Microarray-based Gene Expression Analysis in Cancer Research

Cecilia Laurell

Doctoral Thesis
Royal Institute of Technology
School of Biotechnology
Stockholm 2006
Abstract

Biotechnological inventions during the 20th century have resulted in a wide range of approaches for explorations in the functional genomics field. Microarray technology is one of the recent advances which have provided us with snapshots of which genes are expressed in cells of various tissues and diseases. Methods to obtain reliable microarray data are continuously being developed and improved to meet the demands of biological researchers.

In this thesis microarrays have been used to investigate gene expression patterns in cancer research. Four studies in three different areas were carried out covering adrenocortical tumors, p53 target genes and a comparison of RNA amplification methods.

Adrenocortical tumours are among the most common tumours with an incidence of 7-9%. Malignancy of these tumors is rare. Distinction between malignant and benign tumours is often difficult to establish which makes an improvement of diagnostic approaches important. To elucidate biological processes in adrenocortical tumour development and to examine if there is a molecular signature associated with malignancy, microarray analysis was performed on 29 adrenocortical tumors and four normal specimens. It was possible to classify malignant and benign samples based on the entire expression profile. A number of potential biomarkers was identified which will be further evaluated.

P53 is a gene which is mutated in 50% of all cancers. Functional p53 is a transcription factor which is activated upon cellular stress and DNA damage. Target genes are mainly involved in cell cycle arrest and apoptosis. In solid tumors cells are stressed by hypoxia. To examine which target genes p53 activate under hypoxic conditions a microarray study of the cell lines HCT116p53+/+ and HCT116p53-/- was performed. A set of novel potential p53 target genes was identified while many known target genes were found to be not transcriptionally activated during hypoxia. Follow up which was focused on how p53 affected hypoxia induced apoptosis showed that the death receptor Fas was critical.

When small amounts of tissue are available, amplification of the transcript population is necessary for microarray analysis. A new strategy for amplification based on PCR was evaluated and compared to a commercial in vitro transcription protocol. Both protocols produced reliable results. Advantages with the PCR based method are a lower cost and a high flexibility due to compatibility with both sense and antisense strand microarrays.

Keywords: adrenocortical tumour, apoptosis, cancer, classification, gene expression, microarray, p53, RNA amplification
In Nature’s book of secrecy
A little I can read

William Shakespeare
List of publications


* These authors contributed equally to the work and should therefore be considered as joint first authors.
3.1.1. 3' cDNA Tag Amplification 46
3.1.2 Paper I 48
3.2 Detection of p53 Dependent Genes Induced by Hypoxia 49
3.2.1 P53 - Cancer Gene 49
3.2.2 Regulation of p53 by Hypoxia 50
3.2.3 Paper II 51
3.3 Characterization of a Malignant Expression Signature of Adrenocortical Tumours 53
3.3.2 Adrenal Gland 53
3.3.2 Adrenocortical Tumours 53
3.3.3 Paper III & IV 55

4 Concluding Remarks 57
Abbreviations 59
Acknowledgements 59
References 63
Introduction

Biotechnological inventions during the last half of the previous century, including DNA sequencing for reading of the genetic code, adoption of restriction and ligation enzymes to cut and paste DNA into desired sequences, and amplification of molecules with polymerase chain reaction (PCR), have lead to a revolution in our knowledge of biological sciences (Mullis and Faloona, 1987; Sanger, 1977). A milestone in this respect of scientific progress was the longed-for release of the human genetic code in 2001 which provided a valuable resource for exploration of genetic diseases (Lander, 2001; Venter, 2001).

The central dogma of molecular biology is the flow of information from the heritage material DNA, via messenger carrier molecules RNA, to the final protein products (figure 1) (Crick, 1958). These three families of molecules are the genome, the transcriptome and the proteome. The information overlap between these was from the beginning assumed to be total, but now we know that protein encoding genes only constitute a fraction of the human genome and that most RNA species do not encode proteins. The fact that RNA both can hold information and catalyze chemical reactions has lead to the idea that a RNA world could exist.

Figure 1. A schematic view of information flow in a cell; from DNA, via messenger carrier RNA to proteins, the final products.
Examination of the transcriptome has long time been focused on the protein encoding sequences, thereby leaving a substantial portion in the shadow. The recent years CAGE, tiling microarrays and sequencing technology have shed light on a growing population of non-coding RNAs (Gustincich, 2006). As it has been realized that the non-coding regions are not merely junk sequences but in many cases have important regulatory roles this is now a rapidly growing research field, to find out more about which these new members of the transcriptome are and what they do. A brief description of members of the mammalian transcriptome is given in box 1.

Gene expression analysis often reflects the steady-state levels of mRNA which is the result of a delicate balance between production and decay. Half-lives of unstable transcripts are \( \approx 15 \text{ min} \) while \( > 24 \text{ hours} \) for the most stable ones. Generally mRNAs encoding proteins with important regulatory roles are continuously being transcribed and degraded allowing

### Box 1. The mammalian transcriptome

**mRNA** The average mRNA transcript is 1.5 kb consisting of a protein encoding part in the middle flanked by untranslated regions for regulation of decay and translation. The 3' end has a 250 b polyA tail and the 5' end a guanosine cap structure which both are attributes required for translation.

**miRNA** Micro RNAs are 21-23 b fully processed and binds to mRNAs promoting decay or inhibition of translation by recruiting the RISC complex. They are generated from double-stranded foldback precursors which are cleaved to their mature form by the Dicer (Bartel, 2004a). This is the endogenous process which also is exploited by exogenous interference RNA (Fire, 1998).

**rRNA** Ribosomal RNA constitutes 90% of the RNA and have six different members. These are encoded in multiple copies throughout the genome. Together with RNA binding proteins these form a two unit 3D structure which constitutes a framework for translation of proteins.

**piRNA** Piwi RNAs are 25-31 b and are expressed in developing male germ cells forming a complex with the Piwi and RacQ1 proteins exhibiting DNA helicase activity and is believed to have a role in transcriptional silencing (Girard, 2006; Lau, 2006).

**snRNA** Small nuclear RNAs regulates transcription and play an active role in splicing by providing recognition site for the intron-exon boundaries (O’Gorman, 2006; Valadkhan, 2005).

**snoRNA** Small nucleolar RNAs are 60-300 b and involved in RNA editing promoting methylation or pseudouridylation. They are derived from the introns of pre-mRNA transcripts. About 300 different snoRNAs have been discovered (Mattick and Makunin, 2005).

**tRNA** Transfer RNAs are 73-93 b which form three stemloops one of which has the anticodon for an amino acid. The amino acid is coupled to the 3' end of the tRNA. tRNA supplies the ribosomes with amino acids during translation.
for a transient increase when the decay is switched off. Regulation of the stability is carried out by RNA binding proteins (RBPs) and miRNA binding to AU-rich elements (AREs) in the 3’UTR but also to secondary structures throughout the entire transcript. There are two main pathways for degradation of fully processed mRNA; 3’-5’ decay mediated by cytosolic exosomes and 5’-3’ decay by the Xrn1 enzyme. The first and rate limiting step in both pathways is deadenylation of the polyA tail. This is followed by decapping of the 5’ end for 5’-3’ decay. The exosome pathway has been shown to be the dominant one in mammals (Beelman and Parker, 1995; Meyer, 2004).

Initiation of transcription is regulated by remodelling of the chromatine structure followed by the binding of transcription factors to regions commonly 5’ of the transcription start site recruiting the RNA polymerase complex. To some extent genes are organized into clusters which can share epigenetic control but far less pronounced than the operon structure found in prokaryotes which allows for a complete sharing of regulation (Caron, 2001). Another feature which adds to the complexity of eukaryotic gene expression is alternative splicing (Matlin, 2005). In average each mRNA has three splice variants but there is a large variation and new transcripts are continuously being discovered (Modrek and Lee, 2002). The oncogene MDM2, as an example, has more than 40 splice variants with different functions and regulatory units (Bartel, 2004b).

Gene expression analysis aims at reflecting the mRNA levels but to some extent also mirror the proteome. In this case a quite distorted picture is provided according to the large-scale comparisons between protein and mRNA levels (Griffin, 2002; Gygi, 1999). Some genes are regulated exclusively on protein level while others differ largely in magnitude of the expression. The progress of technologies within the proteomics field can however soon offer us a more detailed map.

The theme of this thesis is applications of microarray technology to global gene expression analysis in cancer research. In the following sections microarray technology is described, followed by its contributions to cancer research in general, and finally a summary of the four studies which are the basis of this thesis.
1 Microarray Technology

1.1 Microarray Platforms

Microarray technology enables analysis of relative abundances of typically 30,000-50,000 different nucleic acid molecules simultaneously. The basic concept of the DNA microarray is to take advantage of that DNA or RNA molecules with complimentary sequences bind to each other, hybridize. The DNA molecules to be measured are immobilized onto a solid support in an organized pattern. Analysis can be performed by hybridising DNA or RNA in the sample to complimentary DNA on the array and then measure the amount. The immobilized DNA is called probe and the molecules binding to it, target. To enable measurements, the target is labelled with molecules making it possible to detect with fluorescence or radioactive radiation, how much has bound. Depending on probe lengths, concentrations and melting temperatures, different specificities and sensitivities can be obtained.

Originally, quantification of a specific DNA molecule in a complex mix by hybridization, was introduced already in 1975 by Southern, but it was not until the 90's high-throughput platforms were launched (Southern, 1975). During the past ten years microarrays have had a revolutionary impact on biomedical research. The parallelization, miniaturization, and automation opportunities offered by chip technology have resulted in that methods examining one gene at a time in many cases have been exchanged by microarrays as the standard tool, generating vast amounts of information with few experiments. The DNA microarray has so far mainly been used for large-scale gene expression analysis, but plays a
central role also in other fields, such as genotyping, epigenetics and promoter analysis (Gresham, 2006; Kaller, 2005; Matsuzaki, 2004; Wilson, 2006). To obtain reliable data, techniques for microarray production, sample preparation and data handling are continuously being developed and improved.

1.1.1 Spotted Microarrays

Spotted microarrays refer to chips where pre-synthesized probes are applied to the array surface by robotic printing. The spot size is typically 100-200 µm. There are mainly three types of probes for gene expression analysis; full length cDNAs, shorter PCR amplified gene specific tags, and single stranded oligonucleotides.

cDNA microarrays: The efforts made to discover new genes by sequencing of cDNA clones in the 1990’s, generated a large amount of cDNA libraries. Sequencing data of these libraries for different organisms and tissues provided a first glance into specific transcriptomes. Expressed sequence tags (ESTs) contribute to 69% of human gene sequences in the GenBank database (Benson, 2006). By amplifying normalized cDNA libraries with PCR, probes suitable for global microarray gene expression analysis were obtained. In 1995 construction of the first microarray was published consisting of 1000 Arabidopsis thaliana genes (Schena, 1995). The first whole genome array was made for Sacharomyces cerviseae and was used in the landmark cell cycle study of Spellman and colleagues (DeRisi, 1997; Schena, 1995; Spellman, 1998).

GST microarrays: A drawback with cDNA arrays is cross-hybridization between gene families and genes with common functional domains. To lower the amount of homologous sequences in the probe set one could selectively avoid these regions by using gene specific primers for gene amplification of gene specific tags (Wirta, 2005). For prokaryotes which lack polyA tail this is an alternative to using shotgun libraries. The procedure requires sequence information of every gene and is costly because of the amount of primers needed to be synthesized. Successful efforts to reduce the primer cost have been made by making efficient primer design, using unique primer pairs for each
gene but allowing each primer to be involved in several pairs. Homology between species enables sharing also between different organisms (Andersson, 2005; Fernandes and Skiena, 2002).

**Oligo microarrays:** A less labour intensive strategy to obtain specific probes is to synthesize an oligonucleotide for each probe (Call, 2001; Kane, 2000; Zhao, 2001). Commercial 40-70mer oligonucleotide sets are available from several companies. The shorter hybridising sequence introduces however an increased susceptibility to polymorphisms.

Besides the commercially available oligo probe sets, there have been numerous developments of methods for optimized probe design which allows for probes adapted to individual requirements (Emrich, 2003; Mrowka, 2002; Nielsen, 2003; Wang, 2002; Wernersson and Nielsen, 2005). The choice of oligo length is not evident and depends on the application. Shorter sequences are cheaper to synthesize and are sometimes necessary to achieve specificity for certain transcripts, while longer ones can ameliorate signal intensities as well as specificity remarkably (Ramdas, 2004). Important features in construction of spotted microarrays besides probe design are spot concentration and morphology, as well as the background intensity. These are affected by the printing buffer used, humidity and temperature, robotic delivery and surface chemistry (Wrobel, 2003).
1.1.2 *In Situ* Synthesized Arrays

**Affymetrix:** The most widespread in situ synthesis technique, commercialized by Affymetrix in the mid 90’s, is based on photolithography technology from the semiconductor industry to direct DNA synthesis on glass slides producing high density oligonucleotide microarrays (GeneChip) (Fodor, 1991; Fodor, 1993; Lockhart, 1996). The procedure is based on synthetic linkers with photoprotected groups attached to the array in a narrow pattern. The probe sequences are elongated stepwise by cyclic addition of dNTPs. To obtain different sequences on distinct positions, probes are alternating being protected and deprotected by light using different photolithographic masks for every step. Probes are designed to give optimal hybridization conditions considering parameters such as melting temperature and sequence. Unspecific and repetitive regions are avoided. Since the technique only allows for synthesis of maximum 25 nucleotides, development of methods to compensate for this shortcoming has been forced. To increase signal to noise ratio as many as 11-20 different probes are synthesized for each mRNA. The redundancy increases the dynamic range, lower cross-hybridization and reduces the number of false positives. The probes are designed to match sequences throughout the whole transcript but with emphasis on the 3’ end. An additional method to filter out false positives is offered by also synthesising an almost identical set of replicate probes. All probes which perfectly matched sequences (PM) have a partner probe which only has one mismatch (MM). These MM probes allow discrimination between real signals and signals due to cross reactivity and other artefacts.

**NimbleGen:** In 2002 NimbleGen Systems launched a new method for high density *in situ* oligosynthesis; the Maskless Array Synthesizer (MAS) (Nuwaysir, 2002). The principle is to create virtual photolithographic masks by using a digital light processor, directing the light to specific alternating probe positions with around 800,000 individually addressable mirrors. Advantages with this new technique were the reduced cost, increased flexibility, longer maximum probe length (90 nt), and the possibility to synthesize the oligonucleotide in both 3’ to 5’ and 5’ to 3’ indirection. The last point enables elongation with DNA polymerases after target hybridization to the probe.
1.1.3 Bead-based Systems

Although the methods described above have shown good results with robust and standardized laboratory management, efforts are being made to develop alternative strategies with increased sensitivity and specificity.

**Illumina:** A bead based array system introduced by Michael and colleagues in 1998 have been used as basis for further developments by Illumina which now has a platform which allows for analysis of 46,000 transcripts with more than 6.5 million features using a randomly ordered array assembly followed by a decoding procedure (Gunderson, 2004; Kuhn, 2004; Michael, 1998). Each feature consists of a bead, 3 µm in diameter, coated with several hundreds of thousands copies of an oligonucleotide containing a 50mer probe sequence and a shorter address sequence used for feature identification.

**454:** The 454 Life Sciences platform was originally developed to speed up whole genome sequencing and is based on sequencing-by-synthesis with pyrosequencing technology (Margulies, 2005; Ronaghi, 1996). Sequencing based gene expression analysis in general has the advantage that novel transcripts can be discovered in addition to providing a high specificity. Briefly, the methodology includes linker ligation and immobilization of fragments onto microbeads (ideally one fragment per bead), followed by emulsion-based PCR and subsequent transfer to a fiberoptic picotiter plate where the pyrosequencing reaction takes place. ATP sulfurylase and luciferase are immobilized onto beads and added once, while the other reagents are efficiently cycled by a liquid handling system, allowing for a maximal read length of 100 b. In 4 hours, 300,000 DNA templates can be sequenced with an accuracy of 99.96%. Several promising methods for transcriptome analysis have emerged based on 454 technology; DeepSAGE which is a hyper efficient method for generation of ditag SAGE libraries, 5’ RATE which sequences 5’ ends thus allowing for mapping of transcription start sites, ultra-high-throughput EST sequencing and also examinations of promotor binding sites and functions of the small non-coding RNAs (Bainbridge, 2006; Gowda, 2006; Ng, 2006; Nielsen, 2006)
1.1.4 Conversions Between Microarray Platforms

The last few years numerous platform comparisons have been published (Barnes, 2005; Kuo, 2002; Mah, 2004; Wang, 2005; Zhu, 2005). Surprisingly these showed that the various platforms do not give concordant results in many cases.

One explanation is the ambiguity of comparing sequences from different parts of the transcript. To be able to combine data from different platforms, it is of high importance that the probe annotations are in agreement. In the beginning GeneChip users had the annotation provided by Affymetrix to rely on since probe sequences were not made public. In 2003 the sequences were released together with an annotation software allowing customers to do more thorough inspections of their expression data (Liu, 2003). Recent studies have indicated that as much as 20% of the Affymetrix probe sequences lack a match in RefSeq db and almost 40% were wrongly annotated (Harbig, 2005; Mecham, 2004a). The accelerated curation of sequences in the databases improves however the robustness of probe design. An advantage with cDNA arrays has so far been that the sequences are based on actual mRNA transcripts. A drawback has been that up to 30 % of the probes in some cases have been misidentified (Watson, 1998). Resequencing of the printing plates followed by reannotation by blasting eliminates this problem. The most reliable strategy to do conversions between probe sets is by sequence-matching (not by annotation) (Mecham, 2004b; Wang, 2002). Other reasons to differences in results are inconsistencies in experimental procedures and data analysis (Shi, 2005b). The largest platform comparison so far with standardized protocols is the recent study by the MicroArray Quality Control (MAQC) project. Illumina Human-6 Expression BeadChips, 48k v1.0, Agilent Whole Human Genome Oligo Microarrays, Affymetrix HG-U133 Plus 2.0 Arrays reached in this evaluation similar results regarding accuracy and sensitivity (Canales, 2006; Shi, 2006).
1.2 Experimental Strategies

1.2.1 RNA Quality Assessment

For analysis of small samples and to waste minimal amounts, microfluidic capillary systems are widely used, which provides the user with concentration, contamination and size information. In addition to employing the ratio 28S/18S as quality measure an artificial neural network (ANN) can be used which is trained to assess the RNA integrity based on the entire size distribution (Schroeder, 2006). A slight degradation, with still visible ribosomal bands, does not affect the results remarkably for most genes (Schoor, 2003). The Microarray Quality Control Consortium uses a RNA Integrity Number (RIN) > 8.0, 28S/18S ratio > 0.9 as quality criteria.

1.2.2 RNA Amplification

To enable microarray analysis of small samples and low abundant transcripts, methods for amplification of the mRNA population have been presented. In general, around 10-20 µg of total RNA is needed for each hybridizations, which corresponds to 1-2 million cells. Tissues of such large sizes are difficult to obtain and are often non-homogenous, containing a variety of cell types with different stages of disease which can confound the biological effect of interest. The possibility to decrease the amount of RNA needed in the cDNA synthesis may enhance specificity in the analysis since a more homogenous tissue sample can be selected. Small tissue samples can be collected both from biopsies with laser microdissection and in vivo using fine needle aspiration. RNA amplification is also required when analyzing cell types difficult to culture. There are two main strategies for RNA amplification; linear amplification based on in vitro transcription and exponential amplification based on PCR.

Linear amplification: In 1990 Eberwine and colleagues introduced the technique of amplifying mRNA by in vitro transcription maintaining relative transcript abundances for low-throughput gene expression analysis in single neurones (Van Gelder, 1990).
Numerous variations of this technique have been applied and evaluated for microarray analysis and are today a standard procedure for RNA amplification in many labs (Baugh, 2001; Dafforn, 2004; Dumur, 2004; Eberwine, 1996; Freeman, 1999; Lindberg, 2006; Luzzi, 2003; Pabon, 2001; Scherer, 2003; Stoyanova, 2004; Wadenback, 2005).

The use of oligonucleotide arrays has increased need for protocols producing sense strand target. Generation of sense or anti sense in vitro transcribed RNA depends on which primer in the cDNA synthesis is tagged with a transcriptase promoter sequence. The original protocol by Eberwine described above has a T7 promoter on the oligodT primer. Since the transcriptase generates RNA with the same sequence as the strand containing the promoter, the RNA produced will be antisense (aRNA). One strategy of solving this is by introducing labelled nucleotides during in vitro transcription and use the aRNA as target (t Hoen, 2003). One can also add a promoter sequence of another transcriptase to the primer for the second strand synthesis, and then either sense or antisense RNA can be transcribed depending on which viral enzyme is used (Brazma, 2001). The three most common transcriptases employed are T7, T3 and SP6, named after which bacteriophages they are cloned from. Another option is to produce unmodified cDNA and then attach the dye with a platinum complex which binds to guanine. A fourth alternative is to produce an unlabelled cDNA strand from the aRNA and then use that as template for second strand synthesis with random primers and labelled nucleotides using Klenow fragment for elongation (Schlingemann, 2005).

**Exponential amplification:** Generally PCR based protocols provide higher amplification yield in shorter time than in vitro transcription and have also the flexibility that either strand can be used as target. The first PCR based cDNA amplification methods were developed for production of full length cDNA probes. The most common approach for this is template-switching PCR (Chenchik, 1996; Herrer, 2000). In order to enable PCR on a full length cDNA, linker sequences for primers are needed in both ends. In the 3’ end, the polyA tail is used, while in the 5’ end, Cs introduced by the reverse transcriptase from Moloney murine leukaemia virus, MMLV-RT, are employed. For the probe application, amplification yield and product length were the most important criteria for the quality of
the procedure. Less attention was paid to whether relative transcript abundances are preserved and reproducibility of the method, which are the key features for a target amplification protocol. Evaluations of the method for this application have however shown results comparable to other strategies (Puskas, 2002; Saghizadeh, 2003). Another possibility is to add a polyA tail with terminal transferase in the end of the cDNA enabling PCR with a single primer containing a oligodT sequence (Iscove, 2002). Since long transcripts are unfavoured in the PCR, strategies for producing short cDNAs are applied. This can be accomplished by decreasing incubation time during reverse transcription and using limiting amounts of nucleotides or by fragmentation through sonication (Hertzberg, 2001; Iscove, 2002; Sievertzon, 2004).

In an experimental set up, amplification and labelling strategies are certainly factors that need to be considered if several methods are used. Since hybridization intensities are altered to a greater extent than ratios, samples that are to be compared need to be handled with the same protocol or in an appropriate experimental design. It is sometimes a good idea to use amplification even though the material is sufficient for microarray expression analysis of the vast majority of genes.

### 1.2.3 Target Labelling

Labelling efficiency of the target molecules is a main issue for transfer of the transcript abundances to detectable reliable signals. In the standardized method for GeneChip technology, biotin labelled cRNA is in vitro transcribed from double stranded cDNA. Prior to hybridization, the cRNA is fragmented to reduce cross hybridization. After the hybridization, streptavidin labelled with phycoerythrin is applied in two stains enabling detection, with a biotinlylated anti-streptavidin antibody staining in between.

For dual channel microarrays the most common fluorophores are cyanines 3 and 5 and Alexa Fluors. These are either introduced by direct or indirect labelling corresponding to using labelled nucleotides in the reverse transcription or using aminoallyl labelled nucleotides in which case the fluorophores are attached after cDNA synthesis by a
diestrification reaction. The latter method is more time consuming but produces higher amounts of cDNA and causes less dye effects.

Signal enhancement may be an alternative when there are limiting amounts of RNA available, the main two strategies being tyramide signal amplification and dendrimer technology. These have proven to give reproducible reliable signals with up to 100–fold amplification (Karsten, 2002; Manduchi, 2002).

### 1.2.4 Hybridization

Hybridization conditions are optimized to produce as specific and high signals as possible compared to the background noise level. Longer incubation time increases specificity pushing the reaction towards equilibrium. Salt increases hybridization while denaturing agents such as formamide decrease it lowering secondary structures among target molecules and allowing lower hybridization temperature which reduces evaporation. To avoid cross hybridization unlabelled repetitive and common sequences are added to the buffer.

### 1.2.5 Scanning

Measurements of how much target has hybridized to each spot are obtained using a high resolution confocal scanner. The fluorophores are excited with laser light corresponding to their optimal excitation wavelength. A photomultiplier tube (PMT) is used for detection. Adjustment of signal intensities can be achieved by changing the PMT setting and/or the laser power to retrieve as high signals as possible without reaching the saturation detection limit. To widen the range several scans at different settings can be combined. To improve accuracy, the lowest signals could be neglected or, alternatively, be subject to a calibration scheme (Bengtsson, 2004; Lyng, 2004; Shi, 2005a).
1.3 Data analysis

A main issue in microarray studies is how to retrieve valuable information from the enormous amount of generated data. Since the field is so new, methods are being developed continuously to meet the demands of biological researchers. The main processes in the data analysis are extraction of spot signals, filtering, normalization, assessment of differential expression, clustering and classification, and placing the study into context of other sources of information, such as biological databases, clinical features and other microarray experiments. There are several reviews which summarize these procedures (Brazma, 2001; Quackenbush, 2006).

1.3.1 Image Analysis

The 16-bit tiff image files obtained from the scanner after completed microarray hybridizations contain a matrix of values ranging from 1 to $2^{16} - 1$ (65,535) describing the signal for each pixel of the slide. To identify which pixels belong to features, image segmentation techniques are required. The first step is to apply a grid to the entire slide determining the position of each feature. Thereafter the shape and size of each spot is calculated. Depending on conformity and circularity of the spot, either adaptive circles or seed growing is used (Adams and Bischof, 1994). For in situ synthesized arrays this step is unnecessary, since features in this case always take the same shape. The intensities for each feature are estimated with the mean or median of pixel values.

1.3.2 Background Correction

The slide background is often used for adjustment of spot foreground intensities and quality acquisition. Subtracting the background to improve quality of feature values is however not entirely accepted since the slide surface surrounding the spots might have different properties than the areas with applied DNA and printing buffer. This difference can to some extent be examined by applying control spots of DNA lacking complimentary sequences in the target. The GenePix software from Axon Instruments estimates the
background with pixels in a fixed area around the spot (Axon, 2005). The median of these is usually used to reduce influence of dust particles. Another method is the morphological opening implemented in the Spot software (Buckley, 2003). Morphological opening is a way of obtaining a smoothed image of the background without any foreground intensities. Trends over the image is estimated and then removed by moving a sliding window twice over the entire image. The first time the pixel in the centre of the window is replaced by the minimum value of all pixels in the window. The next time each pixel is replaced by the maximum value. The window is chosen to be larger than the feature diameter so that spots will disappear in the transformed image. In the Affymetrix software MAS 5.0, the local background estimation is obtained by dividing the slide surface into cells where the median of the background is first calculated. This value is then adjusted to the other cells weighted according to how distant they are (Affymetrix, 2001).

1.3.3 Quality Assessment and Filters

The image analysis software provides the user with a file containing different measurements which can be used for quality assessment, including mean and median of foreground and background intensities, standard deviation, spot diameter and pixel regression ratio. Diagnostic plots revealing systematic trends and artefacts are useful for determining the quality by manual inspection. Spatial trends over the slide surface are visualized by plotting the median background intensities, either one channel at a time or the log2 ratios of both.

Box-plots of the log2 ratios for entire slides or printing-pin blocks within a slide give information of if the log2 ratio distribution is subject to any bias. A ratio-intensity plot, displaying how the log2 ratios vary with increasing intensity is usually a way to get an overall picture of the hybridization results. For single channel arrays a ratio-intensity plot is created by using the average of all slides instead of a second channel. A spread in the leftmost part of the plot reflects the uncertainty of low intensity measurements and bias introduced by the background correction, while artefacts in the high intensity part might
be due to saturation. By applying a set of filters, spots with very weak signals, high background or uneven signals are removed from further analysis not to contaminate intensity values of high quality spots.

The Affymetrix system with 11-20 probe pairs per feature, each probe pair composed by one with perfect matching sequence (PM) and one containing a mismatch (MM), can use the replicates for quality control on several levels. Probes are designed throughout the entire transcript but with emphasis on the 3’ end. An overall measure of the degradation of the RNA sample can be assessed by calculating the ratio between 5’ and 3’ probes. In MAS 5.0 adjustment for cross-hybridization is obtained by subtracting the intensity for the MM probe from the PM probe. Some studies indicate however that it is better to use just the PM intensity (Naef, 2003). It has also been observed that in many cases MM intensity is higher than PM (Irizarry, 2003). Secondly, to get a single value for each transcript a robust average of all the PM-MM differences is calculated using Tukey biweights. Since the variation for a probe across slides has been observed to be lower than the variation within a probe set on a single slide a multiplicative model accounting for probe affinity has been proposed (Li and Wong, 2001). Based on this a robust linear model accounting the infinity effect using a log scale has been proposed by Irizarry. Special measures can be taken in the analysis of amplified target (Cope, 2006).

### 1.3.4 Normalization

To be able to retrieve reliable biological results comparing data from several experiments with different systematic bias, numerous normalization strategies for microarray data have been suggested.

Quantile normalization is a method mainly used for single channel arrays, but is also applied to dual channel data when several distributions are present. The goal is to transform data from different slides (or channels) so that equal distributions are obtained. This is performed by ranking the feature intensities and than calculating the average of all slides for data being on the same rank position. The value for each spot is then replaced with the calculated mean corresponding to the rank on each slide. This means that exactly
the same values will be present in every slide, but assigned to different features. This transformation might seem unrealistic and harsh to the data but has nevertheless proven to perform superior to other methods for Affymetrix GeneChip compared to other methods (Bolstad, 2003).

Locally weighted scatter-plot smoothing, or lowess normalization is the most widely used method for dual channel data. It is calculated by fitting a polynomial to the scatter plot using data in local intensity windows (Cleveland, 1979). Examining the ratio-intensity plot, a curvature shape of the feature values is sometimes observed, revealing an intensity dependence for the ratio distribution. To compensate for this Yang proposed the use of lowess for microarrays which also has been investigated by many others further on (Berger, 2004; Yang, 2001).

Spatial ratio effects are often influenced by the background intensities, and thus dependent on which methods for background correction have been applied. If no background subtraction is used, spatial normalization can compensate this. Uneven hybridization, washing or bleaching may all give contributions to spatial ratio gradients. One possibility is to normalize within each block on the array, which each contains features printed with the same printing pin. This strategy is sometimes referred to as normalization for bias introduced by different print-tips, although these seldom infer any bias. More detailed spatial normalization can be carried out based on the background intensities or on the spatial ratio gradients. As for intensity dependent ratio bias, spatial bias can be corrected for with lowess, fitting a locally weighted polynomial over the slide surface. Normalization between slides can be required also for dual-channel arrays and are then performed with the quantile normalization described above using ratios instead of intensities.

Although these methods are the most widespread, and have shown good results in comparison with others, each data set is unique and has its own tendencies which must be adjusted for. One such situation is when the expression is unevenly altered for a high proportion of genes, for instance when global transcription rates are lowered or if we have a chip with few features. In this case one can use specific control spots printed in different
concentrations on the chip, containing sequences for either exogenous RNA spiked into the target, or housekeeping genes which are supposed to be present to constant levels in all samples. For prokaryotes, genomic DNA can be used as a more robust alternative to housekeeping genes, enabled by their lack of introns.
1.3.5 Experimental Design

The lack of long experience of microarray technology makes it not straightforward to foresee all sources of variation which need to be accounted for to make accurately randomized block experiments with enough replicates to measure the inferences of interest. The common involvement of several research groups at different levels in the experiment also complicates the overview of experimental factors.

Having identified all biological questions of interest in the experimental setting and collected samples in a fashion that these could be answered and not confounded, the large number of technical sources of variation needs to be considered. Several models for estimations of biases have been suggested to adjust for dye-, array-, spatial- and spot effects, as described above. Systematic bias may be removed with normalization methods if only the experimental setup is adequate.

Design of microarray experiments is dependent on the maximum number of hybridizations possible with each sample, the biological and technical variation, and the number of arrays which can be afforded. To reduce the number one might consider pooling of biological RNA samples (Glass, 2005; Peng, 2003; Vasselli, 2003). The risk with mixing several samples on the same array is however that one might not be able to draw any conclusions about the entire population, nor single individuals, since the range of possible expression is so wide. A highly expressed transcript in the pool can be due to over expression in a single sample.

An important factor in choice of design is the need of replication for correct estimation of the variation. Two main types of replicates are considered, biological replicates, referring to patients and cell cultures etc, and technical replicates, performed to estimate variation introduced during laboratory work. Biological replicates are chosen to be representative for the population investigated, while the technical replicates often are just repetitions of the experiment to average out noise and enable balanced designs. The most common types of technical replication are spot and hybridization replicates. The former can be accomplished by printing the same material in different positions on the array to get
spatial variation and possibly at different concentrations to span several intensities, while the latter corresponds to repeating the procedures from reverse transcription of the target and further on.

Comparisons using dual channel arrays can be either direct or indirect, according to figure 2. Direct designs are suitable for comparisons of only two samples, or when the data are paired, where there is no interest in examining variation between samples belonging to different pairs. Indirect comparisons are performed via a reference which is hybridized on the arrays together with one sample at a time. The reference can be either of biological interest, such as time point zero in a time course experiment, or just a tool for comparison. For this purpose commercial RNA is available (Novoradovskaya, 2004). Reference designs are inefficient statistically, not taking full advantage of the possibility to join two samples on the same microarray. A statistically more powerful strategy is the interwoven loop design. An optimal design aiming at minimizing the sum of all variances is obtained by finding the minimum of \((X^TX)^{-1}\), X being the design matrix (Wit and McClure, 2004).

Unequal incorporation of the different dyes causes a dye bias which needs to be considered in the design. To account for the dye effect a in a direct comparison a dye swap experiment is required, with samples labelled in the opposite manner. The dye bias is circumvented using a reference design since all samples to be compared can be labelled with the same dye. Using the interwoven loop design, the dye effect need to be considered, not necessarily by doing a dye swap for each hybridization, but by controlling that the design is dye balanced, meaning that each sample is labelled equally many times with each dye.
Kerr and colleagues proposed an ANOVA model to consider all effects introducing variance in microarray data (Kerr and Churchill, 2001). This eliminates the need of additional normalization methods. The calculations need to be recalculated for every test and are computationally demanding which has had a negative effect on its use. In spite of the statistical inefficiency, reference design has so far been the most popular choice in large studies. The flexibility of being able to expand sample size gradually is an attractive ability.

### 1.3.6 Differentially expressed genes

A core objective in microarray data analysis is to identify genes whose transcript levels have been altered between different conditions. Usually this is presented as a ranking list of genes according to a statistical significance score. In the early days of microarray technology articles were published using fold change as the single parameter to make a decision if a gene had changed or not (Carninci, 2005; Rihn, 2000; Sgroi, 1999). Considering the large number of genes and the small number of replicates the number of false positives will be considerable. Especially features with high variation can obtain a high fold change value. A few of the large number of non-differentially expressed genes will get deviating intensities in some experiments due to noise and thus falsely be assigned as differentially expressed.
Using an ordinary t-test, the variation for each gene will be taken into consideration, so that genes with varying intensities within a condition will not be selected. This improvement is however not enough since there is a risk that some genes will have a small standard deviation by chance. A very small standard deviation will result in a high t-value even though the gene is not differentially expressed. The widespread use of microarrays in concert with lack of foolproof techniques for data analysis has laid the ground to a whole new scientific area of developing new statistical strategies for microarray data. More than hundred articles have been published each presenting a method especially designed for assigning differential expression of microarray data. A major problem is to obtain an estimate of the true variation when there are so few replicates for each measurement. One proposed strategy is to take advantage of that noise tends to decrease with intensity and use a sliding window for calculation of a variation score relative to spots in the same window (Yang, 2002).

Today modified t-tests have become the most widely used approaches. The main purpose of these is to decrease statistical significance of genes with small standard deviation which otherwise would become false positives. The modification normally consists of a small value added to the standard deviation when the t-statistic is calculated. Because of its statistical power and user friendly implementations, the most popular of these methods is significance analysis of microarrays (SAM) (Tusher, 2001). The value added to the standard deviation is chosen to minimize variation of the t-value. An alternative approach which has shown to be equally good is empirical Bayes statistics (Lonnstedt and Speed, 2002; Smyth, 2004). This approach uses a prior estimate of how many genes are differentially expressed and a parameter deciding the importance of this prior distribution will contribute to the posterior probability. A penalized t-statistics can be formed also in this case.

How long ranking list of differentially expressed genes will be selected may be based on variation in the dataset and which follow up studies will be performed on the data, i.e. weighing cost of false positives against the gain of making a revolutionary discovery.
The large number of hypothesis tests due to the high number of genes in each array requires adjustment for multiple testing. To adjust for multiple testing and to control the error rate either the family wise error rate (FWER) or the false discovery rate (FDR) can be calculated. FWER is the probability of finding at least one false positive among the selected genes, while the FDR is the fraction of false positives. FWER is often considered as too conservative for microarray experiments, therefore FDR is the method most widely used.

1.3.7 Clustering and Classification

Clustering is performed to divide the massive amounts of gene expression data into groups based on similarity. This can be accomplished with two different strategies used for two different purposes; unsupervised clustering for exploratory analysis, and supervised clustering, which can be used to create a diagnostic device based on gene expression signatures.

**Unsupervised clustering:** Unsupervised clustering is a way of obtaining a more comprehensible representation of the data set. With reduction techniques, such as principal component analysis (PCA), singular value decomposition (SVD) and multidimensional scaling, it is possible to visualize the data in two or three dimensional space so that the distance relationships can be explored by visual inspection.

For several of these methods missing values are not tolerated. To solve this, one could either remove features with missing values from the data set or alternatively replace them with an estimate. If technical replicates are available the averages of these can be used, otherwise there are several strategies which have been proposed for microarray data including k-nearest-neighbours, single value decomposition and local least squares (Alter, 2000; Kim, 2005; Troyanskaya, 2001).

Clustering can be performed either on genes or samples, or on both, showing relationships between genes and samples which might be of biological relevance. Among the most common methods for exploratory grouping, or clustering, are hierarchical clustering, self-
organising maps (SOM) (Kohonen, 2000; Toronen, 1999) and K-means clustering (Hartigan and Wong, 1979).

Hierarchical clustering is an agglomerative method which produces a dendogram with a bottom-up structure. To start with all data points are treated as separate clusters. The dendogram is then formed by subsequently assigning the two closest clusters together to form new clusters. Distance between clusters can be calculated by using the minimum, maximum or the average distance between samples in two different clusters.

For K-means and SOM, the number of clusters to be formed is predetermined by the user. In K-means clustering, the samples are first randomly assigned to one of the k clusters. The centre of each cluster is obtained by calculating the mean or median of all incorporated expression profiles. The samples are then reassigned to the closest cluster. New cluster locations are calculated from the new cluster members, followed by another reassigning procedure. This procedure is repeated until there are no more changes.

The SOM algorithm is similar to K-means assigning samples to the closest cluster in an iterative process. The difference is that the samples in this case, are mapped onto a two dimensional grid. To determine optimal numbers of clusters the ratio between cluster diameter and distance between clusters can be used.

**Supervised classification:** Supervised classification on the other hand is performed to create a tool which can be used for discrimination of new data. Development of the classifier is based on prior information of how for instance tumour types are connected to appearance of the gene expression profiles. Supervised clustering techniques thus needs a data set with information of the sample labels, for instance if a tumour is malignant or not. The aim is to create a classifier which can classify new unlabelled samples and genes based on the microarray data. These can be divided into machine learning algorithms such as support vector machines (SVM), ANN and k-nearest neighbours (KNN), and statistical linear discriminate analysis.

SVM was introduced by Vapnik in 1995 and applied to microarray data in 2000 (Brown, 2000; Furey, 2000). The algorithm calculates a separating hyperplane with maximized margins to the data points. SVM has been much used in this field compared to the other
methods mainly due to its capability of providing good generality despite the small sample sizes typical for microarray experiments (Furey, 2000).

A weakness for methods based on a molecular signature of thousands of features rather than a small set is that they are difficult to interpret in a biological context and not applicable to lower throughput techniques. Combining these however with feature selection methods makes them more useful and also reduces risk of over-fitting. Filtering and pruning are examples of strategies for feature selection.

Filtering is based on the input data by filtering features based on a specific criterium. Selection can be performed based on a threshold or a number of top genes on a ranking or by evaluating the performance. In the first microarray classification study published, Golub selected a subset of genes from a Signal-to-Noise (S2N) ranking list (Golub, 1999b). Genes with large difference in mean value between groups and small within group standard deviation are selected. A disadvantage with this approach is that many of the genes will be very much correlated containing redundant information.

Pruning on the other hand changes the classifier to reduce its complexity and avoid overfitting. When multilayer networks are used this is a complicated task. A trade off between network performance and complexity is made by minimizing the sum of all weights and/or the number of weights. In the SVM case Guyon and colleagues have shown that ranking weight vector elements according to magnitude, eliminating the smallest weight elements one at a time with consecutive evaluation, outperforms filtering techniques (Guyon, 2002). In contrast to filtering techniques the genes selected by this method do not contain as much redundant information.

1.3.8 Extraction of Biological Information

To find biological reasons why genes appear as differentially expressed or in the same subclusters one tries to find biological features which are overrepresented in these groups. Examples of biological annotations which may be used for this purpose are Gene Ontology terms, Swissprot key words and KEGG pathways (Ashburner, 2000; Kanehisa, 2004). Statistical significance of the enrichment analysis can be calculated with Fischer’s exact
test with adjustment for multiple testing (Al-Shahrour, 2004; Castillo-Davis and Hartl, 2003; Dennis, 2003; Zeeberg, 2003). Gene set enrichment analysis (GSEA) is a recently developed method in which a Kolomogorov-Smirnov statistic is used to evaluate if a ranking list is enriched for a certain gene set at the extremes (top or bottom) (Subramanian, 2005). FDR is used to assign significance and is obtained by comparing to the null distribution which is obtained by permuting samples labels. A similar strategy is the significance analysis of function and expression (SAFE) (Barry, 2005). There are several software which enables comprehensible visualization of gene expression levels in the different pathways such as Pathway Assist, GenMAPP and MAPPFinder (Dahlquist, 2002; Doniger, 2003; Nikitin, 2003). One reason to genes being concordantly expressed is regulation by the same transcription factors. In eukaryotes combinations of transcription factors form cis regulatory modules for a gene to be expressed. The sets of sequence elements in the promoter region which are subject to transcription factor binding are often conserved between species. The co-occurrence of transcription factors in conserved promoter regions can be evaluated with software such as CRÈME, JASPAR and cisRED (Robertson, 2006; Sharan, 2004; Vlieghe, 2006).

### 1.3.9 Software for Data Analysis

The rapid development of methods for data handling puts high requirements on the software development. Besides a large amount of bioinfomatic tools designed to make specific operations there are several commercial (e.g. Genespring, ArrayAssist and Kensington Discovery Environment) and open-source software (Mev 4, Bioconductor) providing complete solutions (Gentleman, 2004; Yang, 2002). The R language for statistical computing has been increasingly popular much due to that a large number of statistical tools are available open-source and that many new methods for microarray data are written in R.
1.3.10 Public Data Bases

The massive data amount generated in a microarray study often exceeds the area of the particular research group who produced it. To accelerate research efficiency it is therefore practical to have data publicly available for others to explore. In the beginning this was primarily realized on the website of each research group or by the journal publishing the study. This spread of data made it rather a time consuming task for other researchers to find and collect the information.

In response to the exponentially increasing number of microarray data sets published and the urge to be able to do meta analysis of data from several data sets, several public repository data bases for storage of microarray data have been launched. The largest one is the Gene Expression Omnibus at NCBI hosting 4493 experiment series (Nov. 2006, http://ncbi.nlm.nih.gov/geo/) (Edgar, 2002). In Europe the most popular database is ArrayExpress, online since 2002, accommodated by the European Bioinformatic Institute (Sarkans, 2005). Today more than 30,000 hybridizations from over 1000 studies are stored here (http://www.ebi.ac.uk/arrayexpress/). To streamline the presentation of microarray studies the Microarray Gene Expression Data society (MGED) has proposed a format which holds the Minimum Information About a Microarray Experiment (MIAME) (Brazma, 2001). Upon publication of a study it is nowadays often required to have the data publicly available in the MIAME format.

1.4 Quality Control

Quality control of microarray experiments can be carried out on several levels. RNA pools with known transcript abundances have been developed to be used for evaluation of laboratory and data management (Thompson, 2005). These can also be used for cross-laboratory and platform comparisons.

To enable control on an analysis of actual samples, spike in RNA can be used which hybridize to special control spots on the microarray. For spotted arrays small sets (~10 probes) of spike in controls which spans a range of intensities and ratios are available from several companies. The External RNA Control Consortium is a volunteer organization of a
large number of companies and academic institutions aiming at developing a platform independent set of 100 spike in sequences which are made publicly available (ERCC, 2005). Affymetrix provides a special data set which can be used to evaluate the performance of methods for data analysis.

A supplementary approach of controlling the results is to examine the expression of genes with another method, mainly quantitative real-time PCR (QRT-PCR) (Canales, 2006; Higuchi, 1993; Lundholm, 2004; Rajeevan, 2001). Discordant results can be due to that different splice variants are measured and cross-hybridization on the microarrays.
2 Microarray-based Gene Expression Analysis in Cancer Research

Cancer is a major health concern in the western world. In Sweden 45,000 people get the diagnosis each year. Traditionally, tumours have been categorized on the basis of histological features. However, the information gained from microscopic examination of staining patterns does not reveal underlying molecular events involved in cancerogenesis and progression. To obtain a deeper understanding of the biology in tumours, adoption of microarrays is a powerful strategy which uncovers the gene expression signatures associated with different phenotypes at the same time.

The multistep process of cancer development starts with a genetic or epigenetic alteration in a cell giving it a growth advantage. As aberrant cells propagate genetic instability is accumulated and abnormal characteristics are gained. Cancer cells are typically self-sufficient in growth signals, insensitive to apoptotic stimuli and have an increased telomerase activity enabling a larger number of cell divisions (Hanahan and Weinberg, 2000). As the tumour grows it gains abilities to break through surrounding tissue and induce angiogenesis to supply the tumour with nutrients and oxygen. The cancer can also spread to other organs distant from the primary tumour. Ninety percent of cancer deaths are caused by secondary tumours, metastasis (Sporn, 1997). For these to arise, malignant cells need to detach from the primary tumour, break through into blood vessels, survive in the circulation, and finally migrate into a new site. Genes that are important for cancer development are commonly called oncogenes and tumour suppressor genes. Genetic alterations of proto-oncogenes which turn them into their active oncogenic form imply a gain of function which promotes tumour growth. Examples of these are growth factors
and their receptors, cell cycle regulators and angiogenic inducers. Tumour suppressor genes on the other hand prohibit malignant growth by inducing actions such as apoptosis or cell cycle arrest. These genes are disabled to give the cell cancerous characteristics. A novel dimension to the mechanisms behind tumor development is given by the miRNA population which also may act as tumor suppressors and oncogenes (Chen, 2005). Cancer is a very heterogenous disease contributing to the complexity of obtaining a general picture of its development and progress. A variety of alterations can end up in similar behaviours. An example of how five distinct events have been linked to specific stages is the famous model for colorectal cancer proposed by Vogelstein and coworkers (Vogelstein, 1988).

2.1 Biomarker Discovery

Search of potential biomarkers based on gene expression profiles generated by microarray experiments have been applied to the major cancer forms during the past ten years, including breast, prostate, leukaemia, ovary, and lung cancer (Adib, 2004; Bhattacharjee, 2001; Dhanasekaran, 2001; Golub, 1999a; Hedenfalk, 2001; Ono, 2000) Genes have been identified whose expression correlates with tumour grade, metastasis, survival and recurrence. Gene expression profiling is commonly the first step in a screening process (Figure 3).

![Figure 3. Example of work flow; potential biomarkers are identified with cDNA microarrays, after verification of clone sequence and expression on protein level, tumour tissues and possibly serum can be used for screening.](image-url)
Follow up studies can include epigenetic examination of specific regions, tissue microarrays of protein expression and serum screening, vastly broadening the collection of diagnostic approaches (Dhanasekaran, 2001; Monni, 2001; Petricoin, 2002). Depending on the context, if the biomarker is to be used in a population screening or to improve diagnosis of a detected tumour, it has to meet different demands. The increasing rate of publications on biomarkers has not yet lead to a similar trend in clinical practice. The number of approved biomarkers by FDA each year has in contrast decreased during the last decade. Many biomarkers are under clinical evaluation and there are high hopes that in the near future the promising findings of the many microarray studies will find their use in clinical applications.

2.2 Tumour Classification

Gene expression profiles can be used to separate tumours into new and well established tumour types. The purpose is that a more precise knowledge of the tumour biology will lead to development of more efficient therapies.

The first study showing the possibility to perform class discovery using gene expression microarrays were a leukaemia study conducted by Golub and colleagues published in 1999 (Golub, 1999a). A classifier was constructed to recognize acute myloid leukaemia and acute lymphoblastic leukaemia. The classifier also happened to make impact clinically since it was able to improve the diagnosis of a patient leading to altered treatment regime. Since then microarrays have been used to examine clinically relevant tumour subtypes in breast, prostate and lung cancer (Alizadeh, 2000; Dhanasekaran, 2001; Hedenfalk, 2001; Lapointe, 2004; Perou, 2000; Sorlie, 2003). In Hedenfalk’s study on hereditary breast adenocarcinomas with mutations in either BRCA1 or BRCA2 it was demonstrated the value of gene expression analysis compared to genotyping. A supervised classifier was able to separate the two groups based on a subset of the genes. Interestingly a spontaneous tumour was classified as a BRCA1 mutation whereas sequencing showed no mutation. The reason was aberrant methylation in the promoter region. In the breast cancer study by
Perou and coworkers expression profiles of 42 breast cancer tumours were subject to hierarchical clustering revealing three relevant groups, one of them being the known subtype Erb-B2+. It was also demonstrated that primary tumours were more similar to their metastasis than to other primary tumours. Sorlie in a study of 115 breast cancers also found that clustering corresponded to biological relevant subtypes, Erb-B2+ was one of these four. Genes selected for clustering were those low variation for samples taken from the same patient with 15 weeks of neoadjuvent therapy inbetween and with large variation between patients. Clustering with these genes on two other independent data sets gave similar results. Three subtypes of prostate cancer tumours were identified in the study by Lapointe of which one was associated with recurrence. Alizadeh performed unsupervised classification on B-cell lymphomas and found that the clusters formed reflected different stages of cancer progression.

Most studies are performed focused on a specific cancer. There is however a gain in collecting data from multiple cancer forms to enable constructure of a classifier which can connect metastasis to tissue of origin. One of the few examples of this is a multiclass cancer classification performed on 14 tumour types and 90 normal tissue samples with a one-versus-all SVM (Ramaswamy, 2001). Accuracy of classification was 78%.

Many studies are carried out in cancer cell lines in vitro which provokes the reliability of the results since the course of events might depend on natural environment of the cells. A large comparison of the NCI60 cell lines to their tissue of origin was therefore carried out which showed high resemblance in gene expression patterns between in vitro and in vivo expression (Tong, 2006).
2.3 Therapeutic Development

Gene expression profiling is an important tool at many stages in the developmental process of therapeutic agents. Initially in the screening step patterns of healthy samples are compared to those of diseased and the differentially expressed genes are searched for potential drug targets. Two to three thousand genes including G protein coupled receptors, nuclear receptors, ion channels and protein kinases are potentially druggable (van Duin, 2003). To determine the importance of selected genes these can be knocked down one at a time by interference RNA (RNAi), antibody or affibody inhibition and then the change can be studied. Once a potential drug well has been developed the response of this can be compared to that of the RNAi. The power of the treatment and possible toxic bireactions can thus be discovered.

NCI60 is a panel of 60 cancer cell lines which have been used for screening of drug sensitivity. To date, the response of more than 100,000 agents have been evaluated with these cell lines. Correlations of gene expression profiles with resistance or sensitivity to several substances have been identified. As an example the rate limiting enzyme in fluoracil metabolism, dihydropyrimidine, was inversely correlated to sensitivity to fluoracil (Scherf, 2000). Supervised classification has been successfully used to also predict chemosensitivity for some compounds (Staunton, 2001).

The connectivity map is a recently published resource which can be used to find connections between chemical agents, diseases and gene expression signatures (Lamb, 2006). One hundred and sixty-four molecules have been studied in one to four cell lines and analyzed with Affymetrix GeneChip arrays. Connections are computed with the GSEA algorithm described in 1.3.9. When querying the connectivity map with a gene expression signature derived from a compound with unknown function, an idea of its effect can be obtained if there are strong connections to agents with known functions.

Another example of the progress of microarray technology within the chemogenomics field is a database with expression profiles of seven different types of rat tissues treated
with approximately 600 different compounds. Effects in liver could as an example be predicted by the gene expression patterns in the database (Thompson, 2005).
3 Present Investigation

3.1 Evaluation of Two Strategies for RNA Amplification

RNA amplification is employed in microarray analysis when the amount of sample is limited. The two main strategies are linear amplification using *in vitro* transcription and exponential amplification with PCR. A broader background of available methods is given in section 1.2.3.

3.1.1. 3' cDNA Tag Amplification

Amplification of full length cDNA with PCR is often criticized for not being able to produce satisfactory amounts of the longest transcripts. In PCR, short fragments are generally favoured. To prevent this size dependent imbalance in the PCR Sievertzon and coworkers introduced a technique based on short 3' end cDNA tags. An outline of this approach is presented in figure 4. First strand cDNA is synthesized using isolated mRNA and a biotinylated and anchored oligodT primer containing a linker with a *NotI* cleavage site. The second strand is obtained by partitioning the mRNA strand into primers with RNaseH, polymerization of DNA with T4 DNA polymerase and ligation with DNA ligase. The double stranded DNA is thereafter collected by precipitation, washed in a gel column and fragmented by sonication. The produced 3' end fragments are immobilized onto streptavidin beads and a linker is added by blunt end ligation. PCR is performed with primers using the *NotI* sequence and the 5' linker. To generate an even size distribution of
the fragments the number of PCR cycles is optimized to achieve an even smear on an agarose gel. Direct labelling is carried out in an asymmetric PCR with labelled nucleotides.

Figure 4. Amplification protocol based on PCR of short 3'end cDNA tags
3.1.2 Paper I

In this study we compared the 3' end cDNA PCR with a commercial kit for linear amplification based on \textit{in vitro} transcription by investigating the difference to unamplified material in microarray experiments. We also used \textit{in vitro} transcribed products on oligo microarrays with a novel labelling kit but with a large drop out due to that the probes were not only in the 3' end. For the comparison RNA from two cancer cell lines was analyzed on spotted cDNA microarrays. As starting material for amplification 100 ng of total RNA was used, which invokes two cycles of \textit{in vitro} transcription. To assure that no confounding effects were introduced by the mRNA isolation in the exponential protocol, unamplified material was analysed both with and without prior mRNA isolation. Amplifications and hybridizations were carried out in two replicates. The comparison was based on the Pearson correlations of log: ratios from hybridizations of the two cell lines between and within methods.

The size distribution of amplified material was between 100 and 600 bp for both amplification protocols. Pearson correlations between amplification replicates showed a better reproducibility for the PCR based method. Comparison to unamplified material showed that both amplification protocols rendered reliable data.
3.2 Detection of p53 Dependent Genes Induced by Hypoxia

3.2.1 P53 - Cancer Gene

P53 has been one of the most extensively studied genes during the last decades. It was independently identified as a cancer gene in 1979 by Levine, Lane and Old (DeLeo, 1979; Lane and Crawford, 1980). The P53 protein attracted the attention of many researchers due to its over-expression in malignant cells, and was also therefore presumed to be an oncogene. After ten years of misinterpretations, Vogelstein and Raymond White finally showed that p53 was mutated in cancer cells and that the wild type protein on the contrary protected against cancer (a tumour suppressor gene) (Baker, 1989).

There are four conserved domains in p53 protein; the N-terminal which is required for transcriptional transactivation (Fields and Jang, 1990), a sequence-specific DNA binding domain (Pavletich, 1993; Wang, 1993), a tetramerization domain near the C-terminal end, and the C-terminal domain which interacts unspecifically with single stranded DNA (Foord, 1991). Mutations prevalent in cancers are localized in the DNA binding domain, disrupting the capability of transactivation of downstream target genes (Hollstein, 1994). Inherited mutations in p53 give rise to the Li-Fraumeni syndrome causing a wide range of cancer forms, predominantly breast carcinoma, soft tissue sarcomas, osteosarcoma, brain tumours and adrenocortical carcinoma (James, 1999).

Wild-type p53 protein binds to specific genomic sites with a consensus binding site 5’-PuPuPuC(A/T)(T/A)GPyPyPy-3’ (el-Deiry, 1992). Whole genome in silico promoter analysis has indicated that as many as 2,500 genes can be induced by p53 (Hoh, 2002).

Most target genes are involved in apoptosis, cell cycle arrest and DNA damage control. P53 has been described as "the guardian of the genome", referring to its role in conserving stability by preventing genome mutations. By inducing cell cycle arrest or apoptosis upon DNA damage, p53 prevents cells with improper genome sequence to propagate. P53 is induced also by other types of stress such as hypoxia, hyperproliferation, and oncogene activation.
P53 is expressed consecutively with a short half life in an unstressed environment. Continuous degradation occurs through ubiquitylation by the ubiquitin E3 ligase MDM2 which also is transcriptionally induced by p53 (constituting a feedback loop). Upon stress, degradation is disrupted and p53 is translocated to the nucleus where a homotetramer is formed to act as a transcriptional activator of downstream target genes. Post translational acetylation and phosphorylation promotes both stability and transcriptional activation.

### 3.2.2 Regulation of p53 by Hypoxia

Hypoxia may be the most common physiological inducer of P53 protein in normal body tissues (An, 1998). Which genes p53 affects under hypoxic conditions is however not well characterized. In hypoxic regions of solid tumours, accumulation of wild type p53 protein is strongly correlated to apoptosis. This leads to selection for loss of p53 which in turn leads to a more aggressive propagation of cancer. The ability of having non-apoptotic regions of hypoxia also stimulates angiogenesis with a consequent re-oxygenation of previously hypoxic areas.

The main molecular mediators of hypoxic response are the transcription factor hypoxia inducible factor-1 HIF1 (Zhong, 1999) and the angiogenic factor, vascular endothelial growth factor (VEGF) (Yancopoulos, 2000). Which roles these play together with p53 is not yet fully understood. It has been shown that HIF1-α is required to stabilize the p53 protein for transactivation under hypoxic conditions (An, 1998) and that high levels of p53 can inhibit HIF-inducible transcription via the coactivator p300 (Blagosklonny, 1998).

In contrast to other cellular stress such as UV- and gamma irradiation, several studies have shown that p53 mediates transrepression, but no transactivation under hypoxic conditions (Hammond and Giaccia, 2005).
3.2.3 Paper II

To examine the p53 dependent response under hypoxic conditions we conducted a microarray gene expression study of the human colon cancer cell lines HCT116 p53+/+ and HCT116 p53-/- exposed to hypoxia. RNA was extracted at 0, 8 and 16 hours and hybridized onto cDNA arrays containing 20,352 probes representing 12,454 genes.

First we wanted to confirm that hypoxia indeed induced accumulation of active p53 protein. This was accomplished by performing Western blots and immunofluorescence staining. To examine p53 dependent apoptosis both cell lines were analyzed by TUNEL staining and flow cytometry showing an increased fraction of apoptotic cells in p53+/+ cells. Based on the microarray experiments we wanted to investigate which genes might be involved in the p53 dependent apoptosis. Among all genes with a functional annotation of apoptosis in the Gene Ontology we found that only two genes were altered in a p53 dependent fashion. These were apoptosis inhibitor BIRC3 which was down regulated, and the death receptor Fas which was up-regulated after 16 hours. To investigate the importance of Fas in p53 dependent apoptosis wtp53+/+ cells were treated with an antibody against Fas in order to block Fas signalling. This resulted in resistance to hypoxia induced apoptosis. The Fas pathway was further investigated by treating wtp53+/+ cells with an antibody against caspase 8 inhibitor. Also this impairment of the Fas signalling pathway reduced apoptosis demonstrating its critical role in hypoxia induced apoptosis. A luciferase reporter assay was used to verify that Fas was direct transcriptionally activated by p53.

Many traditional p53 target genes were not induced by hypoxia which also was confirmed with Northern blots and qRT-PCR. However we found several novel p53 dependently expressed genes, such as ANXA1 which has demonstrated to have a proapoptotic effect in circulating neutrophils, SEL1L which overexpressed in mouse tumour xenografts inhibit growth, and SMURF1 and SMURF2 encoding E3 ubiquitin ligases regulating proteins in the TGFβ signalling pathway. Position of putative p53 binding sites in the promoter regions of these genes were calculated with the p53MH algorithm.
Further work to verify the p53 induction of these potential p53 target genes includes ChIP-chip experiments to investigate p53 binding sites in potential target genes and reproducibility of our findings in other cell lines. Future extension of this work involves examination of how these genes participate in inhibition of tumour development and their potential use in drug development.
3.3 Characterization of a Malignant Expression Signature of Adrenocortical Tumours

3.3.2 Adrenal Gland

The adrenal glands are endocrine organs secreting hormones, which are necessary for life in their role of controlling our response to stress, and regulating levels of blood glucose and blood pressure. They are situated above the kidneys, and have a triangular shape of size 0.5x3x4 cm. The cortex and the medulla are two separate endocrine organs with divergent characteristics and origins. The cortex secretes glucocorticoids, mineralocorticoids, aldosterone and sex hormones, while the medulla is responsible for dopamine, adrenaline and noradrenaline production. Hormones exert their activity by binding to cytosolic receptors which transforms and translocates to the nucleus where they bind to hormone responsive elements in the genome (Dahlman, 1989; Fuller and Young, 2005).

3.3.2 Adrenocortical Tumours

Tumours in adrenocortex are relatively frequent, with an incidence of 9% in autopsy studies (Boushey and Dackiw, 2001). Malignancy of these is rare, the yearly incidence being 0.5 - 2 per million inhabitants, but is associated with a very aggressive behaviour, the five year survival being only 10%, if the tumour is resected, and a mean survival of 12 months if non-resected (Dackiw, 2001; Kjellman, 1999; Sidhu, 2003)). The high frequency of benign tumours (adenomas) together with the aggressiveness of cancers causes a high demand of diagnostic capabilities. With the present methods, distinction between adenomas and cancer is often difficult to establish.
Sometimes tumours cause disturbances in the hormone synthesis machinery, generating overproduction of aldosterone, cortisol and in some cases also sex hormones. Adrenocortex consists of three layers; zona glomerulosa, fasciculata and reticularis. Aldosterone is produced in zona glomerulosa and cortisol and sex hormones in fasciculata. Adenomas with overproduction of cortisol give rise to Cushing’s syndrome which is associated with a characteristic obesity, hypertension and thin, easily bruised skin. Aldosterone producing adenomas, so called aldosteronomas, cause hypertension and hypokalemia. Cancer causes a variety of hormonal defects in 50% of cases. Symptoms of these imbalances can also lead to that the tumour can be detected. Incidentalomas are adenomas with no symptoms (detected by incidence). The increased use of computed tomography and radiologic imaging has improved the detection rate of tumours in the adrenocortex the last years.

Treatment is predominantly carried out by surgical resection. This is performed in case of hormonal overproduction or upon suspicion of malignancy. For the latter, a diameter larger than 4 cm is the main criteria (Kendrick, 2001), although in rare cases tumours as small as 2 cm may metastasize (Barnett, 2000).

Less than ten percent of adrenocortical cancers are familial, the most common syndromes being Hereditary Adrenocortical Carcinoma (ADCC), Multiple Endocrine Neoplasia type 1 (MEN 1), Beckwith-Wiedemann Syndrome (BWS) and Li-Fraumeni Syndrome (LFS). CGH analysis of sporadic tumors have shown that 28% of adenomas and more than 60% of cancers have chromosomal alterations (Kjellman, 1996; Sidhu, 2002). The most common ones in cancers are gains on chromosome 4, 5, 17 and 19 and losses on 1, 2q, 17p and 22. Loss of heterozygosity (LOH) on locus 17p13 is highly associated with malignancy and has a prognostic value. A somewhat lower predictor is LOH on locus 11p15 (Gicquel, 2001).

Few global gene expression profiling studies of sporadic adrenocortical tumours have yet been published. The first report was a study of eleven cancers, four adenomas, three normal adrenocortex and one macronodular hyperplasia using an Affymetrix array with 10,500 genes. Many genes were identified as potential biomarkers, the most prominent
being the already well established gene IGF2 (Giordano, 2003). DeFraipont and coworkers examined the expression of 230 candidate genes in 57 adrenocortical tumours (De Fraipont, 2005). An IGF2 cluster was found to predict malignancy as well as a group of genes involved in steroidegenesis.

### 3.3.3 Paper III & IV

To examine underlying biological mechanisms involved in adrenocortical cancers and to screen for potential biomarkers we conducted a microarray study of different phenotypes of adrenocortical tumours as well as normal adrenocortex. A smaller study was first published with focus on the difference between benign and malignant tumours. The extended study included also normal adrenocortex and was also aimed to explore gene expression giving rise to different hormonal activity.

The sample size in the total study consisted of 17 adenomas of which five were overproducing glucocorticoids, four aldosteronomas and eight incidentalomas, and twelve cancers, including six containing high amounts of necrosis and one with uncertain diagnosis. Necrotic samples were included in the data set exclusively in a classifier and the case with uncertain diagnosis only when performing clustering. Clustering with unsupervised methods revealed that adrenocortical cancers have a molecular signature clearly distinguishable from adenomas. The uncertain case clustered with the adenomas. Analysis of differential expression showed accordingly a high number of genes, 1137, with altered expression levels comparing cancer to all other samples, decreasing to 182 genes for aldosteronomas, 68 for incidentalomas and none for cortisol producing adenomas. In agreement with previous studies IGF2 was one of the most up-regulated genes in cancers.

We also found that there was a group of genes highly correlated to IGF (r > 0.97). Among these genes were not unexpectedly the IGF2R, and interestingly also two proteins in the ubiquitin-proteosome pathway.

Enrichment analysis according to Gene Ontology annotations of differentially expressed genes in cancer showed that mitosis and cell adhesion were biological processes which were overrepresented in concordance with the increased proliferation rate and the
invasive characteristics associated with cancer. Since receptors have a potential as targets for treatment and bioimaging we investigated if there were any of these which could distinguish between the tumour types. IGF2R and FGFR4 could serve as markers based on their gene expression levels. For diagnosis of aldosteronomas VEGFB was the best candidate. Secreted proteins are interesting considering diagnosis based on serum screening. Among these we found IGF2, SEMA7A and HAPLN1. Tumours with high degree of necrosis were excluded from the analysis of differentially expressed genes. These did not show the same degree of up-regulation of the IGF2 cluster. To be able to distinguish also these cancers from adenomas we trained a supervised classifier to address this problem. A support vector machine could based on the expression profiles of all genes classify all tumours correctly into adenomas and cancers.

The continuation of this work will in future studies involve examinations of differences in protein levels as well as more thorough investigations of how the differentially expressed genes are involved in tumourigenesis.
4 Concluding Remarks

Microarrays have taken our knowledge, not least within cancer research, major steps on the road to uncovering what is taking place inside cells under different conditions. The last years tremendous efforts have been made to facilitate microarray analysis, both in terms of laboratory management and data handling. Despite the increasing amount of methods, we might head towards more standardized protocols to facilitate fusion of experiments from different laboratories (Bammler, 2005). Alternatively an increased use of external controls in the experiments can ascertain the quality and provide guidance in the choice analysis pathway. Public databases are valuable resources for continuous future interrogation of data sets. Combination of gene expression data with other types of information such as protein expression and regulatory sequence data open new doors to draw conclusions.

In this thesis microarrays have been used for exploration of molecular signatures of adrenocortical tumours. Clinicians are often faced with difficult discriminations between malignant and benign cases. Development of new tools for diagnosis is urgent to provide more efficient treatments to affected patients. Gene expression profiles provide in themselves a potential basis for diagnosis as well as pave the road to more thorough investigations of genes which constitute biomarker potential and give us an insight in the biological processes involved in adrenocortical tumour development.

The role of the cancer gene p53 was examined with microarrays under hypoxic conditions. Microarray analysis revealed a set of potential novel p53 target genes as well as confirmed that many known target genes were not transcriptionally activated by hypoxia.
Follow up which was focused on how p53 affected hypoxia induced apoptosis showed that the death receptor Fas is a critical gene.

When small amounts of tissue are available, amplification of the transcript population is necessary for microarray analysis. A new strategy for amplification based on PCR 3’end tags was evaluated and compared to a commercial in vitro transcription protocol. Both protocols produced results in agreement with results of unamplified target.
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>adenine</td>
</tr>
<tr>
<td>ANOVA</td>
<td>analysis of variance</td>
</tr>
<tr>
<td>ANXA1</td>
<td>annexin A1</td>
</tr>
<tr>
<td>dNTP</td>
<td>2'deoxyribonuleotide tri-phosphate</td>
</tr>
<tr>
<td>FDA</td>
<td>federal drug administration</td>
</tr>
<tr>
<td>FGFR4</td>
<td>fibroblast growth factor receptor 4</td>
</tr>
<tr>
<td>GSEA</td>
<td>gene set enrichment analysis</td>
</tr>
<tr>
<td>HAPLN1</td>
<td>hyaluronan and proteoglycan link protein 1</td>
</tr>
<tr>
<td>IGF2</td>
<td>insulin growth factor 2</td>
</tr>
<tr>
<td>IGF2R</td>
<td>insulin growth factor 2 receptor</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto encyclopedia of genes and genomes</td>
</tr>
<tr>
<td></td>
<td>mdm2, transformed 3T3 cell double minute 2, p53 binding protein (mouse)</td>
</tr>
<tr>
<td>MDM2</td>
<td></td>
</tr>
<tr>
<td>NCBI</td>
<td>national center for biotechnology information</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>Pu</td>
<td>purine</td>
</tr>
<tr>
<td>Py</td>
<td>pyrimidine</td>
</tr>
<tr>
<td>QRT-PCR</td>
<td>qunatitative real time-polymerase chain reaction</td>
</tr>
<tr>
<td>SEL1L</td>
<td>sel-1 suppressor of lin-12-like (C. elegans)</td>
</tr>
<tr>
<td>SEMA7A</td>
<td>semaphorin 7A</td>
</tr>
<tr>
<td>SMURF1</td>
<td>SMAD specific E3 ubiquitin protein ligase 1</td>
</tr>
<tr>
<td>SMURF2</td>
<td>SMAD specific E3 ubiquitin protein ligase 2</td>
</tr>
<tr>
<td>TGFβ</td>
<td>transforming growth factor beta</td>
</tr>
<tr>
<td>U</td>
<td>uracil</td>
</tr>
<tr>
<td>VEGFB</td>
<td>vascular endothelial growth factor β</td>
</tr>
</tbody>
</table>
Acknowledgements

Så ser man målsnöret och slutet på en resa som varit fylld av minnesvärda upplevelser och lärdomar; motgångar ibland men mestadels medgångar tack vare alla trevliga och skickliga handledare och kollegor jag haft turen att arbeta med. Därför skulle jag vilja rikta ett varmt tack till:

Joakim Lundeberg, handledare, för vetenskaplig vägledning och för att jag fått arbeta med spännande och utvecklande projekt, för din smittande framåtanda och stora engagemang.


Mathias Uhlén för att du satt avdelningen på världskartan och för att din entusiasm för forskningen inspirerar oss alla.

Per-Åke Nygren, Sophia Hober, Stefan Ståhl, och övriga gruppledare för att ni gör avdelningen till en så trevlig arbetsplats att vistas på, där sammanhållningen är stor, vårt växande antal till trots.

Samarbeten:

Martin Bäckdahl och Bertil Hamberger på KS för motivation och inspiration under trevliga möten, för att ni delat med er av era kunskaper och genuina engagemang.

David Velasquez, for your positive attitude which lightened up long days, for being a friend and for translating all protocols to spanish (language lesson).

Janos Geli, Catharina Larsson, Ulla Enberg, Anders Höög, Jacob Odeberg, Magnus Kjellman och Christoffer Juhlin på KI för trevligt binjuresamarbete.

Klas Wiman, Galina Selivanova och Tao Liu på KI för trevliga möten och en lyckad publikation.

Göran Wallin och Mirna Nordling på KS och examensarbetare Roland Sjögren för trevligt samarbete om hyperthyreos.

Fredrik Pontén och cancergruppen i Uppsala för trevliga och lärorika möten, goda lunchmackor, samt konferensen i Sevilla.

Carl Henrik, Emma och Thomas för trevligt projektarbete på ANN-kursen

Kollegor:

B3:1049 Speciellt tack vare alla omtänksamma nuvarande och pensionerade rumskamrater har det varit ett verkligt kul att gå till jobbet, diskussioner om allt från stort till smått, också vetenskapligt, har fylld vardagen med trivsel; Emma, Esther, Karin, Mats, Olof, Ronny och Tove.


Anders för excellent handledning av examensarbetet.

Tjejerna för alla trevliga, middagar, förfester, spelkvällar, träning, och kulturella events.

Alla övriga trevliga prickar på labb och kontor, som bidrar till den goda atmosfären.

Gamla vänner:

Elisabeth, Jenny, Karin, Linda, Maria och Rebecca för allt trevligt under åren på kemi som gick alldeles för snabbt tack vare er, samt alla middagar, kalas och essentiella skärgårdshelger, som ger perspektiv.

Karin, för lång vänskap. Även om man inte ses så ofta längre när man bor i olika länder är det alltid lika kul.

Annelie för många och långa mil i löp- och skidspår, för att du varit en vän i alla väder under åren, för fjällresor, båtbestyr, och allt vi tagit oss igenom tillsammans.

Cia för att du är den du är, den bästa vän man kan ha. Utflykter till Dalarna, vandringar, semesterresor, squash och inte minst räddningsaktionen i skärgården när jag fick soppa torsk. Nicklas också, förstås. För att ni påminner mig om livet utanför KTH.

Min kära familj:

Mormor och morfar, för allt ni gett mig i livet och för att det alltid är lika trivsamt att få komma till er, i Uppsala och på Rindö.

Mina hjältar till bröder: Carl Henrik och Gustaf, för att ni alltid ställer upp, med allt från datakonsultationer till pizza och mycket mer.

Ulla och Carl-Gustaf, mina föräldrar, för att ni alltid funnits där, för allt stöd och uppmuntran ni skänkt genom åren. Ni är fantastiska!
References


