific molecular signatures [57]. Using various statistics including a prediction algorithm, they were able to identify 16 subgroups of patients for which class predictors identified the prognostically important clusters. Chromosomal lesions such as deletions and translocations, but also normal karyotypes associated with the subclasses of AML. Some classes correlated with poor prognosis or could predict overall survival of patients within AML subgroups (refer to figure 12).

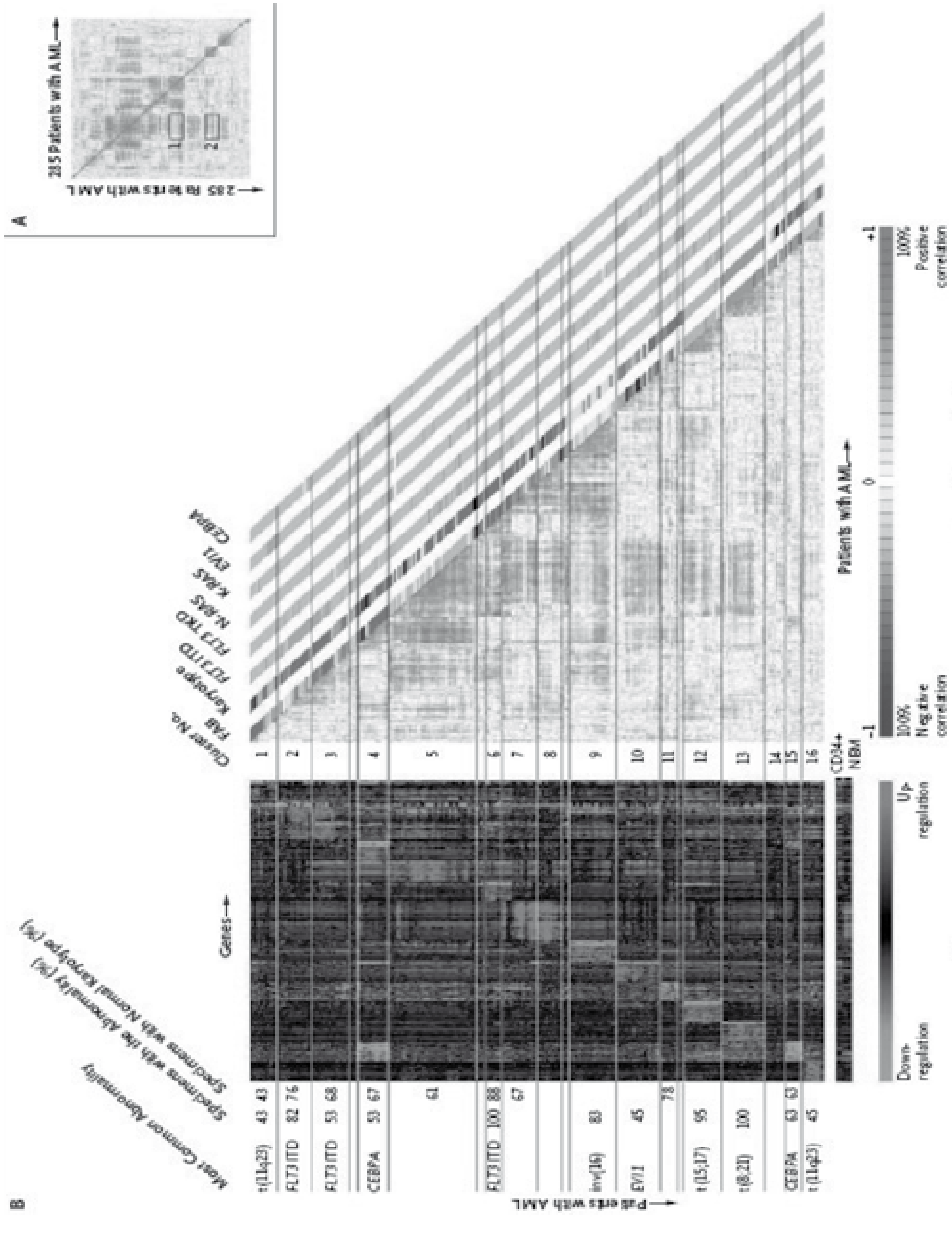

A
285 Patients with AML
285 Patients with AML
1
2
B
Most Common Abnormality
Specimens with the Abnormality (%)
Specimens with Normal Karyotype (%)
Genes
Cluster No.
FAB
Karyotype
FLT3 ITD
FLT3 TKD
N-RAS
K-RAS
EVI1
CEBPA
t(11q23) 43 43 1
FLT3 ITD 82 76 2
FLT3 ITD 53 68 3
CEBPA 53 67 4
61 5
FLT3 ITD 100 88 6
67 7
8
inv(16) 83 9
EVI1 45 10
78 11
t(15;17) 95 12
t(8;21) 100 13
CEBPA 63 63 14
t(11q23) 45 15
16
CD34+ NBM
Patients with AML
Patients with AML
Down-regulation
Up-regulation
Level of Expression
−1
0
+1
100% Negative correlation
100% Positive correlation
Pearson's Correlation Coefficient

Figure 12: Correlation View of Specimens from 285 Patients with AML Involving 2856 Probe Sets (Panel A) and an Adapted Correlation View (2856 Probe Sets) (Right-Hand Side of Panel B), and the Levels of Expression of the Top 40 Genes That Characterized Each of the 16 Individual Clusters (Left-Hand Side of Panel B). In Panel A, the Correlation Visualization tool displays pair-wise correlations between the samples. The colors of the cells relate to Pearson's correlation coefficient values, with deeper colors indicating higher positive (red) or negative (blue) correlations. One hundred percent negative correlation would indicate that genes with a high level of expression in one sample would always have a low level of expression in the other sample and vice versa. Box 1 indicates a positive correlation between clusters 5 and 9 and box 2 a negative correlation between clusters 5 and 12. The red diagonal line displays the intra-individual comparison of results for a patient with AML (i.e., 100 percent correlation). To reveal the patterns of correlation, we applied a matrix-ordering method to rearrange the samples. The ordering algorithm starts with the most highly correlated pair of samples and, through an iterative process, sorts all the samples into correlated blocks. Each sample is joined to a block in an ordered manner so that a correlation trend is formed within a block, with the most correlated samples at the center. The blocks are then positioned along the diagonal of the plot in a similar ordered manner. Panel B shows all 16 clusters identified on the basis of the Correlation View. The French–American–British (FAB) classification and karyotype based on cytogenetic analyses are depicted in the columns along the original diagonal of the Correlation View; FAB subtype M0 is indicated in black, subtype M1 in green, subtype M2 in purple, subtype M3 in orange, subtype M4 in yellow, subtype M5 in blue, and subtype M6 in gray; normal karyotypes are indicated in green, inv(16) abnormalities in yellow, t(8;21) abnormalities in purple, t(15;17) abnormalities in orange, 11q23 abnormalities in blue, 7(q) abnormalities in red, +8 aberrations in pink, complex karyotypes (those involving more than three chromosomal abnormalities) in black, and other abnormalities in gray. FLT3 internal tandem duplication (ITD) mutations, FLT3 mutations in the tyrosine kinase domain (TKD), N-RAS, K-RAS, and CEBPA mutations, and the overexpression of EVI1 are depicted in the same set of columns: red indicates the presence of a given abnormality, and green its absence. The levels of expression of the top 40 genes identified by the significance analysis of microarrays of each of the 16 clusters as well as in normal bone marrow (NBM) and CD34+ cells are shown on the left side. The scale bar indicates an increase (red) or decrease (green) in the level of expression by a factor of at least 4 relative to the geometric mean of all samples. The percentages of the most common abnormalities (those present in more than 40 percent of specimens) and the percentages of specimens in each cluster with a normal karyotype are indicated.

Source: http://content.nejm.org/cgi/content/full/350/16/1617

Many classification algorithms are available to define a set of genes efficiently when using only a sub-selection of genes in the predictor. This may be the top variable genes for example or the results of a class comparison. Selecting the genes to be used in a predictor is very important because the number of genes that are useful to distinguish between classes may be very small relative to the number of genes on the array. The influence of these genes that distinguish between classes may be lost in the total variation of the other genes unless we first select the informative genes in the prediction algorithm [69].

Several multivariate classification methods are available in the BRB-Array tool, including the Compound Covariate Predictor, Diagonal Linear Discriminant Analysis, Nearest Neighbor Predictor, Nearest Centroid Predictor, and Support Vector Machine Predictor. The PAM software [38], developed at Stanford University has also been implemented into the BRB-array tools. Descriptions of these statistical algorithms are given in the BRB-Array manual or can be found on the various websites. Alternatively, most of these algorithms can be utilized within the R statistical environment [38].

1.9.4 Cross-validation

For all class prediction methods, an estimate of how accurately the classes can be predicted by the defined class predictor should be calculated by computing the cross-validated misclassification rate. Leave-one-out cross validation (LOOCV) is often used to overcome this problem of multiple testing. The LOOCV process omits one sample at a time, upon which the entire analysis is repeated from scratch, including determination of which genes are univariately significant on the reduced training sample set. From that gene list, a multivariate predictor is constructed and applied to predict the class of the sample that was omitted. This process is reiterated, leaving out one of the samples at a time. Because of the large number of genes on a microarray and thus candidate predictor variables, it is essential to use cross validation or some similar method to determine whether a defined class predictor indeed predicts accurately. Even with sample sets that do not differ in expression profiles, it is very easy to develop models that predict perfectly when measured in a non cross-validated manner. Such models would be useless for application with independent data sets [70]. LOOCV is an internal validation method for the class predictor. An alternative validation is to have separate prediction and test sets of samples (also called training and validation sets) whereby the predictor is modeled on the predictor set and tested independently on the test set [6].

1.10 Validation of results

There are many caveats to microarray analysis during the planning and experimental stages as well as during statistical analysis and biological interpretation of the results. Technical variation is a very important aspect of microarray analysis. Many experimental steps along the way may introduce unwanted variation, and thus false positives into the final gene list. This means that the genes on the list must be validated in the lab. Procedures such as (transcriptase) RT-PCR and immunohistochemical staining can validate the expression results from microarrays.

1.11 Pattern discovery: ontological classification and pathway analysis

Following statistical analysis, the list of significant genes may reveal what pathways and biological mechanisms distinguish the different classes of samples. The availability of tools such as annotation and pathway knowledge databases has facilitated the interpretation of microarray results. Instead of analyzing the significant genes one at a time, analysis on a global level may show how they all fit together to create a biological story.

The Gene Ontology (GO) consortium has created a database to annotate every gene and gene variant with information relating to how a gene product behaves in a cellular context [71]. Scientists can utilize this ontological annotation system to infer knowledge from large amounts of data [72]. There are three organizing principles of GO: molecular functions of the gene products, their participation in biological processes and their placement in or as cellular components, with any gene classified according to one of each of these principles. GO analysis is a powerful method to gain knowledge of function for the entire human genome, making pattern detection in the data more readily identifiable. A particular cellular process may be affected in the experiment and the knowledge that most of the genes in the dysregulated gene list are involved in that particular function can save a lot of work sorting through large lists of genes. There are many software tools available that can calculate the statistical significance of a particular GO process being over-represented within the significant gene list see Appendix 2 on website.

There are also many pathway analysis tools available today that are invaluable for putting biological meaning to a list of genes resulting from the statistical analysis of microarray data. These tools differ from GO analysis in that you are directly identifying pathways that are involved in the change of gene expression rather than themes of gene annotation. The advancements in genomics mean that we have acquired extensive knowledge to put together the biological

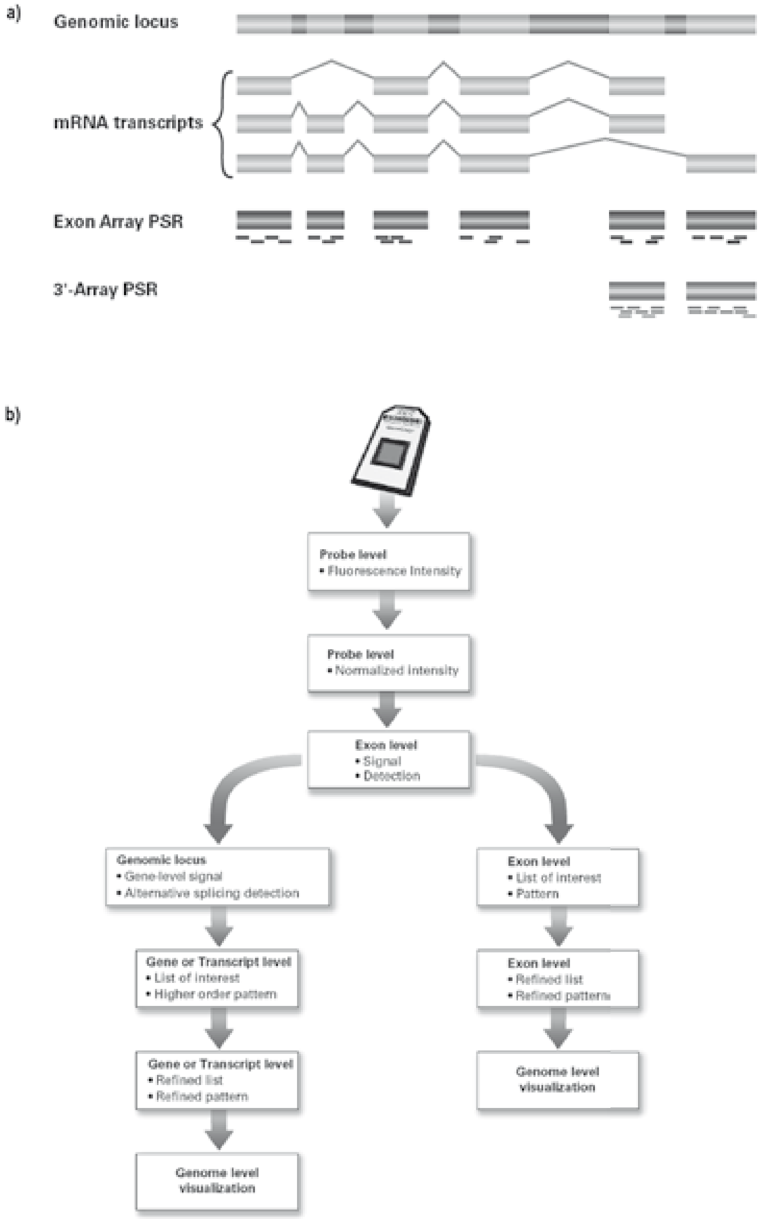


Figure 13: a) Schematic for the coverage of probe sets across the entire length of the transcript. Golden regions are exons whereas the grey regions represent introns that are removed during splicing. The short dashes underneath the exon regions for the exon array and the 3' expression array PSR (Probe Selection Region) indicate individual probes representing that PSR. b) Schematic representation of the exon array analysis workflow.

Figure taken from http://awww.affymetrix.com/technology/manufacturing/index.affx pieces. Pathway analysis tools enable scientists to gain a deeper biological understanding of molecular and cellular mechanisms, which ultimately improves future experimental processes and influences experimental direction and interpretation of results. Table 2 displays a list of very useful pathway analysis tools and Figure 14 illustrates Ingenuity software.

Garinis *et al.* [17], reported a great example of utilizing both gene ontology annotations and pathway analysis to interpret their significant genes. Mice were irradiated with different doses of UV at various time points and the changes in gene expression in skin fibroblasts were analyzed. In their analysis, significant over-representation of GO-classified biological processes in each gene list were calculated using the publicly accessible software EASE [73]. This software compares the number of pertinent genes in a given biological process over the total number of the relevant genes on the array using a Fishers exact test, calculating FDR. They have also used the Ingenuity software [74] to systematically analyze their experimental data in the context of biological pathways. Ingenuity constructs molecular networks centered on the significant genes identified in the study and calculates the probabilistic fit between each network and the observed expression data from the array, assigning a chance probability score to each network. Using both tools for biological discovery, Garinis *et al.* [17] have shown that among UV-absorbing cellular macromolecules, DNA plays the most prominent role in downstream signaling of the damage response. They implicated CPD (cyclobutane pyrimidine dimmers; a predominant lesion caused by short-wavelength UV) -dependent radiation replication products, rather than CPDs themselves, as the primary mediators of the bulk transcriptional response to UV light. Data from these researchers and many others is also being made publicly available for the scientific community to utilize.

1.12 Various types of microarray

http://www.aftymetrix.com outlines the vast number of different microarray platforms that Affymetrix offers encompassing different species but also addressing different biological questions. The most widely used GeneChips are expression arrays (U133-2plus), which have been described throughout this chapter. Gene expression arrays quantify the level of transcripts measured within a sample; however, such arrays only target the sequence at the 3' end of the messenger sequence. A second type of expression array is the exon array (further described in Paragraph 1.12.1), interrogating every exon within a transcript, generating more coverage and the ability to look at events such as alternative splicing. SNP arrays are used to investigate linkage and genetic variation in populations associated with disease [75,76] and are suitable to look at copy number changes in cancers, such as chromosomal deletions and/or amplifications [77-79]. More recently, tiling arrays have been released which not only interrogate every exon with a 35bp resolution but also have the potential to monitor epigenetic regulation such as promoter methylation. A tiling array that has been directed towards known promoter regions, which can interrogate epigenetic regulation and alternative promoter usage, has also been developed. Other commercial companies also offer platforms of microarrays, with slight variations on the Affymetrix technology. Such arrays include protein/antibody arrays and BAC arrays (~1-Mb resolution).

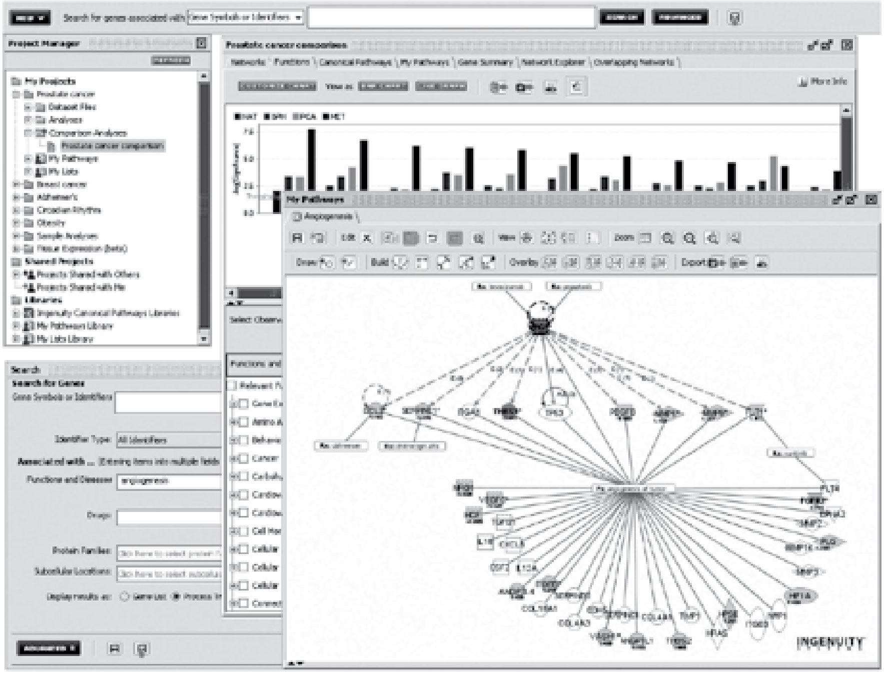


Figure 14: Ingenuity pathway analysis. Ingenuity Pathways Analysis (IPA) 4.0 is a software application that enables identification of biological mechanisms, pathways and functions most relevant to the experimental data sets or genes of interest. Once genomic or proteomic data are generated and statistically analyzed, these datasets can be easily uploaded into the IPA application. The software dynamically computes relevant networks and identifies biological functions or pathways focused on genes of interest. Users can create custom pathways or perform searches for particular targets, disease areas, or biological functions. IPA enables users to gain a deeper biological understanding of molecular and cellular mechanisms.

Figure taken from: http://www.ingenuity.com/docs/Ingenuity_IPA4.0_Dsheet.pdf

1.12.1 Exon arrays

Affymetrix has developed a GeneChip that interrogates transcripts exon-by-exon (HU Exon 1.0 ST). This is the first technology that can investigate the whole-transcriptome on a single array. These arrays can be used to analyze single exon expression, but can also be utilized to look at the transcript expression level by utilizing the annotation associated with each probe set for each exon and calculating the signal intensity of the transcript as a whole (Figure 14). Such a platform can assess thousands of genes simultaneously, permitting large-scale analysis and therefore generating a vast amount of data, causing bioinformatic issues for the regular researcher. The older gene expression microarrays (U133-2plus) profiled 54,000 probesets, whereas the new exon arrays will profile 1.4 million probes. This requires a lot more computer memory to process than the previous arrays as well as handling issues for many of the regular analysis software.

As the microarray technology progressed, limitations in the 3' biased probe set design of the expression arrays were recognized and technology developments were needed to allow coverage of more complex gene transcription. Alternative splicing is an important layer of regulation in eukaryotic gene expression with a qualitative change in the structure of the transcript. A single transcript can have different splice variations that are specific to a certain tissue or disease type, and splice variation of a large number of genes has been implicated in various cell growth and differentiation processes [80-82].

Prior to exon-level expression arrays, other techniques have been used with some success to analyze splice variant expression. These include exon-junction arrays [81], RNA-mediated annealing, selection and ligation (RASL) [83] and digital polony (polymerase colony) exon profiling [84,85]. Recently, arrays containing a combination of exon expression and exon junction probes have also been used to identify alternative splicing events [85,86]. Although all approaches can detect alternative splicing events, many are limited either by screening on a predetermined set of exon-junctions or by screening on a per-gene basis. Various mathematical and statistical methods have also been developed for the analysis of exon array data, mostly based on the assumption that a change in splicing events can appear as a change in expression of one part of the gene versus another [87]. Hu *et al.* identified groups of probes that cluster spatially in the genome with expression levels similar to each other but differing from the average gene expression level [82]. Le *et al.* compared data from two different experimental conditions with data from a third set of experiments, obtained by mixing the two samples in equal portions [88]. The putative splice variants can then be detected from the anti-correlation in the expression levels of the probe sets by the log ratios of the two different samples versus a pool containing both samples. Using this approach they were able to detect a wide variety of tissue-specific alternative splicing events, such as mutually exclusive exons and alternative 3' and alternative 5' splicing, all of which could be validated independently.

The exon clusters represented on the Affymetrix exon array contain sequences from varying sources:

Exons of well-annotated ‘RefSeq’ genes (284,000 core exons).

– Less characterized exons, such as mRNA and ESTs from GenBank and Refseq (523,000 extended exons).
– Exons for which there is no information available in public databases, but have been predicted by gene finder software such as GENSCAN (580,000 full exons).

As part of the pioneering Affymetrix sites at ErasmusMC, we have found that detecting differential splice variants with exon arrays depends of the accurate signal calculation of the metaprobe set (transcript) signal. This can pose a problem when using the full and extended exon probe sets in the analysis as many of these exons are not confirmed as being functional. Only exons expressed above the background threshold, as well as exons in linear relationship with their transcript (i.e. exon signal intensity has a high correlation to the transcript signal intensity) are included as part of the transcript calculations (Chapter 6). Detecting such splice variants in a supervised manner will require knowledge of different subgroups of samples. Applying an algorithm that will calculate the probability of a particular exon being spliced either in or out of a transcript will require further statistics such as a t-test or ANOVA to be applied to give a ranked significance to the results.

The high density Affymetrix exon arrays can be used to study the expression profiles related to many of the following: alternative splicing events (Chapter 6); exon skipping (Chapter 7), intron retention, mutually exclusive exon usage, alternative promoter usage, alternative polyadenylation and alternative splicing donor/acceptor sites with changes over 25 bp. Due to the high coverage of probe sets within each exon, these arrays could also be utilized to identify fusion proteins involved in disease.

In this thesis, various experimental examples using microarray bioinformatics will be shown.

Refer also to http://www-bioinfo.erasmusmc.nl/thesis_peeters for a list of useful tools and websites.

1.13 Bioinformatics bibliography

1. Simon, R., *Bioinformatics in cancer therapeutics-hype or hope?* Nat Clin Pract Oncol, 2005. **2**(5): p. 223.
2. http://www.dictionary.com. [cited.
3. Mullis, K., et al., *Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction.* Cold Spring Harb Symp Quant Biol, 1986. **51 Pt 1**: p. 263-73.
4. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml. [cited.
5. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
6. Christie, J.D., *Microarrays.* Crit Care Med, 2005. **33**(12 Suppl): p. S449-52.
7. Ewis, A.A., et al., *A history of microarrays in biomedicine.* Expert Rev Mol Diagn, 2005. **5**(3): p. 315-28.
8. Peeters, J.K. and P.J. Van der Spek, *Growing applications and advancements in microarray technology and analysis tools.* Cell Biochem Biophys, 2005. **43**(1): p. 149-66.
9. Glass, A., et al., *Representation of individual gene expression in completely pooled mRNA samples.* Biosci Biotechnol Biochem, 2005. **69**(6): p. 1098-103.
10. Jolly, R.A., et al., *Pooling samples within microarray studies: a comparative analysis of rat liver transcription response to prototypical toxicants.* Physiol Genomics, 2005. **22**(3): p. 346-55.
11. Kendziorski, C.M., et al., *The efficiency of pooling mRNA in microarray experiments.* Biostatistics, 2003. **4**(3): p. 465-77.
12. Shih, J.H., et al., *Effects of pooling mRNA in microarray class comparisons.* Bioinformatics, 2004. **20**(18): p.318-25.
13. Zakharkin, S.O., et al., *Sources of variation in Affymetrix microarray experiments.* BMC Bioinformatics, 2005. **6**: p. 214.
14. Verona, E.V., et al., *Transforming Growth Factor-{beta} Signaling in Prostate Stromal Cells Supports Prostate Carcinoma Growth by Up-regulating Stromal Genes Related to Tissue Remodeling.* Cancer Res, 2007. **67**(12): p. 5737-46.
15. Tamura, K., et al., *Molecular features of hormone-refractory prostate cancer cells by genome-wide gene expression profiles.* Cancer Res, 2007. **67**(11): p. 5117-25.
16. Chen, J., *Identification of significant periodic genes in microarray gene expression data.* BMC Bioinformatics, 2005. **6**: p. 286.
17. Garinis, G.A., et al., *Transcriptome analysis reveals cyclobutane pyrimidine dimers as a major source of UV-induced DNA breaks.* Embo J, 2005. **24**(22): p. 3952-62.
18. Yu, K., et al., *A modular analysis of breast cancer reveals a novel low-grade molecular signature in estrogen receptor-positive tumors.* Clin Cancer Res, 2006. **12**(11 Pt 1): p. 3288-96.
19. Rossetti, S., et al., *A distinct epigenetic signature at targets of a leukemia protein.* BMC Genomics, 2007. **8**: p. 38.
20. Grade, M., et al., *Aneuploidy-dependent massive deregulation of the cellular transcriptome and apparent divergence of the Wnt/beta-catenin signaling pathway in human rectal carcinomas.* Cancer Res, 2006. **66**(1): p. 267-82.
21. Rolland, P., et al., *Human leukocyte antigen class I antigen expression is an independent prognostic factor in ovarian cancer.* Clin Cancer Res, 2007. **13**(12): p. 3591-6.
22. Turashvili, G., et al., *Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis.* BMC Cancer, 2007. **7**: p. 55.
23. Harmer, S.L. and S.A. Kay, *Microarrays: determining the balance of cellular transcription.* Plant Cell, 2000. **12**(5): p. 613-6.
24. Millenaar, F.F., et al., *How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results.* BMC Bioinformatics, 2006. **7**: p. 137.
25. Lipshutz, R.J., et al., *High density synthetic oligonucleotide arrays.* Nat Genet, 1999. **21**(1 Suppl): p. 20-4.
26. http://www.affymetrix.com. *Affymetrix.* [cited.
27. Lipshutz, R.J., *Applications of high-density oligonucleotide arrays.* Novartis Found Symp, 2000. **229**: p. 84-90; discussion 90-3.
28. Han, E.S., et al., *Reproducibility, sources of variability, pooling, and sample size: important considerations for the design of high-density oligonucleotide array experiments.* J Gerontol A Biol Sci Med Sci, 2004. **59**(4): p. 306-15.

29. Clarke, J.D. and T. Zhu, *Microarray analysis of the transcriptome as a stepping stone towards understanding biological systems: practical considerations and perspectives.* Plant J, 2006. **45**(4): p. 630-50.

30. http://www.bioconductor.org. [cited.

31. http://www.r-project.org/. [cited.

32. http://www.spotfire.com. [cited.

33. http://www.omniviz.com. [cited.

34. http://www.rosettabio.com/products/resolver/default.htm.

35. Wong, C.L.a.W.H., *DNA-Chip Analyzer (dChip)*.. in *In The analysis of gene expression data: methods and software*, E.G. G Parmigiani, R Irizarry and SL Zeger. . Editor. 2003, Springer.

36. http://biosun1.harvard.edu/complab/dchip/. [cited.

37. Zhong, S., C. Li, and W.H. Wong, *ChipInfo: Software for extracting gene annotation and gene ontology information for microarray analysis.* Nucleic Acids Res, 2003. **31**(13): p. 3483-6.

38. http://linus.nci.nih.gov/BRB-ArrayTools.html. [cited.

39. Verhaak, R.G., et al., *The effect of oligonucleotide microarray data pre-processing on the analysis of patient-cohort studies.* BMC Bioinformatics, 2006. **7**: p. 105.

40. Leung, Y.F. and D. Cavalieri, *Fundamentals of cDNA microarray data analysis.* Trends Genet, 2003. **19**(11): p. 649-59.

41. Szabo, A., et al., *Statistical modeling for selecting housekeeper genes.* Genome Biol, 2004. **5**(8): p. R59.

42. Vandesompele, J., et al., *Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes.* Genome Biol, 2002. **3**(7): p. RESEARCH0034.

43. Kroll, T.C. and S. Wolfl, *Ranking: a closer look on globalisation methods for normalisation of gene expression arrays.* Nucleic Acids Res, 2002. **30**(11): p. e50.

44. Giovanni Parmigiani, E.S.G., Rafael A Irizarry, Scott L Zeger, *The Analysis of Gene Expression Data: Methods and Software*. 2003: Springer.

45. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Res, 2003. **31**(4): p. e15.

46. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-93.

47. http://rmaexpress.bmbolstad.com/. [cited.

48. Fan, W., et al., *A class of models for analyzing GeneChip gene expression analysis array data.* BMC Genomics, 2005. **6**(1): p. 16.

49. Naef, F. and M.O. Magnasco, *Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays.* Phys Rev E Stat Nonlin Soft Matter Phys, 2003. **68**(1 Pt 1): p. 011906.

50. Huber, W., et al., *Variance stabilization applied to microarray data calibration and to the quantification of differential expression.* Bioinformatics, 2002. **18 Suppl 1**: p. S96-104.

51. Draghici, S., *Data Analysis Tools for DNA Microarrays*. 2003, London: Chapman and Hall.

52. http://www.stat.berkeley.edu/users/terry/zarray/Html/log.html.

53. D'Haeseleer, P., *How does gene expression clustering work?* Nat Biotechnol, 2005. **23**(12): p. 1499-501.

54. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.

55. http://genome.ws.utk.edu/. *University of Tennessee Microrray Database (UTMD)*. [cited.

56. Gibbons, F.D. and F.P. Roth, *Judging the quality of gene expression-based clustering methods using gene annotation.* Genome Res, 2002. **12**(10): p. 1574-81.

57. Valk, P.J., et al., *Prognostically useful gene-expression profiles in acute myeloid leukemia.* N Engl J Med, 2004. **350**(16): p. 1617-28.

58. Wright, G.W. and R.M. Simon, *A random variance model for detection of differential gene expression in small microarray experiments.* Bioinformatics, 2003. **19**(18): p. 2448-55.

59. http://www-stat.stanford.edu/~tibs/SAM/. [cited.

60. Benjamini, Y.a.H., Y., *Controlling the false discovery rate; a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.

61. Wettenhall, J.M. and G.K. Smyth, *limmaGUI: a graphical user interface for linear modeling of microarray data.* Bioinformatics, 2004. **20**(18): p. 3705-6.

62. Pounds, S.B., *Estimation and control of multiple testing error rates for microarray studies.* Brief Bioinform, 2006.

7(1): p. 25-36.

63. http://bioinfo.cnio.es/help/. [cited.

64. Gyorffy, B., A. Gyorffy, and Z. Tulassay, *[The problem of multiple testing and solutions for genome-wide studies].* Orv Hetil, 2005. **146**(12): p. 559-63.

65. Hochberg, Y. and Y. Benjamini, *More powerful procedures for multiple significance testing.* Stat Med, 1990. **9**(7): p. 811-8.

66. Bretz, F., J. Landgrebe, and E. Brunner, *Multiplicity issues in microarray experiments.* Methods Inf Med, 2005. **44**(3): p. 431-7.

67. Goeman, J., *Statistical methods for microarray data.*, in *Medical Statistics LUMC*. 2006, Leiden University medical centre.

68. Beer, D.G., et al., *Gene-expression profiles predict survival of patients with lung adenocarcinoma.* Nat Med, 2002. **8**(8): p. 816-24.

69. Simon, R., *Using DNA microarrays for diagnostic and prognostic prediction.* Expert Rev Mol Diagn, 2003. **3**(5): p. 587-95.

70. Simon, R., et al., *Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification.* J Natl Cancer Inst, 2003. **95**(1): p. 14-8.

71. http://www.geneontology.org/. [cited.

72. *The Gene Ontology (GO) project in 2006.* Nucleic Acids Res, 2006. **34**(Database issue): p. D322-6.

73. Hosack, D.A., et al., *Identifying biological themes within lists of genes with EASE.* Genome Biol, 2003. **4**(10): p. R70.

74. http://www.ingenuity.com. [cited; Entrez gene summary].

75. Fitzgibbon, J., et al., *Association between acquired uniparental disomy and homozygous gene mutation in acute myeloid leukemias.* Cancer Res, 2005. **65**(20): p. 9152-4.

76. Tsafrir, D., et al., *Relationship of gene expression and chromosomal abnormalities in colorectal cancer.* Cancer Res, 2006. **66**(4): p. 2129-37.

77. Stark, M. and N. Hayward, *Genome-wide loss of heterozygosity and copy number analysis in melanoma using high-density single-nucleotide polymorphism arrays.* Cancer Res, 2007. **67**(6): p. 2632-42.

78. Miller, L.D., et al., *An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival.* Proc Natl Acad Sci U S A, 2005. **102**(38): p. 13550-5.

79. Wong, K.K., et al., *Genome-wide allelic imbalance analysis of pediatric gliomas by single nucleotide polymorphic allele array.* Cancer Res, 2006. **66**(23): p. 11172-8.

80. Kampa, D., et al., *Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.* Genome Res, 2004. **14**(3): p. 331-42.

81. Johnson, S., et al., *Genomic organisation and alternative splicing of human RIM1, a gene implicated in autosomal dominant cone-rod dystrophy (CORD7).* Genomics, 2003. **81**(3): p. 304-14.

82. Hu, G.K., et al., *Predicting splice variant from DNA chip expression data.* Genome Res, 2001. **11**(7): p. 1237-45.

83. Yeakley, J.M., et al., *Profiling alternative splicing on fiber-optic arrays.* Nat Biotechnol, 2002. **20**(4): p. 353-8.

84. Zhu, J., et al., *Single molecule profiling of alternative pre-mRNA splicing.* Science, 2003. **301**(5634): p. 836-8.

85. Fehlbaum, P., et al., *A microarray configuration to quantify expression levels and relative abundance of splice variants.* Nucleic Acids Res, 2005. **33**(5): p. e47.

86. Nagao, K., et al., *Detecting tissue-specific alternative splicing and disease-associated aberrant splicing of the PTCH gene with exon junction microarrays.* Hum Mol Genet, 2005. **14**(22): p. 3379-88.

87. Sugnet, C.W., et al., *Unusual intron conservation near tissue-regulated exons found by splicing microarrays.* PLoS Comput Biol, 2006. **2**(1): p. e4.

88. Le, K., et al., *Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data.* Nucleic Acids Res, 2004. **32**(22): p. e180.

Chapter 2

Introduction to Cancer

2.0 Introduction to Cancer

2.1 Cancer

In adult individuals, the rate of cell proliferation is essentially proportional to the rate of cell death, maintaining a constant number of cells. Normal cell division is driven by a variety of signals, such as diminished cell-cell contact or stimulation by growth factors. Whatever the initiating signals, they typically trigger a cascade of signaling events that transfer the signal to the nucleus to activate gene transcription and cell proliferation. When the rate of cell proliferation exceeds that of cell death, carcinogenesis is initiated [1]. The initiation of this carcinogenesis begins with the accumulation of genetic and epigenetic alterations in the genome, affecting various pathways.

Cancers in human adults are classified according the cell type in which they arise: epithelial, mesenchymal or hematological. The most frequent are cancers that arise in epithelial cells, commonly referred to as carcinomas. These include most cancers from the breast, prostate, lung and colon. Cancers in mesenchymal cells are referred to as soft tissue tumors, including brain tumors and other central nervous system tumors, sarcomas and blastomas. Hematological or immunological tumors include leukemia's, lymphomas and myelomas. This thesis concerns gene expression microarray analysis of breast cancers and brain tumors and these tumor types will therefore be discussed in more detail.

2.2 Breast cancer

2.2.1 Normal breast histology

Breasts consist of a network of ductal structures that terminate in lobular units, all embedded in an extensive stromal component (Figure 1a, b). Two epithelial cell layers line the ducts and lobules: an inner layer of glandular epithelial (or luminal) cells and an outer layer of myoepithelial (or basal) cells attached to the basement membrane (Figure 1c). The luminal epithelial layer characteristically expresses markers such as luminal cytokeratins (CK7, 8, 18 and 19), estrogen receptor alpha (ER), MUC1 and integrin alpha-6 [2]. The morphology of the cells from this layer is typically cuboidal. The basal epithelial layer characteristically expresses basal cytokeratins (CK5, 14 and 17) as well as vimentin, epidermal growth factor receptor (EGFR), smooth muscle actin (SMA), caldesmon, p63 and CD10. The morphology of these cells is typically spindle, exhibiting features of both epithelial and smooth muscle cells [2].

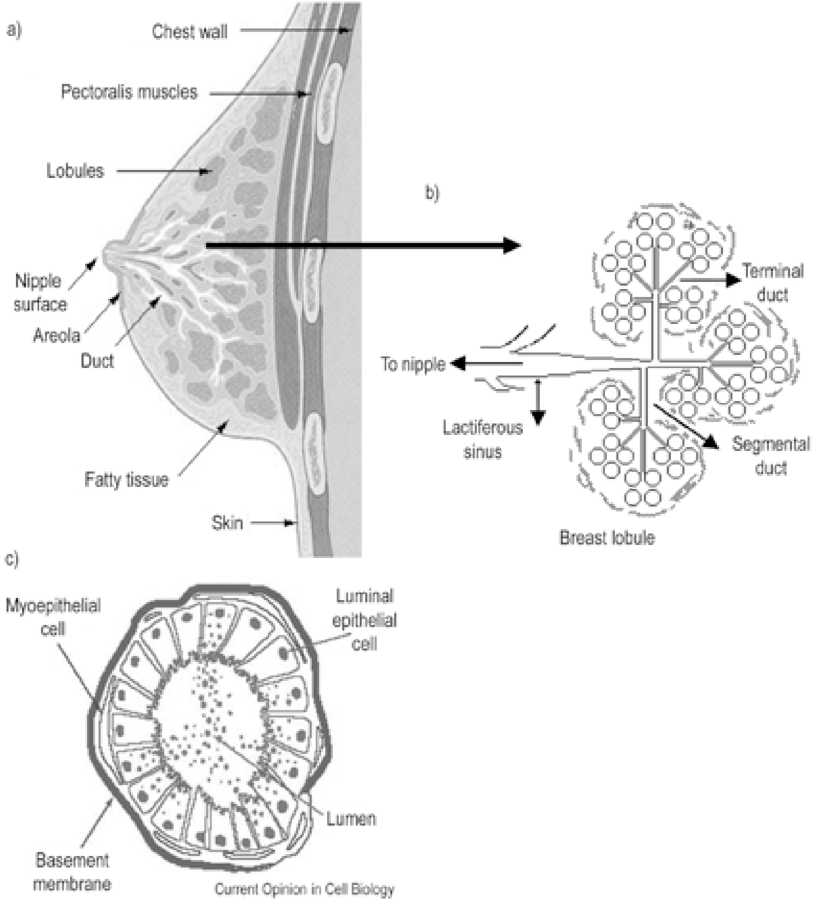


Figure 1: Schematic diagram of the breast showing histology. a) and b) Normal histology of breast tissue consists of a network of ductal structures that terminate in lobular units. Within the lobules are small acini. Lobules are connected to intralobular ductules (segmental ducts) and interlobular (terminal) ducts. Lobules are surrounded by loose connective tissue sensitive to sex hormones. c) The ducts are lined by an inner layer of secretory basal epithelial cells and an outer layer of myoepithelial cells, attached to a basement membrane.

Figures adapted from http://training.seer.cancer.gov/ss_module01_breast/unit02_sec01_anatomy.html. http://www.dailycal.org/printable.php?id=20448

Estrogen signaling is arguably the most important pathway for proliferation regulation of the ductal epithelia in the breast [3]. Estrogen levels vary during the menstrual cycle and during pregnancy, thus controlling epithelial cell proliferation. Upon binding of its estrogen ligand, the estrogen receptor (ER) is translocated to the nucleus where is serves as a transcriptional activator. Estrogen receptors may be formed by three possible combinations of dimers: ERα (αα), ERβ (ββ) and ERαβ (αβ), where the alpha receptor is encoded by the *ESR1* gene at chromosome 6q25.1 and the beta receptor by *ESR2* at 14q23.2. Yet, ERα appears most important for proliferation in normal breast epithelia as well as in breast cancer and is therefore still commonly referred to simply as ER [4].

2.2.2 Malignant breast histology

The majority of breast cancers develop from the epithelial cells in the so-called terminal duct lobular units, where the ducts end in lobules (Figure 1a, b). Similar to the two epithelial layers in normal breast tissues, breast cancers can also be characterized by their expression of basal and luminal cytokeratins. Ellis and colleagues have reported several immunohistochemical studies on a consecutive cohort of over 1800 breast cancers [2,5-8]. Based mainly on their cytokeratin expression patterns, they identified four subtypes of breast cancers:

1) Luminal phenotype (expressing one or more luminal cytokeratins), accounting for 71.4% of breast cancers. These tumors are mostly ER-positive.
2) Combined luminal and basal phenotype (expressing both luminal and basal cytokeratins), accounting for 27.4%. These tumors are mostly ER-negative and frequently express EGFR.
3) Basal phenotype (expressing basal cytokeratins), accounting for 0.8%.
4) Null phenotype (no expression of either luminal or basal cytokeratins), accounting for 0.4%.

Several other groups have classified breast cancers by similar, yet slightly different marker expression patterns. Nielson *et al.* classified breast cancers by 4 antibodies: ERBB2-positive breast cancers; ERBB2-negative and ER-positive breast cancers; ERBB2/ER-negative and EGFR or CK5/6-positive breast cancers; and breast cancers that are negative for all 4 proteins [9]. Livasy *et al.* have reported that the basal profile consisted of being ER and ERBB2 negative and vimentin, EGFR or CK8/18/5/6 positive [10]. Rakha *et al.* proposed that basal breast cancers may be defined solely by their expression of basal cytokeratins, regardless of expression of other markers [5]. Although these breast cancer classifications are similar, their major difference is in the definition of basal breast cancers. Indeed, there is recurrent confusion on basal breast cancers in the scientific literature.

2.2.3 Incidence and risk factors

About one in ten women in the western world develop breast cancer, accounting for one third of the total cancer incidence in women. In the Netherlands, more than 11,000 cases of breast cancer were diagnosed in 2003 [11]. Major risk factors for breast cancer include gender (high frequency of breast cancer in women and only rarely in men); a family history of breast cancer (particularly when carrying a mutant breast cancer susceptibility gene); and age (the incidence of breast cancer doubles with each decade of life with a slower increase after menopause) [12]. Minor risk factors are related to life-style (including diet, weight, alcohol consumption and smoking); estrogen hormone exposure (such as age at menarche/menopause, pregnancy and breast feeding) [13]; and possibly also ER-related genetic susceptibility (including various ER SNPs and ER splice variants and expression level variations) [4,14-16].

2.2.4 Prognosis and therapy

Almost two-thirds of breast cancer patients are diagnosed without evidence of metastasis [17]. These patients have a good prognosis since primary breast cancers can be cured by surgical resection and radiotherapy. However, cancer recurs in about one-quarter of such lymph node-negative patients. Metastasis is potentially life threatening and can only be treated by systemic treatment. Treatment options are guided mainly by tumor grade and size and by histological subtype [17]. Most importantly, patients with ER-positive breast cancers generally have a better prognosis and patients with ER-negative breast cancers tend to have a more aggressive course of disease [3,18]. The prognosis of ER-positive and ER-negative breast cancers is strongly associated with the luminal and basal histological phenotypes of the tumors. Among luminal ER-positive breast cancers, patients that are diagnosed before age 35 years often have a worse prognosis [8]. Among basal phenotype breast cancers, patients with triple negative tumors (ER/PR/ERBB2-negative) but expression of EGFR have a particularly worse prognosis [6,19,20].

ER-positive breast cancers are typically treated with hormonal therapy, either inhibiting estrogen action (with *Tamoxifen*) or inhibiting estrogen production (with aromatase inhibitors) [13,21,22]. ER-negative breast cancers are treated with chemotherapy, often given as a combination of drugs. Although hormonal therapies and chemotherapies all aim at inhibition of cell growth, the latter are directed at more basic cellular proliferation processes and thus involve more side-effects in normal, non-malignant cells [22]. As more becomes known about the molecular causes of breast cancer, targeted therapies will also become available. In patients with ERBB2 over-expressing breast cancers, for example, *Herceptin* is being used to block the activity of the ERBB2 protein [23].

2.3 Brain tumors

2.3.1 Brain tumor pathology

Primary brain tumors are those initiated of cells within the central nervous system (CNS). Gliomas are tumors originating in the neuroglia of the brain or spinal cord and account and are classified according to the type of normal, non-malignant brain cell they resemble astrocytomas, oligodendrogliomas, and a mixed oligoastrocytomas [24]. The WHO (World Health Organization) has also dedicated a grading system to gliomas based on their histopathology: Grade I gliomas are predominantly pediatric tumors; Grade II consists of astrocytomas, oligodendrogliomas and mixed oligoastrocytoma; Grade III consists of anaplastic astrocytoma, anaplastic oligodendroglioma and mixed anaplastic oligoastrocytoma; and grade IV consists of glioblastoma multiforme (GBM) [25,26]. Low-grade tumors (I and II) are slower growing and consist of well differentiated cells [26]. High-grade tumors (III and IV) display little cellular differentiation and thus lack defining histological features. Unfortunately, histological variability is common among gliomas and diagnosis may thus be challenging [26,27].

Ependymomas are rare tumors, representing 5-10% of gliomas [24]. Ependymal cells line the cavities of the brain and the canal containing the spinal cord. Ependymomas usually arise from the floor of the fourth ventricle and typically obstruct the flow of cerebrospinal fluids, causing headache, nausea and/or vomiting, and sometimes hydrocephalus. Ependymomas can also be divided into two main forms; grade II ependymomas; and grade III anaplastic ependyomas [28].

Medulloblastoma is the most common malignant primary brain tumor in childhood representing 30% of pediatric brain tumors and 8-10% of adult brain tumors [29]. Although it is thought that medulloblastomas originate from immature or embryonal cells at their earliest stage of development, the exact cell of origin or “medulloblast” has yet to be identified. Medulloblastomas usually form in the fourth ventricle, between the brainstem and the cerebellum [26, 30]. Today, medulloblastomas are often referred to as primitive neuroectodermal tumor (PNET) rather than glioma.

2.3.2 Incidence and risk factors

Brain tumors account for 1-2% of all adult malignancies [30]. In the Netherlands, over 900 primary brain tumors were diagnosed in 2003 [11] (http://www.ikcnet.nl/page.php?id=1872&nav_id=97). The only established environmental risk factor for the development of gliomas is radiation. Today, most radiation-induced brain tumors (predominantly meningiomas) are caused by radiation to the head received for the treatment of other cancers [31]. Genetic alterations in a number of genes can predispose individuals to a variety of tumors including brain tumors (NF1, NF2, PTEN, P53) and evidence suggests the presence of an ependymoma susceptibility gene on chromosome 22 [26,30].

2.3.3 Prognosis and therapy

Adult individuals diagnosed with a glioma typically have a dismal prognosis. For example, patients with a glioblastoma multiforme, the most common type of glioma, have a median survival of 12 months with less than 4% of patients surviving five years post diagnosis [24,30]. The median survival of patients with an astrocytoma grade II is approximately four years, which is a better prognosis than patients with higher-grade astrocytomas. Low-grade gliomas are well differentiated, slower growing, biologically less aggressive, and portend a better prognosis for the patient. Although challenging, histopathological diagnosis currently is the most reliable guide for prognostication and treatment decisions [28].

The typically poor prognosis of glioma patients is mostly due to the infiltrative nature of these tumors [27]. The high complexity of the brain and its functions makes complete surgical resection of adult gliomas very difficult. Brain tumor treatments therefore frequently involve a combined approach of surgery, radiation therapy and chemotherapy. Surgery can be quite dangerous and is limited to regions that will not damage critical neurological functions. Recent treatments for brain tumors also involve molecularly identified targets, such as *Gefitinib* (*Iressa)* which targets EGFR, but only 10 to 20 percent of patients have a response to such EGFR kinase inhibitors [32]. The mechanism of responsiveness of glioblastomas to these inhibitors is currently unknown. Studies have found that tumors with activating mutations in the *EGFR* kinase domain (such as the *EGFRvIII* splice variant that lacks exons 2-7) are responsive to kinase inhibitor treatments, whereas tumors overexpressing EGFR due to genetic amplification do not respond [33,34]. Mellinghoff *et al*, reported that co-expression of *EGFRvIII* and mutant *PTEN* by glioblastoma cells was associated with responsiveness to EGFR kinase inhibitors [32].

2.4 The genetics of cancer

2.4.1 Accumulation of mutations in several genes

The development of cancer is caused by the accumulation of several genetic changes [35]. In 1969, Ashley was the first to propose and calculate that common cancers may be caused by mutations in a number of genes, varying from 3-7 for specific types of cancer [36,37]. In the early 1990's, Vogelstein proposed a tumor progression model for colorectal cancer, involving the accumulation of mutations in at least seven genes and with a preferred sequence in which these mutations occur. More recently, Vogelstein and colleagues sequenced 13,000 protein-encoding genes in breast cancers and colon cancers and showed that these tumors had at average 93 mutated genes. About ten of these mutations were thought to have driven carcinogenesis whereas the other mutations represent irrelevant passenger mutations [38-40]. Extrapolating these results to all 30,000 genes in the human genome and taking into account that their approach did not detect sizeable deletions, amplifications and complex rearrangements, there may be at least 20 mutant genes that have a critical involvement in the development of each cancer [40].

2.4.2 Somatic and germline mutations

Most mutations in a cancer are of somatic origin. These mutations arise due to errors during the DNA replication process or due to DNA damage by environmental or endogenous carcinogenic agents, such as radiation, tobacco smoke or free oxygen radicals [41]. Mutations may also be inherited through the germline, conferring an increased susceptibility to develop cancer in mutation carriers. About 10% of all human cancers are thought to have a significant involvement of germline mutations. In breast cancer this may be even higher, with 13% of breast cancer patients having one or more first degree relatives with breast cancer, implying an estimated two-fold increased risk of breast cancer for first degree relatives of breast cancer patients [12,42]. Importantly, familial forms of cancer have provided important insights into the molecular events that underlie cancer because often the same genes are involved in both familial and sporadic forms of a particular cancer type.

2.4.3 Cancer genes

The genes that are mutated in cancers normally function in cellular processes that regulate cell proliferation and cell death, but also processes that regulate cell differentiation and cell integrity [43]. The mutations in cancer genes mimic normal activation or inhibition of the signaling events in these processes, thus resulting in constitutive stimulation of cell growth [1]. Cancer genes have been classified according the resulting effect of their mutations: oncogenes, tumor suppressor genes and stability genes.

2.4.3.1 Oncogenes

A proto-oncogene is a normal gene (such as EGFR, ERBB2, RAS and MYC) that may become activated as an oncogene due to mutations. Proto-oncogenes code for proteins that normally stimulate cell division or inhibit cell differentiation. Upon mutation, a proto-oncogene (or its product) becomes a tumor inducing agent; an oncogene [44]. Mutations in oncogenes may arise from abnormal recombination events such as translocations and amplifications, from chromosomal gains, or due to missense mutations in the coding gene sequence.

By definition, mutations in oncogenes are dominant, conferring a 'gain of function' to the encoded protein [43]. An analogy is that growth of a cell is like a driving automobile [45]. A mutation in an oncogene can be seen as a broken accelerator: the car still moves on even when the driver removes his foot from the accelerator.

2.4.3.2 Tumor Suppressor genes

Products of tumor suppressor genes (TSG's) normally have a repressive effect on the regulation of the cell cycle or promote apoptosis. A disruption to these genes can increase the probability of forming a tumor. Mutations in TSG's are recessive, resulting in a 'loss of function' of the encoded protein [43]. A mutation in a TSG is analogous to a dysfunctional brake in an automobile: the car doesn't stop even when the driver attempts to engage it [45]. Due to the recessive nature of TSG's, total inactivation of the encoded function involves mutation of both alleles of the gene, also known as Knudson's two-hit hypothesis [37,46-48]. Bi-allelic inactivation may involve a combination of any of the following modes of inactivation: an intragenic mutation, loss of a substantial chromosomal region that involves the gene (also called loss of heterozygosity; LOH), or epigenetic inactivation through promoter hypermethylation or miRNA mediated gene silencing [37,43,46-49]. An intragenic mutation may involve a nucleotide substitution generating another amino acid residue or a stop codon (missense and nonsense mutations), the deletion or insertion of one or several nucleotides, or a splice site mutation [37,46,48]. LOH may involve loss of an entire chromosome due to failure of the chromosomes to segregate properly at mitosis (non-disjunction) or loss of a smaller chromosomal region or a chromosomal arm due to an unbalanced exchange of genetic material during chromosomal translocation. Although Knudson's two-hit hypothesis is widely accepted, TSG's may also exert a selective growth advantage on a cell when only a single allele is inactivated, a process known as haploinsufficiency [49].

2.4.3.3 Stability genes

Stability genes control the mutation rate in the genome and are therefore often referred to as 'guardians of the genome' [45]. This class of cancer genes includes DNA repair genes involved in mismatch repair (MMR), nucleotide-excision repair (NER) and base-excision repair (BER). Mutations in stability genes thus do not directly control cell proliferation or cell death, but their inactivated or impaired function results in an increased mutation rate in other genes, including oncogenes and TSG's [45,50]. In the absence of an intact MMR system, for example, deficient cells accumulate somatic mutations at a rate some 1000 times faster than normal cells [41]. In the analogy to automobiles, a mutation in a stability gene represents an inept mechanic, resulting in various kinds of technical problems in the car's machinery [45]. Inactivation of stability genes involves mutations similar to those found in TSG's, and their inactivation will be mostly bi-allelic although haploinsufficiency has also been observed [51-53].

2.4.3.4 Epigenetic regulation

The regulation of gene function within cells cannot only be changed by physical changes in the DNA caused by mutation but can also be changed by epigenetic regulation; where changes in gene function occur without the effects being coded in the DNA sequence. Two of these mechanisms; DNA methylation and microRNA's are described below:

DNA methylation and demethylation are important epigenetic mechanisms that regulate changes in the methylation status of cytosine bases (C) within the DNA itself. These mechanisms occur at sites of CpGs clusters called "CpG-islands" in the promoters of genes and have been associated with gene silencing [54,55]. CpG methylation profoundly influences many processes including transcriptional regulation, genomic stability, chromatin structure modulation and X chromosome inactivation to promote genomic integrity and ensure proper temporal and spatial gene expression during development [56]. In contrast to the normal cell, in which approximately 35% of the genome is methylated, the methylation pattern in a cancer cell is disrupted. The bulk of the genome becomes hypomethylated, in particular the normally hypermethylated and silent regions containing repetitive elements are substantially demethylated. Conversely the normally unmethylated CpG island-containing genes often become hypermethylated and silenced. In the field of breast cancer; multiple changes in methylation of cancer related genes have been reported such as *E-cadherin*.

MicroRNA's (miRNA) are very short single-stranded RNA molecules, which are encoded by genes (often found in the introns of genes) that are transcribed from DNA but not translated into protein (non-coding RNA); Mature miRNA molecules are partially complementary to one or more messenger RNA (mRNA) molecules, and their main function is to downregulate gene expression [57]. Numerous miRNA's are deregulated in human cancers, and experimental evidence has indicated that they can play comparible roles as oncogenes or tumor suppressor genes in having an important role in the transformation of malignant cells [57,58]. MiR-21, is an miRNA located at chromosome 17q23 in a chromosomal region frequently amplified in human cancer, in particular breast cancer and glioblastoma's [57,59].

2.4.4 Cancer genes in breast cancer and brain tumors

The Cancer Gene Census has currently listed over 350 genes that have been causally implicated in human cancer [60]. Well known examples of cancer genes that are of relevance to either breast cancer or brain tumors includes the *EGFR* and *ERBB2* oncogenes, *TP53*, *E-cadherin* and *PTEN* TSG's, as well as those TSG's that are well known for their involvement in breast cancer susceptibility: *BRCA1*, *BRCA2* and *CHEK2*.

2.4.4.1 *EGFR* and *ERBB2*

EGFR and *ERBB2* are both members of the ERBB/HER receptor family of receptor tyrosine kinases. EGFR (epithelial growth factor receptor, also known as HER1) is normally involved in the control of cell growth and differentiation [61]. Its transmembrane receptor protein is a tyrosine kinase that is activated by binding of its ligand EGF, or by other members of the EGF family such as TGF-alpha. Activation of EGFR causes it to form homodimers which will then interact and form hetrodimers with other members of the ERBB receptor family (such as ERBB2) [62]. This binding activates a signal transduction to the nucleus, inducing cell proliferation primarily by the MAPK and JNK pathways [33,62]. The *EGFR* gene was shown to be amplified in brain tumors and prostate cancers, and activating intragenic mutations were found in small cell lung cancers and many mutations have been found in gliomas [33,63]. The *EGFRvIII* isoform represents a mutant *EGFR* receptor from which exons 2 through 7 are deleted. The in-frame deletion of exons 2-7 involves the extra-cellular domain of EGFR, resulting in a constitutively active receptor. The *EGFRvIII* isoform was found in almost half of glioblastomas PLEASE CHECK gliomas and NON- Please check small cell lung cancers, always following amplification of the wild-type *EGFR* isoform [64,65].

ERBB2 (also known as HER2 or NEU) has no ligand-binding domain, but binds to other ligand-bound ERBB family members to form heterodimers [61]. Dimerization stabilizes ligand binding and enhances receptor tyrosine kinase-mediated activation of downstream signaling pathways [61]. The *ERBB2* gene is commonly amplified in breast cancers and endometrial cancers, and co-amplified with *GRB7* in gastric cancers and prostate cancers [66]. Activating mutations in the kinase domain of the receptor were found in a small fraction of lung cancers (5-10%) [54,67-69]. *ERBB2* overexpressing breast cancers, as well as gastric cancers and prostate cancers, are known to be more aggressive and more often metastasize [70-72]. *ERBB2* overexpressing breast cancers are also less sensitive to chemotherapy and hormonal therapy [23,73-75].

2.4.4.2 *TP53*

TP53 is perhaps the best-known tumor suppressor gene. TP53 protein is a transcription factor that is activated by phosphorylation upon sensing DNA damage in the G1 phase of the cell cycle [61]. Consequently, the cell cycle is temporarily stalled to allow repair of the damaged DNA, or the cell goes in apoptosis when the DNA is damaged beyond repair. In cancer, mutational inactivation of TP53 function allows undisturbed progress through the G1-S cell cycle checkpoint even in the presence of damaged DNA, generating a favorable environment for the accumulation of mutations in other cancer genes [76,77]. Somatic mutations in *TP53* are found in about half of human cancers, with mutation frequencies varying from 30-70% [78-80]. Interestingly, *TP53* mutations were identified in 30% of ER-positive breast cancers and in 70% of ER-negative breast cancers and mutant TP53 tumors had an unfavorable disease outcome [81]. *TP53* mutations were also identified in about 30% of astrocytoma and secondary glioblastoma multiforme subtypes [79,82]. Importantly, abrogation of the TP53 pathway through homozygous deletion of *p16/p14ARF* was observed in more aggressive high grade glioblastomas and oligodendrogliomas, but not in the lower grade astrocytomas [83], again suggesting an association between TP53 pathway inactivation and worse clinical outcome. Germline *TP53* mutations cause the rare Li-Fraumeni syndrome that is characterized by a wide variety of tumor types, including sarcomas, brain tumors and breast cancers [84]. About half of *TP53* mutations generate a premature termination in the encoded transcripts, resulting in down-regulation of the transcripts and absence of protein expression. The other half is missense mutations that typically locate in the DNA-binding domain of *TP53*, resulting in impaired or no transcriptional activation. As a consequence, the MDM2 feedback loop is no longer activated and the dysfunctional TP53 protein is constitutively expressed [77].

2.4.4.3 *E-cadherin*

E-cadherin (also known as *CDH1*) is another example of a tumor suppressor gene. *E-cadherin* belongs to the family of calcium-dependent adhesion molecules [61]. *E-cadherin* is normally located in the adherence junctions at the surface of epithelial cells, where it maintains cell-cell contacts by interacting with *E-cadherin* proteins of adjacent epithelial cells. The intracellular domain of *E-cadherin* interacts with the actin cytoskeleton via interaction with the cytoplasmic proteins alpha-, beta- and gamma-catenin [85]. Loss of *E-cadherin* is considered a hallmark of epithelial-mesenchymal transitions (EMT) during organogenesis and possibly also during cancer metastasis [86,87]. Loss of *E-cadherin* protein expression was observed for most epithelial cancer types, but inactivating mutations in the gene have only been reported for half of lobular breast cancers and half of diffuse gastric cancers (two cancer subtypes that are morphologically very similar) [88]. Interestingly, most *E-cadherin* mutations in lobular breast cancers cause a premature termination of the encoded proteins, whereas most mutations in diffuse gastric cancers are in-frame deletions [89]. Also, germline *E-cadherin* mutations were frequently found in families with gastric cancer, but not in families with breast cancer [90,91]. Although *E-cadherin* mutations commonly are thought to involve a transition from invasive to metastatic cancer [89], compelling evidence has been suggested that *E-cadherin* mutations are already selected for in benign carcinoma *in situ* lesions in the breast [92].

2.4.4.4 *PTEN*

PTEN is a tumor suppressor gene that is mutated in a wide variety of cancer types, but with particularly high mutation frequencies in prostate cancer and brain tumors. The PTEN protein is a phosphatidylinositol-3,4,5-trisphosphate (PI3) in the kinase pathway. Apart from its phosphatase domain, the PTEN protein has extensive homology to tensin, a protein that interacts with actin filaments at focal adhesions through the negative regulation of the AKT/PKB signaling pathway [93]. PTEN was shown to suppress tumor cell growth by antagonizing the PI3 protein kinase [94]. Although rarely seen in low-grade glial tumors and early-stage prostate cancers, LOH of *PTEN* at 10q23 occurs in 70% of glioblastoma multiforme. Mutations of *PTEN* have been detected in 30% of glioblastoma cell lines, 15-20% of primary glioblastomas, and in 60% of prostate cancers [94]. Importantly, germ line *PTEN* mutations were identified in patients with Cowden disease, a cancer predisposition syndrome that includes an increased risk for brain tumors as well as breast cancer [95].

2.4.4.5 *BRCA1, BRCA2* and *CHEK2*

Three major breast cancer susceptibility genes have been identified*: BRCA1, BRCA2* and *CHEK2*, in addition to several minor susceptibility genes: *ATM, BRIP1* and *PALB2*. Interestingly, each of these susceptibility genes is known to function in the DNA damage response pathway. In response to DNA damage, progression through the cell cycle may be halted at several cell cycle checkpoints. The DNA damage response is activated by phosphorylation of ATM kinase, which then activates *CHEK2* kinase by phosphorylation. *CHEK2*, in turn, may activate TP53 resulting in a halt at the G1-S checkpoint of the cell cycle. Alternatively, *CHEK2* may activate CDC25A or CDC25C to prevent progression through the S- or G-phase of the cell cycle. Finally, its activation of BRCA1 allows repair of damaged DNA. *BRCA1* is part of a large multi-subunit nuclear protein complex known as BASC (BRCA1-associated genome surveillance complex) that among others includes *BRCA2, RAD51, RAD52, DSS1, FANCD2* and PALB2 [61,93]. BASC is thought to be of critical importance in the repair of DNA damage, particularly double-stranded breaks by homologous recombination [61,93]

Germline mutations of *BRCA1* (Breast Cancer susceptibility gene 1) are found in almost 20% of Dutch families with breast cancer and half of families with four or more cases of breast cancer (high-risk families) and/or at least one case of ovarian cancer [12, 96-98]. Yet, *BRCA1* is rarely mutated in sporadic breast cancer cases, so that *BRCA1* mutations are identified in less than 5% of all breast cancers. *BRCA1* is a high-risk breast cancer susceptibility gene, with an estimated 70% lifetime risk for mutation carriers to develop breast cancer. Hundreds of unique *BRCA1* mutations have been described that are located throughout the gene sequence. Roughly 80% of the mutations cause a premature truncation of the encoded proteins, with mutations including nucleotide substitutions, micro-insertions and deletions, but also larger insertions and deletions have been documented [93, 99]. Hypermethylation of the gene promoter has been found in sporadic breast cancers, especially in the presence of LOH at the *BRCA1* locus [22, 100]. It has been estimated that approximately 0.2% of the general population carries a mutation of *BRCA1*, but this rate may be much higher in certain ethnic or geographical populations such as the Ashkenazi Jewish and Icelandic populations [12, 101-103].

Germline mutations of *BRCA2* (Breast Cancer susceptibility gene 2) are found in 6% of Dutch breast cancer families and up to one fifth of high risk breast cancer families or families including at least one case of ovarian cancer [12, 104]. Although *BRCA2* mutations have been associated with male breast cancer, this association is less clear for Dutch breast cancer families [105]. Similar to *BRCA1*, *BRCA2* mutations are rarely identified in sporadic breast cancers [94]. BRCA2 also is a high-risk breast cancer susceptibility gene, with an estimated 50% lifetime risk for mutation carriers to develop breast cancer. Hundreds of unique *BRCA2* mutations have been described, in particular nucleotide substitutions, micro-insertions and deletions [12, 15]. Similar to that of BRCA1, the population frequency of *BRCA2* germline mutations is similar to that of *BRCA1*, an estimated 0.2% [103].

The truncating 1100delC germline mutation of *CHEK2* (Cell cycle checkpoint kinase 2) has been identified in 5% of Dutch breast cancer families and as much as 18% of families with hereditary breast and colorectal cancer [61, 81, 106, 107]. In contrast to *BRCA1* and *BRCA2*, *CHEK2* 1100delC is a low risk breast cancer susceptibility allele, with an estimated 20% lifetime risk developing breast cancer [106-109]. Only a limited number of *CHEK2* germline mutations have been associated with a cancer risk, including the *I157T* variant and the IVS2+1A>G truncating variant. Importantly, the I157T variant has also been associated with a colorectal cancer risk in Finland and Poland however; neither of these variants have been identified in the Dutch population [107,110]. *CHEK2* 1100delC has a frequency of 1% in healthy Dutch individuals, but is present at much lower frequencies in other geographical populations

2.5 Gene expression in cancer

2.5.1 Gene expression, gene mutations and cell biology

Human cancers are diverse in their tissue origin as well as their individual biological and genetic histories. These diversities are reflected by variations in gene expression programs among human cancers. Profiling cancer-specific gene expression programs thus may provide a new basis for the classification of human cancers. With the advent of microarray technology, it became possible to analyze and understand cancer-specific gene expression profiles on a global level instead of a gene-by-gene level. Microarray technology is at the heart of this thesis, with particular focus on gene expression profiling of breast cancers and brain tumors.

There are two main reasons for using microarray technology in oncological research. Firstly, to understand the biology related to particular cancer types or subtypes, their gene mutations and their aberrant (downstream) biological pathways. This is largely exploratory and results from a microarray experiment can be analyzed by using pathways and gene annotations such as Gene Ontology (Chapter 1). Secondly, to classify human cancers according a particular variable: organ type or subtype, patient's prognosis, prediction of treatment response, or site of metastasis. This can be done in two ways: a) by looking purely at the biology associated with a variable or b) classifying tumors, where the biology of the genes involved is not so important as to have reliable genes that can predict the tested variable [111,112]. These analyses correlate clinical or biological data of cancers with their molecular profiles, in order to identify reliable classifiers. The biological analysis presented in this thesis are purely biologiocal related (chapters 3, 4, 5 and 6).

2.5.2 Breast cancer gene expression profiles

Perou and Botstein were the first to use microarray technology to study the biology of human cancers by their intrinsic gene expression program [113]. They were able to distinguish several breast cancer subtypes based on gene expression profiles that correlated with previously identified histological protein expression patterns [2,5-7] (Figure 2 and paragraph 1.5.1). 'Intrinsic' gene signatures were defined that included genes whose differential expression levels could be related to specific histological features of the breast tumors. In a series of follow-up papers, Sorlie and colleagues further refined their intrinsic gene signatures to associate five molecular subtypes of breast cancer with survival data of the patients [111,113-115]. The five subtypes defined by these researchers reflect the inherent cell biology that defines the cluster division of the breast cancer subtypes:

1) 'Luminal A' breast cancers expressing ER: this subtype associated with a favorable prognosis.

2) 'Luminal B' breast cancers expressing ER: this subtype has a less favorable prognosis, in particular for relapse of the disease.

3) 'ERBB2' breast cancers overexpressing ERBB2 and mostly ER negative: this subtype is known for a poor prognosis.

4) 'Basal-like' breast cancers expressing basal cytokeratins 5 and 17, integrin 4 and laminin, but lacking ER, PR and ERBB2 expression: this subtype presented with a more aggressive clinical behavior.

5) 'Normal-like' breast cancers, expressing many genes known to be expressed by adipose tissue and other non-epithelial cell types. These tumors also had strong expression of basal epithelial genes and low expression of luminal epithelial genes.

Molecular profiles have also been associated with other known cancer genes such as TP53 [116,117], BRCA1 [118,119], and EGFR [120]. In such studies, the underlying mutation is presumed to be driving the segregation of the samples.

Other prominent milestones in the application of gene expression microarrays to breast cancer involve the classification of breast cancers according clinical outcome of the patients. Van 't Veer *et al.* [121,122] were the first to define a 70-gene expression signature that predicted the occurrence of metastasis in lymph node-negative breast cancer patients who had been diagnosed before 55 years of age. Similarly, a 21-gene signature was shown to predict metastasis in lymph node-negative patients with ER-positive breast cancer who had received adjuvant hormonal therapy [123]. A 76-gene signature also predicted metastasis in lymph node-negative breast cancer patients who had not received any adjuvant systemic therapy, irrespective of age and ER status [124,125]. Finally, a 44-gene signature has also predicted responsiveness of breast cancers to *Tamoxifen* therapy more accurate than the ER status of the tumors [21]. The ability of microarray technology to identify breast cancer patients who have a more or less favorable prognosis in developing metastasis could guide clinicians in avoiding adjuvant systemic therapy or, alternatively, to choose more aggressive therapeutic options. In this respect, it could also be useful to predict the site of metastasis, as recently was shown for breast cancers that metastasized to the bone [125].

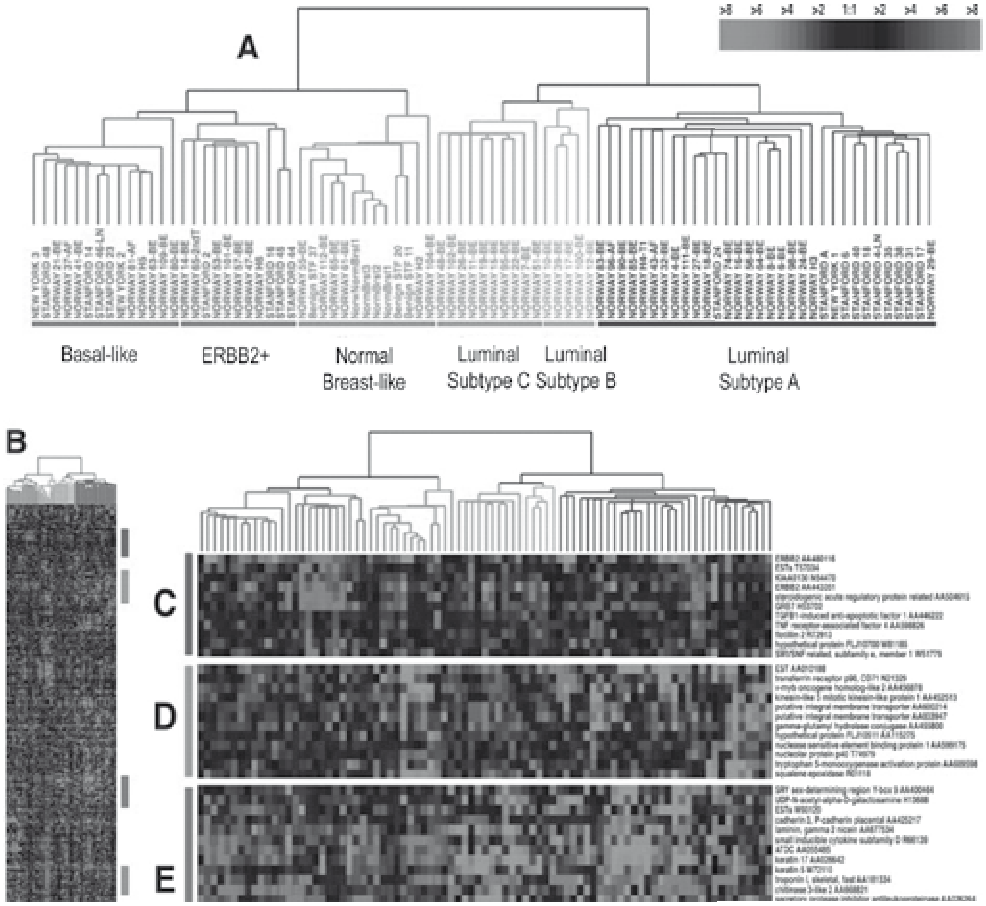


Figure 2: Intrinsic breast cancer subtypes identified by microarray analysis. Gene expression patterns of 85 experimental samples representing 78 carcinomas, three benign tumors, and four normal tissues, analyzed by hierarchical clustering using the 476 cDNA intrinsic clone set. (A) The tumor specimens were divided into five (or six) subtypes based on differences in gene expression. The cluster dendrogram showing the five (six) subtypes of tumors are colored as: luminal subtype A, dark blue; luminal subtype B, yellow; luminal subtype C, light blue; normal breast-like, green; basal-like, red; and ERBB2+, pink. (B) The full cluster diagram scaled down (the complete 456-clone cluster diagram is available as Figure 4). The colored bars on the right represent the inserts presented in C-G. (C) *ERBB2* amplicon cluster. (D) Novel unknown cluster. (E) Basal epithelial cell-enriched cluster. (F) Normal breast-like cluster. (G) Luminal epithelial gene cluster containing ER.

Source: Sorlie, T., et al., Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A, 2001. 98(19): p. 10869-74.

2.5.3 Brain tumor gene expression profiles

Gene expression profiling of brain tumors has been guided primarily by their histological and pathological classification. Brain tumor gene expression profiles have been generated to investigate both the biology and the classification of brain tumors. Looking at biology, Pomeroy *et al.* [126] defined a gene signature that distinguished medulloblastomas from other histologically similar brain tumors and using this classification could predict their therapy response. Importantly, this gene signature revealed that medulloblastomas are biologically distinct from primitive neuroectodermal tumors (PNET's), two subtypes of brain tumors that are often considered a single entity. The medulloblastoma gene expression profile implicated cerebellar granule cells as their cell of origin and revealed an unexpected involvement of the Sonic Hedgehog signaling pathway. Bredel *et al.* have also used gene expression profiling in the biological understanding of human gliomas by applying molecular network knowledge to the analysis of key functions and pathways associated with gliomagenesis [127]. Using a set of 50 human gliomas comprised of various histologies, they have seen via the transcriptional profiles of these tumors that integrin signaling pathway is most significant in the glioblastoma subtype, which is paradigmatic for its strong migratory and invasive behavior. The MYC oncogene was also seen as a major network player in the biological process of gliomagenesis. More specifically, three novel MYC-interacting genes (*UBE2C, EMP1*, and *FBXW7*) with cancer-related functions were identified as network constituents differentially expressed in gliomas, as was *CD151* as a new component of a network that mediates glioblastoma cell invasion [127]. Such biological approaches as Pomeroy *et al.* and Bredel *et al.* have extended existing knowledge about the organizational pattern of gene expression in human gliomas, which can identify potential novel targets for future therapeutic development.

Understanding the biology is of utmost importance in brain tumors, however the classification based on its correlation with clinical parameters is also revealing important information. Classification based on histological subtype and genetic mutations as well as clinical parameters such as response to therapeutic drugs can potentially predict a patient's prognosis. French *et al.* have defined a 16-gene signature that predicted treatment response of oligodendrogliomas and a 103-gene signature for survival of the patients [128]. Interestingly, they were also able to define gene signatures that distinguished oligodendrogliomas with loss of 1p, loss of 19q, or loss of both chromosomal arms. Nutt *et al.* [27] defined a 20-gene signature that appeared to better predict clinical outcome of patients with glioblastomas or high-grade oligodendrogliomas than classical histology. This gene signature also allowed them to classify high-grade gliomas with non-classical histology. Together, these gene expression-profiling studies have shown that microarray technology may be an important tool in the molecular classification of gliomas. This technology can improve the classification of tumor sub groups as well as the correlation of patient's characteristics to make diagnoses and treatment decisions that are more informed.

Perhaps most notable are the findings by French *et al.* that gene expression profiles not only reflect the biology and clinical behavior of gliomas but also their underlying molecular basis. Each subtype of glioma is reflected in its pathological and histological characteristics; however, molecular profiles can further distinguish subtypes based on the underlying transcriptome. These molecular profiles are particularly important for brain tumor patients, as they are in urgent need for new treatment targets.

2.6 Cancer Bibliography

1. http://www.hhmi.org/research/investigators/vogelstein.html.
2. Rakha, E.A., et al., *Morphological and immunophenotypic analysis of breast carcinomas with basal and myoepithelial differentiation.* J Pathol, 2006. **208**(4): p. 495-506.
3. Levin, E.R. and R.J. Pietras, *Estrogen receptors outside the nucleus in breast cancer.* Breast Cancer Res Treat, 2007.
4. Deroo, B.J. and K.S. Korach, *Estrogen receptors and human disease.* J Clin Invest, 2006. **116**(3): p. 561-70.
5. Rakha, E.A., et al., *Breast carcinoma with basal differentiation: a proposal for pathology definition based on basal cytokeratin expression.* Histopathology, 2007. **50**(4): p. 434-8.
6. Rakha, E.A., et al., *Prognostic markers in triple-negative breast cancer.* Cancer, 2007. **109**(1): p. 25-32.
7. Rakha, E.A., et al., *Basal phenotype identifies a poor prognostic subgroup of breast cancer of clinical importance.* Eur J Cancer, 2006. **42**(18): p. 3149-56.
8. Abd El-Rehim, D.M., et al., *Expression of luminal and basal cytokeratins in human breast carcinoma.* J Pathol, 2004. **203**(2): p. 661-71.
9. Nielsen, T.O., et al., *Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma.* Clin Cancer Res, 2004. **10**(16): p. 5367-74.
10. Livasy, C.A., et al., *Phenotypic evaluation of the basal-like subtype of invasive breast carcinoma.* Mod Pathol, 2006. **19**(2): p. 264-71.
11. http://www.ikcnet.nl/page.php?id=1872&nav_id=97.
12. Meijers-Heijboer, H., *Breast cancer susceptibility genes thesis.*
13. Henderson, B.E. and H.S. Feigelson, *Hormonal carcinogenesis.* Carcinogenesis, 2000. **21**(3): p. 427-33.
14. Saji, S., M. Hirose, and M. Toi, *Clinical significance of estrogen receptor beta in breast cancer.* Cancer Chemother Pharmacol, 2005. **56 Suppl 1**: p. 21-6.
15. Hayashi, S.I., et al., *The expression and function of estrogen receptor alpha and beta in human breast cancer and its clinical application.* Endocr Relat Cancer, 2003. **10**(2): p. 193-202.
16. Mirza, S., et al., *Promoter hypermethylation of TMS1, BRCA1, ERalpha and PRB in serum and tumor DNA of invasive ductal breast carcinoma patients.* Life Sci, 2007.
17. http://www.healthcentral.com/breast-cancer/.
18. Rochefort, H., et al., *How to target estrogen receptor-negative breast cancer?* Endocr Relat Cancer, 2003. **10**(2): p. 261-6.
19. Robert Gentleman, V.C., Wolfgang Huber, Rafael Irizarry, Sandrine Dudoit, *Solutions using R and Bioconductor*. 2005, New York: Springer.
20. Finnegan, T.J. and L.A. Carey, *Gene-expression analysis and the basal-like breast cancer subtype.* Future Oncol, 2007. **3**(1): p. 55-63.
21. Jansen, M.P., et al., *Molecular classification of tamoxifen-resistant breast carcinomas by gene expression profiling.* J Clin Oncol, 2005. **23**(4): p. 732-40.
22. Osborne, C., P. Wilson, and D. Tripathy, *Oncogenes and tumor suppressor genes in breast cancer: potential diagnostic and therapeutic applications.* Oncologist, 2004. **9**(4): p. 361-77.
23. Tsuda, H., *HER-2 (c-erbB-2) test update: present status and problems.* Breast Cancer, 2006. **13**(3): p. 236-48.
24. http://www.cbtrus.org/reports//2005-2006/2006report.pdf.
25. Peter C. Burger, M.D.a.B.W.S., *Tumors of Central Nervous System*. Atlas of Tumor Pathology. 1994, Washington DC: Armed Forces Institute of Pathology.
26. Cavenee, P.K.W.K., *Tumors of the Nervous System.* Pathology and Genetics. 2000, Lyon: IARC Press. 314.
27. Nutt, C.L., et al., *Gene expression-based classification of malignant gliomas correlates better with survival than histological classification.* Cancer Res, 2003. **63**(7): p. 1602-7.
28. http://health.enotes.com/cancer-encyclopedia/brain-central-nervous-system-tumors.
29. De Bortoli, M., et al., *Medulloblastoma outcome is adversely associated with overexpression of EEF1D, RPL30, and RPS20 on the long arm of chromosome 8.* BMC Cancer, 2006. **6**: p. 223.
30. Collins, V.P., *Mechanisms of disease: genetic predictors of response to treatment in brain tumors.* Nat Clin Pract Oncol, 2007. **4**(6): p. 362-74.
31. http://www.cancer.org.

32. Mellinghoff, I.K., et al., *Molecular determinants of the response of glioblastomas to EGFR kinase inhibitors.* N Engl J Med, 2005. **353**(19): p. 2012-24.

33. Lynch, T.J., et al., *Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib.* N Engl J Med, 2004. **350**(21): p. 2129-39.

34. Marie, Y., et al., *EGFR tyrosine kinase domain mutations in human gliomas.* Neurology, 2005. **64**(8): p. 1444-5.

35. Kinzler, K.W. and B. Vogelstein, *Lessons from hereditary colorectal cancer.* Cell, 1996. **87**(2): p. 159-70.

36. Ashley, D.J., *The two "hit" and multiple "hit" theories of carcinogenesis.* Br J Cancer, 1969. **23**(2): p. 313-28.

37. Knudson, A.G., Jr., *Mutation and cancer: statistical study of retinoblastoma.* Proc Natl Acad Sci U S A, 1971. **68**(4): p. 820-3.

38. Kaiser, J., *Cancer. First pass at cancer genome reveals complex landscape.* Science, 2006. **313**(5792): p. 1370.

39. http://www.hhmi.org/news/vogelstein20060908.html.

40. Sjoblom, T., et al., *The consensus coding sequences of human breast and colorectal cancers.* Science, 2006. **314**(5797): p. 268-74.

41. Simpson, A.J., *The natural somatic mutation frequency and human carcinogenesis.* Adv Cancer Res, 1997. **71**: p. 209-40.

42. *Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease.* Lancet, 2001. **358**(9291): p. 1389-99.

43. http://www.qub.ac.uk/cm/pat/undergraduate/Basiccancer/.

44. Yokota, J., *Tumor progression and metastasis.* Carcinogenesis, 2000. **21**(3): p. 497-503.

45. Vogelstein, B. and K.W. Kinzler, *Cancer genes and the pathways they control.* Nat Med, 2004. **10**(8): p. 789-99.

46. Knudson, A.G., Jr., *Mutation and cancer in man.* Cancer, 1977. **39**(4 Suppl): p. 1882-6.

47. Knudson, A.G., *Cancer genetics.* Am J Med Genet, 2002. **111**(1): p. 96-102.

48. Knudson, A.G., Jr., H.W. Hethcote, and B.W. Brown, *Mutation and childhood cancer: a probabilistic model for the incidence of retinoblastoma.* Proc Natl Acad Sci U S A, 1975. **72**(12): p. 5116-20.

49. Santarosa, M. and A. Ashworth, *Haploinsufficiency for tumour suppressor genes: when you don't need to go all the way.* Biochim Biophys Acta, 2004. **1654**(2): p. 105-22.

50. Alberts, B., *DNA replication and recombination.* Nature, 2003. **421**(6921): p. 431-5.

51. Yang, G., et al., *Dominant effects of an Msh6 missense mutation on DNA repair and cancer susceptibility.* Cancer Cell, 2004. **6**(2): p. 139-50.

52. Kucherlapati, M., et al., *Haploinsufficiency of Flap endonuclease (Fen1) leads to rapid tumor progression.* Proc Natl Acad Sci U S A, 2002. **99**(15): p. 9924-9.

53. Kolodner, R.D. and G.T. Marsischky, *Eukaryotic DNA mismatch repair.* Curr Opin Genet Dev, 1999. **9**(1): p. 89-96.

54. Gronbaek, K., C. Hother, and P.A. Jones, *Epigenetic changes in cancer.* Apmis, 2007. **115**(10): p. 1039-1059.

55. Herman, J.G., *Epigenetic changes in cancer and preneoplasia.* Cold Spring Harb Symp Quant Biol, 2005. **70**: p. 329-33.

56. Robertson, K.D., *DNA methylation and chromatin - unraveling the tangled web.* Oncogene, 2002. **21**(35): p. 5361-79.

57. Negrini, M., et al., *MicroRNAs in human cancer: from research to therapy.* J Cell Sci, 2007. **120**(Pt 11): p. 1833-40.

58. Looijenga, L.H., et al., *Relevance of microRNAs in normal and malignant development, including human testicular germ cell tumours.* Int J Androl, 2007. **30**(4): p. 304-14; discussion 314-5.

59. Ehrbrecht, A., et al., *Comprehensive genomic analysis of desmoplastic medulloblastomas: identification of novel amplified genes and separate evaluation of the different histological components.* J Pathol, 2006. **208**(4): p. 554-63.

60. http://www.sanger.ac.uk/genetics/CGP/Census.

61. http://www.genecards.com.

62. Oda, K., et al., *A comprehensive pathway map of epidermal growth factor receptor signaling.* Mol Syst Biol, 2005. **1**: p. 2005 0010.

63. Bell, D.W., et al., *Inherited susceptibility to lung cancer may be associated with the T790M drug resistance mutation in EGFR.* Nat Genet, 2005. **37**(12): p. 1315-6.

64. Heimberger, A.B., et al., *The natural history of EGFR and EGFRvIII in glioblastoma patients.* J Transl Med, 2005. **3**: p. 38.

65. Okamoto, I., et al., *Expression of constitutively activated EGFRvIII in non-small cell lung cancer.* Cancer Sci, 2003. **94**(1): p. 50-6.
66. Landgraf, R., *HER2 therapy. HER2 (ERBB2): functional diversity from structurally conserved building blocks.* Breast Cancer Res, 2007. **9**(1): p. 202.
67. Lee, J.W., et al., *ERBB2 kinase domain mutation in the lung squamous cell carcinoma.* Cancer Lett, 2006. **237**(1): p. 89-94.
68. Sonobe, M., et al., *Lung adenocarcinoma harboring mutations in the ERBB2 kinase domain.* J Mol Diagn, 2006. **8**(3): p. 351-6.
69. Stephens, P., et al., *Lung cancer: intragenic ERBB2 kinase mutations in tumours.* Nature, 2004. **431**(7008): p. 525-6.
70. Badache, A. and A. Goncalves, *The ErbB2 signaling network as a target for breast cancer therapy.* J Mammary Gland Biol Neoplasia, 2006. **11**(1): p. 13-25.
71. Borg, A., et al., *ERBB2 amplification in breast cancer with a high rate of proliferation.* Oncogene, 1991. **6**(1): p. 137-43.
72. Looi, L.M., P.L. Cheah, and S.F. Yap, *Correlation between histological grade and c-erbB2 oncoprotein overexpression in infiltrating ductal carcinoma of breast.* Malays J Pathol, 1997. **19**(1): p. 35-9.
73. Yarden, Y., *Biology of HER2 and its importance in breast cancer.* Oncology, 2001. **61 Suppl 2**: p. 1-13.
74. Menard, S., et al., *HER2 as a prognostic factor in breast cancer.* Oncology, 2001. **61 Suppl 2**: p. 67-72.
75. Winston, J.S., J. Ramanaryanan, and E. Levine, *HER-2/neu evaluation in breast cancer are we there yet?* Am J Clin Pathol, 2004. **121 Suppl**: p. S33-49.
76. Smith, N.D., et al., *The p53 tumor suppressor gene and nuclear protein: basic science review and relevance in the management of bladder cancer.* J Urol, 2003. **169**(4): p. 1219-28.
77. Sengupta, S. and C.C. Harris, *p53: traffic cop at the crossroads of DNA repair and recombination.* Nat Rev Mol Cell Biol, 2005. **6**(1): p. 44-55.
78. Olivier, M., et al., *The IARC TP53 database: new online mutation analysis and recommendations to users.* Hum Mutat, 2002. **19**(6): p. 607-14.
79. Petitjean, A., et al., *TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes.* Oncogene, 2007. **26**(15): p. 2157-65.
80. Petitjean, A., et al., *Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database.* Hum Mutat, 2007. **28**(6): p. 622-9.
81. Berns, E.M., et al., *Complete sequencing of TP53 predicts poor response to systemic therapy of advanced breast cancer.* Cancer Res, 2000. **60**(8): p. 2155-62.
82. Kinzler, K.W. and B. Vogelstein, *Cancer-susceptibility genes. Gatekeepers and caretakers.* Nature, 1997. **386**(6627): p. 761, 763.
83. Idoate, M.A. and J. Echeveste, *[Update on the molecular biology of gliomas: towards a pathomolecular classification of gliomas].* Rev Neurol, 2007. **44**(4): p. 217-24.
84. Olivier, M., et al., *Li-Fraumeni and related syndromes: correlation between tumor type, family structure, and TP53 genotype.* Cancer Res, 2003. **63**(20): p. 6643-50.
85. Shimoyama, Y., et al., *Cadherin cell-adhesion molecules in human epithelial tissues and carcinomas.* Cancer Res, 1989. **49**(8): p. 2128-33.
86. Ivanov, D.B., M.P. Philippova, and V.A. Tkachuk, *Structure and functions of classical cadherins.* Biochemistry (Mosc), 2001. **66**(10): p. 1174-86.
87. Noritake, J., et al., *Positive role of IQGAP1, an effector of Rac1, in actin-meshwork formation at sites of cell-cell contact.* Mol Biol Cell, 2004. **15**(3): p. 1065-76.
88. Christofori, G. and H. Semb, *The role of the cell-adhesion molecule E-cadherin as a tumour-suppressor gene.* Trends Biochem Sci, 1999. **24**(2): p. 73-6.
89. Pecina-Slaus, N., *Tumor suppressor gene E-cadherin and its role in normal and malignant cells.* Cancer Cell Int, 2003. **3**(1): p. 17.
90. Guilford, P., et al., *E-cadherin germline mutations in familial gastric cancer.* Nature, 1998. **392**(6674): p. 402-5.
91. Guilford, P.J., et al., *E-cadherin germline mutations define an inherited cancer syndrome dominated by diffuse gastric cancer.* Hum Mutat, 1999. **14**(3): p. 249-55.

92. De Leeuw, W.J., et al., *Simultaneous loss of E-cadherin and catenins in invasive lobular breast cancer and lobular carcinoma in situ.* J Pathol, 1997. **183**(4): p. 404-11.

93. http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene.

94. Thorlacius, S., et al., *Study of a single BRCA2 mutation with high carrier frequency in a small population.* Am J Hum Genet, 1997. **60**(5): p. 1079-84.

95. Hollestelle, A., et al., *Phosphatidylinositol-3-OH kinase or RAS pathway mutations in human breast cancer cell lines.* Mol Cancer Res, 2007. **5**(2): p. 195-201.

96. Ford, D., D.F. Easton, and J. Peto, *Estimates of the gene frequency of BRCA1 and its contribution to breast and ovarian cancer incidence.* Am J Hum Genet, 1995. **57**(6): p. 1457-62.

97. Ford, D., et al., *Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium.* Am J Hum Genet, 1998. **62**(3): p. 676-89.

98. Hall, J.M., et al., *Linkage of early-onset familial breast cancer to chromosome 17q21.* Science, 1990. **250**(4988): p. 1684-9.

99. Deng, C.X., *BRCA1: cell cycle checkpoint, genetic instability, DNA damage response and cancer evolution.* Nucleic Acids Res, 2006. **34**(5): p. 1416-26.

100. Esteller, M., et al., *Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors.* J Natl Cancer Inst, 2000. **92**(7): p. 564-9.

101. Makriyianni, I., et al., *BRCA1:185delAG found in the San Luis Valley probably originated in a Jewish founder.* J Med Genet, 2005. **42**(5): p. e27.

102. Struewing, J.P., et al., *The carrier frequency of the BRCA1 185delAG mutation is approximately 1 percent in Ashkenazi Jewish individuals.* Nat Genet, 1995. **11**(2): p. 198-200.

103. Peto, J., et al., *Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer.* J Natl Cancer Inst, 1999. **91**(11): p. 943-9.

104. Wooster, R., et al., *Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13.* Science, 1994. **265**(5181): p. 2088-90.

105. Kwiatkowska, E., et al., *Somatic mutations in the BRCA2 gene and high frequency of allelic loss of BRCA2 in sporadic male breast cancer.* Int J Cancer, 2002. **98**(6): p. 943-5.

106. Schutte, M., et al., *Variants in CHEK2 other than 1100delC do not make a major contribution to breast cancer susceptibility.* Am J Hum Genet, 2003. **72**(4): p. 1023-8.

107. Meijers-Heijboer, H., et al., *Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations.* Nat Genet, 2002. **31**(1): p. 55-9.

108. *CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies.* Am J Hum Genet, 2004. **74**(6): p. 1175-82.

109. Shaag, A., et al., *Functional and genomic approaches reveal an ancient CHEK2 allele associated with breast cancer in the Ashkenazi Jewish population.* Hum Mol Genet, 2005. **14**(4): p. 555-63.

110. Meijers-Heijboer, H., et al., *The CHEK2 1100delC mutation identifies families with a hereditary breast and colorectal cancer phenotype.* Am J Hum Genet, 2003. **72**(5): p. 1308-14.

111. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-74.

112. Chang, J.C., et al., *Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer.* Lancet, 2003. **362**(9381): p. 362-9.

113. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.

114. Abramovitz, M. and B. Leyland-Jones, *A systems approach to clinical oncology: focus on breast cancer.* Proteome Sci, 2006. **4**: p. 5.

115. Sorlie, T., et al., *Repeated observation of breast tumor subtypes in independent gene expression data sets.* Proc Natl Acad Sci U S A, 2003. **100**(14): p. 8418-23.

116. Voorhoeve, P.M., et al., *A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors.* Adv Exp Med Biol, 2007. **604**: p. 17-46.

117. Yu, K., et al., *A modular analysis of breast cancer reveals a novel low-grade molecular signature in estrogen receptor-positive tumors.* Clin Cancer Res, 2006. **12**(11 Pt 1): p. 3288-96.

118. Hedenfalk, I., et al., *Gene-expression profiles in hereditary breast cancer.* N Engl J Med, 2001. **344**(8): p. 539-48.

119. Hedenfalk, I., et al., *Molecular classification of familial non-BRCA1/BRCA2 breast cancer.* Proc Natl Acad Sci U S A, 2003. **100**(5): p. 2532-7.

120. Angulo, B., et al., *Expression signatures in lung cancer reveal a profile for EGFR-mutant tumours and identify selective PIK3CA overexpression by gene amplification.* J Pathol, 2007.

121. van 't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**(6871): p. 530-6.

122. van 't Veer, L.J., et al., *Expression profiling predicts outcome in breast cancer.* Breast Cancer Res, 2003. **5**(1): p. 57-8.

123. van't Veer, L.J., S. Paik, and D.F. Hayes, *Gene expression profiling of breast cancer: a new tumor marker.* J Clin Oncol, 2005. **23**(8): p. 1631-5.

124. Wang, Y., et al., *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.* Lancet, 2005. **365**(9460): p. 671-9.

125. Smid, M., et al., *Genes associated with breast cancer metastatic to bone.* J Clin Oncol, 2006. **24**(15): p. 2261-7.

126. Pomeroy, S.L., et al., *Prediction of central nervous system embryonal tumour outcome based on gene expression.* Nature, 2002. **415**(6870): p. 436-42.

127. Bredel, M., et al., *Functional network analysis reveals extended gliomagenesis pathway maps and three novel MYC-interacting genes in human gliomas.* Cancer Res, 2005. **65**(19): p. 8679-89.

128. French, P.J., et al., *Gene expression profiles associated with treatment response in oligodendrogliomas.* Cancer Res, 2005. **65**(24): p. 11335-44.

Chapter 3

Growing Applications and Advancements in Microarray Technology and Analysis Tools

Justine K. Peeters and Peter J. van der Spek

Department of Bioinformatics, Erasmus MC, University Medical Centre, Rotterdam

Cell Biochemistry and Biophysics, 2005 43:149-166

Abstract

In today's field of genomics, traditional gene-by-gene approach is not adequate to meet the demand of processing information generated from mapping the complex biology of the human genome. More global views of analyzing the magnitude of information are necessary, such as with microarrays. Microarray technology today is rapidly uncovering broad patterns of genetic activity and showing insight into gene functions, processes, and pathways. With the growing technology, imminent knowledge is being generated looking into transcriptional processes and biological mechanisms from many different organisms and phylogeny. Many tools are being developed to assist with the analysis of such high-throughput data, many applications are being utilized by this technology, and the field is growing and expanding rapidly to accommodate the expanding genomics era.

Chapter
3

Introduction

Ten years ago, microarray technology was known as macroarrays with experiments performed on large membrane sheets spotted with cDNA (1–10,000 genes) for comparative hybridization of RNA species. This technology although an advancement in comparison to classic methods such as Northern and Southern blotting, has moved through to the chip technology of today. This has allowed the exceptional ability to study expression of the entire genome in one experiment, with a quantifiable signal being generated that is directly proportional to the expression level in cells/tissues. It does not, however, take into account the level of translated protein, which in fact does the physical work of the cell. There are a number of different variations on the microarray. As chip technology advances, so does the number of products available to interrogate the human genome, not only at the expression levels as classic microarrays have seen, but also at the physical, transcriptional, and translational levels. As well as the advancement of the physical technology of the microarray chip, new analysis methods, software, and knowledge of their applications are also increasing at a vast rate.

Platforms of Microarray Technology

RNA Expression Arrays

Considering expression analysis, various expression arrays use differing lengths of DNA fragments; cDNA microarrays use approx 200to 500-bp fragments, usually produced by polymerase chain reaction (PCR), and oligonucleotide microarrays use 30- to 70-bp-length oligonucleotides. Both cDNA and oligonucleotide DNA fragments are chemically attached to a glass support and can be represented by up to 10,000+ genes. These arrays are traditionally two color, in which the sample is labeled with red as an example, and the control is labeled with green nucleotides.

These RNA probes can be directly labeled with cy3/cy5 dyes in the reverse transcription labeling protocol (see Figure 1A) or, alternatively, indirectly labeled with fluorescent antibodies conjugated to amino allyl nucleotides, which are directly incorporated into reverse transcription of the RNA. The advantage of indirect labeling methods such as amino allyl incorporation is that no bias is introduced with direct incorporation of the fluorescent dyes, owing to differences in their chemical structures. The amino allyl labeling technique can also be utilized with a reduced amount of RNA, producing greater signal intensity than the directly incorporated fluorophores. These two-color arrays produce a ratio indicating the differential expression between the samples. There are currently a large number of companies producing both cDNA and oligonucleotide arrays that interrogate human genome expression to maximum capacity. GeneChip® technology, a slight variation of oligonucleotide arrays, is produced by Affymetrix (www.affymetrix.com) and is currently in the forefront of microarray technology, along with Agilent Technologies (www.agilent.com). Affymetrix oligonucleotides are 25–30 bp long and are synthesized *in situ* on silicon wafers using a lithographic process. Affymetrix arrays use a standardized biotin labeling protocol (see Figure 1B) and produce an intensity signal, which allows absolute quantification, unlike cDNA/ oligonucleotide arrays. These chips are represented by 10–100,000 array spots representing genes, with potentially up to 4 million in the future with such technology. The RNA expression arrays are the most widely used microarray platform in biological and genomic research today. The applications of these arrays are further covered under the sections Basic Research and Pharmacogenomics.

Single Nucleotide Polymorphism Chips

The recent introduction of Affymetrix 100K SNP chips saw the beginning of a whole new advancement in microarrays. This GeneChip interrogates more than 100,000 single nucleotide polymorphisms (SNPs) known within the population, generating genotypes with a single, easy-to-use mapping assay. This chip comprises two arrays, both with more than 50,000 SNPs utilizing the probe fragments generated with two restriction enzymes, *Hind*III and *Xba*I, in the assay. This array can be used in a number of different applications such as the identification of loss of heterozygosity (LOH), mutation detection, polymorphism analysis, mapping studies, and evolutionary and pharmacogenomic applications. These SNP chips allow a probability-based assessment of SNP copy number, which is a valuable advantage over the traditional genotyping linkage studies.

SNP Chips in Mapping Disease Genes

Shrimpton *et al.* [1] used the GeneChip SNP technology to study the segregation of congenital vertical talus (CVT) (also known as rocker-bottom foot deformity, CVT is a dislocation of the talonavicular joint, with rigid dorsal dislocation of the navicular over the neck of the talus) and Charcot-Marie-Tooth disease (CMT), a distinct heterogeneous group of disorders, with foot abnormalities that typically develop a high-arched claw foot appearance later in life. In their study of a large American family by SNP analysis, in whom both CMT and CVT phenotypes were segregating, they identified a 7-MB critical region on chromosome 2q31 that led to the detection of a single nonsense mutation in the *HOX10* gene. Using this microarray technology, Shrimpton *et al.* [1] could conclude that this mutation in the *HOX10* gene accounted for both CVT and CMT in heterozygotes. Middleton *et al.* 2004 [2] performed linkage studies using the SNP platform on bipolar disorder. As a psychological disorder, it is found to be very heterogeneous and shows strong evidence of heritability [3]. The comparative analysis of 25 extended Portuguese families with the disorder by Middleton *et al.* [2] indicated the presence of significant linkage peaks in chromosomal regions. These regions have previously been uncharacterized by poor coverage of low information content on traditional microsatellite assays. In particular, a region on chromosome 6q22 has been identified with strong genomewide linkage significance scores, indicating that this region may contain genes or elements contributing to the cause of bipolar disorder.

SNP Chips in LOH Studies

Using the SNP platform, LOH has been widely studied. Lieberfarb *et al.* [4] applied the technology and subsequent analysis to automate the definition of statistically valid regions of LOH. In many human cancers, it is predicted that there is an accumulation of genetic events that lead to metastasis. Lieberfarb *et al.* [4] examined whether the heterogeneous nature of prostate cancer can also be based on this nature of accumulated genetic events leading to differences in the disease or whether parallel sets of genetic alterations lead to distinct subtypes of cancer. They assigned LOH (lost or retained) genotypes to prostate cancer samples and organized these samples by hierarchical clustering based on the pattern of this LOH. By comparing all combinations except instances in which retained-retained appeared, this allowed the clustering to be driven primarily by the similarity or difference in a deletion rather than the similarity in retention [4]. Their results indicated that the prostate samples clustered into distinct branches that contained tumors enriched for specific regions of nonoverlap-ping regions of heterozygosity. This suggested the presence of distinct genetic subtypes of prostate cancers that can be defined by LOH status. Studies such as Lieberfarb *et al.*, [4] indicate the strength of the application of SNP microarray technology over older traditional methods, such as microsatellite typing, in which significant results often go undetected, as also demonstrated by Middleton *et al.* [2] in their study of bipolar disorder.

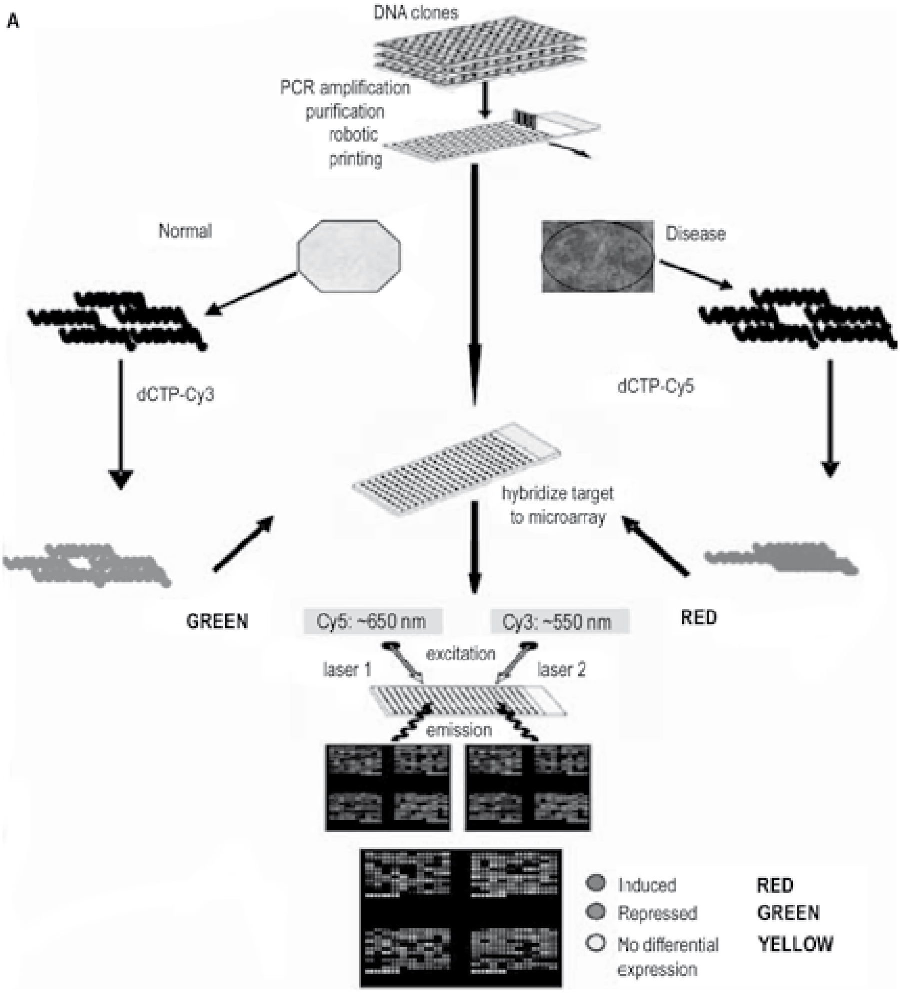
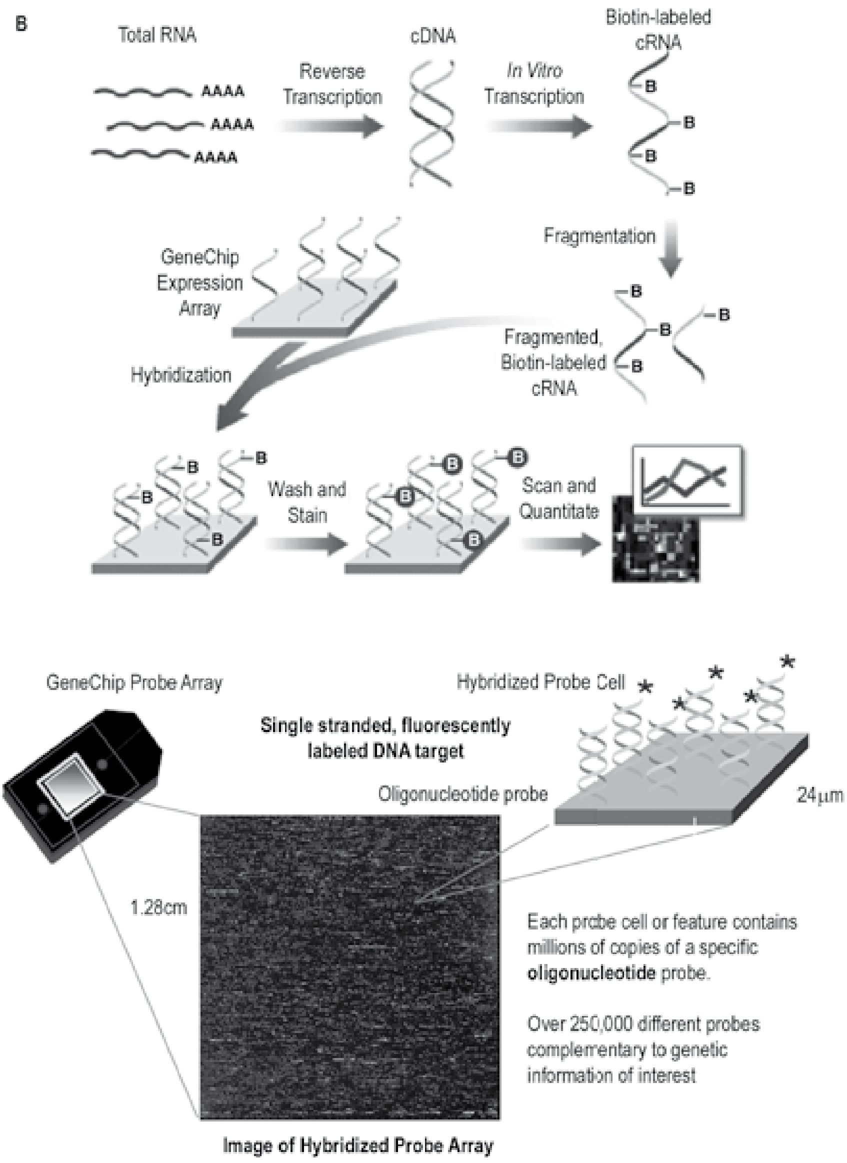


Figure 1. (A) Schematic illustration of cDNA/oligo microarrays. The DNA is first spotted and chemically bound to a slide in an array. RNA is extracted from disease and control tissue/cells and reverse transcribed to make cDNA, which is either directly or indirectly labeled with a fluorophore (Cy3/Cy5 as an example). These cDNA probes are incubated on the microarray slide for a prescribed period of time, the unbound probe is washed off, and the microarray slide is then scanned at the particular laser intensity for the fluorophore to be excited and emit a quantifiable light. These quantified-light emissions are then analyzed as a representation of the expression of the message transcript for that gene to be compared with therespective genes on the control. (B) Schematic illustration of Affymetrix GeneChip Oligo microarrays. The oligos are synthesized directly onto the microarray slide by a lithographic process. RNA is extracted from disease and control tissue/cells and reverse transcribed to make cDNA, which is subsequently *in vivo*


B
Total RNA
AAAA
AAAA
AAAA
Reverse Transcription
cDNA
In Vitro Transcription
Biotin-labeled cRNA
B
B
B
B
Fragmentation
GeneChip Expression Array
Fragmented, Biotin-labeled cRNA
Hybridization
Wash and Stain
Scan and Quantitate
GeneChip Probe Array
Hybridized Probe Cell
Single stranded, fluorescently labeled DNA target
Oligonucleotide probe
24µm
1.28cm
Each probe cell or feature contains millions of copies of a specific oligonucleotide probe.
Over 250,000 different probes complementary to genetic information of interest
Image of Hybridized Probe Array

transcribed with biotin-labeled nucleotides to produce cRNA. These cRNA probes are fragmented and incubated on the microarray slide for a prescribed period of time, and the unbound probe is washed off and stained with biotin-conjugated antibodies, which can emit a fluorescent light. The microarray slide is then scanned at the particular laser intensity for the flourophore to be excited, and the light emissions are quantified and analyzed as a representation of the expression of the message transcript for that gene. Figure adapted from www.affymetrix.com.

Bacterial Artificial Chromosome Arrays

Spectral Genomics (www.spectralgenomics. com) has produced a microarray platform that utilizes bacterial artificial chromosomes (BACs) representing each of the chromosomes (1-Mb resolution) for comparative genomic hybridization (CGH). Reliable and high-resolution detection of changes in copy number of genomic DNA is essential in diagnosing patients for various diseases such as mental retardation linked to trisomy and chromosome imbalances in cancer. CGH makes it possible to screen the whole genome for genomic aberrations. Traditional CGH performed on metaphase chromosomes has a comparatively low resolution (5–10 Mb) with that of the human BAC microarrays. Such a high-resolution platform works very well for the detection of cryptic chromosome rearrangements and compared to traditional mapping techniques is very useful for size mapping of the aberrations, which facilitates the phenotype-genotype correlation [5]. CGH arrays such as those produced by Spectral Genomics are likely to be offered as a genetic test in clinical diagnostic laboratories in the near future. Currently Signaturechip (www.signaturegenomics.com), with its human genome microarrays, is offering diagnostic CGH services. Its chip, which is not commercially available, can simultaneously assay the genome at selected loci with greater resolution than the available conventional cytogenetic testing. Signaturechip's diagnostic service allows for testing of unbalanced translocations, deletions, and duplications of subtelomeric imbalances with more than 125 clinical aberrations tested in one chip. DeLeeuw *et al.* [6] used CGH microarray technology to genomically characterize mantle cell lymphoma (MCL), an aggressive non-Hodgkin's lymphoma. Characteristically, MCL is known to be caused by a t(11;14)(q13;q32) translocation, which is found in virtually all patients; however, research suggests that this single translocation is insufficient to result in the lymphoma. Using such a high-resolution technique as BAC arrays, DeLeeuw *et al.* [6] also defined 13 novel regions, including small intragene deletions that may further identify novel dominant oncogenes and tumor suppressor genes.

Protein/Antibody Arrays

Several companies have developed protein and/or antibody arrays, and these arrays are currently being used for the identification of protein expression. Because RNA expression does not predict subsequent protein expression, these protein arrays are a useful tool in the follow-up of expression arrays. Antibody arrays are also being used in immunological applications such as for human allergies and the monitoring of patient-specific antibodies [7]. A large number of disease-related parameters can be simultaneously monitored using protein microarrays. These arrays are also being used for research into the effects of SNPs and small mutations in the human genome and the translated protein. For example, many thousands of allelic variants of the p53 gene involved in cancer have been described; however, few of the protein products of this gene have been functionally characterized. The application and analysis of protein arrays can provide insight into such functions. Boutell *et al.* [8] quantified in parallel the effects of mutations and polymorphisms on the DNA-binding function of the p53 oncoprotein using a protein microarray, allowing subclassification according to their functional effect. In addition, Sun *et al.* [9] researched the use of protein arrays for parallel detection of tumor markers as a tool for tumor detection both in cancer patients and in screening cancer in asymptomatic populations with high risk.

Exon Arrays

Future directions in microarray platforms include exon arrays. Studies have estimated that 30 to 60% of genes undergo alternative splicing, an important regulatory mechanism often controlled by developmental or tissue-specific factors, and often overlooked or missed when analyzing whole genome expression arrays [10]. In generating expression microarray data on an exon-by-exon basis, rather than measuring on a per gene basis, genetic changes such as alternative splicing can be identified and recognized for their role in disease generation.

Intergenic Arrays

A newer platform of microarray technology is the intergenic arrays. Chromosome regions between annotated open reading frames may contain shorter expressed sequences that have not yet been identified. The discovery of such intergenic expression may lead to the discovery of novel transcripts. Affymetrix (www. affymetrix.com) has begun to produce GeneChip arrays in bacterial species that interrogate the intergenic regions, both in expression arrays and in antisense arrays. These arrays are becoming increasingly important in many areas of biology, specifically when combined with pull-down transcription factor assays.

Recent applications of Microarray Technology

Each of the chip technology platforms exploits the human genome to generate information at a number of different levels in various mechanisms, and as this new technology rapidly increases, a growing number of fields are applying it.

Pharmacogenomics

Pharmaceutical companies are widely utilizing microarray technology to increase the certainty in drug development via improved target identification and a better understanding of the mechanisms of action. In the application of microarray technology, drug effects and drug targets can be studied for their effects over time and in differing tissues with the goal of selecting the optimal drug therapy and dosage for each patient. Watters and McLeod [11] reviewed the importance of applying microarray technology in pharmaoncology, because genetic polymorphisms in drug-metabolizing enzymes and other biochemical molecules are responsible for much of the interindividual differences in the efficiency and toxicity of many chemotherapy agents today. These polymorphisms can affect the outcome of cancer treatment, and knowledge from SNP and expression microarray data can help clinicians to predict patient-treatment relationships and response to therapy. Such information can be used with precedence in future cancer diagnostics and treatments as well as in revealing novel therapeutic targets. It is important to support the clinical trials with microarray technology. Roche (www.roche-diagnostics.com) currently has a P450 chip that measures DNA markers for predicting patient response to many common drugs. The chip detects variations in DNA that are known to affect genes such as cytochrome P450 that control the body's mechanisms for processing drugs. This chip produced by Roche is the first chip using Affymetrix technology that meets the standards for clinical use []12]. Presently, a validated high-throughput clinical microarray environment is being established at Erasmus MC (Rotterdam, The Netherlands) in collaboration with Affymetrix to serve such an important purpose.

Forensics

Since the introduction of DNA polymorphism analysis techniques to forensics, older methodologies such as restriction fragment length polymorphism fingerprinting are rapidly being taken over by more advanced techniques. Analyzing DNA by PCR using minisatellite markers and small tandem repeats has been utilized in recent years. However, with the advancement of microarray technology, in conjunction with the identification of genomewide SNPs, it is possible to obtain as much genetic information as quickly as possible in order to enable rapid individual identification [13,14].

Basic Research

Epidemiology

Microarrays have also been transforming the field of genetic epidemiology. Infectious outbreaks can be monitored or genotypic variations can be determined that underlie disease susceptibility

[15]. Because epidemiology relies on the high-throughput collation of numerous results, microarrays have rapidly become a powerful tool in this area owing to their efficient production of large-scale results.

Cardiovascular

Because cardiac anomalies and cardiac disease are common causes of morbidity in both children and adults, microarray application is commonly used for chromosomal mapping and identification of genes involved in the primary etiology of cardiac disease as well as identification of significant risk factors for the development and advancement of such diseases. Konstantinov *et al.* [16] applied microarray technology to research into the neonatal myocardial stress response during cardiac surgery and found that from the transcriptional profile a compensatory antidisease transcriptional response occurs in the neonatal heart. Using such microarray profiles, various pathways can also begin to be mapped out in these fields of cardiology and vascular research. A serious problem for cancer survivors is radiation-induced vascular injury [17]. By understanding the molecular sequence of events that causes such problems, treatments following or more caution in radiotherapy can be taken. Like cancer research, the progression of and susceptibility for cardiac anomalies and disease can also be monitored using microarray technology, for the benefit of understanding common disease and defects as well as the drug discovery and treatment within this field.

Oncology and disease classification

In the oncology research field, microarrays are used to study diagnostics as well as the progression of disease and heterogeneity to treatment response. Cancer classifications have primarily been based on the morphological appearance of the tumor, but this has serious limitations, because histopathology is insufficient to predict disease progression and clinical outcome. To overcome this, many research groups have begun to apply microarray technology to identify particular pathological subgroups of disease that can predict patient survival and treatment outcomes.

Disease classification not only for cancer has become an important component in downstream microarray analysis. The classification can be divided into two areas: class discovery and class prediction. Class discovery refers to redefining previously unrecognized tumor subtypes and class prediction refers to the assignment of particular tumor samples to the already defined subclass based on a selection of significant genes [18]. Based on this classification, Beer *et al.* [19] identified a set of genes that can predict survival in early stage lung carcinoma. This group also described and delineated a high-risk group that may benefit from adjuvant or supplementary therapy, whereby a pharmacological or immunological agent can be added to the treatment to increase or aid its effect or that of the antigenic response. More recently, advanced statistical tools have been applied to these class discovery and predictions in basic research. Multiple myeloma has been studied by numerous cancer research groups using microarray technologies. Claudio *et al.* [20] confirmed the morphological homogeneity of multiple myeloma. Using microarray disease classification techniques, they also established that although multiple myeloma is morphologically homogeneous, there are underlying differences in individual tumor gene expression patterns that correlate with the heterogeneity of disease severity. Such underlying patterns include immunoglobin translocations and other structural genetic changes that both classify and impact patients' prognosis of cancer. Golub *et al.* [18] used sophisticated statistical methods to automatically classify new cases of acute leukaemia into those arising from lymphoid precursors (acute lymphoblastic leukemia) or from myeloid precursors (acute myeloid leukemia [AML]). More specifically and advanced in the area of AML, Bullinger *et al.* [21] and Valk *et al.* [22] with very large microarray data sets were able to identify subgroups of patients with AML on the basis of molecular signatures and disease classification. Valk *et al.* [22] used gene expression profiles from 286 patients with AML to identify and determine the prognostic significance of AML cases with specific molecular signatures. Using various advanced statistical techniques and visualization tools available today, such as the OmniViz software SAM (Significant Analysis of Microarrays, developed by Stanford) and PAM (Prediction Analysis for Microarrays), they identified 16 subgroups. Genes from these subgroups could be identified as class predictors to identify such prognostically important clusters. These subclasses of AML were defined by various chromosomal lesions such as translocations but also those with normal karyotypes. Some of these unique classes when coupled with extensive clinical data correlated with the prognosis of a poor treatment outcome and could predict overall survival among patients within AML subgroups including that with a neutral karyotype (see Figure 2).

Developments in Microarray Analysis Tools

As well as the progression of microarray technology, there has been great progress in the analysis and mining of microarray data, including the class discovery and prediction as used extensively by Valk *et al.* [22]. Up until recently, microarray studies have been primarily descriptive, rather than analytical, and many have focused primarily on cell culture, rather than primary patient material, which in itself is a disadvantage to analysis, because genetic "noise" may obscure underlying reproducible expression patterns significant to disease [18]. Recently, more awareness of the technology has brought about rigorous experimental designs including subject material, protocols, and statistical procedures in analysis. The application of more systematic approaches to analyze the data based on the simultaneous expression monitoring of the human genome has also been applied to this growing field.

Microarray Statistics

Biologists today have begun teaming up with mathematicians and statisticians to increase the rigor of experimentation and address the problems associated with the manipulation of large data sets [23]. In 2001, Vingron [24], in a bioinformatics editorial, specified the need to adopt a more statistical way of thinking. Today, there are numerous microarray analysis packages with advanced statistical tools. These tools for analyzing data have also become more advanced in their techniques and methods to separate and distinguish variability. This includes the significant variability among samples and also the biological and technical variability, or "noise." The significant variability among samples can thus be cleanly analyzed further with downstream tools and subsequent advanced statistics. One such analysis package is Bioconductor (see www.bioconductor.org) This tool is a user integration of the R language, a widely used open source language and environment for statistical computing and graphics (see www.r-project.org) The Bioconductor package provides software for analyzing microarray and other genomic data, and associating it in real time to biological metadata from Web databases such as GenBank, LocusLink, and PubMed. Bioconductor also provides implementations for a broad range of sophisticated statistical and graphic techniques, including linear and nonlinear modeling, cluster analysis, prediction, resampling, survival analysis, and time-series analysis.

Stanford University has developed several tools for the use in statistical microarray analysis. The SAM tool (as used by Valk *et al.* [22]) is supervised learning software that uses a statistical algorithm to determine differentially expressed genes in a microarray. SAM correlates gene expression data to a wide variety of clinical parameters including treatment diagnosis categories, survival time, and time trends. It provides estimates of false discovery rates for multiple testing using the FDR and q-value methods as presented in Storey [25]. This tool was originally built as an Excel add-in but has since been integrated into the OmniViz software program and Bioconductor package.

The PAM tool, like SAM, was developed by Stanford and is class prediction software for genomic expression data mining within a gene list. This tool can classify genes as class predictors via the nearest shrunken centroid method [26]. It provides a list of significant genes whose expression characterizes each diagnosis class. The PAM tool has also been integrated into the Bioconductor package.

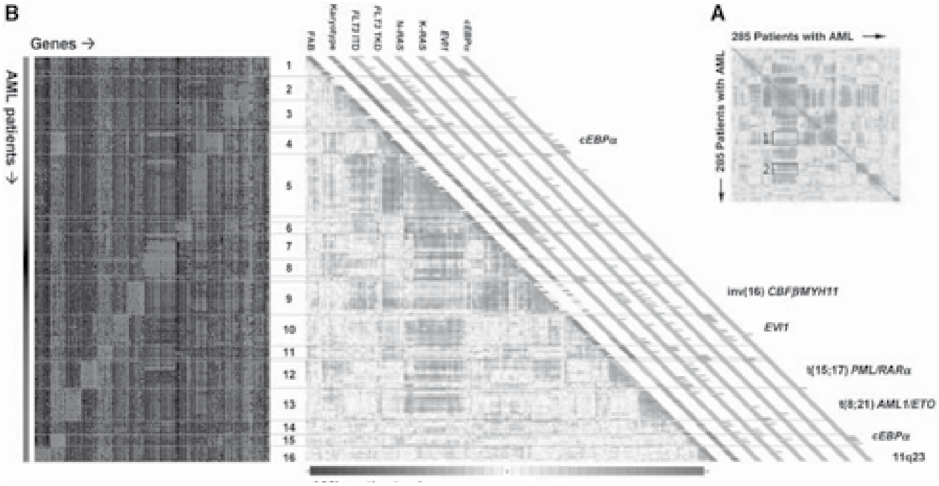


Figure 2. Advanced statistical and visualization techniques used to identify and classify prognostically significant subgroups of AML. Correlation View of specimens from 285 patients with AML involving 2856 probe sets (A) and adapted Correlation View (2856 probe sets) (B, right), and levels of expression of top 40 genes that characterized each of the 16 individual clusters (B, left). In (A), the Correlation Visualization tool displays pairwise correlations between the samples. The colors of the cells relate to Pearson's correlation coefficient values, with deeper colors indicating higher positive (red) or negative (blue) correlations. One hundred percent negative correlation would indicate that genes with a high level of expression in one sample would always have a low level of expression in the other sample and vice versa. Box 1 indicates a positive correlation between clusters 5 and 9 and box 2 a negative correlation between clusters 5 and 12. The red diagonal line displays the intraindividual comparison of results for a patient with AML (i.e., 100% correlation). To reveal the patterns of correlation, we applied a matrix-ordering method to rearrange the samples. The ordering algorithm starts with the most highly correlated pair of samples and, through an iterative process, sorts all the samples into correlated blocks. Each sample is joined to a block in an ordered manner so that a correlation trend is formed within a block, with the most correlated samples at the center. The blocks are then positioned along the diagonal of the plot in a similar ordered manner. In (B) are shown all 16 clusters identified on the basis of the Correlation View. The FrenchAmerican-British (FAB) classification and karyotype based on cytogenetic analyses are depicted in the columns along the original diagonal of the Correlation View. FAB subtype M0 is indicated in black, subtype M1 in green, subtype M2 in purple, subtype M3 in orange, subtype M4 in yellow, subtype M5 in blue, and subtype M6 in gray; normal karyotypes are indicated in green, inv(16) abnormalities in yellow, t(8;21) abnormalities in purple, t(15;17) abnormalities in orange, 11q23 abnormalities in blue, 7(q) abnormalities in

red, +8 aberrations in pink, complex karyotypes (those involving more than three chromosomal abnormalities) in black, and other abnormalities in gray. FLT3 internal tandem duplication (ITD) mutations, FLT3 mutations in the tyrosine kinase domain (TKD), N-RAS, K-RAS, and CEBPA mutations, and the overexpression of EVI1, are depicted in the same set of columns: red indicates the presence of a given abnormality and green its absence. The levels of expression of the top 40 genes identified by the significance analysis of microarrays of each of the 16 clusters as well as in normal bone marrow (NBM) and CD34+ cells are shown on the left side. The scale bar indicates an increase (red) or decrease (green) in the level of expression by a factor of at least 4 relative to the geometric mean of all samples. The percentages of the most common abnormalities (those present in more than 40% of specimens) and the percentages of specimens in each cluster with a normal karyotype are indicated. (Adapted from ref. 21.) (See also page 34 Chapter 1: figure 12)

Microarray Data Storage

As the microarray technology and analysis expands and advances, so does the information that is generated. The shear amount of data that needs to be analyzed and stored is astounding. One company that has addressed this problem, with the help of a useful statistical package, is Rosetta Resolver (see www.rosettabio.com). It utilizes an Oracle database within the package that enables the handling of data volume with the flexibility the analysis requires.

Visualization Tools

In recent years the rapid growth of Internet technology has led to the development of powerful visualization and data manipulation tools for microarray data. Companies such as Spotfire (www.spotfire.com), Inxight (www. inxight.com), and OmniViz (www.omniviz. com) are setting the field for such tools. The OmniViz software program can visualize multidimensional clustering profiles, including correlation plots (which were adapted for presentation of the advanced statistics in Valk *et al.* [22]), galaxy clusters, treescapes for hierarchical and k-means clustering, and coloration/heat maps. The high-powered algorithms within this software enable it to mine external databases such as PubMed and OMIM for text relating to the annotations of one's microarray gene list (see Figure 3). With so much information within such public databases, text mining is becoming a very powerful tool for the future integration of downstream and multiple-source analysis of microarrays.

Pathway Interrogation Tools

As the microarray technology expands and the availability of expression data and the ability to mine such large data sets become available, there is greater recognition of the need and potential to analyze data beyond the simple expression profile. One of the major advantages of microarray expression technology is the ability to uncover biological, biochemical, and metabolic pathways quite rapidly, because the biological function of a gene also provides an extra dimension in which to extract information from a microarray data set. Multidimensional analysis can uncover the major players in disease initiation, and progression, as well as the downstream effects of such genetic aberrations. It can also identify already orphan diseases and those caused by microdeletions as well as begin to identify the biochemical pathways that give a disease phenotype.

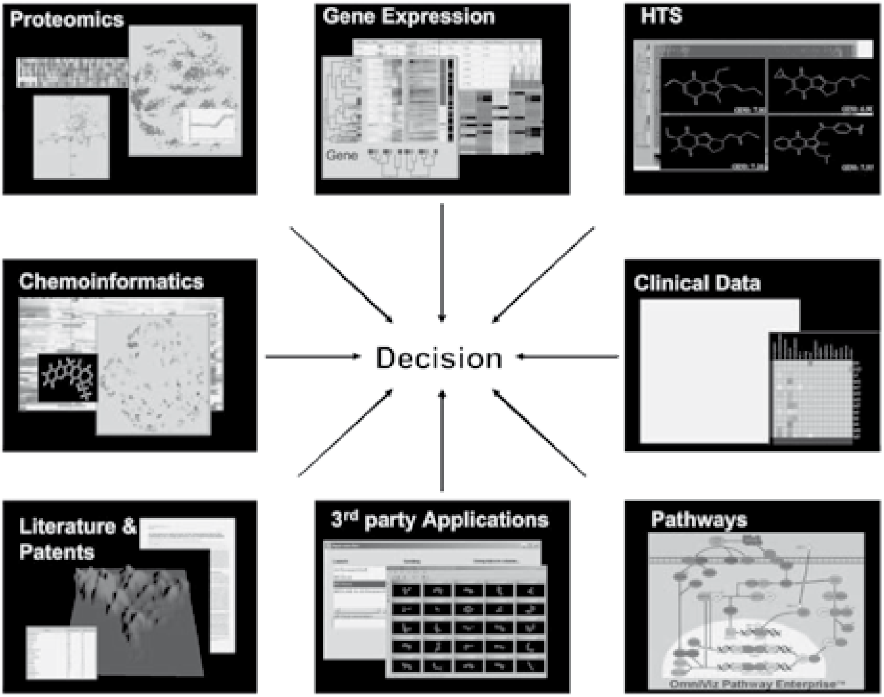


Figure 3. OmniViz decision-making applications. Schematic diagram illustrating various applications of OmniViz software tools to microarray analysis and visualization of data.

There are a large number of open-source and commercially available databases to constitutively mine in parallel with personal interpretation of microarray results. There are also a number of

Web-based bioinformatics tools used today to assist with uncovering such information, such as Ingenuity, which is an application that enables biologists to discover, visualize, and explore therapeutically relevant networks that are significant to their experimental results such as gene or protein expression array data sets. It comprises a database of biological networks created from millions of individually modeled relationships among proteins, genes, complexes, cells, tissues, drugs, and diseases (see www.ingenuity.com) (see Figure 4). Biocarta (see www.brocarta.com/genes/index.asp) and Pathway Assist by Ariadne genomics (see www.ariadnegenomics. com/products/pathway/html) also have dynamic graphic models of how genes interact. OmniViz Pathway Enterprise provides an enterprisewide solution for drawing and analyzing pathways, as well as interrogating pathway data from disparate sources. These tools are highly useful for biological pathway analysis of microarray data, both immediate and downstream.

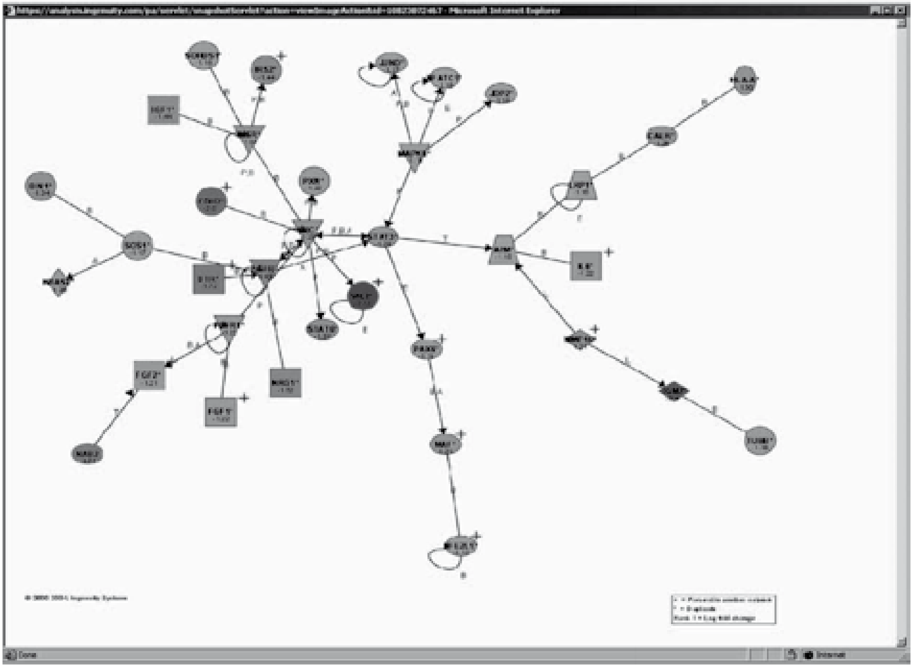


Figure 4. View of an Ingenuity knowledge base network. Ingenuity pathway tools provide an easy to use network explorer visualization to view pathway query results from microarray analysis. The network viewer is color coded and node shaped for distinct results such as an enzyme, kinase, cytokine, transmembrane receptor, or transporter. The viewer also illustrates whether a particular gene was up- or downregulated in the input list and whether it binds, inhibits, or acts on the neighboring genes within the pathway.

Analysis was performed using web tools at www.ingenuity.com**.**

Gene Ontologies

A recently expanded collaboration, The Gene Ontology (GO) Consortium, has helped researchers take a great leap forward in this expanded analysis (see www.geneontology. org). The GO consortium aims to compose a structured terminology describing various aspects of biology that are shared by all living forms in order to annotate genetic data in a consistent way [27]. Three ontology hierarchies are offered: biological process, molecular function, and cellular component. These hierarchies enable researchers to connect their protein and/or genetic data to a GO term, which will allow functional analysis in the aforementioned areas of biology together with the gene expression profiles [28,29]. Smid and Dorssers [29] developed a tool called Go-Mapper, in which GO terms are weighted using the actual measured levels of expression of all associated genes. This quantitatively links gene expression terms to gene expression levels for multiple experiments in an automated way. This tool can also be useful in conjunction with pathway analysis. With massive data sets and heterogeneous results, both pathways and GO tools can predict the significant differential changes in an experiment at the conceptual level and give lead to uncover the impact on underlying biochemical and molecular pathways affected. An understanding of biological pathways does not come from the analysis of a single experiment but from libraries of several experiments. These experiments can be the researcher's own but tend to be quite costly, or they can be from the acquisition and integration of many of the public genome-scale expression databases such as Microarray Gene Expression Database (see www.mged.org/ workgroups/MIAME/ miame/html), Array-Express (see www.ebi.ac.uk/ arrayexpress), and Stanford Microarray Database (see http://genome-www5.stanford. edu/), to name a few. The integration of data originating from different sources, such as sequence data, expression data, CGH data, literature, and chemical structures, should all be linked for efficient data analysis [30].

Current Limitations of a Growing Technology

Comparisons of Microarray Experiments and Platforms

Some major criticized points and downfalls of such new high-throughput technology are the effective management and interpretation of large data sets and the comparison of multiple microarray data sets. Microarray experiments also have their own subtleties and nuances that can make even identical experiments different (see www.affymetrix.com/ community/ wayahead/index.affx). Researchers must be aware that the integration of different microarrays and microarray platforms in the comparison of data such as that represented in public databases or experimental replicates may include various drawbacks for analysis. The difference in probe sequences or spotted material may produce differing signal intensities owing to cross-hybridization and discrimination of differing splice variants as well as the systematic variation in experiments such as labeling difference efficiencies, quantification of RNA, and labeling/experimental protocols. In addition, temporal differences in running the experiments may cause unnecessary variation in data sets, thus obscuring significant results. Three critical aspects of successful experimental design that have been defined by the Tumor Analysis Best Working Group are (1) using sufficient biological replicates, (2) making comparisons between equivalent tissue types, and (3) standardizing tissue sampling and storing procedures. Not applying such criteria to a microarray experiment can generate sufficient noise to hide away the significant results (see www.affymetrix.com/ community/wayahead/index.affx). Such critical aspects can also create problems in the comparison of expression data from public open source microarray databases.

Use of Open Source Information and Tools

Tools such as public databases and software programs for analysis and pathway mapping will not provide all the pieces to the puzzle of downstream analysis to one's experiments, because they are only ever as powerful as the information within them. Although there is an increasing amount of knowledge in databases and analysis tools, if the genetic connection is not already known, these tools will not assist in this form of discovery, and interpretation of the data will need to be made done manually.

Target Validation

Expression target validation is also a very important step in this microarray pipeline. Unfortunately, the amounts of messenger RNA may not always reflect the amounts of protein, and the expression of a protein may not always have a physiological consequence; therefore, labor-intensive *in situ* hybridization and immunohistochemistry experiments are required.

Microarray Genome Coverage

Although there are numerous arrays on the market today, and many new genes are currently being functionally characterized, there are still a large number of genes/families of genes that are underrepresented on microarrays. One such area, as reviewed by Comelli *et al.* [31], is that of glycosyltransferases. There are currently many opportunities to produce custom arrays for such a narrowed field of genomics; however, these can be more expensive than the regular commercially produced arrays. As the genomics field rapidly expands, as it has been doing over the past 4 yr, the genetic representation on the microarray chips will broaden and, thus, extend the possibility for novel discoveries and subsequently the public knowledge databases used to assist analysis. Affymetrix now has a full genome chip for both human and murine (U133A 2plus) that covers almost all of the human genome known today.

Standardized Statistical Measures

The standardized application of statistical measures to analyze microarray data is also very important, and the numerous statistical programs that are now available to analyze and interpret microarray data can create another source of confusion in an experiment. Dr. Warrington from the Tumor Analysis Best Practices Working Group 2004 (www. affymetrix.com/community/wayahead/index .affx) states that she has encountered up to 50% variation between comparisons of two different statistical algorithms, so developing a standardized way to compare gene expression from one array to another is essential, as well as finding a statistical method that is best suited for one's data.

MIAME Guidelines as a Solution

The microarray public has formed certain guidelines in undertaking microarray experiments for the purpose of publication in MIAME, which stipulates the Minimum Information About a Microarray Experiment that is needed to enable interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment (www. mged.org/workgroups/MIAME/miame.html). These MIAME guidelines can facilitate others in running their experiments and analyzing their data. These guidelines have also proven to be a useful tool in downloading, organizing, and understanding previously published data and the ability to be able to compare one's own microarray results with those in public databases. Future MIAME guidelines on the statistical research of data interpretation may one day be stringent enough to have all experimenters comply with a consistent analysis. Although statistics/statistical analysis packages for microarrays are continually improving, it is difficult to conclude objectively that one is more significantly reliable than another. Because microarrays are continually produced with the possibility of inter/intravariation, a microarray is dependent on the distribution of one's data set in order to extract the correct information reliably.

Future of Microarray Technology

Slowly the microarray field is shifting to nano- and septotechnology. Research and development in large companies of this technology are reviewing the possibility of combining a number of arrays in one experiment, so as to study the effects of alterations at the transcriptional level of responses to a variety of stimuli on a single chip (see Figure 2A). Nanochip technology, led by Nanochip, produces a tiny silicon chip powering the experiment by electricity. Each electronic microarray contains 100 test sites laid out in a geometric grid. Each test site can be controlled electronically from the system's onboard computer. A permeation layer coated on the chip acts as the protective interface between the electrically active surface and the biological test environment. Molecules, including DNA and RNA, have natural positive and negative charges. With electricity, it is possible to facilitate both rapid movement to and from, and concentration at, designated test sites on the chip. Current applications performed on the NanoChip® array include SNPs, short tandem repeats, insertions, deletions, and other mutation analyses (see www.nanogen. com/products/nanochip_micro.htm).

Future directions of microarray might also move away from the single-layer platform, which is limited to solid-phase kinetics, to the use of beads (microspheres), solution-phase kinetics, which will give the substrate more surface area to which it can bind. Companies such as Luminex® (www.luminexcorp.com/01_xMAPTechnology/index.html), with its xMAP technology, are leading this technology at present. First, Luminex uses 5.6-μ polystyrene beads, called microspheres. These beads are internally dyed with red infrared fluorophores, and using differing ratios of the fluorophores the beads can be divided into 100 distinct sets. Each bead set can be coated with a reagent specific to a particular bioassay, allowing the capture and detection of specific analytes from a sample. Within the Luminex analyzer, lasers excite the internal dyes that identify each microsphere particle, and also any reporter dye captured during the assay. Many readings are made on each bead set, further validating the results. In this way, xMAP technology allows multiplexing of up to 100 unique assays within a single sample, both rapidly and precisely. The surface chemistry on the beads allows simple chemical coupling of reagents such as antibodies, oligonucleotides, peptides, or receptors, which allows the wide potential for applications such as allergy testing, autoimmune, cancer, cardiac, and metabolic marker identification that defines disease classification, genotyping, and infectious disease antibody testing to name a few.

Conclusion

The merging of robotics, biotechnology, and computer sciences, as well as the completion of genome-sequencing efforts for several organisms, has resulted in groundbreaking changes in the way biomedical research is conducted. Biological researchers have traditionally examined functional genetic information to elucidate fundamental cellular processes and unravel the etiology of human disease. In today's postgenome era, scientists are drowning in data trying to control high-throughput experimental platforms, and understand the millions of interrelations among proteins, small molecules, and phenotypes. It is now possible to manufacture high-density arrays of specified DNA sequences that include every known gene of an organism on a single glass slide. Genomics, informatics, and automation will play increasingly important roles as discovery tools in the basic biological sciences, as well as in diagnostic and therapeutics within the clinical field. Many tools are continually being developed in the microarray field, in both technology and analysis, and the opportunity to apply these technologies to many different fields within bioscience is amazing. Scientists are becoming more aware of microarrays' potential to exploit their research, and as knowledge increases so do the awareness and possible solutions of the limitations microarrays may currently still hold.

Acknowledgments

We thank Dr. Hari S. Sharma, Department of Pharmacology, MC Erasmus for helpful suggestions while preparing the manuscript. Our thanks also go to Peter Valk, Department of Haematology, for allowing us to use Figure 2; Michael J. Moorhouse for detailed review of the manuscript; and Mark Ott and Jeff Saffer for their contribution to Figure 3.

References

1. Shrimpton, A. E., Levinsohn, E. M., Yozawitz, J. M., et al. (2004) *A HOX gene mutation in a family with isolated congenital vertical talus and Charcot-Marie-Tooth disease.* Am. J. Hum. Genet. **75**, 92–96.
2. Middleton, F. A., Pato, M. T., Gentile, K. L., et al. (2004) *Genomewide linkage analysis of bipolar disorder by use of a high-density singlenucleotide-polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22.* Am. J. Hum. Genet. **74**, 886–897.
3. Faraone, S. V. and Tsuang, M. T. (2003) **Heterogeneity and the genetics of bipolar disorder.** Am. J. Med. Genet. **123C** 1–9.
4. Lieberfarb, M. E., Lin, M., Lechpammer, M., et al. (2003) *Genome-wide loss of heterozygosity analysis from laser capture microdissected prostate cancer using single nucleotide polymorphic allele (SNP) arrays and a novel bioinformatics platform dChipSNP.* Cancer Res. **63**, 4781–4785.
5. Schoumans, J., Anderlid, B. M., Blennow, E., Teh, B. T., and Nordenskjold, M. (2004) *The performance of CGH array for the detection of cryptic constitutional chromosome imbalances.* J. Med. Genet. **41**, 198–202.
6. De Leeuw, R. J., Davies, J. J., Rosenwald, A., et al. (2004) *Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes.* Hum. Mol. Genet. **13**, 1827–1837.
7. Harwanegg, C. and Hiller, R. (2004) *Protein microarrays in diagnosing IgE-mediated diseases: spotting allergy at the molecular level.* Expert Rev. Mol. Diagn. **4**, 539–548.
8. Boutell, J. M., Hart, D. J., Godber, B. L., Kozlowski, R. Z., and Blackburn, J. M. (2004) *Functional protein microarrays for parallel characterisation of p53 mutants.* Proteomics **4**, 1950–1958.
9. Sun, Z., Fu, X., Zhang, L., Yang, X., Liu, F., and Hu, G. (2004) *A protein chip system for parallel analysis of multi-tumor markers and its application in cancer detection.* Anticancer Res. **24**, 1159–1165.
10. Wang, H., Yang, U., Lee, C., and Blume, J. (2004) *Pacific Symposium on Biocomputing* 9:3–4 (www.Affymetrix.com).
11. Watters, J. W. and McLeod, H. L. (2003) *Cancer pharmacogenomics: current and future applications.* Biochim. Biophys. Acta **1603**(2), 99–111.
12. Branca, M. (2003) *Roche unveils two pharmacogenomic developments.* Bio. IT World (online).
13. Fukushima, H. (1999) *Forensic DNA analysis- past and future.* Nippon Hoigaku Zasshi *53*, 276–284.
14. Radtkey, R., Feng, L., Muralhidar, M., et al. (2000) *Rapid, high fidelity analysis of simple sequence repeats on an electronically active DNA microchip.* Nucleic Acids Res. **28**, E17.
15. Dalma-Weiszhausz, D. D., Chicurel, M. E., and Gingeras, T. R. (2002) *Microarrays and genetic epidemiology: a multipurpose tool for a multifaceted field.* Genet. Epidemiol. **23**, 4–20.
16. Konstantinov, I. E., Coles, J. G., Boscarino, C., et al. (2004) *Gene expression profiles in children undergoing cardiac surgery for right heart obstructive lesions.* J. Thorac Cardiovasc. Surg. **127**, 746–754.
17. Kruse, J. J., te Poele, J. A., Russell, N. S., Boersma, L. J., and Stewart, F. A. (2004) *Microarray analysis to identify molecular mechanisms of radiation-induced microvascular damage in normal tissues.* Int. J. Radiat. Oncol. Biol. Phys. **58**, 420–426.
18. Golub, T. R., Slonim, D. K., Tamayo, P., et al. (1999) *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science **286**, 531–537.
19. Beer, D. G., Kardia, S. L., Huang, C. C., et al. (2002) *Gene-expression profiles predict survival of patients with lung adenocarcinoma.* Nat. Med. **8**, 816–824.
20. Claudio, J. O., Masih-Khan, E., and Stewart, A.K. (2004) *Insights from the gene expression pro Peeters and Van der Spek filing of multiple myeloma.* Curr. Hematol. Rep. **3**, 67–73.
21. Bullinger, L., Dohner, K., Bair, E., et al. (2004) *Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukaemia.* N. Engl. J. Med. **350**, 1605–1616.
22. Valk, P. J., Verhaak, R. G., Beijen, M. A., et al. (2004) *Prognostically useful gene-expression profiles in acute myeloid leukemia.* N. Engl. J. Med. **350**, 1617–1628.
23. Hardiman, G. (2004) *Microarray platforms- comparisons and contrasts.* Pharmacogenomics **5**, 487–502.
24. Vingron, M. (2001) *Bioinformatics needs to adopt statistical thinking.* Bioinformatics **17**(5), 389, 390.
25. Storey, J. D. (2002) **A direct approach to false discovery rates.** J. Roy. Stat. Soc. B *64*, 479–498.

26. Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002) *Diagnosis of multiple cancer types by shrunken centroids of gene expression.* Proc. Natl. Acad. Sci. USA **99**, 6567–6572.
27. Harris, M. A., Clark, J., Ireland, A., Lomax, J., et al. (2004) *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Res. **32**(Database issue), D258–D261.
28. Ashburner, M., Ball, C. A., Blake, J. A., et al. (2000) *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat. Genet. **25**, 25–29.
29. Smid, M. and Dorssers, L. C. (2004) *GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate gene ontology terms.* Bioinformatics **20**, 1–8.
30. Stubbs, A. and van der Spek, P. (2003) *Micro-array bioinformatics, in Nature Encyclopedia of the Human Genome,* (Cooper, D. N., ed.), Macmillan (UK) and Nature Publishing (NY) groups, pp. 912–917.
31. Comelli, E. M., Amado, M., Head, S. R., and Paulson, J. C. (2002) *Custom microarray for glycobiologists: considerations for glycosyltransferase gene expression profiling.* Biochem. Soc. Symp. **69**, 135–142.

Chapter 4

Epigenetic silencing and mutational inactivation of *E-cadherin* associate with distinct breast cancer subtypes

Antoinette Hollestelle[1]; ***Justine K. Peeters*[3]*;** Marcel Smid[1]; Leon Verhoog[2]; Pieter J. Westenend[4]; Mieke Timmermans[1]; Alan Chan[5]; Jan G.M. Klijn[1]; Peter J. van der Spek[3]; John A. Foekens[1]; Michael A. den Bakker[2]; and Mieke Schutte[1*]

[1]Department of Medical Oncology,

[2] Pathology Josephine Nefkens Institute, Erasmus University Medical Center, 3000 DR Rotterdam, The Netherlands

[3] Department of Bioinformatics, Erasmus University Medical Center, 3000 DR Rotterdam, The Netherlands

[4] Pathology Laboratory, Albert Schweitzer Hospital, 3317 NL Dordrecht, The Netherlands

[5] PamGene International B.V., 5200 BJ 's-Hertogenbosch, The Netherlands

Submitted to Cancer Cell 2008

Summary

Loss of E-cadherin expression has been observed for most human epithelial tumor types, but *E-cadherin* gene mutations have only been identified in half of lobular breast cancers and diffuse gastric cancers. We have investigated *E-cadherin* inactivation in a collection of 41 human breast cancer cell lines. We identified ten *E-cadherin* gene mutants and twelve cell lines with epigenetic *E-cadherin* silencing due to promoter methylation.

E-cadherin mutants were quite distinct from cell lines with silenced *E-cadherin*, including a vastly different intrinsic gene expression program that also involved *E-cadherin* transcriptional repressors, a rounded versus spindle cell morphology, and resemblance to luminal versus basal breast cancers. Loss of wild-type *E-cadherin* expression was causative for the rounded cell morphology but not for the spindle cell morphology, further implying that the two modes of *E-cadherin* inactivation are fundamentally different. A 3-protein spindle cell signature defined on breast cancer cell lines with spindle cell morphology indeed associated with clinical breast cancers of the basal subtype, where E-cadherin protein loss was particularly pronounced in a pathological subtype of metaplastic breast cancer. Importantly, metaplastic breast cancers are typified by transdifferentiated components, suggesting that E-cadherin's role in epithelial mesenchymal transitions may be restricted to breast cancers of the basal subtype. Our evidence for two biologically distinct modes of *E-cadherin* inactivation challenges the paradigm that mutational inactivation and epigenetic silencing of tumor suppressor genes are functionally similar. It also may explain recurrent controversies in E-cadherin research and calls for re-evaluation of functional E-cadherin studies as well as the clinical outcome of patients with E-cadherin-negative breast cancers.

Significance

Many breast cancers lack E-cadherin protein expression, but *E-cadherin* gene mutations have only been identified in the lobular subtype of breast cancers. While lobular breast cancers have a relatively favorable clinical outcome, it has been puzzling that this was not true for E-cadherin negative breast cancers at large. Here, we found that genetic inactivation of the *E-cadherin* gene through mutation is biologically and clinically distinct from epigenetic inactivation of E-cadherin through promoter hypermethylation. These results may explain recurrent discrepancies in both biological and clinical E-cadherin research. But perhaps even more important is that they challenge the paradigm that gene mutation and promoter hypermethylation are similar means to an end in cancer gene inactivation.

Introduction

The mammary gland consists of secretory acini connected by a tree of branching ducts, embedded in a stromal compartment. The epithelium that lines the ductal system is composed of two layers, a luminal epithelial layer and a basal myoepithelial layer. The luminal epithelial cells are cuboidal-shaped and form a polarized continuous layer that lines the lumen. The basal myoepithelial cells are typically spindle or stellate-shaped and reside between the luminal epithelial layer and the basement membrane, forming a nearly continuous layer in the ducts and a discontinuous basket-like structure around the lobular acini. Apart from their morphological appearance and localization, luminal epithelial cells are distinguished from basal myoepithelial cells by their cytokeratin (CK) protein expression profiles: luminal cells express luminal or simple cytokeratins CK7, CK8, CK18 and/or CK19 and basal myoepithelial cells express basal or stratified cytokeratins CK5, CK14 and/or CK17, although it has been noted that luminal cells may also express basal cytokeratins [1-6]. Several other proteins are differentially expressed by the two epithelial layers, such as epithelial membrane antigen (EMA, also known as MUC1), estrogen receptor alpha (ERα) and the progesterone receptor (PR) in luminal epithelial cells and CD10/CALLA, smooth muscle actin (SMA), p63 and the epidermal growth factor receptor (EGFR) in basal myoepithelial cells [7-15]. A series of seminal studies on isolated mammary epithelial cell populations have formed the basis for an epithelial differentiation model in which CK5+ committed stem cells differentiate into bipotent precursor cells (CK5+, CK8/18+) that are positioned suprabasal in the luminal epithelium and produce fully differentiated cells of both the luminal epithelial lineage (CK8/18+) and the basal myoepithelial lineage (SMA+) [16-21]. The existence of a common precursor for luminal and basal mammary epithelia unmistakably illustrates the intricate alliance and relatedness of the epithelial and myoepithelial cell layers in the mammary gland.

Most breast cancers arise in the terminal ductal lobular unit. Pathological classification of breast cancers is based on cytological and architectural features. Ductal type breast carcinoma constitutes about two-thirds of breast cancers and lobular breast cancer accounts for 10-15%. Less prevalent pathological subtypes include tubular, mucinous (colloid), medullary and metaplastic breast cancers [22,23]. Breast cancers can also be classified based on their gene expression profiles. Five intrinsic subtypes of breast cancer were defined based on the intrinsic gene set: luminal A, luminal B, ERBB2+, basal-like and normal-like [24-26]. Cytokeratin protein expression has become another major determinant in breast cancer classification. Combined analyses revealed that some 60-80% of primary invasive breast cancers express only luminal cytokeratins, 20-40% express both luminal and basal cytokeratins, and a minority expresses only basal cytokeratins or is negative for both luminal and basal markers (each <1%) [3,27-29]. Luminal subtype breast cancers, whether defined by intrinsic subtype or by expression of only luminal cytokeratins, frequently express ER and patients have a relatively favorable prognosis. Basal subtype breast cancers more often have the triple-negative phenotype (ER/PR/ERBB2-negative) and their clinical outcome tends to be worse [30,31]. There is however no clear consensus on the definition of this histological subtype of breast cancer (reviewed in [32]) Some have argued that expression of any basal cytokeratin defines basal breast cancers [33], whereas others suggested expression of basal CK14 as being diagnostic [34]. Perou and colleagues defined an immunohistochemical test that identified with high specificity breast cancers of the basal-like intrinsic subtype as defined by the intrinsic gene subset. Using a 4-protein signature, they classified four groups of breast cancers: ERBB2 overexpressing (ERBB2+), luminal (ERBB2- and ER+), basal-like (ERBB2/ER- and CK5/6+ and/or EGFR+), and a negative group that lacks expression of all four proteins [35]. A major discrepancy among these and other definitions lies in breast cancers that express basal cytokeratins as well as ER, which may be as much as one-third of all basal cytokeratin expressing breast cancers [33]. It is likely that a consensus on the definition of basal breast cancers will only be reached once such definition proves clinically highly relevant.

E-cadherin is a major determinant in maintaining epithelial cell integrity. The E-cadherin transmembrane protein is expressed in the adherence junctions of epithelial cells and mediates homophilic cell-cell adhesion between E-cadherin molecules on adjacent cells. The intracellular domain of E-cadherin interacts with either β-catenin or γ-catenin, which are mutually exclusive in the E-cadherin-catenin protein complex. α-Catenin proteins, in their turn, interact with either the actin cytoskeleton or with β-catenin or γ-catenin in a dynamic fashion [36-44]. E-cadherin is a suppressor of invasion and loss of E-cadherin has been noted for most human epithelial tumor types [45-47]. Although this suggests a tumor suppressor function of E-cadherin in multiple tumor types, inactivating *E-cadherin* gene mutations have only been identified in breast cancers and gastric cancers. Importantly, mutations were found in about half of lobular breast cancers and in about half of diffuse gastric cancers [48-51], but not in other subtypes of breast cancer or gastric cancer. Lobular breast cancer and diffuse gastric cancer are both characterized by a typical pathological appearance of diffusely growing, rounded cells with scant cytoplasm. *E-cadherin* gene mutations thus appear to have a profound effect on cell morphology. Less expected was the absence of *E-cadherin* gene mutations among carcinomas from other anatomical sites, or in the remaining breast cancers and gastric cancers. Loss of E-cadherin expression in these carcinomas was suggested to involve transcriptional silencing in association with methylation of CpG islands in the *E-cadherin* promoter region

[52-55]. In this respect, several transcriptional repressors of E-cadherin have been identified: SNAIL, SLUG, SIP1, δEF1, E47, and TWIST [55-62]. Expression of E-cadherin transcriptional repressors has indeed been observed for various carcinoma types and has been associated with a more aggressive clinical course [55]. To gain further insight in loss of E-cadherin expression in tumorigenesis, we studied a model of 41 human breast cancer cell lines. We provide evidence that inactivation of *E-cadherin* through mutation of the gene is biologically distinct from epigenetic silencing associated with promoter methylation, where gene mutations associated with a rounded cell morphology and promoter methylation with a spindle cell morphology. We also show that these different modes of *E-cadherin* inactivation associate with luminal and basal clinical breast cancers, respectively, particularly those of lobular and metaplastic pathology.

Results

A rounded cell morphology typifies *E-cadherin* mutant breast cancer cell lines

We used a human breast cancer cell line model to investigate the biological basis of *E-cadherin* inactivation. An earlier *E-cadherin* sequence analysis [63] was now extended to all 41 breast cancer cell lines from our collection, revealing two more *E-cadherin* mutant cell lines in addition to the eight mutants that already had been identified (Table 1; mutations are detailed in Table S1A in the Supplemental Data). Together, the collection contains one cell line with deletion of the major part of the *E-cadherin* gene, six cell lines with *E-cadherin* mutations that result in premature stopcodons and three mutants with an in-frame deletion. All ten mutants had lost the other *E-cadherin* allele, consistent with the tumor suppressor function of *E-cadherin* [48,49].

Strikingly, all ten *E-cadherin* mutant cell lines grow with rounded cells with scant cytoplasm (Figure 1). These rounded cells may grow in clusters of cells, varying from grape-like bunches to so-called Indian files of cells, or as single cells, that are either attached to adherent cells or freely floating in the culture medium. The cell cultures typically also contain adherent cells that grow as epithelial sheets with diminished cell-cell adhesion, with a cell line-specific percentage of adherent cells that varies from less than ten percent to over ninety percent of the cell population. In addition to the ten *E-cadherin* mutant cell lines, two *E-cadherin* wild-type cell lines also have this rounded cell morphology. The presence of small rounded cells with scant cytoplasm is a cytological characteristic of the lobular pathological subtype of human breast cancer. Also, *E-cadherin* gene mutations are identified in half of lobular cancers, but not in other subtypes of breast cancer [48,49]. The twelve breast cancer cell lines with the rounded cell morphology thus appear to resemble lobular breast cancers. Consistent with this notion, four *E-cadherin* mutant cell lines from our collection were known to be derived from breast cancers with lobular characteristics (EVSA-T, MDA-MB-134VI, MDA-MB-330 and SUM44PE [64-66] and personal communication Dr. S.P. Ethier).

Apart from these rounded cell lines, two other major morphology groups were apparent among the breast cancer cell lines (Table 1 and Figure 1). First, fifteen "epithelial" cell lines grow in sheets of adherent epithelial cells or in spheroid-like cell clusters. Second, thirteen "spindle" cell lines grow rather similar to fibroblasts, yet with extensions that are less pronounced than those of fibroblasts. Cell line DU4475 could not be assigned to either of these three morphology groups. DU4475 is an atypical breast cancer cell line in that it carries an *APC* gene mutation [67] and has constitutive Wnt signaling activation [63]. All 29 non-rounded cell lines had *E-cadherin* wild-type genes, whether epithelial, spindle or unclassified.

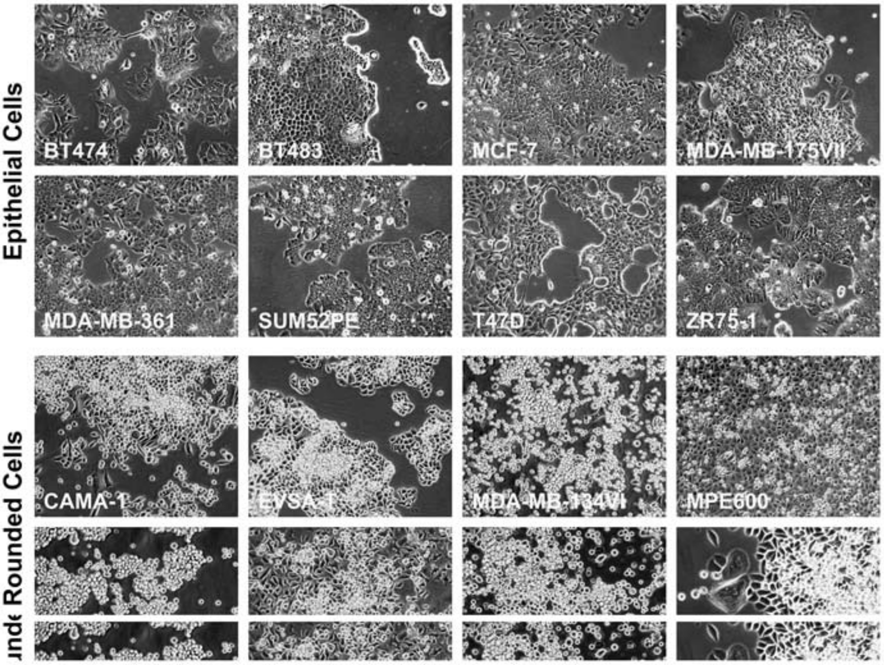


Figure 1. Morphology of Human Breast Cancer Cell Lines. Eight examples are given for each of three morphology groups: Epithelial cells that all have wild-type *E-cadherin* genes and apparently normal E-cadherin protein expression; Rounded cells that all have mutant *E-cadherin* genes; and spindle cells that all have methylation at the *E-cadherin* promoter region.

A spindle cell morphology typifies breast cancer cell lines with *E-cadherin* promoter methylation

We evaluated the effects of the *E-cadherin* gene mutations by expression analysis. *E-cadherin* transcripts were detected by duplex RT-PCR with the *HPRT* housekeeper, and E-cadherin proteins were detected by western blotting using HECD-1 antibody (Table 1 and Figure 2A and 2B). Twelve breast cancer cell lines had low or no detectable *E-cadherin* transcript levels. These cell lines included four of the seven truncating *E-cadherin* mutant cell lines. As expected, none of these four cell lines, nor any of the other three truncating *E-cadherin* mutants expressed detectable levels of E-cadherin proteins. All three in-frame *E-cadherin* mutant cell lines, however, expressed *E-cadherin* transcripts and proteins at apparently normal levels. The eight other breast cancer cell lines with low or no detectable *E-cadherin* transcripts indeed also did not express E-cadherin proteins. Remarkably, these eight cell lines all grow with the spindle cell morphology (Table 1). Five other breast cancer cell lines from our collection also have the spindle cell morphology, but these five cell lines all expressed *E-cadherin* transcripts and proteins. Note that all spindle cell lines had *E-cadherin* wild-type genes. E-cadherin proteins were also expressed at apparently normal levels in all epithelial and unclassified cell lines. Together, these observations suggest that silencing of *E-cadherin* expression is not always due to mutation of the gene. In fact, the typical spindle cell morphology of the silenced *E-cadherin* wild-type cell lines suggests that their mechanism of *E-cadherin* inactivation may be biologically distinct from inactivation through *E-cadherin* gene mutation, as the latter always correlates with a rounded cell morphology.

Four of the spindle cell lines had reportedly silenced *E-cadherin* gene expression in association with methylation of CpG islands in its promoter region (Hs578T, MDA-MB-231 and MDA-MB-435s; [53, 68]). To investigate the extent of methylation inactivation in our breast cancer cell line model, we determined methylation-associated silencing of *E-cadherin* by azacytidine treatment of the cell lines combined with methylation-specific PCR (MSP) for CpG islands 1 and 3 (CpG1 and CpG3, respectively) [69] (Table 1 and Figure 2C and 2D). MSP for CpG3, that is located nearest to the transcription start site, identified methylation in all eight E-cadherin-negative spindle cell lines and weak methylation in a single E-cadherin-expressing spindle cell line. No methylation was detectable in twenty-nine other cell lines (Figure 2D). CpG1 MSP products were detected for twelve of the thirteen spindle cell lines, albeit that the methylation levels of cell lines that were also methylated at CpG3 appeared somewhat higher. Accordingly, the same twelve spindle cell lines also showed upregulation of *E-cadherin* in the azacytidine assay, whereas transcript expression was not restored in any of the four cell lines that had no or low E-cadherin expression due to mutation of the *E-cadherin* gene (Figure 2C). Thus,

Table 1. Molecular Status of Human Breast Cancer Cell Lines with Respect to E-cadherin and Related Proteins. Breast cancer cell lines are organized by their morphology and then by their E-cadherin status. E-cadherin gene mutations are detailed in Table S1 in the Supplemental Data and E-cadherin methylation has been indicated for CpG islands 1 and 3 in the promoter region. nd, not determined; -, negative or absent; ±, low or barely detectable expression; +, clearly detectable expression.

Breast cancer cell line	Cell morphology	Allelic loss at 16q	*E-cadherin* gene sequence	*E-cadherin* methylation	*E-cadherin* transcript	*E-cadherin* protein	β-catenin protein	E47 protein	SIP1 transcript	δEF1 transcript	SLUG transcript	TWIST transcript	SNAIL transcript
HCC1937	epithelial	no	wild-type	-	+	+	+	-	-	+/-	+	+/-	+
MCF-7	epithelial	no	wild-type	-	+	+	+	-	-	+/-	+	+/-	+
BT474	epithelial	no	wild-type	-	+	+	+	-	-	+/-	+	+	+
UACC812	epithelial	no	wild-type	-	+	+	+	-	-	+/-	+	+	+
BT483	epithelial	no	wild-type	-	+	+	+	-	-	-	+	+	+
MDA-MB-175VII	epithelial	no	wild-type	-	+	+	+	-	-	-	+	+	+
MDA-MB-361	epithelial	no	wild-type	-	+	+	+	-	-	-	+	+	+
MDA-MB-415	epithelial	no	wild-type	-	+	+	+	-	-	-	+	-	+
SUM52PE	epithelial	yes	wild-type	-	+	+	+	-	-	-	+	-	+
SUM190PT	epithelial	no	wild-type	nd	+	+	+	-	nd	nd	+	nd	+
SUM225CWN	epithelial	yes	wild-type	nd	+	+	+	-	nd	nd	+	nd	+
SUM185PE	epithelial	yes	wild-type	-	+	+	+	-	-	-	+/-	+	+
T47D	epithelial	yes	wild-type	-	+	+	+	-	-	-	-	+	+
ZR75-1	epithelial	yes	wild-type	-	+	+	+	-	-	-	-	+	+
UACC893	epithelial	no	wild-type	-	+	+	+	-	-	+/-	-	-	+
MDA-MB-330	rounded	no	wild-type	-	+	+	+	-	-	+	-	+	+/-
MDA-MB-468	rounded	yes	wild-type	-	+	+	+	-	-	-	+/-	+/-	+
CAMA-1	rounded	yes	in-frame mutant	-	+	+	+	-	-	+/-	+/-	+	+
EVSA-T	rounded	yes	in-frame mutant	-	+	+	+	-	-	-	-	+	+
MPE600	rounded	yes	in-frame mutant	-	+	+	+	-	-	-	+/-	+	+

OCUB-F/-M	rounded	yes	truncating mutant	-	+	-	+/-	-	-	+/-	+/-	+	+/-
SUM44PE	rounded	yes	truncating mutant	nd	+	-	+/-	-	-	+/-	+/-	+	+
MDA-MB-134VI	rounded	yes	truncating mutant	-	+	-	-	-	-	+/-	+/-	+	+
SK-BR-5	rounded	yes	truncating mutant	-	+/-	-	-	-	-	-	+/-	+	+
MDA-MB-453	rounded	yes	truncating mutant	-	+/-	-	+/-	-	-	-	+/-	-	+/-
ZR75-30	rounded	yes	truncating mutant	-	+/-	-	-	-	-	-	+/-	+	+
SK-BR-3	rounded	yes	deletion mutant	-	-	-	+/-	-	-	-	+/-	+/-	+
BT549	spindle	yes	wild-type	CpG1/CpG3	-	-	+	+	+	+	+/-	+	+
Hs578T	spindle	no	wild-type	CpG1/CpG3	-	-	+	+	+	+	+	+	+
MDA-MB-435s	spindle	yes	wild-type	CpG1/CpG3	-	-	+	+	+	+	+	+	+
MDA-MB-436	spindle	no	wild-type	CpG1/CpG3	-	-	+	+	+	+	+	+	+
SUM159PT	spindle	no	wild-type	CpG1/CpG3	-	-	+	+	+	+	+	+	+
SUM1315MO2	spindle	no	wild-type	CpG1/CpG3	-	-	+	+	+	+	+	+	+/-
MDA-MB-157	spindle	no	wild-type	CpG1/CpG3	+/-	-	+	+	+	+	+	+	+
MDA-MB-231	spindle	yes	wild-type	CpG1/CpG3	+/-	-	+	+	+	+	+	-	+/-
SK-BR-7	spindle	no	wild-type	CpG1/CpG3?	+	+	+	+	+/-	+	+/-	+	+
SUM102PT	spindle	no	wild-type	CpG1	+	+	+	+/-	+	+	+	+	+
SUM149PT	spindle	no	wild-type	CpG1	+	+	+	-	+	+/-	+	+	+
SUM229PE	spindle	no	wild-type	CpG1	+	+	+	+/-	+/-	+/-	+	+	+/-
BT20	spindle	yes	wild-type	-	+	+	+	-	-	+	+	-	+/-
DU4475	other	no	wild-type	-	+	+	+	-	-	+	+	-	+

E-cadherin promoter methylation associates with a spindle cell morphology albeit that complete down regulation of *E-cadherin* transcript and protein expression involves methylation at both CpG1 and CpG3.

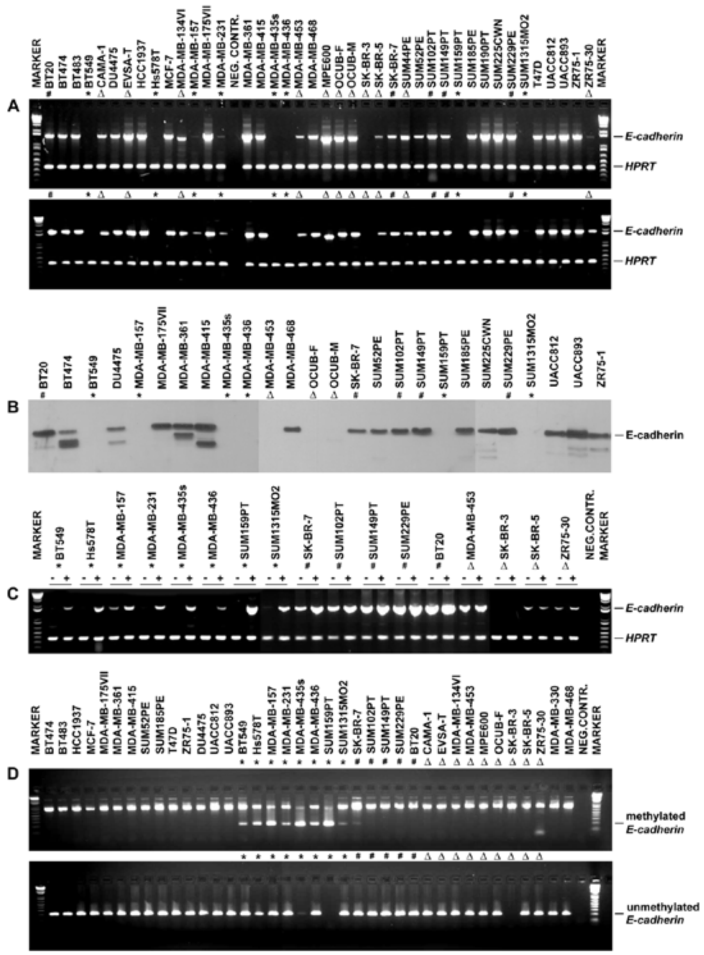


Figure 2. *E-cadherin* Expression and Methylation Analysis of Human Breast Cancer Cell Lines. (A) *E-cadherin* transcript expression by duplex RT-PCR with the *HPR*T housekeeper, using primers directed at overlapping 5' and 3' fragments of the *E-cadherin* transcript (top and bottom panel, respectively). Low or barely detectable transcript levels were identified in cell lines with truncating *E-cadherin* mutations or in cell lines with *E-cadherin* CpG1 and CpG3 methylation. (B) E-cadherin protein expression by western blotting with HECD-1 antibody that is directed at an extracellular epitope. Spindle cell lines with *E-cadherin* methylation at CpG1 and CpG3

do not express E-cadherin proteins, whereas spindle cell lines with methylation at only CpG1 have detectable protein expression. (C) *E-cadherin* transcript expression by duplex RT-PCR with the *HPRT* housekeeper, upon azacytidine methylation assays. + and -, cells were cultured in the presence or absence of 5-aza-2-deoxycytidine. *E-cadherin* transcript expression was induced upon azacytidine treatment in *E-cadherin* methylated cell lines but not in *E-cadherin* mutant cell lines. (D) Methylation-specific PCR for CpG3 at the *E-cadherin* promoter region, using primers specific for methylated or unmethylated CpG3 (top and bottom panel, respectively). Whereas 12 of 13 spindle cell lines had methylation at CpG1, only eight or perhaps nine spindle cell lines had methylation at CpG3. Loss of E-cadherin protein expression was observed only for cell lines with methylation at both CpG islands (Table 1). Δ, Rounded cell lines with a mutant *E-cadherin* gene; #, Spindle cell lines with E-cadherin protein expression *, Spindle cell lines without E-cadherin protein expression; Neg. control, template-negative amplification reaction; Marker, 1-kb ladder (Invitrogen).

Ectopic *E-cadherin* expression restores an epithelial cell morphology in *E-cadherin* mutant cells, but not in *E-cadherin* methylated cells

We investigated causality of E-cadherin expression in determining cell morphology, by ectopic expression of human *E-cadherin* wild-type cDNA in the *E-cadherin* mutant cell line SK-BR-3 and in the *E-cadherin* methylated cell line MDA-MB-231. Cell line SK-BR-3 has a homozygous deletion of the major part of the *E-cadherin* gene and grows as rounded cells, whereas cell line MDA-MB-231 has an *E-cadherin* wild-type gene and grows with spindle cells (Table 1 and Figure 1). Neither of the two cell lines express detectable levels of *E-cadherin* transcripts and proteins (Figure 2A and B), rendering them suitable models for *E-cadherin* reconstitution experiments. Also, SK-BR-3 expresses low levels of β-catenin proteins while MDA-MB-231 expresses normal levels of β-catenin proteins, even though both cell lines express β-*catenin* transcripts at apparently normal levels. Notably, all seven cell lines with truncating *E-cadherin* mutations had reduced β-catenin protein levels, whereas all thirteen spindle cell lines expressed β-catenin proteins at apparently normal levels – irrespective of their E-cadherin protein expression levels (Table 1). Cell lines SK-BR-3 and MDA-MB-231 were also reconstituted with *E-cadherin* delEx9 cDNA, that contains an in-frame deletion of exon 9 of the gene, and with the empty vector. The delEx9 deletion had been identified in breast cancer cell line MPE600 (Table S1A in the Supplemental Data) and in several clinical specimens of diffuse gastric cancer [50, 63, 70], suggesting that this deletion is of functional relevance in human tumorigenesis. Importantly, MPE600 cells still express E-cadherin proteins and β-catenin proteins, rendering

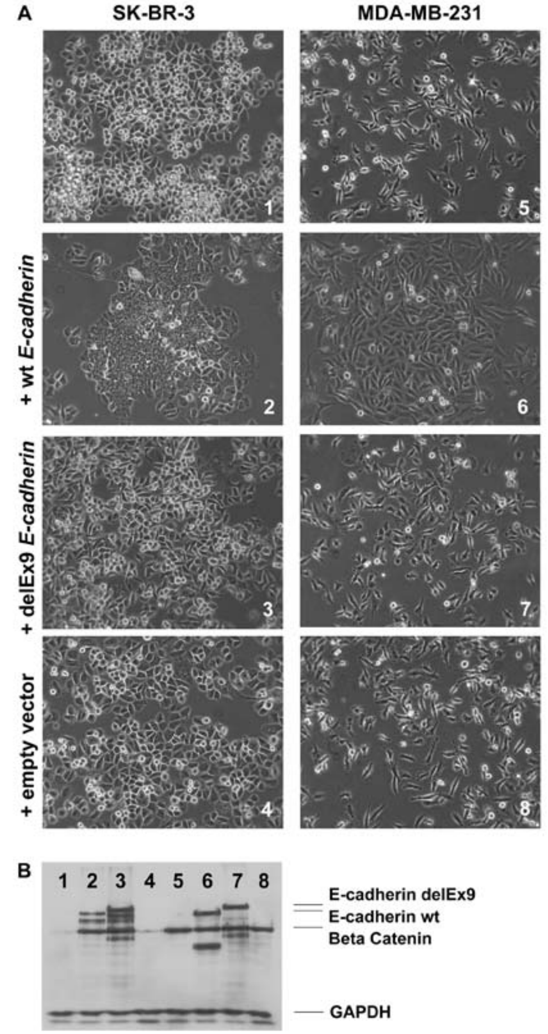


Figure 3. E-cadherin Reconstitution Experiments. Human *E-cadherin* cDNA was stably expressed in two E-cadherin null breast cancer cell lines. SK-BR-3 has lost protein expression due to a large genomic homozygous deletion of *E-cadherin* and MDA-MB-231 had lost protein expression due to *E-cadherin* promoter methylation. Expression constructs contained wild-type *E-cadherin* or mutant *E-cadherin* with an in-frame deletion of exon 9 (delEx9). (A) Conversion from rounded cell morphology to epithelial cell morphology was observed for SK-BR-3 cells transfected with the wild-type *E-cadherin* construct but not with delEx9 *E-cadherin* construct or the empty vector. MDA-MB-231 cells remained spindle-shaped irrespective of the construct

used. (B) E-cadherin and β-catenin protein expression by Western blotting, using GAPDH as housekeeper control. Note that expression of wild-type or delEx9 *E-cadherin* cDNA also resulted in expression of β-catenin proteins in SK-BR-3. MDA-MB-231 cells had not lost β-catenin protein expression, nor had any of the other *E-cadherin* methylated cell lines (Table 1) [1-4], SK-BR-3 cells untransfected, or stably transfected with wild-type *E-cadherin*, delEx9 *E-cadherin*, or empty vector [5-8], lbid. for MDA-MB-231 cells.

the delEx9 cDNA an appropriate negative control. We generated independent stable *E-cadherin* transfectants for each of four reconstitution combinations (2 cell lines x 2 *E-cadherin* constructs), by G418-selective growth under limiting dilution conditions (resulting in monoclonality with P<0.05). All of about 25 clones that were collected from each reconstitution combination were analyzed for cell morphology and for E-cadherin and β-catenin protein expression by immunohistochemistry using C-terminal antibodies.

Three SK-BR-3 clones that had been transfected with *E-cadherin* wild-type cDNA expressed both E-cadherin and β-catenin proteins and all three clones had converted from the rounded cell morphology to the epithelial cell morphology (Figure 3). Six *E-cadherin* delEx9 SK-BR-3 clones also expressed both E-cadherin and β-catenin proteins, but none of them had converted to the epithelial cell morphology, suggesting that a wild-type *E-cadherin* gene was pivotal for the morphology conversion (Figure 3). None of the other SK-BR-3 clones (wild-type, delEx9, and empty vector) expressed E-cadherin or β-catenin proteins, and neither had they converted to the epithelial cell morphology. These results imply that inactivation of *E-cadherin* through mutation of the gene is causal in determining the rounded cell morphology of the mutant breast cancer cell lines.

Reconstitution of the spindle-shaped *E-cadherin* methylated MDA-MB-231 cell line yielded six *E-cadherin* wild-type clones and eight *E-cadherin* delEx9 clones that expressed E-cadherin proteins (and retained β-catenin protein expression). The *E-cadherin* wild-type clones grew less dispersed than the untransfected MDA-MB-231 cell line or the clones reconstituted with the *E-cadherin* delEx9 cDNA or empty vector. Yet, all MDA-MB-231 clones retained the spindle cell morphology, whether they were reconstituted with *E-cadherin* wild-type or delEx9 cDNA, and whether or not they expressed E-cadherin proteins (Figure 3). Thus, silencing of (wild-type) *E-cadherin* gene expression in association with promoter methylation is not a major determinant for the spindle cell morphology. In fact, the observation that several spindle cell lines are not methylated at CpG island 3 at the *E-cadherin* promoter and also still express E-cadherin proteins, suggests that loss of *E-cadherin* expression is secondary to the morphological differentiation status of the cells.

Expression of SIP1 and E47 repressors, but not SLUG, SNAIL, δEF1 and TWIST, is restricted to breast cancer cell lines with the spindle cell morphology

Conversion from an epithelial cell morphology to a spindle-shaped or fibroblast-like cell morphology had reportedly been associated with transcriptional repression and downregulation of *E-cadherin* [56-62]. We therefore analyzed the breast cancer cell lines for expression of the E-cadherin transcriptional repressors *SNAIL*, *SLUG*, *SIP1*, δ*EF1* and *TWIST* by qRT-PCR, and E47 by western blotting (Table 1). All six E-cadherin repressors were expressed in most of the thirteen spindle cell lines (in 11, 12 or all spindle cell lines). In contrast to *SIP1* and E47, δ*EF1*, *TWIST*, *SLUG* and *SNAIL* were also expressed in a substantial proportion of non-spindle cell lines (in 11, 21, 23 and all 28 non-spindle cell lines). However, the expression levels of δ*EF1*, *TWIST* and *SLUG*, but not of *SNAIL*, typically were somewhat higher in the spindle cell lines compared to the non-spindle cell lines (with average Ct values of 26 vs. 34, 25 vs. 28, 22 vs. 30, and 28 vs. 27, for δ*EF1*, *TWIST*, *SLUG* and *SNAIL*; Table 1). Thus, expression of *SIP1* and E47 repressors, and to a lesser extent the other E-cadherin repressors, strongly associates with the spindle cell morphology, albeit that this does not always result in significant downregulation of *E-cadherin*.

Genetic *E-cadherin* inactivation occurs in luminal breast cancer cell lines and methylation-associated silencing in basal breast cancer cell lines

Our analyses thus far strongly suggested that genetic mutation of *E-cadherin* is fundamentally distinct from transcriptional silencing of *E-cadherin* through promoter methylation, where the former is typified by the rounded cell morphology and the latter by the spindle cell morphology. To conclusively resolve whether these two mechanisms of *E-cadherin* inactivation indeed involve different biological pathways, we determined gene expression profiles of 36 breast cancer cell lines using Affymetrix U133A microarrays. Unsupervised Pearson correlation, in which samples are positioned according to their overall similarity in gene expression profiles, revealed two main clusters of cell lines, whether the correlation was calculated from a log2GM <-2 and >2 probe subset or log2GM <-3 and >3 probe subset (5527 and 2000 probe sets, respectively; data shown for log2GM <-2 and >2 in Figure 4A). The lower cluster included all 13 spindle cell lines, a single epithelial cell line and a single rounded cell line (HCC1937 and MDA-MB-468, respectively). The upper cluster included all other epithelial and rounded cell lines intermingled, but none of the spindle cell lines (Figure 4).In fact, the lower cluster could be further subdivided into a major subgroup of ten spindle cell lines (8 E-cadherin negative and 2 E-cadherin positive), and a minor subgroup of three spindle cell lines (all E-cadherin positive) and the HCC1937 and MDA-MB-468 cell lines. It is important to note that morphological classification of HCC1937 was somewhat ambiguous, as this cell line grows with spindle features at low cell densities but has an epithelial cell morphology at higher cell densities. The unambiguous division of the spindle cell lines from the non-spindle cell lines, without any supervision on the samples, indicates that the spindle cell lines have a gene expression program that is very different from that of the non-spindle cell lines. This implies that the spindle cell lines have a distinct differentiation status that first of all is characterized by its morphological appearance and only secondary by epigenetic silencing of *E-cadherin* expression.

Gene expression profiling of clinical breast cancers had defined five intrinsic subtypes of breast cancer [24-26]. When we classified the breast cancer cell lines based on the intrinsic gene set, the minor spindle cell subgroup was identified as the 'basal-like' intrinsic subtype and the major spindle cell subgroup was identified as 'normal-like' (Figure 4B). The epithelial/rounded cell line cluster classified as 'luminal', with a minor 'ERBB2' subgroup that included four of nine cell lines with ERBB2 overexpression. As the intrinsic gene set had been defined on clinical breast cancers, classification of the cell lines according the intrinsic molecular subtypes not only implied that these subtypes were determined by gene expression of the tumor cells in the clinical specimens (in stead of non-malignant stromal cells) but also that breast cancer cell lines are a relevant model to study human breast cancer.

Perou and colleagues recently also defined a 4-protein signature to identify breast cancers of the basal-like intrinsic subtype by immunohistochemistry [35]. Analysis of the breast cancer cell lines with this 4-protein signature revealed a strong correlation between classification based on the 4-protein signature and that based on the intrinsic gene set, although the intrinsic normal-like cell lines were classified as basal-like, five more cell lines were classified as ERBB2+, and two cell lines are classified as negative (SUM185PE and MDA-MB-435s; Figure 4B). Yet, we were able to distinguish intrinsic normal-like from intrinsic basal-like breast cancer cell lines by their absence of basal and luminal cytokeratins (CK5 and CK8/18) and E-cadherin protein expression (Figure 4B). Most important, these results indicate that methylation-associated loss of E-cadherin protein expression resides in basal-like breast cancer cell lines whereas loss of E-cadherin protein expression due to mutation of the gene is restricted to luminal breast cancer cell lines, implying that these two mechanisms of E-cadherin inactivation are biologically distinct.

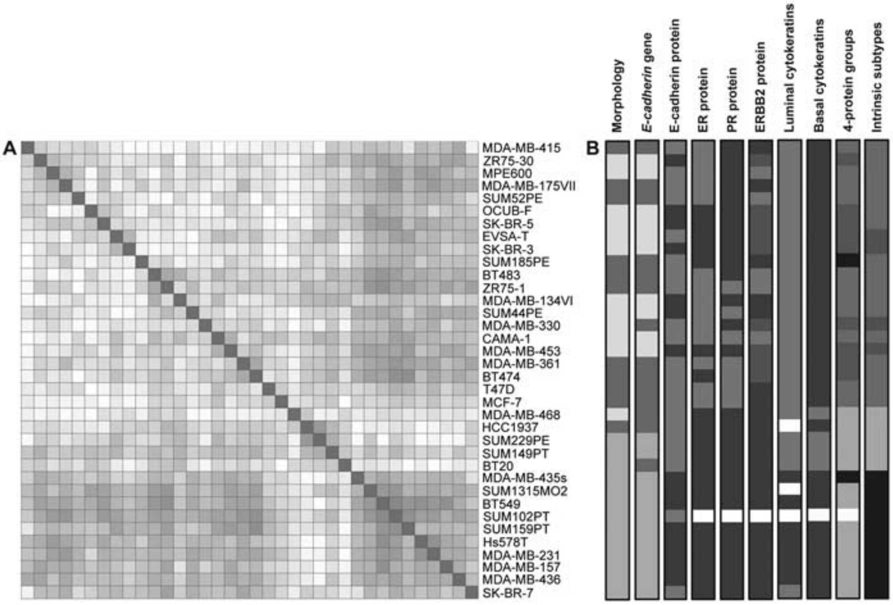


Figure 4: Pearson correlation and Expression of Breast Cancer Associated Proteins in Human Breast Cancer Cell Lines. (A) Pearson correlation plot based on the log2GM <-2 and >2 subset (5527 probe sets). The Pearson correlation coefficient algorithm positions samples according their overall similarity in gene expression, where red indicates high overall similarity (positive correlation coefficient) and blue indicates low similarity (negative correlation coefficient). (B) Various characterizations of the cell lines indicated that the upper cluster in the Pearson correlation plot contains the epithelial and rounded cell lines intermingled whereas the lower cluster contains all spindle cell lines and a single epithelial cell line and a single rounded cell line (HCC1937 and MDA-MB-468). This lower cluster included two subgroups that by the intrinsic gene set classified as basal-like and normal-like intrinsic subtypes, where all E-cadherin-negative spindle cell lines classified as normal-like. The lower cluster classified as basal breast cancers by the 4-protein signature. Color coding morphology column: green, epithelial morphology; yellow, rounded cell morphology; orange, spindle cell morphology. *E-cadherin* gene column: green, wild-type *E-cadherin* gene; yellow, mutant *E-cadherin* gene; orange, methylated *E-cadherin* gene. E-cadherin protein, ER protein, PR protein, ERBB2 protein, luminal cytokeratins and basal cytokeratins columns: red, protein expression; blue, no protein expression; brown, protein overexpression. 4-protein groups column: green, luminal group; brown, ERBB2+ group; black, negative group; orange, basal-like group. Intrinsic subtypes column: green, luminal subtype; brown, ERBB2+ subtype; orange, basal-like subtype; black, normal-like subtype.

In contrast to lobular breast cancers, frequent loss of E-cadherin protein expression in metaplastic breast cancers is not associated with mutational inactivation of the gene

E-cadherin gene mutations were reportedly identified in breast cancers with lobular pathology [48,49]. Consistent with our cell line data, lobular breast cancers are of the luminal subtype of breast cancers. Here, we report methylation-associated silencing of *E-cadherin* in basal breast cancer cell lines, suggestive for another pathological subtype of clinical breast cancers with frequent loss of E-cadherin protein expression. To search for this putative breast cancer subtype in an unbiased manner, we analyzed 166 primary breast cancers of a variety of pathological subtypes by immunohistochemistry. Tissue microarrays included ductal breast cancers (24 ER+ and 22 ER-) and lobular, mucinous, tubular and medullary breast cancers (28, 20, 16 and 22 cases, respectively). Whole sections were analyzed for 34 metaplastic breast tumors because of their characteristic morphological heterogeneity. As expected, loss of E-cadherin protein expression was frequently observed among lobular breast cancers (82%, Figure 5B). Interestingly, loss of E-cadherin protein expression was also observed for half of metaplastic breast cancers, whereas protein loss was found for only 20% of mucinous and 14% of medullary breast cancers and never exceeded 10% of breast cancers from other pathological subtypes. Evaluation of the clinical breast cancers for the 4-protein signature [Nielsen, 2004 #28 indicated that all 28 lobular breast cancers were indeed of the luminal group whereas the vast majority of metaplastic breast cancers were basal-like (94%, and 6% negative group; Figures 5A and 6). The duality that we observed for E-cadherin loss among the breast cancer cell lines was thus reproduced in clinical breast cancers from the lobular and metaplastic pathological subtypes.

To confirm that mutational inactivation of *E-cadherin* associates with lobular breast cancer and epigenetic silencing with metaplastic breast cancer, we screened all E-cadherin-negative clinical breast cancers with DNA available for mutations in the *E-cadherin* gene, by direct sequencing (Figure 5B). We identified ten mutants among the 23 luminal breast cancers and all mutations predicted premature truncation of the encoded E-cadherin proteins (detailed in Table S1B in Supplemental Data). All mutations were identified in breast cancers with lobular histology (Figure 5B). Importantly, no *E-cadherin* mutations were detected among 17 basal-like and a single negative breast cancer. We also addressed *E-cadherin* promoter methylation by MSP of CpG1 and CpG3, but methylation was observed in all samples, irrespective of their pathological subtype. Evaluation of *E-cadherin* promoter methylation in primary cancer specimens was reported to be severely hampered by the inevitable presence of leukocytes with *E-cadherin* methylation [71]. Indeed, we even detected *E-cadherin* methylation in all of eight dissected primary breast cancer samples (with <5% leukocytes), effectively precluding analysis of *E-cadherin* promoter methylation in clinical breast cancers. We therefore took advantage of the characteristic gene expression program that we observed for the spindle cell lines, as an indirect but more comprehensive measure for the differentiation program that involves epigenetic *E-cadherin* silencing. A spindle cell gene signature of 1144 probe sets was determined by significance analysis of microarrays (SAM) [72] of the thirteen spindle cell lines versus all epithelial and rounded cell lines, except for HCC1937 and MDA-MB-468, with <1 falsely called positive probe set. The list of differentially expressed genes from this spindle cell gene signature included genes known to be associated with mammary myoepithelial cells and/or the putative CD44+ breast cancer stem cell ([7,8,13-15,24-26,73]; The gene list is provided in Table S2 in Supplemental Data). This transcript-based spindle cell signature was translated into a protein-based signature to allow screening of clinical breast cancers by immunohistochemistry. A 3-protein spindle cell signature of Caldesmon-1, Caveolin-1 and Vimentin (over)expression correctly classified all 25 non-spindle breast cancer cell lines and 75% of 13 spindle cell lines, thus validating the transcript-to-protein translation (Table S3 in Supplemental Data). From the 166 clinical breast cancers, 35 classified with the spindle cell protein signature and these included 33 basal-like and 2 negative group breast cancers (Figures 5 and 6). Importantly, 29 of the 35 spindle cell cases were of metaplastic pathology, and the 16 spindle cell cases with loss of E-cadherin protein expression were all metaplastic breast cancers. We thus have identified metaplastic breast cancers as a second pathological subtype with frequent loss of E-cadherin protein expression that, in contrast to lobular breast cancers, is not characterized by *E-cadherin* gene mutations.

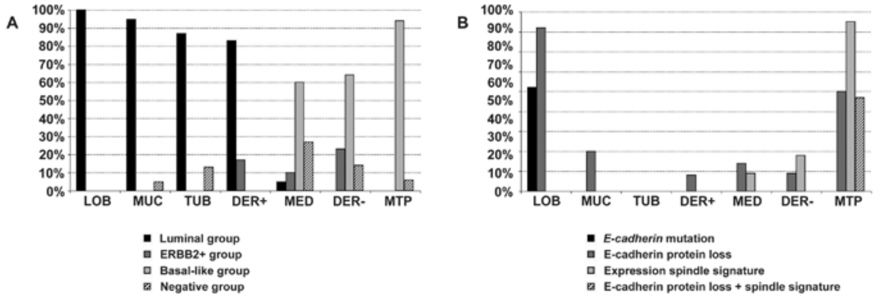


Figure 5. Molecular Characterization of Clinical Breast Cancers. Classification of seven pathological subtypes of clinical breast cancers by (A) the 4-protein signature and (B) *E-cadherin* status and our 3-protein spindle cell signature. Pathological breast cancer subtypes: LOB, lobular; MUC, mucinous; TUB, tubular; DER+, ductal ER-positive; MED, medullary; DER-, ductal ER-negative; MTP, metaplastic. Similar to breast cancer cell lines, the spindle cell signature associated with basal breast cancers, particularly metaplastic breast cancers. Mutational inactivation of *E-cadherin* was detected only in lobular breast cancers.

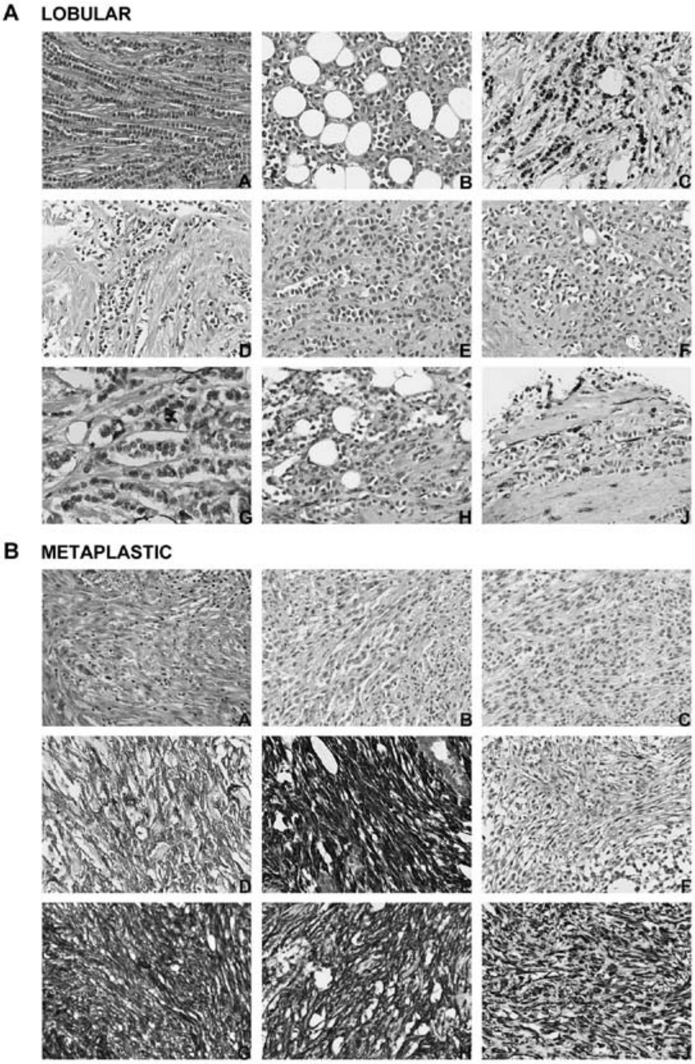


Figure 6. Protein Expression in Clinical Breast Cancers. Examples of immunohistochemical analysis of (A) lobular breast cancers and (B) metaplastic breast cancers. Microscopic views: A, HE-staining; B, ERBB2; C, ER; D, EGFR; E, CK5; F, E-cadherin; G, Caveolin-1; H, Caldesmon; J, Vimentin.

Discussion

Mutational inactivation of *E-cadherin* is distinct from epigenetic silencing

By studying a model of 41 human breast cancer cell lines, we have shown that cell lines with *E-cadherin* gene mutations resemble breast cancers of the luminal subtype, whereas cell lines with epigenetic silencing of *E-cadherin* in association with promoter methylation resemble basal breast cancers. The differences in differentiation status between both subtypes of breast cancer cell lines were reflected by their different morphological appearance (rounded cells versus spindle cells) and by their very distinct intrinsic gene expression program. This duality in E-cadherin loss was also observed in clinical breast cancers. E-cadherin loss of protein expression was frequent among lobular breast cancers and metaplastic breast cancers (of luminal and basal subtype, respectively), but inactivating *E-cadherin* gene mutations were only identified in lobular tumors whereas the 3-protein spindle cell signature was pronounced among metaplastic tumors. In cancer, genetic mutation and epigenetic silencing in association with promoter methylation are widely accepted as two mechanistic means to the same end: inactivation of a tumor suppressor gene. Here we have shown that, at least for *E-cadherin*, this may not always be true.

The identification of two distinct modes of *E-cadherin* inactivation may very well explain recurrent discrepancies in E-cadherin scientific literature. E-cadherin “null” cell lines may not always generate consistent results in functional studies when defined simply by loss of E-cadherin protein expression. E-cadherin’s acclaimed role in cancer invasion, for example, has for long been based on the papers by Frixen *et al.* and Vleminckx *et al.* [45,46] in which cancer invasion was being investigated by experimental manipulation of *E-cadherin* methylated cell lines and spindle-shaped v-*ras* transformed MDCK cells. In contrast, when *E-cadherin* was inactivated by targeted intragenic deletion in mammary epithelium, mice developed invasive breast cancers only upon concurrent *p53* inactivation [74]. Importantly, these *E-cadherin* mutant tumors were of the lobular pathological subtype and thus represent the very first murine model that faithfully replicates human *E-cadherin* mutational inactivation. In humans, it has been shown that *E-cadherin* mutations already arise in premalignant carcinoma *in situ* lesions, also implying that mutational inactivation of *E-cadherin* is an early event that does not yet bring about cancer invasion [75]. Yet, loss of E-cadherin protein expression has been associated with a worse clinical outcome of cancer patients [76]. Although apparently contradictory, all of these observations are likely to be correct. The data reported here imply that the flaw lies in the interpretation of results. It is now imperative to revisit functional E-cadherin studies as well as clinical studies on the prognosis of patients with E-cadherin-negative breast cancers, with careful dissection of the mode of E-cadherin inactivation. Our 3-protein spindle cell signature makes this both feasible and doable.

Breast cancer cell lines are a valid model to study human breast cancer

Cancer cell lines provide the unique opportunity to perform sensitive and in-depth characterizations on an unlimited source of purely tumor cells. For example, expression of E-cadherin transcriptional repressors could be analyzed in detail, revealing widespread expression in spindle cell lines, but unexpectedly, also some expression in non-spindle cell lines. It thus appears that the spindle cell differentiation program involves a rather complex concerted action of E-cadherin repressors. Also, determination of *E-cadherin* promoter methylation is essentially precluded in clinical cancer samples due to the inevitable presence of methylated leukocytes in these specimens [71]. But most important, the breast cancer cell lines allowed us to comprehensively investigate their gene expression program. We identified four of the intrinsic subtypes among the cell lines [24,25,26]. Our results were highly similar to those obtained by Neve *et al.* albeit that they designated the basal-like and normal-like intrinsic subtypes as "basal A" and "basal B" [77]. In fact, our data suggest that their nomenclature may indeed be more appropriate. It should be kept in mind that our breast cancer cell line collection and Gray's partially-overlapping collection both have an overrepresentation of normal-like/basal B cell lines and perhaps also of cell lines with lobular characteristics. Nevertheless, classification of the cell lines according the established intrinsic subtypes of clinical breast cancers provides the most conclusive proof that breast cancer cell lines are indeed a valid model to study human breast cancer.

Epigenetic silencing of *E-cadherin* and its role in EMT is restricted to basal breast cancers, particularly metaplastic breast cancers

Our reconstitution experiments revealed that mutation of the *E-cadherin* gene was causative in the conversion from an epithelial to rounded cell morphology, but epigenetic silencing of *E-cadherin* expression was not causative in spindle cell morphology. This latter observation was rather unexpected since loss of E-cadherin expression is considered the hallmark of epithelial mesenchymal transitions (EMT), i.e. transdifferentiation of epithelial cells into spindle-shaped cells of presumed mesenchymal origin [55]. E-cadherin's role in EMT was further challenged by its expression in five of the thirteen spindle cell lines from our collection, where all E-cadherin-negative cell lines were of the normal-like intrinsic subtype and all basal-like intrinsic subtype cell lines were E-cadherin-positive. Extensive characterization of our collection of breast cancer cell lines revealed that the basal-like and normal-like cell lines do share many characteristics, including their spindle cell morphology and associated gene expression program, the triple-negative phenotype (i.e., ER/PR/ERBB2-negative), widespread expression of E-cadherin transcriptional repressors, and a gene mutation spectrum that includes *p16/ARF* deletion, *RB1* and *BRCA1* mutation and mutational activation of the RAS pathway (see accompanying paper by Nagel *et al.*). Yet, loss of expression of both luminal and basal cytokeratins distinguished normal-like from basal-like breast cancer cell lines, apart from their differential E-cadherin protein expression. It may be that breast cancers of the basal-like subtype arise from luminal breast cancers by an EMT-like transdifferentiation, and that normal-like breast cancers then represent cancers with fully completed EMT. Alternatively, luminal breast cancers may arise from the luminal ductal epithelium in the breast whereas basal-like subtype breast cancers arise from basal ductal epithelium. Normal-like breast cancers then again appear transdifferentiated basal-like breast cancers. Although somewhat irrelevant for this line of thought, it may also be that the phenotypic subtype of breast cancers is dictated by their gene mutation profile in stead of their cellular origin (see accompanying paper by Nagel *et al.*). Either way, breast cancers of the basal-like and normal-like intrinsic subtypes appear to constitute two ends of a spectrum of basal breast cancers – with complete EMT being signified by loss of expression of luminal and basal cytokeratins as well as E-cadherin proteins. Similar to the cell lines, we observed loss of E-cadherin protein expression in a substantial fraction of clinical breast cancers of the basal-like group, particularly among those of the metaplastic pathological subtype (20 of 59 basal-like tumors, of which 16 were metaplastic). And again there was a strong but not exclusive association of the spindle cell signature with E-cadherin-negative basal group breast cancers (Figure 5). It appears that the spindle cell signature also heralds an EMT-like transdifferentiation in clinical breast cancers even before these cancers loose E-cadherin protein expression. It is notable that metaplastic breast cancers as a group are heterogeneous and may include a wide spectrum of histological components, including spindle cells, squamous epithelial cells, matrix-producing cells and cartilage components, all typical characteristics of EMT. The question arises whether all basal breast cancers are susceptible to EMT or whether this is restricted to basal breast cancers with the spindle cell signature. The answer may come from analysis of a larger and consecutive cohort of clinical breast cancers. Yet, the recent identification of CD44+ progenitor-like cells concurrent with more differentiated CD24+ luminal cells in individual breast cancers is noteworthy, particularly so because the CD44+ gene expression profile was highly reminiscent to our spindle cell gene signature [73]. Thus although our data strongly suggest that E-cadherin's role in EMT is restricted to basal breast cancers, there may also be wide-spread transitions involving luminal breast cancer cells.

Experimental procedures

Breast Cancer Samples

The 41 human breast cancer cell lines used in this study are listed in Table 1. Cell lines EVSA-T, MPE600, and SK-BR-5/7 were kind gifts of Dr. N. de Vleesschouwer (Institut Jules Bordet, Brussels, Belgium), Dr. H.S. Smith (California Pacific Medical Center, San Francisco, CA) and Dr. E. Stockert (Sloan-Kettering Institute for Cancer Research, New York, NY), respectively. The SUM cell lines were generated in the Ethier laboratory (available at http://www.asterand.com). Cell line OCUB-F was obtained from Riken Gene Bank (Tsukuba, Japan), and all other cell lines were obtained from ATCC (Manassas, VA). All cell lines were cultured in RPMI supplemented with 10% fetal bovine serum. All cell lines are unique and monoclonal as shown by extensive analysis of nearly 150 polymorphic microsatellite markers [78].

Formalin-fixed paraffin-embedded primary breast cancer specimens were obtained from the Department of Pathology archive at Erasmus MC in Rotterdam, selected from diagnosis years 1990 through 2005. Twenty-four ER-positive and 22 ER-negative ductal breast cancers and 28 lobular, 20 mucinous, 16 tubular and 22 medullary breast cancers were selected. The 34 metaplastic breast cancers were obtained from various hospitals from the Rotterdam/Dordrecht area. All tumor samples were re-evaluated by pathologists with a special interest in breast pathology (MdB and PJW) with metaplastic breast cancers being diagnosed according to criteria based on the WHO classification [22]. Tissue microarrays were constructed by punching three 0.6-mm cores from representative areas of each tumor and transferring them into a recipient paraffin block by using an ATA27 automated tissue microarrayer (Beecher Instruments, Sun Prairie, WI).

The Medical Ethical Committee at Erasmus MC has approved the study, which was carried out according the Code of Conduct of the Federation of Medical Scientific Societies in the Netherlands.

Gene Mutation Analysis

E-cadherin (*CDH1*; Genbank #Z13009) mutations were identified by direct sequencing of PCR-amplified genomic sequences of exons 2 through 16, using an ABI 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA). All mutations were confirmed by sequencing of an independently amplified DNA template. Mutation analysis of 26 breast cancer cell lines had been reported previously [63]. Allelic loss of *E-cadherin* was determined by PCR-based microsatellite analysis using markers *D16S421*, *D16S496*, *D16S2621* and *D16S2624*, where a homozygous allele pattern for all four markers was interpreted as allelic loss, with $P<0.01$ [78].

Promoter Methylation Analysis

For azacytidine assays, exponentially growing cells were treated for three days with 10 µM filter-sterilized 5-aza-2'-deoxycytidine (Sigma, Steinheim, Germany). On the fourth day, cells were washed with PBS at 37°C, harvested by lysis in the flask, and total RNA was isolated. As a control, cultures without 5-aza-2'-deoxycytidine were taken along. *E-cadherin* transcript expression was determined by duplex RT-PCR with the *HPRT* housekeeper.

Methylation-specific PCR (MSP) was performed on bisulphate-treated genomic DNA using the EZ Methylation Kit (Zymed, Orange, CA). MSP for *E-cadherin* CpG islands 1 and 3 was done by using reported primers [69] under our standard PCR conditions [63].

Expression Analysis

Duplex reverse transcriptase (RT)-PCR for *E-cadherin* and *HPRT* was done using the Qiagen (Hilden, Germany) one-step RT-PCR kit and gene-specific exonic primers.

Quantitative (q)RT-PCR was performed on an ABI 7700 Taqman Analyzer (Applied Biosystems), using cDNA templates generated with oligo-dT and random hexamer primers and SYBR Green PCR Master mix (Applied Biosystems). All qRT-PCR Ct-values were normalized according *HPRT* and *PBDG* housekeeper Ct-values. For *SNAIL*, *SLUG*, *SIP-1*, *δEF-1* and *TWIST*, normalized Ct-values <30 were scored expressed (+), Ct-values from 30 through 35 were scored weakly expressed (±), and Ct-values >35 were scored not expressed (-). All primer sequences are provided in Table S4 in the Supplemental Data. Specificity of primers for *SNAIL, SLUG*, *SIP-1*, *δEF-1* and *TWIST* was confirmed by sequencing of the amplification products.

Western blotting was performed as described [79], using the following antibodies: mouse anti-E-cadherin (BD Biosciences (Franklin Lakes, NJ), clone 36 and Zymed, clone HECD-1), mouse anti-β-Catenin (BD Biosciences, clone 14), mouse anti-E2A (Santa Cruz (Santa Cruz, CA), clone Yea), mouse anti-GAPDH (Chemicon (Billerica, MA), clone MAB374) and rabbit-anti-mouse horse radish peroxidase conjugated secondary antibodies (DAKO, Glostrup, Denmark).

Immunohistochemistry was performed by autostaining slides with Chemmate Envision+ kit (DAKO). Slides were first boiled in Tris-EDTA buffer pH=9 (Klinipath, Duiven, The Netherlands) for 20 minutes to unmask antigens. Primary monoclonal antibodies were: Caldesmon-1 (Novocastra (Newcastle upon Tyne, UK), clone TD107), Caveolin-1 (BD Biosciences, clone 2297), CDH1 (DAKO, clone NCH-38), CK5 (Novocastra, clone XM26), CK8-18 (Biogenex (San Ramon, CA), clone NCL5D3), EGFR (DAKO, EGFR pharmDx™ Kit), ER (DAKO, clone 1D5), ERBB2 (DAKO, HercepTest^{TM}), PR (DAKO, clone PgR 636) and Vimentin (DAKO, clone V9).

Transfections

Human *E-cadherin* wild-type and delEx9 cDNA's were RT-PCR amplified from breast cancer cell lines T47D and MPE600. The cDNAs were cloned into pcDNA3.0-Neo expression vector (Invitrogen, Paisley, Scotland) and inserts were verified by sequencing and restriction endonuclease digestions. Transfections were performed using Fugene-6 transfection reagent (Roche, Basel, Switzerland) and cells were grown at limiting dilutions in 96-well plates under G418 selection (Invitrogen). All transfection clones were evaluated for morphological growth pattern and for E-cadherin and β-catenin protein expression, and checked for presence of the correct *E-cadherin* cDNA and genotype of the mother cell line.

Gene Expression Profiling

Breast cancer cell lines were grown to optimal cell densities. The culture medium of the cells was changed 16-20 hours before harvesting (at 37°C). RNA was isolated using the RNeasy kit (Qiagen) upon lysing the cells in the culture flask. DNAseI-treated RNA was antisense biotinylated using the MEGAScript T7 labeling kit (Ambion, Foster City, CA) and Affymetrix U133A microarrays were hybridized according Affymetrix GeneChip Manual, both performed on commercial basis by ServiceXS (Leiden, the Netherlands). Intensity values for all genes were scaled using the global normalization factor as specified by GCOS 1.1, and further normalized with Omniviz software 3.6 (Biowisdom, Maynard, MA). Intensity values <30 were set to 30. Differential gene expression was based on log2 transformed distances to the geometric mean for each probe set. Unsupervised Pearson correlations were performed on <-2 and >2 and <-3 and >3 log2 geometric mean subsets, resulting in 5527 and 2000 probe sets respectively. Patterns of correlation were revealed by applying a matrix-ordering method that sorts samples into correlated blocks, resulting in highly similar plots and identical groupings for log2GM <-2 and >2 or log2GM <-3 and >3 probe subsets.

The Stanford intrinsic gene set for clinical breast cancers [24] was translated into an Affymetrix intrinsic gene set, including 451 probe sets from the Stanford list of 496 genes (Smid *et al.*, manuscript submitted for publication). Breast cancer cell lines were classified for their intrinsic subtype by this Affymetrix intrinsic gene set, using average distance linkage hierarchical clustering with non-centered correlation as distance metric [80].

The transcript-based spindle cell signature was determined by Significance Analysis of Microarrays (SAM; [72]) within Omniviz software package. The criteria in identifying the top 1144 genes with significant differential expression between the spindle and non-spindle breast cancer cell lines were: falsely called median <1, false discovery rate <1, and q-values <1%. This spindle cell signature was validated by qRT-PCR analysis of nine signature genes

(*EMP3, FXYD3, SPDEF, VIM, RAB25, CLDN7, BSPRY, TACSTD1* and *ARHGAP8*). Affymetrix gene expression levels of these genes correlated very well with their qRT-PCR expression levels for all 36 breast cancer cell lines, with average Spearman correlation coefficient of 0.85 (range 0.77-0.91; p<0.001). Primer sequences are provided in Table S4B in the Supplemental Data.

The protein-based spindle cell signature was defined by selection of 16 monoclonal antibodies based on their significance in the transcript-based spindle cell signature, reported relevance for breast cancer and availability. Nine antibodies that proved reliable in immunohistochemistry were evaluated for their ability to distinguish spindle breast cancer cell lines from non-spindle cell lines, demanding 100% specificity for non-spindle cell lines to minimize false positive rates. Five antibodies met these criteria (Table S3 in the Supplemental Data), of which Caveolin-1, Caldesmon-1 and Vimentin were selected for their high specificity for spindle cell lines (100%, 73% and 73%, respectively).

Supplemental Data

The Supplemental Data include four supplemental tables and can be found with this article and on line at http://www-bioinf.erasmusmc.nl/thesis_peeters.

Acknowledgements

We appreciate the technical assistance of Dr. Anieta Sieuwerts, Thierry van de Wetering, and members of the pathology immunohistochemistry labs. We thank Prof. Wolter Oosterhuis and Dr. John Martens for insightful discussions. This work was supported by grants from KWF Dutch Cancer Society and Erasmus MC Mrace.

References

1. Moll, R., et al., *The catalog of human cytokeratins: patterns of expression in normal epithelia, tumors and cultured cells.* Cell, 1982. **31**(1): p. 11-24.
2. Debus, E., K. Weber, and M. Osborn, *Monoclonal cytokeratin antibodies that distinguish simple from stratified squamous epithelia: characterization on human tissues.* Embo J, 1982. **1**(12): p. 1641-7.
3. Nagle, R.B., et al., *Characterization of breast carcinomas by two monoclonal antibodies distinguishing myoepithelial from luminal epithelial cells.* J Histochem Cytochem, 1986. **34**(7): p. 869-81.
4. Smith, G.H., T. Mehrel, and D.R. Roop, *Differential keratin gene expression in developing, differentiating, preneoplastic, and neoplastic mouse mammary epithelium.* Cell Growth Differ, 1990. **1**(4): p. 161-70.
5. Su, L., P.R. Morgan, and E.B. Lane, *Expression of cytokeratin messenger RNA versus protein in the normal mammary gland and in breast cancer.* Hum Pathol, 1996. **27**(8): p. 800-6.
6. Page, M.J., et al., *Proteomic definition of normal human luminal and myoepithelial breast cells purified from reduction mammoplasties.* Proc Natl Acad Sci U S A, 1999. **96**(22): p. 12589-94.
7. Petersen, O.W. and B. van Deurs, *Characterization of epithelial membrane antigen expression in human mammary epithelium by ultrastructural immunoperoxidase cytochemistry.* J Histochem Cytochem, 1986. **34**(6): p. 801-9.
8. Petersen, O.W., P.E. Hoyer, and B. van Deurs, *Frequency and distribution of estrogen receptor-positive cells in normal, nonlactating human breast tissue.* Cancer Res, 1987. **47**(21): p. 5748-51.
9. Gusterson, B.A., et al., *Identification of myoepithelial cells in human and rat breasts by anti-common acute lymphoblastic leukemia antigen antibody A12.* J Natl Cancer Inst, 1986. **77**(2): p. 343-9.
10. Skalli, O., et al., *A monoclonal antibody against alpha-smooth muscle actin: a new probe for smooth muscle differentiation.* J Cell Biol, 1986. **103**(6 Pt 2): p. 2787-96.
11. Zhang, R.R., et al., *A subset of morphologically distinct mammary myoepithelial cells lacks corresponding immunophenotypic markers.* Breast Cancer Res, 2003. **5**(5): p. R151-6.
12. Barbareschi, M., et al., *p63, a p53 homologue, is a selective nuclear marker of myoepithelial cells of the human breast.* Am J Surg Pathol, 2001. **25**(8): p. 1054-60.
13. Coleman, S., G.B. Silberstein, and C.W. Daniel, *Ductal morphogenesis in the mouse mammary gland: evidence supporting a role for epidermal growth factor.* Dev Biol, 1988. **127**(2): p. 304-15.
14. Moller, P., et al., *Expression of epidermal growth factor receptor in benign and malignant primary tumours of the breast.* Virchows Arch A Pathol Anat Histopathol, 1989. **414**(2): p. 157-64.
15. Santini, D., et al., *Immunocytochemical expression of epidermal growth factor receptor in myoepithelial cells of the breast.* Appl Immunohistochem Mol Morphol, 2002. **10**(1): p. 29-33.
16. O'Hare, M.J., et al., *Characterization in vitro of luminal and myoepithelial cells isolated from the human mammary gland by cell sorting.* Differentiation, 1991. **46**(3): p. 209-21.
17. Pechoux, C., et al., *Human mammary luminal epithelial cells contain progenitors to myoepithelial cells.* Dev Biol, 1999. **206**(1): p. 88-99.
18. Bocker, W., et al., *Common adult stem cells in the human breast give rise to glandular and myoepithelial cell lineages: a new cell biological concept.* Lab Invest, 2002. **82**(6): p. 737-46.
19. Petersen, O.W., et al., *Epithelial progenitor cell lines as models of normal breast morphogenesis and neoplasia.* Cell Prolif, 2003. **36 Suppl 1**: p. 33-44.
20. Gudjonsson, T., et al., *Isolation, immortalization, and characterization of a human breast epithelial cell line with stem cell properties.* Genes Dev, 2002. **16**(6): p. 693-706.
21. Boecker, W. and H. Buerger, *Evidence of progenitor cells of glandular and myoepithelial cell lineages in the human adult female breast epithelium: a new progenitor (adult stem) cell concept.* Cell Prolif, 2003. **36 Suppl 1**: p. 73-84.
22. Tavassoli, F.A., et al., *Pathology and genetics of tumours of the breast and female genital organs*. World Health Organization classification of tumours. 2003, LyonOxford: International Agency for Research on Cancerh; Oxford University Press (distributor). 432 p.
23. Rosen, P.P., *Rosen's breast pathology*. 2nd ed. 2001, Philadelphia: Lippincott Williams & Wilkins. xviii, 1004 p.
24. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.
25. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-74.

26. Sorlie, T., et al., *Repeated observation of breast tumor subtypes in independent gene expression data sets.* Proc Natl Acad Sci U S A, 2003. **100**(14): p. 8418-23.

27. Wetzels, R.H., et al., *Basal cell-specific and hyperproliferation-related keratins in human breast cancer.* Am J Pathol, 1991. **138**(3): p. 751-63.

28. Malzahn, K., et al., *Biological and prognostic significance of stratified epithelial cytokeratins in infiltrating ductal breast carcinomas.* Virchows Arch, 1998. **433**(2): p. 119-29.

29. Abd El-Rehim, D.M., et al., *Expression of luminal and basal cytokeratins in human breast carcinoma.* J Pathol, 2004. **203**(2): p. 661-71.

30. Korsching, E., et al., *Cytogenetic alterations and cytokeratin expression patterns in breast cancer: integrating a new model of breast differentiation into cytogenetic pathways of breast carcinogenesis.* Lab Invest, 2002. **82**(11): p. 1525-33.

31. Jones, C., et al., *CGH analysis of ductal carcinoma of the breast with basaloid/myoepithelial cell differentiation.* Br J Cancer, 2001. **85**(3): p. 422-7.

32. Gusterson, B.A., et al., *Basal cytokeratins and their relationship to the cellular origin and functional classification of breast cancer.* Breast Cancer Res, 2005. **7**(4): p. 143-8.

33. Rakha, E.A., et al., *Breast carcinoma with basal differentiation: a proposal for pathology definition based on basal cytokeratin expression.* Histopathology, 2007. **50**(4): p. 434-8.

34. Fulford, L.G., et al., *Specific morphological features predictive for the basal phenotype in grade 3 invasive ductal carcinoma of breast.* Histopathology, 2006. **49**(1): p. 22-34.

35. Nielsen, T.O., et al., *Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma.* Clin Cancer Res, 2004. **10**(16): p. 5367-74.

36. Takeichi, M., *Cadherin cell adhesion receptors as a morphogenetic regulator.* Science, 1991. **251**(5000): p. 1451-5.

37. Kemler, R., *From cadherins to catenins: cytoplasmic protein interactions and regulation of cell adhesion.* Trends Genet, 1993. **9**(9): p. 317-21.

38. Cowin, P., *Unraveling the cytoplasmic interactions of the cadherin superfamily.* Proc Natl Acad Sci U S A, 1994. **91**(23): p. 10759-61.

39. Gumbiner, B.M., *Cell adhesion: the molecular basis of tissue architecture and morphogenesis.* Cell, 1996. **84**(3): p. 345-57.

40. Semb, H. and G. Christofori, *The tumor-suppressor function of E-cadherin.* Am J Hum Genet, 1998. **63**(6): p. 1588-93.

41. Berx, G. and F. Van Roy, *The E-cadherin/catenin complex: an important gatekeeper in breast cancer tumorigenesis and malignant progression.* Breast Cancer Res, 2001. **3**(5): p. 289-93.

42. Jamora, C. and E. Fuchs, *Intercellular adhesion, signalling and the cytoskeleton.* Nat Cell Biol, 2002. **4**(4): p. E101-8.

43. Yamada, S., et al., *Deconstructing the cadherin-catenin-actin complex.* Cell, 2005. **123**(5): p. 889-901.

44. Drees, F., et al., *Alpha-catenin is a molecular switch that binds E-cadherin-beta-catenin and regulates actin-filament assembly.* Cell, 2005. **123**(5): p. 903-15.

45. Frixen, U.H., et al., *E-cadherin-mediated cell-cell adhesion prevents invasiveness of human carcinoma cells.* J Cell Biol, 1991. **113**(1): p. 173-85.

46. Vleminckx, K., et al., *Genetic manipulation of E-cadherin expression by epithelial tumor cells reveals an invasion suppressor role.* Cell, 1991. **66**(1): p. 107-19.

47. Perl, A.K., et al., *A causal role for E-cadherin in the transition from adenoma to carcinoma.* Nature, 1998. **392**(6672): p. 190-3.

48. Berx, G., et al., *E-cadherin is a tumour/invasion suppressor gene mutated in human lobular breast cancers.* Embo J, 1995. **14**(24): p. 6107-15.

49. Berx, G., et al., *E-cadherin is inactivated in a majority of invasive human lobular breast cancers by truncation mutations throughout its extracellular domain.* Oncogene, 1996. **13**(9): p. 1919-25.

50. Becker, K.F., et al., *Exon skipping in the E-cadherin gene transcript in metastatic human gastric carcinomas.* Hum Mol Genet, 1993. **2**(6): p. 803-4.

51. Becker, K.F., et al., *E-cadherin gene mutations provide clues to diffuse type gastric carcinomas.* Cancer Res, 1994. **54**(14): p. 3845-52.

52. Yoshiura, K., et al., *Silencing of the E-cadherin invasion-suppressor gene by CpG methylation in human carcinomas.* Proc Natl Acad Sci U S A, 1995. **92**(16): p. 7416-9.
53. Graff, J.R., et al., *E-cadherin expression is silenced by DNA hypermethylation in human breast and prostate carcinomas.* Cancer Res, 1995. **55**(22): p. 5195-9.
54. Graff, J.R., et al., *Methylation patterns of the E-cadherin 5' CpG island are unstable and reflect the dynamic, heterogeneous loss of E-cadherin expression during metastatic progression.* J Biol Chem, 2000. **275**(4): p. 2727-32.
55. Peinado, H., D. Olmeda, and A. Cano, *Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype?* Nat Rev Cancer, 2007. **7**(6): p. 415-28.
56. Batlle, E., et al., *The transcription factor snail is a repressor of E-cadherin gene expression in epithelial tumour cells.* Nat Cell Biol, 2000. **2**(2): p. 84-9.
57. Cano, A., et al., *The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression.* Nat Cell Biol, 2000. **2**(2): p. 76-83.
58. Hajra, K.M., D.Y. Chen, and E.R. Fearon, *The SLUG zinc-finger protein represses E-cadherin in breast cancer.* Cancer Res, 2002. **62**(6): p. 1613-8.
59. Comijn, J., et al., *The two-handed E box binding zinc finger protein SIP1 downregulates E-cadherin and induces invasion.* Mol Cell, 2001. **7**(6): p. 1267-78.
60. Eger, A., et al., *DeltaEF1 is a transcriptional repressor of E-cadherin and regulates epithelial plasticity in breast cancer cells.* Oncogene, 2005. **24**(14): p. 2375-85.
61. Perez-Moreno, M.A., et al., *A new role for E12/E47 in the repression of E-cadherin expression and epithelial-mesenchymal transitions.* J Biol Chem, 2001. **276**(29): p. 27424-31.
62. Yang, J., et al., *Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis.* Cell, 2004. **117**(7): p. 927-39.
63. van de Wetering, M., et al., *Mutant E-cadherin breast cancer cells do not display constitutive Wnt signaling.* Cancer Res, 2001. **61**(1): p. 278-84.
64. Engel, L.W. and N.A. Young, *Human breast carcinoma cells in continuous culture: a review.* Cancer Res, 1978. **38**(11 Pt 2): p. 4327-39.
65. Cailleau, R., et al., *Breast tumor cell lines from pleural effusions.* J Natl Cancer Inst, 1974. **53**(3): p. 661-74.
66. Cailleau, R., M. Olive, and Q.V. Cruciger, *Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization.* In Vitro, 1978. **14**(11): p. 911-5.
67. Schlosshauer, P.W., et al., *APC truncation and increased beta-catenin levels in a human breast cancer cell line.* Carcinogenesis, 2000. **21**(7): p. 1453-6.
68. Lombaerts, M., et al., *E-cadherin transcriptional downregulation by promoter methylation but not mutation is related to epithelial-to-mesenchymal transition in breast cancer cell lines.* Br J Cancer, 2006. **94**(5): p. 661-71.
69. Graff, J.R., et al., *Mapping patterns of CpG island methylation in normal and neoplastic cells implicates both upstream and downstream regions in de novo methylation.* J Biol Chem, 1997. **272**(35): p. 22322-9.
70. Hiraguri, S., et al., *Mechanisms of inactivation of E-cadherin in breast cancer cell lines.* Cancer Res, 1998. **58**(9): p. 1972-7.
71. Lombaerts, M., et al., *Infiltrating leukocytes confound the detection of E-cadherin promoter methylation in tumors.* Biochem Biophys Res Commun, 2004. **319**(2): p. 697-704.
72. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
73. Shipitsin, M., et al., *Molecular definition of breast tumor heterogeneity.* Cancer Cell, 2007. **11**(3): p. 259-73.
74. Derksen, P.W., et al., *Somatic inactivation of E-cadherin and p53 in mice leads to metastatic lobular mammary carcinoma through induction of anoikis resistance and angiogenesis.* Cancer Cell, 2006. **10**(5): p. 437-49.
75. Vos, C.B., et al., *E-cadherin inactivation in lobular carcinoma in situ of the breast: an early event in tumorigenesis.* Br J Cancer, 1997. **76**(9): p. 1131-3.
76. Siitonen, S.M., et al., *Reduced E-cadherin expression is associated with invasiveness and unfavorable prognosis in breast cancer.* Am J Clin Pathol, 1996. **105**(4): p. 394-402.
77. Neve, R.M., et al., *A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes.* Cancer Cell, 2006. **10**(6): p. 515-27.
78. Harkes, I.C., et al., *Allelotype of 28 human breast cancer cell lines and xenografts.* Br J Cancer, 2003. **89**(12): p. 2289-92.

79. Wasielewski, M., et al., *Thirteen new p53 gene mutants identified among 41 human breast cancer cell lines.* Breast Cancer Res Treat, 2006. **99**(1): p. 97-101.

80. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.

81. Antonarakis, S.E., *Recommendations for a nomenclature system for human gene mutations. Nomenclature Working Group.* Hum Mutat, 1998. **11**(1): p. 1-3.

82. den Dunnen, J.T. and S.E. Antonarakis, *Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion.* Hum Mutat, 2000. **15**(1): p. 7-12.

83. Becker, K.F., et al., *Identification of eleven novel tumor-associated E-cadherin mutations. Mutations in brief no. 215. Online.* Hum Mutat, 1999. **13**(2): p. 171.

84. Risinger, J.I., et al., *Mutations of the E-cadherin gene in human gynecologic cancers.* Nat Genet, 1994. **7**(1): p. 98-102.

Table S1. ***E-cadherin*** **Gene Mutations Identified in Breast Cancer Samples**

A. *E-cadherin* mutations identified in 10 of 41 breast cancer cell lines

Breast cancer cell line	Location of alteration	E-cadherin gene sequence	E-cadherin transcript sequence	Predicted protein effect	Type of mutation
SK-BR-3	Exon 1 - 12	c.1_1936del1936	not detectable	not expressed	deletion exons 1-12 (NI)
OCUB-F/-M	Exon 2	c.49_163del115	r.49_163del115	p.V17X	deletion exon 2 (PT)
SK-BR-5	IVS4	c.532-1G>C	r.532_547del16	p.I178TfsX32	deletion 1st 16 bases exon 5 (PT)
EVSA-T	IVS5	c.687+1delGT	r.646_687del42	p.V216_T229del	deletion last 42 bases exon 5 (IF)
MDA-MB-134VI	Exon 6	c.688_832del145	r.688_832del145	p.L230EfsX4	deletion exon 6 (PT)
ZR75-30	Exon 6	c.727G>T	r.727G>U	p.E243X	nonsense (PT)
MPE600	Exon 9	c.1138-21_1138del22	r.1138_1320del183	p.Y380_K440del	deletion exon 9 (IF)
SUM44PE	Exon 9	c.1269delT	r.1269delU	p.F423LfsX8	deletion of 1 base (PT)
CAMA-1	IVS11	c.1712-1G>A	r.1566_1712del147	p.Y523_G571del	deletion exon 11 + 1st base exon 12 (IF)
MDA-MB-453	Exon 12	c.1913G>A	r.1913G>A	p.W638X	nonsense (PT)

B. *E-cadherin* mutations identified in 6 of 14 clinical breast cancers of lobular pathology

Breast tumor	Subtype	Location of alteration	E-cadherin gene sequence	Predicted protein effect	Type of mutation
1	Lobular	Exon 2	c.55_67del13	p.S19RfsX33	deletion of 13 bases (PT)
2	Lobular	Exon 2	c.115_116insGTAGT	p.F39CfsX19	insertion of 5 bases (PT)
3	Lobular	Exon 2	c.152_163+11del23	unclear	deletion of 23 bases including splice site (PT?)
4	Lobular	Exon 3	c.229_241del13	p.G78X	deletion of 13 bases (PT)
5	Lobular	Exon 3	c.377_378insC	p.P127AfsX41	insertion of 1 base (PT)
6	Lobular	Exon 5	c.595_596insA	p.T199NfsX10	insertion of 1 base (PT)
7	Lobular	Exon 7	c.897_898ins26	p.I299MfsX3	insertion of 26 bases (PT)
8	Lobular	Exon 7	c.972_973insA	p.V325SfsX25	insertion of 1 base (PT)
9	Lobular	Exon 15	c.2398delC	p.R800AfsX16	deletion of 1 base (PT)
10	Lobular	Exon 15	c.2398delC	p.R800AfsX16	deletion of 1 base (PT)

E-cadherin / *CDH1* gene mutations are numbered according Genbank accession Z13009 and nomenclature is according Nomenclature Working Group recommendations and http://www.hgvs.org/mutnomen/index.html). *E-cadherin* gene sequences are determined by sequencing of PCR-amplified genomic templates. For the breast cancer cell lines, transcripts sequences were determined by sequencing of RT-PCR-amplified cDNA templates and protein expression was detected for all three in-frame mutants but not for the other mutants. The clinical breast cancers were selected for loss of E-cadherin protein expression and >90% of the coding sequence was analyzed for alterations. Mutations in OCUB-F, MPE600 and breast tumor 5 were previously identified in clinical cancer specimens. NI, no initiation site for translation; PT, premature termination of translation; IF, in-frame deletion from the transcript; PT?, likely premature termination of translation as E-cadherin protein expression is lost.

Table S2: See website http://www-bioinf.erasmusmc.nl/thesis_peeters

Table S3: Validation of Spindle Cell Signature in Human Breast Cancer Cell Line**s.** A tissue microarray of the breast cancer cell lines was used to validate nine spindle cell signature proteins by immunohistochemistry. The antibodies were required to accurately classify all of 25 non-spindle cell lines and as much as possible of the 13 spindle cell lines. The three best performing antibodies were selected for the 3-protein spindle cell signature (CALD1, CAV1 and VIM).

Gene	Antibody	% spindle cell lines correctly classifed	% non-spindle cell lines correctly classifed
CAV1	Clone 2297, BD Transduction Laboratories	100%	100%
VIM	Clone V9, DAKO	73%	100%
CALD1	Clone TD107, Novocastra	73%	100%
CDH2	Clone 6G11, DAKO	64%	100%
HAI-1	Clone 169417, R&D systems	64%	100%
PML	Clone PG-M3, Santa Cruz Biotechnology	73%	96%
JUP	Clone 15, BD Transduction Laboratories	45%	92%
EGFR	EGFR pharmDx™ Kit, DAKO	91%	88%
CK19	Clone RCK108, DAKO	64%	88%

Table S4: Primer Sequences

A. Primers for duplex RT-PCR

Gene	Primer	Sequence 5’ -> 3’
E-cadherin	R-F1	TTGCGGAAGTCAGTTCAGAC
	R-MPE-R1	CAAAGTCCTCGGACACTTC
	R-MPE-F1	ACAGGAACACAGGAGTCATC
	R-R1	GCAACGTGATTTCTGCATTTC
HPRT	243b	GTGGGGTCCTTTTCACCAG
	244b	TATGGACAGGACTGAACGTC

B. Primers for qRT-PCR

Gene	Primer	Sequence 5’ -> 3’
HPRT	Fh1	TATTGTAATGACCAGTCAACAG
	1	GGTCCTTTTCACCAGCAAG
PBDG	F	CATGTCTGGTAACGGCAATG
	R	GTACGAGGCTTTCAATGTTG
SNAIL	F1	CGAGCTGCAGGACTCTAATC
	R1	TGGGGCGCCAGGACAGAG
SLUG	F1	AGCGAACTGGACACACATAC
	R1	GGTAGCTGGGCGTGGAATG
SIP-1	F2	AACACCCCTGGCACAACAAC
	R2	CTCCAGTTTTCTTTTGGCAAAG
δEF-1	F1	AAGAGAAGGGAATGCTAAGAAC
	R1	TGTTGTTGTAGAAACTCTTCAAC
TWIST	FA	GGACAAGCTGAGCAAGATTC
	RA	TTATCCAGCTCCAGAGTCTC
EMP3	F1	ACGAGGAGGTCTCTTCTATG
	R1	GCCAGGGCGAAGCAGTATC
FXYD3	F1	AGCGCTCTGACATGCAGAAG
	R1	ACAGAACCCCAGCGCAGATG
PDEF	F4	GCCCCTGGGTGGGGATG
	R2	TCCTTGTTGAGCCACCTAATG
RAB25	F2	GCCAAGATGGGGAATGGAAC
	R2	TGTCGTGGCTGAACTCATTG
B-SPRY	F2	TCTGGGCAACGGCGGTTC
	R2	GTGGCAGCCAGGGCATTG
TACSTD1	F2	GGGGCCCTCCAGAACAATG
	R2	CAATGATGATCCAGTAGGTTC
Claudin-7	F1	CCATGTTTGTGGCCACGATG
	R1	GCCAAACTCATACTTAATGTTG
ARHGAP8	F2	CCGTGAACTTTGACGACTAC
	R2	CTCTGGGAGGCTCCGTAAG
Vimentin	FX	CAGATTCAGGAACAGCATGTC
	RX	TCCAGCCTTTCCAGGAACA

Chapter 5

Gene expression profiling assigns *CHEK2* 1100delC breast cancers to the luminal intrinsic subtypes

***Justine K. Peeters*[1*]**; Jord H.A. Nagel[2*]; Marcel Smid[2]; Anieta M. Sieuwerts[2]; Marijke Wasielewski[2]; Vanja de Weerd[2]; Anita M.A.C. Trapman-Jansen[2]; Ans van den Ouweland[3]; Henk Portengen[2]; Hennie Brüggenwirth[3]; Wilfred van IJcken[4]; Jan G.M. Klijn[2]; Peter J. van der Spek[1]; John A. Foekens[2]; John W.M. Martens[2]; Mieke Schutte[2]; and Hanne Meijers-Heijboer[3,2,5]

[1] Department of Bioinformatics,

[2] Department of Medical Oncology,

[3] Department of Clinical Genetics, and

[4] Erasmus Center for Biomics, Erasmus University Medical Center, Rotterdam, The Netherlands; and

[5] Department of Clinical Genetics, VU Medical Center, Amsterdam, The Netherlands

[*] These authors contributed equally to this work.

Submitted to Cancer Cell

Abstract

CHEK2 1100delC is a moderate-risk cancer susceptibility allele that confers a high breast cancer risk in a polygenic setting. Gene expression profiling of *CHEK2* 1100delC breast cancers may reveal clues to the nature of the polygenic *CHEK2* model and its genes involved. Here, we report global gene expression profiles of a cohort of 155 familial breast cancers, including 26 *CHEK2* 1100delC mutant tumors. A 40-gene *CHEK2* signature was defined that significantly associated with *CHEK2* 1100delC breast cancers. The identification of a *CHEK2* gene signature implies an unexpected biological homogeneity among the *CHEK2* 1100delC breast cancers. In addition, all 26 *CHEK2* 1100delC tumors classified as luminal intrinsic subtype breast cancers, with 8 luminal A and 18 luminal B tumors. This biological make-up of among *CHEK2* 1100delC breast cancers suggests that a relatively limited number of additional susceptibility alleles are involved in the polygenic *CHEK2* model. Identification of these as-yet-unknown susceptibility alleles should be aided by clues from the 40-gene *CHEK2* signature.

Introduction

At least ten percent of breast cancers arise within a familial clustering of multiple breast cancers. Inherited mutations of the *BRCA1* or *BRCA2* genes are identified in about one-quarter of the families with breast cancer (OMIM 113705 and 600185) [1-5]. Female carriers of mutant *BRCA1* or *BRCA2* genes have a lifetime risk of 50 to 85% to develop breast cancer, classifying both genes as high-risk breast cancer susceptibility genes. In 2002, we and others identified the *CHEK2* gene as the third breast cancer susceptibility gene (OMIM 604373) [5-9]. A single founder mutation, *CHEK2* 1100delC, was identified in about 5% of families with breast cancer that did not have mutations in either *BRCA1* or *BRCA2*. In contrast to *BRCA1* and *BRCA2*, *CHEK2* 1100delC was estimated to confer only a moderate 20 to 25% risk to develop breast cancer for female mutation carriers. Although this classified *CHEK2* 1100delC as a moderate-risk breast cancer susceptibility allele, the mutation was found to be particularly prevalent among families with a high-risk breast cancer inheritance pattern, with mutation frequencies rising to over 20% among families with four or more cases of breast cancer [5-7,9,10]. Segregation of the *CHEK2* 1100delC mutation with the cancer phenotype typically was incomplete in the high-risk breast cancer families, suggesting the inheritance of an additional breast cancer susceptibility allele or alleles in these families. Independent investigations indeed have implied that a vast amount of non-*BRCA1*/*BRCA2* familial breast cancers likely arise within a context of polygenic breast cancer susceptibility, where multiple moderate-risk or low-risk susceptibility alleles act in concert to confer a high risk to develop breast cancer [11,12]. By now, several other moderate-risk breast cancer genes have been identified, including the *ATM*, *BRIP1*, *PALB2* and *CASP8* genes, and recently also six low-risk loci, and each of them appeared to operate in a polygenic setting [13-17]. An intriguing question is whether these moderate-risk and low-risk susceptibility genes operate in a polygenic setting wherein each risk allele may act in concert with any other risk allele, or are there also risk alleles that are less promiscuous and operate with a limited set of risk alleles? The issue certainly is not trivial, as polygenic breast cancers likely would be far more biologically homogeneous in the latter setting and thus be anticipated to have a more predictable clinical outcome.

Historically, breast cancers had been classified by their expression of the estrogen and progesterone hormone receptors (ER and PGR) and the ERBB2/HER2/NEU receptor. An important breakthrough in breast cancer classification came with the advent of microarray technology, allowing genome wide expression analysis of a tumor sample. Seminal gene expression profiling studies by Sørlie, Perou and their colleagues have revealed that breast cancers might be classified by their global gene expression program, distinguishing two subsets of breast cancers among ER-positive tumors (luminal A and B), two subsets among ER-negative tumors (basal-like and normal-like) and the ERBB2 subset being mainly ER negative [18-20]. These intrinsic subtypes were shown to be relevant in prognosis and prediction of clinical outcome of breast cancer patients [20,21], although not as powerful as gene signatures that had been defined based on prognosis or therapy responses of patients [22-34]. Gene expression profiles have also been associated with genetic alterations present in breast cancers, including alterations of *BRCA1*, *ERBB2*, *p53* and *MYC* 20,35-38. Classification of *BRCA1* breast cancers as being predominantly of the basal-like intrinsic subtype [20] was particularly instrumental because it implied biological homogeneity among *BRCA1* tumors.

Here, we have generated global gene expression profiles of a cohort of 155 familial breast cancers, including *BRCA1*, *BRCA2* and *CHEK2* mutant tumors. We specifically investigated whether or not *CHEK2* 1100delC breast cancers are biologically homogeneous as this may provide clues to the nature of polygenic *CHEK2* cancer susceptibility and its genes involved.

Results

Definition of a 40-Gene Signature for *CHEK2* 1100delC Breast Cancers

Global gene expression profiles of 155 fresh-frozen primary familial breast cancers were generated by using Affymetrix U133 Plus 2.0 GeneChips. Familial breast cancer cases were classified by the presence of an oncogenic germline *BRCA1* or *BRCA2* mutation (n=47 and n=6), by the presence of the *CHEK2* 1100delC founder mutation (n=26), or by a family history of breast cancer when no mutations were detected in either gene (n=76). Unsupervised clustering of the 155 tumors by Pearson correlation based on the top 10% variably-expressed probe sets (n=5,467) revealed two highly distinct clusters of tumor samples, designated "hormone receptor-positive" and "hormone receptor-negative" (HR-pos and HR-neg; Figure 1; Reproducibility measure R=0.95). Based on microarray transcript expression levels, 96% of the 100 tumors in the HR-pos cluster were considered positive for expression of Estrogen Receptor alpha (*ESR1*) compared with 9% of the 55 tumors in the HR-neg cluster (Fisher's exact P<0.0001 and *ESR1* cut-off 1,000; Figure 1, Table 1, and Supporting Information (SI) Table S1), strongly suggesting that the molecular dichotomy among the breast cancers was related to their ER status. In concordance with this notion and with literature reports, univariate t-testing revealed that the differential gene expression programs between the two tumor clusters included not only *ESR1* but also its downstream target genes, such as *PGR*, *FOXOA1*, *GATA3*, *TFF3*, *NAT1* and *XBP*. Remarkably, all 26 *CHEK2* 1100delC tumors and all six *BRCA2* tumors located in the HR-pos cluster, whereas 36 of the 47 *BRCA1* tumors located in the HR-neg cluster (Figure 1 and Table 1). This cluster division among mutation-positive familial breast cancers was in concordance with their ER status, known to be predominantly ER-positive for *CHEK2* 1100delC and *BRCA2* tumors and predominantly ER-negative for *BRCA1* tumors [39-42].

To determine the *CHEK2* 1100delC gene expression signature, we performed supervised class comparison of the 26 *CHEK2* 1100delC tumors with all 129 non-*CHEK2* tumors from the familial breast cancer cohort. We have evaluated several class comparison and class prediction methods but the generated gene signatures were all strongly dominated by *ESR1* response genes. The assignment of all 26 *CHEK2* 1100delC breast cancers to the HR-pos tumor cluster suggested molecular homogeneity among these tumors, albeit that it was unclear whether this homogeneity reflected their ER-positive hormone receptor status, their *CHEK2* mutation status, or both. We therefore restricted the supervised analysis to the 100 HR-pos familial breast cancers from the cohort. Univariate t-test comparison of the 26 *CHEK2* 1100delC tumors with the 74 non-*CHEK2* HR-pos breast cancers now allowed identification of a *CHEK2* signature of 40 differentially expressed genes, represented by 43 probe sets (P<0.001 and global test P=0.03; SI Table S1). Unsupervised hierarchical clustering of all 100 HR-pos breast cancers based on

Table 1: Receptor Expression Among Familial Breast Cancers. Receptor transcript expression cut-offs were 1,000 for *ESR1* (probe set 205225_at), 20 for *PGR* (probe set 208305_at) and 3,700 for *ERBB2* (probe set 216836_s_at).

Tumors by mutation status	Hormone receptor-positive tumors							Hormone receptor-negative tumors						
	No.	ER		PGR		ERBB2		No.	ER		PGR		ERBB2	
		+ve	(%)	+ve	(%)	+ve	(%)		+ve	(%)	+ve	(%)	+ve	(%)
CHEK2 1100delC	26	26	(100)	23	(88)	16	(62)	0	-	-	-	-	-	-
BRCA1	9	8	(89)	7	(78)	4	(44)	38	3	(8)	14	(37)	3	(8)
BRCA2	6	6	(100)	5	(83)	2	(33)	0	-	-	-	-	-	-
Non-mutant	59	56	(95)	53	(90)	23	(39)	17	2	(12)	5	(29)	2	(12)
All tumors	100	96	(96)	88	(88)	45	(45)	55	5	(9)	19	(35)	5	(9)

Table 2: Gene Expression Among Familial Breast Cancers. Gene transcript expression cut-offs were <-1.0 and >+1.0 of the log-2 geometric mean expression level among all 155 familial breast cancers.

Tumors by		*CHEK2*		*BRCA1*		*BRCA2*		*p53*	
mutation status	No.	<-1.0 (%)	>+1.0 (%)	<-1.0 (%)	>+1.0 (%)	<-1.0 (%)	>+1.0 (%)	<-1.0 (%)	>+1.0 (%)
CHEK2 1100delC	26	16 (61)	0	2 (8)	4 (15)	4 (15)	1 (4)	2 (8)	2 (8)
BRCA1	47	0	8 (17)	8 (17)	2 (4)	2 (4)	14 (30)	16 (34)	9 (19)
BRCA2	6	0	1 (17)	0	4 (66)	0	0	3 (50)	0
Non-mutant	76	7 (9)	8 (10)	8 (10)	6 (8)	14 (18)	6 (8)	4 (5)	9 (11)
All tumors	155	23 (15)	17 (11)	18 (12)	16 (10)	20 (13)	21 (13)	25 (16)	20 (13)

the 40-gene *CHEK2* signature correctly assigned 23 of 26 *CHEK2* 1100delC breast cancers to a single cluster branch, reflecting a sensitivity of 88% (Figure 2A). Clustering of the HR-pos breast cancers based on the *CHEK2* signature was highly significant and reproducible (Fisher's exact P<0.0001 and Reproducibility measure R=0.73). The robustness of the *CHEK2* signature was also evaluated by simulating signature identification based on 26 randomly selected HR-pos breast cancers instead of *CHEK2* 1100delC tumors. Three simulation experiments revealed gene signatures of 9, 5 and 3 differentially expressed genes, and none of the three gene signatures were significant by Fisher's exact testing. Taken together, these analyses indicate that the 40-gene *CHEK2* signature significantly associated with *CHEK2* 1100delC breast cancers.

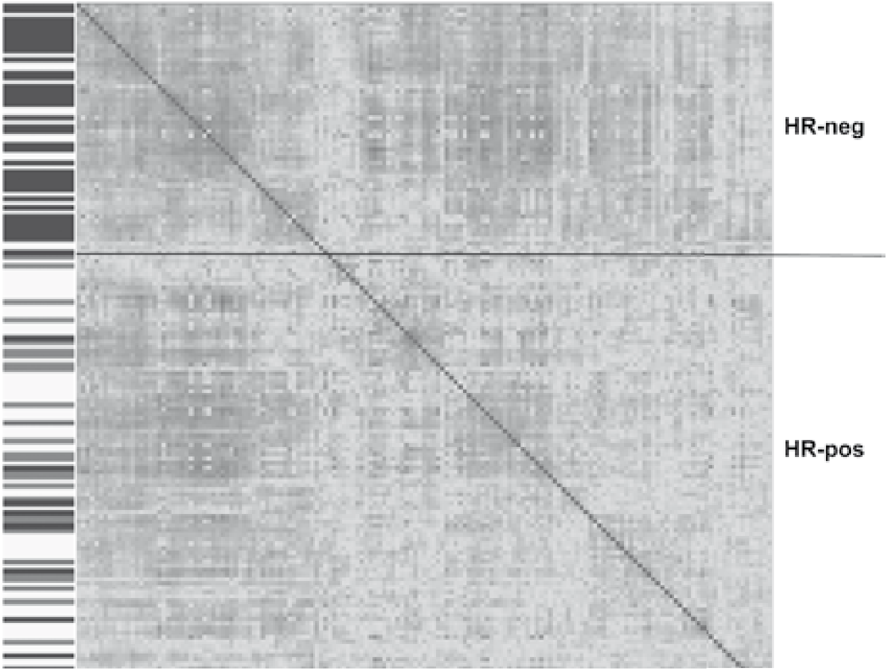


Figure 1: Unsupervised Pearson Correlation Matrix of 155 Familial Breast Cancers. The correlation visualization tool displays pair wise correlations between tumors numbered 1 through 155, where red cells indicate positive correlation and blue cells indicate negative correlation. The matrix revealed two main clusters of breast cancers, containing 100 and 55 tumor samples and designated HR-pos and HR-neg, respectively. Color-coding mutation status: Red, *CHEK2* 1100delC tumors; Blue, *BRCA1* tumors; Green, *BRCA2* tumors; Yellow, Non-mutant tumors. Color-coding *ESR1* transcript expression: Red, positive; Blue, negative; based on microarray data with a cut-off of 1,000 for *ESR1* (probe set 205225_at).

Unsupervised hierarchical clustering of the 100 HR-pos familial breast cancers based on the 40-gene *CHEK2* signature assigned six of the nine HR-pos *BRCA1* breast cancers to the *CHEK2* tumor cluster (Figure 2A). This was somewhat unexpected because the *CHEK2* signature had been defined by comparison to all non-*CHEK2* HR-pos breast cancers – including the nine *BRCA1* breast cancers – and might suggest a functional relationship between the BRCA1 and CHEK2 proteins. To further evaluate this putative relationship, we also defined a gene signature by t-test comparison of the 26 *CHEK2* 1100delC tumors with all HR-pos breast cancers but with exclusion of the nine HR-pos *BRCA1* breast cancers. Interestingly, the *CHEK2*-minus-*BRCA1* signature included considerably more differentially expressed genes than the *CHEK2* signature (69 versus 40 genes with an overlap of 37 genes; global test P=0.008; SI Table S2), implying that the *CHEK2*-minus-*BRCA1* classification yielded biologically more homogeneous tumor clusters. After all, the more distinct two sample groups, the more genes that are expressed exclusively in either sample group, while increasing heterogeneity within a sample group diminishes the number of genes whose expression typifies that group. The biological homogeneity was also reflected by the improved performance of the *CHEK2*-minus-*BRCA1* signature over the *CHEK2* signature in clustering the 100 HR-pos breast cancers, with two more *CHEK2* 1100delC breast cancers and two more *BRCA1* breast cancers that located to the *CHEK2* tumor cluster (Figure 2A and 2B; R=0.73). Although these results suggest shared oncogenic functions for the two breast cancer genes, our experiments are not conclusive and a larger HR-pos *BRCA1* breast cancer cohort is required to validate the putative functional link between CHEK2 and *BRCA1*. In anticipation thereof, we have focused further analyses to the 40-gene *CHEK2* signature.

Evaluation of the 40 genes from the *CHEK2* signature by Ingenuity pathway analysis revealed that the *CHEK2* gene itself was the top most differentially expressed gene (SI Table S2). Indeed, 16 of 23 breast cancers with *CHEK2* transcript expression levels below half the average expression level also carried the *CHEK2* 1100delC mutation (Table 2). The most prominent gene from the *CHEK2* signature known to be involved in CHEK2 functions was *RECQ5L*, a member of the RecQ family of DNA helicases that also includes *RECQL*, *RECQL4*, *BLM* and *WRN* [43]. The RecQ helicases are involved in processing of aberrant DNA structures that arise during DNA replication and repair, where RECQL5's function is thought to be in mitotic recombination events. Importantly, germline mutations in the *RECQL4*, *BLM* and *WRN* genes each give rise to ageing disorders with an increased cancer risk: Rothmund-Thomson, Bloom and Werner syndromes, respectively [44-46]. Other genes from the *CHEK2* signature had been associated with BRCA1's function in maintenance of a normal, inactive X chromosome, including the non-coding *XIST* gene and the polycomb group family member *EED* [47-50]. Ingenuity pathway analysis of the 40 genes from the *CHEK2* signature also identified "Cell cycle

G2/M DNA damage checkpoint regulation" as the most prominent canonical pathway that associated with the signature, consistent with the well-defined kinase function of CHEK2 in the G2 phase of the cell cycle [51,52]. "Fibroblast Growth Factor signaling" and "p53 signaling" also featured as top biological networks associated with the *CHEK2* signature, which again was not surprising because FGF signaling has previously been associated with breast cancer susceptibility [17,53-55] whereas the p53 protein is a well known downstream phosphorylation target of *CHEK2* kinase [51,52]. The concordance of the functional assignments of the *CHEK2* signature genes with current views on *CHEK2* kinase function and its suggested interaction with *BRCA1* and p53 further supports the validity of the 40-gene *CHEK2* signature.

CHEK2 1100delC Tumors Are Luminal Subtype Breast Cancers

Unsupervised hierarchical clustering of the 100 HR-pos familial breast cancers based on the top 10% variably-expressed probe sets divided the *CHEK2* 1100delC tumors over two different tumor clusters, with 17 *CHEK2* 1100delC tumors in one cluster and 9 in the other cluster (Figure 2C). A dichotomy among hormone receptor-positive breast cancers had also been observed by Sørlie *et al.* [18-2]0. Therefore, we classified all 155 familial breast cancers from our cohort according the intrinsic gene signatures defined by Sørlie *et al.* Of the 496 genes from their intrinsic gene signatures, 451 could be mapped to the Affymetrix GeneChips. Hierarchical clustering based on these genes identified 32 (21%) luminal A subtype tumors, 66 (42%) luminal B and 57 (37%) basal-like subtype tumors among the 155 familial breast cancers and, notably, no ERBB2 or normal-like breast cancers (Table 3). In concordance with previous reports, all 38 HR-neg *BRCA1* tumors and a single HR-pos *BRCA1* tumor classified as basal-like subtype breast cancers [20]. Importantly, the *CHEK2* 1100delC tumors were all classified as luminal subtype breast cancers (with 8 luminal A and 18 luminal B tumors), suggesting considerable biological homogeneity among *CHEK*2 1100delC breast cancers.

Table 3: Intrinsic Subtypes Among Familial Breast Cancers. Familial breast cancers had been clustered based on 451 unique Affymetrix probe sets that mapped to the intrinsic genes defined by Sørlie, Perou and colleagues [18-20].

Tumors by			**Intrinsic subtypes**			
mutation status	**No.**	**Luminal A**	**Luminal B**	**ERBB2**	**Basal-like**	**Normal-like**
CHEK2 1100delC	26	8	18	0	0	0
BRCA1	47	1	7	0	39	0
BRCA2	6	1	5	0	0	0
Non-mutant	76	22	36	0	18	0
All tumors	155	32	66	0	57	0

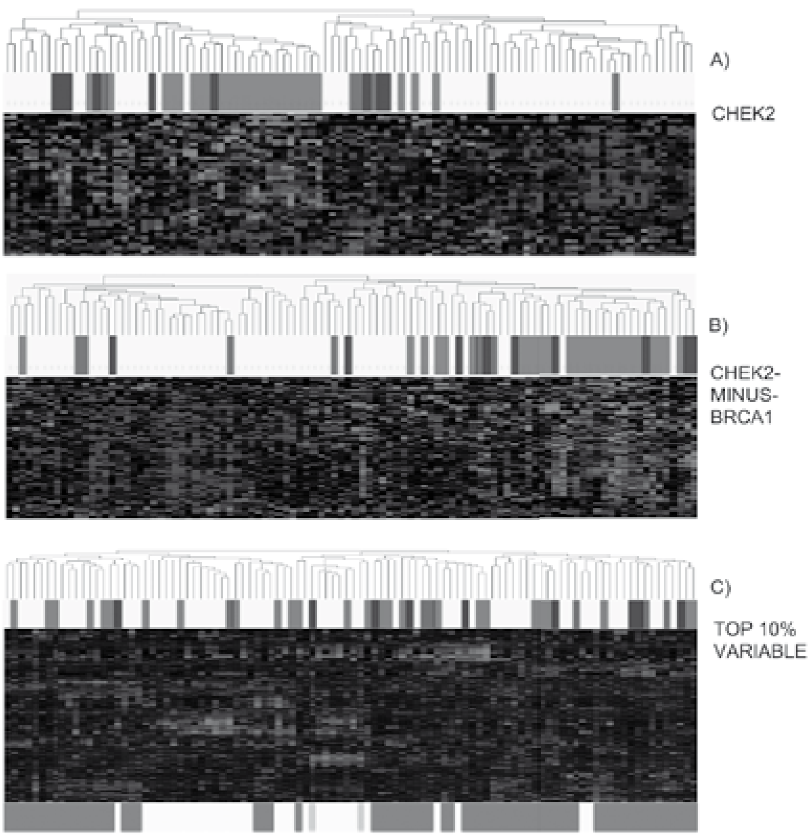


Figure 2: Hierarchical Clustering of 100 HR-pos Familial Breast Cancers. Gene expression heatmap, with red and green cells indicating overexpression and underexpression of probe sets relative to the log-2 transformed geometrical mean expression among all samples. A, Clustering based on the 40-gene *CHEK2* signature (represented by 43 probe sets); B, Clustering based on the 69-gene *CHEK2*-minus-*BRCA1* signature (represented by 71 probe sets); C, Unsupervised clustering based on the top 10% variably expressed probe sets (n=5,467). Color-coding mutation status: Red, *CHEK2* 1100delC tumors; Blue, *BRCA1* tumors; Green, *BRCA2* tumors; Yellow, Non-mutant tumors.

The division of *CHEK2* 1100delC tumors over the two luminal breast cancer subtypes only partially coincided with their observed cluster divisions upon unsupervised hierarchical clustering of the 100 HR-pos tumors based on the top 10% variably-expressed probe sets: 7 of 8

CHEK2 1100delC tumors of the luminal A subtype were found in the major "top 10%" cluster branch, whereas the 18 luminal B subtype tumors equally divided over both cluster branches (Fisher's exact P=0.08). It is important to note that the "top 10%" clustering involved 5,467 probe sets which were variably expressed among the 100 HR-pos breast cancers and thus likely also reflects tumor biology unrelated to *CHEK2* 1100delC mutation status. So even though the 40-gene *CHEK2* signature implies homogeneity among *CHEK2* 1100delC breast cancers, there apparently still exists some degree of heterogeneity among these tumors.

Discussion

Hormone Receptor Status of Breast Cancers May Confound Gene Signatures

We here have defined a 40-gene *CHEK2* signature that was significantly associated with *CHEK2* 1100delC breast cancers. However, definition of the *CHEK2* signature had not been possible without restriction of the analyses to the so-called hormone receptor-positive (HR-pos) cohort of familial breast cancers because of confounded expression of *ESR1* response genes. We successfully circumvented the issue by classifying each breast cancer as either HR-pos or HR-neg based on their global gene expression program (Figure 1) and then defined the *CHEK2* signature by using only HR-pos breast cancers. Although there was a strong concordance between this hormone receptor status classification and ER status of the tumors, we believe that using the full set of probes on the microarray is more reliable than using only *ESR1* transcript or ER protein expression data (Figure 1, with 96% and 9% *ESR1*-positive breast cancers in either tumor cluster). Indeed, hormone receptor status of breast cancers not only depends on ER status but also on PGR status and likely also on other factors such as FOXOA1, GATA3, TFF3, NAT1 and XBP. The global gene expression program of breast cancers includes all of these known and unknown biological factors that determine their hormone receptor-related biology. Exploitation of the complete expression data set therefore conceivably results in more accurate hormone receptor classification of breast cancers.

The *CHEK2* Oncogenic Pathway Includes *p53* and Likely Also *BRCA1*

Functional pathway analysis of the 40 genes from the *CHEK2* signature identified p53 signaling among the top biological networks associated with the signature. This was not an unexpected result as p53 is a well known phosphorylation target of *CHEK2* kinase [51,52,56-58]. A function of both proteins in the same oncogenic pathway would predict that most *CHEK2* 1100delC tumors carry wild-type *p53* alleles since a *p53* mutation would not confer a further selective survival advantage to the tumorigenic cells. Indeed, *p53* transcript expression was average among the *CHEK2* 1100delC breast cancers (Table 2). In contrast, and consistent with their predominantly *p53* mutant status [59-63], *p53* transcript expression was lower in a substantial fraction of the *BRCA1* and *BRCA2* breast cancers (Table 2). In fact, it could be that the apparent clustering of *BRCA1* tumors with the *CHEK2* 1100delC tumors in the hierarchical clustering based on the 40-gene *CHEK2* signature simply reflects their p53 pathway deficiency (Figure 2A). Alternatively, the *BRCA1* protein may also partake in the *CHEK2* oncogenic pathway. Indeed, *BRCA1* is another known phosphorylation target of *CHEK2* kinase [64,65] and the *CHEK2* signature also included the two BRCA1-associated genes *XIST* and *EED*. In addition, we and others had observed that *BRCA1* mutation carriers are less likely to also carry the *CHEK2* 1100delC mutation [5,6,8,9], again hinting to a functional association between *CHEK2* and *BRCA1*. Either way, the convergence of the functional pathway analysis on the well-documented *CHEK2* functions in cell cycle control and DNA damage responses [51,52], that typically also include *p53* and *BRCA1* proteins, is rather impressive and illustrates the central role of these processes in oncogenesis in the mammary gland.

CHEK2 1100delC Breast Cancers Are of the Luminal Intrinsic Subtypes

The *CHEK2* 1100delC mutation is thought to confer breast cancer susceptibility in concert with another as-yet-unknown susceptibility allele or alleles [5,6,8,9]. Therefore, identification of a gene signature that significantly associated with *CHEK2* 1100delC breast cancers implied an unexpected biological homogeneity among these tumors. The very existence of a *CHEK2* signature suggests that the *CHEK2* 1100delC mutation substantially contributes to the oncogenesis of *CHEK2* 1100delC breast cancers. It is notable that there still appears to be biological heterogeneity among *CHEK2* 1100delC breast cancers, among others involving their classification as luminal A or luminal B subtype breast cancers (Table 3). Their heterogeneity may reflect differences among the additional susceptibility alleles present in *CHEK2* 1100delC breast cancers or differences in the epithelial cell compartment from which the tumors originated. Still, the classification of all 26 *CHEK2* 1100delC tumors as breast cancers of the luminal subtypes suggests that only a limited number of additional susceptibility alleles are operative in the polygenic *CHEK2* model or, in case of still a substantial number of additional susceptibility alleles, that these alleles partake in only a few highly similar oncogenic pathways. Perhaps the most encouraging implication is that we now – with a 40-gene *CHEK2* signature in hand – may be able to identify the additional susceptibility alleles in the polygenic *CHEK2* model in a not too far future.

Methods

Breast Cancer Samples

Fresh-frozen female primary breast cancers were all selected from the Rotterdam Medical Oncology Tumor (RMOT) bank. Familial breast cancers were identified by linking records of tumor specimens present in the RMOT bank with records of breast cancer patients registered at the Rotterdam Family Cancer Clinic. All familial cases had been screened for mutations in *BRCA1* and *BRCA2* and for the *CHEK2* 1100delC mutation [6,66]. Additional *CHEK2* 1100delC cases had been identified by genetic screening of 1,706 RMOT cases that were unselected for a family history of cancer [8]. Three *CHEK2* 1100delC breast cancers have been excluded from the study because they had deleted the mutant allele, rendering the involvement of the *CHEK2* 1100delC mutation in the oncogenesis of these tumors uncertain. Together, the familial breast cancer cohort included 26 *CHEK2* 1100delC tumors, 47 *BRCA1* tumors, 6 *BRCA2* tumors, as well as 76 non-*BRCA1*/*BRCA2*/*CHEK2* 1100delC tumors designated "non-mutant tumors" (Table 1). The non-mutant breast cancer cases all were from a family with at least two breast cancer cases in first or second degree relatives of which at least one had been diagnosed before age 60 years. The Medical Ethical Committee at Erasmus MC has approved the study, which was carried out according the Code of Conduct of the Federation of Medical Scientific Societies in the Netherlands.

Mutation Screening

Screening for the *CHEK2* 1100delC mutation was performed by allele-specific oligonucleotide (ASO) hybridization as previously described [66]. Mutation-positive samples were confirmed by amplification of *CHEK2* exons 9-14 by long-range PCR, followed by nested PCR amplification of exon 10 [66]. Generated *CHEK2* templates were sequenced by using the Big Dye Terminator Cycle Sequencing kit (Applied Biosystems, Foster City, CA) and sequence reactions were analyzed on an ABI 3100 Capillary Sequencer. *BRCA1* and *BRCA2* mutation screens entailed screening of the complete coding sequences of the genes and intron/exon boundaries as well as screening for all known Dutch founder deletions, as previously described [6].

Gene Expression Profiling

Total RNA was isolated from cryostat sections using RNAzol B (Campro Scientific, Veenendaal, the Netherlands) and RNA quality and quantity was evaluated on an Agilent Bioanalyzer. Antisense biotinylated RNA was prepared and hybridized to Affymetrix U133 Plus 2.0 GeneChips, according to the manufacturer's guidelines (Affymetrix, Santa Clara, CA).

Microarray Data Processing

Intensity values were scaled to an average value of 100 per GeneChip by global scaling normalization, using the R Bioconductor “mas” algorithm (www.bioconductor.org – v1.7; www.r-project.org – v2.4.0). Intensity values below 30 were set at 30. The geometric mean of intensity values among all samples was calculated for each probe set and intensity values of each sample were then determined relative to the geometric mean and base-2 log transformed using Omniviz (Biowisdom, Maynard, MA).

Pearson Correlation Matrix

Omniviz package version 3.6 was used to calculate pair wise Pearson correlation coefficients based on overall gene expression of samples (Biowisdom, Maynard, MA), where all probe sets with variable expression in at least one sample were selected. The Omniviz Visualisation Tool was used to order and visualize a matrix of sample correlations.

Class Comparison

Differentially expressed genes between *CHEK2* 1100delC tumors and non-*CHEK2* HR-pos tumors were identified from among the top 20% variably-expressed probe sets (n=10,935) by using an univariate t-test with 1,000 permutations and univariate *P* value <0.001. We also performed a global test to determine differences between expression profiles of *CHEK2* 1100delC tumors and non-*CHEK2* HR-pos tumors, by permuting the class labels. The global test significance level represents the proportion of 1,000 permutations that resulted in at least as many genes as the original gene signature at significance level P<0.001. As a control, the same analysis was performed trice for 26 randomly selected HR-pos tumors.

Unsupervised Hierarchical Clustering

The NCI Biometric Research Branch BrB-Array Tool was used to perform unsupervised hierarchical cluster analysis (http://linus.nci.nih.gov/BRB-ArrayTools.html), using average distance linkage and centered correlation measures. Clustering was performed based on differentially-expressed probe sets identified by class comparisons or based on the top 10% variably-expressed probe sets for unsupervised analyses. The robustness of cluster reproducibility was calculated by perturbing the expression data with Gaussian noise and re-clustering 100 times and then measuring the similarity of the new clusters to the original clusters [67]. The thus obtained R measure represents the proportion of sample pairs that remained clustered together, averaged over all sample pairs and 100 perturbations. Fisher’s exact testing was used to determine the significance of distributions of the tumor samples over clusters.

Biological Pathway Analysis

The 40-gene *CHEK2* signature was evaluated for biological function and network interactions by using Ingenuity pathway analysis software (www.ingenuity.com). Genes from the *CHEK2* signature were ranked by their Ingenuity score, reflecting their involvement in biological networks that were recurrent among the signature genes, where biological networks are unique but not mutually exclusive. Ingenuity also identifies the most relevant biological processes among the signature genes by gene ontology analysis and calculates the significance of association of signature genes with canonical pathways and diseases.

Classification in Intrinsic Molecular Subtypes

To enable classification of the familial breast cancers in intrinsic molecular subtypes reported by Sørlie *et al.* [18-20], we transformed the reported intrinsic gene signatures to Affymetrix intrinsic gene signatures. Therefore, Genbank accession numbers of the 496 genes from the intrinsic gene signatures and Affymetrix ID annotation numbers from the U133 Plus 2.0 GeneChips were linked to Unigene HS numbers, allowing mapping of 451 unique Affymetrix probe sets. Unsupervised hierarchical clustering of all 155 familial breast cancers based on the intrinsic gene signature was performed as described above, except that non-centered correlation metric was used in stead of centered correlation in order to more accurately replicate analyses reported by Sørlie *et al.* [18-20].

Acknowledgements

We are grateful to the breast cancer patients and their clinicians for participation in this research. We thank Marion Meier-van Gelder, Mieke Timmermans, Miranda Arnold, Anneke Goedheer, Roberto Rodriguez-Garcia, Wendy van der Smissen and Anja de Snoo for their technical assistance. We also thank Wim van Putten and Antoinette Hollestelle for insightful discussions. This work was supported by grants DDHK 2002-2687 and DDHK 2003-2862 from the Dutch Cancer Society KWF, and in part by the Netherlands Genomics Initiative / Netherlands Organisation for Scientific Research (NWO). H.M.-H. is a fellow from the NWO Vidi Research Program.

References

1. Miki Y, Swensen J, Shattuck-Eidens D, et al. *A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1*. Science 1994;**266**(5182):66-71.
2. Wooster R, Bignell G, Lancaster J, et al. *Identification of the breast cancer susceptibility gene BRCA2.* Nature 1995;**378**(6559):789-92.
3. Tavtigian SV, Simard J, Rommens J, et al. *The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds.* Nat Genet 1996;**12**(3):333-7.
4. Thompson D, Easton D. *The genetic epidemiology of breast cancer genes.* J Mammary Gland Biol Neoplasia 2004;**9**(3):221-36.
5. Walsh T, King MC. *Ten genes for inherited breast cancer.* Cancer Cell 2007;**11**(2):103-5.
6. Meijers-Heijboer H, van den Ouweland A, Klijn J, et al. *Low-penetrance susceptibility to breast cancer due to CHEK2*1100delC in noncarriers of BRCA1 or BRCA2 mutations.* Nat Genet 2002;**31**(1):55-9.
7. Vahteristo P, Bartkova J, Eerola H, et al. *A CHEK2 genetic variant contributing to a substantial fraction of familial breast cancer.* Am J Hum Genet 2002;**71**(2):432-8.
8. *Consortium CBCC-C. CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies.* Am J Hum Genet 2004;**74**(6):1175-82.
9. Nevanlinna H, Bartek J. *The CHEK2 gene and inherited breast cancer susceptibility.* Oncogene 2006;**25**(43):5912-9.
10. Oldenburg RA, Kroeze-Jansema K, Kraan J, et al. *The CHEK2*1100delC variant acts as a breast cancer risk modifier in non-BRCA1/BRCA2 multiple-case families.* Cancer Res 2003;**63**(23):8153-7.
11. Peto J. B*reast cancer susceptibility-A new look at an old model.* Cancer Cell 2002;**1**(5):411-2.
12. Smith P, McGuffog L, Easton DF, et al. *A genome wide linkage search for breast cancer susceptibility genes.* Genes Chromosomes Cancer 2006.
13. Renwick A, Thompson D, Seal S, et al. *ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles.* Nat Genet 2006;**38**(8):873-5.
14. Seal S, Thompson D, Renwick A, et al. *Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles.* Nat Genet 2006;**38**(11):1239-41.
15. Rahman N, Seal S, Thompson D, et al. *PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene.* Nat Genet 2007;**39**(2):165-7.
16. Cox A, Dunning AM, Garcia-Closas M, et al. *A common coding variant in CASP8 is associated with breast cancer risk.* Nat Genet 2007;**39**(3):352-8.
17. Easton DF, Pooley KA, Dunning AM, et al. *Genome-wide association study identifies novel breast cancer susceptibility loci.* Nature 2007;**447**(7148):1087-93.
18. Perou CM, Sorlie T, Eisen MB, et al. *Molecular portraits of human breast tumours.* Nature 2000;**406**(6797):747-52.
19. Sorlie T, Perou CM, Tibshirani R, et al. *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A 2001;**98**(19):10869-74.
20. Sorlie T, Tibshirani R, Parker J, et al. *Repeated observation of breast tumor subtypes in independent gene expression data sets.* Proc Natl Acad Sci U S A 2003;**100**(14):8418-23.
21. Fan C, Oh DS, Wessels L, et al. *Concordance among gene-expression-based predictors for breast cancer.* N Engl J Med 2006;**355**(6):560-9.
22. van 't Veer LJ, Dai H, van de Vijver MJ, et al. *Gene expression profiling predicts clinical outcome of breast cancer.* Nature 2002;**415**(6871):530-6.
23. van de Vijver MJ, He YD, van't Veer LJ, et al. *A gene-expression signature as a predictor of survival in breast cancer.* N Engl J Med 2002;**347**(25):1999-2009.
24. Ahr A, Karn T, Solbach C, et al. *Identification of high risk breast-cancer patients by gene expression profiling.* Lancet 2002;**359**(9301):131-2.
25. Huang E, Cheng SH, Dressman H, et al. *Gene expression predictors of breast cancer outcomes.* Lancet 2003;**361**(9369):1590-6.

26. Sotiriou C, Neo SY, McShane LM, et al. *Breast cancer classification and prognosis based on gene expression profiles from a population-based study.* Proc Natl Acad Sci U S A 2003;**100**(18):10393-8.
27. Ma XJ, Wang Z, Ryan PD, et al. *A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen.* Cancer Cell 2004;**5**(6):607-16.
28. Paik S, Shak S, Tang G, et al. *A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.* N Engl J Med 2004;**351**(27):2817-26.
29. Ramaswamy S, Ross KN, Lander ES, Golub TR. *A molecular signature of metastasis in primary solid tumors.* Nat Genet 2003;**33**(1):49-54.
30. Chang JC, Wooten EC, Tsimelzon A, et al. *Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer.* Lancet 2003;**362**(9381):362-9.
31. Wang Y, Klijn JG, Zhang Y, et al. *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.* Lancet 2005;**365**(9460):671-9.
32. Jansen MP, Foekens JA, van Staveren IL, et al. *Molecular classification of tamoxifen-resistant breast carcinomas by gene expression profiling.* J Clin Oncol 2005;**23**(4):732-40.
33. Foekens JA, Atkins D, Zhang Y, et al. *Multicenter Validation of a Gene Expression-Based Prognostic Signature in Lymph Node-Negative Primary Breast Cancer.* J Clin Oncol 2006.
34. Hess KR, Anderson K, Symmans WF, et al. *Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer.* J Clin Oncol 2006;**24**(26):4236-44.
35. Hedenfalk I, Duggan D, Chen Y, et al. *Gene-expression profiles in hereditary breast cancer.* N Engl J Med 2001;**344**(8):539-48.
36. Dressman MA, Baras A, Malinowski R, et al. *Gene expression profiling detects gene amplification and differentiates tumor types in breast cancer.* Cancer Res 2003;**63**(9):2194-9.
37. Miller LD, Smeds J, George J, et al. *An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival.* Proc Natl Acad Sci U S A 2005;**102**(38):13550-5.
38. Adler AS, Lin M, Horlings H, Nuyten DS, van de Vijver MJ, Chang HY. *Genetic regulators of large-scale transcriptional signatures in cancer.* Nat Genet 2006;**38**(4):421-30.
39. Verhoog LC, Brekelmans CT, Seynaeve C, et al. *Survival and tumour characteristics of breast-cancer patients with germline mutations of BRCA1.* Lancet 1998;**351**(9099):316-21.
40. Lakhani SR, Van De Vijver MJ, Jacquemier J, et al. *The pathology of familial breast cancer: predictive value of immunohistochemical markers estrogen receptor, progesterone receptor, HER-2, and p53 in patients with mutations in BRCA1 and BRCA2.* J Clin Oncol 2002;**20**(9):2310-8.
41. de Bock GH, Schutte M, Krol-Warmerdam EM, et al. *Tumour characteristics and prognosis of breast cancer patients carrying the germline CHEK2*1100delC variant.* J Med Genet 2004;**41**(10):731-5.
42. Schmidt MK, Tollenaar RA, de Kemp SR, et al. *Breast cancer survival and tumor characteristics in premenopausal women carrying the CHEK2*1100delC germline mutation.* J Clin Oncol 2007;**25**(1):64-9.
43. Hickson ID. *RecQ helicases: caretakers of the genome.* Nat Rev Cancer 2003;**3**(3):169-78.
44. Kitao S, Shimamoto A, Goto M, et al. *Mutations in RECQL4 cause a subset of cases of Rothmund-Thomson syndrome.* Nat Genet 1999;**22**(1):82-4.
45. Ellis NA, Groden J, Ye TZ, et al. *The Bloom's syndrome gene product is homologous to RecQ helicases.* Cell 1995;**83**(4):655-66.
46. Yu CE, Oshima J, Fu YH, et al. *Positional cloning of the Werner's syndrome gene.* Science 1996;**272**(5259):258-62.
47. Ganesan S, Silver DP, Greenberg RA, et al. *BRCA1 supports XIST RNA concentration on the inactive X chromosome.* Cell 2002;**111**(3):393-405.
48. Richardson AL, Wang ZC, De Nicolo A, et al. *X chromosomal abnormalities in basal-like human breast cancer.* Cancer Cell 2006;**9**(2):121-32.
49. Wang J, Mager J, Chen Y, et al. *Imprinted X inactivation maintained by a mouse Polycomb group gene.* Nat Genet 2001;**28**(4):371-5.
50. Silva J, Mak W, Zvetkova I, et al. *Establishment of histone h3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes.* Dev Cell 2003;**4**(4):481-95.

51. Zhou BB, Elledge SJ. *The DNA damage response: putting checkpoints in perspective.* Nature 2000;**408**(6811):433-9.

52. Bartek J, Falck J, Lukas J. *CHK2 kinase - a busy messenger.* Nat Rev Mol Cell Biol 2001;2(12):877-86.

53. Moffa AB, Tannheimer SL, Ethier SP. *Transforming potential of alternatively spliced variants of fibroblast growth factor receptor 2 in human mammary epithelial cells.* Mol Cancer Res 2004;**2**(11):643-52.

54. Adnane J, Gaudray P, Dionne CA, et al. *BEK and FLG, two receptors to members of the FGF family, are amplified in subsets of human breast cancers.* Oncogene 1991;**6**(4):659-63.

55. Greenman C, Stephens P, Smith R, et al. *Patterns of somatic mutation in human cancer genomes.* Nature 2007;**446**(7132):153-8.

56. Hirao A, Kong YY, Matsuoka S, et al. *DNA damage-induced activation of p53 by the checkpoint kinase chk2.* Science 2000;**287**(5459):1824-7.

57. Chehab NH, Malikzay A, Appel M, Halazonetis TD. *Chk2/hCds1 functions as a DNA damage checkpoint in G(1) by stabilizing p53.* Genes Dev 2000;**14**(3):278-88.

58. Shieh SY, Ahn J, Tamai K, Taya Y, Prives C. *The human homologs of checkpoint kinases Chk1 and Cds1 (Chk2) phosphorylate p53 at multiple DNA damage-inducible sites.* Genes Dev 2000;**14**(3):289-300.

59. Crook T, Crossland S, Crompton MR, Osin P, Gusterson BA. *p53 mutations in BRCA1-associated familial breast cancer.* Lancet 1997;**350**(9078):638-9.

60. Crook T, Brooks LA, Crossland S, et al. *p53 mutation with frequent novel condons but not a mutator phenotype in BRCA1- and BRCA2-associated breast tumours.* Oncogene 1998;**17**(13):1681-9.

61. Phillips KA, Nichol K, Ozcelik H, et al. *Frequency of p53 mutations in breast carcinomas from Ashkenazi Jewish carriers of BRCA1 mutations.* J Natl Cancer Inst 1999;**91**(5):469-73.

62. Schuyer M, Berns EM. *Is TP53 dysfunction required for BRCA1-associated carcinogenesis?* Mol Cell Endocrinol 1999;**155**(1-2):143-52.

63. Wasielewski M, Elstrodt F, Klijn JG, Berns EM, Schutte M. *Thirteen new p53 gene mutants identified among 41 human breast cancer cell lines.* Breast Cancer Res Treat 2006;**99**:97-101.

64. Cortez D, Wang Y, Qin J, Elledge SJ. *Requirement of ATM-dependent phosphorylation of brca1 in the DNA damage response to double-strand breaks.* Science 1999;**286**(5442):1162-6.

65. Lee JS, Collins KM, Brown AL, Lee CH, Chung JH. *hCds1-mediated phosphorylation of BRCA1 regulates the DNA damage response.* Nature 2000;**404**(6774):201-4.

66. Meijers-Heijboer H, Wijnen J, Vasen H, et al. *The CHEK2 1100delC mutation identifies families with a hereditary breast and colorectal cancer phenotype.* Am J Hum Genet 2003;**72**(5):1308-14.

67. McShane LM, Radmacher MD, Freidlin B, Yu R, Li MC, Simon R. *Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data.* Bioinformatics 2002;**18**(11):1462-9.

Table S1: The 40-gene CHEK2 1100delC signature. Probe sets that represent identical genes; # probe sets that are not shared with the *CHEK2*-minus-*BRCA1* signature (see Table S2); P-value, parametric P-value; FDR, false discovery rate; GM *CHEK2* / non-*CHEK2*, geometric mean of expression intensities among *CHEK2* 1100delC versus non-*CHEK2* mutant tumors with their ratio given between brackets.

Probe set	Gene symbol	Gene description	P-value	FDR	*CHEK2* / non-*CHEK2* GM (ratio)
210416_s_at	CHEK2	Cell cycle checkpoint kinase 2	< 1e-07	< 1e-07	0.6 / 1.2 (0.5)
229388_at		cDNA clone IMAGE:5295564	4.0E-07	0.002	1.6 / 0.9 (1.8)
236098_at	RECQL5	RecQ protein-like 5	9.6E-06	0.035	1.6 / 0.9 (1.8)
1557383_a_at		cDNA FLJ38112 fis, clone D3OST2002272	1.8E-05	0.042	1.8 / 0.8 (2.3)
231369_at	ZNF333	Zinc finger protein 333	2.6E-05	0.042	1.5 / 0.9 (1.7)
229795_at		Transcribed locus	2.8E-05	0.042	1.7 / 0.8 (2.1)
1555976_s_at	MRCL3	Myosin regulatory light chain 3	3.1E-05	0.042	1.6 / 0.9 (1.8)
223773_s_at	C1orf79 *	Chromosome 1 ORF 79	3.2E-05	0.042	1.5 / 0.9 (1.7)
215292_s_at	MKL1	Megakaryoblastic leukemia 1	3.8E-05	0.042	0.6 / 1.2 (0.5)
226444_at	SLC39A10	Solute carrier family 39, member 10 (zinc transporter)	3.9E-05	0.042	1.6 / 0.8 (2.0)
1553292_s_at	FLJ25006	Hypothetical protein FLJ25006	4.9E-05	0.045	1.5 / 0.9 (1.7)
1557242_at	MICA	MHC class I polypeptide-related sequence A	5.0E-05	0.045	1.5 / 0.9 (1.7)
208798_x_at	GOLGA8A	Golgi autoantigen, golgin subfamily a, 8A	7.8E-05	0.065	1.7 / 0.8 (2.1)
1555878_at	RPS24	Ribosomal protein S24	9.4E-05	0.073	1.7 / 0.8 (2.1)
209889_at	SEC31L2	SEC31-like 2 (S. cerevisiae)	0.00011	0.077	1.5 / 0.9 (1.7)
228997_at	TRSPAP1	tRNA selenocysteine associated protein 1	0.00013	0.087	1.7 / 0.8 (2.1)
227793_at		cDNA clone IMAGE:4361039	0.00014	0.087	1.8 / 0.8 (2.3)
214605_x_at	GPR1	G protein-coupled receptor 1	0.00016	0.093	1.5 / 0.9 (1.7)
221728_x_at	XIST *	X (inactive)-specific transcript	0.00016	0.093	1.7 / 0.9 (1.9)
227154_at	IGSF21	Immunoglobin superfamily, member 21	0.00022	0.118	0.5 / 1.2 (0.4)
1555960_at	HINT1	Histidine triad nucleotide BP 1	0.00024	0.118	1.5 / 0.9 (1.7)

241402_at	TSEN54	tRNA splicing endonuclease 54 (S. cerevisiae)	0.00024	0.118	1.6 / 0.9 (1.8)
211497_x_at	NKX3-1	NK3 transcription factor related 1 (Drosophila)	0.00027	0.124	0.7 / 1.1 (0.6)
217317_s_at	HERC2P2	HECT domain and RLD 2 pseudogene 2	0.00029	0.124	1.4 / 0.9 (1.6)
217593_at	ZNF447	Zinc finger protein 447	0.00030	0.124	1.5 / 0.9 (1.7)
235535_x_at	LOC283788	Hypothetical protein LOC283788	0.00032	0.124	1.5 / 0.9 (1.7)
202127_at	PRPF4B	Pre-mRNA processing factor 4 homolog B (yeast)	0.00032	0.124	1.4 / 0.9 (1.6)
228030_at		Transcribed locus	0.00033	0.124	1.5 / 0.9 (1.7)
223774_at	C1orf79 *	Chromosome 1 ORF 79	0.00033	0.124	1.5 / 0.9 (1.7)
209403_at	TBC1D3	LOC 653380 / 653498	0.00036	0.129	1.7 / 0.9 (1.9)
220798_x_at	PRG2	Plasticity-related gene 2	0.00037	0.131	0.7 / 1.1 (0.6)
210656_at	EED	Embryonic ectoderm development	0.00039	0.134	1.5 / 0.9 (1.7)
211599_x_at	MET	MET proto-oncogene (HGFR)	0.00066	0.217	0.7 / 1.1 (0.6)
230959_at	LOC644450#	Hypothetical protein LOC644450	0.00072	0.231	1.4 / 0.9 (1.6)
221792_at	RAB6B *#	RAB6B, member RAS oncogene family	0.00077	0.234	0.7 / 1.1 (0.6)
227671_at	XIST *	X (inactive)-specific transcript	0.00077	0.234	1.5 / 0.9 (1.7)
209007_s_at	C1orf63	Chromosome 1 ORF 63	0.00079	0.234	1.4 / 0.9 (1.6)
226305_at	LYNX1#	Ly6 / Neurotoxin 1	0.00084	0.241	0.7 / 1.2 (0.6)
225259_at	RAB6B *#	RAB6B, member RAS oncogene family	0.00088	0.242	0.7 / 1.1 (0.6)
226363_at	ABCC5	ATP-binding cassette, sub-family C (CFTR/MRP), member 5	0.00089	0.243	1.5 / 0.9 (1.7)
225786_at	LOC284702	Hypothetical protein LOC284702	0.00096	0.248	1.4 / 0.9 (1.6)
210425_x_at	GOLGA8B	Golgi autoantigen, golgin subfamily a, 8B	0.00096	0.248	1.5 / 0.9 (1.7)
221989_at	RPL10	LOC 284393 / 285176 /389342 / 39364 /644039 /647074	0.00097	0.248	1.4 / 0.9 (1.6)

Table S2: The 69-gene *CHEK2*-minus-*BRCA1* Signature. *, Probe sets that represent identical genes; # probe sets that are not shared with the CHEK2 signature (see Table S2); P-value, parametric P-value; FDR, false discovery rate; GM *CHEK2* / non-*CHEK2*, geometric mean of expression intensities among *CHEK2* 1100delC versus non-*CHEK2* mutant tumors with their ratio given between brackets.

Probe set	Gene symbol	Gene description	P-value	FDR	*CHEK2* / non-*CHEK2* GM (ratio)
210416_s_at	CHEK2	Cell cycle checkpoint kinase 2	< 1e-07	< 1e-07	0.6 / 1.1 (0.5)
229388_at		cDNA clone IMAGE:5295564	2.0E-07	0.001	1.6 / 0.8 (2.0)
223773_s_at	C1orf79 *	Chromosome 1 ORF 79	2.4E-06	0.007	1.5 / 0.8 (1.8)
1555976_s_at	MRCL3	Myosin regulatory light chain 3	2.6E-06	0.007	1.6 / 0.8 (2.0)
202127_at	PRPF4B	Pre-mRNA processing factor 4 homolog B (yeast)	4.6E-06	0.010	1.4 / 0.8 (1.7)
1553292_s_at	FLJ25006	Hypothetical protein FLJ25006	5.6E-06	0.010	1.5 / 0.8 (1.8)
226444_at	SLC39A10	Solute carrier family 39, member 10 (zinc transporter)	6.7E-06	0.010	1.6 / 0.8 (2.0)
209889_at	SEC31L2	SEC31-like 2 (S. cerevisiae)	1.2E-05	0.013	1.5 / 0.8 (1.8)
236098_at	RECQL5	RecQ protein-like 5	1.2E-05	0.013	1.6 / 0.9 (1.7)
225786_at	LOC284702	Hypothetical protein LOC284702	1.4E-05	0.013	1.4 / 0.8 (1.7)
208798_x_at	GOLGA8A	Golgi autoantigen, golgin subfamily a, 8A	1.4E-05	0.013	1.7 / 0.8 (2.1)
1555878_at	RPS24	Ribosomal protein S24	1.6E-05	0.014	1.7 / 0.8 (2.1)
229795_at		Transcribed locus	1.7E-05	0.014	1.7 / 0.8 (2.1)
215292_s_at	MKL1	Megakaryoblastic leukemia 1	1.8E-05	0.014	0.6 / 1.2 (0.5)
1555960_at	HINT1	Histidine triad nucleotide BP 1	2.4E-05	0.016	1.5 / 0.8 (1.8)
1557383_a_at		cDNA FLJ38112 fis, clone D3OST2002272	2.5E-05	0.016	1.8 / 0.8 (2.2)
227793_at		cDNA clone IMAGE:4361039	2.5E-05	0.016	1.8 / 0.7 (2.5)
217317_s_at	HERC2P2	HECT domain and RLD 2 pseudogene 2	3.1E-05	0.018	1.4 / 0.8 (1.7)
228997_at	TRSPAP1	tRNA selenocysteine associated protein 1	3.3E-05	0.018	1.7 / 0.8 (2.1)
232504_at	#		3.3E-05	0.018	1.4 / 0.8 (1.7)
231369_at	ZNF333	Zinc finger protein 333	4.5E-05	0.023	1.5 / 0.9 (1.6)
228990_at	C1orf79 *	Chromosome 1 ORF 79	4.9E-05	0.024	1.4 / 0.8 (1.7)

210656_at	EED	Embryonic ectoderm development	5.0E-05	0.024	1.5 / 0.8 (1.8)
223774_at	C1orf79 *	Chromosome 1 ORF 79	7.5E-05	0.033	1.5 / 0.8 (1.8)
1557242_at	MICA	MHC class I polypeptide-related sequence A	7.6E-05	0.033	1.5 / 0.9 (1.6)
221989_at	RPL10	LOC 284393 / 285176 / 389342 / 39364 / 644039 / 647074	9.0E-05	0.037	1.4 / 0.8 (1.7)
221728_x_at	XIST	X (inactive)-specific transcript	9.6E-05	0.037	1.7 / 0.9 (1.8)
235535_x_at	LOC283788	Hypothetical protein LOC283788	9.8E-05	0.037	1.5 / 0.8 (1.8)
224610_at	STX5 #	Syntaxin 5	9.9E-05	0.037	1.4 / 0.8 (1.7)
220798_x_at	PRG2	Plasticity-related gene 2	0.00011	0.038	0.7 / 1.2 (0.5)
214048_at	MBD4 #	Methyl-CpG binding domain protein 4	0.00014	0.048	1.4 / 0.8 (1.7)
213672_at	MARS #	Methionine-tRNA synthetase	0.00016	0.056	1.4 / 0.8 (1.7)
209007_s_at	C1orf63	Chromosome 1 ORF 63	0.00020	0.067	1.4 / 0.8 (1.7)
226363_at	ABCC5	ATP-binding cassette, sub-family C (CFTR / MRP), member 5	0.00021	0.067	1.5 / 0.8 (1.8)
244786_at	C14orf62 #	Chromosome 14 ORF 62	0.00022	0.067	1.5 / 0.8 (1.8)
213212_x_at	LOC161527 #	Hypothetical protein LOC161527	0.00022	0.067	1.4 / 0.8 (1.8)
228030_at		Transcribed locus	0.00023	0.067	1.5 / 0.8 (1.8)
227517_s_at	C1orf155 #	Chromosome 1 ORF 155	0.00025	0.071	1.5 / 0.8 (1.8)
235564_at	ZNF117 #	Zinc finger protein 117 (HPF9)	0.00025	0.071	1.4 / 0.8 (1.7)
214163_at	#		0.00028	0.075	1.5 / 0.8 (1.8)
214605_x_at	GPR1	G protein-coupled receptor 1	0.00030	0.075	1.5 / 0.9 (1.6)
221768_at	SFPQ #	Splicing factor proline / glutamine-rich	0.00030	0.075	1.3 / 0.9 (1.4)
210425_x_at	GOLGA8B	Golgi autoantigen, golgin subfamily a, 8B	0.00030	0.075	1.5 / 0.8 (1.8)
227154_at	IGSF21	Immunoglobin superfamily, member 21	0.00031	0.075	0.5 / 1.2 (0.4)
222018_at	NACA #	Nascent-polypeptide-associated complex alpha polypeptide	0.00031	0.075	1.4 / 0.8 (1.7)
240594_at	CDYL #	Chromodomain protein, Y-like	0.00032	0.075	1.3 / 0.9 (1.4)
241402_at	TSEN54	tRNA splicing endonuclease 54 homolog (S. cerevisiae)	0.00034	0.078	1.6 / 0.9 (1.7)

235315_at	TSC22D1 #	TSC22 domain family, member 1	0.00034	0.078	1.5 / 0.9 (1.6)
222040_at	HNRPA1 #	Heterogeneous nuclear ribonucleoprotein A1	0.00036	0.078	1.3 / 0.8 (1.6)
209403_at	TBC1D3	LOC 653380 / 653498	0.00036	0.078	1.7 / 0.8 (2.1)
236314_at	#	Transcribed locus, weakly similar to NP_055301.1 AD7c-NTP	0.00038	0.081	1.3 / 0.8 (1.6)
226316_at	#	cDNA clone IMAGE:5295896	0.00042	0.084	1.4 / 0.8 (1.7)
205933_at	SETBP1 #	SET binding protein 1	0.00043	0.084	0.8 / 1.2 (0.6)
211497_x_at	NKX3-1	NK3 transcription factor related, locus 1 (Drosophila)	0.00043	0.084	0.7 / 1.1 (0.6)
211599_x_at	MET	MET proto-oncogene (HGFR)	0.00043	0.084	0.7 / 1.2 (0.5)
227350_at	HELLS #	Helicase, lymphoid-specific	0.00044	0.084	1.4 / 0.8 (1.7)
218750_at	JOSD3 #	Josephin domain containing 3	0.00044	0.084	1.4 / 0.8 (1.7)
1555920_at	CBX3 #	Chromobox homolog 3 (HP1 gamma homolog, Drosophila)	0.00045	0.084	1.4 / 0.8 (1.7)
231697_s_at	TMEM49 #	Transmembrane protein 49	0.00047	0.088	1.4 / 0.8 (1.7)
214683_s_at	CLK1 #	CDC-like kinase 1	0.00049	0.090	1.4 / 0.8 (1.7)
218990_s_at	SPRR3 #	Small proline-rich protein 3	0.00053	0.094	1.4 / 0.9 (1.5)
227168_at	FLJ25967 #	Hypothetical gene supported by AK098833	0.00053	0.094	1.6 / 0.8 (2.0)
229437_at	BIC #	BIC transcript	0.00055	0.095	1.5 / 0.8 (1.8)
218922_s_at	LASS4 #	LAG1 longevity assurance homolog 4 (S. cerevisiae)	0.00065	0.111	0.7 / 1.1 (0.6)
216407_at	VAC14 #	Vac14 homolog (S. cerevisiae)	0.00066	0.111	0.7 / 1.1 (0.6)
228387_at	#	cDNA DKFZp686B0610	0.00075	0.123	1.4 / 0.9 (1.5)
228111_s_at	DNAH1 #	Dynein, axonemal, heavy polypeptide 1	0.00075	0.123	1.4 / 0.9 (1.5)
227971_at	NRK #	Nik related kinase	0.00076	0.123	1.8 / 0.8 (2.2)
201295_s_at	WSB1 #	WD repeat and SOCS box-containing 1	0.00088	0.138	1.4 / 0.9 (1.5)
232090_at	DNM3 #	Dynamin 3	0.00088	0.138	1.5 / 0.8 (1.8)
217593_at	ZNF447	Zinc finger protein 447	0.00091	0.141	1.5 / 0.9 (1.6)

Chapter 6

Identification of differentially regulated splice-variants and novel exons in glial brain tumors using exon expression arrays

Pim J. French[1]; ***Justine Peeters***[2]; Sebastiaan Horsman[2]; Elza Duijm[1]; Martin J. van den Bent[1]; Theo M. Luider[1]; Johan M. Kros[3]; Peter van der Spek[2] and Peter A. Sillevis Smitt[1]

[1] Deptartment of Neurology, Erasmus MC, Rotterdam, The Netherlands

[2] Bioinformatics, Erasmus MC, Rotterdam, The Netherlands and

[3] Pathology, Erasmus MC, Rotterdam, The Netherlands

Cancer Research 2007;67 (12):5635-42.

Abstract

Aberrant splice variants are involved in the initiation and/or progression of glial brain tumors. We therefore set out to identify splice variants that are differentially expressed between histologic subgroups of gliomas. Splice variants were identified using a novel platform that profiles the expression of virtually all known and predicted exons present in the human genome. Exon-level expression profiling was done on 26 glioblastomas, 22 oligodendrogliomas, and 6 control brain samples. Our results show that Human Exon arrays can identify subgroups of gliomas based on their histologic appearance and genetic aberrations. We next used our expression data to identify differentially expressed splice variants. In two independent approaches, we identified 49 and up to 459 exons that are differentially spliced between glioblastomas and oligodendrogliomas, a subset of which (47% and 33%) were confirmed by reverse transcription-PCR (RT-PCR). In addition, exon level expression profiling also identified >700 novel exons. Expression of ~67% of these candidate novel exons was confirmed by RT-PCR. Our results indicate that exon level expression profiling can be used to molecularly classify brain tumor subgroups, can identify differentially regulated splice variants, and can identify novel exons. The splice variants identified by exon level expression profiling may help to detect the genetic changes that cause or maintain gliomas and may serve as novel treatment targets.

Introduction

Diffuse gliomas are the most common primary central nervous system tumors in adults [1,2], and it is estimated that 43,800 new patients are diagnosed in 2005 with a primary brain tumor in the United States.Based on their histologic appearance, gliomas can be divided into astrocytic tumors, pure oligodendroglial tumors, and mixed oligoastrocytic tumors according to standard WHO classification [3]. Despite advances in neurosurgery, chemotherapy, and radiotherapy, the prognosis for most glioma patients remains dismal [4,5].

There is strong evidence that aberrant splice isoforms are involved in the initiation and/or progression of glial brain tumors [6]. For example, glioblastomas with epidermal growth factor receptor (EGFR) amplification frequently (32 of 48) express *EGFRvIII*, a tumor-specific, ligand-independent, constitutively active isoform of the *EGFR* that lacks exons 2 to 7 [7]. Expression of this splice variant can induce glioma formation in mice [8] and is associated with response to *EGFR* kinase inhibitors in human[9]. Other, (activating) aberrant *EGFR* splice variants are also frequently observed in gliomas [10]. In addition, many nervous system cancer– related spice variants were identified using a gene-centric [11–16] or bioinformatical approach screening public domain databases [17].

Because aberrant splice isoforms are involved in the initiation and/or progression of glial brain tumors, we initiated a screen to identify splice variants expressed in gliomas. Our screen was done by profiling the expression of virtually all known and predicted exons in the human genome (1.4 million). Splice variants were then calculated from the expression level of exons relative to its transcript. Our results indicate that exon level expression profiling can classify brain tumor subgroups based on their histologic appearance, can identify differentially regulated splice variants, and can identify novel exons.

Materials and Methods

Samples

All glioma samples were derived from patients treated within the Erasmus MC. Patient data, histologic diagnosis, and chromosomal aberrations are summarized in Supplementary Table S1. Samples were collected immediately after surgical resection, snapped frozen, and stored at -80°C. All samples were visually inspected on 5-μm H&E–stained frozen sections by the neuropathologist (J.M.K.). We selected 48 glioma samples including (*a*) classic oligodendrogliomas with loss of heterozygosity (LOH) on 1p and 19q (*n*=22, of which 20 WHO

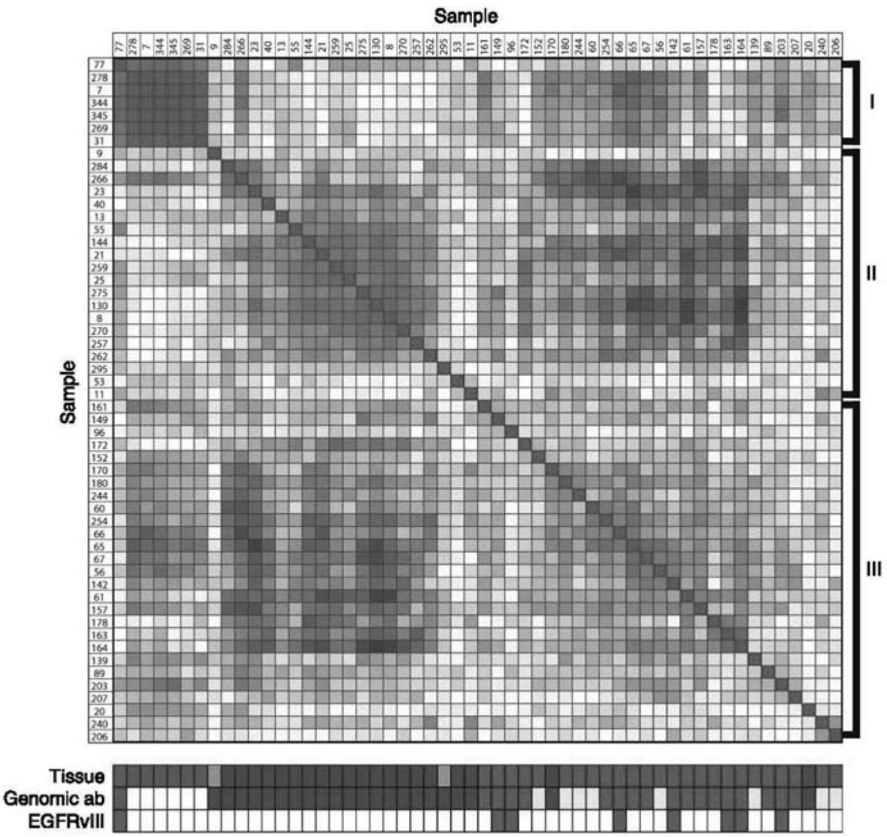


Figure 1. Correlation plot of all samples. Samples are plotted against each other as Pearson's correlation to determine the degree of similarity based on expressed exons. All exons with 4-fold expression difference from the geometric mean are included in the clustering. Red, high correlation; blue, low correlation. Below the correlation plot is a graphic representation of histologic and patient data. *Tissue.* Origin of sample: control cortex; 🟩 anaplastic oligodendroglioma (WHO grade III); 🟩 oligodendroglioma (WHO grade II); and 🟥 glioblastoma. Genomic aberrations. Genomic aberrations of the sample: 5 control sample; LOH on 1p and 19q, no amplification of *EGFR*; 🟥 no LOH on 1p and 19q but amplification of *EGFR*; no LOH on 1p and 19q, no amplification of *EGFR*. *EGFRvIII*: expression of *EGFRvIII* as determined by RT-PCR: 5 no expression; 🟥 expression. Subgroups identified by Pearsons's correlation plot (right; I–III).

grade III and 2 WHO grade II; ref. 3); (*b*) primary glioblastoma with EGFR amplification (*n*=18); and (*c*) secondary glioblastoma without EGFR amplification (*n*=8). Six control brain samples from patients with no history of neurologic disease were also included. All but one sample

(GBM 77) contained >70% tumors. Tissue adjacent to the inspected sections was subsequently used for nucleic acid isolation. Microsatellite analysis on 1p and 19q and amplification of the EGFR were done as described [18].

Nucleic acid isolation, cDNA synthesis, and array hybridization

Total RNA and genomic DNA was isolated from 20to 40 cryostat sections of 40-μm thickness (50–100 mg) using Trizol (Invitrogen) according to the manufacturer's instructions (see also ref. 18). Total RNA was then further purified on RNeasy mini columns (Qiagen). RNA quality was assessed on a Bioanalyser (Agilent). High-quality RNA (i.e., RNA integrity number >7.0; ref. 19) was used for our experiments. rRNA reduction, first round doublestrand–cDNA synthesis, cRNA synthesis, second round single-strand (ss)– cDNA synthesis, ss-cDNA fragmentation, and labeling was done according to the Affymetrix GeneChip Whole-Transcript Sense Target–Labeling Assay manual. Affymetrix Human Exon 1.0ST microarrays were hybridized overnight with 5-μg biotin–labeled ss-cDNA.

Data analysis

Signal intensity estimate and P value for each probe set were extracted from the arrays in Affymetrix ExACT 1.0software using the PLIER and DABG algorithm, respectively. PLIER expression data were normalized using the quantile method in *R* statistical software v2.2.1. DABG *P* values allow calculation of false positive and negative probe sets at various PLIER expression level cutoff values. The results are summarized in Supplementary Figure S1 and show that a PLIER expression level of 30is close to the cutoff that results in the least amount of falsely called probe sets at DABG P values of <0.05. A higher cutoff level close to PLIER expression 70 seems to result in the least amount of falsely called probe sets at the more stringent DABG *P* value of <0.01. All values were then imported into Omniviz v3.9 (Omniviz) software for further analysis. For each probe set, the geometric mean of the hybridization intensities of all samples from the patients was calculated with expression values of <30 set to 30 (close to the optimal cutoff with least amount of falsely called probe sets at DABG *P* value of <0.05).

The expression level of each probe set in every sample was determined relative to the geometric mean and logarithmically transformed (base 2 of scale) to ascribe equal weight to gene expression levels. Deviation from the geometric mean reflects differential probe set expression. Pearson's correlation plots were generated using all probe sets that differed 4-fold from the geometric mean in at least one sample (97175 probe sets in total, Figure 1) or with DABG *P*<0.01 in at least five samples (yielding virtually identical similar results, data not shown). Ordering of samples is done according to the algorithm present in Omniviz software as described [20]. This method reveals patterns of homologous samples based on Pearson's correlation. The ordering algorithm sorts all samples into correlated blocks through an iterative process and starts with the most highly correlated pair of samples. Each sample is joined to a block, resulting in a correlation trend within a block. The most correlated samples are at the center of each block. The blocks are then positioned along the diagonal of the plot in a similar ordered manner.

Splice variant detection

We used pattern-based correlation (PAC) as an algorithm to identify differentially regulated splice variants. PAC predicts the expression of a probe set in a given sample by the product of its metaprobe set level (a metaprobe set is a collection of probe sets that belong to the same transcript; the metaprobe set level is the calculated transcript level based on the expression level of these probe sets) and the probe set/transcript ratio of all samples:

$$\mathrm{Exp}_{a,c} = \mathrm{Tr}_{b,c}\mathrm{Ex}_{ave\text{-}a} / \mathrm{Tr}_{ave\text{-}b}$$

where $\mathrm{Exp}_{a,c}$ is the predicted expression of probe set a in sample c, $\mathrm{Tr}_{b,c}$ is the calculated metaprobe set level of transcript b (of which probe set a is part) in sample c, $\mathrm{Ex}_{ave\text{-}a}$ is the measured expression average of probe set a in all samples, and $\mathrm{Tr}_{ave\text{-}b}$ is the expression average of transcript b in all samples. In absence of alternative splicing or when a similar ratio of alternative splicing is observed in all samples, the predicted expression value should be identical to the measured PLIER expression levels:

$$\mathrm{Exm}_{a,c}\text{-}\mathrm{Exp}_{a,c} = 0$$

where $\mathrm{Exm}_{a,c}$ is the measured PLIER expression data from the array. Any deviation from 0in this formula is a predictor for alternative splicing: negative values predict the exon is spliced out in a given sample; positive values predict the exon is spliced in. PAC values were calculated using $\log^2$ transformed expression data.

Because splice variant detection requires an accurate estimation of metaprobe sets, we used two independent approaches to calculate metaprobe set levels. The first metaprobe set levels were calculated using ExACT 1.0software based on probe sets determined by Affymetrix. The second metaprobe set calculations required two iterations: We first determined metaprobe set levels by averaging all probe sets with PLIER expression levels >30, >50, or >80. We next hypothesized that differentially spliced exons will result in a metaprobe set level that is lower than when calculated using constitutive exons only. For example, an exon that is spliced out in subgroup A can reduce its metaprobe set level so that constitutive exons are identified as exons that are differentially spliced-out in subgroup B. Therefore, transcript levels should be calculated only using constitutively incorporated (i.e., not differentially spliced between defined subgroups) exons. We defined those constitutive exons (probe sets) as those that are highly correlated (correlation coefficient >0.7, >0.8, or >0.9) with the first round transcript calculations. A total of five metaprobe set calculations were done using cutoff values: (*a*) PLIER 50, correlation 0.8; (*b*) PLIER 30, correlation 0.8; (*c*) PLIER 80, correlation 0.8; (*d*) PLIER 50, correlation 0.7; and (*e*) PLIER 50, correlation 0.9. This two-step metaprobe set calculation not only excludes differentially spliced exons but also excludes "nonlinear" probe sets (probe sets that are outside the linear detection range of arrays) and "a-specific" probe sets (probe sets that bear no relation to its transcript).

Statistical analysis was done using standard t tests. Identical filtering and statistical analysis was done on 10 randomized groups to test for type I errors and estimate the false-discovery rate.

Reverse transcription-PCR

Candidate differentially regulated splice variants identified by PAC analysis were analyzed by reverse transcription-PCR (RT-PCR) to confirm differential regulation. All RT-PCR experiments were done on cDNA that was independently reverse transcribed from the cDNA that was used for array hybridization. rRNA (0.5 μg)-depleted (ribominus RNA) total RNA (the remainder of RNA that was used for array hybridization) was reverse transcribed for 1 h at 42°C in the presence of 200 units of Superscript II, 50 ng T7-(N)6 primers, 0.5 mmol/L deoxynucleotide triphosphates, 10mmol/L DTT, and RNase inhibitor. Primers were designed using Primer3[5] and are listed in Supplementary Table S2. Amplified PCR products from novel exon analysis were sequence verified using the Big Dye Terminator Cycle Sequencing kit (Applied Biosystems). Reactions were run on an ABI 3100 genetic analyzer.

Results

Human Exon arrays performance and unsupervised clustering

In this study, we performed exon level expression profiling to identify differentially expressed splice variants in glial brain tumors. Profiling was done using Human Exon 1.0 Arrays (Affymetrix), a novel platform that determines the expression of virtually all exons present in the human genome. These arrays are designed to target all well-annotated (RefSeq) exons (*core* exons), less well-characterized exons [e.g., derived from unique EST sequences that are not included in the RefSeq database (*extended* exons)] and all predicted exons (*full* exons) for which no expression data is present in public domain databases. In total, ~1.4 million probe sets (a set of up to four oligonucleotide probes that examines the expression of a single exon) are spotted on Human Exon 1.0 arrays: 284,000 *core*, 523,000 *extended*, and 580,000 *full* probe sets. Multiple probe sets may be directed against the same exon, thus, allowing identification of alternative splice-acceptor or splice-donor sites. Exon arrays also allow calculation of whole-transcript levels based on the expression level of probe sets that belong to the same transcript. Calculated transcript levels are called metaprobe set levels. In our experiments, the DABG significant expression (P<0.01) of 23.7 ± 4.5% of all 1.4 million probe sets were detected. *Core* exons are detected at higher signal intensities than *extended* and *full* exons (Supplementary Figure S2). Individual sample performance for all array quality control variables is stated in Supplementary Table S3.

This platform has thus far not been characterized, and we therefore first validated the performance of these arrays using unsupervised clustering analysis. Unsupervised clustering was done using probe sets with PLIER expression levels of >30that differed 4-fold from the geometric mean in at least one sample (Figure 1). A first subgroup (I) consists of all control samples and GBM 77, a sample that contained a low amount (<10%) of tumor. A second subgroup (II) consists of most (2 0of 22) of the oligodendrogliomas with LOH on 1p and 19q. The final subgroup (III) predominantly (25 of 27) consists of glioblastomas but also includes two oligodendrogliomas with 1p and 19q LOH (OD20 and OD170). Interestingly, OD20 also did not cluster with the majority of oligodendrogliomas with 1p/19q LOH using expression profiling on HU133 plus 2 microarrays [18]. Identical subgroups were identified by principle components analysis, using all core probe sets or core metaprobe sets (Supplementary Figure S2). Unsupervised clustering therefore indicates that exon expression profiling can identify brain tumor subgroups based on their histologic appearance. Our data therefore confirm the observation that histologically defined glioma subgroups are molecularly distinct (for review, see ref. 21) and indicates that, on a global scale, this novel platform performs similar to other expression profiling platforms.

Identification of differentially regulated splice variants

We next examined whether Human Exon arrays can detect glioma subgroup-specific splice variants. The identification of splice variants was done using PAC. PAC values represent a predicted level of expression for each probe set. Therefore, differences between PAC and

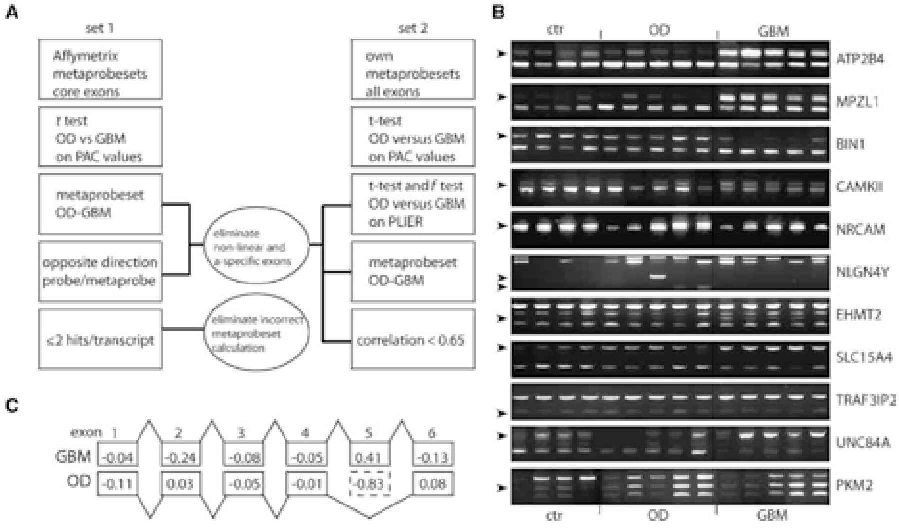

A
set 1
Affymetrix metaprobesets core exons
f test OD vs GBM on PAC values
metaprobeset OD-GBM
opposite direction probe/metaprobe
≤2 hits/transcript
eliminate non-linear and a-specific exons
eliminate incorrect metaprobeset calculation
set 2
own metaprobesets all exons
t-test OD versus GBM on PAC values
t-test and f test OD versus GBM on PLIER
metaprobeset OD-GBM
correlation < 0.65
B
ctr
OD
GBM
ATP2B4
MPZL1
BIN1
CAMKII
NRCAM
NLGN4Y
EHMT2
SLC15A4
TRAF3IP3
UNC84A
PKM2
C
exon 1 2 3 4 5 6
GBM -0.04 -0.24 -0.08 -0.05 0.41 -0.13
OD -0.11 0.03 -0.05 -0.01 -0.83 0.08

Figure 2: Identification of differentially expressed splice variants. A, summary of filtering steps used to identify 49 and 254 to 459 candidate differentially expressed exons, see also Table 1. B, RT-PCR of identified candidates using exon-spanning primers. *ATP2B4, CaMKII, NLGN4Y*, and *UNC84A* were confirmed hits identified in set 1. *BIN1, MPZL1*, and *NRCAM* were confirmed hits from sets 1 and 2. Other candidates were confirmed from set 2. In *NLGN4Y,* an exon 5' to the exon identified by PAC also shows alternative splicing, although this exon (exon 3) does not seem to be differentially expressed between oligodendrogliomas and glioblastomas. Top arrowhead, transcripts lacking only exon 4; bottom arrowhead, transcripts lacking both exons 3 and 4. RT-PCR products of *PKM2* were digested with pstI: the differentially spliced exon is mutually exclusive with a 5' exon of identical length. This exon however does not contain a pstI restriction site. C, model of alternative splicing of *MPLZ1*. In oligodendrogliomas, exon 5 is spliced out, identified by PAC analysis, and confirmed by RT-PCR. PAC values are stated in the represented exons. *OD*, oligodendrogliomas; *GBM*, glioblastomas.

expression values are indicative for alternative splicing. Negative values predict that the exon is, compared with the other 53 samples, being spliced out. However, PAC requires a complete linearity of all probe sets within a single transcript: if a transcript is up-regulated 2-fold in one subgroup, all of the probe sets that belong to this transcript should be up-regulated exactly 2-fold. Any probe set that does not exhibit this linearity in expression detection (nonlinear probe sets) or bear no correlation whatsoever with its native transcript (a-specific probe sets) will be identified as a false positive differentially spliced candidate. Examples of such nonlinear and a-specific probe sets are shown in Supplementary Figure S3. Any strategy to identify differentially expressed splice variants therefore requires filtering out nonlinear and a-specific probe sets.

Table 1: Filtering steps used to identify candidate differentially expressed exons

First strategy	Probe sets	PAC values	PAC *t* test	Transcr GMB ~ OD	Diff exp ex-tr	<3 ex/tr
Affy metaprobesets core	286,000	188,419	7,776	5,934	65	49
Random sampling	286,000	188,419	16	16	0	0
Range			6-35	6-35		
FDR			0.20%	0.30%		

Second strategy	Probe	PAC	PAC	PLIER	PLIER	Correlation	Expr.	Overlap
	sets	values	*t* test	*t* test	*f* test			
PLIER 50 corr 0.8	1,400,000	622,971	33,580	7,754	4,696	443	414	
Random sampling	1,400,000	622,971	2,012	20	7.5	1.8		
Range			1150–3203	2–56	0–20	0–7		
FDR			6%	0.25%	0.15%	0.40%		
PLIER 50 corr 0.9	1,400,000	338,970	22,190	5,364	3,199	267	254	208/254
PLIER 50 corr 0.7	1,400,000	850,382	52,820	9,826	5,737	492	459	388/459
PLIER 30 corr 0.8	1,400,000	701,535	34,799	7,016	4,034	418	390	360/390
PLIER 80 corr 0.8	1,400,000	567,410	32,294	8,358	4,989	418	397	346/397

NOTE: Our first strategy made use of core exons only using metaprobe sets predetermined by Affymetrix. For our second strategy, we calculated PAC values using recalculated metaprobe set expression levels (as outlined in Materials and Methods) with metaprobe set levels and the subsequent PAC values being recalculated at various probe set inclusion criteria. PAC values represent the number of probe sets in which PAC values could be calculated, omitting all probe sets with absent metaprobe set levels. *Transcript GBM ~ OD*: all probe sets in which metaprobe set levels differed <3-fold between oligodendrogliomas and glioblastomas. *Diff exp ex-tr*: remaining candidates were further selected by probe sets in which the direction of expression is differential between probe sets and metaprobe sets. If the average probe set level expression in OD>GBM, then the average metaprobe set expression should be OD<GBM and vice versa. This filter is likely to exclude many true positive candidates but will also rigorously exclude most nonlinear and a-specific candidates. <3 *ex/tr*: all probe sets with three or more candidates within a single transcript were excluded because these are likely to be false positive candidates due to incorrect metaprobe set calculation. *Correlation*: probe sets with high correlation between probe set and metaprobe set expression were excluded (correlation coefficient > 0.65). This filter is based on the hypothesis that regulated splice variants are expected to havean exon/transcript correlation that is less than constitutively incorporated exons. Overlap: number of candidates that were also identified using PLIER 50, correlation 0.8.

We adopted two independent strategies to identify candidate splice variants that are differentially regulated between oligodendrogliomas and glioblastomas. In the first strategy, we calculated PAC values for every probe set in all samples using metaprobe sets predetermined by Affymetrix. For our second strategy, we calculated PAC values using recalculated metaprobe set expression levels (see Materials and Methods) with metaprobe set levels (and subsequent PAC values) derived at varying PLIER expression level and/or correlation coefficient cutoff values. We then aimed to exclude nonlinear and a-specific probe sets using the filtering steps outlined in Figure 2 and Table 1. These filtering steps resulted in final set of 49 (first strategy) and 254 to 459 candidate differentially regulated splice variants (second strategy). Table 1 summarizes the results at each step in our strategy to identify candidate splice variants. Supplementary Table S4 contains a list of all candidates.

To estimate the false discovery rate, we randomly assigned a group number to each tissue sample and then repeated the filtering and statistical analysis (Table 1). This scrambling procedure was repeated 10 times and failed to identify any candidate splice variant in the first strategy and 1.8 candidates splice variants (range, 0–7) in the second strategy.

Altering the variables used for metaprobe set calculation often resulted in significant overlap between candidates identified: many candidates identified at cutoff values PLIER 50, and correlation coefficient 0.8 are also found when the PLIER expression cutoff is reduced to 30 (88%), increased to 80 (83%), or the correlation cutoff is reduced to 0.7 (93%). In contrast, increasing the correlation cutoff to 0.9 results in a set of candidates that contains only 50% of the probe sets identified by PLIER 50correlation 0.8 with 46 additional probe sets identified.

We did RT-PCR using exon spanning primers to confirm the differential expression of candidate splice variants. RT-PCR was done on 15 candidates from the first screen and 21 candidates from the second screen (PLIER 50, correlation 0.8). RT-PCR candidates were randomly selected from the total number of candidates but omitted candidates with alternative 5'- or3'-end exons. We confirmed 7 of 15 (47%) from the first screen and 7 of 21 (33%) from the second analysis (Figure 2). Three of the confirmed candidates were identified in both analysis; the total number of differentially expressed splice variants equaled 11. All differentially expressed splice variants belonged to the *core* probe set list. Public domain databases (EMSEMBL, UCSC, HOLLYWOOD) also indicated that most (9 of 11) RT-PCR confirmed candidates are subject to alternative splicing. It is possible that the percentage of regulated splice variants is higher than the RT-PCR–confirmed 47% to 33%: rare splice variants or splice variants that show only minor differential regulation may not have been detected by RT-PCR. Nevertheless, our results show that exon level expression profiling can identify splice variants that are differentially regulated between histologically defined subgroups of gliomas.

Identification of novel exons

We finally examined whether Human Exon arrays can be used to identify novel exons. We screened for novel exons using the *full* probe set list (580,000 probe sets) because all *full* exons lack evidence for expression in public domain databases. *Full* probe sets are composed of exons that can be predicted (e.g., based on the presence of consensus splice acceptor and donor sites) and of sequences that are conserved between human, mouse, and rat. Candidate novel exons met the following criteria (see Figure 3): (*a*) show significant expression (PLIER expression levels ≥50);(*b*) are part of a core metaprobe set as many *full* probe sets are part of poorly characterized and single-exon transcripts; and (*c*) should have a high (>0.8) correlation coefficient with its metaprobe set (i.e., the probe set is highly expressed in those samples in which the metaprobe set is highly expressed). These criteria resulted in a final set of 715 *full* probe sets as candidate novel exons. More candidates are identified using less stringent criteria (exon/transcript correlation ≥0.7, identifies 1482 *full* exons). *In silico* analysis of the first 158 *full* probe sets confirmed that 127 of 158 (80%) are indeed novel exons; they are not present in the RefSeq database and no spliced EST has thus far been identified. Of the remaining probe sets, 18 of 158 (11%) were incorrectly annotated and are in fact part of a RefSeq gene, and 13 of 158 (8%) were identified as part of (rare) spliced ESTs.

We next used RT-PCR to verify that candidate novel exons are indeed expressed as part of a known gene. Primers were designed to span >2 kb intronic sequence to exclude false positives due to amplification of genomic DNA or pre-mRNA sequence. RT-PCR confirmed the expression of 6 of 9 (67%) *full* exons, for which no expression data is present in public domain databases (Figure 3B). These PCRs were done using one of the primers within the novel exon. We used direct sequencing to confirm that the novel exons are indeed expressed as part of a known transcript and not due to amplification of a-specific products (Figure 3C). In all cases, products that contain the (RefSeq) known flanking exons and the novel exon were identified. Furthermore, direct sequencing enabled us to confirm the presence of consensus splice acceptor/donor sequences surrounding the novel exons.

RT-PCR also confirmed the expression of 3 of 3 (100%) full exons that, in public domain databases, were part of rare spliced ESTs. All three exons could be identified in all examined samples. For KDHRBS2 and DTNA, RT-PCR was done using exon-spanning primers; for PDE1C, RT-PCR was done with the forward primer in the candidate novel exon because the novel exon may represent a novel 5' exon. Identification of transcripts that have incorporated the novel exon using exon-spanning primers suggests that a significant percentage of transcripts have incorporated the full exon in adult brain (Figure 3B).

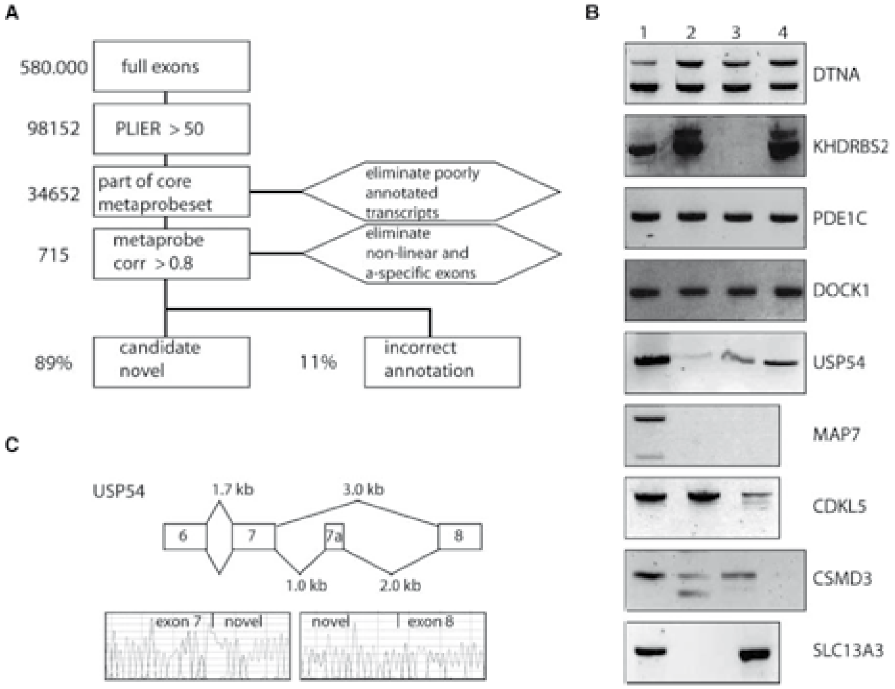


Figure 3: Identification of novel exons by exon level expression profiling. (A), filtering steps used to identify 715 candidate novel exons. Candidate novel exons are expressed (PLIER) >50 as part of a well-characterized transcript and have a correlation coefficient of >0.8 with its transcript. (B), RT-PCR of a subset of identified candidates on independent samples (lanes 1–4). *DTNA, KHDRBS2*, and *PDE1C* were identified as part of a rare splice variant in public domain databases. Expression of *DTNA* and *KHDRBS2 full* exons was confirmed using exon spanning primers, other full exons were confirmed using one primer within the candidate novel exon. Products were sequence verified to exclude a-specific amplifications. (C),model of splicing of the nove lidentified exon in *USP54*. Direct sequencing confirmed the presence of the novel exon expressed as part of *USP54*.

Discussion

In this study, we did exon level expression profiling on a set of glial brain tumors. To our knowledge, we are among the first to describe the use of Human Exon 1.0arrays as an expression profiling platform. Our results show that Human Exon arrays can identify subgroups of gliomas based on their histologic appearance and genetic aberrations, can identify differentially expressed splice variants, and can identify novel exons.

The molecular subgroups identified using exon level expression profiling is highly similar to the subgroups that are identified in other studies using 3' biased expression profiling [18, 22–27]. Our data therefore confirm the observation that histologically defined glioma subgroups are molecularly distinct (for review, see ref. 21). Furthermore, the similarity in glial tumor classification indicates that, at least on a global scale, this novel platform performs similar to other expression-profiling platforms.

The additional complexity of exon level expression profiling over transcript-level expression profiling is the ability to identify splice variants that are differentially expressed between tumor subgroups. Our data indicate that the identification of differentially expressed splice variants requires rigorous filtering steps to exclude nonlinear and a-specific probe sets. In the two independent approaches adopted by us, we identified 49 and 254 to 459 candidate splice variants that are differentially expressed between OD and GBM. The list of candidates differs significantly between the two approaches. Furthermore, candidates identified by our second approach (recalculated metaprobe set level) are dependent on the inclusion criteria used to recalculate metaprobeset levels. It remains to be determined which variables are optimal for spice variant detection. However, all candidate lists generated by our second approach contain a similar percentage of known splicing events (~12%; range, 10.4–13.8%; see Supplementary Table S4) as determined by screening public domain databases on a subset of candidates.

RT-PCR confirmed the differential regulation of a subset of these candidate splice variants. The select number of differentially expressed splice variants identified by us may reflect the similarity in splice variant expression between OD and GBM. Indeed, a limited number (591) of differentially expressed splice variants between mouse brain and immune tissue were identified by Ule and coworkers using exon-junction arrays [28]. In contrast, experimental evidence exists for the regulated expression of a large number of splice variants: many splice variants show some degree of tissue specificity [29-31]. It is therefore also possible that the strong filtering used in this study has led to the identification of only a subset of differentially regulated splice variants.

The differential expression of splice variants between two tumor subtypes may be caused by a differential expression of proteins that regulate alternative splicing. Indeed, a large number of proteins have been identified to play a role in the regulation of alternative splicing (for review, see refs. 32-34). However, the expression of glioma subgroup-specific splice variants may also be a result of genetic changes. For example, glioblastomas with *EGFR* amplifications frequently carry an intragenic deletion of exons 2 through 7, resulting in expression of the tumor specific, constitutively active *EGFRvIII* isoform [35]. Such aberrant splice isoforms have been shown to play a role in the initiation and/or progression of glial brain tumors [6]. Identifying glioma-specific splice variants may therefore help identify the causative genetic changes of glial brain tumors.

Apart from exon expression arrays, other techniques have been used to analyze splice variant expression. These include exon- junction arrays [36], RNA-mediated annealing, selection and ligation [37] and digital polony (polymerase colony) exon profiling [38]. Recently, arrays containing a combination of exon expression and exon junction probes have also been used to identify alternative splicing events [39,40]. Although all approaches can detect alternative splicing events, many are limited either by screening on a predetermined set of exon-junctions or screening on a per-gene base. Our data shows that exon expression profiling is a suitable alternative for genome-wide screening of regulated splicing events between two distinct subgroups.

Our study has also identified 715 full exons that are expressed as part of a well-annotated transcript. In silico analysis (screening public domain databases) of a subset of candidates indicated that charges. This article must therefore be hereby marked advertisement in accordance 80% are indeed novel exons; they are not present in the RefSeq database and no spliced EST has thus far been identified. We confirmed the expression of ~67%, suggesting a total of ~446 (0.78*0.8*715) novel exons are expressed as part of a well-annotated transcript. Candidates that were not confirmed by RT-PCR (33%) may be falsely identified, for example when the exon array detects unspliced, pre-mRNA species (see e.g., ref. 41). The majority (5 of 6) of RT-PCR confirmed novel exons are expressed in normal adult human brain, indicating they are not aberrant, cancer-specific splice isoforms. Furthermore, most (5 of 6) of the RT-PCR confirmed novel exons result in changes at the protein level: the novel exons are often found within the protein coding region.

Many of the full probe sets on the Human Exon arrays are based on evolutionary sequence conservation between human, mouse, and rat. Other studies have also found novel exons based on such sequence conservation. For example, ~ 150candidate novel human exons were identified in a screen based on the expression of ESTs in mouse/rat [42]. Furthermore, a bioinformatical approach using sequence conservation has identified up to 2,300 novel, rodent-specific exons [43]. In a separate study, bioinformatical analysis based on exon expression profiles from adult mouse tissue has suggested the presence of a large number (40–70,000) of novel exons [44]. Although our study identified fewer novel exons, both studies argue for the presence of novel exons in human/mouse genomes and that such novel exons can be identified using exon expression profiling.

In summary, our results indicate that exon level expression profiling can be used to molecularly classify brain tumor subgroups, can identify differentially regulated splice variants, and can identify novel exons. See also appendices 3 and 4.

References

1. Legler JM, Ries LA, Smith MA, et al. *Cancer surveillance series [corrected]: brain and other central nervous system cancers: recent trends in incidence and mortality.* J Natl Cancer Inst 1999;**91**:1382–90.
2. Macdonald DR. *New frontiers in the treatment of malignant glioma.* Semin Oncol 2003;**30**:72–6.
3. Kleihues P, Cavenee WK. *World Health Organization Classification of Tumours of the Nervous System.* Lyon: WHO/IARC; 2000.
4. Hegi ME, Diserens AC, Gorlia T, et al. *MGMT gene silencing and benefit from temozolomide in glioblasto ma.* N Engl J Med 2005;**352**:997–1003.
5. Ohgaki H, Kleihues P. *Epidemiology and etiology of gliomas.* Acta Neuropathol Berl 2005;**109**:93–108.
6. Grabowski PJ, Black DL. *Alternative RNA splicing in the nervous system.* Prog Neurobiol 2001;**65**:289–308. 162 Suppl:S17–21.
7. Nagane M, Lin H, Cavenee WK, Huang HJ. *Aberrant receptor signaling in human malignant gliomas: mech anisms and therapeutic implications.* Cancer Lett 2001;
8. Holland EC, Hively WP, DePinho RA, Varmus HE. *A constitutively active epidermal growth factor receptor cooperates with disruption of G1 cell-cycle arrest pathways to induce glioma-like lesions in mice.* Genes Dev 1998;**12**:3675–85.
9. Mellinghoff IK, Wang MY, Vivanco I, et al. Mole-16. Xu X, Joh HD, Pin S, et al. *Mole cular determinants of the response of glioblastomas to EGFR kinase inhibitors.* N Engl J Med 2005;**353**: 2012–24.
10. Frederick L, Wang XY, Eley G, James CD. *Diversity and frequency of epidermal growth factor receptor mutations in human glioblastomas.* Cancer Res 2000;**60**: 1383–7. 3013–23.
11. Shi E, Kan M, Xu J, Wang F, Hou J, McKeehan WL. *Control of fibroblast growth factor receptor kinase signal transduction by heterodimerization of combina torial splice variants.* Mol Cell Biol 1993;**13**:3907–18.
12. Wang F, Kan M, Yan G, Xu J, McKeehan WL. *Alternately spliced NH2-terminal immunoglobulin-like Loop I in the ectodomain of the fibroblast growth factor (FGF) receptor 1 lowers affinity for both heparin and FGF-1.* J Biol Chem 1995;**270**:10231–5.
13. Chunduru S, Kawami H, Gullick R, Monacci WJ, Dougherty G, Cutler ML. *Identification of an alterna tively spliced RNA for the Ras suppressor RSU-1 human gliomas.* J Neurooncol 2002;**60**:201–11.
14. Simon M, Koster G, Ludwig M, et al. *Alternative splicing of the p15 cdk inhibitor in glioblastoma multiforme.* Acta Neuropathol Berl 2001;**102**:167–74.
15. Uno M, Oba-Shinjo SM, de Aguiar PH, et al. *Detection of somatic TP53 splice site mutations in diffuse astrocytomas.* Cancer Lett 2005;**224**:321–7.
16. *Expression of multiple larger-sized transcripts for several genes in oligoden drogliomas: potential markers for glioma subtype.* Cancer Lett 2001;**171**:67–77.
17. Hui L, Zhang X, Wu X, et al. *Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment*. Oncogene 2004;**23**:
18. French PJ, Swagemakers SMA, Nagel JHA, et al. *Gene expression profiles associated with treatment response in oligodendrogliomas.* Cancer Res 2005;**65**: 11335–44.
19. Schroeder A, Mueller O, Stocker S, et al. *The RIN: an RNA integrity number for assigning integrity values to RNA measurements.* BMC Mol Biol 2006;**7**:3.
20. Valk PJ, Verhaak RG, Beijen MA, et al. *Prognostically useful gene-expression profiles in acute myeloid leuke mia.* N Engl J Med 2004;**350**:1617–28.
21. Mischel PS, Cloughesy TF, Nelson SF. *DNA-micro-in array analysis of brain cancer: molecular classification for therapy.* Nat Rev Neurosci 2004;**5**:782–92.
22. Rickman DS, Bobek MP, Misek DE, et al. Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. Cancer Res 2001;61:6885–91.
23. Godard S, Getz G, Delorenzi M, et al. Classification of human astrocytic gliomas on the basis of gene expression: a correlated group of genes with angiogenic activity emerges as a strong predictor of subtypes. Cancer Res 2003;63:6613–25.
24. Nutt CL, Mani DR, Betensky RA, et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res 2003;63:1602–7.

25. Freije WA, Castro-Vargas FE, Fang Z, et al. *Gene expression profiling of gliomas strongly predicts survival.* Cancer Res 2004;**64**:6503–10.

26. Liang Y, Diehn M, Watson N, et al. *Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme.* Proc Natl Acad Sci U S A 2005;**102**:5814–9.

27. Phillips HS, Kharbanda S, Chen R, et al. *Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis.* Cancer Cell 2006;**9**:157–73.

28. Ule J, Ule A, Spencer J, et al. *Nova regulates brain-specific splicing to shape the synapse.* Nat Genet 2005; **37**:844–52.

29. Xu Q, Modrek B, Lee C. *Genome-wide detection of tissue-specific alternative splicing in the human transcriptome.* Nucleic Acids Res 2002;**30**:3754–66.

30. Yeo G, Holste D, Kreiman G, Burge CB. *Variation in alternative splicing across human tissues.* Genome Biol 2004;**5**:R74.

31. Pan Q, Shai O, Misquitta C, et al. *Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform.* Mol Cell 2004;**16**:929–41.

32. Black DL. *Mechanisms of alternative pre-messenger RNA splicing.* Annu Rev Biochem 2003;**72**:291–336.

33. Kalnina Z, Zayakin P, Silina K, Line A. *Alterations of pre-mRNA splicing in cancer.* Genes Chromosomes Cancer 2005;**42**:342–57.

34. Graveley BR. *Sorting out the complexity of SR protein functions.* RNA 2000;**6**:1197–211.

35. Frederick L, Eley G, Wang XY, James CD. *Analysis of genomic rearrangements associated with EGRFvIII expression suggests involvement of Alu repeat elements.* Neuro-oncol 2000;**2**:159–63.

36. Johnson JM, Castle J, Garrett-Engele P, et al. *Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.* Science 2003;**302**:2141–4.

37. Yeakley JM, Fan JB, Doucet D, et al. *Profiling alternative splicing on fiber-optic arrays.* Nat Biotechnol 2002;**20**:353–8.

38. Zhu J, Shendure J, Mitra RD, Church GM. *Singlemolecule profiling of alternative pre-mRNA splicing.* Science 2003;*301*:836–8.

39. Fehlbaum P, Guihal C, Bracco L, Cochet O. *A microarray configuration to quantify expression levels and relative abundance of splice variants.* Nucleic Acids Res 2005;**33**:e47.

40. Nagao K, Togawa N, Fujii K, et al. *Detecting tissue-specific alternative splicing and disease-associated aberrant splicing of the PTCH gene with exon junction microarrays.* Hum Mol Genet 2005;**14**:3379–88.

41. O'Connor V, Genin A, Davis S, et al. *Differential amplification of intron-containing transcripts reveals long term potentiation-associated up-regulation of specific Pde10A phosphodiesterase splice variants.* J Biol Chem 2004;**279**:15841–9.

42. Chen FC, Chen CJ, Ho JY, Chuang TJ. *Identification and evolutionary analysis of novel exons and alternative splicing events using cross-species EST-to-genome comparisons in human, mouse and rat.* BMC Bioinformatics 2006;**7**:136.

43. Nekrutenko A. *Identification of novel exons from rat-mouse comparisons.* J Mol Evol 2004;**59:**703–8.

44. Frey BJ, Mohammad N, Morris QD, et al. *Genome-wide analysis of mouse transcripts using exon micro-arrays and factor graphs.* Nat Genet 2005;**37**:991–6.

Chapter 7

Exon expression arrays as a tool to identify new cancer genes

Mieke Schutte[1]; Fons Elstrodt[1]; Elza Duijm[2]; Jord H.A. Nagel[1]; Antoinette Hollestelle[1]; Marijke Wasielewski[1]; ***Justine Peeters*[3]**; Peter van der Spek[3]; Peter A. Sillevis Smitt[2] and Pim J. French[2]

[1] Department of Medical Oncology, Josephine Nefkens Institute, Erasmus University Medical Center, Rotterdam, The Netherlands
[2] Department of Neurology, Josephine Nefkens Institute, Erasmus University Medical Center, Rotterdam, The Netherlands
[3] Department of Bioinformatics, Josephine Nefkens Institute, Erasmus University Medical Center, Rotterdam, The Netherlands

Submitted to Oncogene, 2008

Abstract

Identification of genes that are causally implicated in oncogenesis is a major goal in cancer research. An estimated 10-20% of the cancer-related gene mutations result in skipping of one or more exons in the encoded transcripts represent. Here we report on a strategy, termed Outlier Exon Screening (OES), to screen in a global fashion for such exon skipping events. OES uses exon-level expression profiles to identify outlier exons that are markedly lower expressed than predicted based on the expression level of their transcripts. As a proof-of-principle, we tested the OES strategy on human cancer samples of which the complete coding sequence of eight cancer genes had been screened for mutations. OES detected all seven exon-skipping mutants among 12 cancer cell lines. OES also identified exon-skipping mutants in clinical cancer specimens although detection was compromised due to heterogeneous transcript expression. Importantly, OES reduced the number of candidate genes or exons for subsequent mutational analysis by two to three orders of magnitude, with one-quarter of the identified outlier exons representing truly skipped exons. The performance of OES was further evaluated by analysis of 112 selected outlier exons. Two novel exon skipping events, two novel base changes and 21 previously reported base changes (SNPs) were identified. The single nucleotide substitutions were apparently detected because their affinity to the complementary probe set sequence on exon array is reduced. The identification of known and novel exon skipping events and base changes confirms the suitability of OES to identify candidate cancer genes.

Introduction

Cancer is driven by mutations in genes that control the proliferation of cells, their survival and their integrity. Screens aimed at identifying such cancer genes often use chromosomal location and/or functional properties to select candidates genes for sequencing [1-4]. Although many candidate cancer gene loci have been identified, a significant gene search effort remains to find the corresponding cancer gene. Other gene search strategies have focused on aberrant gene expression patterns to identify candidates. For example, gene mutants that result in premature termination codons were identified by screening for genes that were specifically expressed following chemical inhibition of nonsense mediated RNA decay [5]. Furthermore, fusion genes in prostate cancer were identified by screening for outliers in a large cohort of gene-expression profiles [6].

Human cancer gene mutations frequently result in the skipping of one or several exons from the encoded transcripts [7-9]. Exon-skipping mutations may be caused by nucleotide substitutions within the consensus splice sites or by deletions that span entire exons. In addition, exon-skipping mutations may be caused by relatively small intragenic insertions, deletions or duplications. Even though exon-skipping mutations represent an estimated 10-20% of all cancer-related gene mutations [4,9-12], no high throughput method has been available to screen for such mutations. Here, we describe Outlier Exon Screening (OES) as an approach to identify candidate cancer genes by screening for exon-skipping events in a global fashion. OES identified outlier exons can then be subject to detailed sequence analysis to identify underlying genetic mutations. As a proof-of-principle, we demonstrate the efficacy of the OES strategy on previously identified exon-skipping mutations in breast cancer cell lines and in clinical brain tumor samples. Our study also demonstrates that OES can identify novel exon skipping events and genetic changes in known cancer genes and in randomly-selected OES-identified outlier exons.

Results

Outlier Exon Screening (OES)

We developed a new approach to screen for exon-skipping events in human cancer samples, termed Outlier Exon Screening (OES). Briefly, exon-level expression profiles are generated using Affymetrix Human Exon Arrays, which determine the expression level of virtually all exons present in the human genome. The PAC (PAttern-based Correlation) algorithm is used to calculate the predicted expression level of each exon (or probe set), taking into account the overall expression of the transcript in that particular sample as also the exon/transcript expression ratio as determined from all samples. OES then identifies outlier exons by subtracting the PAC-predicted expression level of exons from their measured expression level, with OES values equaling zero when the measured expression level of an exon was similar to its predicted expression level (formulated in detail under Methods). OES effectively normalizes the variability in gene expression levels between samples and, in a single sample, normalizes the variability in signal intensity between probe sets of the same transcript (Figure 1).

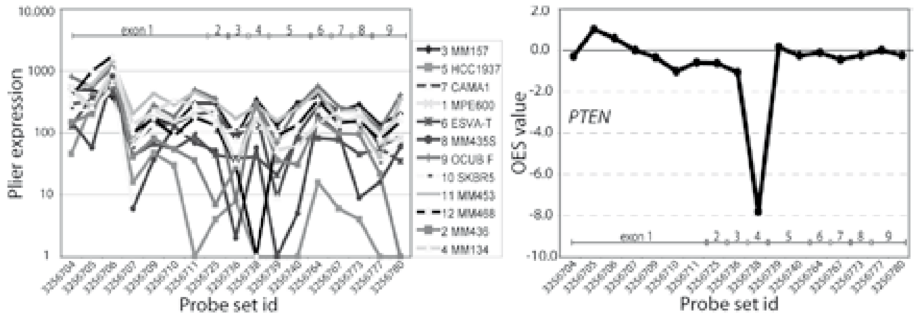


Figure 1: Outlier Exon Screening detection of an exon-skipping *PTEN* mutant (A) Normalized expression data of all exons within the *PTEN* gene. Each exon probe set is represented by a dot in the solid line; multiple probe sets may be directed against the same exon. The variability of expression levels between samples and within exons of a single sample makes it difficult to identify of exon skipping mutations. (B) OES normalizes the variability in gene expression levels between samples and, in a single sample, the variability in signal intensity between probe sets of the same transcript. OES calculation therefore allows rapid detection of skipping of *PTEN* exon 4 in breast cancer cell line MDA-MB-468 due to a *PTEN* c.253+1G>T splice site mutation that we previously had identified [16].

OES detects exon-skipping events in breast cancer cell lines

We tested the feasibility of the OES strategy on a panel of 12 human breast cancer cell lines that had been screened for mutations in seven tumor suppressor genes: *BRCA1, CDH1, MAP2K4, PTEN, p16, p53* and *RB1* [13-17], and unpublished results). Mutation analysis was performed by sequencing of the complete coding sequences of the genes and analysis of all mutations on both genomic gene fragments and transcripts. Together, the 12 cell lines contained seven gene mutants that should be detectable by OES, as they resulted in the skipping of eight exons from among four tumor suppressor genes (mutations are detailed in Supplementary Table 1). We have explored the OES strategy at different cut-off levels, identifying outlier exons that were expressed less than 16-fold, 8-fold, 4-fold, 2.8-fold and 2.5-fold than their predicted expression level (i.e. OES values of -4.0, -3.0, -2.0, -1.5 and -1.3, respectively). Outlier exons were identified without prior knowledge of the mutation data.

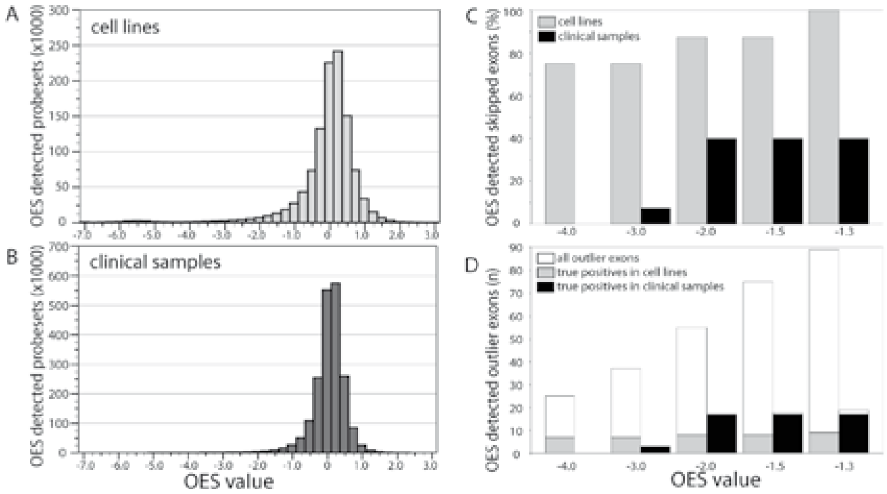


Figure 2: Performance of Outlier Exon Screening to detect exon-skipping mutants. (A) and (B) Total number of OES-detected outlier probe sets from among 290,000 core probe sets in 12 breast cancer cell lines and in 14 glioblastomas, respectively. (C) Number of skipped exons detected by OES as a percentage of all eight skipped exons present in the breast cancer cell lines, or as a percentage of the 36 skipped *EGFR* exons present in the glioblastomas (see Table 1). (D) Total number of outlier exons (true plus false positives) and number of true positive outlier exons detected by OES among the seven tumor suppressor genes and the *EGFR* oncogene. True positive outlier exons include all OES detected skipped exons and two missense mutations (*PTEN* c.274G>C in CAMA1, *MAP2K4* c.551C>G in MDA-MB-134VI).

From the total of 3.4 million core probe sets that we assayed for the 12 cell lines (290,000 core probe sets per sample), OES identified 21,151 (0.6%) outlier probe sets at OES value -4.0 and 94,590 (2.8%) outlier probe sets at OES value -1.3 (Figure 2A). OES of the seven fully characterized tumor suppressor genes in the 12 cell lines involved analysis of 1200 exons (1752 probe sets). OES correctly detected six of the eight skipped exons when using OES value -4.0, seven skipped exons were detected at OES value -2.0 and all eight skipped exons were detected at OES value -1.3 (Figure 2C). Importantly, the number of false positive outlier exons was substantially reduced at OES value -4.0 as compared to OES value -1.3, resulting in an increase of the true positive rate from 9% to 24% of the identified outlier exons (Figure 2D). In this respect, it is important to note that reduction of the number of false positive candidate genes may initially be far more beneficial for a gene search project than accurate identification of all true positive outlier exons. Together, our results show that the OES strategy is reliable in detecting exon-skipping mutants in cancer cell lines.

OES performance in samples with heterogeneous transcript expression

Similar to other genetic screening methods, OES is most suited to detect homozygous genetic changes. For example, the presence of 50% wild-type transcript results in a 2 fold lower exon level expression observed than predicted with corresponding OES value of -1.0. The somewhat compromised detection of skipped exons at OES value -4.0 as compared to OES value -1.3 (i.e. six vs. all eight skipped exons) in our panel of breast cancer cell lines therefore may have been caused by the expression of a second aberrant transcript that still includes (part of) the exon. Indeed, a second *CDH1* transcript length of minor intensity was detected in CAMA-1 (Figure 3A), the splice site mutant that had been detected only at OES value -1.3.

To further asses the performance of OES in samples with heterogeneous (wild-type and mutant) transcript expression, we performed OES on 14 clinical glioblastoma specimens (selected to contain >70% tumor nuclei) that had genomic amplifications of the *EGFR* oncogene. Glioblastomas with *EGFR* amplifications frequently carry an intragenic deletion of exons 2 through 7, resulting in expression of the constitutively active *EGFRvIII* isoform [8,21]. However, glioblastomas expressing the *EGFRvIII* isoform also frequently express wild-type *EGFR* transcripts. This heterogeneous *EGFR* expression is related to amplification of the *EGFR* locus prior to the deletion of exons [22], although non-malignant cells in the glioblastoma specimens may also express *EGFR*. Of the fourteen glioblastoma samples used in this study, six expressed *EGFRvIII* (a total of 36 skipped exons) of which five also expressed significant levels of wild-type *EGFR* transcripts as determined by quantitative Real-Time PCR (qPCR) (Figure 3B) (insufficient RNA remained of the sixth sample with *EGFRvIII* expression to perform qPCR).

From the total of 4.1 million core probe sets that we assayed for these 14 samples (290,000 core probe sets per sample), OES identified 1,646 (0.04%) outlier probe sets at OES value -4.0 and 39,936 (1.0%) outlier probe sets at OES value -1.3 (Figure 1B). Outlier exons in glioblastoma samples were identified without prior knowledge of the mutation data. OES thus identified three to ten-fold less outlier exons in the glioblastomas as compared to the breast cancer cell lines (Figure 1A). This smaller number of outlier exons in the glioblastomas may be related to their homogeneous histopathology and their highly similar gene expression profiles [18,23], to the presence of non-neoplastic cells in the tumor samples, or may reflect sampling biases due to small cohort sizes.

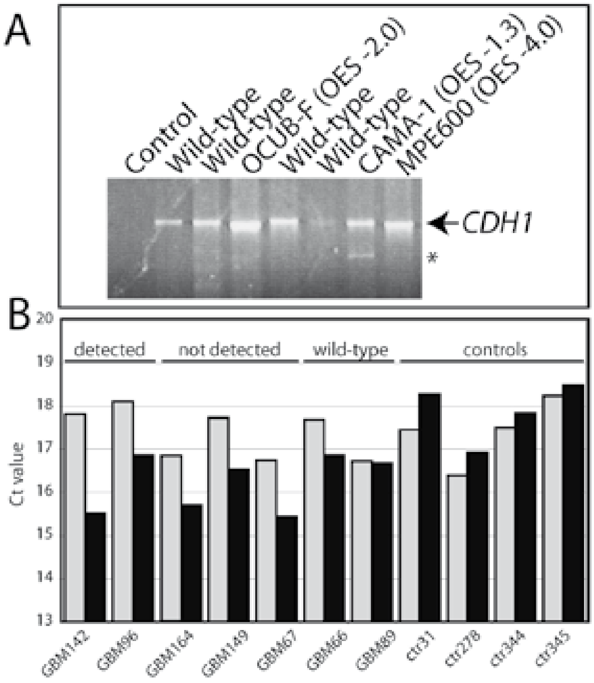


Figure 3: Compromised Outlier Exon Screening due to heterogeneous transcript expression. Skipping of *CDH1* exon 11 in breast cancer cell line CAMA-1 was only detected at OES value -1.3, likely due to expression of a second aberrant transcript variant (*) that was detected by conventional RT-PCR. (B) Expression of *EGFR* transcripts was detected in glioblastoma samples by Real-Time RT-PCR, using primers designed to anneal inside the exon 2-7 deletion region of the EGFRvIII isoform (gray bars) or outside the deletion region (black bars). Differences in Ct values between the two transcript fragments are indicative for *EGFRvIII* isoform expression levels. All five samples with the *EGFRvIII* isoform also expressed significant amounts of wild-type *EGFR* transcripts, likely compromising outlier detection by OES (indicated by “detected” and “not detected”). Wild-type, samples with normal transcripts; Controls, non-malignant brain specimens.

OES of the *EGFR* gene in the 14 glioblastomas involved the analysis of 392 exons (434 probe sets). OES detected 2/6 EGFRvIII expressing tumors (12 of the 36 skipped exons) at OES values -2.0 and lower (Figure 2C). Of the two glioblastomas with *EGFRvIII* that had been detected by OES, one had significantly (i.e. >5 fold) more mutant than wild-type *EGFR* transcripts; Ct value difference >2 between qPCR fragments inside (measuring only wild-type *EGFR* transcripts) and outside (measuring both wild-type and *EGFRvIII* transcripts) the EGFR exon 2-7 deletion region (Figure 3B). The other glioblastoma had a similar expression level difference between wt and *EGFRvIII* transcripts (similar Ct value difference of ~1.5) as the three glioblastomas that had not been detected by OES, but had lower overall *EGFR* transcript levels. It appears that OES detection of the *EGFRvIII* isoform is determined by the overall expression level of *EGFR* transcripts in combination with the ratio of *EGFRvIII* and wild-type *EGFR* transcripts, where samples with too high *EGFR* transcript levels may escape OES detection due to saturation of the probe sets involved. These results show that the OES strategy can detect exon-skipping mutants in clinical cancer specimens if the ratio mutant/wild-type transcript level is high and when probe sets are within the linear detection range of the microarray.

OES performance in detecting recurrent outlier exons

OES performance can also be challenged by recurrent outlier exons. Such frequently skipped exons will result in an underestimation of the exon/transcript ratio in the OES algorithm and so increase OES values. We therefore evaluated the performance of OES in detecting recurrent outlier exons by reiterated replacement of *EGFRvIII* expressing samples with samples that expressed only wild-type *EGFR* (Figure 4A). When 6/14 samples express *EGFRvIII*, a deletion of exons 2-7 is not OES detected in GBM67. OES values indeed decreased with decreasing ratios of wild-type versus mutant samples. However, the decrease was relatively small and resulted in the identification of only one of the six deleted exons once the ratio had dropped to 1 mutant among 14 samples. We also simulated OES detection of recurrent mutations with two breast cancer cell lines, of which HCC1937 had skipped *RB1* exon 22, and we were already able to identify the mutant from among two samples up to even five mutants from among six samples (Figure 4B). These simulation experiments indicate that OES performs well in identifying recurrent exon-skipping mutations.

Detection of nucleotide substitutions and novel genetic changes by OES

The performance of OES was further evaluated by analysis of outlier exons selected from all candidates at OES value ≤-2.0 in breast cancer cell lines (n=44) and clinical glioblastoma samples (n=68) respectively. Sequence analysis of PCR amplified outlier exons identified 2 novel exon skipping events and 2 novel genetic base changes in glioblastoma samples, as well as a number of previously reported base changes (homozygous SNPs) in breast cancer cell lines (n=5) and glioblastomas (n=16).

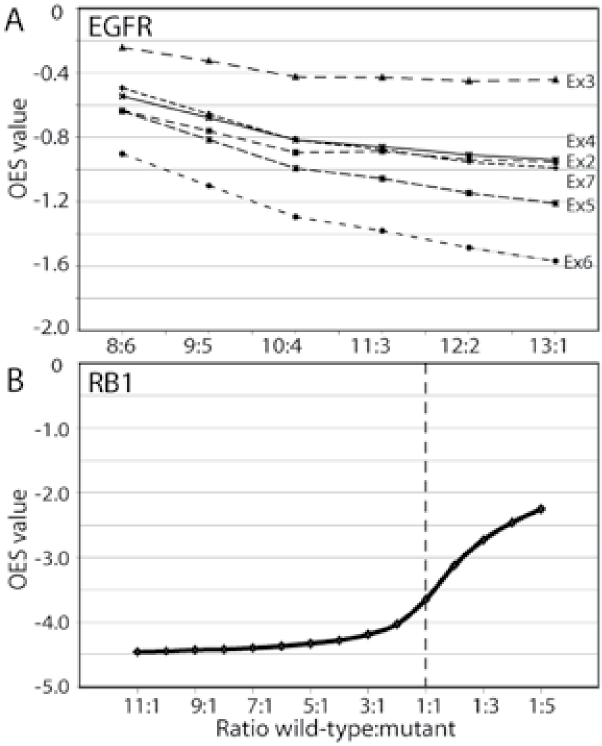
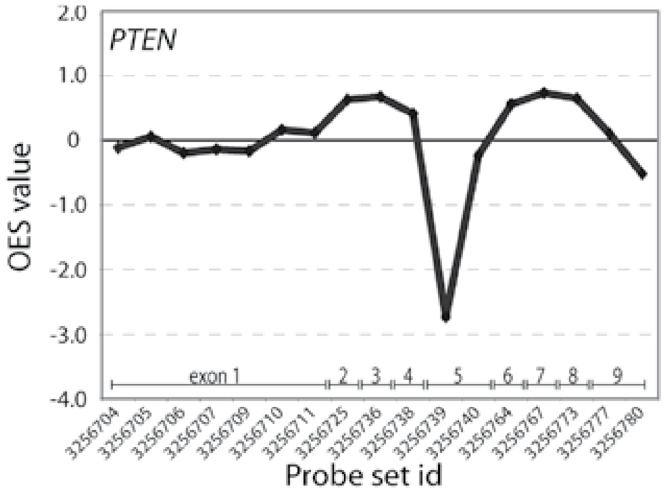


Figure 4: Identification of nucleotide substitutions by OES. (A) OES predicts skipping of the 5' end of *PTEN* exon 5 in the CAMA1 breast cancer cell line. This cell line contains a nucleotide substitution within the identified exon. This base change does not induce exon skipping but is centrally located within all three probes of the probe set (B). The central location suggests this mutation causes a reduced affinity to the probes on the exon-array.

The majority of genetic changes identified by OES were single nucleotide changes, both in breast cancer cell lines (5 known SNPs) and in glial brain tumors (2 novel base changes, 16 known SNPs). Moreover, two out of ten previously identified oncogenic point mutations that did not induce exon skipping events were also OES detected in our cohort of breast cancer cell lines: *MAP2K4* c.551C>G in MDA-MB-134VI and *PTEN* c.274G>C in CAMA-1; [15,16] (Figure 5). Single nucleotide mismatches have been used to define hybridization specificity on other Affymetrix microarray platforms. By analogy, single nucleotide substitutions in cancer may also cause reduced hybridization to the probes on the microarray and thus be detected as outlier exons by OES. Indeed, all of the OES detected base changes and SNPs were centrally localized within the probe set selection region and overlap with several of its individual probes (Figure 5).



```
prs seq:   CAATATCCTTTTGAAGACCATAACCCAC
probe  1   CAATATCCTTTTGAAGACCATAACC
       2     ATATCCTTTTGAAGACCATAACCCA
       3      TATCCTTTTGAAGACCATAACCCAC
CAMA1 seq:CAATATCCTTTTGAACACCATAACCCAC
```

Figure 5: Performance of Outlier Exon Screening to detect recurrent outlier exons. (A) Simulation experiment to determine OES performance in detecting recurrent exon-skipping events among clinical glioblastoma samples, where mutant samples express the *EGFRvIII* isoform with deletion of exons 2 through 7. The cohort of 14 glioblastomas included six mutant samples that were replaced by wild-type samples through reiteration, based on their position from left to right in Figure 3B. Deletion of *EGFR* exon 6 in sample GBM67 was detected only as unique mutant sample. (B) Simulation experiment to determine OES performance in detecting recurrent exon-skipping events among breast cancer cell lines, using the wild-type cell line CAMA-1 and the *RB1* exon 22 deletion mutant HCC1937. The two cell lines were analyzed under various cohort sizes, with either the wild-type or the mutant cell line as single sample. The mutant sample was still detected at OES value -2.0 with five recurrent mutants among six samples. The average expression level of RB1 exon 22 dropped below PLIER 50 when more than five mutants were simulated, precluding OES analysis (see Materials and Methods).

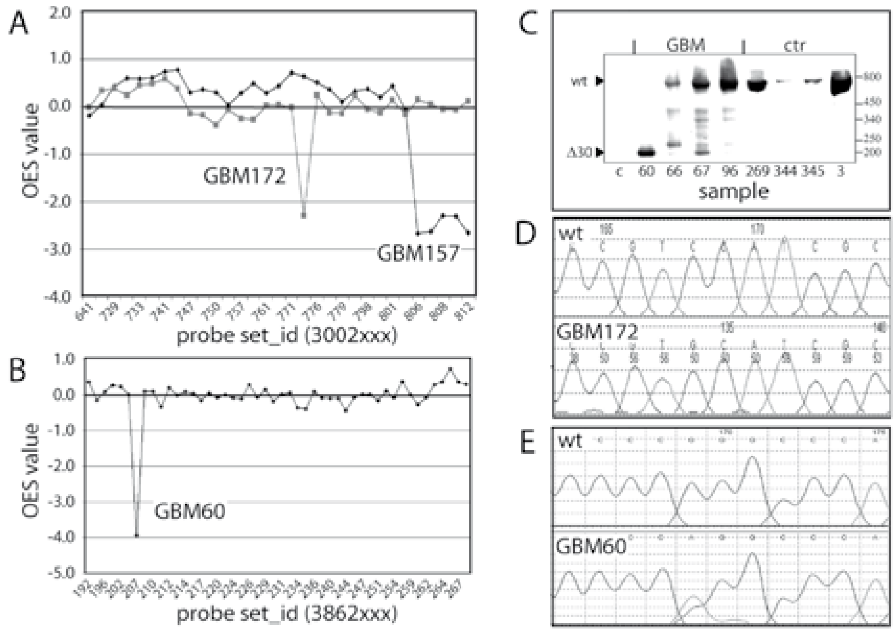


Figure 6: OES Identification of novel genetic changes (A) OES detection of novel genetic changes in EGFR. OES predicted skipping of the last four exons of GBM157 and in the 5' end of exon 17 in GBM172. Real-time PCR on genomic DNA confirmed a deletion of in GBM157 (not shown) and direct sequencing identified a single base change in GBM172 (D). (B) OES predicts skipping of exon 30 in the *FCGBP* gene in GBM60. (C) RT-PCR confirmed the exon skipping event in GBM60; other tumors did not show this exon skipping. (E) Confirmation of an OES predicted change in the *TLE2* gene in GBM60. The nucleotide substitution overlaps with individual probes of the probe set.

One of the identified novel exon skipping was predicted to result in a deletion of the four 3'-end exons of *EGFR* (Figure 6A). We confirmed this deletion using semiquantitative PCR on genomic tumor DNA. Compared to the 5' end of the EGFR locus in GBM157, the 3' end showed less (ΔCt -2.5) amplification whereas other samples showed equal amplification between the 5' and 3' end of the gene (ΔCt 0.3 ± 1.9). Similar 3v deletions in *EGFR* have been observed previously in gliomas (21). The second confirmed exon-skipping event predicted by OES would result in a deletion of exon 30 in the *FCGBP* cDNA (Figure 6B). This deletion was confirmed by RT-PCR and sequence analysis (Figure 6C). Novel identified single base changes include a single base change 1934C>G (s645c) in the *EGFR* gene, (Figure 6A and D), and a single base change 946G>A (g316r) in the *TLE2* gene (Figure 6E).

In summary, the novel exon skipping events and base changes identified by analysis of a select set of outlier exons confirms the suitability of OES to identify candidate cancer genes.

Discussion

We have developed Outlier Exon Screening (OES) to screen for cancer gene mutations that cause exon skipping in the encoded transcripts. We demonstrate that OES correctly detected all of seven previously identified exon-skipping mutants in breast cancer cell lines and two of six mutants in clinical glioblastoma samples. Importantly, OES identified a number of novel genetic changes, including those affecting splicing, that previously had gone undetected. A significant number of nucleotide substitutions that are located within the probe set selection region are also OES detected (Figure 5).Our results thus classify OES as a reliable approach to screen for candidate cancer genes in a global fashion.

Gene expression profiling at the level of individual exons has only recently become feasible through the release of exon arrays. Here, we have explored the efficacy of OES to identify exon-skipping mutants, but the strategy may also be used to deduce the primary structure of gene transcripts [18]. It is important to note that the OES algorithm, detailed under Materials and Methods, is in essence a simple formula that compares measured exon expression levels with their predicted expression levels that in turn are calculated by the publicly available PAC algorithm. The OES algorithm is independent of array platform or organism, allowing application of the OES strategy in a wide variety of biological systems. Several algorithms for exon-level expression profiling are commercially available, including Stratagene ArrayAssist (www.stratagene.com), Partek Genomics Suite (www.partek.com) and Genomatix Suite (www.genomatix.de). Although each of these software packages is relatively straight-forward, important advantages of OES are that it allows detection of unique outlier exons without any prior knowledge of the encoding gene or its transcript structure and that it does not require predefined subgroups of samples with differential expression of the outlier exons.

As with any global screening strategy, OES has its preconditions for detecting outlier exons. First and foremost, identification of outlier exons requires their transcript expression level to be within the linear detection range of the exon array, which is determined by their transcript expression level as well as the hybridization efficiency and specificity of the probe sets involved. The constituency of the test samples is another consideration, particularly when both mutant and wild-type transcripts may be expressed. For example, the breast cancer cell line cohort included two splice site mutants that escaped detection by OES because each had a second transcript length of major intensity that resulted from cryptic splicing (*BRCA1* c.5396+1G>A in MDA-MB-436 (13) and *p16* c.150+2T>C in MDA-MB-436 (Nagel and Schutte, manuscript in preparation). Furthermore, OES detection of the *EGFRvIII* transcript isoform in clinical glioblastomas was determined by the overall expression level of *EGFR* transcripts, that was near the limits of linear detection in all five *EGFRvIII* glioblastomas, but also by the ratio of the *EGFRvIII* isoform versus wild-type *EGFR* transcripts (Figure 3B). A corollary is that OES performance may be compromised in detecting an outlier exon when wild-type transcripts represent more than one-fourth of all transcripts of that particular gene, which could be the case in tumor samples with less than 75% neoplastic cells. However, expression levels of mutant and wild-type alleles typically are disproportional to their allele frequency and detection by OES thus again is determined by the (relative) expression level of the outlier transcript. OES therefore performs best in the absence of wild-type transcript expression. Homozygous transcripts are predominantly found among tumor suppressor genes, where often one allele is mutated accompanied by loss of the other allele.

The influence of allele ratios was further stressed in our simulations of recurrent outlier detection by OES: The *EGFRvIII* isoform in GBM67 was detected only once it was present as a unique outlier among 14 samples, whereas it had not been detected in our original OES screen that included five other *EGFRvIII* expressing glioblastomas (Figure 4A). However, this sub optimal OES performance appeared not related to the recurrence of outliers, as recurrent outliers were easily identified among cell lines – even when present in five out of six cell lines (Figure 4B). The simulation experiments also revealed that two cell lines were sufficient to reliably detect outlier exons and that more than eight cell lines did not further improve OES performance, whereas for clinical tumor samples ten samples appeared the minimum but twenty would be preferred (Fiure 4).

How efficient might OES be in detecting mutations in cancer genomes? Recent comprehensive surveys for somatic gene mutations in cancer genomes suggested that, on average, a breast cancer genome contains 12 likely oncogenic mutations (excluding mutants due to sizeable deletions as these would not have been detected) [4,11,24,25]. In the functionally selected subset of protein kinases, the mutation frequency was estimated to be 0.19 and 0.32 mutations/Mb of DNA in breast cancer and glioma samples respectively [4]. Based on our mutation analysis of seven tumor suppressor genes in 41 breast cancer cell lines, where sizable deletions amounted to one-quarter of the mutations identified, we estimate that a breast cancer genome contains at least three sizeable deletions in addition to the 12 likely oncogenic mutations. By extension, OES is anticipated to detect about three oncogenic mutations in an average breast cancer genome if one screens the same set of well-annotated genes as Sjöblom *et al.*

This classifies OES as a highly efficient screening method. A mutation analysis of all outlier exons in a single breast cancer cell line would involve 1,763 amplification and sequencing reactions on a single sample (all outliers at OES values <-4.0). This number of reactions is comparable to a mutation analysis of three genes with each 12 exons, in 50 samples. Importantly, whereas one may end up empty handed upon three whole gene screens, sequencing of OES-identified outliers promises the identification of perhaps as much as three oncogenic mutants.

Material and Methods

Samples

Our collection of 41 publicly-available human breast cancer cell lines had been subjected to mutational screens of seven tumor suppressor genes: *BRCA1* (Breast Cancer Susceptibility Gene 1; OMIM 113705), *CDH1* (E-cadherin; OMIM 192090), *MAP2K4* (MAP Kinase Kinase 4, a.k.a. *MKK4*; OMIM 601335), *PTEN* (Phosphatase and Tensin Homolog; OMIM 601728), p16 (CDK4-inhibitor, a.k.a. *INK4A, CDKN2A*; OMIM 600160), *p53* (Tumor Protein p53; OMIM 191170) and RB1 (Retinoblastoma Susceptibility Gene 1; OMIM 180200) [13-17] (Nagel and Schutte, manuscript in preparation). Mutational analysis involved sequencing the entire coding region of these genes on genomic DNA as well as analysis of the resulting transcript. The twelve breast cancer cell lines used for this study were: CAMA-1, EVSA-T, HCC1937, MDA-MB-134VI, MDA-MB-157, MDA-MB-435s, MDA-MB-436, MDA-MB-453, MDA-MB-468, MPE600, OCUB-F and SK-BR-5. Clinical glioblastoma specimens were frozen in liquid nitrogen immediately upon surgical resection from patients at Erasmus University Medical Center, as described elsewhere [18]. Pathological review revealed at least 70% tumor nuclei for each specimen. Mutation analysis of the *EGFR* oncogene (Epidermal Growth Factor Receptor; OMIM 131550) in the glioblastomas was performed by conventional RT-PCR and subsequent sequencing of transcripts from samples with *EGFR* amplifications. *EGFR* transcript expression was quantified by Real-Time RT-PCR, using primers that amplified exons 2-3 or exons 22-23 and thus allowed discrimination of wild-type *EGFR* transcripts and the *EGFRvIII* isoform.

Exon-level expression profiling

Total RNA was isolated using the Qiagen RNeasy kit for the breast cancer cell lines and using Trizol followed by RNeasy for the glioblastoma specimens [19]. RNA quality was assessed sing the Agilent Bioanalyser, requiring RNA integrity >7.0 [20]. All further processing of the samples was performed according the Affymetrix GeneChip Whole Transcript (WT) Sense Target

Labeling Assay. Affymetrix GeneChip Human Exon 1.0 ST Arrays were used to determine the expression level of virtually all exons present in the human genome (1.4 million probe sets covering >1 million exon clusters). For this study, we used expression data of the 290,000 core probe sets that are supported by putative full-length mRNA from e.g. the RefSeq database. Signal processing was performed after sketch normalization by using Affymetrix ExACT 1.2.1 software and the PLIER algorithm, described in Affymetrix GeneChip Exon Array Whitepaper "Gene Signal Estimates from Exon Arrays" and Technote "Guide to Probe Logarithmic Intensity Error (PLIER) Estimation" (www.affymetrix.com/support/technical).

Exon Screening (OES)

Predicted exon expression levels were calculated by using the PAC algorithm, described in Whitepaper "Alternative Transcript Analysis Methods for Exon Arrays", where the predicted expression level of the exon (Exon-pr) equals the overall expression of its transcript in that sample (Transcript-m: the meta probe set expression level) multiplied by the average expression level of that exon among all samples (Exon-ave) and divided by the average overall expression of the transcript among all samples (Transcript-ave), all 2-logarithm transformed. In formula:

2log [Exon-pr] = 2log [Transcript-m] * 2log [Exon-ave] / 2log [Transcript-ave].

OES values were calculated by subtracting the predicted expression level of the exon in that sample from its measured expression level (Exon-m), again with 2-logarithm transformation:

OES value = 2log [Exon-m] / 2log [Exon-pr].

Meta probe set expression levels were calculated using all core probe sets of a transcript with PLIER signal estimates >50. To enrich for probe sets with significant expression above background, PAC values were calculated using exons and transcripts that had PLIER signal estimates >50 [18]. Identification of outlier exons was performed without prior knowledge of the mutation data.

Acknowledgements

This work was supported by the Susan G. Komen Breast Cancer Foundation, the Dutch Cancer Society Koningin Wilhelmina Fonds and Erasmus MC Mrace.

References

1. Hahn SA, et al. (1996) *DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1* Science **271**, 350-3.
2. Li J, et al. (1997) *PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer.* Science **275**, 1943-7.
3. Samuels Y, et al. (2004) *High frequency of mutations of the PIK3CA gene in human cancers* Science **304**, 554.
4. Greenman C, et al. (2007) *Patterns of somatic mutation in human cancer genomes* Nature **446**, 153-8.
5. Huusko P, et al. (2004) *Nonsense-mediated decay microarray analysis identifies mutations of EPHB2 in human prostate cancer* Nat Genet **36**, 979-83.
6. Tomlins SA, et al. (2005) *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer* Science **310**, 644-8.
7. Becker KF, et al. (1994) *E-cadherin gene mutations provide clues to diffuse type gastric carcinomas* Cancer Res **54**, 3845-52.
8. Frederick L, Eley, G, Wang, XYJames, CD (2000) *Analysis of genomic rearrangements associated with EGRFvIII expression suggests involvement of Alu repeat elements* Neuro-oncol **2**, 159-63.
9. Faustino NACooper, TA (2003) *Pre-mRNA splicing and human disease* Genes Dev **17**, 419-37.
10. Futreal PA, et al. (2004) *A census of human cancer genes* Nat Rev Cancer **4**, 177-83.
11. Sjoblom T, et al. (2006) *The consensus coding sequences of human breast and colorectal cancers* Science **314**, 268-74.
12. Blencowe BJ (2006) *Alternative splicing: new insights from global analyses* Cell **126**, 37-47.
13. Elstrodt F, et al. (2006) *BRCA1 mutation analysis of 41 human breast cancer cell lines reveals three new deleterious mutants* Cancer Res **66**, 41-5.
14. van de Wetering M, et al. (2001) *Mutant E-cadherin breast cancer cells do not display constitutive Wnt signaling* Cancer Res **61**, 278-84.
15. Su GH, et al. (2002) *Mutation rate of MAP2K4/MKK4 in breast carcinoma* Hum Mutat **19**, 81.
16. Hollestelle A, et al. (2007) *Phosphatidylinositol-3-OH Kinase or RAS Pathway Mutations in Human Breast Cancer* Cell Lines Mol Cancer Res **5**, 195-201.
17. Wasielewski M, et al. (2006) *Thirteen new p53 gene mutants identified among 41 human breast cancer cell lines Breast* Cancer Res Treat **99**, 97-101.
18. French PJ, et al. (2007) *Identification of differentially regulated splice variants and novel exons in glial brain tumors using exon expression arrays* Cancer Res **67**, 5635-42.
19. French PJ, et al. (2005) *Gene expression profiles associated with treatment response in oligodendrogliomas* Cancer Res **65**, 11335-44.
20. Schroeder A, et al. (2006) *The RIN: an RNA integrity number for assigning integrity values to RNA measurements* BMC Mol Biol **7**, 3.
21. Frederick L, Wang, XY, Eley, GJames, CD (2000) *Diversity and frequency of epidermal growth factor receptor mutations in human glioblastomas* Cancer Res **60**, 1383-7.
22. Nishikawa R, et al. (2004) *Immunohistochemical analysis of the mutant epidermal growth factor, deltaEGFR, in glioblastoma* Brain Tumor Pathol **21**, 53-6.
23. Mischel PS, Cloughesy, TFNelson, SF (2004) *DNA-microarray analysis of brain cancer: molecular classification for therapy* Nat Rev Neurosci **5**, 782-92.
24. Balakrishnan A, et al. (2007) *Novel Somatic and Germline Mutations in Cancer Candidate Genes in Glioblastoma, Melanoma, and Pancreatic Carcinoma* Cancer Res **67**, 3545-3550.
25. Wood LD, et al. (2007) *The genomic landscapes of human breast and colorectal cancers* Science **318**, 1108-13.

Chapter 8

Discussion

8.0 Discussion

8.1 Microarray applications to oncology

The application of microarray technology to oncological research has brought about a new age of molecular classification of tumors, resulting in the identification of diverse subtypes based on underlying gene transcription and providing insights into the prediction of disease prognosis and response to therapy. The introduction to this thesis has outlined microarray technology including experimental design and analysis, as well as relevant aspects from oncology. Following this, applications of microarray technology and its interpretation in oncology have been described in subsequent chapters. The results presented in this thesis contribute to piecing together the pieces of the puzzle in understanding two types of cancer: breast cancer and brain tumors.

8.2 Considerations on microarray technology

The great advance in our ability to profile the human genome with microarray technology has not come without limitations. Microarray experiments require careful planning and design, from the initial lab procedures through to the complete analysis pipeline. In the following sections, various technical considerations are being discussed on the applications of microarray technology that are described in this thesis.

8.2.1 Sample variability

Cancers are inherently heterogeneous, even without taking in account the inevitable infiltration of normal cells and surrounding stroma. Accordingly, microarray-generated profiles of tumors are heterogeneous, resulting in much variation in results. This is always going to be a limitation of microarray technology and as much as the improvements to the technology to detect and measure accurate signal can improve, tumor samples will always be biologically heterogeneous. Sample size can be increased to reduce biological variability. However, oncological studies have mainly been restricted to analyzing fresh frozen tumor specimens, which are not often readily available. With limited numbers of samples, the influence of false positives and negatives in the analysis must be considered and solid conclusions can only be generated with independent validation studies and/or laboratory experiments, again increasing the need for samples.

The possibility to profile formalin-fixed paraffin-embedded (FFPE) tumors by microarrays has partially overcome this limitation. FFPE tumor samples are more readily available and, importantly, these samples tend the have much better clinical annotation.

Standardizing laboratory steps such as sample preparation, labeling and hybridization are also important issues to address in interpreting variation within and across microarray experiments. Most methods for RNA amplification and labeling rely on high quality RNA samples. However, RNA samples may become partially degraded due to inherent nucleases, handling, age, and method of preservation or extraction. RNA from FFPE tumor samples are particularly susceptible to degradation related to the preservation process. With the technological improvements in RNA labeling protocols, profiling of FFPE tumor samples can now also produce reproducible, good quality results [1,2] . Such technological consistency is important for microarray platforms to be a reliable tool for future diagnostics in medicine. Again, tumors are inherently heterogeneous, putting an extra challenge on the analysis and interpretation of microarray data and making the detection of signal over noise all the more important.

8.2.2 Technical variability: the probes

With current expression microarray technology, expression changes in abundant transcripts can be reliably detected although not necessarily the magnitude of the changes [3]. The ability to quantify amounts of transcript is limited to 1-3 copies of mRNA per cell and even less for rare and low abundance genes. The accurate detection of such low abundance transcripts is difficult, causing a major problem in reproducibility both within and between microarrays [3]. Although most expression microarray analysis is at the summarized transcript level, it is also important to pay attention to the individual probes within a probe set. It is not unusual that probes directed against different regions of the same gene produce different signal intensity values implying that the measured signal intensity may not always be proportional to the absolute concentration of a transcript [3,4]. This may in part be due to sub-optimal probe design or choice and/or incorrect probe annotations [5]. A reasonably effective technique to minimize technical variation between probe intensities is Perfect-Match/Mis-Match (PM/MM) probe set design from Affymetrix. A limitation of this design is that it does not detect mismatch probes that for some reason inappropriately hybridized, with the summarization of the probe set giving a higher expression value for the transcript than its actual expression [6]. Disproportional signal values may also be due to differences in hybridization efficiency between probes, which typically are related to the probe sequence. Although signal strength can be increased by a longer probe length, this may compromise probe specificity due to non-specific or cross hybridization [3,7,8]. Draghici *et al.*

have proposed that labeling with quantum dots further increases the sensitivity of transcript detection without loss of specificity, which would indeed evolve microarray technology [9]. Technical variation between probes (and thus transcripts) can be minimized by normalization procedures in the pre-processing of the chips. Such limitations, however must be taken into account when making solid conclusions and supporting the need for validation studies.

8.2.3 Reproducibility: different platforms and multiple array comparison

The earlier microarray platforms did not have sufficient probe capacity to profile the entire human genome. Differences in gene signatures from different studies could thus be driven by the genes represented on the platforms rather than by true biological differences [10]. It is therefore important that the same sets of genes are being analyzed when comparing different microarray platforms, but even when the same genes are represented on different platforms, the probe selection regions may differ and thus generate differences in signal intensities. Finally, inadequate performance of gene signatures in follow-up studies may not be related to differences in microarray platforms, but simply to the lack of an independent validation set or appropriate cross validation methods in the original study. This suggests high sensitivity/specificity ratios due to the inevitable variation among sample sets that had not been taken into account (see also chapters 1.3 and 1.9.3-4). The best validation of a gene signature is profiling independent sample sets, preferably collected at other institutions.

External variation between datasets can contribute to differences in differentially expressed gene sets; thus what may be a classifier for one dataset may not hold for the next dataset of independent tumors. Besides technological limitations, this can also be due to heterogeneity among the profiled tumors as previously described. An important comparison of different gene signatures was reported by Fan *et al.* [11]. In this study, a single sample set of 295 breast cancers was analyzed with five established prognostic gene-expression-based models: intrinsic subtypes [12,13]; 70-gene profile of good versus bad prognosis model [14]; wound response model [15]; recurrence score [16,17]; and the two-gene ratio [18]. Even though the five gene signatures consisted of different gene sets, four of the five gene signatures showed significant agreement in the prognosis of individual patients and thus are probably tracking a common set of biological phenotypes. The question remains why the gene signatures differ and which is the most reliable gene signature. Will the final classifier that goes into the clinic be based on a series of gene signatures, or will there be a single gene signature that out performs the others [19]?

With increasing knowledge of the complex biology underlying cancer, it seems evident that the most reliable diagnosis, prognosis and prediction will be based on a combination of platforms that interrogate biology at the DNA, RNA and protein levels. It will also be of increasing importance to address epigenetic changes in tumors, such as methylation-associated and miRNA-associated silencing of gene expression. In this respect, numerous miRNAs have now been shown to be deregulated in human cancers, and over half of known miRNA's are located at sites of recurrent deletions or amplifications [20]. With this information, it is even more apparent that unraveling the underlying biology requires a systems biology approach; in this case integration of CGH or SNP microarray data with miRNA data. Also, miRNA expression in breast cancer correlated with specific breast cancer histopathologic features, such as estrogen and progesterone receptor expression and tumor stage [21]. MicroRNA's are an important piece of the biological puzzle that is cancer, yet only a single piece of information. In order to correctly determine how they define the transcriptome as well as all other mechanisms of regulation that work in conjunction with miRNAs, the entire orchestra will need to be measured to get a complete picture in which to base genetic subtyping, diagnosis treatment options and prognosis as well as the discovery of potential drug targets. Still today only little is known about the specific involvement of non-coding genes and the microarray technology as it stands is not capable of capturing the genetic signature of the tumor including both coding and non-coding transcriptome and epigenetic regulation.

8.2.4 Analytical variability

Each microarray dataset is different in regards to both its distribution of signal intensity, variation and the underlying interacting biology that is being detected. There are many ways to analyze microarray data. There is however not a single algorithm that has been shown to be superior over others and they all may give global answers to the experimental hypothesis. Normalization pre-processing can efficiently be performed with over 20 different algorithms, each transforming the data slightly different. Given the magnitude of data, the analysis of microarrays will also be affected by errors due to multiple comparisons [22]. Many statistical pipelines consider this issue, with adjustment of the false discovery rate being most widely used. There are also many different statistical procedures and selection criteria to define gene signatures or classifiers, resulting in different gene lists even when using the exact same data set. Microarray data analysis may need to be more regulated for microarray technology to become a conclusive tool. For implementation of microarray technology in the clinical setting, it is important to be able to distinguish between noise and signal under all circumstances as well as to have consistent sensitivity and specificity for a particular gene signature, whatever the objective of the analysis [23].

8.3 Focused microarray analysis

Although the general procedure for microarray analysis follows similar steps of class comparison, class prediction, multiple testing and FDR detection procedures, as described in chapter 1, there is no ‘rule of thumb’ as such. Each microarray experiment requires its own particular analytical approach related to the sample set and its underlying biology, the research question and hypothesis, or the microarray platform being used. This often involves extra levels of microarray analysis, as exemplified in various chapters in this thesis. The profiling of *CHEK2* 1100delC mutant breast cancers, for example, was hampered by the strong gene expression program associated with estrogen receptor (ER) status (Chapter 5). ER status is a problem in most breast cancer profiling experiments, because the ER profile is nearly always dominant over other profiles. To unmask the *CHEK2* profile, we restricted the microarray analysis to the subgroup of ER-positive tumors. Determination of the ER subgroups was performed by unsupervised correlation based cluster analysis, which resulted in a division of samples that was >85% consistent with ER protein expression data by ELISA, immunohistochemistry and ligand binding assays. This modification in microarray analysis allowed identification of the *CHEK2* profile underlying the initial ER-dominated transcriptional profile.

Analysis of the exon-level expression arrays required considerations on a technical level. The exon array comprises the full, extended and core exons (paragraph 1.12.1 and chapters 6 and 7). Exon arrays surpass the classical 3’- probe directed expression arrays by having the potential to identify unusual transcriptional phenomena within individual transcripts. In oncological research, an important application is the detection of differential splice variants (Chapter 6), gene mutations that cause exon skipping (Chapter 7), and the detection of aberrant fusion proteins. When using exon arrays similar to the classical expression arrays, the full dataset on all exons can be used. The more specific applications of exon arrays may however require technical modifications. Identification of splice variants and outlier exon screening (OES), for example, depends on accurate calculation of the overall expression level of the transcript to predict skipping of single exons by correlating its predicted expression level with the measured level. Not all exons represented on the exon array however have biological relevance; especially the extended and full exons that include predicted exons, which may not even be a part of a real transcript let alone exist as a true exon. Inclusion of these exons in the microarray analysis will create problems in the PAC calculations because they generate a false representation of the summarized transcript signals intensity. Exclusion of these exons indeed allowed us to calculate accurate representation of the transcript signal and detect differential splice variants between cancer subtypes (Chapter 6).

8.4 Considerations on microarray applications in oncology

Within this thesis, the following papers were presented:

8.4.1 Epigenetic inactivation of E-cadherin by methylation is distinct from genetic inactivation by mutation

Mutations in the *E-cadherin* gene were reported in lobular breast cancers but not in other breast cancer subtypes. By studying 41 human breast cancer cell lines, we identified ten cell lines with genetic mutations of *E-cadherin* and twelve cell lines with *E-cadherin* promoter hypermethylation (Chapter 4). Interestingly, all cell lines with genetic mutation of *E-cadherin* grew with rounded cells whereas all cell lines with epigenetic *E-cadherin* inactivation grew with spindle cells, suggesting that genetic and epigenetic inactivation of *E-cadherin* involves distinct biological pathways. Microarray gene expression analysis then associated the spindle cell lines with an ER-negative gene signature with loss of several cytokeratins and upregulation of the mesenchymal marker vimentin. A 3-protein spindle cell signature identified the spindle cell signature in 34 ER- clinical breast cancers, of which 28 were metaplastic breast cancers, but not in any of the ER+ clinical breast cancers, including 21 lobular breast cancers. Half of metaplastic breast cancers had lost E-cadherin protein expression and 75% of lobular breast cancers, but gene mutations were only found in the lobular breast cancers. We conclude genetic and epigenetic inactivation of *E-cadherin* involves two distinct biological pathways that associate with different breast cancer subtypes: lobular and metaplastic breast cancers.

Global gene expression profiling using microarrays can give a complete molecular picture of a cancer: not only the initiating events but also the entire cascade of events that make up a molecular tumor portrait. This is valuable in situations such as *E-cadherin* inactivation. Loss of E-cadherin has been associated with lobular breast cancers, however a global view at the transcriptome demonstrated that loss of *E-cadherin* by methylation is in fact associated with a metaplastic breast cancer subtype, which are predominantly basal breast cancers. Information such as this will be important in drug discovery and treatment of cancer, as mode-dependent inactivation of the protein is associated with two different subtypes of cancer. In the future, medicine can step towards a more personalized route because of our increasing knowledge of molecular disease markers. This may result in a custom-tailored breast cancer chip, with all informative genes for breast cancer spotted onto it. A single experiment could then determine disease and subtype, as well as treatment response and survival prognosis.

8.4.2 A gene signature is associated with *CHEK2* 1000delC mutations in breast cancer

The *CHEK2* 1100delC mutation is a breast cancer susceptibility allele that confers a low breast cancer risk. We determined a *CHEK2* gene expression profile, using a cohort of 155 familial breast cancers (Chapter 5). Although *CHEK2* 1100delC is a low risk breast cancer allele, *CHEK2* 1100delC families typically display a high-risk cancer inheritance pattern. It has therefore been postulated that *CHEK2* works in conjunction with another susceptibility allele in a 'multi-genic' model, possibly even requiring multiple modifiers [24]. The *CHEK2* 1100delC mutation appears mutually exclusive with *BRCA1* mutations in breast cancer patients and BRCA1 is a phosphorylation target of CHEK2 in the DNA repair pathway.

All 26 *CHEK2* 1100delC mutated samples in the cohort showed an ER-positive transcriptional profile. The analysis was therefore focused on the 100 ER-positive samples in the cohort. A class comparison analysis was performed between the *CHEK2* mutant samples versus all other samples. Given that the two might be functionally related, the class comparison analysis was also performed for the *CHEK2* samples versus the rest of the samples but minus those with a *BRCA1* mutation. The results of our *CHEK2* profiling could suggest that there may be multiple subtypes underlying the established *CHEK2* profile, as seen in the distribution of the *CHEK2* tumors within two major arms in the unsupervised clustering of the gene expression values. On initial observation this division could be seen to be related to the luminal A and B division of molecular subtypes [12,13,25], however the statistics does not support this theory and CHEK2 1100delC tumors are of both of the Luminal A and B subtypes.

The class comparison analysis gave a statistically significant 40-gene signature that associated with the *CHEK2* 1100delC mutation. The gene list from the analysis minus the BRCA1 samples was significantly more extensive (69 genes) with a lower FDR, suggesting that the similarity between the *BRCA1* and *CHEK2* profiles was indeed high (37 of 40 genes from the *CHEK2* signature were also in the minus-BRCA1 signature) and may have made the detection of significantly differentially expressed genes more difficult. By excluding the *BRCA1* samples from the analysis, the unsupervised clustering of *CHEK2* samples using the gene profile was in fact tighter than with the *BRCA1* samples being included in the 'non-CHEK2' class. Within the gene signatures two known *BRCA1* associated genes were present and an over representation of genes associated with TP53 biological networks.

The unsupervised clustering of the top variable genes would most likely reflect the underlying tumor biology not related to the mutation status. Given previous research and literature, it was expected that the division of tumors would coincide with known molecular subtypes. The fact that the *CHEK2* tumors did not coincide with one molecular subtype such as Luminal A or B suggests that there is still some heterogeneity among them unrelated to the molecular subtype. The presence of a 40 gene signature again implies biological homogeneity among the samples, however clustering of the samples using these genes did not produce a distinct group of *CHEK2* tumors associated with the gene signature confirming that they are not completely homogeneous. This heterogeneity may be related to the additional susceptibility modifiers in the proposed 'multi-genic' model and can suggest a further division of the *CHEK2* tumors into subtypes dominated by these modifiers. The close distribution of the *CHEK2* tumors among the two arms of the clustering may suggest that there could be in fact a limited number of gene modifiers that are associated with *CHEK2* in this model thus driving the clustering division. This also suggests that the modifiers could in fact lie within the gene signature. We are currently investigating whether these subtypes may be associated with different *CHEK2* modifiers. Because the *CHEK2* 1100delC mutation is not the only predisposing factor for breast cancer, detecting a *CHEK2* profile requires a relatively large cohort of tumors.

The clustering together of *CHEK2* mutant breast cancers with *BRCA1* mutants also suggests a correlation in their molecular profile; that they may in fact functionally work together. The presence of two known *BRCA1* associated genes in the gene signature also suggests this relationship, This is interesting as *BRCA1* mutant breast cancers are – in contrast to *CHEK2* 1100delC tumors – predominantly ER-negative and from the basal subtype of breast cancers. This however needs to be further confirmed with a bigger BRCA1 cohort, expanding out to also look at ER-negative breast cancers.

Results from the biological function analysis of the gene signature, suggest that *CHEK2* and *TP53* function in the same oncogenic pathway. Genes that partake in the same oncogenic pathway would not be mutated in the same tumor as the second mutation would not give any greater selective survival advantage. This would predict that *CHEK2* 1100delC tumors would carry a wild-type *TP53* allele. Indeed, the *TP53* expression values supported this theory with *CHEK2* samples having average *TP53* transcription. In a majority of *BRCA1* samples, *TP53* expression was considerably lower. The clustering of *BRCA1* samples with the *CHEK2* samples could reflect their homogeneity of oncogenic function; however it could also merely reflect the underlying *TP53* mutation status.

Knowing the molecular profiles associated with *CHEK*2 1100delC mutations as well as discovering its modifiers may assist in determining prognosis of patients and improve our ability to provide them a successful treatment.

8.4.3 Exon arrays identify differentially expressed splice variants in brain tumors

The novel exon-level expression microarrays measure the expression of virtually all known and predicted exons present in the human genome. We have performed exon array analysis on 26 glioblastomas, 22 oligodendrogliomas and 6 control brain samples, using a pattern based correlation algorithm (Chapter 6). We were able to classify the brain tumors by splice variants that were differentially expressed between oligodendrogliomas and glioblastomas. These results demonstrate the advancements in microarray technology, with reliable detection of splice variants that play a key role in the biology of brain tumors. Perhaps as much as 20% of oncogenic mutations may affect the splicing of transcripts; an important phenomenon to detect when measuring gene expression in tumors, but was beyond the capabilities of the classical 3'-directed expression arrays. Detecting aberrant splicing events with exon array technology will increase our understanding of the molecular events involved in the initiation and/or progression of brain tumors as well as the regulation of downstream targets in the different subtypes, which can provide an important basis for drug discovery.

8.4.4 Exon arrays identify exon-skipping mutations in breast cancer cell lines and brain tumors

Exon-skipping mutations represent an estimated 10-20% of all cancer-related gene mutations and as much as half of all human disease gene mutations. We have developed OES to identify gene mutants that cause exon skipping in the encoded transcripts, using exon arrays (chapter 7). OES detected all seven exon-skipping mutants among 12 breast cancer cell lines and three of seven mutants among 14 brain tumors. OES reduced the number of candidate genes or exons by two to three orders of magnitude, increasing the number of truly skipped exons up to 24% of identified outlier exons. The OES approach does not directly identify disease genes, but it provides a method for screening candidate genes. OES has great advantages over long screening procedures involving positional cloning, linkage/population analysis and rigorous sequencing. In a single microarray experiment, the number of genes from a candidate gene list can be narrowed down from hundreds to tens. Importantly, OES identifies a candidate exon instead of a gene, as the mutation will often lie within or flanking the outlier exon.

8.5 The future of microarrays applications in oncology and final conclusions

Microarray technology is an invaluable tool for the future. The combination of computer science, mathematics and biology in bioinformatics, allows scientists to process and analyze large volumes of high dimensional data from high throughput technologies such as microarrays. It is crucial to cancer and many other complex diseases that the integration of such data using a bioinformatics, systems biology approach can lead to a detailed understanding of disease. It is also very important to know and understand the capacities of the technology and its limitations. The efficient integration of genomics (at both DNA and RNA levels), epigenomics (including both promotor methylation and microRNA's) and proteomics will need to be met with advances in development of tools and analysis software in order to integrate this data with its large differences in measurement levels. Combining such data allows a technique- independent validation of results [23]. This is very important as basing biological conclusions solely on a single platform can pose a problem when it is translated into the clinic. Genomics and epigenomics may not directly translate into the proteome that is being researched due to variations of cellular processing of mRNA and posttranslational modifications [59]. On the DNA level, SNP arrays have been a useful platform in biological research, including cancer. Using this type of array, both genomic polymorphisms and DNA copy number can be established. This technique can be used to establish information on chromosomal gains, losses or loss of heterozygosity. On the epigenomics level, microarrays are continuously being developed to interrogate the expression of non-coding RNA's such as miRNA's and exon and tiling arrays by Affymetrix can address the phenomenon's of alternative splicing, alternative promoter usage, promoter methylation and gene silencing. On the protein level, mass spectrometry is also a fast growing area that looks on a global level at the protein content and protein levels in cells [26,27]. This technique is still quite new and like microarrays must be developed further to confidentially identify the proteome present in a biological sample. Currently, mass spectrometry reliably identifies only one to five percent of most abundant proteins present in a sample, however this is changing rapidly with great technological and analytical advances to the technique [23].

It has only become possible in the past decade to mine large amounts of high-dimensional data on human cancers. Our rapidly increasing knowledge of specific genes and proteins as well as biological pathways that are associated with carcinogenesis has provided opportunities to improve diagnosis and to develop targeted cancer therapies. Currently, cancer diagnosis is based largely on histopathology. Microarray gene expression profiling determines the transcriptome of a tumor and may thus help to uncover the underlying genetic mutations, pathways and functions affected in the tumor. Tumors that appear heterogeneous may in fact be rather homogeneous in their genetic make-up, allowing further distinguishing tumor subtypes and thus improving diagnosis. Global information on tumor biology may also improve the prognosis of cancer patients, the prediction of their clinical outcome, or open up new avenues in the development of targeted treatment of cancer patients. For example, an interesting possibility of miRNA's is their ability to affect multiple targets [28]. Whereas this is a disadvantage in siRNA/shRNA technology due to unpredictable off-targets, miRNA targets are far more specific. The multiple targeting of miRNA's thus is more likely to re-establish an entire network of cell function.

The ability to accurately capture a tumors genetic signature is priceless, as is the information it contains for the development of individualized diagnosis and treatment of disease [29]. The ideal of providing such personalized medicine for cancer in the future however, has not been fully met for several reasons:

- Cancer-related pathways are complicated, with frequent cross talk between them, rendering our true comprehension of cancer biology still rather limited. The more the biological web is uncovered, the more we become aware how complex carcinogenesis is. In addition, we have only just begun to unravel other biological aspects e.g. the involvement of miRNA's and differentially expressed splice variants.
- Patients with similar clinical and pathological features in their tumors are still heterogeneous in other aspects, causing clinical outcome to be variable and unpredictable. Further research is needed to understand the genetic and metabolic background of a patient, and how they metabolize drugs, handle side effects of treatments and what may affect their chances of tumor metastasis. Expression of pharmacogenetic targets is generally not confined to cancer cells and, hence, unexpected (or expected) toxicities may be observed. A patient's genetic composition may also influence resistance to drugs such as that seen in breast cancer patients treated with *Tamoxifen* [30]. This information cannot be obtained solely through the biological analysis of tumor specimens, but also requires additional evaluation of a patient's clinical characteristics.

The advancements in microarray technology have opened up many different avenues of research**.** The investigations reported in this thesis provide both information and evidence on the value of microarray expression profiling in tumors. Microarray technology is of great value in understanding the underlying biology of cancer and in the diagnosis, prognosis and prediction of clinical outcome of patients with cancer. In this thesis, I have demonstrated that microarray technology combined with bioinformatic analysis and *in silico* research allows high throughput molecular research on breast cancers and brain tumors. The microarray era is still evolving, with the emergence of more sophisticated technology as well as new statistical methods being introduced. It can be foreseen that expression profiling offers added informative value and may replace certain current routine diagnostic approaches (immunophenotyping, cytogenetics or molecular diagnostics) within the near future. As it stands today, however, further optimization and standardization of microarrays and their analysis is required before they can be used reliably for clinical decision making.

8.6 Discussion bibliography

1. Coudry, R.A., et al., *Successful application of microarray technology to microdissected formalin-fixed, paraffin-embedded tissue.* J Mol Diagn, 2007. **9**(1): p. 70-9.
2. Penland, S.K., et al., *RNA expression analysis of formalin-fixed paraffin-embedded tumors.* Lab Invest, 2007. **87**(4): p. 383-91.
3. Draghici, S., et al., *Reliability and reproducibility issues in DNA microarray measurements.* Trends Genet, 2006. **22**(2): p. 101-9.
4. Leiske, D.L., et al., *A comparison of alternative 60-mer probe designs in an in-situ synthesized oligonucleotide microarray.* BMC Genomics, 2006. **7**: p. 72.
5. Perez-Iratxeta, C. and M.A. Andrade, *Inconsistencies over time in 5% of NetAffx probe-to-gene annotations.* BMC Bioinformatics, 2005. **6**: p. 183.
6. Draghici, S., *Data Analysis Tools for DNA Microarrays*. 2003, London: Chapman and Hall.
7. Shippy, R., et al., *Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations.* BMC Genomics, 2004. **5**(1): p. 61.
8. Relogio, A., et al., *Optimization of oligonucleotide-based DNA microarrays.* Nucleic Acids Res, 2002. **30**(11): p. e51.
9. Liang, R.Q., et al., *An oligonucleotide microarray for microRNA expression analysis based on labeling RNA with quantum dot and nanogold probe.* Nucleic Acids Res, 2005. **33**(2): p. e17.
10. Larkin, J.E., et al., *Independence and reproducibility across microarray platforms.* Nat Methods, 2005. **2**(5): p. 337-44.
11. Fan, C., et al., *Concordance among gene-expression-based predictors for breast cancer.* N Engl J Med, 2006. **355**(6): p. 560-9.
12. Sorlie, T., et al., *Repeated observation of breast tumor subtypes in independent gene expression data sets.* Proc Natl Acad Sci U S A, 2003. **100**(14): p. 8418-23.
13. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.
14. van 't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**(6871): p. 530-6.
15. Chang, H.Y., et al., *Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival.* Proc Natl Acad Sci U S A, 2005. **102**(10): p. 3738-43.
16. Wang, Y., et al., *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.* Lancet, 2005. **365**(9460): p. 671-9.
17. van 't Veer, L.J., et al., *Expression profiling predicts outcome in breast cancer.* Breast Cancer Res, 2003. **5**(1): p. 57-8.
18. Ma, X.J., et al., *A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen.* Cancer Cell, 2004. **5**(6): p. 607-16.
19. Reis-Filho, J.S., C. Westbury, and J.Y. Pierga, *The impact of expression profiling on prognostic and predictive testing in breast cancer.* J Clin Pathol, 2006. **59**(3): p. 225-31.
20. Calin, G.A. and C.M. Croce, *Chromosomal rearrangements and microRNAs: a new cancer link with clinical implications.* J Clin Invest, 2007. **117**(8): p. 2059-66.
21. Iorio, M.V., et al., *MicroRNA gene expression deregulation in human breast cancer.* Cancer Res, 2005. **65**(16): p. 7065-70.
22. Christie, J.D., *Microarrays.* Crit Care Med, 2005. **33**(12 Suppl): p. S449-52.
23. Verhaak, R., *Gene expression profiling of acute myeloid leukemia*, in *Haematology*. 2006, ErasmusMC: Rotterdam.
24. Meijers-Heijboer, H., *Breast cancer susceptibility genes thesis*.
25. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-74.
26. Aebersold, R., *Quantitative proteome analysis: methods and applications.* J Infect Dis, 2003. **187 Suppl 2**: p. S315-20.
27. Aebersold, R. and M. Mann, *Mass spectrometry-based proteomics.* Nature, 2003. **422**(6928): p. 198-207.

28. Negrini, M., et al., *MicroRNAs in human cancer: from research to therapy.* J Cell Sci, 2007. **120**(Pt 11): p. 1833-40.

29. Osborne, C., P. Wilson, and D. Tripathy, *Oncogenes and tumor suppressor genes in breast cancer: potential diagnostic and therapeutic applications.* Oncologist, 2004. **9**(4): p. 361-77.

30. Jansen, M.P., et al., *Molecular classification of tamoxifen-resistant breast carcinomas by gene expression profiling.* J Clin Oncol, 2005. **23**(4): p. 732-40.

Summary

Samenvatting

Acknowledgements

Curriculum vitae

Publications

Abbreviations

SAMENVATTING/SUMMARY

De microarray technologie is een vooruitgang in het genomische veld. Met deze techniek kunnen expressie niveaus van duizenden genen van een enkele individu tegelijk gemeten worden. Hierdoor kan afwijkende genexpressie gedetecteerd worden door patiënten of behandelde patiënten te vergelijken met controles. Het op grote schaal detecteren van genexpressie kan toegepast worden in vele biologische en medische werkvelden, bijvoorbeeld oncologie. Deze nieuwe technieken zijn het begin van een nieuw tijdperk high-throughput laboratorium experimenten en functionele studies. Deze vooruitgang kan een beter begrip van biologie mogelijk maken en de klinische diagnose van ziekten verbeteren, tot uiteindelijk een op de patiënt afgestemd medicijn. Door deze dynamische verandering in onderzoek is er een vraag naar bioinformatica ontstaan. Dit veld combineert wiskunde, statistiek, computerwetenschap en biologie om biologische vraagstukken op te lossen, meestal op moleculair niveau. Om bioinformatica toe te passen in microarray technologie in high-throughput genexpressie studies is het ontwikkelen en gebruiken van statistische applicaties nodig; om signaal van ruis te onderscheiden, voor de statistische detectie van veranderde genexpressie en voor de interpretatie van relevante biologische patronen in het experiment (**Hoofdstuk 2**).

Dit proefschrift, voornamelijk gericht op oncologie (**Hoofdstuk 1**), is geschreven als referentie voor wetenschappers die van plan zijn om microarray experimenten toe te passen en deze data te analyseren.

In de introductie zijn experimentele procedures van begin tot eind beschreven als een referentie om wetenschappers te begeleiden om op juiste wijze een microarray experiment te plannen en om ze te wijzen op de methoden en applicaties die beschikbaar zijn voor het analyseren en het beantwoorden van hun experimentele hypothese. Het tweede deel van dit proefschrift presenteert verschillende applicaties van microarrays in oncologie, waarin de gebruikte analyse technieken worden geschetst in de refererende hoofdstukken. Microarray technologie is een applicatie van onschatbare waarde in het veld van oncologie, vanwege het complexe genetische karakter. Het in kaart brengen van het hele transcriptoom kan een inzicht geven in de biologie achter een ziekte en patronen van transcriptie en regulatie mechanismen onthullen die betrokken zijn bij de initiatie en progressie van de ziekte. Daarnaast kan dit ook inzicht geven in de reactie op behandeling en de prognose van de patiënt.

Hoofdstuk 3 presenteert een review artikel die de groei in microarray technologie en analyse applicaties en zijn succesvolle toepassingen in vele biomedische disciplines benadert. Het in kaart brengen van genexpressie bij oncologie heeft grote potentie, omdat tumoren ontstaan vanuit een serie genetische en epigenetische mutaties. Dit oncologie onderzoek

is van groot belang voor het complete begrip van de onderliggende biologische kenmerken van heterogene tumoren en tumorsubtypen en voor het verbeteren van de mogelijkheden om kanker te voorkomen, detecteren en behandelen.

In **hoofdstuk 4** beginnen de peer-reviewed experimentele artikelen. In borstkanker-onderzoek is met 3' expressie arrays gevonden dat inactivatie van het tumor suppressor gen E-cadherin, door zowel genetische als epigenetische mechanismen, betrokken is bij twee verschillende biologische netwerken *in vitro* die geassocieerd zijn met lobulaire en metaplastische borstkanker subtypen. Met 3' expressie arrays hebben we ook een moleculair profiel geïdentificeerd dat geassocieerd is met CHEK21000delC, een mutatie die gevonden wordt in familiaire borstkanker en gekoppeld is met een verhoogd risico tot het ontwikkelen van borstkanker (**Hoofdtuk 5**).

In de neuro-oncologie hebben we een algoritme aangepast voor de analyse van exon expressie arrays voor het identificeren van differentieel gereguleerde splice varianten in centraal zenuwstelsel glioma's (zowel oligodendroglioma als glioblastoma histologische subtypen), die betrokken zijn in de initiatie en/of progressie van deze tumoren (**Hoofdstuk 6**). In deze analyse zijn ook nieuwe exonen gevonden die niet eerder geassocieerd zijn met bekende transcripten in publieke databases.

Met deze exon array data hebben we een nieuwe analyse strategie ontwikkeld voor het identificeren van humane ziekte genen; in het bijzonder splice mutaties (**Hoofdstuk 7**). Voor dit algoritme hebben we gebruikt gemaakt van bekende borstkankercellijnen en biopten van centraal zenuwstelsel tumoren. Deze aanpak is zeer bruikbaar in de globale screening van exon data voor ziektegenen, omdat dit het screenen beperkt tot individuele exonen in plaats van grote lijsten van positionele en functionele kandidaat genen. Al de gedemonstreerde profielen uit dit proefschrift zijn voorbeelden van bioinformatica toepassingen in microarray analyse en ieder profiel is een belangrijke stap in het begrijpen van de onderliggende biologie van verschillende subtypen van tumoren.

Concluderend beschrijft en illustreert dit proefschrift de bestaande micro array technologie en de bioinformatica toepassingen, rekening houdend met de bestaande status van de technologie en zijn beperkingen.

SUMMARY/SAMENVATTING

The microarray technology is a new advancement in the genomics field, which involves measuring the expression level of thousands of genes simultaneously from a single individual/sample. This technology can be used to detect dysregulated genes by comparing the gene expression in diseased/treated and normal samples. Measuring gene expression on a global scale using a microarray is applicable to many areas of biology and medicine such as oncology and these new technologies are paving the way for a new era of high-throughput laboratory experiments and transcriptome analysis. These advances can enable a better understanding of true biology as well as improve clinical diagnosis of disease to bring forth 'personalized medicine'. Such a dynamic change in research has created a need for bioinformatics. This field combines the mathematics, statistics, computer science and biology to solve biological problems usually on the molecular level. Bioinformatics applied to microarray technology in high-throughput gene expression studies involves developing and utilizing statistical tools to separate signal from noise, the statistical detection of altered gene expression and the interpretation of relevant biological patterns within the experiment (**Chapter 2**).

This thesis, primarily focused in the context of oncology (**Chapter 1**), has been written as a reference tool for scientists planning to run, and analyze microarray experiments. Within the introduction, experimental procedures from the start to finish have been described as a reference to educate scientists into the correct way of thinking to efficiently plan a microarray experiment as well as making them aware of the methods and tools available to analyze and answer their experimental hypothesis. The second part of this thesis presents various applications of microarrays in oncology, in which the analysis techniques used are outlined in the reference chapters. Microarray technology is an invaluable tool in the field of oncology as cancer is a complex genetic disease. Profiling the whole transcriptome can provide a snapshot of the true biology and disclose patterns of transcription and regulatory mechanisms involved in the initiation and progression of disease as well as the response to treatment and prognosis of the patient.

Chapter 3 presents a review paper addressing the growth in microarray technology and analysis tools and its successful applications in many biomedical disciplines. The field of oncology has great potential for the application of gene expression profiling as cancer arises from a series of genetic and epigenetic mutations. Such research in the field of oncology is of great importance to gaining a complete understanding of the underlying biological characteristics of heterogeneous tumors and tumor subtypes, as well as improving the possibilities to prevent, detect and treat cancer.

Chapter 4 begins the peer-reviewed experimental papers. In the breast cancer field, we have identified using 3' expression arrays, that inactivation of the tumor suppressor gene E-Cadherin by both genetic and epigenetic mechanisms involves two distinct biological pathways *in vitro* that associate with lobular and metaplastic breast cancer subtypes. Using the same platform of microarray, we have also identified a molecular profile associated with CHEK2 1000delC, a mutation found in familial breast cancer linked with an increased risk of developing breast cancer (**Chapter 5**).In the field of neuro-oncology, we have adapted an algorithm from the analysis of exon expression arrays to identify differentially regulated splice variants in CNS gliomas (both oligodendroglioma and glioblastoma histological subtypes), which are involved in the initiation and/or progression of such tumors (**chapter 6**). The analysis has also revealed novel exons not previously associated in public databases with known transcripts. Also using the exon arrays, we have also developed a new analysis strategy to identify human disease genes; specifically exon skipping mutations (**chapter 7**). This algorithm was developed using known breast cancer cell lines as well as solid specimens from CNS tumors. This approach is very useful in the global screening of exon data for disease genes as it limits the screens to individual exons rather than having to mine large lists of positional or functional candidate genes. All of the profiles demonstrated in this thesis are examples of the bioinformatic approaches to microarray analysis and each profile is an important step in understanding the complete picture of underlying biology in various subtypes of cancer.

In conclusion, this thesis aims to describe and illustrate the current microarray technology and the bioinformatics applications associated with its application and data analysis, taking into consideration the current status of the technology as well as its limitations.

DANKWORD / ACKNOWLEDGEMENTS

I would like to thank all the people who supported me (and until the very end) on this journey; being not only the PhD but also the living and adapting to a country so familiar from my Dutch background, yet so strange and foreign. My journey here in this country has definitely been an experience. Learning to ride my 'oma fiets' in dutch wind; racing to get to work with my feet peddling hard but the bike traveling backwards; not to mention the rain! Bier, bitterballen, great laughs and conversation on the terraces over the years. There have been so many people who have experienced this along the way with me, both directly and indirectly. My parents; whom I have so much love and respect for respecting me and all that I have ever done. Writing words in this acknowledgement section will not suffice a way to express my love and gratitude for you. Time away from you is the hardest thing of all.

My supervisors who have given me the opportunity and experience of completing my PhD. Professor Peter van der Spek and Dr Mieke Schutte. Mieke, thank you so much for helping, supporting and seeing me through to the end of my thesis. Thank-you for giving me so much of your time- unconditionally. It has been a pleasure and I will always remember the many hours of conversation in your office. You taught me so much about myself, and the professional scientific world.

Thanks also to Professor Dirk Bootsma who passed on in numerous conversations: his scientific and wordly wisdom and gave me a lot of encouragement.

The Bioinformatics department, whom have seen me through my trials and tribulations of being a PhD student. My paranymphomaniacs; Karlijn, you have been my pillar of strength, a listening ear, my social landmark with so much wisdom. Laura, your artistic bubbly flair always put a smile on my face. Rachel, your big friendly eyes always let me know that everything would be alright. You were there too many times to count. Thanks guys for all your support and those times of "shakin' our butts." Having the 3 of you there over the years helped pull me through.

Mirijam, you never said 'No' to helping no matter how small the problem was and you always did everything with a smile. You are more than IT support for the department. Anton, you always had great words of wisdom and seemed to put things right. Marijana, your giggle always put a smile on my face. Lennard, thanks for all the music to get me through my days! Tjeerd and Bas, thank you so much for your programming support; your help always made the work a lot easier to handle.

Many friends outside the department have also seen me through, whom I would like to give special thanks. Roel; one of my greatest friends, or more like a stubborn brother here in NL. You have continued to be a valued friend in Boston. You always gave as good as you got. Karl Brand; my Aussie baseline. Thanks (Taa) for everything; the pure Occa in you was my light (who needs 5000 watts of industrial light shining in your face during the winter mornings!). Michael Moorehouse; thanks for all the gripes about the Dutch culture. You always had an answer to my worldly questions, always had a selection of pens, torches, laser pointers and navigation systems at my disposal. I wouldn't have had the time I did without your quirks! Marcel Smid; my mate from the JNI, whom I always enjoyed to lunch with. You have helped and given me a lot of microarray wisdom that I very much appreciate. Special thanks also for the idea of stelling IX. Others from the JNI whom it was always a pleasure to work with; Pim French, Antoinette Hollestelle and the rest of the Medical Oncology laboratory, and John Martens. A special mention also to Gert-Jan van der Geijn, whom although I never got the pleasure to directly work with has been a great lunch-breakfast friend and a big part of my social crew outside of ErasmusMC. You have become a very valued friend of mine. A special thank-you also to Damian Melles (The paper machine!). It was great to work with you. You always worked with a smile! Thank-you for all the publications.

My vast array of friends in Rotterdam, both Dutch and International. Astrid van Dijk; there are too many great times to mention. You were a hard one to crack but its one of my greatest achievements! Haa (Arequipa, Arequipa!). I can see we will be friends for a long time to come. Paul; also too many great times to mention. All the amazing conversations and you have taught me so much about the true meaning of being Dutch. You friendship means a lot to me. Ori, Inigo ('El Presidento'), Samantha, Vladimiros, Duygu, Marga, Bibiana, Bia, Satish, Eva, Tamas, Krisztina, JC, Lotte, Stef, Monique, Fiona, Silvia, Jane, Anouka, Jari, Susan, Rianne, Stephanie, Zoe, Richard, Jeroen, Lu Lu and Yin-Wah…all my friends in Rotterdam whom I have had 'pub sessions' with over the years.

Gabey and Bez; two of my greatest friends who understand my Aussie soul. It has been busy times, but it means a lot to know you are here in Netherlands experiencing this crazy land with me. Gabey, your integration here is my inspiration.

My friends at home in Melbourne, Australia whom I didn't speak to as much as I would have loved, It is so hard to communicate from the other side of the world immersed in a different culture, but there were a few people who no matter what happens and even in periods of quiet communication, I know they will always be a friend. Astrid and Paul; no matter where I am I can feel their spirit. You guys mean the world to me. Christine; one of the strongest people I know. Nobody knows me like you do and you have been there to support me along the road-most roads I have taken! Gareth; I will always miss our walks, Townhall lunches and beers; there is no colleague/friend that could still ever replace you. You are the person to thank for this PhD as you were the one to teach me the tricks and push me on my way and for that I am grateful. I hope to work with you again one day. Paul and Lisa, Tic and Josh, Jobba and Gus. These are also the people who after being away for many years I know will always be my friends.

The professional people I have met along the way, many of who have given me a lot of help and inspiration. Deon Venter; whom I must also thank for giving me the opportunity to move into the microarray world. The opportunity you gave me allowed me to be where I am today.
Peter Farlie and Don Newgreen; who gave me the confidence that I should be in science and the potential to succeed! Gracia Mancini, thank you for your professional support.
Affymetrix; All the Affymetrix team from both Europe and U.S.A have always been a great help and have also given me many great opportunities. Mike Levielt. Thank-you for all your advice over the past years. It has been great to work with you. Thank-you for all the post PhD coaching. Steve Lincoln, Geoff Scopes, Edwin deVries and Stephan Scrooten. Ingenuity; Brian Dron, Adam Corner. It was fantastic to meet and work with you both. There were always big smiles on your faces which makes a big difference. Omniviz; Geoff Scopes, Spotfire; Steven Narding, NBIC; Victor de Jager whom was once my colleague in bioinformatics.
Thank-you to Peter Valk from the Department of Hematology who has given me the opportunity to continue on in a Post-Doc position at ErasmusMC.

There are always people that you forget in writing down these acknowledgements, however I want everyone to know my appreciation for everything that brought me to where I am today. No matter how big or small your contribution or in what form, I greatly appreciate it!

I’m not afraid of storms, for I’m learning to sail my ship. (Louisa May Alcott)

Those who lose dreaming are lost. (Australian Aboriginal)

Believe nothing merely because you have been told it.
Do not believe what your teacher tells you merely out of respect for the teacher.
But whatever, after due examination and analysis,
you find to be kind, conducive to the good, the benefit, the welfare of all beings -
that doctrine believe and cling to, and take it as your guide. (Buddha)

If you are going through hell; keep going. (Winston Churchill)

CURRICULUM VITAE

Justine Kate Peeters was born in Melbourne, Australia on June 30th 1975. After finishing her Victorian Certificate of Education (VCE) at Killester College, Melbourne in 1992, she started a Bachelor of Science degree (BSc) at Monash University, Melbourne, Australia. As part of this education she followed majors in genetics and psychology. In 2000, an Honors year was completed with first class at the Center for Reproduction and Development, Monash University, Melbourne Australia under the supervision of Professor Melanie Pritchard and Professor Ismail Kola. Her thesis was entitled "The role of ELF5 in cancer". In 2003, she obtained a Masters degree (MSc) with a Helen Schut Foundation scholarship, under the supervision of Dr Peter Farlie and Dr Don Newgreen at Murdoch Children's Research Institute, Melbourne Australia. Her thesis was entitled "Elucidating a functional role for YPEL1". In September 2003 she started her PhD degree in the Department of Bioinformatics, ErasmusMC Rotterdam Netherlands, under the supervision of Professor Peter van der Spek and Dr Mieke Schutte. The author will continue her career as a Post Doctoral Researcher in the Department of Hematology at ErasmusMC in the group of Peter Valk, focusing her efforts in microarray bioinformatics.

LIST OF PUBLICATIONS

Presented in this thesis

Antoinette Hollestelle, ***Justine K. Peeters***, Pieter J. Westenend, Thierry van de Wetering, Leon Verhoog, Allan Chan, Jan G.M. Klijn, Peter J. van der Spek, Michael A. den Bakker and Mieke Schutte. *Epigenetic inactivation of E-cadherin is associated with a rare histological subtype of metaplastic breast cancer.*. Submmitted to Cancer Cell; 2008

P.J. French, ***J.K. Peeters***, S. Horsman, E. Duijm, M.J. van den Bent, T.M. Luider, J.M. Kros, P. van der Spek and P. Sillevis Smitt. *Identification of novel exons and differentially regulated splice-variants in glial brain tumors using exon expression arrays.* Cancer Res. 2007 Jun 15;67(12):5635-42. PMID: 17575129

Justine K. Peeters, Jord H.A. Nagel, Marcel Smid, Anieta M. Sieuwerts, Marijke Wasielewski, Vanja de Weerd, Anita M.A.C. Trapman-Jansen, Ans van den Ouweland, Henk Portengen, Hennie Brüggenwirth, Wilfred van IJcken, Jan G.M. Klijn, Peter J. van der Spek, John A. Foekens, John W.M. Martens, Mieke Schutte, and Hanne Meijers-Heijboer. *Gene expression profiling assigns CHEK2 1100delC breast cancers to the luminal intrinsic subtypes.* Submitted to Cancer Cell, 2008.

Peeters JK, Van der Spek PJ. *Growing applications and advancements in microarray technology and analysis tools.* Cell Biochem Biophys. 2005;43(1):149-66. Review. PMID: 16043891

Mieke Schutte, Fons Elstrodt, Elza Duijm, Jord H.A. Nagel, Antoinette Hollestelle, Marijke Wasielewski, ***Justine Peeters***, Peter van der Spek, Peter A. Sillevis Smitt & Pim J. French. *Exon expression arrays as a tool to identify human disease genes.* Final re-submission with Human Mutation. January 2007

Melles DC, Gorkink RF, Boelens HA, Snijders SV, ***Peeters JK,*** Moorhouse MJ, van der Spek PJ, van Leeuwen WB, Simons G, Verbrugh HA, van Belkum A. *Natural population dynamics and expansion of pathogenic clones of Staphylococcus aureus.* Journal of Clinical Investigation. 2004 Dec;114(12):1732-40. PMID: 15599398

Subsequent PhD publications not presented in this thesis

Jord H.A. Nagel, Antoinette Hollestelle, Marcel Smid, Suzanne Lam, Fons Elstrodt, Marijke Wasielewski, Ser Sue Ng, Pim J. French, ***Justine K. Peeters***, Marieke Rozendaal, Muhammad Riaz, Ellen Zwarthoff, Amina Teunisse, Joerg Volkland, Peter J. van der Spek, Jan G.M. Klijn, Stephen P. Ethier, Hans Clevers, Aart G. Jochemsen, Michael A. den Bakker, John A. Foekens, John W.M. Martens, and Mieke Schutte. *Distinct gene mutation profiles among luminal and basal type breast cancer cell lines.* Submitted to Cancer Cell, 2008.

Aparna Duggirala, Prashanth Kenchappa, Savitri Sharma, ***Justine K Peeters***, Niyaz Ahmed, Prashant Garg, Taraprasad Das, Seyed E Hasnain . *High-resolution genome profiling differentiated Staphylococcus epidermidis strains isolated from patients with ocular infections and normal individuals.* Invest Ophthalmol Vis Sci. 2007 Jul;48(7):3239-45. PMID: 17591894

Melles DC, Pauw E, van den Boogaard L, Boelens HA, Peters J, ***Peeters JK***, Witsenboer H, van Leeuwen WB, Verbrugh HA, van Belkum A, Nouwen JL. *Host-microbe interplay in persistent Staphylococcus aureus nasal carriage in HIV patients.* Microbes Infect. 2007 Nov 9 *PMID: 18248760*

Melles DC, Tenover FC, Kuehnert MJ, Witsenboer H, ***Peeters JK***, Verbrugh HA, van Belkum A. *Overlapping Population Structures of Nasal Isolates of Staphylococcus aureus from Healthy Dutch and American Individuals.* J Clin Microbiol. 2008 Jan;46(1):235-41. PMID: 17977984

Melles DC, van Leeuwen WB, Snijders SV, Horst-Kreft D, ***Peeters JK***, Verbrugh HA, van Belkum A. *Comparison of multilocus sequence typing (MLST), pulsed-field gel electrophoresis (PFGE), and amplified fragment length polymorphism (AFLP) for genetic typing of Staphylococcus aureus.* J Microbiol Methods. 2007 May;69(2):371-5. PMID: 17346834

J. P. Hays, R. Gorkink, G. Simons, ***J. K. Peeters***, K. Eadie, C. M. Verduin, H. Verbrugh and A. van Belkum. *High-throughput amplification fragment length polymorphism (htAFLP) analysis identifies genetic lineage markers but not complement phenotype-specific markers in Moraxella catarrhalis.* Clin Microbiol Infect 2007; 13: 55–62. PMID: 17184288

Melles DC, D. Bogaert, RFJ Gorkink, ***JK Peeters***, MJ Moorhouse, A Ott, WB vanLeeuwen, G Simons, HA Verbrugh, PWM Hermans. *Nasopharyngeal co-colonization with Staphylococcus aureus and streptococcus pneumoniae in children is bacterial genotype independent.* Microbiology, March 2007. PMID: 17322188

Melles DC, Gorkink RF, Boelens HA, Snijders SV, ***Peeters JK***, Moorhouse MJ, van der Spek PJ, van Leeuwen WB, Simons G, Verbrugh HA, van Belkum A. *Panton-Valentine leucocidin genes in Staphylococcus aureus.* Emerg Infect Dis. 2006 Jul;12(7):1174-5. PMID: 16848048

van Leeuwen WB, Melles DC, Alaidan A, Al-Ahdal M, Boelens HA, Snijders SV, Wertheim H, van Duijkeren E, ***Peeters JK***, van der Spek PJ, Gorkink R, Simons G, Verbrugh HA, van Belkum A.. *Host- and tissue-specific pathogenic traits of Staphylococcus aureus.* Journal Bacteriol. 2005 Jul;187(13):4584-91. PMID: 15968069

Juliëtte A. Severin, Endang Sri Lestari, Kuntaman Kuntaman, Damian C. Melles, Martijn Pastink, ***Justine K. Peeters***, Susan V. Snijders, Usman Hadi , D. Offra Duerink, Alex van Belkum, Henri A. Verbrugh. *Unusually High Prevalence of Panton-Valentine Leukocidin Genes among Methicillin-Sensitive Staphylococcus aureus carried in the Indonesian Population.* Accepted for Publication to Journal of Clinical Microbiology, 2008

Published Interviews

Global view of gene expression analysis. Interview with ***Justine Peeters*** and Pim French, ErasmusMC. Published in Australian Life Scientist Journal. 19/12/2007
http://www.biotechnews.com.au/index.php/id;735583971;fp;4;fpid;1017

Discovery of Novel Splice Variations Improves Glial Tumor Classification Erasmus Medical Center's Pim French and ***Justine Peeters*** talk with Noam Shomron at MIT about using exon arrays to study alternative splicing in glioma.
http://www.microarraybulletin.com/community/article.php?p=226

Book Chapters

Introduction to Gene Expression Profiling: Microarray Technology, Experimental Design and Analysis. ***Justine K. Peeters*** and Peter Valk. Edited by Willem B. van Leeuwen. School of Molecular Medicine, Rotterdam. To be published 2008

Other publications not completed within PhD

Peter Farlie., Candice Reid., Stephen Wilcox., ***Justine Peeters.,*** Gullveig Reid., Don Newgreen. *Ypel1: a novel nuclear protein that induces an epithelial-like morphology in fibroblasts.* Genes to Cells, 6:619-629, 2001.
PMID: 11473580

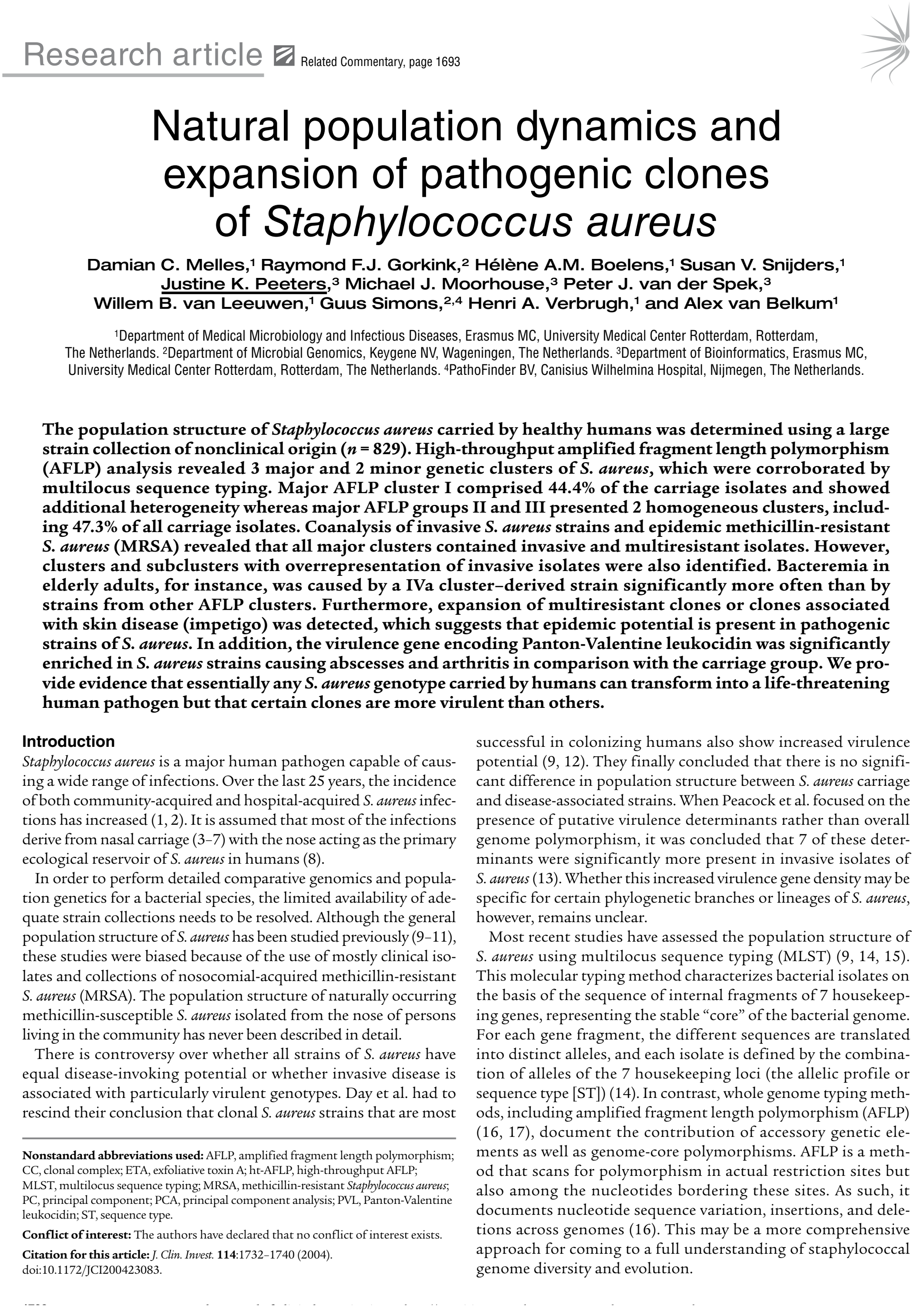
Susan Ramus, Gareth Price, ***Justine Peeters***, Katrina Bell, Mervyn Thomas, Timothy Littlejohn, Melanie de Silva, John Ciciulla, Anne-Marie Hutchins, Jane E. Armes, Deon J. Venter. *Variability of molecular pathogenetic pathways in early-onset and familial breast cancers.* Currently in writing.

Gareth Price, ***Justine Peeters***, Tiffany Cowie, John Ciciulla, Alexis Mahoney, Natalia Yarovaya, Anne-Marie Hutchins Jane E. Armes, Deon J. Venter. *Identification of candidate oncogenes defining different pathways of molecular pathogenesis in gynaecological cancers.* Currently in writing.

ABBREVIATIONS

.TIFF	Tagged Image File Format
AML	Acute Myeloid Leukemia
ANOVA	Analysis of Variance
cDNA	Complimentary Deoxyribonucleic acid
CV	Coefficient of Variation
DNA	Deoxyribonucleic Acid
EASE	Expression Analysis Systematic Explorer
EST	Expressed Sequence Tag
GCOS	Gene Chip Operating Software
GO	Gene Ontology
GUI	Graphical User Interface
IM	Ideal Match
IPA	Ingenuity Pathway Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
LOOCV	Leave One Out Cross Validation
MAS	Microarray Suite
miRNA	Micro Ribonucleic Acid
MM	Mis Match
NCHGHR	National Centre for Human Genome Research
PAM	Prediction Analysis of Microarray
PCA	Principle Component Analysis
PCR	Polymerase Chain Reaction
PM	Perfect Match
PSR	Probe Selection Region
RMA	Robust Multi Average
RNA	Ribonucleic Acid
RT-PCR	Reverse Transcription Polymerase Chain Reaction
RVM	Random Variance Model
SAGE	Serial Analysis Gene Expression
SAM	Significant Analysis of Microarray
siRNA	Silencing Ribonucleic Acid
SNP	Single Nucleotide polymorphism
SVM	Support Vector Machine
VSN	Variance Stabilizing Normalization

APPENDIX 1: Further Applications of Cluster Analysis



Natural population dynamics and expansion of pathogenic clones of *Staphylococcus aureus*

Damian C. Melles,[1] Raymond F.J. Gorkink,[2] Hélène A.M. Boelens,[1] Susan V. Snijders,[1] Justine K. Peeters,[3] Michael J. Moorhouse,[3] Peter J. van der Spek,[3] Willem B. van Leeuwen,[1] Guus Simons,[2,4] Henri A. Verbrugh,[1] and Alex van Belkum[1]

[1]Department of Medical Microbiology and Infectious Diseases, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [2]Department of Microbial Genomics, Keygene NV, Wageningen, The Netherlands. [3]Department of Bioinformatics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [4]PathoFinder BV, Canisius Wilhelmina Hospital, Nijmegen, The Netherlands.

The population structure of *Staphylococcus aureus* carried by healthy humans was determined using a large strain collection of nonclinical origin (*n* = 829). High-throughput amplified fragment length polymorphism (AFLP) analysis revealed 3 major and 2 minor genetic clusters of *S. aureus*, which were corroborated by multilocus sequence typing. Major AFLP cluster I comprised 44.4% of the carriage isolates and showed additional heterogeneity whereas major AFLP groups II and III presented 2 homogeneous clusters, including 47.3% of all carriage isolates. Coanalysis of invasive *S. aureus* strains and epidemic methicillin-resistant *S. aureus* (MRSA) revealed that all major clusters contained invasive and multiresistant isolates. However, clusters and subclusters with overrepresentation of invasive isolates were also identified. Bacteremia in elderly adults, for instance, was caused by a IVa cluster–derived strain significantly more often than by strains from other AFLP clusters. Furthermore, expansion of multiresistant clones or clones associated with skin disease (impetigo) was detected, which suggests that epidemic potential is present in pathogenic strains of *S. aureus*. In addition, the virulence gene encoding Panton-Valentine leukocidin was significantly enriched in *S. aureus* strains causing abscesses and arthritis in comparison with the carriage group. We provide evidence that essentially any *S. aureus* genotype carried by humans can transform into a life-threatening human pathogen but that certain clones are more virulent than others.

Introduction

Staphylococcus aureus is a major human pathogen capable of causing a wide range of infections. Over the last 25 years, the incidence of both community-acquired and hospital-acquired *S. aureus* infections has increased (1, 2). It is assumed that most of the infections derive from nasal carriage (3–7) with the nose acting as the primary ecological reservoir of *S. aureus* in humans (8).

In order to perform detailed comparative genomics and population genetics for a bacterial species, the limited availability of adequate strain collections needs to be resolved. Although the general population structure of *S. aureus* has been studied previously (9–11), these studies were biased because of the use of mostly clinical isolates and collections of nosocomial-acquired methicillin-resistant *S. aureus* (MRSA). The population structure of naturally occurring methicillin-susceptible *S. aureus* isolated from the nose of persons living in the community has never been described in detail.

There is controversy over whether all strains of *S. aureus* have equal disease-invoking potential or whether invasive disease is associated with particularly virulent genotypes. Day et al. had to rescind their conclusion that clonal *S. aureus* strains that are most successful in colonizing humans also show increased virulence potential (9, 12). They finally concluded that there is no significant difference in population structure between *S. aureus* carriage and disease-associated strains. When Peacock et al. focused on the presence of putative virulence determinants rather than overall genome polymorphism, it was concluded that 7 of these determinants were significantly more present in invasive isolates of *S. aureus* (13). Whether this increased virulence gene density may be specific for certain phylogenetic branches or lineages of *S. aureus*, however, remains unclear.

Most recent studies have assessed the population structure of *S. aureus* using multilocus sequence typing (MLST) (9, 14, 15). This molecular typing method characterizes bacterial isolates on the basis of the sequence of internal fragments of 7 housekeeping genes, representing the stable "core" of the bacterial genome. For each gene fragment, the different sequences are translated into distinct alleles, and each isolate is defined by the combination of alleles of the 7 housekeeping loci (the allelic profile or sequence type [ST]) (14). In contrast, whole genome typing methods, including amplified fragment length polymorphism (AFLP) (16, 17), document the contribution of accessory genetic elements as well as genome-core polymorphisms. AFLP is a method that scans for polymorphism in actual restriction sites but also among the nucleotides bordering these sites. As such, it documents nucleotide sequence variation, insertions, and deletions across genomes (16). This may be a more comprehensive approach for coming to a full understanding of staphylococcal genome diversity and evolution.

Nonstandard abbreviations used: AFLP, amplified fragment length polymorphism; CC, clonal complex; ETA, exfoliative toxin A; ht-AFLP, high-throughput AFLP; MLST, multilocus sequence typing; MRSA, methicillin-resistant *Staphylococcus aureus*; PC, principal component; PCA, principal component analysis; PVL, Panton-Valentine leukocidin; ST, sequence type.

Conflict of interest: The authors have declared that no conflict of interest exists.

Citation for this article: *J. Clin. Invest.* **114**:1732–1740 (2004). doi:10.1172/JCI200423083.

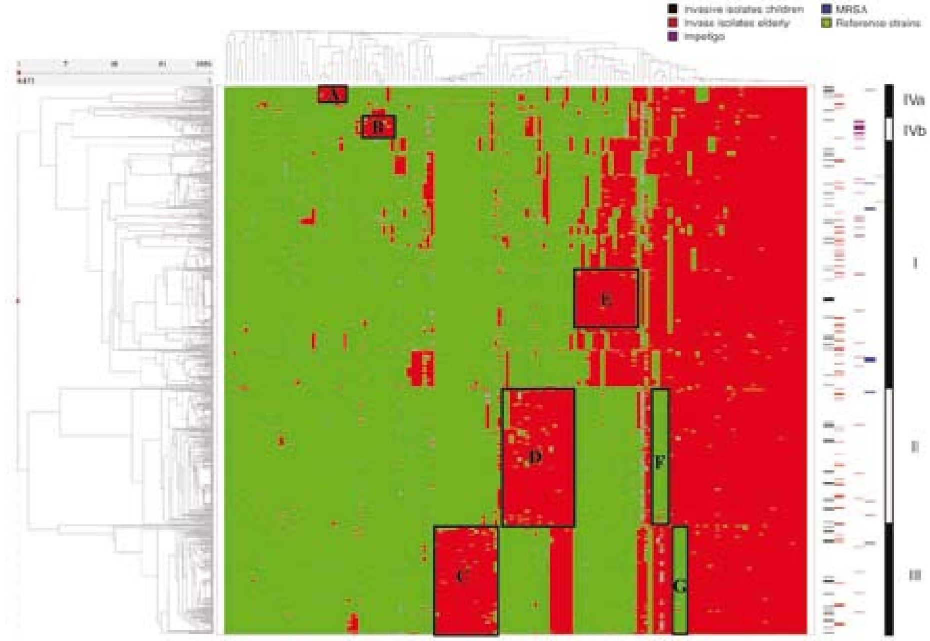


Figure 1

2D hierarchical clustering of the 1,056 *S. aureus* strains. The green/red figure represents 155,232 binary outcomes generated by ht-AFLP with 147 marker fragments. Marker absence corresponds with green and marker presence with red. The dendrogram on the *y* axis represents the phylogenetic clustering of the 1,056 strains. The dendrogram on the *x* axis shows the clustering of the 147 AFLP markers, many of which segregate in specific groups. These groups are cluster specific, and some of these groups are shown as boxes in the figure (A through G). The colored, striped bars on the right represent the distribution of the invasive strains (children and elderly adults), the impetigo isolates, the MRSA strains, and the reference strains. The carriage strains (*n* = 829) are not pointed out separately. In conjunction with PCA 3 major (I, II, III) and 2 minor (IVa, IVb) branches were identified; these are represented by the black and white bar on the right of the figure.

We collected over 2,000 nonclinical *S. aureus* isolates from various groups of healthy individuals in the Dutch community over an extended period of time, creating a unique population-based strain collection. To assess differences in the virulence potential of various strains of *S. aureus*, insight into the natural (nonclinical) population structure is essential. We here present whole genome scanning by high-throughput AFLP (ht-AFLP) of a random selection of these *S. aureus* strains (*n* = 829). Strains were derived from the anterior nares of healthy children (1–18 years) and healthy elderly adults (>55 years) from the Rotterdam area (The Netherlands) (18, 19). To investigate genomic differences between these carriage strains and invasive isolates, contemporary isolates from blood, deep-seated, and soft-tissue infections from the same geographical area were included (*n* = 164). Furthermore, a collection of international epidemic MRSA strains (*n* = 21) (20) and *S. aureus* isolates from Rotterdam children with nonbullous impetigo (*n* = 40) were included (21). The prevalence of the clinically relevant *mecA* and Panton-Valentine leukocidin (PVL) genes was determined for all strains as well.

Results

Genetic diversity of S. aureus. Using the set of 1,056 *S. aureus* strains, a total of 155,232 AFLP fragments were generated, covering 147 different marker fragments per strain. These outcomes are visualized in Figure 1. The dendrogram on the left (*y* axis) reveals bacterial clustering. The bar on the right of this figure delineates the presence of 3 major (I, II, III) and 2 minor (IVa, IVb) branches, as identified by principal component analysis (PCA) (Figure 2, A and B). Unsupervised cluster analysis of the 1,056 strains (Figure 3) clearly demonstrates that the AFLP data represent 2 distinct, homogenous major clusters (II and III) and, conversely, several other smaller subclusters that could be divided into 1 major and 2 minor AFLP clusters by PCA (I, IVa, IVb). MLST analysis indeed revealed additional heterogeneity in AFLP group I, identifying different clonal complexes (CCs) [CC5, CC8, CC15 (Figure 4)]. CCs are defined as clusters of closely related STs where single differences in the allelic profile are tolerated (9). In contrast, AFLP clusters II and III harbor single CCs, CC30 and CC45, respectively. These 2 major CCs embrace almost half (47.3%) of all carriage isolates. Clusters IVa and

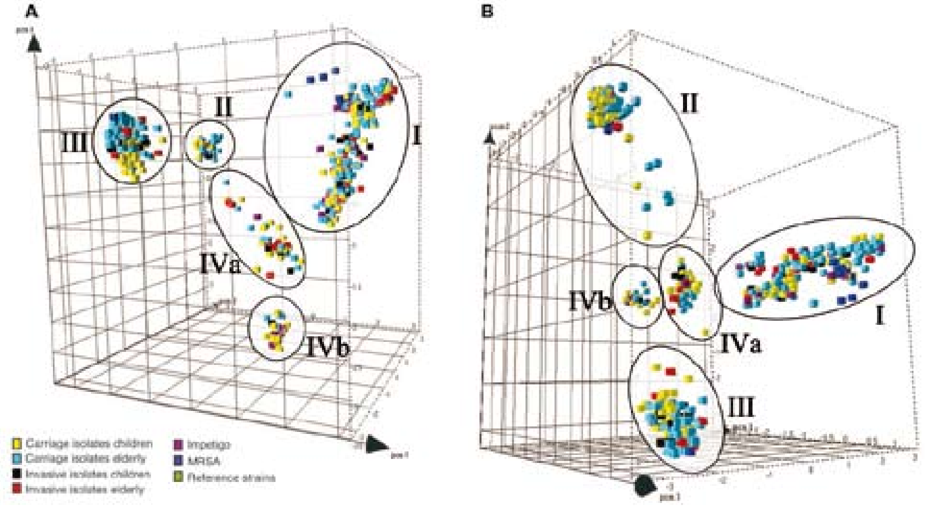


Figure 2
PCA of the 1,056 *S. aureus* strains. The different cubes (plotted in 3D space), colored according to the source, represent every strain in the study. Each axis represents the score calculated for that strain on each PC. The distribution of the strains is shown from 2 different angles (**A** and **B**). The 5 circles indicate the different phylogenetic AFLP clusters.

IVb are associated with CC22 and CC121, respectively (Figure 4). The unsupervised correlation-analysis showed that minor AFLP group IVa also consists of different subclusters (Figure 3).

The horizontal dendrogram in Figure 1 shows the clustering of the 147 AFLP markers, many of which segregate in separate groups. These groups are cluster specific; some of these are shown as boxes A–G (Figure 1). The colored bars on the right represent the distribution of the invasive strains (in children and elderly adults), the impetigo isolates, the MRSA strains, and the reference strains across the dendrogram. The central area of this figure best defines the 5 phylogenetic lineages. The clustered green and red boxes show components of genomic diversity within the 5 *S. aureus* subpopulations. The distributions of the carriage isolates as well as the invasive isolates are also visualized in Figure 3.

Carriage versus disease-causing strains of S. aureus. Among the *S. aureus* isolates from healthy individuals with nasal carriage, 3 major genetic clusters (I, II, III) could be identified, comprising 760 of the 829 (91.7%) carriage strains. Two minor clusters (IVa, IVb) embrace the remaining 69 (8.3%) carriage strains. All 5 clusters coherently contain carriage strains isolated from children as well as elderly adults (Figure 5). The distribution of the pediatric and geriatric strains across the 5 AFLP clusters was overlapping (Figure 3). However, healthy children more often carried strains from cluster IVa+b than did healthy elderly adults (Fisher's exact test, $P < 0.0001$), whereas cluster I strains were slightly overrepresented in elderly adults (Fisher's exact test, $P = 0.03$). Although carriage isolates from children and elderly adults were equally divided in AFLP cluster II (Figure 5), a clone strongly associated with carriage isolates in children could be identified (Fisher's exact test, $P < 0.0001$) (area between the dotted lines in Figure 3).

Contemporary invasive *S. aureus* strains ($n = 164$) from children and elderly adults were distributed across all phylogenetic branches and widely scattered across the AFLP dendrogram (Figures 1 and 3). The population structure of contemporary carriage isolates and invasive isolates in the same geographical area appeared to be strongly overlapping (Figure 5). However, bacteremia in elderly adults was significantly more often caused by a IVa cluster–derived strain (Fisher's exact test, $P = 0.0095$). Additional analysis of AFLP cluster IVa showed that this group consists of distinct subclusters (Figure 3). Therefore it would be incorrect to define this minor cluster as a single invasive clone. Apparently, several subclusters, plotted next to each other by 3D (AFLP cluster IVa), contain proportionately more bacteremia isolates from elderly adults than carriage isolates from the same group. Statistical analysis of major cluster I revealed 2 subclusters potentially associated with invasive disease. Subcluster If (Figure 3) contains proportionately more bacteremia isolates from elderly adults ($n = 5$; 5.6%) in comparison with carriage isolates from elderly adults ($n = 6$; 1.4%) (Fisher's exact test, $P = 0.027$). Subcluster Ia shows overrepresentation of invasive isolates from both children and elderly adults ($n = 11$; 6.7%) in comparison with carriage isolates from both groups ($n = 20$; 2.4%) (Fisher's exact test; $P = 0.01$) (Figure 3).

The invasive strains from children in major cluster I are associated with hospital-acquired disease (Fisher's exact test, $P = 0.01$) (data not shown). No significant difference was found in the distribution of isolates from individuals with invasive community-acquired disease versus invasive hospital-acquired disease in the other clusters.

All 5 clusters contain *S. aureus* strains isolated from children with community-acquired nonbullous impetigo. The distribution is shown in Figures 1 and 5. Compared to pediatric carriage strains,

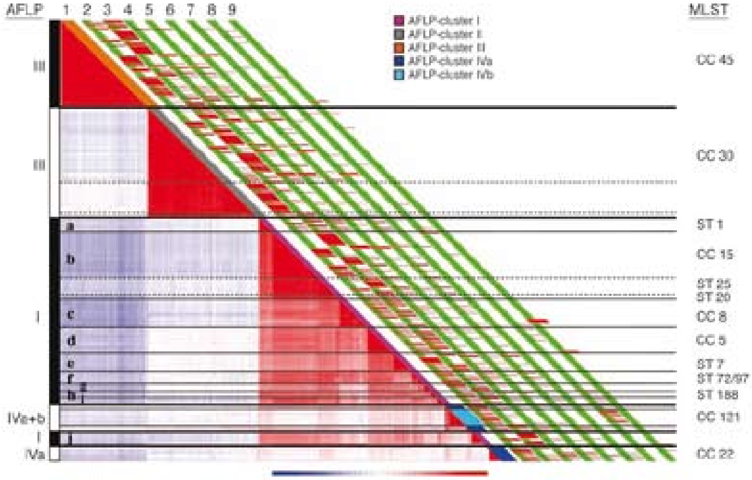


Figure 3

Cluster analysis of the 1,056 *S. aureus* strains using OmniViz. The cells in the correlation visualization are colored by Pearson's correlation coefficient values with deeper colors indicating higher positive (red) or negative (blue) correlations. The scale bar (underneath the figure) indicates 100% correlation (red) toward 100% anticorrelation (blue). In order to reveal correlation patterns, a matrix-ordering method was applied to rearrange the samples. The OmniViz correlation view generated with 1,056 strains was adapted so that descriptive (clinical) parameters could be plotted directly adjacent to the original diagonal. The black and white bar on the left indicates the 5 AFLP groups based on PCA. This figure shows additional subclustering in major group I (a–j) as well as in minor group IVa, indicated by several lines. The dotted lines identify blocks of minimal changes in 1 cluster or subcluster. The corresponding MLST data (see also Figure 4) are shown on the right side of the figure. The distributions of the strains from different origins are visualized as red lines in the diagonal red and green bars of the figure (numbered 2–9). Variable 1 indicates the different AFLP clusters based on PCA; 2, carriage isolates, children (n = 400); 3, carriage isolates, elderly adults (n = 429); 4, invasive isolates, total (n = 164); 5, invasive isolates, children (n = 74); 6, invasive isolates, elderly adults (n = 90); 7, invasive isolates, children (deep-seated and soft-tissue infections) (n = 18); 8, impetigo isolates (n = 40); 9, MRSA (n = 21).

impetigo isolates were more frequently found in cluster IVb (Fisher's exact test, $P < 0.0001$) and less frequently in cluster II (Fisher's exact test, $P = 0.01$) (Figure 5), suggesting clonal expansion of a certain genotype associated with impetigo (22).

MecA- and PVL-positive strains. The 21 international epidemic MRSA strains are present in several clusters and subclusters of *S. aureus* (Ic, Id, Ij, II, III, IVa). This strain collection comprises epidemic MRSA from Belgium, Finland, France, Greece, Spain, Germany, and the United Kingdom (20). AFLP cluster I contained disproportionately more MRSA isolates as compared to the other clusters (Fisher's exact test, $P = 0.006$) (Figure 5), suggesting that many of these epidemic MRSA strains are derived from a common cluster I ancestor. Notably, most of the 21 MRSA strains are located in subcluster Ic (n = 9; 42,9%; Fisher's exact test, $P < 0.0001$)].

None of the Dutch carriage and clinical *S. aureus* isolates (n = 1033) included in this study harbor the *mecA* gene, which is consistent with the reported low MRSA prevalence in the Dutch population (23). Four PVL-positive *S. aureus* strains (1.0%) were found in the pediatric carriage group (n = 400) and 1 (0.2%) in the elderly adults carriage group (n = 429). Three of 146 (2.1%) blood-culture isolates carried the PVL gene, 2 of which derived from pediatric patients. Seven of the 18 (38.9%) invasive strains isolated from deep-seated or soft-tissue infections in children were PVL positive. There was no significant difference in the presence of PVL when comparing the carriage isolates and invasive blood-culture isolates. In contrast, *S. aureus* strains causing abscesses and arthritis were significantly enriched in the presence of PVL (38.9%) in comparison with the pediatric carriage group (1.0%) (Fisher's exact test, $P < 0.0001$) and in comparison with the pediatric bacteremia isolates (3.6%) (Fisher's exact test, $P = 0.0005$). All impetigo strains (n = 40) were PVL negative.

Sequence assessment of AFLP markers. To determine the origin of genetic polymorphism, the nucleotide sequence of a set of 81 AFLP markers was established, 60 of which were located in cluster-specific marker boxes (Figure 1, boxes A–G). Nineteen of the remaining 21 markers were not cluster specific and were present in almost all 1,056 *S. aureus* strains (red area on the right side of box G, Figure 1). Interestingly, 66 (81.5%) of the 81 markers showed homology to all 3 completed (MW2, Mu50, N315) and 4 unfinished (252, 476, COL, NCTC 8325) genome sequences of *S. aureus*. Conversely, 7 (8.6%) of the 81 markers appear to be absent in all 3 completed genomes (MW2, Mu50, N315) of *S. aureus* (see Supplemental Table 1; supplemental material available at http://www.jci.org/cgi/content/full/114/12/1732/DC1). Of these 7 marker fragments, 3 showed homology to the *S. aureus* strain 252 epidemic MRSA-16 (EMRSA-16), 1 to bacteriophage Φ exfoliative toxin A (ETA) DNA, and 1 to *S. aureus* TY4 exfoliative toxin B (ETB) plasmid DNA. Only 2 did not match with any GenBank entry and may represent novel (or hypermutable) *S. aureus* genome segments. The 15 markers that

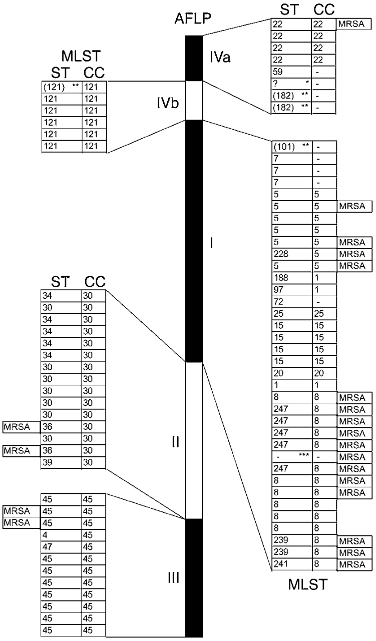


Figure 4
AFLP analysis versus MLST analysis of *S. aureus*. The bar in the center of the figure represents the 1,056 strains divided into the 5 phylogenetic AFLP clusters (similar to those defined at the right side of Figure 1). MLST data is shown for 77 *S. aureus* strains, which are spread over the different AFLP clusters. The order of the MLST sequence types in this figure is determined by the location of the strain in the AFLP dendrogram (Figure 1). *Unknown ST; **6 of the 7 loci are similar to the particular ST; ***data not available.

appear to be absent in at least 1 of the known genome sequences of *S. aureus* may be associated with (larger) deletions, the existence of which has been described before (24).

The cluster-specific marker boxes are associated with different AFLP clusters. Box A is associated with a part of cluster IVa, B with cluster IVb, C (present) and G (absent) with cluster III, D (present) and F (absent) with cluster II, and box E with cluster I (Figure 1). The majority (78.0%) of the markers present in boxes A–D (not present in cluster I) showed homology to the reference strains Mu50 and N315, which are located in AFLP cluster I. This suggests that these markers become cluster specific by point mutations rather than genomic rearrangement (deletions or insertions). This is also supported by additional data (Supplemental Figure 1), which show that cluster-defining markers in each box are widely spread across the *S. aureus* chromosome.

Cluster-specific box B, associated with cluster IVb (with overrepresentation of impetigo-isolates), comprises both virulence genes bacteriophage Φ ETA DNA and ETB plasmid DNA (Supplemental Table 1).

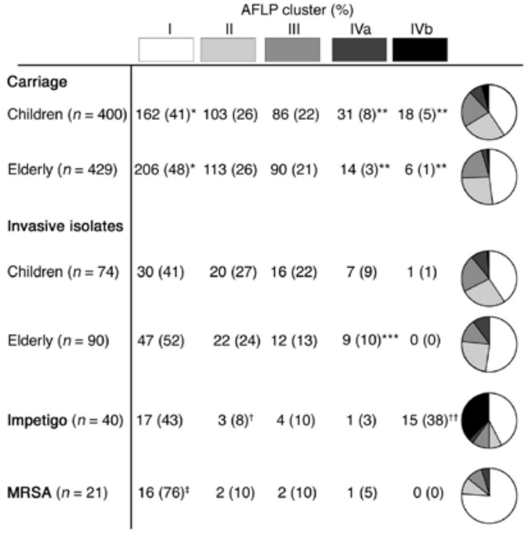
Discussion

The incidence of *S. aureus* bacteremia is rising and has more than doubled over the past 25 years in some Western European countries. This increase coincides with a growing rate of community-acquired disease (in proportion to hospital-acquired disease) and the epidemic emergence of nosocomial MRSA strains (2). Elderly adults are most frequently affected, particularly those with additional predisposing risk factors. Although a part of this effect may be explained by host susceptibility and population aging, this does not fully explain the current and drastic rise in the number of infections. However, little is known about possible bacterial determinants and whether or not these are associated with changes in the virulence of *S. aureus*.

We previously showed that AFLP analysis using optimal enzyme and primer combinations is an excellent tool for assessing genetic polymorphism in the clonal microorganism *M. tuberculosis* (17). For *S. aureus* we used the enzyme combination *Mbo*I-*Csp*6I, which resulted in a fingerprint of about 70 polymorphic AFLP fragments well distributed within the size range of 100 bp to 600 bp in a single AFLP reaction. However, a potential limitation of this AFLP approach is in the randomness of the restriction sites for *Mbo*I and *Csp*6I. For instance, genomic islands with underrepresentation of these restriction sites will not be fully scanned for polymorphism. However, if currently known genomic sequences of *S. aureus* (Mu50 and N315) are analyzed by computer for the occurrence of these sites, the average number of fragments (useful for AFLP) generated per genome is 4.373, and the average length of the fragments is 200 bp. This suggests that coverage is indeed random and, in this respect, AFLP provides more of a whole genome–scanning approach than MLST, for instance.

In the present study, the population structure of *S. aureus*, isolated from the nose of healthy individuals in the Rotterdam area (The Netherlands), has been determined. Using ht-AFLP, we analyzed 147 polymorphic markers for 1,056 *S. aureus* strains. Two large unbiased strain collections of a nonclinical origin were used. These collections were obtained from children (<19 years) and elderly adults (>55 years) with nasal carriage of *S. aureus*. ht-AFLP analysis revealed the existence of 3 major (I, II, III) and 2 minor (IVa, IVb) phylogenetic branches. Major AFLP group I could be subdivided into 10 different subclusters, indicating its heterogeneity. In contrast, major clusters II and III showed a very homogenous nature.

Carriage isolates were not completely randomly distributed over these clusters. AFLP cluster IVa+b represents more carriage strains from children (isolated in 2002) than carriage strains from elderly adults (isolated from 1997–1999). A clonal expansion associated with carriage isolates in children was also observed in AFLP group II. Conversely, AFLP group I embraces more carriage strains from elderly adults. Apparently, a certain degree of bacterial population heterogeneity exists between the 2 groups included. Whether the different sampling moments or the different age categories are fundamental to the observed differences is currently unclear. Danish studies from the late 1970s demonstrated that waves of phage types of *S. aureus* go through human populations (25, 26). Such a phenomenon may explain our current observations. However, the proportionate distribution of the major phylogenetic branches (I, II, III) within the overall population of *S. aureus*, isolated from humans, appears to be fairly stable over time and comparable for children and elderly adults.



	AFLP cluster (%) I	II	III	IVa	IVb
Carriage					
Children (n = 400)	162 (41)*	103 (26)	86 (22)	31 (8)**	18 (5)**
Elderly (n = 429)	206 (48)*	113 (26)	90 (21)	14 (3)**	6 (1)**
Invasive isolates					
Children (n = 74)	30 (41)	20 (27)	16 (22)	7 (9)	1 (1)
Elderly (n = 90)	47 (52)	22 (24)	12 (13)	9 (10)***	0 (0)
Impetigo (n = 40)	17 (43)	3 (8)†	4 (10)	1 (3)	15 (38)††
MRSA (n = 21)	16 (76)‡	2 (10)	2 (10)	1 (5)	0 (0)

Figure 5

Distribution of *S. aureus* strains in the 5 phylogenetic branches. *Overrepresentation of carriage in elderly adults (Fisher's exact test, P = 0.01); **overrepresentation of carriage in children (Fisher's exact test, P < 0.0001); ***proportionately more bacteremia-associated strains from elderly adults as compared to carriage strains from the same group (3% vs. 10%; Fisher's exact test, P = 0.0095); †proportionately fewer impetigo-associated strains as compared to carriage in children (8% vs. 26%; Fisher's exact test, P = 0.01); ††overrepresentation of impetigo-associated strains as compared to carriage in children (38% vs. 5%; Fisher's exact test, P < 0.0001); ‡proportionately more MRSA strains as compared to all carriage isolates (76% vs. 44%; Fisher's exact test, P < 0.006).

The 5 AFLP clusters identified in this study match with the major CCs as defined by MLST (http://www.mlst.net/). These MLST-based CCs have been defined by studying carriage, invasive, and MRSA isolates mainly from the United Kingdom. The top 5 CCs in the MLST database are CC8, CC30, CC5, CC22, and CC45 (9, 15, 27, 28). We studied more than 1,000 strains isolated in the Rotterdam region (The Netherlands), and we identified essentially the same CCs. Apparently, these clonal clusters have spread successfully in the United Kingdom and The Netherlands and probably worldwide. All large-scale molecular typing studies of nonclinical isolates of *S. aureus* have been performed using geographically biased strain collections, including our present analysis. However, considering the overlap in MLST types and the similarity in prevalence of certain major clonal clusters, it is supposed that geographical bias is not a confounding factor. An ongoing analysis of strains derived from Indonesian carriers corroborated this hypothesis. The Indonesian strains clustered in the same groups (I to IV), although there was a difference in the relative numbers of isolates per cluster. No new AFLP clusters were identified (Melles et al., unpublished data).

ht-AFLP clusters II and III, identical to MLST CCs 30 and 45, respectively, account for almost half (47%) of all carriage isolates in our population, which suggests that these 2 CCs have evolved to be very successful in colonizing humans.

The above-mentioned observations suggest that virulent and nonvirulent strains are probably not fundamentally different from each other: clinical isolates and MRSA from international sources fall into the same main clusters as carriage isolates. In addition, invasive *S. aureus* strains, mainly blood-culture isolates (89%), were found to be widely distributed across all 5 AFLP groups in this study. This suggests that strains from each of the genetic clusters are essentially able to cause invasive disease. On the other hand, 1 minor AFLP cluster (IVa) and 2 AFLP subclusters (Ia and If) contained proportionately more invasive isolates. Direct cross-infection is not a likely explanation for this finding because these epidemiologically nonrelated invasive strains were isolated from patients in different medical departments over a period of several years. There have been controversial reports on the existence of hypervirulent lineages of *S. aureus* (9, 12), but our data suggest that not all *S. aureus* strains share the same invasive potential. This is not in agreement with recent findings from the Oxford (United Kingdom) region where MLST data suggest that virulence may not be associated with clonal lineages (15). Rather, a relationship between genetic background and disease type is thought to be primarily dependent on the presence of certain toxin genes only (29).

Table 1

Number of *Staphylococcus aureus* strains included in this study

	Children (n)	**Elderly adults (n)**	**Total (n)**
Carriage	400	429	829
Invasive — blood culture			
Hospital acquired	43	68	111
Community acquired	13	22	35
Invasive — deep-seated or soft-tissue infection			
Hospital acquired	4	–	4
Community acquired	12	–	12
Unknown	2	–	2
Impetigo	40	–	40
MRSA	–	–	21
Reference strains	–	–	2
Total	514	519	1056

S. aureus strains isolated in the Rotterdam area from patients with nonbullous impetigo showed less clonal diversity than bacteremia-associated strains. Although each of the 5 clusters was found to contain impetigo-derived isolates, AFLP cluster IVb clearly represented significantly more isolates causing impetigo. An explanation for this clonal expansion in impetigo could be the facile spread of this disease. A recent study by Koning et al. concerning nonbullous *S. aureus* impetigo concluded that a combination of staphylococcal virulence and resistance genes determines the development and course of nonbullous impetigo (22).

The 21 international epidemic MRSA strains included in this study were spread across several lineages, indicating that methicillin resistance has developed in distinct phylogenetic subpopulations of *S. aureus*, which has been described before (10, 30). MLST studies have placed most of the MRSA in 5 major clusters: CC5, CC8, CC22, CC30, and CC45 (10, 27, 31, 32). Figure 4 shows similar dissemination for strains from a nonclinical origin as well. Notably, computer algorithms used to solve the *S. aureus* population structure based on MLST data (eBURST; ref. 28) generate clustering identical to that found in the AFLP Spotfire and OmniViz analyses. This provides a solid experimental and mathematical framework for all conclusions drawn from our AFLP data.

We have also assessed the prevalence of *mecA* and the PVL genes in the carriage and disease-causing populations of *S. aureus*. All carriage isolates (*n* = 829) were *mecA* negative, corroborating data showing the insignificant spread of MRSA in the Dutch community (23). Also, all clinical isolates (*n* = 204) were *mecA* negative, as opposed to proportions of above 10% in many of the other European countries, including those sharing borders with The Netherlands (33). PVL is a toxin associated with skin infections (furuncles), community-acquired MRSA infections, and necrotizing pneumonia (31, 34, 35). PVL prevalence in a *S. aureus* population of nonclinical isolates has never been studied accurately. We found a very low prevalence of 0.6% in a large (*n* = 829) carriage collection. In this study PVL was carried in 2.1% of blood-culture isolates. However, a significantly higher prevalence of PVL (38.9%) was found in *S. aureus* strains causing abscesses and arthritis. This is in agreement with the proposed involvement of PVL in severe and invasive (soft-tissue) staphylococcal infections (31, 35).

Our sequence analysis for the clustered AFLP markers suggested that genetic diversity among clusters is primarily caused by point mutation rather than by large-scale deletions or insertions (15). Ultimate proof for this hypothesis should be provided by detailed physical mapping and large-scale sequencing studies, however. Furthermore, we provide indirect proof that 7 genome sequences quite accurately represent the genetic potential of *S. aureus* as a species; only 4 of 81 marker sequences did not match with the 7 known *S. aureus* whole genome sequences.

In conclusion, we have solved the population structure of *S. aureus* of nonclinical origin. Three major and 2 minor phylogenetic branches were identified in our geographically restricted group. Inclusion of invasive *S. aureus* strains and international-epidemic MRSA revealed that within all major clusters, invasive and multiresistant isolates could be identified. However, clusters and subclusters with overrepresentation of bacteremia-associated isolates were identified. Expansion of multiresistant clones or clones associated with skin disease (impetigo) was observed as well. We suggest that essentially any *S. aureus* genotype that is carried by humans can transform into a life-threatening human pathogen, but strains from some clonal lineages are more virulent than others.

Methods

Bacterial strains. Two strain collections provided nonclinical *S. aureus* carriage isolates from healthy individuals. These collections were obtained from 2 study cohorts involving children and elderly adults. In addition, various clinical isolates were included. Contemporary invasive *S. aureus* strains, isolated from children and elderly adults from the same geographic region, were cultured from normally sterile sites in hospitalized patients with clinical signs of *S. aureus* infection. Community-acquired invasive disease was defined as isolation of *S. aureus* from patients within 48 hours of admission; hospital-acquired was defined as isolation of *S. aureus* 48 hours or longer after admission. The different subcollections are described in more detail below.

In total, 3,198 children from Rotterdam (The Netherlands), aged between 1 and 19 years and participating in the national 2002 Meningococcal Vaccination Campaign, were enrolled (19). A team of 10 research nurses and medical doctors obtained a single nasopharyngeal swab per child at the time of vaccination. *S. aureus* was isolated from 1,116 children. All isolates were stored at –80°C in broth containing glycerol. A random sample of 400 *S. aureus* carriage isolates was drawn.

The second collection originated from a community-based prospective study of elderly adults in Rotterdam (The Netherlands) (18). From 3,851 persons aged over 55 years, nasal swab cultures were obtained between April 1, 1997, and December 31, 1999. *S. aureus* strains were isolated from 1,043 elderly adults. All isolates were stored at –80°C in glycerol containing broth. A random sample of 429 carriage isolates was drawn.

Seventy-four clinical *S. aureus* isolates were retrospectively collected from children with invasive *S. aureus* disease identified in Sophia Children's University Hospital (Rotterdam, The Netherlands) (2000–2002). Fifty-six isolates derived from blood cultures and 18 isolates were obtained from deep-seated (arthritis; *n* = 4) or soft-tissue (abscess; *n* = 14) infections. Ninety clinical isolates from elderly adults (>55 years) were obtained from persons with *S. aureus* bacteremia identified in Erasmus MC (Rotterdam, The Netherlands) (1997–1999). Forty *S. aureus* strains obtained from lesions of children suffering from impetigo were randomly drawn from a collection described by Koning et al. (21).

Twenty-one international epidemic MRSA strains were obtained from the HARMONY collection (http://www.harmony-microbe.net), described by Murchan et al. (20). Finally, we included 2 reference strains. N315 is an MRSA strain isolated in 1982, and Mu50 is an MRSA strain with reduced susceptibility to vancomycin isolated in 1997. For both strains genome sequences have been determined (36). In total, 1,056 *S. aureus* strains were included (Table 1).

Cultures, DNA isolation, and detection of mecA and PVL genes. Bacteria were grown overnight at 37°C on Columbia III agar (BD) supplemented with 5% sheep blood. Three to 5 colonies were suspended in TEG buffer (25 mM Tris, 10 mM EDTA, 50 mM glucose) containing lysostaphin (50 μg/ml) and incubated at 37°C for 1 hour. DNA was extracted with the MagNA Pure LC DNA Isolation Kit III (Bacteria, Fungi) using the MagNA Pure LC Instrument (Roche Diagnostics) and stored at –20°C. We determined the presence of the *mecA* and PVL genes by PCR (34, 37).

AFLP. AFLP analysis has been performed as described by Van den Braak et al. (17). Using the predictive software package Recomb (Keygene NV) (38), the optimal enzyme and primer combinations were selected. Bacterial DNA was digested with the enzymes *Mbo*I and *Csp*6I, and the linker oligonucleotide pair for *Mbo*I (5′-CTCGTAGACTGCGTACC-3′ and 5′-GATCGGTACGCAGTCTAC-3′) and for *Csp*6I (5′-GACGATGAGTCCTGAC-3′ and 5′-TAGTCAGGACTCAT-3′) were ligated. Subsequently, a nonselective preamplification was performed using the *Mbo*I primer (5′-GTAGACTGCGTACCGATC-3′) and *Csp*6I primer (5′-GACGATGAGTCCTGACTAC-3′). In the final amplification, a ^{33}P-labeled *Mbo*I primer containing 1 selective nucleotide (either +C or +G) and a *Csp*6I primer containing 2 selective nucleotides (+TA) were used. Amplified material was analyzed using standard polyacrylamide slabgels and subsequent autoradiography. Marker fragments were scored and a binary table scoring marker fragment absence (0) or presence (1) was constructed.

After excision of some selected AFLP fragments from dried gels, reamplification followed by double-strand sequence analysis was performed (17, 38). The sequence of several additional fragments was determined by computer analysis. The size of the fragments in combination with the selective nucleotides of the AFLP primers facilitated adequate mapping of the fragments on the staphylococcal genome sequence. These fragments were further analyzed by BLAST searching (http://www.ncbi.nlm.nih.gov/BLAST/) (39) against the 3 completed (MW2, NC_003923.1; Mu50, NC_002758.1; N315, NC_002745.2) and 4 unfinished (252, NC_002952; 476, NC_002953; COL, NC_002951; NCTC 8325, NC_002954) genomic sequences of *S. aureus*. The BLAST results enabled computer-mediated genomic localization and gene annotations of the AFLP fragments.

MLST. MLST was carried out for 56 *S. aureus* strains using DNA arrays (40). The selected strains were equally distributed across the AFLP dendrogram by selecting 1 out of 10 carriage or invasive strains isolated from children, going from top to bottom through the AFLP dendrogram (Figure 1). MLST data for the 21 epidemic MRSA strains are available at the MLST home page (http://www.mlst.net/) (20).

Data analysis. The method used for 2D clustering of the AFLP data was agglomerative (successive) hierarchical. This was performed using the unweighted pair group method with arithmetic mean (UPGMA). The similarity metric used was Tanimoto (Spotfire DecisionSite 7.2; Spotfire), which defines similarity for binary data (0 and 1) based on the number of positive attributes that 2 records have in common. The resulting dendrogram was ordered by average value.

PCA is a standard multivariate method used to reduce the dimensional space of the data to its principal components (PCs) (41, 42). PCA aims to reduce a large number of variables that explain most of the variation in the data (43). It is basically a rotation of axes after centering data to the means of the variables, the rotated axes being the PCs, which are linear combinations of the original variables. The PC computation is displayed as a 3D scatter plot in which the position along the axes shows the PCA score of the strain. PCA was used to identify subgroups of AFLP clusters as hidden by 2D representation of hierarchical clustering. The distribution of the strains in the 5 phylogenetic branches was defined on the basis of PCA. Hierarchical cluster analysis and PCA were performed using Spotfire DecisionSite 7.2 software.

The OmniViz package (OmniViz Inc.) was used to perform and visualize the results of unsupervised cluster analysis in a correlation visualization. This correlation visualization tool displays pairwise correlations among the samples calculated by Pearson's correlation coefficient (44). In order to reveal correlation patterns, a matrix-ordering method is applied to rearrange the samples. The ordering algorithm starts with the most correlated sample pair and, through an iterative process, sorts all the samples into correlated blocks. Each sample is joined to a block in an ordered manner so that a correlation trend is formed within a block with the most correlated samples at the center. The blocks are then positioned along the diagonal of the plot in a similar ordered manner. As the resultant visualization is symmetrical about the diagonal, half the matrix display is replaced by appropriate clinical data.

To compare the distribution of strain categories in different phylogenetic lineages, Fisher's exact test was used. A 2-sided *P* value of less than 0.05 was considered significant.

Acknowledgments

We gratefully acknowledge all participants, personnel, and researchers of the nasopharyngeal carriage study in healthy Dutch children and the community-based prospective cohort study in elderly adults for participating in and facilitating this study. Furthermore, we thank Alewijn Ott for statistical advice. The research described in this communication has in part been facilitated by a grant provided by the Dutch Ministry of Economic Affairs (BTS 00145). AFLP is a registered trademark of Keygene NV, and the AFLP technology is covered by patents (US006045994A, EP0534858B1) and patent applications owned by Keygene NV.

Received for publication August 19, 2004, and accepted in revised form October 19, 2004.

Address correspondence to: Damian C. Melles, Erasmus MC, University Medical Center Rotterdam, Department of Medical Microbiology & Infectious Diseases, Room L-313, Dr Molewaterplein 40, 3015 GD Rotterdam, The Netherlands. Phone: 31-10-463-3510; Fax: 31-10-463-3875; E-mail: d.melles@erasmusmc.nl.

1. Steinberg, J.P., Clark, C.C., and Hackman, B.O. 1996. Nosocomial and community-acquired *Staphylococcus aureus* bacteremias from 1980 to 1993: impact of intravascular devices and methicillin resistance. *Clin. Infect. Dis.* **23**:255–259.
2. Staphylococcus Laboratory, Statens Serum Institut. 2003. Annual report on *Staphylococcus aureus* bacteraemia cases 2001. Staphylococcus Laboratory, National Center for Antimicrobials and Infection Control, Statens Serum Institut. Copenhagen, Denmark. 9 pp.
3. Luzar, M.A., et al. 1990. *Staphylococcus aureus* nasal carriage and infection in patients on continuous ambulatory peritoneal dialysis. *N. Engl. J. Med.* **322**:505–509.
4. Yu, V.L., et al. 1986. *Staphylococcus aureus* nasal carriage and infection in patients on hemodialysis. Efficacy of antibiotic prophylaxis. *N. Engl. J. Med.* **315**:91–96.
5. Nguyen, M.H., et al. 1999. Nasal carriage of and infection with *Staphylococcus aureus* in HIV-infected patients. *Ann. Intern. Med.* **130**:221–225.
6. Kluytmans, J., van Belkum, A., and Verbrugh, H. 1997. Nasal carriage of *Staphylococcus aureus*: epidemiology, underlying mechanisms, and associated risks. *Clin. Microbiol. Rev.* **10**:505–520.
7. von Eiff, C., Becker, K., Machka, K., Stammer, H., and Peters, G. 2001. Nasal carriage as a source of *Staphylococcus aureus* bacteremia. Study group. *N. Engl. J. Med.* **344**:11–16.
8. Moss, B., Squire, J.R., and Topley, E. 1948. Nose and skin carriage of *Staphylococcus aureus* in patients receiving penicillin. *Lancet.* **1**:320–325.
9. Day, N.P., et al. 2001. A link between virulence and ecological abundance in natural populations of *Staphylococcus aureus*. *Science.* **292**:114–116.
10. Enright, M.C., et al. 2002. The evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA). *Proc. Natl. Acad. Sci. U. S. A.* **99**:7687–7692.
11. Grundmann, H., et al. 2002. Determining the genetic structure of the natural population of *Staphylococcus aureus*: a comparison of multilocus sequence typing with pulsed-field gel electrophoresis, randomly amplified polymorphic DNA analysis, and phage typing. *J. Clin. Microbiol.* **40**:4544–4546.
12. Day, N.P., et al. 2002. Retraction. *Science.* **295**:971.
13. Peacock, S.J., et al. 2002. Virulent combinations of adhesin and toxin genes in natural populations of *Staphylococcus aureus*. *Infect. Immun.* **70**:4987–4996.
14. Enright, M.C., Day, N.P., Davies, C.E., Peacock, S.J., and Spratt, B.G. 2000. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. *J. Clin. Microbiol.* **38**:1008–1015.
15. Feil, E.J., et al. 2003. How clonal is *Staphylococcus aureus*? *J. Bacteriol.* **185**:3307–3316.
16. Vos, P., et al. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* **23**:4407–4414.
17. van den Braak, N., et al. 2004. A new high-throughput AFLP approach for identification of new genetic polymorphism in the genome of the clonal microorganism *Mycobacterium tuberculosis*. *J. Microbiol. Methods.* **56**:49–62.
18. Hofman, A., Grobbee, D.E., de Jong, P.T., and van den Ouweland, F.A. 1991. Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. *Eur. J. Epidemiol.* **7**:403–422.
19. Bogaert, D., et al. 2004. Colonisation by *Streptococcus pneumoniae* and *Staphylococcus aureus* in healthy children. *Lancet.* **363**:1871–1872.
20. Murchan, S., et al. 2003. Harmonization of pulsed-field gel electrophoresis protocols for epidemiological typing of strains of methicillin-resistant *Staphylococcus aureus*: a single approach developed by consensus in 10 European laboratories and its application for tracing the spread of related strains. *J. Clin. Microbiol.* **41**:1574–1585.
21. Koning, S., et al. 2002. Fusidic acid cream in the treatment of impetigo in general practice: double blind randomised placebo controlled trial. *BMJ.* **324**:203–206.
22. Koning, S., et al. 2003. Severity of nonbullous *Staphylococcus aureus* impetigo in children is associated with strains harboring genetic markers for

exfoliative toxin B, Panton-Valentine leukocidin, and the multidrug resistance plasmid pSK41. *J. Clin. Microbiol.* **41**:3017–3021.

23. Wertheim, H.F., et al. 2004. Low prevalence of methicillin-resistant *Staphylococcus aureus* (MRSA) at hospital admission in the Netherlands: the value of search and destroy and restrictive antibiotic use. *J. Hosp. Infect.* **56**:321–325.
24. Robinson, D.A., and Enright, M.C. 2004. Evolution of *Staphylococcus aureus* by large chromosomal replacements. *J. Bacteriol.* **186**:1060–1064.
25. Schonheyder, H., Jensen, K.T., Pers, C., Korsager, B., and Rosdahl, V.T. 1992. Spread of *Staphylococcus aureus* strains of phage-type 95 in Denmark 1968-1989. *J. Hosp. Infect.* **20**:25–34.
26. Rosendal, K., Jessen, O., Faber, V., and Bentzon, M.W. 1983. Frequency, phage types and antibiotic resistance of *Staphylococcus aureus* isolated from blood cultures in Denmark 1975-1981. *Scand. J. Infect. Dis. Suppl.* **41**:19–29.
27. Robinson, D.A., and Enright, M.C. 2004. Multilocus sequence typing and the evolution of methicillin-resistant *Staphylococcus aureus*. *Clin. Microbiol. Infect.* **10**:92–97.
28. Feil, E.J., Li, B.C., Aanensen, D.M., Hanage, W.P., and Spratt, B.G. 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J. Bacteriol.* **186**:1518–1530.
29. Jarraud, S., et al. 2002. Relationships between *Staphylococcus aureus* genetic background, virulence factors, agr groups (alleles), and human disease. *Infect. Immun.* **70**:631–641.
30. Fitzgerald, J.R., Sturdevant, D.E., Mackie, S.M., Gill, S.R., and Musser, J.M. 2001. Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proc. Natl. Acad. Sci. U. S. A.* **98**:8821–8826.
31. Vandenesch, F., et al. 2003. Community-acquired methicillin-resistant *Staphylococcus aureus* carrying Panton-Valentine leukocidin genes: worldwide emergence. *Emerging Infect. Dis.* **9**:978–984.
32. Robinson, D.A., and Enright, M.C. 2003. Evolutionary models of the emergence of methicillin-resistant *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **47**:3926–3934.
33. Veldhuijzen, I., Bronzwaer, S.L., Degener, J., Kool, J.L., and EARSS participants. 2000. European Antimicrobial Resistance Surveillance System (EARSS): susceptibility testing of invasive *Staphylococcus aureus*. *Euro Surveill.* **5**:34–36.
34. Lina, G., et al. 1999. Involvement of Panton-Valentine leukocidin-producing *Staphylococcus aureus* in primary skin infections and pneumonia. *Clin. Infect. Dis.* **29**:1128–1132.
35. Gillet, Y., et al. 2002. Association between *Staphylococcus aureus* strains carrying gene for Panton-Valentine leukocidin and highly lethal necrotising pneumonia in young immunocompetent patients. *Lancet.* **359**:753–759.
36. Kuroda, M., et al. 2001. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. *Lancet.* **357**:1225–1240.
37. Murakami, K., et al. 1991. Identification of methicillin-resistant strains of staphylococci by polymerase chain reaction. *J. Clin. Microbiol.* **29**:2240–2244.
38. Reijans, M., et al. 2003. Quantitative comparison of cDNA-AFLP, microarrays, and GeneChip expression data in *Saccharomyces cerevisiae*. *Genomics.* **82**:606–618.
39. Cummings, L., et al. 2002. Genomic BLAST: custom-defined virtual databases for complete and unfinished genomes. *FEMS Microbiol. Lett.* **216**:133–138.
40. van Leeuwen, W.B., et al. 2003. Multilocus sequence typing of *Staphylococcus aureus* with DNA array technology. *J. Clin. Microbiol.* **41**:3323–3326.
41. Krzanowski, W.J. 2000. *Principles of multivariate analysis: a user's perspective.* Oxford University Press. New York, New York, USA. 586 pp.
42. Turner, J.D., et al. 2003. Th2 cytokines are associated with reduced worm burdens in a human intestinal helminth infection. *J. Infect. Dis.* **188**:1768–1775.
43. Quinn, G.P., and Keough, M.J. 2002. *Experimental design and data analysis for biologists.* Cambridge University Press. Cambridge, United Kingdom. 520 pp.
44. Valk, P.J., et al. 2004. Prognostically useful gene-expression profiles in acute myeloid leukemia. *N. Engl. J. Med.* **350**:1617–1628.

APPENDIX 2: Microarray Reference Table

Refer to website http://www-bioinf.erasmusmc.nl/thesis_peeters

APPENDIX 3: Discovery of Novel Splice Variations Improves Glial Tumor Classification

AFFYMETRIX
MICROARRAY
BULLETIN

EXPRESSION
VOLUME 2 ■ ISSUE 3
AMB REPRINT

www.microarraybulletin.com
Summer 2006

Discovery of Novel Splice Variations Improves Glial Tumor Classification

Erasmus Medical Center's Pim French and Justine Peeters talk with Noam Shomron at MIT about using exon arrays to study alternative splicing in glioma

By Megha Satyanarayana

Researchers at Erasmus Medical Center in the Netherlands have discovered expression profiles with distinct splice variants that more accurately classify two forms of glial cancers.

They made the splicing discovery using new microarrays that analyze over 1.4 million probe sets spanning all known and predicted exons; their findings may help clinicians more accurately diagnose the multiple classes and variable prognoses of brain cancer.

The team, led by Pim French, a postdoctoral fellow in the department of neurology at the Erasmus Medical Center and the Josephine Nefkens Institute, tested 28 glioblastomas, 20 oligodendrogliomas and 6 control brain samples with the new GeneChip® Human Exon 1.0 ST Array. They discovered that both types of tumors have a distinct pattern of alternative splicing in addition to a distinct gene expression profile.

French believes that the development of a splicing-based expression profile for the different types of glial tumors will provide a further level of certainty to histological analyses and may identify causative genetic changes.

"I would like to run a chip for every tumor that comes in, because I think you really need expression data to know what molecular subgroup a tumor is in to better aid the pathologist," said French. "Cancer is complex, and not caused by a single gene mutation. So, you really do have to have a global view of what is happening in the cells in order to get a good picture of disease

Pim French and Justine Peeters of Erasmus Medical Center

Pim French is a post-doctoral fellow at the Josephine Nefkens Insitute and department of neurology at the Erasmus Medical Center in Rotterdam, Netherlands. He received his Ph.D. at Erasmus Medical Center in cell biology, and completed a post-doctoral fellowship at the National Institute for Medical Research in London. His main focus is using microarray technology to study neurological cancers and disorders, to aid pathology and to find causative genetic mutations.

and ultimately this involves studying splicing isoforms."

Justine Peeters led the bioinformatics effort to analyze the glioma exon array data. She is finishing her Ph.D. studies in the analysis of microarray data, with further applications to cancer, in the laboratory of Prof. Peter van der Spek at Erasmus. Her thesis project will provide an educational resource for performing expression, SNP and exon microarray experiments, from planning the experiment to analyzing the resulting data. French and Peeters' work has recently been submitted for publication at *PNAS*.

"These arrays give us just a whole other level of analysis, really," said Peeters. "Instead of just having your transcript measured with expression arrays, we can look at the differentially represented exons and then translate this information into changes in the transcripts. So, even though you have the same overall information, exon arrays are more of a true measure of gene expression."

French and Peeters recently spoke to Noam Shomron, a post-doctoral fellow in the laboratory of Chris Burge at MIT about new applications for studying alternative splicing at the whole genome level. One of Noam's research fields focuses on using novel computational tools for studying alternative splicing events in general and for identifying unannotated alternatively spliced exons.

The three discussed:

- The advantages of exon arrays over other expression arrays in classifying gliomas
- Testing, troubleshooting and validating exon array data
- The potential application of exon arrays to clinical diagnosis and prognosis

Exon arrays vs. other expression arrays in glioma diagnosis

Shomron: Gliomas are noted for their aberrations in alternative splicing. How were you studying glioma gene expression before and could you describe your current approach?

French: Originally, we had been studying global gene expression in gliomas using GeneChip® Human Genome U133 and then using that information to supplement our pathological diagnoses. We could identify molecular subgroups of gliomas and classify them based on their expression profile.

Our current approach is to use exon arrays to look at the contribution of splicing to tumor biology. We strongly believe that exon arrays will help us find causative genetic changes in cancer. For example, exon arrays readily detect expression of a pathological splice variant of EGFR. Exon arrays should also lead to the identification of fusion genes because the exon probes would cover the joined breakpoints. Such breakpoints are hard to find using the old U133 arrays.

Shomron: So, have you completely shifted to exon arrays or are you still using both?

French: We are still using both because we don't want to change protocols for ongoing experiments. Furthermore, most of the gene expression information we see is in transcripts rather than in differentially expressed splice variants. So, for identifying molecular subgroups of gliomas, we use expression arrays, but I think we will shift towards using exon arrays, because in the end, you just get more data from them.

Shomron: But can you successfully retrieve expression data from the exon array as well?

French: Yes, you can and in fact, quite easily and very well.

Peeters: The probe sets used in the U133 Plus 2.0 arrays are 3-prime biased, so, they're not really a true presentation of what is really happening with transcripts. If you have a splice variant influencing the biology of what you are looking at, this will most likely be missed with a selection of probes interrogating only the 3' end of the gene. This is where exon arrays are more powerful in reflecting the true biology, as every exon whether it is located in the 3 prime or 5 prime end of the gene is represented in the array. You can utilize these arrays in looking at an exon by exon approach or translate this information into transcript intensities.

Shomron: So, when do you think you will completely shift to using exon arrays?

French: It mainly depends on funding for running such a large number of arrays. Also, it is slightly more difficult to handle exon arrays than the U133 Plus 2.0 arrays. So, there's a little more training involved.

Peeters: And previously you were dealing with 54,000 probe sets. Now we have 1.4 million probe sets. It creates a problem for computational analysis—available computer memory and available software.

Shomron: Can you elaborate on any overlapping clusters and correlations you found from using these two array platforms?

French: The two platforms correlate extremely well. The subgroups that we get with U133 Plus 2.0 arrays are identical to the subgroups that we get with the exon arrays. That was an important step in saying how well they work. Here's an example. We have two oligodendroglial tumor samples that cluster with glioblastoma samples on the exon arrays — they are outliers. On the U133 Plus 2.0 arrays, we ran the same two samples and we found that one of the two samples also clustered with glioblastomas as opposed to oligodendromas. The subgroups clustered very well and some of the same samples are outliers in both studies.

Testing and validating glioma expression studies

Shomron: In the exon array, how many positive controls of known splicing isoforms did you look at in order to gain some confidence that your arrays worked well before you proceeded for further analysis?

Peeters: We looked at EGFR variant III, which has been seen in glioblastomas containing EGFR amplification. *EGFR*, of course, is known to be differentially spliced in glioma. We also looked at other causative splice isoforms in other tissues such as breast cancer. In developing an algorithm to detect splice variants, we used this as a positive control as a guide to look at step-by-step filtering procedures. As part of this algorithm, we had to filter out exons which have a non-linear relationship with transcript as well as non-informative exons which are also represented on the chip. Signal from such exons can skew your results and lead to many false positives.

Shomron: In your studies, you note that there was a correlation in the subgroups defined through expression analysis and through histology. Was that surprising?

Peeters: Well no. Gene expression basically directs morphology and thus, histology, but a pathologist cannot see what gene expression tells us. Therefore, the differences are quite informative for follow-up and treatment. It's possible that the underlying transcript I am seeing through expression studies can also predict drug response. Looking at the transcript, you can actually predict, even though they may look the same histologically, how the patient can respond to treatment or what type of treatment you can actually give.

> "Looking at the transcript, you can actually predict, even though they may look the same histologically, how the patient can respond to treatment or what type of treatment you can actually give."

Shomron: Do you always perform histology in addition to arrays?

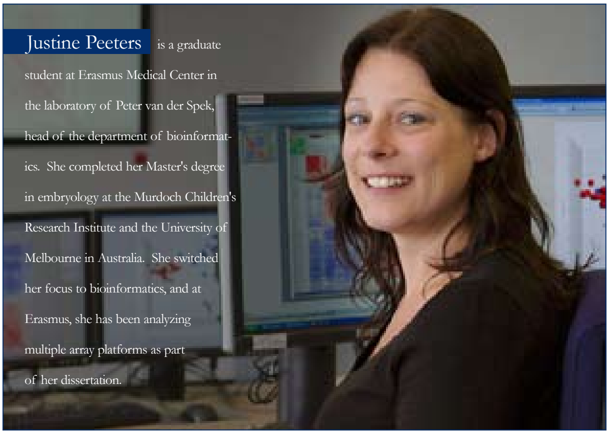
French: Not all samples are of sufficient quality to run on our arrays. Of course, for samples that are good quality, we do the comparison with histology.

Shomron: You've identified hundreds of novel exons in gliomas that are not supported by current databases. Have you confirmed them through RT-PCR?

French: Yes. This is our protocol: We design two primer pairs, both of which contain one primer in the putative novel exon and one in a known exon. In designing the primers, we make sure that there is a large intronic region in between, to avoid amplifying genomic DNA. We run the assay and get bands of about the expected size. We then sequence the bands and indeed, you get the product that you would expect. With the sequence, we refer to genomic databases and we often find consensus splice acceptor and donor sites.

Shomron: And what is the success rate? What is the false positive and negative of those RT-PCRs?

French: We have identified around 700, of which about 80 percent truly represent novel exons. we can confirm something like 60 to 70 percent of candidates. The 30-40 percent we cannot confirm may be due to incorrect primer design because you might have a completely differently spliced transcript. This



Justine Peeters is a graduate student at Erasmus Medical Center in the laboratory of Peter van der Spek, head of the department of bioinformatics. She completed her Master's degree in embryology at the Murdoch Children's Research Institute and the University of Melbourne in Australia. She switched her focus to bioinformatics, and at Erasmus, she has been analyzing multiple array platforms as part of her dissertation.

Noam Shomron is a post-doctoral fellow in the laboratory of Chris Burge at MIT. He completed his Ph.D. at Tel Aviv University in Israel and his Master's Degree at the Hebrew University in Jerusalem. While a graduate student, he studied splicing variations caused by specific proteins in human cells, and as a post-doctoral fellow, among other projects, is applying his knowledge of alternative splicing to study splicing in whole genomes using microarrays.

means that the number of novel exons may even be higher than we can confirm by RT-PCR.

Shomron: That's terrific. Now, let's talk technically about non-linear exons — exons that are expressed at levels outside the linear detection range of the arrays, or exons that bear no relation to its putative transcript. You mentioned that after comparing the exon array results and RT-PCR you recalculated PAC, the probability of the splicing events, based on metaprobe sets lacking "non-linear" exons. How much data was discarded as "non-linear"?

French: If you look at the 1.4 million probe sets that are on the array, only a few percent are nonlinear or nonspecific exons. But, if you try to identify regulated splice variants, then you narrow down your search to a few thousand candidate exons. So the smaller set of candidates are now biased with these nonlinear exons, and you end up with quite a high percentage — up to 80 to 90 percent in first pass, without any filtering steps.

Shomron: Then does your filtering process actually retain a large number of nonlinear exons?

French: Yes, we find it difficult to completely filter them out. But in identifying regular splice variants you ask that every probe set within the gene of interest behaves in a completely linear fashion. So if the gene is upregulated two-fold in one sample or subgroup, all the individual exons of that gene should be upregulated exactly two-fold on the array. Any probe set that does not show this two-fold upregulation will be identified as a false positive candidate.

Peeters: And also, in identifying splice variants, you have to translate all of your exon probe sets into your transcript signal intensity in order to calculate differences in expression. This is a whole additional level of analysis as compared to the U133 Plus 2.0 expression arrays. What we used was an adapted correlation based PAC algorithm in our calculations of the splice variants. Including these nonlinear exons in your initial calculation of the transcripts to be utilized in the PAC algorithm disrupts the true correlation between your exon and your transcript, which complicates things.

Shomron: And, you verify them by quantitative or nonquantitative RT-PCR?

French: We use nonquantitative RT-PCR, because differences are pretty obvious. For ones that we could confirm, we have a verification success rate of about 40 percent. RT-PCR using primers that span the spliced exon gives you a semi-quantitative result anyway, because you see the ratio of inclusion and exclusion of the spliced exon.

Shomron: I think the next stage after identifying which exons are represented is finding the difference in magnitude and the varied interplay or network between the different isoforms. That is probably something for the future after users feel confident with their exon arrays.

Future studies and using exon arrays for diagnostics

Shomron: Will your results from these arrays take you into follow up studies of individual genes or back to running more arrays and clustering?

Peeters: I think we would definitely use both approaches. Following up on these individual genes to look at their role in the development of the subtypes of brain tumors is definitely going to be interesting, but we also would be interested in running some more arrays and also reassessing the arrays that we have already run with alternative analysis techniques. We believe that we can identify causative changes using these exon arrays and hopefully identify more molecular subtypes or subgroups of brain tumors. We would like to be able to include some more histological subgroups and see whether again we can find differentially regulated splicing and even causative changes that cause the errant splicing event. So, yes there is a lot of information in these exon arrays that we can utilize.

Shomron: I agree that there is an advantage of looking at the global gene expression/isoform picture, rather than just one or two genes at a time. Do you think that splicing isoform clusters will eventually replace expression clusters in diagnosis and prognostics?

Peeters: Yes, I do think exon arrays will be the future of diagnostics and prognosis, because they do have a lot more information than the older expression arrays. As I have mentioned before, we can look specifically at different isoforms of transcripts that may be involved in patient response to drug treatment and the specific diagnosis of disease.

At the moment, it is fresh, new analysis for us. So, the most important thing for us is to see if we can use this platform to identify differential expressed splice variants and be able to confirm them in the lab.

Shomron: Last question — if you could set up the ultimate diagnostic laboratory what would it look like?

French: You would use exon and SNP arrays to identify which molecular subgroup the tumor belongs to. These subgroups would aid in guiding therapy and will also give prognostic information for the patient. I also strongly believe that future therapies will focus on the molecular aberrations of tumors. Therefore, I would also like to perform an array experiment that would sequence all of the known tumor suppressors and oncogenes. This way, you know that in patient A you have genes X, Y, and Z mutated. Such molecular knowledge will be of high importance to guide future therapies. I think that would be the ultimate diagnostic lab.

Shomron: Yes, I definitely agree with you. Use microarrays to analyze everything, collect the data, then look at part of it now, and then maybe reanalyze a few years later, when there are a few more revelations and a few more diagnostic tests and genes or mechanisms identified. An example would be regulation by microRNAs, a concept which has gained ground only in the past few years.

French: You never know what you are going to get from such large data sets, but there is a lot of information encoded in the expression data and most of it we do not really know at this point. It's amazing. We are only scratching the surface of what you can do with these arrays.

AFFYMETRIX
MICROARRAY
BULLETIN

Editorial Staff

Wes Conard, *Editor-in-Chief*
wes_conard@affymetrix.com
Tommy Broudy, *Managing Editor*
thomas_broudy@affymetrix.com
Rachel Shreter, *Editor*
rachel_shreter@affymetrix.com
Kamalia Dam, *Associate Editor*
Stacey Ryder, *Associate Editor*
Daniel Noble, *Copy Editor*
Marva Maida, *Contributing Designer*

FOR MORE INFORMATION

Contacts

- Pim French, Ph.D.
Josephine Nefkens Institute
Department of Neurology
Erasmus MC
PO Box 1838
3000 DR Rotterdam
Netherlands
p.french@erasmusmc.nl

- Justine Peeters
Department of Bioinformatics
ErasmusMC
Postbus 1738,
3000DR Rotterdam
Netherlands
j.peeters@erasmusmc.nl

- Noam Shomron, Ph.D.
Department of Biology
Massachusetts Institute of Technology
77 Massachusetts Ave 68-217,
Cambridge, MA 02139
United States of America
nshomron@mit.edu
http://web.mit.edu/nshomron/www

Companies

- Affymetrix, Inc. - http://affymetrix.com

Organizations

- Erasmus Medical Center - http://www.erasmusmc.nl/
- Josephine Nefkens Institute - http://www.eur.nl/fgg/pathol/
- Massachusetts Institute of Technology - http://www.mit.edu

People

- Chris Burge, Ph.D., MIT - http://genes.mit.edu/chris/
- Peter van der Spek, Ph.D., Erasmus Medical Center - http://www.erasmusmc.nl/bioinformatics/people/staff.shtml

Further Reading

- French PJ, Swagemakers SM, Nagel JH, Kouwenhoven MC, Brouwer E, van der Spek P, Luider TM, Kros JM, van den Bent MJ, Sillevis Smitt PA. Gene expression profiles Associated with Treatment Response in Oligodendrogliomas. *Cancer Res.* 2005 Dec 15;65(24):11335-44.
- Peeters JK, Van der Spek PJ. Growing applications and advancements in microarray technology and analysis tools. *Cell Biochem Biophys.* 2005;43(1):149-66.

APPENDIX 4: Global View of Gene Expression Analysis

AUSTRALIAN LifeScientist

Global view of gene expression analysis

It is now believed that more than 60 per cent of genes are alternatively spliced, a final nail in the coffin, if it needed one, of the one-gene-one-protein hypothesis.

Since the inception of microarray-based expression profiling, microarrays have used a 3' (three prime) biased labeling system. To date, this approach has made significant contributions towards the better understanding of the mechanisms governing many diseases.

However, it is now understood that more than 60 per cent of genes are alternatively spliced, the process in which the exons in pre-mRNAs are spliced out, changing the transcript's sequence and contributing to the hundreds of thousands of transcript isoforms that help to define the biology of the system being studied.

Of specific importance to human disease is the fact that up to 50 per cent of disease causing mutations may result in transcript splicing alternations and 20 per cent of cancer-causing mutations can result in exon-skipping events.

Unfortunately, the traditional 3' approach to expression profiling is limited by assumptions that the 3' end of each gene is clearly defined and that each transcript has an intact Poly-A tail. As such, 3' expression profiling cannot detect alternative spliced transcripts with the same 3' end, transcripts from the same gene locus with a different 3' tail, non-polyadenylated transcripts, genomic deletions and other genomic alterations (see figure 1).

Microarray pioneer Affymetrix has now developed a new Whole Transcript Assay (WT Assay) and exon-based microarrays that are designed to reveal a level of transcriptional complexity not previously detectable with conventional 3' microarray expression approaches.

The main purpose of the WT Assay is to provide a more complete and accurate picture of overall gene expression. The assay uses a random priming method for generating labeled sense targets throughout the entire length of RNA transcripts, thereby negating the limitations of 3' expression profiling.

“There are two types of arrays that are compatible with the new assay - the GeneChip Exon 1.0 ST and Gene 1.0 ST arrays, which are available for human, mouse and rat,” says Dr Robert Henke, chief scientific officer of Millennium Science, the Australian distributor for Affymetrix.

“The Exon 1.0 ST Arrays provide three levels of expression information from a single array - differential gene-level expression, differential exon-level expression, and discovery of alternative splice events.”

Henke says some of the alternative splicing events detectable by exon arrays include exon skipping, differential isoform expression, alternative 5’ transcriptional start sites, truncated transcripts, genomic deletions and translocations, and detection of transcripts with undefined or non-poly adenylated 3’ends.

“Exon arrays target over one million exons with about four probes per exon, providing an average of 40 probes spread across the entire length of each targeted gene,” Henke says. “Through the use of this exon-centric probe content, researchers have the ability to analyse both alternative splicing and differential expression of individual exons within each gene.

“For gene-level expression analysis, the full set of probes spanning the entire transcript are used to generate a single data value that represents the expression level of all transcripts generated by the gene.”

Splice variations and glioblastomas

Techniques such as whole transcript assays and exon arrays are unveiling alterations in exon usage that may play a critical role in diseases such as cancer.

For example, researchers at Millennium Pharmaceuticals in the US have used Affymetrix’s Exon 1.0 ST Array to discover specific exon cassettes within the CD44 gene that are expressed in primary colon cancer cell lines, but are absent from metastatic colon cancer and Hela cell lines. These results suggest that CD44 splice variants might serve as diagnostic or prognostic markers for colon cancer.

Exon arrays are also being used to distinguish glioblastomas from oligodendrogliomas in order to help clinicians better diagnose brain cancer. Dr Pim French and Dr Justine Peeters from the Erasmus Medical Centre in the Netherlands have used exon arrays to compare different tumour samples and identify novel exon-skipping events and associated genes.

They have performed an analysis of differentially regulated splice variants and novel exons in glial brain tumours.

“Splice variants have a key role in biology,” Peeters says. “Both tissue and developmental stage-specific alternative splicing contributes to significant protein diversity.”

This team’s primary aim it its research project is to explore how disease-related deregulation of splicing might be critical in pathogenesis and contribute to disease diversity and complexity.

“Compared to other primary tumour types, there is not so much known about the underlying molecular causes that contribute to the onset, progression and treatment of gliomas,” Peeters says. “By identifying splice variants that are differentially regulated between histological subgroups we can uncover more of the biology involved in these aggressive tumours.”

Greater coverage

The Erasmus Medical Centre was a major test site for Affymetrix Exon 1.0 ST Array and has previously published a study in Cancer Research on glial tumours using the Affymetrix U133 Plus 2.0 expression arrays (a 3’ focused microarray).

According to Peeters, the Exon 1.0 ST array provides greater genome coverage, as well as the possibility of detecting regulatory mechanisms such as exon skipping, intron retention and alternative promoter usage. The Exon array also demonstrated the ability to identify and characterize glial tumour subgroups based on different analyses methods.

“We were able to identify and molecularly separate these subgroups based on both the expression of the exons, as well as the associated transcript expression,” she says.

“We were able to detect differentially regulated splice variants, novel exons and possible translocated transcripts and we have also been able to detect exon skipping mutations.”

For Pim French, the key goal wasn’t to find more markers for specific subtypes of glial brain tumours but to utlise the potential of such arrays to allow researchers to find causal genetic changes, like the pathological splice variant of EGFR.

According to French, a large proportion of glial brain tumours have a genetic deletion within the epidermal growth factor receptor (EGFR) locus.

“This deletion results in the expression of a pathological splice variant that is constitutively active,” he says. “This splice variant has been demonstrated to play a role in tumour formation and is associated with response to EGFR inhibitors.”

Using the Exon array, “I was most excited to identify pathological splice variants like the one in EGFR. In fact, we found a few in that gene we were not aware of.

“Such pathological splice variants will not be detected with other expression profiling platforms.”

“The information contained on the exon array is greater than the older 3’ arrays and such genome coverage give more possibility of answering more diverse biological questions,” Peeters says. “Having run disease-related samples on the arrays gives endless possibilities of remining the data.”

Author: Kate McDonald.
Interview published in Australian Life Scientist: 19th February 2008.