Transcript profiling of small tissue samples using microarray technology

Maria Sievertzon

KTH Biotechnology

Royal Institute of Technology
Department of Biotechnology
Stockholm 2005

© Maria Sievertzon

Department of Biotechnology
Royal Insitute of Technology
AlbaNova University Center
SE-106 91 Stockholm
Sweden

Printed at Universitetsservice US AB
Box 700 14
SE-100 44 Stockholm
Sweden

ISBN 91-7283-989-9

Maria Sievertzon (2005). Transcript profiling in small tissue samples using microarray technology. Department of Biotechnology, AlbaNova University Center, Royal Institute of Technology Stockholm, Sweden
ISBN 91-7283-989-9

Abstract

Through a number of biological, technological and computational achievements during the 20th century and the devoted work of hundreds of researchers the sequence of the human and other genomes are now available in public databases. The current challenge is to begin to understand the information encoded by the DNA sequence, to elucidate the functions of the proteins and RNA molecules encoded by the genes as well as how they are regulated. For this purpose new technologies within the area of functional genomics are being developed. Among those are powerful tools for gene expression analysis, such as microarrays, providing means to investigate when and where certain genes are used.

This thesis describes a method that was developed to enable gene expression analysis, on the transcriptome level, in small tissue samples. It relies on PCR amplification of the 3'-ends of cDNA (denoted 3'-end signature tags). PCR is a powerful technology for amplification of nucleic acids, but has not been used much for transcript profiling since it is generally considered to introduce biases, distorting the original relative transcript levels. The described method addresses this issue by generating uniformly sized representatives of the transcripts/cDNAs prior to amplification. This is achieved through sonication which, unlike restriction enzymes, does not require a specific recognition sequence and fragments each transcript randomly. The method was evaluated using cDNA microarrays, Affymetrix™ oligonucleotide arrays and real-time quantitative PCR. It was shown to perform well, yielding transcript profiles that correlate well to the original, unamplified material, as well as being highly reproducible.

The developed method was applied to stem cell biology. The variability in gene expression between different populations of cultured neural stem cells (neurospheres) was investigated. It was shown that neurospheres isolated from different animals or passaged to different degrees show large fluctuations in gene expression, while neurospheres isolated and cultured under identical conditions are more similar and suitable for gene expression analysis. A second study showed that withdrawing epidermal growth factor (EGF) from the culture medium when treating the cells with an agent of interest has profound effects on gene expression, something which should be taken into consideration in future neurosphere studies.

© Maria Sievertzon

Keywords: Gene expression analysis, transcriptomics, microarray analysis, 3'-tag signature amplification, neural stem cells

Till er som jag älskar
och som fortfarande ger mig
glädje och inspiration

Pappa, Magnus
Mormor och Morfar
Åsa

List of publications

This thesis is based on the following papers, which are referred to in the text by the corresponding Roman numerals:

I Hertzberg, M., **Sievertzon, M.,** Aspeborg, H., Nilsson, P., Sandberg, G. and Lundeberg, J. (2001). cDNA microarray analysis of small tissue samples using a cDNA tag target amplification protocol. *The plant journal* 25(5):585-591

II **Sievertzon, M.,** Agaton, A., Nilsson, P. and Lundeberg J. (2004). Amplification of mRNA populations by a cDNA tag strategy. *BioTechniques* 36(2):253-259

III **Sievertzon, M.,** Wirta, V., Mercer, A., Meletis, K., Erlandsson, R., Wikström, L., Frisén, J. and Lundeberg, J. Transcriptome analysis in primary neural stem cells using a tag cDNA amplification method. Submitted.

IV **Sievertzon, M.,** Wirta, V., Mercer, A., Frisén, J. and Lundeberg, J. Confounding effects of epidermal growth factor (EGF) in the study of pituitary adenylate cyclase-activating polypeptide (PACAP) activation of primary neural stem cell proliferation. Submitted.

Table of contents

INTRODUCTION

PRESENT INVESTIGATION

INTRODUCTION

1. **From genome to function**

Deoxyribonucleic acid (DNA), the hereditary material within our cells (Avery, MacLeod et al. 1944), is fascinating material. The majority of human cells contain approximately two metres (!) of DNA, divided and tightly packed into 46 chromosomes. The molecule consists of four slightly different building blocks (denoted A, C, G and T), joined together end-to-end in a long stretch, a DNA sequence. This surprisingly simple DNA structure contains all the necessary information to create and sustain life. Using X-ray diffraction patterns obtained by Rosalind Franklin, James Watson and Francis Crick laid the foundations for understanding how this could be done in 1953, when they described the double helical structure of DNA, which showed how the information can be copied and transferred to later generations (Watson and Crick 1953; Watson and Crick 1953). In 1958 Francis Crick also postulated the central dogma. This was a milestone for molecular biology since it describes how DNA can be copied, and how specific portions of DNA sequence (genes) are translated into the active players of the cells, the proteins, via messenger RNA (also called transcripts) (Crick 1958). The entire DNA, RNA and protein contents of a cell or organism are called the genome, transcriptome and proteome respectively.

During the following decades the research field of molecular biology exploded, greatly facilitated by the discovery and development of several new biotechnical tools, including restriction enzymes (Smith and Wilcox 1970), DNA sequencing techniques (Maxam and Gilbert 1977) and the polymerase chain reaction (PCR) (Saiki, Scharf et al. 1985; Mullis and Faloona 1987). The technological advances eventually led to proposals to sequence the entire human genome. The Human Genome Project (HGP) was launched in 1990, with the goal of completely sequencing the human genome within 15 years. The genomes of other model organisms were also sequenced in parallel, to facilitate comparative genomics as well as research concerning the respective organisms, and to help the further development of high-throughput sequencing and sequence analysis. In 2001 a first draft version of the human genome sequence was released (Lander, Linton et al. 2001) (Venter, Adams et al. 2001), and in April 2003, two years earlier than anticipated, HGP spokespersons announced that the sequence was finished (covering about 99% of the genome). The sequence is now available in databases at the National Center for Biotechnology Information, NCBI (http://www.ncbi.nlm.nih.gov/genome/seq/), the University of California at Santa Cruz (http://genome.ucsc.edu/), and Ensembl (http://www.ensembl.org/). The data revealed that less than 2% of the 3 billion (10^9) basepair long sequence consists of coding sequence (exons), while ~24% consists of introns (non-coding intragenic sequence) and ~75% is intergenic DNA. The function of the intergenic sequence, of which repetitive sequences account for roughly 50%, is poorly understood. The coding DNA is predicted to contain less than 30 000 protein-coding genes (2004) (Pennisi 2003). This is a surprisingly low number, considering that earlier estimates ranged up to 120 000 genes (Liang, Holt et al. 2000). However, the complexity of the cellular components is increased by the processes of alternative splicing of RNA (see below) and post-translational modification of proteins.

Following the completion of the human and other genome sequences, we are now entering what is often called the "post-genomic era". The entire blueprint of the basis of life is now available, and the remaining task is to deduce the biological function of its constituents; the genes and their products. The research community is thus turning its attention towards the area of functional genomics. Using the genome sequence and a collection of high-throughput technologies the characteristics of many hundreds to thousands of genes and proteins are being studied in parallel. Functional genomics covers diverse matters, such as the times, tissues and levels at which specific genes are expressed (RNA and protein expression profiling), the cells and sub-cellular locations in which proteins exert their effects (localisation studies), their interactions with other molecules (protein-protein and protein-DNA interactions), protein structure and phenotypic characterisation after gene silencing. Tremendous amounts of demanding but exciting work lie ahead, before we elucidate the functions of all our genes and understand their role in human physiology and disease. The words of Winston Churchill, spoken in 1942 after three years of war, capture well the start of the post-genomic era:

> *"Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning."*

This thesis describes technologies used for the branch of functional genomics called transcriptomics; the study of all, or almost all, transcripts in an organism or cell. In particular it describes how differentially expressed genes can be identified in small tissue samples, such as micro-dissected tissue, or cells that can only be isolated or cultured in limited amounts. The other main branch of functional genomics, proteomics, could bring researchers closer to understanding events and processes in specific cells. The proteins are the ultimate actors of the cell, being the catalysts of most cellular activities and participating as building blocks in cellular structures (although some RNAs also have very important catalytic and other functions). The proteins are, however, complex structures with a wide range of different characteristics, making the parallel study of large numbers of proteins difficult. Although high-throughput technologies for the study of proteins are currently being developed and have started to be more widely used, they are generally still more costly than transcriptome analyses, and cannot investigate as many gene products in parallel. Consequently much genome-wide research has been directed towards the more easily studied transcriptome, which can be used as an indicator of the proteomic state of the cell. Transcriptome and proteome technologies will undoubtedly complement each other in efforts to deduce the function of all our genes.

2. **The eukaryotic transcriptome**

The human genome is estimated to contain approximately 20–25 000 genes, all of which are present in every cell. However, at any given time and condition, each cell only expresses (transcribes and translates) approximately 10 000–15 000 of those genes (Yamamoto, Wakatsuki et al. 2001) (Jongeneel, Iseli et al. 2003). Most of the expressed genes are needed to carry out the basic functions of cells in general, but a small proportion of them contribute to the unique characteristics of each different type of cell. Each gene is expressed at different levels and it has been estimated that a total of 200 000–300 000 transcripts are present in a single cell at any given time (Bishop, Morton et al. 1974). A few "abundant" genes are expressed at very high levels and their transcripts constitute around 20% of the total coding transcripts. A few hundred genes are expressed at intermediate levels, and the vast majority of genes are "rare", expressed at levels of one or a few copies per cell. Through a highly regulated system the cell can increase or decrease the expression of certain genes in response to external and/or internal stimuli. Sometimes this regulatory system can be distorted, leading to deviant responses and, perhaps, eventually disease. The transcriptome contains mainly non-coding RNA, such as ribosomal RNA (rRNA), transfer RNA (tRNA), small nucleolar RNAs (snRNA) and certain other, recently discovered, small RNA species that have been implicated to have important regulatory functions (e.g microRNA and small interfering RNA) (Eddy 2001). Only the transcripts encoding proteins, the mRNAs, will be further discussed.

Transcription of DNA into mRNA is initiated by the binding of transcription factors and RNA polymerase to specific sequences in the promoter region of a gene. The RNA polymerase then moves along the gene sequence and "reads" information in the DNA to synthesise a complementary RNA copy, by successively incorporating new nucleotides in the 5' to 3' direction. At the end of the gene the polymerase encounters a stop signal sequence, forcing it to terminate transcription, and is released from the DNA. During this process the mRNA is processed in several ways. At the 5' end a 5' cap structure is formed by the addition of a 7-methylguanosine, bound by an unusual 5',5'-triphosphate linkage, and specific methylation of the first nucleotide residues. The cap seems to protect the RNA from degradation and is also needed for binding to the ribosome for translation initiation. At the 3' end the transcript is cleaved at specific polyadenylation sites and then immediately polyadenylated by the addition of 20–250 adenosine monophosphates (AMPs). Most genes have multiple polyadenylation sites, utilised to differing extents (Jongeneel, Iseli et al. 2003). Both the 5' cap and the poly(A) tail are decorated by specific proteins, which probably help protect the RNA from degradation.

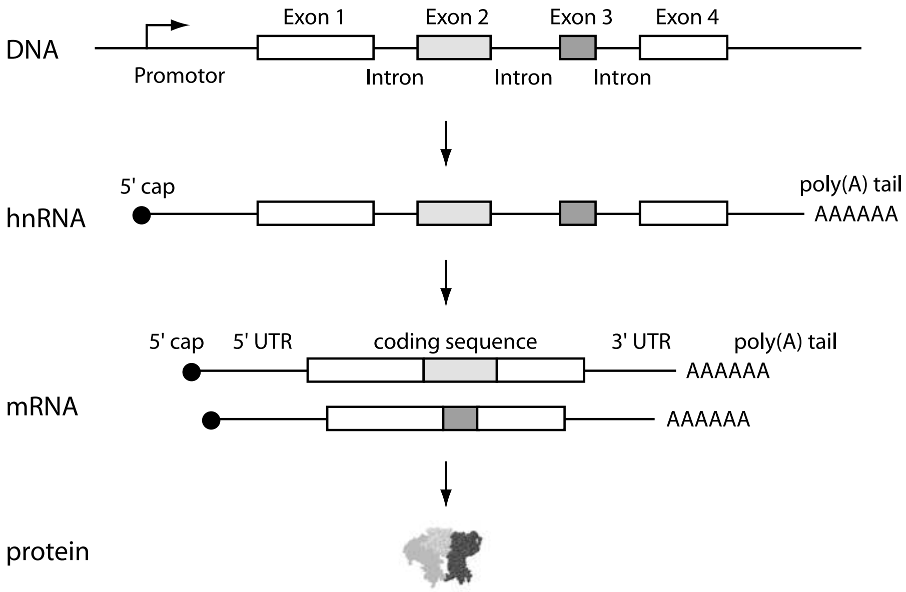

Figure 1. Processing of a protein coding gene. DNA is transcribed into heteronuclear RNA (hnRNA), which is processed into mature messenger RNA (mRNA) through 5' capping, polyadenylation and splicing. Two alternatively spliced mRNAs from the same gene are shown. The mature mRNA is transferred from the nucleus to the cytoplasm, where it is translated into protein.
UTR = untranslated region

While still in the nucleus, the transcript is further modified by a process called splicing. In eukaryotes the genes are not usually coded in one long uninterrupted sequence, but rather in short stretches (100–1000 bp) of coding DNA (exons) interrupted by stretches (50–20 000bp) of non-coding DNA (introns). This exon-intron structure is copied into the transcript via the RNA polymerase. In the splicing process a RNA-protein complex, called the spliceosome, connects the ends of two exons at certain splice sites, and catalyses the formation of exon-exon bonds and the removal of introns (for an overview of the details of this process see (Maniatis and Tasic 2002)). By splicing at different points ("alternative splicing") the splicing machinery can connect different combinations of exons. A transcript can thus exist as several different splice variants, encoding proteins that may have functional differences ranging from subtle to profound. Alternative splicing is accordingly believed to have a very important role in expanding protein diversity. It has been estimated that between 35 and 59% of the human genes are alternatively spliced, with an average of approximately three distinct transcripts per gene (Mironov, Fickett et al. 1999) (Modrek, Resch et al. 2001) (Lander, Linton et al. 2001). Many of these splicing events only occur in specific tissues, at specific developmental stages and under certain conditions, probably reflecting their functional relevance. Indeed, alternative splicing is known to have profound effects on normal cellular fate as well as disease states in many cases (Graveley 2001) (Nissim-Rafinia and Kerem 2002). Alternative splicing is regulated by proteins (hnRNPs and SR proteins) that bind to certain splicing enhancer and silencing sequences within the pre-mRNA, thereby stimulating or repressing the splicing of the corresponding exons and introns (Maniatis and Tasic 2002). The cell regulates the splicing by altering the composition of the splice regulation proteins. This is done by enhancing their synthesis, relocating them outside the nucleus or to nucleolar bodies, or by phosphorylation. The mature mRNA consists of a 5' cap, a 5' untranslated region (UTR; with a mean length of about 300 basepairs), a coding sequence (average length, 1340 bp), a 3' UTR (average length, 770 bp) and a poly(A) tail (Lander, Linton et al. 2001). It is then transported to the cytoplasm, where it serves as a template for protein production until it is degraded.

Tight control of the transcript concentration of each gene is necessary to maintain appropriate cellular functions. Transcript levels are balanced via the regulation of transcription initiation and the rate of degradation. Transcription initiation is regulated by proteins that bind to the promoter sequence upstream of the coding sequence, as well as to enhancer sequences that can be located at different distances, both upstream and downstream, of the coding sequences (Levine and Tjian 2003). The binding of these proteins to the DNA affects the subsequent binding of the RNA polymerase and other proteins needed for transcription. Transcription is also regulated by the structure of the underlying chromatin (a protein structure that efficiently packages DNA in the nucleus) (Ehrenhofer-Murray 2004) (Kadam and Emerson 2002). This process involves chromatin remodelling complexes, which move the nucleosomes or alter their structure and thus increases the accessibility of the DNA, and histone-modifying enzymes, which modify the histones and other proteins by acetylation, deacetylation, phosphorylation, ubiquitination and methylation, thereby altering their structure and affinity for other chromatin proteins. The chromatin structure is also involved in further gene regulation processes, such as imprinting and gene silencing.

Degradation of mRNA in both yeast and higher eukaryotes, is initiated by the removal of the 3' poly(A) tail by deadenylase enzymes (Wilusz, Wormington et al. 2001). In yeast, once the poly(A) shortening is complete the 5' cap is removed by a decapping enzyme, and then the rest of the transcript is degraded by 5' and 3' exonucleases. The decay mechanism is not as well understood, but probably similar, in mammals. It is a highly controlled process, regulated by several sequence elements that either promote (destabiliser elements) or inhibit (stabiliser elements) decay, in concert with transcript binding proteins. The A+U-rich elements (AREs) comprise one class of such elements. AREs can have slightly different sequences, are found in the 3' untranslated region (3'-UTR) and are mainly destabilising. A number of different ARE-binding proteins can interact with these sequences and influence transcript stability, translation and the subcellular localisation of the mRNA (Wilusz, Wormington et al. 2001) (Bevilacqua, Ceriani et al. 2003). A special mRNA decay pathway is the nonsense-mediated decay (NMD), which ensures that mRNA with premature stop codons (resulting from processes such as mutations or mis-splicing) are eliminated (Schell, Kulozik et al. 2002). Non-coding RNAs (from introninc and intergenic regions, as well as pseudogenes and anti-sense sequences to coding genes) have also gained increasing attention as having potentially important regulatory functions in a number of different cellular processes, e.g. transcript decay (Mattick 2004). For example, microRNA (miRNA) and small interfering RNAs (siRNAs) have been shown to be post-transcriptional regulators (He and Hannon 2004). Pri-miRNAs are encoded in the genome and processed into miRNA, 21–25 nucleotides long, that bind with near but imperfect complementarity to multiple sites within the 3'-UTR of certain transcripts. This creates secondary structures within the 3'-UTR that can repress subsequent translation. Similarly, siRNAs, also 21–25 nucleotides long, bind with perfect complementarity to their targets, thereby inducing their cleavage and degradation.

The following sections describe methods that are available for investigating the protein coding parts of the transcriptome (the mRNA). One should keep in mind that transcript profiles do not always correlate well to their corresponding proteomes, although the correlations are better for induced changes (Griffin, Gygi et al. 2002) than for absolute amounts (Gygi, Rochon et al. 1999) (Futcher, Latter et al. 1999). Nevertheless, they provide invaluable clues concerning the identity and function of genes involved in normal and disease processes.

3. **Methods for transcript profiling**

The transcriptome reflects the state and intrinsic properties of a cell. Although most expressed genes are involved in the basic functions of the cell, such as metabolism and maintenance of cellular structures, small proportions of the genes carry out the specific functions of a particular cell type. In addition, genes are turned on and off in response to external stimuli, at different stages of cell differentiation and as a result of disease. These specifically expressed genes reflect important differences between cells and can give insight into the molecules and functional processes that are involved in certain cellular functions. In addition, the identification of these transcribed sequences is an important tool for finding and annotating genes and coding sequences within the genome (Marra, Hillier et al. 1998), a task that can otherwise be very tedious, especially for mammalian genomes, where the coding sequences account for only a small percentage of the total DNA. Therefore, a collection of methods have been developed to identify these differentially expressed genes and measure the level of their respective transcripts in different cells, generating so-called expression or transcript profiles. Some of the methods are “global”, indicating that they measure all, or almost all, transcripts within the transcriptome. In contrast “selective” methods aim at identifying only those genes which have altered or differential expression.

3.1 Global approaches for transcript profiling

Protein coding transcripts, in the form of mRNA, can easily be isolated from cells, tissues or extracted RNA by capturing their poly(A) tails using complementary poly(T) oligonucleotides attached to a solid support. The RNA molecule, however, is a very unstable molecule that is difficult to handle because of its rapid degradation. The mRNA population is therefore often reverse transcribed into complementary DNA (cDNA): single or double stranded DNA molecules with complementary sequences to the mRNA. Reverse transcription of a whole population of mRNAs generates a cDNA pool, a collection of cDNAs with (theoretically) the same distribution as the original mRNA. The content of the cDNA pool can subsequently be exploited by cloning (to separate the individual cDNAs from each other) and sequencing the obtained cDNA library (see the EST sequencing, SAGE and MPSS sections, below), or by hybridising it to previously identified sequences on a microarray.

3.1.1 EST sequencing

The first high-throughput sequencing of a cDNA library was reported in 1991 by Adams and co-workers (Adams, Kelley et al. 1991), and the term expressed sequence tag (EST) was coined. EST sequencing is performed by single-pass sequencing of a cDNA cloned into a vector, yielding a 200–700 bp sequence. The sequencing can be done either randomly (along the whole cDNA sequence) (Adams, Kelley et al. 1991) (Adams, Dubnick et al. 1992) or, more commonly, from the 3' or 5' ends of a directionally cloned cDNA library (Matsubara and Okubo 1993). An EST is thus a partial sequence of a transcribed gene, which can be used as a "tag" for identifying that particular transcript. The advantage of the random sequencing strategy is that the obtained sequences often represent the coding sequences of the transcripts. Through homology searches of various nucleotide and protein databases information on the transcript can thus be obtained, which can give clues about its protein counterpart's function. Since the 5' parts of the cDNAs are often truncated due to incomplete reverse transcription 5' ESTs also often give information on the coding sequence of the transcript. 3' ESTs, however, often span the 3' untranslated region (UTR). This part of the transcript is less evolutionarily conserved than the coding sequence and therefore is more unique to its transcript, making it the most suitable part for transcript identification. Furthermore, 3' sequencing will yield sequences from a more defined position of the transcript, facilitating tag counting. Combining 5' and 3' EST sequencing of a cDNA clone obviously has the benefit of yielding both coding sequence information and a unique identifier for the gene (Williamson 1999).

The initial objective of EST sequencing projects was to speed up the human genome project (HGP) and to initiate processes of gene identification, mapping genes onto genome sequences (Wilcox, Khan et al. 1991) (Hudson, Stein et al. 1995) and identifying exon-intron borders and possible alternative splicing events (Adams, Kelley et al. 1991) (Marra, Hillier et al. 1998). It was also soon found that EST sequencing could be used for transcript profiling (Okubo, Hori et al. 1992) (Adams, Kerlavage et al. 1993) (Matsubara and Okubo 1993). The cDNA clones in a cDNA library theoretically have the same distribution as the original mRNA population and thus EST sequencing of a large number of clones will give a statistical overview of the transcripts expressed in the cell or tissue type. Often 3' ESTs are generated for this purpose because of their transcript uniqueness and defined position (see above). The transcript profile obtained from one cDNA library can then be compared to the profile obtained from another library, facilitating identification of differentially expressed genes. The outcome of this type of transcript profiling largely depends on the number of ESTs sequenced (Audic and Claverie 1997). Even when thousands of clones are sequenced, the sample numbers will generally be too low for robust statistical analysis of the abundance of rare transcripts. In a SAGE study (see below) by Zhang and co-workers it was discovered that 86% of all transcripts were present at less than five copies per cell, and even when sequencing 300 000 tags there was an 8% probability of not detecting transcripts present at a level of three copies per cell (Zhang, Zhou et al. 1997).

Thus, the advantages of EST sequencing as a transcript profiling method are that novel genes can be discovered, the tags are long enough for certain identification and it gives a very good representation of the transcripts present in a cell or tissue. The clones obtained from EST projects can also be used for further transcript profiling using microarray techniques (see below). The main drawback of EST analysis is the very large scale sequencing effort required, making it a very laborious, time consuming and costly method.

Several large EST sequencing projects have been initiated, including the "Bodymap" project in Osaka, Japan (Okubo and Matsubara 1997), and the Merck Gene Index project (MGIP), a collaboration between Merck, the IMAGE (Integrated Molecular Analysis of Genomes and their Expression) consortium and Washington University (Williamson 1999). Sequences from these and other projects are deposited in public databases such as dbEST (http://www.ncbi.nlm.nih.gov/dbEST/index.html) and Bodymap (http://bodymap.ims.u-tokyo.ac.jp/). dbEST is the largest, currently containing over six million human ESTs, four million mouse ESTs and several million ESTs from other organisms. The sequences are also clustered according to their sequences and partitioned into a non-redundant set of gene-oriented clusters, available in the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene). UniGene today contains a little more than 50 000 human clusters, giving a rough estimate of the number of human genes, although some clusters probably represent the same, but alternatively spliced, gene products.

3.1.2 Serial Analysis of Gene Expression (SAGE)

Another tag sequencing strategy for transcript profiling, serial analysis of gene expression or SAGE, was described in 1995 (Velculescu, Zhang et al. 1995). Similarly to EST sequencing, SAGE involves counting cDNA tags to yield a statistical representation of the transcripts within a cDNA library. The tags are, however, much shorter and each clone contains up to 50 tags, facilitating a much higher sequencing throughput and thereby reducing both time and cost. This allows more extensive transcript profiling than EST sequencing as hundreds of thousands of transcripts can be measured within each library (Zhang, Zhou et al. 1997).

In SAGE the cDNA is first cleaved with an "anchoring enzyme" (AE), a restriction enzyme that recognises a 4 bp sequence. The 3' ends (biotinylated through the polyT primer) are bound to streptavidin beads and linkers are ligated to the anchoring enzyme overhang. Within the linkers there is a recognition site for a second enzyme, the "tagging enzyme" (TE), which is a type IIS restriction enzyme that cleaves the DNA at a certain distance downstream of its recognition site. The linkers also contain sequences that can be used in subsequent PCR. Cleavage with the tagging enzyme yields short tags, ~14 bp in the original approach (Velculescu, Zhang et al. 1995), which can be ligated to each other, forming ditags.

The ditags are amplified by PCR, using linker specific primers, and then concatemerised into long, continuous stretches of DNA which are cloned into a vector and then sequenced. Each tag thus contains a 9–17 bp long tag-specific sequence (depending on the restriction enzymes used), plus a 4 bp recognition site for a specific AE, and is located at a defined position within the 3' end of the original transcript. The sequences of such defined SAGE tags for different transcripts are contained within SAGE reference databases (http://www.ncbi.nlm.nih.gov/SAGE/), facilitating identification of the obtained tag sequences.

The greatest advantage with the SAGE methodology is the high throughput it allows. However, the method has also been associated with several drawbacks and limitations, some of which have been solved over the years (Yamamoto, Wakatsuki et al. 2001). The main concern is the short length of the tags, 9–10 bp in the original approach, that are tag specific. In theory 4^9 (262 144) different transcripts can be distinguished by sequencing 9 bp sequences, provided the nucleotide distribution is random throughout the genome. Given that the estimated number of human genes is 20 000–30 000, 9 bp should be plenty for discriminating of all of the genes. However, the genome sequence is not random, containing conserved regions for instance, such as those shared by common domains and gene families. Thus, multiple genes can share the same tag, and there have also been instances in which a single gene has multiple tags, due to alternative splicing and alternative polyadenylation sites. An extreme case was observed by Ishii *et al*, in which one sequenced tag corresponded to 22 different UniGene clusters (Ishii, Hashimoto et al. 2000). In addition, these short tags make SAGE very sensitive to sequencing errors and single nucleotide polymorphisms (SNPs) (Silva, De Souza et al. 2004). Substituting even a single base will give the tag a new identity, associated with another, known or presumably novel, transcript. Consequently, many SAGE projects have excluded rare, unmappable tags, and thus maybe real but rare novel transcripts. Several groups have tried to overcome these problems by developing SAGE methods with longer tags, using alternative restriction enzymes, e.g. a method developed by Ryo and co-workers that generates 14 bp tags (Ryo, Kondoh et al. 2000) and the LongSAGE method developed by Saha and co-workers that generates 17 bp tags (Saha, Sparks et al. 2002).

Another concern has been that some transcripts may lack the recognition site for the anchoring enzyme used, leading to their absence in the tag library, a problem that could be overcome by generating libraries using different restriction enzymes (Yamamoto, Wakatsuki et al. 2001) (Unneberg, Wennborg et al. 2003). The SAGE methodology has also been subject to technical limitations. For example, large amounts of starting RNA are required in the standard protocol (2.5–5 µg of mRNA). This has led to development of methods such as MicroSAGE (requiring <10 cells) (Datson, van der Perk-de Jong et al. 1999), SAGE-lite (100 ng total RNA) (Peters, Kassam et al. 1999), SADE (50 000 cells) (Virlon, Cheval et al. 1999) and miniSAGE (1 µg total RNA, without PCR amplification) (Ye, Zhang et al. 2000). Other problems have included difficulties in separating the tags within a ditag and linker dimerisation during PCR. Nevertheless, despite these drawbacks, SAGE is a very powerful technique for transcript profiling and has provided a wealth of valuable expression data in a number of different studies (for a comprehensive review see (Yamamoto, Wakatsuki et al.

2001). Tens of millions of SAGE tags from various organisms are currently deposited in the public SAGEmap database (http://www.ncbi.nlm.nih.gov/SAGE/) and the gene expression omnibus database (GEO) (http://www.ncbi.nlm.nih.gov/projects/geo/).

Apart from gene expression studies SAGE has also been used for the discovery of novel genes and physical mapping of the human genome. Because of its power to measure very rare transcripts it has the potential to detect genes or exons that no other method (such as EST sequencing) is able to detect. Indeed, several SAGE gene mapping projects predict the number of expressed sequences in the human genome to be ten-fold higher than predictions based on EST comparisons and data acquired in the HGP and Celera human genome projects (Boheler and Stern 2003) (Chen, Sun et al. 2002) (Saha, Sparks et al. 2002). The novel transcripts found by these means may originate from alternatively spliced transcripts, non-coding transcripts that may have regulatory functions, or novel genes (Chen, Sun et al. 2002). The confirmation of such novel transcripts, found through SAGE tags, has been facilitated by the development of a technique generating longer 3' cDNA fragments from these tags (GLGI) (Chen, Rowley et al. 2000). However, due to its sensitivity to sequencing errors the SAGE approach to novel gene discovery is still controversial.

3.1.3 Massively Parallel Signature Sequencing (MPSS)

A method based on a similar principle to SAGE is massively parallel signature sequencing, MPSS, first described by Brenner *et al* in 2000 (Brenner, Johnson et al. 2000) (Reinartz, Bruyns et al. 2002). Like SAGE, MPSS involves counting 16–20 bp signature tags to establish a transcription profile for a cell or tissue type. The tags are sequenced using microbead arrays instead of cloning them into concatemers and sequencing them by standard sequencers. Each tag is bound to a microbead which has a fixed position in a flow cell (array). By stepwise cleavage of the tags and addition of fluorescently labelled adaptors with known sequences the sequences of the tags are deduced. The great advantage of MPSS compared to SAGE is the number of tags that can be sequenced. Each microbead array can hold millions of beads, facilitating the sequencing of hundreds of thousands to millions of tags from each sample. The greatest disadvantage of the method is the shortness of the tags, as in SAGE. Thus, as with SAGE, it can be difficult to distinguish between the transcripts of similar genes when using MPSS. However, due to the vast amount of tags sequenced, the data obtained are statistically much more robust, making the measurement of rare transcripts more reliable and the method less sensitive to sequencing errors. Also, the method requires that all transcripts contain a certain restriction enzyme recognition site, most commonly for *Dpn*II, and it requires rather large amounts of input RNA (100 µg total RNA).

Using MPSS Jongeneel *et al* generated 10 million (!) signature tags from each of two different cell lines (Jongeneel, Iseli et al. 2003). This is a truly redundant coverage of the estimated 200 000–300 000 transcripts believed to be present in a cell (Bishop, Morton et al. 1974). From their data it was estimated that each cell line expresses between 10 000 and 15 000 genes. The majority of tags mapped to known transcripts (65%). A smaller number (6.6%) mapped to introns or close to exons of known genes, suggesting that more sequences than are known today are indeed expressed. A substantial proportion (20%) of the tags, all unmappable, was believed to have been generated from sequencing errors or polymorphisms. A number of tags occurred at multiple locations in the genome, making them impossible to discriminate and identify. The study also found the utilisation of alternative polyadenylation sites to be quite extensive (on average 1.32 sites/gene), and perhaps also tissue-specific. Although a rather costly method, MPSS has also been used in studies to investigate the expression profile of *Arabidopsis thaliana* (Hoth, Ikeda et al. 2003) (Meyers, Tej et al. 2004) (Meyers, Vu et al. 2004), the expression profile of embryonic stem cells (Brandenberger, Khrebtukova et al. 2004) and non-coding micro RNA in *A. thaliana* (Wang, Reyes et al. 2004).

3.1.4 Microarray technology

Among the most powerful, versatile and widely used methods for gene expression analysis is high-density DNA microarray analysis. The origin of this approach is somewhat controversial (Ekins and Chu 1999) (Weeraratna, Nagel et al. 2004), but it was conceptually and practically developed during the late 1980s and early 1990s (Ekins, Chu et al. 1989) (Ekins 1989) (Augenlicht, Wahrman et al. 1987) (Lennon and Lehrach 1991) (Southern, Case-Green et al. 1994) (Zhao, Hashida et al. 1995). Array techniques for gene expression analysis are based on the immobilisation of hundreds to tens of thousands of distinct DNA sequences (probes) on a solid support, generating a two-dimensional array of spots/features, where each feature represents a certain gene, exon or splice-variant. When a labelled RNA or cDNA sample (target) is applied to the DNA array the probes within a feature will capture the target molecules through sequence complementation. The strength of the label signal (often fluorescent) from the captured targets reflects the abundance of that target within the hybridised sample. A great many platforms and versions of the microarray technology, with different characteristics and applications, have been developed. This thesis will describe the two most commonly used platforms for gene expression analysis, oligonucleotide microarrays/GeneChips® and cDNA microarrays, and briefly mention some recent advances and new applications within the field.

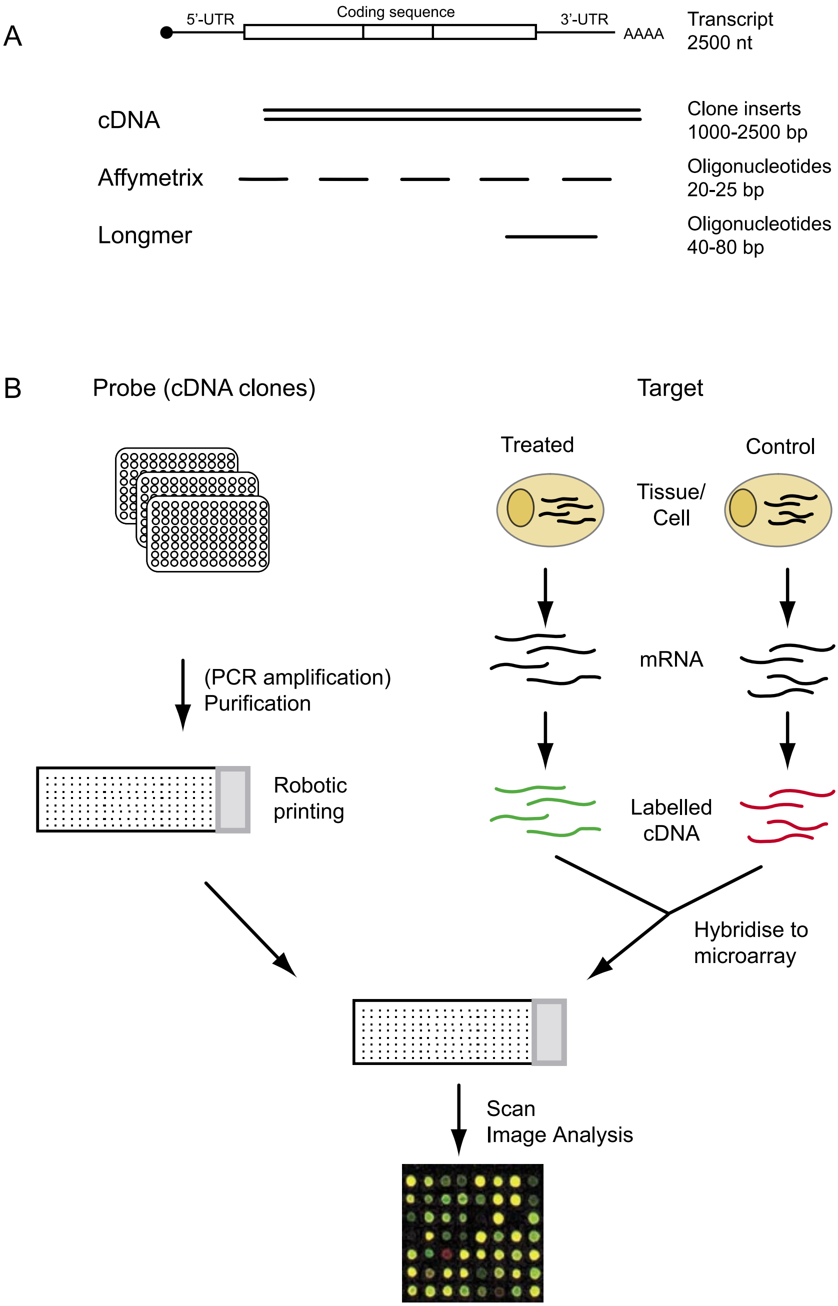

Figure 2. Microarray technology; A) Different microarrray platforms, B) Principles of cDNA microarray technology.

3.1.4.1 Oligonucleotide microarrays / GeneChip® technology

Array manufacturing

High-density DNA microarrays were first manufactured on a large, commercial scale by the biotechnology company Affymetrix (Lockhart, Dong et al. 1996) (Lipshutz, Fodor et al. 1999), whose GeneChip® arrays are made by chemically synthesizing oligonucleotides directly on a solid surface, using a photolithographic method similar to techniques used in the production of computer chips (Fodor, Read et al. 1991) (Pease, Solas et al. 1994). Briefly, synthetic linkers modified with photosensitive protecting groups are attached to a glass surface. Using a photolitographic mask, light is then directed to specific areas on the surface to remove the protection groups from the exposed linkers. Bi-functional deoxynucleosides (adenosine, cytidine, guanosine or thymidine) are added one at a time to the surface, resulting in chemical coupling to the de-protected sites. Another mask is used to direct light to and de-protect other sites, new deoxynucleosides are added and the process is repeated until the desired length of oligonucleotide is synthesised (4 x N cycles of photoactivation and nucleoside addition are required for an N nucleotide long sequence). A 1.28 x 1.28 cm array can include over a million different oligonucleotide sequences. For gene expression purposes the oligonucleotides are generally 25 bases long and each transcript is represented by 11–20 such probes. The probe sequences are ideally spread throughout the gene sequence, generally being more concentrated at the 3'-end. The probe sequences used are selected on the basis of gene and EST data from public databases according to a number of criteria; most importantly that they should be unique for the gene (avoiding, for example, characteristic sequences of gene families) and relatively uniform in their hybridisation properties. In addition, each perfect match (PM) probe is paired with a mismatch (MM) probe, an identical probe except for a single base difference in a central position. The MM probes act as specificity controls and allow for subtraction of background and cross-hybridisation. The use of multiple independent probes for each gene greatly improves signal-to-noise ratios, improves the accuracy of RNA quantitation (averaging and outlier rejection), increases the dynamic range and reduces the rate of false positives and miscalls (Lipshutz, Fodor et al. 1999).

Target preparation, labelling and hybridisation

Samples to be hybridised to GeneChip® microarrays are prepared following standard protocols. Briefly, double-stranded cDNA is synthesised using a oligo(dT) primer with a T7 promoter sequence. The cDNA is then transcribed *in vitro* in the presence of biotinylated ribonucleotide analogs to generate tagged, complementary, amplified RNA, aRNA. Before hybridisation the cRNA is fragmented, to reduce the formation of secondary structures and yield products that anneal to the short probes on the array surface. After hybridisation and washing of the array the target is labelled using an antibody amplification staining procedure involving streptavidin (binding biotin) and phycoerythrin as a fluorescent reporter. Confocal laser scanning is then used to read the fluorescent signals, generating a digital image of probe-specific fluorescent intensities.

Data analysis

Many aspects of data analysis apply to all microarray platforms and will be discussed under cDNA microarray technology. Here, specific aspects of GeneChip® data analysis will be briefly discussed. As with all experiments, a carefully planned experimental design is essential. An important aspect of experimental design when planning a microarray experiment is the design of the array itself. Questions that should be considered may include the following. Are the genes of interest present on the array? Do I want to study changes in the expression of a specific set of genes associated with the biological process under investigation, e.g. apoptosis, or do I want a more global view of the changes in gene expression? What negative and positive controls should I include? Since the manufacture of GeneChips® requires individual photo-litographic masks for each array, the flexibility in the design of the array is limited, although custom designed arrays can now be purchased from Affymetrix (www.affymetrix.com).

Once the fluorescent signal intensities for each probe have been obtained, algorithms are needed to generate a qualitative and quantitative measure of the corresponding gene expression. These measures are calculated from data from all the PM and MM probes for each gene. Affymetrix provides analysis software with empirically and statistically developed algorithms; Microarray Suite (MAS) 4.0 and 5.0. However, the use of both PM and MM data, and other aspects of this approach, have been criticised and several other academically developed methods of data processing are also available (Cope, Irizarry et al. 2004) (Gautier, Cope et al. 2004).

Advantages and disadvantages

GeneChip® arrays have several advantages over cDNA microarrays (see below). Since the probe sequences are designed, redundant sequences within the genome can be avoided, and the probes have relatively uniform hybridisation properties. It is also possible to include predicted genes, which are not present in cDNA libraries, on the array. In addition, the use of standardised protocols and a variety of controls present on the arrays (including the PM and MM probe pairs) makes the generation of data very consistent and highly reproducible. It also enables data obtained in different laboratories to be compared. GeneChips® appear to have a higher dynamic range than cDNA microarrays (Barrett and Kawasaki 2003), and since there is no need to handle large numbers of cDNA clones and PCR products there is no risk of mixing up clones or feature identities. However, since lithographic masks are required to produce the arrays they are costly, restricting both the degree of replication that can be used and the number of academic users who can afford them. The array production procedure also restricts their flexibility, since the addition or removal of probes is quite complex and costly. Another disadvantage is that only already known sequences can be represented on the arrays, making them unsuitable for discovering unknown or unpredicted genes, or for studying organisms with poorly characterised genomes.

Recently, longmer oligonucleotide (50–70 bases long) arrays, combining the advantages of controlled probe design with the higher probe specificity and ease of production associated with cDNA microarrays (see below), have been produced. These will also be described below, following the discussion of cDNA microarrays.

3.1.4.2 cDNA microarrays

Cheaper alternatives to the GeneChip® arrays are cDNA microarrays, which were developed through academic efforts at a number of different laboratories during the early 1990s, and then made generally available by the groups of Patrick Brown and David Botstein at Stanford University (Schena, Shalon et al. 1995) (Shalon, Smith et al. 1996) (DeRisi, Iyer et al. 1997). The Brown group also provided freely available directions on how to make a microarray robot and all the necessary protocols to perform a microarray experiment. Today, diverse protocols for array fabrication, sample labelling and hybridisation are available, but the scope of this thesis only allows coverage of some of their basic principles. Comprehensive reviews covering technical, data analysis and conceptual aspects are also available (Nature Genetics 1999) (Lockhart) (Schulze and Downward 2001) (Nature genetics 2002) (Weeraratna, Nagel et al. 2004).

Experimental design
To conduct a successful microarray experiment careful planning of the experiment is essential. Aspects considered should include the array design, the target samples that are to be used, how well the experiment should be replicated in order to yield the required level of statistical significance and the experimental design, in which it is decided how the samples should be labelled and co-hybridised.

The fabrication of cDNA microarrays relies on the physical existence of cDNA clones. The arrays may be focused on transcripts associated with a particular type of tissue, cell, chromosome, or function (e.g. signalling molecules, cytokines or apoptotic mediators), or may be more global, representing all or most of the transcriptome (Forster, Roy et al. 2003) (Weeraratna, Nagel et al. 2004). A focused array may be designed to yield detailed information on specific transcripts, e.g. different splice variants, whereas global arrays may be more suitable for a broader analysis. In addition, a set of control clones should be included on the array. These should preferably (if space allows) include replicate spots printed at different positions on the array, to ensure reproducibility and hybridisation quality. Negative controls can include repetitive DNA, poly(A) DNA, inter-genic DNA, empty spots and non-homologous sequences from other organisms, to measure possible cross-reactivity and non-specific fluorescence. Also, positive controls can be included, e.g. for measuring spiked RNA for normalisation purposes (see below).

It is also important to select appropriate samples for the study (Forster, Roy et al. 2003) (Weeraratna, Nagel et al. 2004). A common problem is the difficulty of obtaining a pure cell population. Most tissues consist of a mixture of cell types and are often infiltrated by different blood cells. In some cases such mixtures could be relevant, but it is often difficult to obtain representative samples and collect equivalent samples from different patients. This problem may be solved by using techniques such as laser capture microdissection or fluorescence activated cell sorting (FACS), or the researcher may consider using cultured primary cells or cell lines. If a diseased and the corresponding normal state is to be compared, the disease progression has to be considered and the relevant disease stage chosen. Similarly, if a treated vs. non-treated comparison is to be made, the time point(s) and relevant concentration(s) of any drugs/compounds used must be decided upon.

Statistical issues are also crucial considerations when planning a microarray experiment. Microarray experiments are very sensitive to many parameters, and factors such as tissue sampling, RNA preparation, labelling and hybridisation introduce variations in the data that could distort the measured levels of gene expression. In order to compensate for this, different levels of replication are necessary (Lee, Kuo et al. 2000) (Pan, Lin et al. 2002) (Churchill 2002) (Dobbin, Shih et al. 2003). Biological replicates will account for genetically and environmentally induced variations between individuals and samples. Technical replicates (repeated hybridisations with the same biological sample) average out variation introduced during RNA extraction, labelling and hybridisation. Duplicate spots of the same probe are also often included on the array to account for measurement errors caused by artefacts during printing or hybridisation. In general, biological replication is preferred over technical replication, since biological replication will allow broader conclusions to be drawn, and will also cover the technical variation introduced, unless (of course) the research is of a technical nature or if only limited numbers of samples are available. Assaying each sample multiple times will clearly be preferable to single measurements. Another option to deal with biological variation is to pool individual samples before hybridisation. The idea then is to average out differences between samples and simultaneously reduce the number of arrays needed for hybridisation, thus reducing costs. Pooling does, however, mean that valuable information on sample variability will be lost, the variance components cannot be well established and outliers (e.g. from poorly prepared samples) cannot be identified (Kendziorski, Irizarry et al. 2004) (Dobbin, Shih et al. 2003). Also, technical replication and the use of multiple slides are still necessary. In conclusion, when the number of available arrays is large pooling should be avoided, but if the number of samples exceeds the number of arrays pooling could be advantageous.

Lastly, an experimental design, describing how the samples should be labelled and co-hybridised, must be decided upon. In principle three main types of designs have been described and used: direct comparisons, reference design and loop design (Churchill 2002) (Yang and Speed 2002) (Dobbin, Shih et al. 2003) (Kerr and Churchill 2001). The most commonly used type of design is the reference design, which offers great flexibility in that multiple samples can be compared and samples can be added or withdrawn from the study without disturbing the rest of the samples. The reference could be a biologically relevant sample or a completely different sample, preferably containing as many different transcripts as possible (Dudley, Aach et al. 2002) (Sterrenburg, Turk et al. 2002). The reference design, however, introduces more variability into the data than a direct comparison, since the comparisons are made via the reference, and thus direct comparison can be more appropriate in cases where only subtle changes in gene expression are expected. Dye swap hybridisations (where the samples are reciprocally labelled in two different hybridisations) should always be included in a direct design, to avoid confounding dye effects with treatment effects. Loop designs are also an option if multiple samples are to be compared (Kerr and Churchill 2001). They are more cost efficient than reference designs since there is no need to put resources into labelling and hybridisation of reference. They are, however, sensitive to array fall outs and the data obtained can be difficult to analyse for a non-statistician. Other types of design have also been proposed (Yang and Speed 2002).

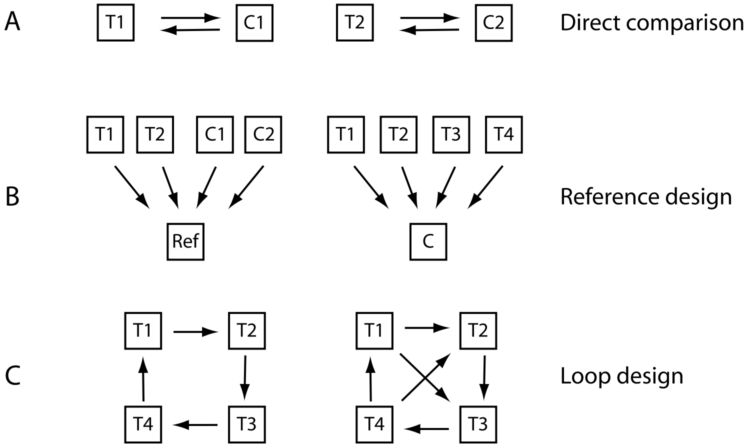

Figure 3. Some examples of experimental designs for microarray experiments. Arrowheads indicate labelling with Cy5, the tails indicate labelling with Cy3. Samples can either be compared directly, as in (A), or indirectly, as in (B, left), or a mixture of both. Dye swap experiments are indicated by two opposite arrows.
T = treatment/disease or similar, C = control/normal or similar, ref = reference sample

Array manufacture

Once the array design has been decided, the cDNA clones to be spotted can be obtained from large clone collections generated by EST or genomic sequencing efforts (e.g. the Integrated Molecular Analysis of Genomes and their Expression consortium, IMAGE), or from locally produced cDNA libraries. In order to obtain the required concentration of DNA each clone needs to be amplified. This is done using PCR, preferably with universal primers corresponding to the vector sequences of the particular cDNA library(ies). The amplified products are analysed by gel electrophoresis, purified and re-suspended in an appropriate spotting solution. Keeping track of all the clones through this process is a vital and non-trivial problem, and re-sequencing of random clones to ensure the correct identity of clones is advisable. The purified, concentrated DNA fragments are then positioned in an ordered pattern onto a solid surface using a robotic arraying device. The most commonly used surfaces for cDNA microarrays today are aminosilane or polylysine coated glass slides, although nitrocellulose or nylon membranes can also be used. Successful printing, generating small, uniform spots with high concentrations of probes, depends on the printing technicque and spotting solution used. Printing techniques are based either on contact (using printing pins) or non-contact (using piezo-electrical deposition) principles. Piezo-electrical deposition generally generates small, homogenous spots, whereas the results of contact printing depend largely on the quality of the printing pins. Successful printing also requires controlled environmental conditions, such as optimised air humidity, temperature and the absence of dirt and dust particles. The DNA within the created spots is fixed onto the array surface by randomly cross-linking the DNA backbone to the surface, using heat or ultraviolet (UV) radiation. Alternatively, 5'-aminoacylated primers can be used in the PCR, facilitating covalent cross-linking, to for example aldehyde-based coatings. In theory this approach will make the probe more readily available for binding to the target.

Target preparation, labelling and hybridisation

Because the printing procedures produce variations in spot morphology and quality between individual arrays, glass cDNA microarrays are hybridised to two differently labelled target samples simultaneously to obtain relative measurements of gene expression (in contrast to Affymetrix™ oligonucleotide microarrays, which are hybridised to one sample at a time). In most studies the fluorescent dyes Cy3 and Cy5 are used. First, total or mRNA is isolated from the respective samples and checked for purity and quality. The samples are then labelled, directly or indirectly, through the reverse transcription of poly(A) RNA into cDNA. Direct labelling is accomplished either by using a labelled poly(T) primer or (more commonly since more labels are incorporated) labelled nucleotides. However, the fluorescent groups are often bulky and therefore more efficient, indirect labelling strategies have been developed. These are based on the incorporation of chemically modified nucleotides during cDNA synthesis, followed by subsequent coupling of dyes to the modified nucleotides. The indirect labelling strategies are more laborious, but often yield stronger signals and are often less expensive than direct labelling. In a commonly used strategy, aminoallyl modified nucleotides are incorporated into the cDNA, which are then coupled to Cy3- or Cy5-esters (Randolph and Waggoner 1997) (Schroeder, Peterson et al. 2002). Other strategies to improve labelling efficiency and increase signal strength include the use of alternative reverse transcriptases (SuperScript™ and FluoroScript™ from Invitrogen and CyScribe™ from Amersham Biosciences), alternative fluorophores (Wildsmith, Archer et al. 2001) and signal amplification systems such as the dendrimer-based 3DNA™ Submicro system from Genisphere®, (Stears et al 2000, Physiol genomics 3:93–99) and the tyramide signal amplification (TSA) system from PerkinElmer® (Karsten, Van Deerlin et al. 2002). Despite these efforts to increase the fluorescent signal from the labelled target the amount of starting RNA required to perform one hybridisation is quite high, at present around 10–20 µg of total RNA. This is equivalent to millions of cells, which is not always easy, or even possible, to obtain. Therefore, several RNA and cDNA amplification strategies have been developed. These include amplification of antisense RNA, aRNA, using *in vitro* transcription from a T7 promoter and PCR based methods, as discussed later in this thesis.

Once the target samples have been labelled they are mixed, dissolved in a hybridisation buffer and applied to the cDNA microarray. Hybridisation is performed in special humidified chambers, which are either placed in a warm water bath or automated hybridisation stations for 16–20 hours. To avoid cross-hybridisation, but allow specific duplex formation between the target and probe, the hybridisation conditions are optimised in terms of temperature and buffer composition. To reduce background, pre-treatment of the slides is necessary, e.g. by blocking reactive groups with non-fluorescent biomolecules, such as bovine serum albumin (BSA). The hybridisation buffers commonly include denaturing agents such as formamide and sodium dodecyl sulphate (SDS), as well as salts and agents such as COT1-DNA and poly(A)-DNA that block repetitive sequences. After hybridisation the arrays are carefully washed in several steps with increasing stringency, to wash away any unbound DNA and dirt particles, while retaining any specifically bound targets on the array.

3.1.4.3 cDNA microarrays – data analysis

Image acquisition and processing

After hybridisation, fluorescence intensities are obtained by scanning the arrays with a confocal laser scanner. A wide variety of scanners are available on the market, with different properties, limitations and advantages. The scanner settings, and thus the obtained signal strengths, could potentially influence the quality of the downstream data analysis, and thus different strategies to obtain optimal settings have been proposed (Forster, Roy et al. 2003) (Yang, Buckley et al. 2002). Generally, the settings are adjusted so that the brightest pixels are just below the saturation level, since it has been found that appropriate normalisation will correct for any bias introduced during scanning (Yang, Dudoit et al. 2002). The scanner produces two 16-bit TIFF images, one for each channel (Cy3/Cy5), that represent the fluorescence associated with each pixel on the array. From these images the signal intensities for each spot, and from the surrounding background, are extracted. This is done in three steps; (i) approximate localisation of the spot (gridding), (ii) spot and background identification (segmentation), and (iii) extraction of foreground (spot) and background intensities. Spot quality measures may also be included. A number of different image analysis programs are available, using different algorithms for these procedures, most of which provide a combination of automatic and manual gridding options.

The choice of segmentation algorithm can strongly influence the outcome of a microarray experiment, since it defines which pixels will be included in the foreground and background measures. The most commonly used segmentation methods can be divided into four groups: fixed circle segmentation, adaptive circle segmentation, adaptive shape segmentation and histogram segmentation (Yang, Buckley et al. 2002), although several other methods have also been proposed (Glasbey and Ghazal 2003). Fixed circle segmentation fits a circle with constant diameter to all the spots in the image. Since the spot size often varies to quite a high degree over the slide this is clearly not a good option. For some spots, low intensity background pixels will be included in the foreground and, similarly, other spots will include high intensity pixel values in the background measurement. In the adaptive circle segmentation method the circle diameter is adapted to the size of the spot, solving some of those problems. However, spots are rarely perfectly circular, but can exhibit oval, donut or other shapes. This is accounted for by the adaptive shape segmentation methods, which identify "seed" pixels for the spot and background, and extends the respective regions until the pixel intensities fall or rise significantly. In the histogram segmentation methods histograms of pixel values within a certain area around the spots are formed. The foreground and background pixels are then defined in various ways, e.g. according to whether their values are higher or lower than certain threshold values, higher or lower than a certain percentile of the pixel values, or some other statistical definition.

After the segmentation has been done the signal intensities for the foreground and background are calculated. The simplest, and most commonly used, method for calculating foreground intensity is to compute the mean or median pixel intensity for the spot. Although this could also be done for the background, it is more sensitive to dust specks and other artefacts, and therefore other alternatives have also been proposed (Yang, Buckley et al. 2002) (Glasbey and Ghazal 2003) (Brown, Goodwin et al. 2001). In addition, a decision must be made as to whether or not the background intensity should be subtracted from the foreground. In many cases background adjustment increases the variability and reduces the precision of the spot intensity measurement. The rationale for background subtraction, that the unspecific fluorescence is as strong within the spot as outside it, might not even be valid. The glass slides are treated both before and after printing in order to avoid unspecific binding of target DNA to the glass surface. It thus seems likely that glass surface occupied by DNA (the spots) has different properties than surface area that is not occupied by DNA. Generally, the choice of background adjustment method applied, if any, has a larger impact on the obtained intensities than the various segmentation approaches.

In addition to calculating the foreground and background intensities, most image analysis software also provides quality measurements of the spots. These can include statistics on the within-spot pixel-to-pixel intensities, spot size, signal-to-noise ratios and degree of pixel saturation (Brown, Goodwin et al. 2001) (Wang, Ghosh et al. 2001). They can also be used for subsequent data filtration and analysis.

Filtration and log-transformation

Some of the largest sources of noise within microarray data are the processes of array fabrication, hybridisation and image acquisition. These processes can yield granular and donut shaped spots, fabrication inconsistencies, highly variable background fluorescence and other artefacts. Consequently, some spots, designated poor quality spots, will not reflect the true signal ratios between the two target samples. In order to obtain reliable data for the normalisation procedure and identification of differentially expressed genes these poor quality spots should be removed. Filtration can be done using a number of different criteria, for instance (to name just a few): the spot's size, signal-to-noise ratio and level of saturation (Wang, Ghosh et al. 2001), within-spot pixel-to-pixel variation (Brown, Goodwin et al. 2001), the variation between repeatedly spotted clones (Tseng, Oh et al. 2001) (Jenssen, Langaas et al. 2002), weak spots that are indistinguishable from background (Yang, Ruan et al. 2001) and the difference between a spot's mean and median intensities (Tran, Peiffer et al. 2002). In addition, since cDNA microarray experiments are used to investigate relationships between samples, the obtained data are often presented as ratios between the sample intensities. In order to present and treat up- and down-regulated genes equally these ratios are also log-transformed (most commonly using base-2 logarithms).

Normalisation

Throughout a microarray experiment systematic bias will also be added to the data due to factors such as differences in the physical properties and/or incorporation efficiencies of the dyes, different concentrations of the hybridised target samples, scanner settings, and systematic differences between plates or cDNA libraries in the probe preparation processes and between clones or arrays during the printing process. The result is an uneven distribution of the Cy3 and Cy5 intensities. In order to remove such variation from the data, leaving only the biologically relevant information, the data have to be normalised (Smyth and Speed 2003) (Quackenbush 2002). Within-slide normalisation is done to achieve an appropriate balance between the intensities of the Cy3 and Cy5 channels. In some cases it may also be necessary to perform between-slide normalisation or scaling (see below). Within-slide normalisation is done in two steps. The first is to identify a sufficient number of non-differentially expressed genes that are present on the array. In the second these genes are used for normalisation, employing one of a number of available normalisation strategies. Even though a range of different normalisation options are available in a user-friendly format, it can take some experience to select the most appropriate method confidently. This must be done with caution, to avoid over- or under-fitting the data, thereby distorting the true biological information within it.

For the majority of microarray experiments all genes on the array can be included in the normalisation, based on the assumptions that among the many thousands of genes present on the array only a small proportion will be differentially expressed and/or that there is symmetry in the up- and down-regulation of genes (Yang, Dudoit et al. 2001). When smaller or more focused arrays are used, or when the target samples strongly differ these assumptions may not be valid. One may then use a smaller subset of genes that are known to be uniformly expressed for normalisation. One option is to use so-called "housekeeping genes"; genes that are involved in basic metabolic functions of the cell and thus are believed to be constantly expressed across a variety of conditions. There are, however, some problems with this approach. Although often believed to be constantly expressed, their expression can vary substantially. In addition they are often highly expressed, and thus cannot be used for normalisation across the lower intensity ranges. Another, similar, approach is to find genes that are constantly expressed across the material under investigation. One such approach is the rank-invariant selection of genes described by Tseng and co-workers (Tseng, Oh et al. 2001). A fourth option is to use controls for normalisation. These should ideally span the entire intensity range and be present at equal levels in both samples. One such option is to spot DNA from an unrelated organism on the array and then spike target samples with corresponding DNA in equal amounts. This approach can be used to correct for experimental bias, but not for possible differences in target sample concentration. Another control option is to spot the same probe in a dilution series on the array. For yeast experiments, genomic DNA (which should hybridise to all mRNA species in the two target samples) can be used for this purpose. However, this is not feasible in studies of higher eukaryotic genomes (e.g. mammalian genomes) because of their complexity. A conceptually similar approach was suggested by Yang and co-workers, who used a pool of all probes on the array, a microarray sample pool, MSP, for spotting in a titration series (Yang, Dudoit et al. 2002). In general, the use of all genes for normalisation offers the greatest reliability, because of the great number of genes covering all intensity ranges and spatial distributions over the slide. Subsets of genes should only be used if necessary.

In the first methods used for normalisation of microarray data it was assumed that the Cy3 and Cy5 intensities were related by a constant factor across the entire intensity range. In these so-called global normalisation methods, a constant normalisation factor was used to shift the mean of the intensity ratios of all the genes on an array to one. However, it was soon realised that the Cy5 and Cy3 dye biases, and thus the signal ratios, were dependent on spot intensity. This effect can easily be visualised in an M vs. A plot, where $M = \log_2(Cy5/Cy3)$ and $A = 0.5\log_2(Cy5*Cy3)$. A number of different intensity-dependent normalisation methods have been developed to adjust for this effect. One of the simplest is an iterative linear regression method proposed by Finkelstein and co-workers (Finkelstein, Gollub et al. 2001). It does, however, assume that the signal intensities and ratios obtained are linearly related, something which is not always true and often has to be corrected for. The most commonly used normalisation method, taking non-linearity into account, is the lowess normalisation proposed by Yang and co-workers (Yang, Dudoit et al. 2001) (Yang, Dudoit et al. 2002). In this approach, a robust scatter plot smoother (lowess; a weight function that de-emphasises the contribution of outlier spots) is used to perform a series of local regressions, one for each point in the M vs. A scatter plot. The user can define the fraction, or span, of the data that is closest in terms of intensity to the spot being predicted, to be used for the regression at each point. A large span will result in inefficient normalisation, while a small span will over-fit the data. By default the span is normally set to 0.4 (40% of the data). The lowess normalisation method has been shown to perform well in a number of different studies (Quackenbush 2002) (Tseng, Oh et al. 2001) (Xie, Jeong et al. 2004). A similar approach, using a different local regression method instead of lowess, was also proposed by Kepler *et al* (Kepler, Crosby et al. 2002). Other normalisation strategies include methods based on ANOVA (Kerr, Martin et al. 2000) (Wolfinger, Gibson et al. 2001), single value decomposition (SVD) (Alter, Brown et al. 2000) and Bayesian approaches (Newton, Kendziorski et al. 2001).

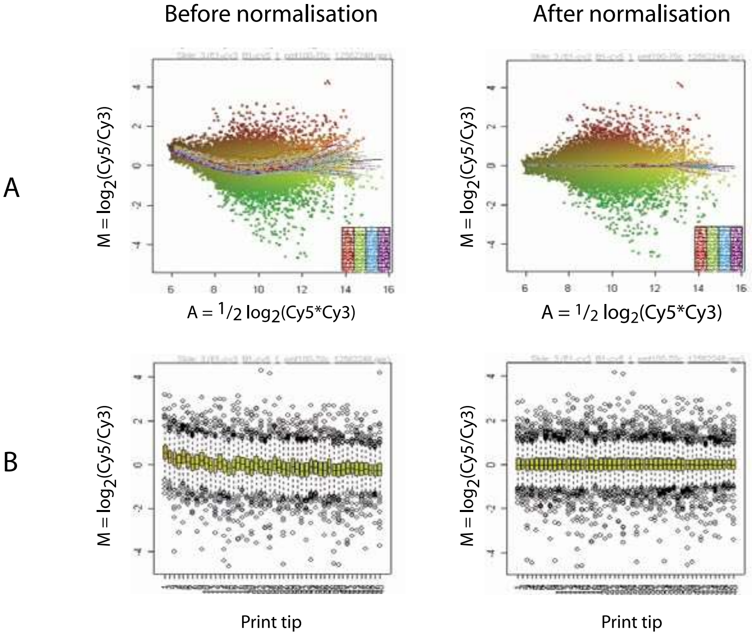

Figure 4. Visualisation of cDNA microarray data. MA plots (A) and box plots (B) can be used to visualise artefacts within the data. In the example intensity dependence of the expression ratios is observed as a "banana" shaped data cloud in the MA plot. Spatial, or print tip, artefacts can be visualised through lowess lines for each print tip group of genes, as in (A), or by box plots of the M-values for each print tip group, as in (B). The artefacts can be corrected for through for example print tip lowess normalisation (right).

Microarray data also often show non-biologically relevant trends related to the location of the probes on the array surface, due to differences in the length or opening of the print tips, variability in the slide surface, uneven hybridisation or other factors. These spatial trends can be identified through a number of different diagnostic plots (Smyth and Speed 2003). Often the probes printed by the same print tip, and therefore belonging to the same grid on the array, are grouped. These print-tip groups serve as proxies for the spatial effects. Spatial image plots, showing the background intensities or un-normalised M-values in colour code over a representation of the array surface, can be useful. Another option is to make an MA plot with separate lowess curves for the different print-tip groups. If the shape or levels of the lowess curves differ some kind of spatial normalisation is probably needed. Alternatively, side-by-side box plots of the print-tip groups can reveal differences in their M-value distribution. If spatial trends are identified in the data, normalisation can be performed separately for each of the individual print-tip groups. Most of the above mentioned normalisation methods can be used, although print-tip lowess is the most common (Yang, Dudoit et al. 2001) (Xie, Jeong et al. 2004). After normalisation the log-ratios of the different print-tip groups will be centred around zero, but the spread of the M-values can vary substantially (which can easily be seen using print-tip group box plots). If so, scaling between the print-tip groups may be required. However, this can introduce more noise, and should not be done unless necessary (Yang, Dudoit et al. 2001) (Xie, Jeong et al. 2004). Similarly, after within-slide normalisation the spread of the M-values can vary substantially between slides, and between-slide normalisation/scaling could be necessary (Yang, Dudoit et al. 2001) (Xie, Jeong et al. 2004). Normalisation and subsequent data analysis can be performed using the freely available statistical environment R ((R Development Core Team); http://www.R-project.org) and associated packages developed for microarray analysis, e.g. Bioconductor (Gentleman, Carey et al. 2004), LIMMA (Smyth 2004) and Aroma (Bengtsson).

Identification of differentially expressed genes

In the early days of microarray experiments few, if any, replicates were included in the experimental design, and a fixed fold-change (or M-value) cut-off was used to define genes that were differentially expressed (DE). This approach does not take into account the variability of the M-values and thus many genes with high fold-changes but poor-quality data were mistakenly identified as being DE, while genes with reliable data but low fold changes were missed. Slight improvements were made by "borrowing" variability data from other genes on the array, either globally (using all genes) or, better, locally, using a sliding window over genes sorted by their level of signal intensity (Yang, Chen et al. 2002). It has since become increasingly evident that replication substantially improves the quality of the obtained data, allowing an estimate of variability to be obtained for each gene (Lee, Kuo et al. 2000) (Kerr and Churchill 2001), and many sophisticated statistical approaches, based on different models, assumptions regarding error distribution and ways to estimate the errors, have been developed, only a few of which will be discussed in this thesis. These statistical tests generally involve two steps: calculating a test statistic and determining its corresponding significance.

One way to rank genes according to their differential expression when two samples (e.g. control vs. treated) are to be compared, is to use an ordinary t-test ($t = M / (s/\sqrt{n})$, where M is the average log-ratio, s is the standard deviation of the M-value and n is the number of replicates) (Dudoit, Yang et al. 2002). However, this approach is not ideal, because among the thousands of genes measured some will have very small values of s, by chance, and thus will have large t-values even if they are not DE. This problem can be solved by using a penalised t-test, where a constant is added to the denominator of t, thus making it less sensitive to small variances. Examples include the significance analysis of microarrays, SAM, developed by Tusher *et al* (Tusher, Tibshirani et al. 2001) and a non-parametric empirical Bayes method developed by Efron *et al* (Efron, Tibshirani et al. 2001). Another similar approach is the parametric empirical Bayes method described by Speed and Lönnstedt and further developed by Smyth and co-workers (Lönnstedt and Speed 2002) (Smyth, Michaud et al. 2003) (Smyth 2004). The empirical Bayes methods provide convenient ways to address the problem of "many genes, few replicates", by effectively borrowing information across genes and thus obtaining better variance estimates. This global information, or prior distribution, is combined with the gene-wise observed data to obtain posterior odds for a gene being DE. A weighting factor, or hyperparameter, will determine the relative influence of the prior distribution and the gene-wise observations on the obtained test statistic. The hyperparameter may be based on the number of replicates used (Long, Mangalam et al. 2001). The t-test approach could be further generalised for analysing more than two samples by using ANOVA-based approaches (Kerr and Churchill 2001) (Kerr, Martin et al. 2000) (Long, Mangalam et al. 2001) (Wolfinger, Gibson et al. 2001). Other statistical methods that have been used for identifying differentially expressed genes include regression modelling (Thomas, Olson et al. 2001), likelihood ratio testing (Ideker, Thorsson et al. 2000) and another empirical Bayesian approach (Newton, Kendziorski et al. 2001).

Having calculated the test statistics and ranked them accordingly, the next step is to choose a threshold value, above which the genes will be regarded as differentially expressed. An important aspect here is the need to control for multiple testing. Because large numbers of genes are studied simultaneously there is a risk of identifying false positives. The most stringent approach to multiple testing is to control the family-wise error rate (the probability of there being at least one false positive). The Bonferroni single-step adjustment of p-values is the most conservative of these, single-step meaning that all genes are corrected for equally, regardless of their p-values. Less stringent is the Holm's step-down adjustment, making successively smaller adjustments for higher p-values. Even more general and less conservative is the Westfall and Young step-down adjustment, which in addition takes into account the dependence structure between the genes (Dudoit, Yang et al. 2002). Controlling for the family-wise error rate may in any case be unnecessarily stringent for microarray experiments, since it is probably more desirable to include a number of false positives rather than risk discarding some truly significant genes. False discovery rate adjustment has therefore been proposed as an even less stringent and more appropriate alternative (Benjamini and Hochberg 1995) (Reiner, Yekutieli et al. 2003).

Data mining

The results of the pre-processing steps (data acquisition, filtration, normalisation and identification of DE genes) of a microarray experiment are usually one or more lists of differentially expressed genes. The goal then is to extract biologically meaningful information from these lists. A variety of tools are available for this data mining phase of a microarray experiment, the optimal choice depending largely on the question(s) being asked.

One of the most common approaches for data mining is the use of clustering (nicely reviewed in (Quackenbush 2001). This refers to a wide variety of methods for organising the genes into groups with roughly similar expression patterns. Unsupervised methods, requiring no other information than the expression data, include hierarchial clustering (Eisen, Spellman et al. 1998) (Spellman, Sherlock et al. 1998), k-means clustering (Tavazoie, Hughes et al. 1999), self-organising-maps, SOM (Tamayo, Slonim et al. 1999) (Toronen, Kolehmainen et al. 1999) and the related principal components analysis (Raychaudhuri, Stuart et al. 2000). One of the purposes of all of these methods is to assign potential functions to genes that have not yet been fully characterised. The basic assumption is then that genes with similar expression patterns are likely to be functionally related (due to so-called "guilt-by-association"). Although not a rigorous assumption, the approach has proven to be successful in many cases (Cho, Campbell et al. 1998) (Hughes, Marton et al. 2000). Cluster analysis can also lead to the identification of new transcription factors involved in the transcription of a co-regulated set of genes, and their binding motifs within promoter sequences (Roth, Hughes et al. 1998) (Bussemaker, Li et al. 2001). A third application where clustering has been used is for classifying new disease subclasses, based on the associated expression profiles (Alizadeh, Eisen et al. 2000). If, however, the classes are already known (e.g. tumour vs. normal tissue) and one wants to discriminate between them at the gene expression level, it is far more efficient to use discrimination, or supervised, methods (Golub, Slonim et al. 1999). These methods require both expression data and class assignments as inputs. They include a training phase with samples whose classes are already known, and a testing phase in which the algorithm uses information from the training set to predict classes in an uncharacterised data set. Supervised methods include easily applied methods such as linear discriminant analysis, nearest-neighbour classifiers, classification trees (Dudoit, Fridlyand et al. 2002), and more advanced methods such as the use of neural networks and support vector machines (Brown, Grundy et al. 2000).

Another way to mine the data from microarray experiments is to look for biological themes within the generated gene lists, based for instance on gene ontology (GO) annotation, a controlled vocabulary for describing a gene's molecular function, its involvement in different biological processes and the cellular locations where its effects are exerted (Ashburner, Ball et al. 2000). The vocabulary, developed by the Gene Ontology Consortium, includes GO terms that can be used for all eukaryotes and can be included in various databases containing information about genes and their products. A number of different software packages can link the genes within microarray gene lists to appropriate GO terms and use different statistics to calculate their possible over-representation within the data. These include EASE (http://apps1.niaid.nih.gov/david/; (Hosack, Dennis et al. 2003)), MAPPfinder (http://www.genmapp.org; (Doniger, Salomonis et al. 2003)) and GoMiner (http://discover.nci.nih.gov/gominer/; (Zeeberg, Feng et al. 2003)).

A related way to mine microarray data is to look for overrepresentation of genes involved in certain known or predicted cellular pathways. Programs such as GeneSpring (Silicon Genetics, CA, USA) map differentially expressed genes to known pathways, collected in the KEGG (Kyoto Encyclopedia of Genes and Genomes) database (Ogata, Goto et al. 1999), and calculate scores of their overrepresentation. Other programs, such as PathwayAssist (Ariadne Genomics, MD, USA) use advanced scientific text mining tools to automatically extract biological findings from scientific literature and to build networks of molecular interactions. These networks can then be used in a similar fashion as the KEGG pathways in GeneSpring. GeneSpring and PathwayAssist belong to a large group of available software packages which offer user-friendly ways to perform normalisation, statistical analysis, cluster and discrimination analysis and other data mining tools (see also GeneSight (BioDiscovery, CA, USA) and the freely available packages ExpressionProfiler (EMBL-EBI, Hinxton, UK) and GeneCluster (the Broad institute, MA, USA). Less user friendly, but highly flexible and versatile tools for microarray data analysis have also been developed for the freely available statistical platform R (http://www.r-project.org) (Gentleman, Carey et al. 2004) (Smyth 2004) (Bengtsson).

3.1.4.4 Microarrays – perspectives

In addition to the GeneChip® oligonucleotide and cDNA arrays described above a new approach, in which long (40–80 bases) oligonucleotides are used as probes, has been developed in the last few years. The first such arrays were synthesised by an ink-jet printing method (Hughes, Mao et al. 2001). Recently, pre-fabricated oligonucleotides have been used and spotted similarly to cDNA microarrays (Beaucage 2001), and whole genome longmer oligonucleotide sets for printing are now available from commercial vendors. The long oligonucleotide arrays combine properties of both cDNA and GeneChip® arrays and offer several advantages: the greater length of the probes enables higher specificity in the hybridisation than the shorter GeneChip® probes, and makes them less sensitive to phenomena such as single nucleotide polymorphisms. Every probe has the same length and approximately the same melting temperature, making the hybridisation more homogenous, and thus enabling greater control of the hybridisation. The oligonucleotides are single stranded, so they do not require a denaturation step and cannot re-naturate and decrease hybridisation efficiency. In addition, the researcher has greater control over the transcript (sense or anti-sense) the probe hybridises to. Last, but not least, the manual design of the oligonucleotides offers greater control over the probes than cDNA probes and gives opportunities to study features such as different specific splice variants.

The use of microarrays has also extended beyond the study of overall gene expression. For transcriptome studies, for example, special arrays have been designed to study alternative splicing (Hu, Madore et al. 2001) (Clark, Sugnet et al. 2002) (Johnson, Castle et al. 2003) (Yeakley, Fan et al. 2002). Microarrays have also been used to monitor transcript degradation (Bernstein, Khodursky et al. 2002) (Fan, Yang et al. 2002) and evaluate the expression of noncoding RNAs (Cawley, Bekiranov et al. 2004). For genome studies exon and “tiling” arrays have been used to experimentally validate and refine computational gene predictions (Shoemaker, Schadt et al. 2001) (Kapranov, Cawley et al. 2002). Microarrays have also been used for SNP genotyping (Patil, Berno et al. 2001) and to detect changes in DNA sequence copy number (Kashiwagi and Uchida 2000), for mapping the origins of replication in *Saccharomyces cerevisiae* (Raghuraman, Winzeler et al. 2001) and *Sulfolobus* (Lundgren, Andersson et al. 2004), identifying the binding sites for transcription factors on a genome-wide level (Iyer, Horak et al. 2001) (Bulyk, Huang et al. 2001), and characterising epigenetic modifications of the chromatin (van Steensel and Henikoff 2003).

As described above, the microarray techniques are highly diverse, including multiple array formats and a wide variety of protocols for fabrication, target sampling and data analysis. This has led to difficulties in comparing and interpreting the sometimes very divergent results from similar studies performed at different laboratories. The need for standardised and informative presentation and exchange of microarray data has thus gradually been recognised, and in 1999 the Microarray Gene Expression Data (MGED) Society was formed, with the aim of drafting proposals for such standards. The resulting proposals, the Minimum Information About a Microarray Experiment (MIAME) standards, were published in 2001 (Brazma, Hingamp et al. 2001), describing the minimum information required to ensure that microarray data can be easily interpreted and that results derived from their analysis can be independently verified. The MIAME standards have also facilitated the development of databases for storage and exchange of microarray data, including the European database ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) and the American Gene Expression Omnibus, GEO (http://www.ncbi.nlm.nih.gov/geo/).

3.2 Selective approaches for transcript profiling

In addition to the global gene expression methods, in which all expressed transcripts are studied, techniques for identifying only genes that are differentially expressed have also been developed. The most commonly used selective approaches are based on either fingerprint assays (where a "fingerprint", such as transcript-containing bands on a gel, is generated for each sample to be compared) or subtraction hybridisation (where signals from common transcripts are subtracted from one sample by signals from the transcriptome of the other sample, leaving only signals from the unique transcripts of the first sample to be analysed).

3.2.1 Differential Display (DD) and RNA arbitrarily primed PCR (RAP-PCR)

One of the first PCR-based methods to be developed for gene expression analysis was differential display (DD) (Liang and Pardee 1992). This method relies on reverse transcription and PCR amplification of the 3'-ends of transcripts using a set of different primer pairs. For each primer pair a subset, typically 50–100, of the cDNA population will be amplified. This sub-population of cDNA fragments, denoted an RNA fingerprint, can be separated into bands of distinct sizes on a polyacrylamide gel. Through side-by-side comparison of the gels obtained from two samples, bands representing differential expression (i.e. bands that are only present in one of the samples) can be discerned. DNA from the corresponding bands can then be isolated, cloned and sequenced for gene identification. By repeating the procedure with many different primer pairs fingerprints covering the whole transcriptome can be obtained. A closely related method to DD is RNA arbitrarily primed PCR (RAP-PCR) (Welsh and McClelland 1991) (Welsh, Chada et al. 1992). In DD one of the primers of primer pair is directed towards the polyA tail, generating 3'-biased fragments, whereas in RAP-PCR both primers have an arbitrary sequence, enabling the generation of fragments throughout the gene sequence and also from non-polyadenylated RNA, such as prokaryotic RNA.

The major advantage with DD is the small amount of starting material required (a few ng of mRNA). In addition, novel transcripts can be identified, both up- and down- regulated genes are analysed in the same experiment and more than two samples or treatments can be compared simultaneously (McClelland, Mathieu-Daude et al. 1995) (Liang 2002). The main drawback has been that DD generates a large number of false positives (up to >70% of the discovered transcripts), leading to the need for laborious and time consuming optimisations and confirmatory work (Sompayrac, Jane et al. 1995) (Debouck 1995). However, efforts have been made to refine and improve DD, resolving some of these problems (Liang 1998)

(Matz and Lukyanov 1998) (Rosok, Odeberg et al. 1996). Similarly to other methods it is also difficult to detect rare transcripts by DD (McClelland, Mathieu-Daude et al. 1995), although improved strategies for that purpose have also been developed (Ralph, McClelland et al. 1993).

3.2.2. Representational difference analysis (RDA)

The first subtraction hybridisation assays for analysing of differential gene expression appeared in the mid 1980's to 90's (Hedrick, Cohen et al. 1984) (Duguid, Rohwer et al. 1988) (Hara, Kato et al. 1991) (Wang and Brown 1991) (Zeng, Gorski et al. 1994). In general they involve hybridisation of cDNA from one population (the tester) to an excess of cDNA from another population (the driver). The unique, unhybridised cDNAs in the tester sample are then separated from the hybridised, common cDNAs, using various approaches. For example, single stranded driver cDNA can be labelled with biotin, and after hybridisation with single stranded, unlabelled tester the hybrids can be removed using streptavidin affinity resin (Duguid, Rohwer et al. 1988). These methods have been successfully used to isolate numerous important genes, but nevertheless have a number of drawbacks. They often require relatively large amounts of starting RNA, although the introduction of PCR has dramatically increased the sensitivity. Greater problems include the inefficiency of obtaining low abundance transcripts, and the facts that they often involve multiple subtraction steps, are technically difficult, labour intensive and often unreliable.

In 1995 the PCR-coupled subtractive method representational difference analysis (RDA) was introduced (Hubank and Schatz 1994). The method was developed from a protocol originally used for detecting genome differences (Lisitsyn and Wigler 1993). RDA does not require physical separation of the hybridisation products. Instead, the products of interest (the differentially expressed transcripts) are amplified by PCR, which is much easier to do. First, so-called representations, amplified cDNA representing the original cDNA, from two samples are generated. This is done by restriction enzyme digestion (using a "four-cutting" enzyme) followed by linker ligation and PCR amplification. The linkers are then removed and new linkers are ligated onto the tester sample. An excess of driver is mixed and hybridised to the tester sample. This generates a mixture of double stranded hybridisation products; driver-specific sequences with no linkers, driver-tester hybrids with linker on one strand and tester-specific sequences with linkers at both ends. In a PCR with linker-specific primers the tester specific fragments will thus be exponentially amplified while the hybrids will be linearly amplified and the driver-specific fragments not amplified at all. The resulting product is termed the first difference product (DP). To further enrich the tester-specific fragments the procedure is repeated several times, with successively more stringent hybridisation conditions (increased ratios of driver to tester). Both up-regulated and down-regulated genes can be identified by interchanging the tester and driver samples.

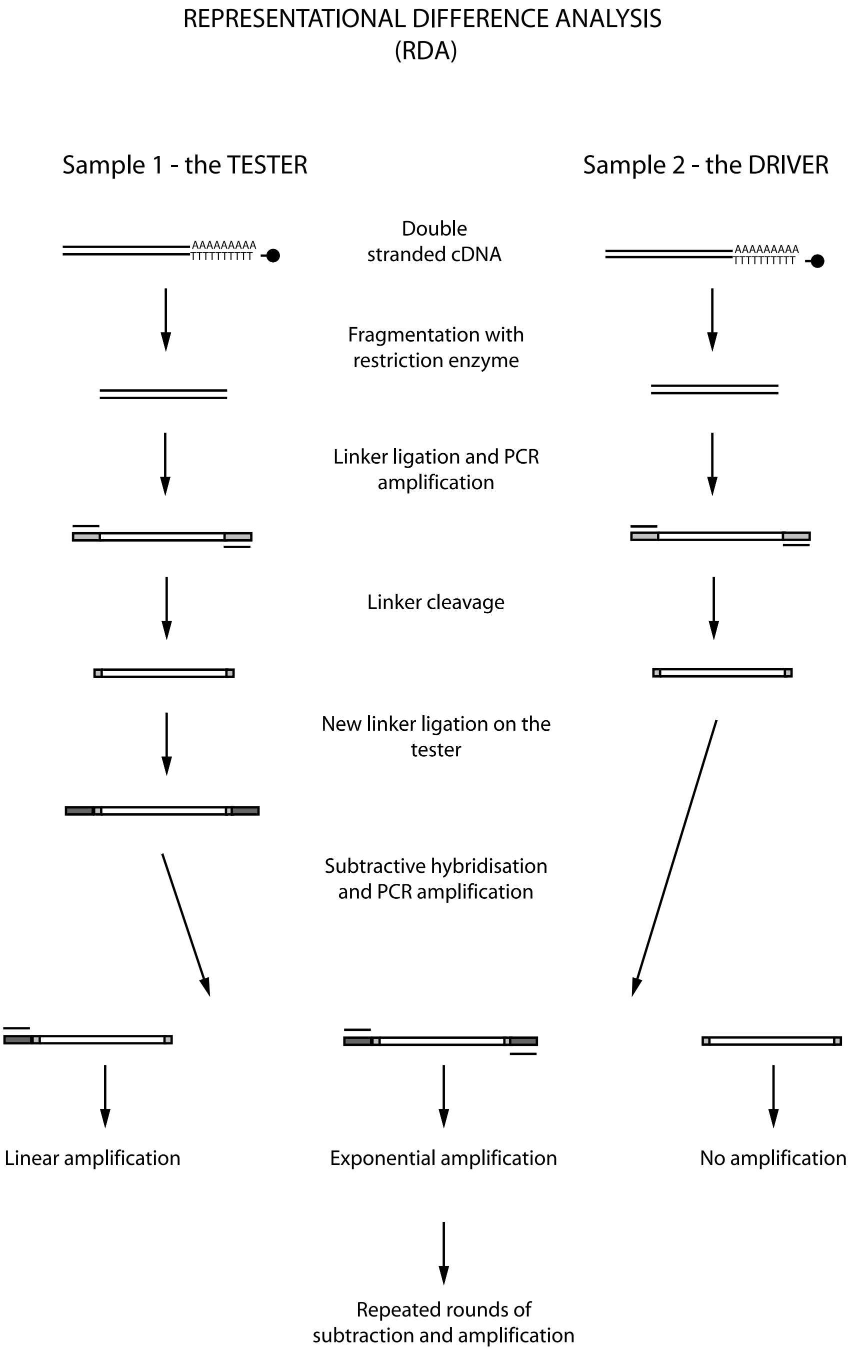
REPRESENTATIONAL DIFFERENCE ANALYSIS
(RDA)
Sample 1 - the TESTER
Sample 2 - the DRIVER
AAAAAAAAA
TTTTTTTTT
Double stranded cDNA
AAAAAAAAA
TTTTTTTTT
Fragmentation with restriction enzyme
Linker ligation and PCR amplification
Linker cleavage
New linker ligation on the tester
Subtractive hybridisation and PCR amplification
Linear amplification
Exponential amplification
No amplification
Repeated rounds of subtraction and amplification

Figure 5. Schematic overview of cDNA RDA. (Kindly provided by Stina Boräng)

The main advantage of RDA is its efficiency in enriching differentially expressed genes, yielding few false positives. Furthermore, by modifying the tester:driver ratio, genes with different magnitudes of differential expression can be identified. Because of the amplification of the original cDNA to generate the representations, RDA does not require large amounts of starting material (a few µg of total RNA). Limitations of the method include the use of restriction enzymes. Transcripts with several recognition sites for the enzyme will yield multiple fragments, while transcripts with one or no recognition sites will be lost. For the original distribution of transcript levels to be maintained, the representations must also be carefully generated and tedious PCR titration may be needed. Also, the protocol is mainly suitable for the comparison of similar samples, where few genes are expected to be differentially expressed, and is not efficient for the detection of rare transcripts.

3.2.3 Suppression Subtractive Hybridisation (SSH)

The main criticism raised against RDA, as well as the previous subtraction hybridisation methods, is their inability to simultaneously measure both rare and abundant mRNA species. In 1996 Luda Diatchenko and co-workers introduced a method called suppression subtractive hybridisation (SSH), which addresses that problem (Diatchenko, Lau et al. 1996). This is done by including a normalisation step simultaneously to the subtractive hybridisation. Suppression PCR is also included in the process, facilitating separation of the hybridisation products.

In SSH the tester and driver cDNA populations are first digested by restriction enzymes. The tester is divided into two identical halves, each ligated to separate sets of linkers containing long, inverted terminal repeats. The subtractive hybridisation is then carried out in three steps (Diatchenko, Lau et al. 1996) (Gurskaya, Diatchenko et al. 1996). During the first steps, the cDNA fragments of interest that will "survive" the whole process, and eventually be identified as differentially expressed, are those that remain single-stranded (ss). In the first step excess driver is added to each tester sample. During this step double-stranded (ds) molecules will form from cDNAs present in both the tester and driver populations (heterohybrids) as well as from cDNAs only present in the tester or in excess in the driver (homohybrids). Due to the second-order kinetics of hybridisation the hybridisation will be much more effective for abundant than for rare molecules. Consequently, the ss concentration of both abundant and rare transcripts will become approximately equalised. In step two freshly denatured driver is added to the hybridisation mixture. This will remove even more of the molecules present in both the tester and driver populations from the ss fraction. After steps one and two, molecules that are still single-stranded represent transcripts unique to the tester (with linkers), at normalised levels, as well as a small portion of transcripts that are also present in the driver (without linkers). In the third step the two hybridisation samples are mixed, the remaining ss molecules can hybridise and the tester-unique transcripts form ds molecules asymmetrically flanked with two different linkers. In the subsequent PCR these molecules will be exponentially amplified. However, tester homohybrids, formed mainly by abundant transcripts, are symmetrically flanked by the same linker. Containing long, inverted terminal repeats these molecules will turn back and anneal on themselves during PCR,

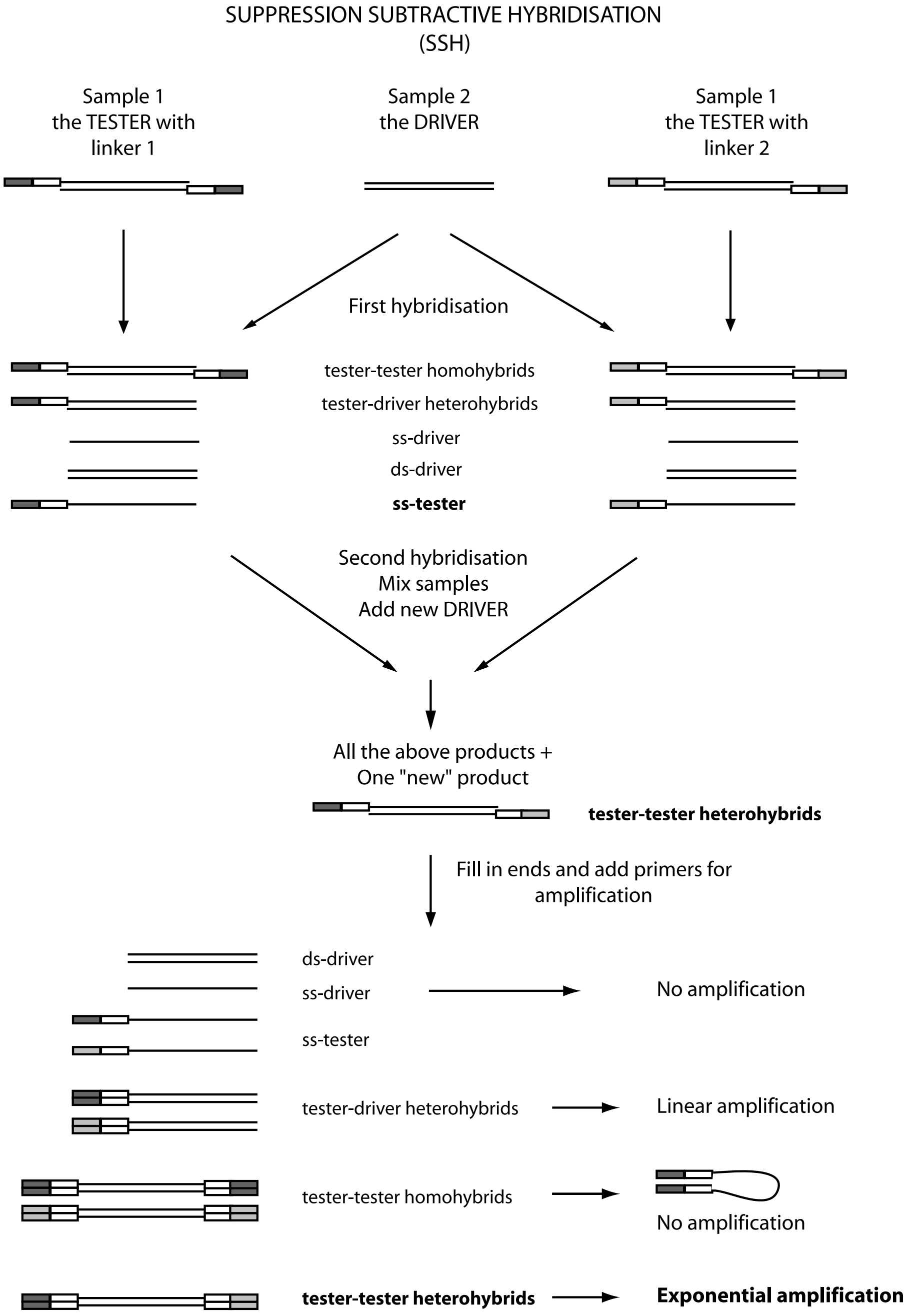

Figure 6. Schematic overview of the SSH method. (Modified from picture provided by Stina Boräng)

forming panhandle-like structures that will prevent them from amplifying (suppression PCR) (Siebert, Chenchik et al. 1995). All other hybridisation products will contain only one linker or no linker at all, thus they will be linearly amplified or not amplified at all during PCR.

The main advantages with SSH are its ability to simultaneously detect abundant and rare differentially expressed genes, and its relatively simple procedure. One disadvantage is that relatively high amounts of starting material are needed (1–2 µg of mRNA). Like many other methods SSH, also relies on restriction enzymes, so transcripts lacking the recognition site will be lost during the analysis. Another problem associated with SSH is that it can generate many false positives, although efforts to solve that problem have been made (Rebrikov, Britanova et al. 2000).

Combining selective methods such as RDA and SSH with microarrays has also been shown to be a promising approach for gene expression studies (Yang, Ross et al. 1999) (Geschwind, Ou et al. 2001) (Andersson, Unneberg et al. 2002).

4. **Transcript profiling of small tissue samples**

Microarray technology approaches have become the most widely used techniques for analysing global gene expression, due to their capacity to measure thousands of transcripts in parallel and to their simplicity. One of the limitations in the early days of these techniques was that they required significant amounts of RNA; the earliest microarray experiments required 100 µg of total RNA or more. This amount has since been lowered to 10–20 µg of total RNA, which still corresponds to millions of cells. In investigations of many biological issues such numbers of cells are far from obtainable. For example, one might want to study very small tissues or only certain cell types isolated from heterogeneous tissues, corresponding to tens to thousands of cells. In many cases it is impossible or inappropriate to grow and expand the cell type of interest in culture. In order to reduce the required amount of starting material, a number of different amplification strategies have therefore been developed. These can be divided into three main categories: amplification of the target label signal, linear amplification of the transcriptome using *in vitro* transcription, and exponential amplification of the corresponding cDNA using PCR.

4.1 Amplification of the target label signal

One way to lower the amount of material needed for detection on a microarray is to increase the signal strength from each molecule. This can be done by increasing the number of labels that are incorporated into the target or by amplifying the signal post hybridisation. In 2000 Stears and co-workers described a signal amplification system based on so-called 3DNA dendrimers (Stears, Getts et al. 2000), which are now commercially available from Genisphere® Inc (http://www.genisphere.com/). Dendrimers are large, branched, spherical complexes, formed by partially double-stranded oligonucleotides that are labelled with predefined numbers (hundreds) of Cy3 or Cy5 molecules. During first-strand cDNA synthesis of the target, the 5'-end of the oligo(dT) primer contains a sequence which is complementary to the dendrimer. After hybridisation this sequence is used to capture the dendrimer, thereby labelling the target. The advantages of this approach is that the labelling does not require incorporation of modified nucleotides and is not sequence dependent (unlike direct labelling), and it is faster and easier to perform than other amplification strategies (see below). This approach reduces the requirements for starting RNA to 1 µg of total RNA.

Another signal amplification method is the MICROMAX™ tyramide signal amplification system supplied by PerkinElmer® (http://las.perkinelmer.com/) (Adler, Broadbent et al. 2000) (Karsten, Van Deerlin et al. 2002). Tyramide signal amplification has been used for a long time to yield high sensitivity in immunohistochemistry studies. For microarray purposes one target cDNA sample is first labelled with biotin and the other with fluorescein-modified nucleotides. Post hybridisation the signal is then amplified through enzymatic reactions, by conjugating of streptavidin-horseradish peroxidase (HPR) and adding of Cy5-tyramide, or antifluorescein-HPR antibody and Cy3-tyramide, respectively. The method requires 0.5–1 µg of starting total RNA, could potentially introduce sequence and label biases, and takes a few hours to perform.

Radioactive labelling, using two different isotopes for the two target samples, has also been proposed as it increases the sensitivity and reduces the amount of starting material required to 100 ng of mRNA (approx. 2 µg of total RNA) (Salin, Vujasinovic et al. 2002). A simple and less expensive method was described by Xiang et al (Xiang, Kozhich et al. 2002). By using amine-modified random hexamer primers in combination with aminoallyl modified nucleotides in the cDNA synthesis, and then coupling the amine-groups to the fluorescent dyes, they attained a sensitivity that allowed the use of just 1 µg of total RNA as starting material.

Although these methods are relatively easy to perform and avoid possible biases associated with target amplification many of them are rather expensive, and the 1 µg requirement of starting total RNA still corresponds to tens to hundreds of thousands of cells.

4.2 RNA amplification using *in vitro* transcription

Currently the most widely used amplification strategy for microarray analysis is *in vitro* transcription of the target samples, a method first described by Eberwine and co-workers in the early 1990s (Van Gelder, von Zastrow et al. 1990) (Eberwine, Yeh et al. 1992). This involves cDNA synthesis using an oligo(dT) primer with a T7 RNA promoter sequence in the 5'-end. Following first and second strand cDNA synthesis the promoter facilitates linear amplification of antisense RNA (aRNA), resulting in a 10^2–10^3 fold increase in starting material. Repeated cDNA synthesis can be performed, facilitating a second step of RNA amplification which results in a 10^5–10^6 fold increase (Eberwine, Yeh et al. 1992). Minor modifications have since been suggested to improve the original protocol (Wang, Miller et al. 2000) (Baugh, Hill et al. 2001) (Xiang, Chen et al. 2003). Validation of the T7 RNA amplification protocol suggests that the method is 3'-biased (i.e. produces 5' truncated transcripts) and introduces some distortion of the original transcript representation. However, these distortions are systematic and not too severe, making the comparison of two amplified samples highly reproducible and feasible (Wang, Miller et al. 2000) (Baugh, Hill et al. 2001) (Hu, Wang et al. 2002) (Zhao, Hastie et al. 2002). The great power of the method is its sensitivity; less than 500 ng of starting total RNA is routinely used, but as few as 1000 cells (Luo, Salunga et al. 1999), 10 cells (Xiang, Chen et al. 2003) or even a single cell (!)(Kamme, Salunga et al. 2003) have also been successfully used. It is, however, time-consuming, labour-intensive and costly, making it less suitable for high-throughput applications.

4.3 Exponential amplification of cDNA using PCR

PCR-based strategies for amplification have a number of potential advantages; they usually yield vast amounts of material through the exponential amplification, they are fast and easy to perform and they are more cost effective than the abovementioned methods. However, the development of PCR-based strategies for amplification of target samples to be analysed by microarrays has been hampered by the fact that PCR is generally considered to introduce biases, distorting the original relative transcript abundances. These biases are believed to be mainly due to differences in the efficiency of amplifying transcripts of different lengths, with different primer sequences or at different abundance levels (there are greater risks that rare genes will not be primed in the first cycles). Nevertheless, some PCR-based methods for target amplification have been developed.

A method that can be used in attempts to avoid the abovementioned problems was developed for small-scale expression analysis by Iscove and co-workers, and later improved and evaluated for use in combination with microarrays (Brady and Iscove 1993) (Iscove, Barbara et al. 2002). It involves introducing oligo(dT) sequences in both the 3'- and 5'-ends of the cDNA, by using an oligo(dT) primer in the first strand synthesis and adding an oligo(dA) tail using a terminal transferase. Subsequent PCR can then be performed using a single oligo(dT)-containing primer. The length of the cDNAs are kept homogenous by limiting the deoxynucleotide concentration and time of reverse transcription, thereby limiting the cDNA synthesis to only a few hundred bases of the 3'-end. The method was shown to be highly reproducible and to yield higher correlations than *in vitro* transcription to results obtained with unamplified targets (Iscove, Barbara et al. 2002). As little as 10 pg of total RNA, corresponding to a single cell, could be used as starting material.

The SMART™ (Switch Mechanism At 5'-end of RNA Transcription) approach was originally developed by CLONTECH for generating of full-length cDNA libraries (http://www.bdbiosciences.com/clontech/; (Endege, Steinmann et al. 1999) (Zhu, Machleder et al. 2001). The method is based on the template-switching effect; when it reaches the 5'-end of the RNA template, the reverse transcriptase applied is able to add a few extra, non-template nucleotides to the newly synthesised cDNA strand. A template-switch oligonucleotide is added to the reaction, which anneals to the extra nucleotides and enables the reverse transcriptase to switch template and continue cDNA synthesis. In this way the template-switch oligonucleotide sequence, in combination with a sequence added through the oligo(dT) primer, can be used for PCR amplification of the full-length cDNA. The method has been evaluated (Vernon, Unger et al. 2000) (Gonzalez, Zigler et al. 1999) and compared to *in vitro* transcription amplification (Puskas, Zvara et al. 2002) (Petalidis, Bhattacharyya et al. 2003). It introduces slight changes in the transcript ratios, compared to the results obtained by conventional labelling, but with careful optimisation of the number of cycles used in the PCR it is highly reproducible and identifies differentially expressed genes that are often missed by other methods. In one study (Puskas, Zvara et al. 2002) the *in vitro* transcription performed better than the SMART™ technique, while in another the latter gave better results (Petalidis, Bhattacharyya et al. 2003). A great advantage with the SMART™ approach is that the full-length amplification enables alternative splicing to be investigated. As little as 50 ng of starting total RNA has been successfully used with the method.

Other PCR-based amplification strategies include a method in which both target samples are co-amplified in the same tube before separation (Makrigiorgos, Chakrabarti et al. 2002), a semi-linear, asymmetric PCR method (Smith, Underhill et al. 2003) and methods combining T7 *in vitro* transcription and PCR (Aoyagi, Tatsuta et al. 2003) (Castle, Garrett-Engele et al. 2003).

5. **Verification strategies**

Many of the described transcript profiling methods require verification by independent methods. Microarray results are, for example, inherently noisy and before one continues time-consuming and costly research on particular genes their differential expression needs to be confirmed. Traditional methods include northern blot (Alwine, Kemp et al. 1979) and *in situ* hybridisation (Gall and Pardue 1969). A newer, and considerably more sensitive, method is based on real-time quantitative PCR (Heid, Stevens et al. 1996) (Gibson, Heid et al. 1996). Using different chemistries fluorophores are incorporated into the newly synthesised DNA during PCR. The increased fluorescence is measured in real time using a photosensitive detector connected to the PCR instrument and the number of cycles required to reach a certain threshold is determined. The results for individual genes are normalised against a gene that is known to be expressed at similar levels in the samples compared, and then the relative expression levels for a certain gene in the different samples can be determined. The method requires careful template titration, primer design and replication to avoid unspecific amplification artefacts, but if carefully planned and performed it is extremely sensitive and powerful.

Objectives

The aim of the work underlying this thesis was to develop a novel target amplification protocol for analysing gene expression in small tissue samples. The method is easy to use, inexpensive and can be scaled for use in high-throughput gene expression analysis of multiple samples in parallel. Papers I and II present the method and describe experiments performed to evaluate its performance using cDNA microarrays (Paper I) and Affymetrix oligonucleotide arrays (Paper II). The amplification protocol was then applied to gene expression analysis of cultured adult neural stem cells (neurospheres). Paper III describes an investigation of the variation in gene expression between different populations of neurospheres, which verified that replicates of neurospheres are similar enough for further studies. Gene expression in neurospheres was then monitored following treatment with the proliferative agent adenylate cyclase-activating polypeptide (PACAP), as presented in Paper IV.

6. **Development of a cDNA tag amplification strategy (I and II)**

As mentioned in section 4, some kind of amplification of either the target label signal or the target itself is usually needed for gene expression analysis of small tissue samples. The mentioned amplification strategies all have attractive features as well as drawbacks. Enhancing the label signal strength minimises the risk of biased target amplification, but the methods are usually rather costly and do not increase the sensitivity sufficiently for analysing minute amounts of sample. T7 RNA polymerase *in vitro* transcription, on the other hand, enables very small samples to be analysed, but is labour intensive, time consuming, and thus difficult to use for high-throughput purposes. PCR amplification techniques provide efficient, exponential amplification and are fast and easy to use. The use of PCR amplification techniques has, however, been impeded by the risk of skewing the relative levels of transcripts, since they tend to amplify short templates more efficiently than longer ones (McCulloch, Choong et al. 1995).

Papers I and II describe a PCR-based amplification method that was developed to be easy to use and to allow automation (thus facilitating high-throughput analysis of multiple samples), while minimising the risk of size-biased amplification. Size normalisation of the transcriptome is achieved by sonicating the cDNA population, generating cDNA tags that vary between 100 and 600 bp in length. Similar results can be obtained using restriction enzymes, but due to the non-random nature of the DNA sequence, restriction enzymes cut with uneven frequencies, and some transcripts may even lack certain restriction sites, thereby preventing their identification. In addition, in contrast to restriction enzymes the sonication fragments each transcript randomly, further minimising amplification bias. The whole amplification procedure is illustrated in Figure 7. The cDNA is 3'-labelled with biotin by incorporating a biotinylated oligo(dT) primer (which also contains a *Not*I restriction site) in the cDNA synthesis. This is used, following sonication, to capture the 3'-end fragments (denoted 3'-end signature tags) on streptavidin-coated magnetic beads, while all non-biotinylated fragments are removed. Thus, only the 3'- ends – the most unique part of the transcripts – will be amplified. The isolation of short 3'-end signature tags may also yield representative tags for transcripts with strong secondary structures that hinder full-length extension. While still immobilised, adaptors containing a PCR primer site are ligated to the 5'-end of the signature tags, which are subsequently removed from the solid phase through *Not*I restriction, and finally PCR amplified using a *Not*I-oligo(dT) primer and an adaptor primer. The amplified 3'-end signature tags can be cloned and sequenced or fluorescently labelled through asymmetric PCR and hybridised to microarrays.

In Paper I the developed amplification strategy was evaluated using cDNA microarrays. Transcripts in two different plant tissues (xylem and phloem) were compared on arrays containing 192 hybrid aspen cDNA clones spotted in triplicate. Expression ratios obtained using different target preparation protocols were then compared. These included traditional labelling (using reverse transcriptase) of 1 µg unamplified mRNA or 100 µg total RNA, as well as labelling (using Taq DNA polymerase) of amplified 3'-end signature tags from 1 µg or 100 ng of starting total RNA. The expression ratios obtained with amplified signature tags showed good correlation to the ratios obtained with unamplified RNA (with correlation coefficients > 0.95 for all hybridisations except one, with a correlation coefficient of 0.93). For replicate hybridisations with amplified 3'-end signature tags the correlation coefficients were even higher (0.96 when starting with 1 µg total RNA and 0.94 when starting with 100 ng), showing the high reproducibility of the method. It was also shown that a 1.6- to 2-fold difference in expression ratios was detected with 99% confidence when comparing replicate hybridisations with unamplified targets. Slightly worse, but comparable results (a 2.1-fold difference) were obtained for comparisons of hybridisations with amplified and unamplified targets. A self-to-self hybridisation with amplified 3'-end signature tags from xylem tissue on a larger array (2995 clones) further demonstrated the reproducibility of the amplification protocol. It showed that 1.7-fold changes in expression levels can be detected with 99% confidence.

A more detailed description and broader evaluation of the 3'-end signature tag amplification protocol was presented in Paper II. Optimisation of the sonication and amplification steps were first demonstrated, starting with a single, 1500 bp long, cDNA clone. The size distribution of the obtained fragments was monitored, confirming that the sonication yielded a quite homogenous fragment population (100–600 bp) and that the size distribution was maintained during the amplification. The whole protocol was then applied to a more complex cDNA population, starting with total RNA from HeLa cells, again confirming the expected size distribution after sonication and amplification. To confirm that the relative transcript levels are maintained throughout the protocol and that the amplification is not biased, a series of hybridisations were made using Affymetrix oligonucleotide arrays with probe sets for 5 600 genes. Unamplified, full-length cDNA was labelled and hybridised according to standard Affymetrix protocols and compared to hybridisations with amplified 3'-end signature tags from 10, 20 and 30 cycles. The correlation coefficients for results from

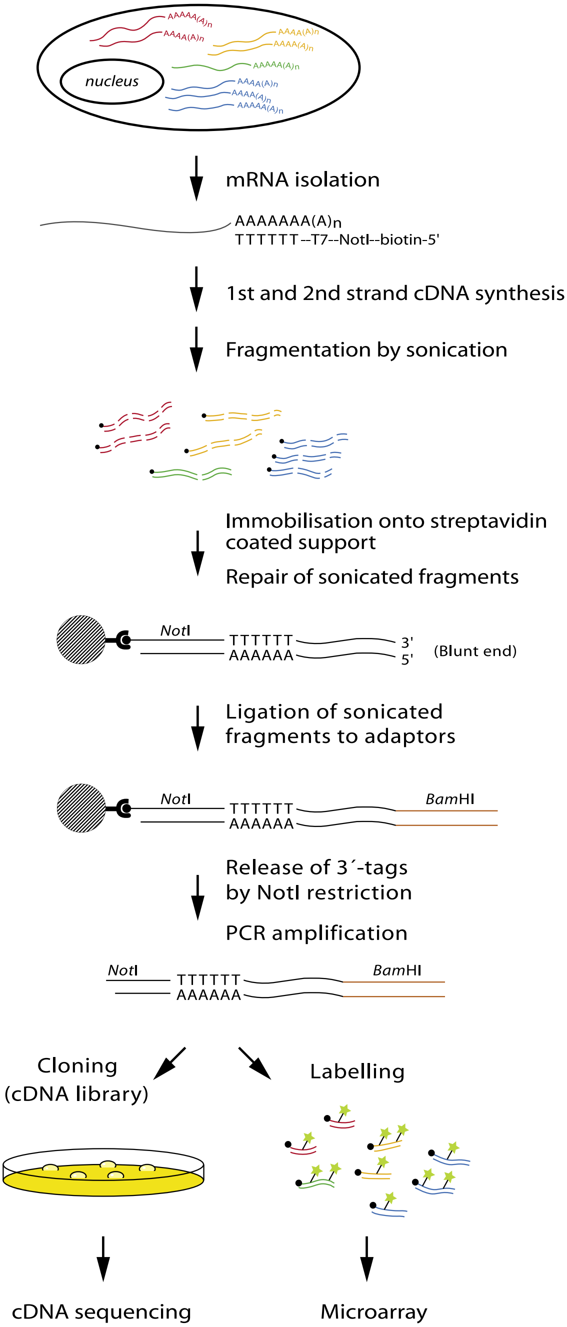
nucleus
AAAA(A)n
AAAAA(A)n
mRNA isolation
AAAAAAA(A)n
TTTTTT--T7--NotI--biotin-5'
1st and 2nd strand cDNA synthesis
Fragmentation by sonication
Immobilisation onto streptavidin coated support
Repair of sonicated fragments
NotI
TTTTTT
AAAAAA
3'
5'
(Blunt end)
Ligation of sonicated fragments to adaptors
NotI
TTTTTT
AAAAAA
BamHI
Release of 3´-tags by NotI restriction
PCR amplification
NotI
TTTTTT
AAAAAA
BamHI
Cloning (cDNA library)
Labelling
cDNA sequencing
Microarray

Figure 7. A schematic overview of the 3'-end signature tag amplification strategy.

the amplified signature tags and the full-length cDNA were 0.93, 0.89 and 0.85 for 10, 20 and 30 cycles, respectively. It was, however, realised that the expression data obtained from Affymetrix hybridisations are calculated with an algorithm that uses information from all the probes representing the respective genes. These probes are distributed along the transcript sequence, thus covering both the 3'- and more 5'-parts. The data from full-length cDNA, covering all probes, will thus yield different results than data obtained with 3'-end signature tags (Figure 8). Unfortunately, at that time there was no easy way to exclude the results from the 5'-probes to facilitate a more equal comparison. Instead, the results from the most modestly amplified sample (10 cycles) was compared to the results from 20 and 30 cycles, yielding correlation coefficients of 0.96 and 0.94, respectively. Unbiased amplification was further confirmed using quantitative real-time PCR. The relative transcript levels from 13 genes were monitored by running RTQ-PCR on full-length, unamplified cDNA as well as amplified 3'-end signature tags from 10, 20 and 30 cycles. The results further confirmed that relative transcript levels are maintained during least 20 cycles of amplification.

The results in Papers I and II show that the developed amplification strategy is non-biased and reproducible, and can be used for reliable global transcript profiling. The method has facilitated subsequent studies in which as little as 10 ng of total RNA (corresponding to approximately 1000 cells) was used as starting material (see below).

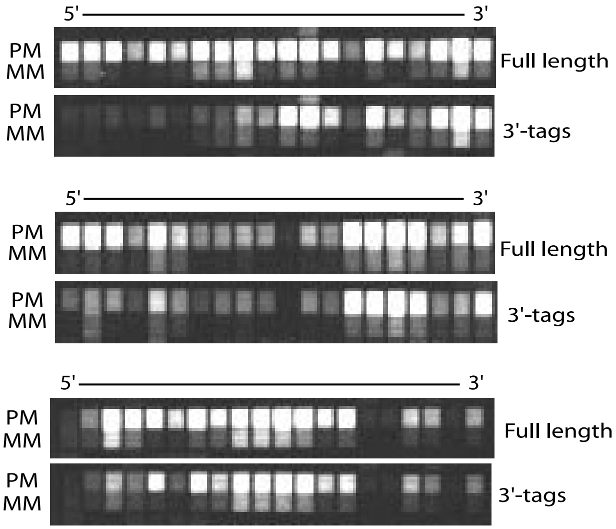

Figure 8. Hybridisation of full-length cDNA and 3'-end signature tags to Affymetrix^{TM} arrays yields different 5'-probe hybridisation patterns.

7. **Transcript profiling of adult neural stem cells**

7.1 Adult neural stem cells

Just a decade ago the adult brain was thought to be a fairly rigid, non-plastic structure that could not regenerate. Although mitotic activity in the brain had been observed as early as in the 1960's (Altman and Das 1965), the importance of this proliferation was not realised. Once dead, it was believed, the neurons of the brain and spinal cord could not be regenerated. This view was challenged in the mid-1990's when it was shown that cells isolated from crude dissections of the medullary centre of the adult mouse forebrain could proliferate in vitro and subsequently differentiate into neurons and astrocytes (Reynolds and Weiss 1992). It was not known, however, if the cells were precursor cells or if they had the broader potential of stem cells. The first evidence that the cells in the mouse striatum were stem cells was reported by Gritti *et al* in 1996 (Gritti, Parati et al. 1996). Since then a series of studies have shown that neural stem cells (NSCs) are present in the adult brain in several different mammalian species, including rat, mouse and human (Palmer, Takahashi et al. 1997) (Johansson, Momma et al. 1999) (Palmer, Schwartz et al. 2001). These NSCs contribute to the continuous replacement of neuronal cells within specific regions of the brain throughout adulthood.

Adult neural stem cells are multipotential, being able to differentiate into all the main cell types of the central nervous system; neurons, astrocytes and oligodendrocytes, while retaining the ability to multiply (self-renew). In recent years they have been proposed to exhibit even greater developmental ability (plasticity), also being able to differentiate into red and white blood cells (Bjornson, Rietze et al. 1999) skeletal muscle (Galli, Borello et al. 2000) and endothelial cells (Wurmser, Nakashima et al. 2004), and contribute to brain, heart, lung, stomach, intestine and liver tissues when transplanted into the developing embryo (Clarke, Johansson et al. 2000), although these findings have been highly debated. In addition, neural stem cells can be isolated from the adult mammalian brain and expanded *in vitro* (Gottlieb 2002). Taken together these characteristics of the adult neural stem cells, in addition to being remarkable and truly fascinating in their own right, have made them promising materials for use in cell replacement therapies of the central nervous system and possibly other tissues (Lindvall, Kokaia et al. 2004) (Uchida, Buck et al. 2000). They have thus attracted immense interest within the research community.

To be regarded as a stem cell a cell must possess two properties: (i) it has to be multipotent, that is able to give rise to all the different cell types of the organ in which it is located, and (ii) it must be able to self-renew for prolonged periods of time, in theory for the lifetime of the organism. Cells with both of these properties are found in the subventricular zone (SVZ) of the lateral ventricles in the mammalian brain and in the subgranular layer of the dentate gyrus of the hippocampus (McKay 1997) (Temple and Alvarez-Buylla 1999) (Momma, Johansson et al. 2000) (Gage 2000), although the stem cell nature of the latter population is uncertain (Morshead and van der Kooy 2004). The origin of the stem cells in the SVZ is not known, but since the SVZ is a remnant of the embryonic germinal neuroepithelium it is close at hand to believe that they are cells that are restrained from differentiating during development and thus retained in their original state (Clarke 2003). The identity of the NCS has not been fully established either. It has been proposed that they are differentiated ependymal cells, lining the ventricular wall, that express the intermediate filament protein nestin, as well as the astrocyte marker glial fibrillary acidic protein, GFAP (Johansson, Momma et al. 1999). The more prevalent view is that they are astrocyte-like cells in the underlying 2–3 cell layer thick subventricular zone, which also express GFAP (Doetsch, Caille et al. 1999) (Chiasson, Tropepe et al. 1999). An electron microscope study of mouse brain by Doetsch *et al* revealed that there are four main cell types in the SVZ (Figure 9) (Doetsch, Garcia-Verdugo et al. 1997). Type A cells are neuronal precursors, also known as neuroblasts, which migrate from the SVZ towards the olfactory bulb and express PSA-NCAM. Type B cells are slowly dividing/quiescent astrocytes that express GFAP. Type C cells are fast-proliferating, immature precursor cells that express Dlx2, and the fourth type are the ependymal cells. Doetsch *et al* proposed that the slowly dividing type B cells are the long-term self-renewal stem cells, which give rise to the more rapidly-proliferating precursor type C cells, which in turn give rise to the neuroblast type A cells. It is possible that both the subventricular and ependymal cell populations contain cells with stem cell properties, especially as they share some common properties (Johansson, Svensson et al. 1999) (Momma, Johansson et al. 2000). The new neurons that are born within the SVZ migrate along the rostral migratory stream, gradually differentiating until they integrate as fully mature neurons within the olfactory bulb.

In vivo, adult stem cells need to be tightly regulated to maintain tissue homeostasis and to participate in controlled regeneration after injury. This regulation is mediated by both cell-intrinsic and extrinsic factors, in what is known as the stem cell niche (Spradling, Drummond-Barbosa et al. 2001) (Doetsch 2003). Through a number of different factors the niche controls the stem cell's fate, determining whether it will proliferate, remain quiescent, become apoptotic or differentiate. Common components of a stem cell niche include connections to and signals from the surrounding somatic cells, extracellular matrices (ECMs) and the epigenetic state of the cell itself, controlling the signals the cell is able to respond to. The stem cell niche in the SVZ consists of the four cell types described above (A, B and C type and ependymal cells) and some additional structures. The astrocyte-like stem cells (type B) surround the type A neuroblasts and form a tube in which they can migrate towards the olfactory bulb. Rapidly dividing type C cells are scattered along the chain of migrating neuroblasts. Blood vessels are common in the SVZ and extend into a specialized basal lamina, which has extensive contacts with all types of SVZ types. Occasionally a type B astrocyte extends a process between the ependymal cells to contact the lateral ventricle.

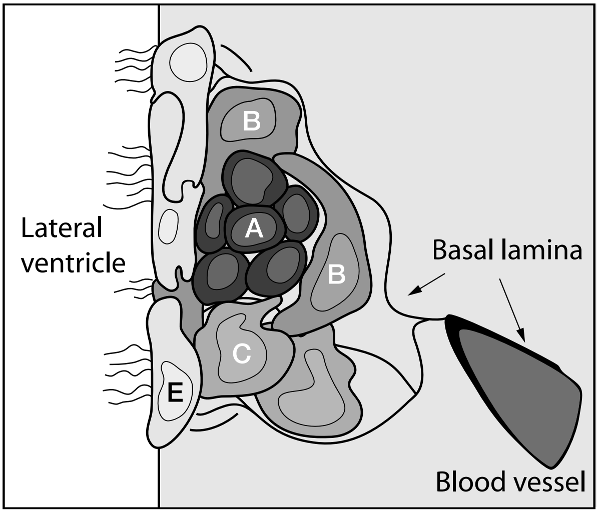

Figure 9. The neural stem cell niche. For details, see text.
A = type A cell (PSA-NCAM$^+$ neuroblast)
B = type B cell (GFAP$^+$, slowly dividing/quiescent)
C = type C cell (Dlx2$^+$, fast-proliferating)
E = ependymal cell
Picture modified from (Doetsch, Curr Opin Genet Dev 2003)

Thus, it seems that the NCS can receive signals from multiple sources. The ECM and cells surrounding the NCS are likely to transmit short-distance signals, while endothelial cells lining the blood vessels, the ependymal cells and stem cell protrusions connecting with the cerebrospinal fluid in the lateral ventricle are likely to transmit more long-distance signals. Stem cell can receive signals from various sources in their surrounding, including secreted signals such as chemokines and growth factors, and can also be influenced by diverse interactive cell-cell features, such as membrane-bound receptors and ligands. The ECM and basal lamina provide spatial cues within the stem cell niche, and are attached to carbohydrates that can potentiate ligand activity. Another important factor that influences the fate of the NSC is its epigenetic/chromatin state. By specific modification of the core histones the chromatin can be transformed between an "open", actively transcribed state, a repressed state or a silent state. In this way the cell can be restrained from expressing certain genes and gene families and thus become unable to respond to signals that it could respond to at an earlier stage. This provides a means for regulating development by, for instance, sequentially inhibiting and activating different transcription factors at different stages of differentiation (Doetsch 2003).

Some extracellular factors influencing the growth and fate of neural stem cells have been identified. These include both mitogens, that control cell proliferation, and morphogens, that control differentiation, and factors that act as both. They can also act in different ways, depending on the stage of development, the position within the central nervous system and their co-regulation with other factors. Some of these factors include epidermal growth factor (Egf), basic fibroblast growth factor (Fgf2), brain-derived neurotrophic factor (Bdnf) and transforming growth factor-α (Tgfa) (Rossi and Cattaneo 2002) (Lindvall, Kokaia et al. 2004), bone morphogenetic proteins (BMPs) and sonic hedgehog (SHH) (Panchision and McKay 2002), Notch (Frisen and Lendahl 2001), Wnt (Patapoutian and Reichardt 2000) and Ephrins (Holmberg and Frisen 2002).

One way to obtain further knowledge about these cells is to explore their gene expression profiles in different conditions and differentiation stages.

7.2 Transcript profiling of cultured adult neural stem cells (neurospheres) (III)

Adult neural stem cells can be isolated from the mature mammalian brain and expanded *in vitro* in the presence of growth factors, where they form floating aggregates of cells called neurospheres (Reynolds and Weiss 1992). Neurospheres are derived from one clonally expanded NSC or progenitor cell at a particular stage of maturation (Suslov, Kukekov et al. 2002). As the original NSC or progenitor cell proliferates new cells are formed that adhere to each other and gradually differentiate towards neural or glial fates. Neurospheres are thus complex structures that consist of multiple cell types of varying degrees of maturation. They have a dense extracellular matrix and the cells connect extensively through cell-cell contacts.

Electron-microscopy studies of rat fetal striatum EGF-expanded neurospheres (Lobo, Alonso et al. 2003) have shown that they consist of two types of cells, electron-dense and electron-lucent cells, both of which could be healthy, apoptotic or necrotic. Healthy cells were often connected to each other by adherence junctions and often displayed pseudopodia. They also engulfed and phagocytosed apoptotic and necrotic cells. Neurospheres from adult human brain have been similarly characterised, revealing the same type of heterogeneous, complex structure (Kukekov, Laywell et al. 1999).

After plating the neurospheres to solid support and withdrawing the growth factors, the cells of the neurospheres start differentiating into all neural cell types (neurons, astrocytes and oligodendrocytes), demonstrating that at least the initial cell of the sphere is multipotent (Gritti, Parati et al. 1996). They have thus gained attention as an *in vitro* model of neurogenesis. It is not known how well this *in vitro* process of neurogenesis resembles the situation *in vivo*, but challenging a cell *in vitro* will nevertheless unveil some of its developmental properties and potentials. By subjecting neurospheres to different microenvironments (e.g. by adding particular drugs or factors) one can discover factors and mechanisms that make them proliferate or differentiate in certain directions, e.g. to become neurons of a particular type. Neurospheres may also have the potential to become an important source of cells for cell replacement therapies for different neurological diseases. Therefore understanding their characteristics and the ways they can be manipulated is of great interest.

Paper III describes a basic transcript profiling study of neurospheres, designed to determine whether they could be used in future microarray studies. Since neurospheres are such heterogeneous structures, their heterogeneity might be reflected in their gene expression. Neurospheres have shown phenotypic variability both on the individual neurosphere level (Suslov, Kukekov et al. 2002) and the population level; spheres from different parts of the brain (Ostenfeld, Joly et al. 2002) and passaged different number of times (Morshead, Benveniste et al. 2002) showing different characteristics. It is also possible that gene expression may vary in neurospheres isolated from different animals and cultured in separate bottles, which could be confounded with treatment effects in future microarray studies. In order to determine the level of heterogeneity in gene expression between different neurosphere populations, neurospheres from different isolations, different passages and identical but separate culture flaks, were compared. Since it was desirable to limit the number of passages, limited numbers of cells and amounts of RNA were obtained, and thus the described amplification method was used.

Seven different comparisons were made. First the technical variation was evaluated, by co-hybridising two separately amplified 3'-end signature tags from the same RNA pool (technical replicates). Two comparisons were made to evaluate the variability in gene expression between identical, but separate cultures (culture replicates). One comparison was made to measure the difference between neurospheres from successive passages and two were made to measure the difference between neurospheres from different isolations (i.e. from separate animals). Finally, in order to compare the results from the neurosphere replicates with results from two truly different samples, and to confirm that we could detect true differentially expressed genes, undifferentiated neurospheres were compared to spheres induced to differentiate by withdrawing the growth factors from the culture medium, plating on a solid support and adding serum. Two replicates and two dye-swap hybridisations were performed for each comparison, giving four hybridisations for each comparison in total. The technical reproducibility and variation between the different neurosphere populations were evaluated by comparing the number and extent of differentially expressed genes identified in each comparison, visualised through a series of plots. The 3'-end signature amplification was shown to perform excellently and yielded highly reproducible results. Even with rather stringent criteria for identification of differentially expressed genes, no such genes were found for the technical replicates, and the correlation coefficient between the two replicate amplifications was as high as 0.99. As expected, the variability was higher for the different neurosphere replicates. Neurospheres from different isolations and different passages showed the largest variation in gene expression (181 and 383 differentially expressed genes, respectively. The large difference between different passages was a little surprising, since the two samples differed by only one passage (passage 1 vs. passage 2). Neurospheres from the same isolation and same passage, cultured in parallel, were not as divergent in their gene expression: one comparison yielding 27 and the other 82 differentially expressed genes. The magnitude of the differential expression was also relatively low. It was concluded that parallel cultures show some fluctuations in gene expression, but the difference is sufficiently low to enable identification of differentially expressed genes when neurospheres are exposed to different microenvironments. Furthermore, the biological noise could be reduced with by including more biological replicates in future experiments. The results from the comparison of undifferentiated neurospheres and serum-differentiated cells showed the highest number of differentially expressed genes (748). Many of the most highly differentially expressed genes were genes that were expected to be detected, e.g. myelin-related genes and genes related to transmitter substance signalling. Also, when the differentially expressed genes were grouped according to their biological function the neurospheres were found to be enriched in genes involved in the mitotic cell cycle, whereas the differentiated cells were enriched in genes involved in neurogenesis, synaptic transmission and development, as expected.

7.3 Transcript profiling of PACAP stimulated cultured adult neural stem cells (IV)

As previously mentioned, hopes have been raised that neural stem cells could be used as a new source of cells for cell-replacement therapies to treat various neurological diseases (Lindvall, Kokaia et al. 2004) (Uchida, Buck et al. 2000). Cell replacement with fetal mesencephalic or striatal tissue has previously been shown to lead to functional improvement in patients with Parkinson's disease and Huntington's disease (Bachoud-Levi, Remy et al. 2000) (Kordower, Freeman et al. 1995) (Bjorklund and Lindvall 2000). However, the use of fetal cells is hampered by a number of obstacles, in addition to ethical concerns. For instance, fetal tissue is only available in limited quantities, and the cells are mostly postmitotic and cannot be readily expanded. Furthermore, these cell populations are heterogeneous and their purity and viability cannot be reliably controlled. These attributes perhaps explain the variation in functional outcome observed after their transplantation. In contrast, stem cells represent sources of cells that are more readily obtainable, expandable and that could potentially be maintained as more homogeneous, pure cell populations.

The neural stem cells could either be induced to proliferate and differentiate endogenously, while still in the brain, or expanded and differentiated in culture before being transplanted into the damaged site. Both strategies have been evaluated with limited, but promising, success (Nakatomi, Kuriu et al. 2002) (Studer, Tabar et al. 1998). In order to achieve successful, functional recovery a deep understanding of the factors and mechanisms influencing NSC proliferation and differentiation, both *in vivo* and *in vitro*, is needed and effective strategies to isolate, expand and differentiate these cells into appropriate and specific phenotypes must be developed. Some factors that can stimulate proliferation and regulate neurogenesis have been discovered (see above), but too little is known about these processes as yet.

Recently it was reported that pituitary adenylate cyclase-activating polypeptide (PACAP) promotes neural stem cell proliferation both *in vivo* and *in vitro* (Mercer, Ronnholm et al. 2004). PACAP is a pleiotropic neuropeptide that belongs to the vasoactive intestinal peptide (VIP)/secretin/glucagon family of peptides (reviewed in (Arimura 1998) and (Vaudry, Gonzalez et al. 2000). It exists in two forms, PACAP27 and PACAP38, which share an identical 27-amino acid N-terminus. In the brain PACAP functions as a neurotransmitter and neuromodulator. In addition, it acts as neurotrophic factor that may play an important role during development, and appears to stimulate neuronal survival in the adult brain. PACAP also has important functions outside the central nervous system. For example, it stimulates release of insulin from pancreatic ß-cells, has a regulatory role in the maturation of germ cells, and also exerts important effects in the respiratory and cardiovascular systems. PACAP binds to three receptors of the VIP receptor family, a family of G-protein coupled receptors. PACAP and VIP both bind to the VIP receptors VIPR1 and VIPR2 while PACAP alone binds to the PACAP receptor 1 (PAC1). At least eight subtypes of PAC1 exist, resulting from alternative splicing, and each subtype is coupled to specific signalling pathways and functions. Five of the splice variants differ from one another by the absence (short variant) or presence of either one or two cassettes named hip and hop (27–28 amino acids long), present in the third cytoplasmic loop, the primary site of interaction between the receptor and G proteins. The different splice variants exhibit different patterns of adenylate cyclase and phospholipase C (PLC) stimulation.

The effects exerted by PACAP depend on the temporal and spatial distribution of PACAP and its different receptors. Both PACAP and PAC1 are widely expressed in the central nervous system. PAC1 is expressed at high levels in ventricular zones during the development of rodents, and in areas of neurogenesis in the adult central nervous system (Jaworski and Proctor 2000) (Mercer, Ronnholm et al. 2004), suggesting that PAC1 may be involved in adult neurogenesis. Furthermore, PAC1 is expressed in neural stem cells isolated from the lateral ventricle wall of adult mice and cultured *in vitro* as neurospheres (Mercer, Ronnholm et al. 2004). When supplied to the culture medium in the cited study, PACAP induced a pronounced proliferative response in the neurospheres. This effect appeared to involve the PLC and protein kinase C (PKC) signalling pathway.

Paper IV describes an experiment in which transcriptional changes underlying the proliferative effects of PACAP on neurospheres were investigated using microarrays and the described 3'-end signature tag amplification protocol. Three different comparisons were made, all of which used undifferentiated neurospheres, ordinarily cultured in medium supplied with epidermal growth factor (EGF), as a control sample. In the first comparison the control neurospheres were compared to neurospheres induced to proliferate by replacing the EGF-supplemented medium with PACAP-supplemented medium. In the second comparison, used as a differentiation control, the control neurospheres were compared to neurospheres induced to differentiate by replacing the EGF-supplemented medium with calf serum-supplemented medium and plating the spheres on a solid support. In the third comparison, used as a proliferation control, the control neurospheres were compared to neurospheres induced to proliferate by replacing the EGF-supplemented medium with medium supplemented with a small molecule acting as a transmembrane receptor agonist (TMR agonist). All samples were cultured in two parallel cultures in order to account for the biological noise seen in study III. In addition, control hybridisations were made with 3'-end signature tag replicates from one sample and with control neurospheres grown in parallel cultures, to evaluate the technical and biological noise in the set up.

The 3'-end signature tag amplification method again showed high reproducibility, while slightly noisier data were obtained from the parallel culture replicates, confirming the findings presented in Paper III. In contrast, the three treatment comparisons detected a much higher number of differentially expressed genes and greater magnitude of differential expression. 814 genes showed differential expression after PACAP treatment, 604 genes were differentially expressed in the neurosphere vs. serum differentiated control comparison, and 735 genes were differentially expressed in the neurosphere vs. TMR agonist proliferation control comparison. When comparing the three gene lists it was discovered that a majority of the genes (435) were differentially expressed in each of three the different comparisons. This was highly surprising, especially considering the major differences between proliferating and differentiating cells, which should theoretically be reflected at the gene expression level. A likely explanation is that the removal of EGF from the culture medium has a large effect on gene expression in the cells, leading to confounding results between the effect of EGF removal and the effect of the particular treatment. When the 435 common genes were grouped according to their biological functions an enrichment of genes involved in the cell cycle and DNA replication was seen in the EGF-supplemented control neurospheres, while themes such as neurogenesis, organogenesis and development were overrepresented in the treated, EGF-withdrawn samples, further providing further support for the hypothesis. Consequently, when studying the effects of different substances on neurospheres these findings should be taken into account in future neurosphere studies.

8. **Concluding remarks**

The objective of the work presented in this thesis was to develop, evaluate and apply a new method for cDNA amplification, enabling transcript analysis of small tissue samples such as laser capture microdissected tissues or rare cells that cannot be readily expanded in culture. The method is based on size normalisation of the transcriptome prior to exponential amplification, to avoid size biased artefacts. The method has shown excellent performance in all studies, yielding transcript profiles that are similar to those obtained with unamplified material as well as highly reproducible data. The use of 3'-end cDNA tags enables reliable identification of the expressed genes. However, the method cannot be used for studies requiring full-length cDNA sequences, such as investigations of alternative splicing, where methods such as SMART™ might be more appropriate. The developed method is inexpensive, easy to perform and can be used for high-throughput purposes. Recently this protocol has been automated using magnetic bead robotics that further minimises manual variations and misstakes. Up to 48 samples can be amplified in parallel, also facilitating larger biological studies that rely on amplification protocols.

In the work underlying this thesis the amplification method was applied to stem cell biology. It was first used to show that cultured neural stem cells (neurospheres) display substantial differences in gene expression when isolated from different animals or passaged to different degrees. However, if careful experimental design is applied, expanding them under identical conditions and using appropriate biological replicates, they are suitable for gene expression analysis. It was also shown that withdrawing epidermal growth factor (EGF) from the culture medium has profound effects on gene expression, which should be taken into consideration in future neurosphere studies.

Acknowledgements

Jag skulle varmt vilja tacka en mängd människor som gjort det möjligt för mig att få ihop denna avhandling, på ett jobbmässigt och, inte minst, mänskligt plan:

Joakim Lundeberg, min handledare, för att du, även om du håller på att gå under själv för att du har så mycket att göra, alltid har tid. Och för att du alltid lyckas se och ta fram det positiva i resultat som inte är så uppenbart bra till en början ☺. Tack också för ditt personliga engagemang i dina doktorander och för många trevliga middagar, nyårsaftnar, semestrar mm.

Mathias Uhlén, för din entusiasm och postitiva anda, som sprider sig ner genom hela DNA corner.

Stefan Ståhl, för din smittande glädje och värme. Det var din öppna famn då jag hade det lite tufft som fick mig att stanna kvar på KTH!

ALLA andra i DNA corner, gamla som nya, för att ni gör det till ett så fantastiskt ställe att jobba på! Ett särskilt stort tack till:

Sophia, Per-Åke och alla andra PI's för ert engagemang i både forskningen och era doktorander, och för att ni skapar en arbetsplats som präglas av entusiasm, strålande forskning, hjälpsamhet och en massa socialt skoj.

Christin, Cissi, Fredrik Sterky, Jacob och Åsa för strålande handledarskap under ex-jobb, sommarjobb och andra projekt!

Blåkullagänget; Lotta, Tove, Stina och senare Kicki och Anders, för fantastiska pysselkvällar med påskkärringar, kycklingar, spöken och pumpor, fågelfrukostar, glöggkvällar, konferensresor (inklusive strålande shopping) och andra galna påhitt. Ni har verkligen gjort det roligt att gå till jobbet!

Microarraygruppen; Peter, Valtteri, Anders, Tove, Henrik Aspeborg, Annelie, Angela, Anna W, Cecilia, Johan, Marcus, Esther, Afshin, Emmelie, Max, Jenny, Fredrik och Christian för allt samarbete, hjälp och spännande diskussioner!

Alla forskningsingenjörer och tekniker; Kicki, Annelie, Bahram, Anders Holmberg, Anna W, Angela, Ulla, Jenny, Annika, Lasse, Fredrik, Tobbe, Pia, Rebecca, Emma m.fl under åren, administrativ personal; Monica, Mona, Tina och Inger och datorsupport; Erik, Anders Michael och Danko, för att ni bidrar med andra perspektiv och för allt ovärderligt arbete ni gör. Utan er skulle det inte fungera! Ett särskilt stort tack till Anna Westring för allt lysande arbete med stamcellsprojekten, Bahram för din stora hjälpsamhet med sekvensning och annat,och Angela som kämpade länge med C-peptidklonerna (även om de inte hann komma med i slutänden...).

Esther och Johan, för allt ert arbete i nämnda C-peptidprojekt och för att ni var så trevliga att ha som ex-jobbare!

Luciakören; det har varit SÅ roligt och SÅ mysigt!

Ridgänget; Kicki, Malin, Marko och Lisa, för en väldigt rolig, om än alldeles för kort, tid!

Lunchgänget; Åsa, Cissi, Cristina och Lisa för alla roliga och avkopplande små diskussisoner som varit så välbehövliga mitt på da'n!

Mina rumskamrater; Anna B, Cilla, Tobbe, Mikaela och Caroline för att ni haft sådant tålamod och visat sådan hänsyn under de här månaderna av avhandlingsskrivande!

Och alla andra för er kreativitet och energi både när det gäller forskning och saker som semmelbak, julfestspynt, -lekar och –uppträdanden, och andra sociala tillställningar!

Alla samarbetspartners i Umeå, på Astra Zeneca, NeuroNova och Karolinska sjukhuset under åren; Ett särskilt tack till Anders Thelin, för all hjälp med Affymetrix-försöken, även när vi var tvungna att labba mitt i natten, och för att det alltid känts så positivt att komma ner till Göteborg! Tack också Alex, för all hjälp med upplägg av försök och pek-skrivande, och Zhihui och Alexandra för all tid ni lagt ner på cellodling!

Alla mina medförfattare. Ett särskilt stort jättetack till Valtteri för all hjälp med dataanalys, för spännande diskussioner och för att du ägnade så mycket tid åt sista peket, annars hade det inte gått!

Mina vänner utanför jobbet;
Universitetsgänget; Jessica, Maria, Eva och Emmelie för alla roliga år under utbildningen, och för att det fortfarande är så himla mysigt att träffa er! Jessica, för att du är en fantastiskt vän som alltid har tid och alltid lyssnar, och för alla barnvagnspromenader! Maria, för att du drog med mig på mitt livs äventyr till Afrika – jag blir fortfarande alldeles glad när jag tänker på det. Och, naturligtvis, för att du är en sådan bra och sprudlande kompis!

Ullis och Sofia, för att ni varit ett så enormt stöd då tillvaron känts lite tyngre, och för allt skoj och allt allvar vi upplevt under våra resor till Frankrike och Italien! Tack även Ted för all din värme, alla spännande samtal och för att du är en av mina största inspirationskällor!

Lotta, jeanskepan 2 ☺, för alla roliga och tokiga saker vi gjort och upplevt. Även om vi inte ses så ofta längre uppskattar jag verkligen din humor och din värme!

Anna B, för allt roligt under forskarskoletiden, alla värmande samtal och för att det är så kul att träffa dig och Mats!

Maria och Joel, för alla fikastunder, barnprat och för att våra döttrar är så fantastiskt söta tillsammans!

Anna G, för att det känts så kul att gå till jobbet när du är där, för alla roliga saker vi gjort under konferenser, middagar, fester, målarkvällar, skulpturdagar, ridturer etc, och, inte minst, för allt tjejsnack som rört allt här i livet!

Åsa, inte bara för att du är världens bästa svägerska, utan för att du även är världens bästa vän. Hade nog inte klarat tillvaron utan allt vi gjort och utan alla samtal med dig!

Lotta, för att din glädje och omtänksamhet fick mig på fötter igen! Tack för allt genom de här åren, för allt du dragit med mig på, för både din och din familjs glädje, värme och öppenhet och för en massa massa annat ☺.

Min kära familj; Mamma och Rolf, för att ni alltid ställt upp (inte minst med passning av Tyra under den här tiden!) och gett ert stöd och uppmuntran, och för att det alltid är så skönt att komma hem. Morfar, för att du alltid trodde på mig även om du inte alltid förstod vad det var bra för det jag gjorde. Micke, för att du är världens bästa (om än tröttaste) storebror, som alltid ställt upp då jag verkligen behövt det! Magnus, för att din glädje och positiva attityd, din nyfikenhet och dina funderingar över allt här i livet fortfarande ger mig inspiration!

Min extra-syster ☺, Jenny, för fantastisk hjälp med layoutandet av denna avhandling, och för att du är så mysig!

Peter, Gaby, Sara, Moa och Anders, för att ni är en så go' och glad familj med stor,
öppen famn!

Min nya, lilla familj; Tyra, min lilla solstråle, för att det inte finns något ljuvligare än att komma hem till dina uppsträckta armar. Lillen, som hållt mig sällskap under långa och stressiga skrivstunder. Niklas, för hjälp med framsida och annan grafik, för att du fått lov att vara vaken hela nätter och nästan stupat för att jag skulle få ihop det här, och för att du är den du är! Mitt allt!

References

(1999). *Nature genetics* 21(suppl): 1-60.

(2002). *Nature genetics* 32(suppl): 461-552.

(2004). Finishing the euchromatic sequence of the human genome. *Nature* 431(7011): 931-45.

Adams, M. D., M. Dubnick, et al. (1992). Sequence identification of 2,375 human brain genes. *Nature* 355(6361): 632-4.

Adams, M. D., J. M. Kelley, et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252(5013): 1651-6.

Adams, M. D., A. R. Kerlavage, et al. (1993). 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nat Genet* 4(3): 256-67.

Adler, K., J. Broadbent, et al. (2000). MICROMAX™: A highly sensitive system for differential gene expression on microarrays. Microarray Biochip Technology. M. Scheena. Natick, MA, Eaton Publication: 221-230.

Alizadeh, A. A., M. B. Eisen, et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403(6769): 503-11.

Alter, O., P. O. Brown, et al. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A* 97(18): 10101-6.

Altman, J. and G. D. Das (1965). Autoradiographic and histological evidence of postnatal hippocampal neurogenesis in rats. *J Comp Neurol* 124(3): 319-35.

Alwine, J. C., D. J. Kemp, et al. (1979). Detection of specific RNAs or specific fragments of DNA by fractionation in gels and transfer to diazobenzyloxymethyl paper. *Methods Enzymol* 68: 220-42.

Andersson, T., P. Unneberg, et al. (2002). Monitoring of representational difference analysis subtraction procedures by global microarrays. *Biotechniques* 32(6): 1348-50, 1352, 1354-6, 1358.

Aoyagi, K., T. Tatsuta, et al. (2003). A faithful method for PCR-mediated global mRNA amplification and its integration into microarray analysis on laser-captured cells. *Biochem Biophys Res Commun* 300(4): 915-20.

Arimura, A. (1998). Perspectives on pituitary adenylate cyclase activating polypeptide (PACAP) in the neuroendocrine, endocrine, and nervous systems. *Jpn J Physiol* 48(5): 301-31.

Ashburner, M., C. A. Ball, et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1): 25-9.

Audic, S. and J. M. Claverie (1997). The significance of digital gene expression profiles. *Genome Res* 7(10): 986-95.

Augenlicht, L. H., M. Z. Wahrman, et al. (1987). Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro. *Cancer Res* 47(22): 6017-21.

Avery, O. T., C. M. MacLeod, et al. (1944). *Journal of Experimental Medicine* 79: 137-158.

Bachoud-Levi, A. C., P. Remy, et al. (2000). Motor and cognitive improvements in patients with Huntington's disease after neural transplantation. *Lancet* 356(9246): 1975-9.

Barrett, J. C. and E. S. Kawasaki (2003). Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. *Drug Discov Today* 8(3): 134-41.

Baugh, L. R., A. A. Hill, et al. (2001). Quantitative analysis of mRNA amplification by in vitro transcription. *Nucleic Acids Res* 29(5): E29.

Beaucage, S. L. (2001). Strategies in the preparation of DNA oligonucleotide arrays for diagnostic applications. *Curr Med Chem* 8(10): 1213-44.

Bengtsson, H. aroma - An R Object-oriented Microarray Analysis environment. [http://www.maths.lth.se/publications/].

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

Bernstein, J. A., A. B. Khodursky, et al. (2002). Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci U S A* 99(15): 9697-702.

Bevilacqua, A., M. C. Ceriani, et al. (2003). Post-transcriptional regulation of gene expression by degradation of messenger RNAs. *J Cell Physiol* 195(3): 356-72.

Bishop, J. O., J. G. Morton, et al. (1974). Three abundance classes in HeLa cell messenger RNA. *Nature* 250(463): 199-204.

Bjorklund, A. and O. Lindvall (2000). Cell replacement therapies for central nervous system disorders. *Nat Neurosci* 3(6): 537-44.

Bjornson, C. R., R. L. Rietze, et al. (1999). Turning brain into blood: a hematopoietic fate adopted by adult neural stem cells in vivo. *Science* 283(5401): 534-7.

Boheler, K. R. and M. D. Stern (2003). The new role of SAGE in gene discovery. *Trends Biotechnol* 21(2): 55-7; discussion 57-8.

Brady, G. and N. N. Iscove (1993). Construction of cDNA libraries from single cells. *Methods Enzymol* 225: 611-23.

Brandenberger, R., I. Khrebtukova, et al. (2004). MPSS profiling of human embryonic stem cells. *BMC Dev Biol* 4(1): 10.

Brazma, A., P. Hingamp, et al. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 29(4): 365-71.

Brenner, S., M. Johnson, et al. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 18(6): 630-4.

Brown, C. S., P. C. Goodwin, et al. (2001). Image metrics in the statistical analysis of DNA microarray data. *Proc Natl Acad Sci U S A* 98(16): 8944-9.

Brown, M. P., W. N. Grundy, et al. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A* 97(1): 262-7.

Bulyk, M. L., X. Huang, et al. (2001). Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc Natl Acad Sci U S A* 98(13): 7158-63.

Bussemaker, H. J., H. Li, et al. (2001). Regulatory element detection using correlation with expression. *Nat Genet* 27(2): 167-71.

Castle, J., P. Garrett-Engele, et al. (2003). Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. *Genome Biol* 4(10): R66.

Cawley, S., S. Bekiranov, et al. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116(4): 499-509.

Chen, J., M. Sun, et al. (2002). Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. *Proc Natl Acad Sci U S A* 99(19): 12257-62.

Chen, J. J., J. D. Rowley, et al. (2000). Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. *Proc Natl Acad Sci U S A* 97(1): 349-53.

Chiasson, B. J., V. Tropepe, et al. (1999). Adult mammalian forebrain ependymal and subependymal cells demonstrate proliferative potential, but only subependymal cells have neural stem cell characteristics. *J Neurosci* 19(11): 4462-71.

Cho, R. J., M. J. Campbell, et al. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 2(1): 65-73.

Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nat Genet* 32 Suppl: 490-5.

Clark, T. A., C. W. Sugnet, et al. (2002). Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* 296(5569): 907-10.

Clarke, D. L. (2003). Neural stem cells. *Bone Marrow Transplant* 32 Suppl 1: S13-7.

Clarke, D. L., C. B. Johansson, et al. (2000). Generalized potential of adult neural stem cells. *Science* 288(5471): 1660-3.

Cope, L. M., R. A. Irizarry, et al. (2004). A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* 20(3): 323-31.

Crick, F. H. (1958). The biological replication of Macromolecules. *Symp Soc Exp Biol* XII: 138.

Datson, N. A., J. van der Perk-de Jong, et al. (1999). MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue. *Nucleic Acids Res* 27(5): 1300-7.

Debouck, C. (1995). Differential display or differential dismay? *Current Opinion in Biotechnology* 6: 597-599.

DeRisi, J. L., V. R. Iyer, et al. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278(5338): 680-6.

Diatchenko, L., Y. F. Lau, et al. (1996). Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc Natl Acad Sci U S A* 93(12): 6025-30.

Dobbin, K., J. H. Shih, et al. (2003). Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. *J Natl Cancer Inst* 95(18): 1362-9.

Doetsch, F. (2003). A niche for adult neural stem cells. *Curr Opin Genet Dev* 13(5): 543-50.

Doetsch, F., I. Caille, et al. (1999). Subventricular zone astrocytes are neural stem cells in the adult mammalian brain. *Cell* 97(6): 703-16.

Doetsch, F., J. M. Garcia-Verdugo, et al. (1997). Cellular composition and three-dimensional organization of the subventricular germinal zone in the adult mammalian brain. *J Neurosci* 17(13): 5046-61.

Doniger, S. W., N. Salomonis, et al. (2003). MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol* 4(1): R7.

Dudley, A. M., J. Aach, et al. (2002). Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc Natl Acad Sci U S A* 99(11): 7554-9.

Dudoit, S., J. Fridlyand, et al. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* 97: 77-87.

Dudoit, S., Y. H. Yang, et al. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 12: 111-140.

Duguid, J. R., R. G. Rohwer, et al. (1988). Isolation of cDNAs of scrapie-modulated RNAs by subtractive hybridization of a cDNA library. *Proc Natl Acad Sci U S A* 85(15): 5738-42.

Eberwine, J., H. Yeh, et al. (1992). Analysis of gene expression in single live neurons. *Proc Natl Acad Sci U S A* 89(7): 3010-4.

Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nat Rev Genet* 2(12): 919-29.

Efron, B., R. Tibshirani, et al. (2001). Empirical Bayes analysis of a micoarray experiment. *Journal of the American Statistical Association* 96: 1151-1160.

Ehrenhofer-Murray, A. E. (2004). Chromatin dynamics at DNA replication, transcription and repair. *Eur J Biochem* 271(12): 2335-49.

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

Ekins, R., F. Chu, et al. (1989). High specific activity chemiluminescent and fluorescent markers: their potential application to high sensitivity and 'multi-analyte' immunoassays. *J Biolumin Chemilumin* 4(1): 59-78.

Ekins, R. and F. W. Chu (1999). Microarrays: their origins and applications. *Trends Biotechnol* 17(6): 217-8.

Ekins, R. P. (1989). Multi-analyte immunoassay. *J Pharm Biomed Anal* 7(2): 155-68.

Endege, W. O., K. E. Steinmann, et al. (1999). Representative cDNA libraries and their utility in gene expression profiling. *Biotechniques* 26(3): 542-8, 550.

Fan, J., X. Yang, et al. (2002). Global analysis of stress-regulated mRNA turnover by using cDNA arrays. *Proc Natl Acad Sci U S A* 99(16): 10611-6.

Finkelstein, D. B., J. Gollub, et al. (2001). Iterative linear regression by sector. Methods of microarray data analysis. Papers from CAMDA 2000. S. M. Lin and K. F. Johnson, Kluwer Academic: 57-68.

Fodor, S. P., J. L. Read, et al. (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science* 251(4995): 767-73.

Forster, T., D. Roy, et al. (2003). Experiments using microarray technology: limitations and standard operating procedures. *J Endocrinol* 178(2): 195-204.

Frisen, J. and U. Lendahl (2001). Oh no, Notch again! *Bioessays* 23(1): 3-7.

Futcher, B., G. I. Latter, et al. (1999). A sampling of the yeast proteome. *Mol Cell Biol* 19(11): 7357-68.

Gage, F. H. (2000). Mammalian neural stem cells. *Science* 287(5457): 1433-8.

Gall, J. G. and M. L. Pardue (1969). Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *Proc Natl Acad Sci U S A* 63(2): 378-83.

Galli, R., U. Borello, et al. (2000). Skeletal myogenic potential of human and mouse neural stem cells. Nat Neurosci 3(10): 986-91.

Gautier, L., L. Cope, et al. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20(3): 307-15.

Gentleman, R., V. Carey, et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* 5(10): R80.

Geschwind, D. H., J. Ou, et al. (2001). A genetic analysis of neural progenitor differentiation. *Neuron* 29(2): 325-39.

Gibson, U. E., C. A. Heid, et al. (1996). A novel method for real time quantitative RT-PCR. *Genome Res* 6(10): 995-1001.

Glasbey, C. A. and P. Ghazal (2003). Combinatorial image analysis of DNA microarray features. *Bioinformatics* 19(2): 194-203.

Golub, T. R., D. K. Slonim, et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286(5439): 531-7.

Gonzalez, P., J. S. Zigler, Jr., et al. (1999). Identification and isolation of differentially expressed genes from very small tissue samples. *Biotechniques* 26(5): 884-6, 888-92.

Gottlieb, D. I. (2002). Large-scale sources of neural stem cells. *Annu Rev Neurosci* 25: 381-407.

Graveley, B. R. (2001). Alternative splicing: increasing diversity in the proteomic world. *Trends Genet* 17(2): 100-7.

Griffin, T. J., S. P. Gygi, et al. (2002). Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. *Mol Cell Proteomics* 1(4): 323-33.

Gritti, A., E. A. Parati, et al. (1996). Multipotential stem cells from the adult mouse brain proliferate and self-renew in response to basic fibroblast growth factor. *J Neurosci* 16(3): 1091-100.

Gurskaya, N. G., L. Diatchenko, et al. (1996). Equalizing cDNA subtraction based on selective suppression of polymerase chain reaction: cloning of Jurkat cell transcripts induced by phytohemaglutinin and phorbol 12-myristate 13-acetate. *Anal Biochem* 240(1): 90-7.

Gygi, S. P., Y. Rochon, et al. (1999). Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* 19(3): 1720-30.

Hara, E., T. Kato, et al. (1991). Subtractive cDNA cloning using oligo(dT)30-latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells. *Nucleic Acids Res* 19(25): 7097-104.

He, L. and G. J. Hannon (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet* 5(7): 522-31.

Hedrick, S. M., D. I. Cohen, et al. (1984). Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. *Nature* 308(5955): 149-53.

Heid, C. A., J. Stevens, et al. (1996). Real time quantitative PCR. *Genome Res* 6(10): 986-94.

Holmberg, J. and J. Frisen (2002). Ephrins are not only unattractive. *Trends Neurosci* 25(5): 239-43.

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

Hoth, S., Y. Ikeda, et al. (2003). Monitoring genome-wide changes in gene expression in response to endogenous cytokinin reveals targets in Arabidopsis thaliana. *FEBS Lett* 554(3): 373-80.

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

Hu, L., J. Wang, et al. (2002). Obtaining reliable information from minute amounts of RNA using cDNA microarrays. *BMC Genomics* 3(1): 16.

Hubank, M. and D. G. Schatz (1994). Identifying differences in mRNA expression by representational difference analysis of cDNA. *Nucleic Acids Res* 22(25): 5640-8.

Hudson, T. J., L. D. Stein, et al. (1995). An STS-based map of the human genome. *Science* 270(5244): 1945-54.

Hughes, T. R., M. Mao, et al. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* 19(4): 342-7.

Hughes, T. R., M. J. Marton, et al. (2000). Functional discovery via a compendium of expression profiles. *Cell* 102(1): 109-26.

Iscove, N. N., M. Barbara, et al. (2002). Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. *Nat Biotechnol* 20(9): 940-3.

Ishii, M., S. Hashimoto, et al. (2000). Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics* 68(2): 136-43.

Iyer, V. R., C. E. Horak, et al. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409(6819): 533-8.

Jaworski, D. M. and M. D. Proctor (2000). Developmental regulation of pituitary adenylate cyclase-activating polypeptide and PAC(1) receptor mRNA expression in the rat central nervous system. Brain Res Dev Brain Res 120(1): 27-39.

Jenssen, T. K., M. Langaas, et al. (2002). Analysis of repeatability in spotted cDNA microarrays. *Nucleic Acids Res* 30(14): 3235-44.

Johansson, C. B., S. Momma, et al. (1999). Identification of a neural stem cell in the adult mammalian central nervous system. *Cell* 96(1): 25-34.

Johansson, C. B., M. Svensson, et al. (1999). Neural stem cells in the adult human brain. *Exp Cell Res* 253(2): 733-6.

Johnson, J. M., J. Castle, et al. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302(5653): 2141-4.

Jongeneel, C. V., C. Iseli, et al. (2003). Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. *Proc Natl Acad Sci U S A* 100(8): 4702-5.

Kadam, S. and B. M. Emerson (2002). Mechanisms of chromatin assembly and transcription. *Curr Opin Cell Biol* 14(3): 262-8.

Kamme, F., R. Salunga, et al. (2003). Single-cell microarray analysis in hippocampus CA1: demonstration and validation of cellular heterogeneity. *J Neurosci* 23(9): 3607-15.

Kapranov, P., S. E. Cawley, et al. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296(5569): 916-9.

Karsten, S. L., V. M. Van Deerlin, et al. (2002). An evaluation of tyramide signal amplification and archived fixed and frozen tissue in microarray gene expression analysis. *Nucleic Acids Res* 30(2): E4.

Kashiwagi, H. and K. Uchida (2000). Genome-wide profiling of gene amplification and deletion in cancer. *Hum Cell* 13(3): 135-41.

Kendziorski, C., R. A. Irizarry, et al. (2004). To pool or not to pool: a question of microarray experimental design. Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 46.

Kepler, T. B., L. Crosby, et al. (2002). Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biol* 3(7): RESEARCH0037.

Kerr, M. K. and G. A. Churchill (2001). Experimental design for gene expression microarrays. *Biostatistics* 2(2): 183-201.

Kerr, M. K. and G. A. Churchill (2001). Statistical design and the analysis of gene expression microarray data. *Genet Res* 77(2): 123-8.

Kerr, M. K., M. Martin, et al. (2000). Analysis of variance for gene expression microarray data. *J Comput Biol* 7(6): 819-37.

Kordower, J. H., T. B. Freeman, et al. (1995). Neuropathological evidence of graft survival and striatal reinnervation after the transplantation of fetal mesencephalic tissue in a patient with Parkinson's disease. *N Engl J Med* 332(17): 1118-24.

Kukekov, V. G., E. D. Laywell, et al. (1999). Multipotent stem/progenitor cells with similar properties arise from two neurogenic regions of adult human brain. *Exp Neurol* 156(2): 333-44.

Lander, E. S., L. M. Linton, et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409(6822): 860-921.

Lee, M. L., F. C. Kuo, et al. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci U S A* 97(18): 9834-9.

Lennon, G. G. and H. Lehrach (1991). Hybridization analyses of arrayed cDNA libraries. *Trends Genet* 7(10): 314-7.

Levine, M. and R. Tjian (2003). Transcription regulation and animal diversity. *Nature* 424(6945): 147-51.

Liang, F., I. Holt, et al. (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat Genet* 25(2): 239-40.

Liang, P. (1998). Factors ensuring successful use of differential display. *Methods* 16(4): 361-4.

Liang, P. (2002). A decade of differential display. *Biotechniques* 33(2): 338-44, 346.

Liang, P. and A. B. Pardee (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* 257(5072): 967-71.

Lindvall, O., Z. Kokaia, et al. (2004). Stem cell therapy for human neurodegenerative disorders-how to make it work. *Nat Med* 10 Suppl: S42-50.

Lipshutz, R. J., S. P. Fodor, et al. (1999). High density synthetic oligonucleotide arrays. *Nat Genet* 21(1 Suppl): 20-4.

Lisitsyn, N. and M. Wigler (1993). Cloning the differences between two complex genomes. *Science* 259(5097): 946-51.

Lobo, M. V., F. J. Alonso, et al. (2003). Cellular characterization of epidermal growth factor-expanded free-floating neurospheres. *J Histochem Cytochem* 51(1): 89-103.

Lockhart, D. J., H. Dong, et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 14(13): 1675-80.

Long, A. D., H. J. Mangalam, et al. (2001). Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. *J Biol Chem* 276(23): 19937-44.

Lundgren, M., A. Andersson, et al. (2004). Three replication origins in Sulfolobus species: synchronous initiation of chromosome replication and asynchronous termination. *Proc Natl Acad Sci U S A* 101(18): 7046-51.

Luo, L., R. C. Salunga, et al. (1999). Gene expression profiles of laser-captured adjacent neuronal subtypes. *Nat Med* 5(1): 117-22.

Lönnstedt, I. and T. Speed (2002). Replicated microarray data. *Statistica Sinica* 12: 31-46.

Makrigiorgos, G. M., S. Chakrabarti, et al. (2002). A PCR-based amplification method retaining the quantitative difference between two complex genomes. *Nat Biotechnol* 20(9): 936-9.

Maniatis, T. and B. Tasic (2002). Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* 418(6894): 236-43.

Marra, M. A., L. Hillier, et al. (1998). Expressed sequence tags--ESTablishing bridges between genomes. *Trends Genet* 14(1): 4-7.

Matsubara, K. and K. Okubo (1993). cDNA analyses in the human genome project. *Gene* 135(1-2): 265-74.

Mattick, J. S. (2004). RNA regulation: a new genetics? *Nat Rev Genet* 5(4): 316-23.

Matz, M. V. and S. A. Lukyanov (1998). Different strategies of differential display: areas of application. *Nucleic Acids Res* 26(24): 5537-43.

Maxam, A. M. and W. Gilbert (1977). A new method for sequencing DNA. *Proc Natl Acad Sci U S A* 74(2): 560-4.

McClelland, M., F. Mathieu-Daude, et al. (1995). RNA fingerprinting and differential display using arbitrarily primed PCR. *Trends Genet* 11(6): 242-6.

McCulloch, R. K., C. S. Choong, et al. (1995). An evaluation of competitor type and size for use in the determination of mRNA by competitive PCR. *PCR Methods Appl* 4(4): 219-26.

McKay, R. (1997). Stem cells in the central nervous system. *Science* 276(5309): 66-71.

Mercer, A., H. Ronnholm, et al. (2004). PACAP promotes neural stem cell proliferation in adult mouse brain. *J Neurosci Res* 76(2): 205-15.

Meyers, B. C., S. S. Tej, et al. (2004). The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. *Genome Res* 14(8): 1641-53.

Meyers, B. C., T. H. Vu, et al. (2004). Analysis of the transcriptional complexity of Arabidopsis thaliana by massively parallel signature sequencing. *Nat Biotechnol* 22(8): 1006-11.

Mironov, A. A., J. W. Fickett, et al. (1999). Frequent alternative splicing of human genes. *Genome Res* 9(12): 1288-93.

Modrek, B., A. Resch, et al. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res* 29(13): 2850-9.

Momma, S., C. B. Johansson, et al. (2000). Get to know your stem cells. *Curr Opin Neurobiol* 10(1): 45-9.

Morshead, C. M., P. Benveniste, et al. (2002). Hematopoietic competence is a rare property of neural stem cells that may depend on genetic and epigenetic alterations. *Nat Med* 8(3): 268-73.

Morshead, C. M. and D. van der Kooy (2004). Disguising adult neural stem cells. *Curr Opin Neurobiol* 14(1): 125-31.

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

Nakatomi, H., T. Kuriu, et al. (2002). Regeneration of hippocampal pyramidal neurons after ischemic brain injury by recruitment of endogenous neural progenitors. *Cell* 110(4): 429-41.

Newton, M. A., C. M. Kendziorski, et al. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol* 8(1): 37-52.

Nissim-Rafinia, M. and B. Kerem (2002). Splicing regulation as a potential genetic modifier. *Trends Genet* 18(3): 123-7.

Ogata, H., S. Goto, et al. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 27(1): 29-34.

Okubo, K., N. Hori, et al. (1992). Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat Genet* 2(3): 173-9.

Okubo, K. and K. Matsubara (1997). Complementary DNA sequence (EST) collections and the expression information of the human genome. *FEBS Lett* 403(3): 225-9.

Ostenfeld, T., E. Joly, et al. (2002). Regional specification of rodent and human neurospheres. *Brain Res Dev Brain Res* 134(1-2): 43-55.

Palmer, T. D., P. H. Schwartz, et al. (2001). Cell culture. Progenitor cells from human brain after death. *Nature* 411(6833): 42-3.

Palmer, T. D., J. Takahashi, et al. (1997). The adult rat hippocampus contains primordial neural stem cells. *Mol Cell Neurosci* 8(6): 389-404.

Pan, W., J. Lin, et al. (2002). How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biol* 3(5): research0022.

Panchision, D. M. and R. D. McKay (2002). The control of neural stem cells by morphogenic signals. *Curr Opin Genet Dev* 12(4): 478-87.

Patapoutian, A. and L. F. Reichardt (2000). Roles of Wnt proteins in neural development and maintenance. *Curr Opin Neurobiol* 10(3): 392-9.

Patil, N., A. J. Berno, et al. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294(5547): 1719-23.

Pease, A. C., D. Solas, et al. (1994). Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci U S A* 91(11): 5022-6.

Pennisi, E. (2003). Bioinformatics. Gene counters struggle to get the right answer. *Science* 301(5636): 1040-1.

Petalidis, L., S. Bhattacharyya, et al. (2003). Global amplification of mRNA by template-switching PCR: linearity and application to microarray analysis. *Nucleic Acids Res* 31(22): e142.

Peters, D. G., A. B. Kassam, et al. (1999). Comprehensive transcript analysis in small quantities of mRNA by SAGE-lite. *Nucleic Acids Res* 27(24): e39.

Picard-Riera, N., B. Nait-Oumesmar, et al. (2004). Endogenous adult neural stem cells: limits and potential to repair the injured central nervous system. *J Neurosci Res* 76(2): 223-31.

Puskas, L. G., A. Zvara, et al. (2002). RNA amplification results in reproducible microarray data with slight ratio bias. *Biotechniques* 32(6): 1330-4, 1336, 1338, 1340.

Quackenbush, J. (2001). Computational analysis of microarray data. *Nat Rev Genet* 2(6): 418-27.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nat Genet* 32 Suppl: 496-501.

R Development Core Team R: A language and environment for statistical computing. [http://www.R-project.org].

Raghuraman, M. K., E. A. Winzeler, et al. (2001). Replication dynamics of the yeast genome. *Science* 294(5540): 115-21.

Ralph, D., M. McClelland, et al. (1993). RNA fingerprinting using arbitrarily primed PCR identifies differentially regulated RNAs in mink lung (Mv1Lu) cells growth arrested by transforming growth factor beta 1. *Proc Natl Acad Sci U S A* 90(22): 10710-4.

Randolph, J. B. and A. S. Waggoner (1997). Stability, specificity and fluorescence brightness of multiply-labeled fluorescent DNA probes. *Nucleic Acids Res* 25(14): 2923-9.

Raychaudhuri, S., J. M. Stuart, et al. (2000). Principal components analysis to summarize microarray experiments: application to spoulation time series. *Pac. Symp. Biocomput.* 2000: 455-466.

Rebrikov, D. V., O. V. Britanova, et al. (2000). Mirror orientation selection (MOS): a method for eliminating false positive clones from libraries generated by suppression subtractive hybridization. *Nucleic Acids Res* 28(20): E90.

Reinartz, J., E. Bruyns, et al. (2002). Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Brief Funct Genomic Proteomic* 1(1): 95-104.

Reiner, A., D. Yekutieli, et al. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19(3): 368-75.

Reynolds, B. A. and S. Weiss (1992). Generation of neurons and astrocytes from isolated cells of the adult mammalian central nervous system. *Science* 255(5052): 1707-10.

Rosok, O., J. Odeberg, et al. (1996). Solid-phase method for differential display of genes expressed in hematopoietic stem cells. *Biotechniques* 21(1): 114-21.

Rossi, F. and E. Cattaneo (2002). Opinion: neural stem cell therapy for neurological diseases: dreams and reality. *Nat Rev Neurosci* 3(5): 401-9.

Rossi, F. and E. Cattaneo (2002). Opinion: neural stem cell therapy for neurological diseases: dreams and reality. *Nat Rev Neurosci* 3(5): 401-9.

Roth, F. P., J. D. Hughes, et al. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* 16(10): 939-45.

Ryo, A., N. Kondoh, et al. (2000). A modified serial analysis of gene expression that generates longer sequence tags by nonpalindromic cohesive linker ligation. *Anal Biochem* 277(1): 160-2.

Saha, S., A. B. Sparks, et al. (2002). Using the transcriptome to annotate the genome. *Nat Biotechnol* 20(5): 508-12.

Saiki, R. K., S. Scharf, et al. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230(4732): 1350-4.

Salin, H., T. Vujasinovic, et al. (2002). A novel sensitive microarray approach for differential screening using probes labelled with two different radioelements. *Nucleic Acids Res* 30(4): e17.

Schell, T., A. E. Kulozik, et al. (2002). Integration of splicing, transport and translation to achieve mRNA quality control by the nonsense-mediated decay pathway. *Genome Biol* 3(3): REVIEWS1006.

Schena, M., D. Shalon, et al. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270(5235): 467-70.

Schroeder, B. G., L. M. Peterson, et al. (2002). Improved quantitation and reproducibility in Mycobacterium tuberculosis DNA microarrays. *J Mol Microbiol Biotechnol* 4(2): 123-6.

Schulze, A. and J. Downward (2001). Navigating gene expression using microarrays--a technology review. *Nat Cell Biol* 3(8): E190-5.

Shalon, D., S. J. Smith, et al. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* 6(7): 639-45.

Shoemaker, D. D., E. E. Schadt, et al. (2001). Experimental annotation of the human genome using microarray technology. *Nature* 409(6822): 922-7.

Siebert, P. D., A. Chenchik, et al. (1995). An improved PCR method for walking in uncloned genomic DNA. *Nucleic Acids Res* 23(6): 1087-8.

Silva, A. P., J. E. De Souza, et al. (2004). The impact of SNPs on the interpretation of SAGE and MPSS experimental data. *Nucleic Acids Res* 32(20): 6104-10.

Smith, H. O. and K. W. Wilcox (1970). A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. *J Mol Biol* 51(2): 379-91.

Smith, L., P. Underhill, et al. (2003). Single primer amplification (SPA) of cDNA for microarray expression analysis. *Nucleic Acids Res* 31(3): e9.

Smyth, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3(1): Article 3.

Smyth, G. K., J. Michaud, et al. (2003). The use of within-array duplicate spots for assessing differential expression in microarray experiments. *Submitted:* http://www.statsci.org/smyth/pubs/dupcor.pdf.

Smyth, G. K. and T. Speed (2003). Normalization of cDNA microarray data. *Methods* 31(4): 265-73.

Sompayrac, L., S. Jane, et al. (1995). Overcoming limitations of the mRNA differential display technique. *Nucleic Acids Res* 23(22): 4738-9.

Southern, E. M., S. C. Case-Green, et al. (1994). Arrays of complementary oligonucleotides for analysing the hybridisation behaviour of nucleic acids. *Nucleic Acids Res* 22(8): 1368-73.

Spellman, P. T., G. Sherlock, et al. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 9(12): 3273-97.

Spradling, A., D. Drummond-Barbosa, et al. (2001). Stem cells find their niche. *Nature* 414(6859): 98-104.

Stears, R. L., R. C. Getts, et al. (2000). A novel, sensitive detection system for high-density microarrays using dendrimer technology. *Physiol Genomics* 3(2): 93-9.

Sterrenburg, E., R. Turk, et al. (2002). A common reference for cDNA microarray hybridizations. *Nucleic Acids Res* 30(21): e116.

Studer, L., V. Tabar, et al. (1998). Transplantation of expanded mesencephalic precursors leads to recovery in parkinsonian rats. *Nat Neurosci* 1(4): 290-5.

Suslov, O. N., V. G. Kukekov, et al. (2002). Neural stem cell heterogeneity demonstrated by molecular phenotyping of clonal neurospheres. *Proc Natl Acad Sci U S A* 99(22): 14506-11.

Tamayo, P., D. Slonim, et al. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* 96(6): 2907-12.

Tavazoie, S., J. D. Hughes, et al. (1999). Systematic determination of genetic network architecture. *Nat Genet* 22(3): 281-5.

Temple, S. and A. Alvarez-Buylla (1999). Stem cells in the adult mammalian central nervous system. *Curr Opin Neurobiol* 9(1): 135-41.

Thomas, J. G., J. M. Olson, et al. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res* 11(7): 1227-36.

Toronen, P., M. Kolehmainen, et al. (1999). Analysis of gene expression data using self-organizing maps. *FEBS Lett* 451(2): 142-6.

Tran, P. H., D. A. Peiffer, et al. (2002). Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. *Nucleic Acids Res* 30(12): e54.

Tseng, G. C., M. K. Oh, et al. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res* 29(12): 2549-57.

Tusher, V. G., R. Tibshirani, et al. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98(9): 5116-21.

Uchida, N., D. W. Buck, et al. (2000). Direct isolation of human central nervous system stem cells. *Proc Natl Acad Sci U S A* 97(26): 14720-5.

Unneberg, P., A. Wennborg, et al. (2003). Transcript identification by analysis of short sequence tags--influence of tag length, restriction site and transcript database. *Nucleic Acids Res* 31(8): 2217-26.

Van Gelder, R. N., M. E. von Zastrow, et al. (1990). Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc Natl Acad Sci U S A* 87(5): 1663-7.

van Steensel, B. and S. Henikoff (2003). Epigenomic profiling using microarrays. *Biotechniques* 35(2): 346-50, 352-4, 356-7.

Wang, E., L. D. Miller, et al. (2000). High-fidelity mRNA amplification for gene profiling. *Nat Biotechnol* 18(4): 457-9.

Wang, X., S. Ghosh, et al. (2001). Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res* 29(15): E75-5.

Wang, X. J., J. L. Reyes, et al. (2004). Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. *Genome Biol* 5(9): R65.

Wang, Z. and D. D. Brown (1991). A gene expression screen. *Proc Natl Acad Sci U S A* 88(24): 11505-9.

Watson, J. D. and F. H. Crick (1953). Genetical implications of the structure of deoxyribonucleic acid. *Nature* 171(4361): 964-7.

Watson, J. D. and F. H. Crick (1953). The structure of DNA. *Cold Spring Harb Symp Quant Biol* 18: 123-31.

Vaudry, D., B. J. Gonzalez, et al. (2000). Pituitary adenylate cyclase-activating polypeptide and its receptors: from structure to functions. *Pharmacol Rev* 52(2): 269-324.

Weeraratna, A. T., J. E. Nagel, et al. (2004). Gene expression profiling: from microarrays to medicine. *J Clin Immunol* 24(3): 213-24.

Velculescu, V. E., L. Zhang, et al. (1995). Serial analysis of gene expression. *Science* 270(5235): 484-7.

Welsh, J., K. Chada, et al. (1992). Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Res* 20(19): 4965-70.

Welsh, J. and M. McClelland (1991). Genomic fingerprinting using arbitrarily primed PCR and a matrix of pairwise combinations of primers. *Nucleic Acids Res* 19(19): 5275-9.

Venter, J. C., M. D. Adams, et al. (2001). The sequence of the human genome. *Science* 291(5507): 1304-51.

Vernon, S. D., E. R. Unger, et al. (2000). Reproducibility of alternative probe synthesis approaches for gene expression profiling with arrays. *J Mol Diagn* 2(3): 124-7.

Wilcox, A. S., A. S. Khan, et al. (1991). Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome. *Nucleic Acids Res* 19(8): 1837-43.

Wildsmith, S. E., G. E. Archer, et al. (2001). Maximization of signal derived from cDNA microarrays. *Biotechniques* 30(1): 202-6, 208.

Williamson, A. R. (1999). The Merck Gene Index project. *Drug Discov Today* 4(3): 115-122.

Wilusz, C. J., M. Wormington, et al. (2001). The cap-to-tail guide to mRNA turnover. *Nat Rev Mol Cell Biol* 2(4): 237-46.

Virlon, B., L. Cheval, et al. (1999). Serial microanalysis of renal transcriptomes. *Proc Natl Acad Sci U S A* 96(26): 15286-91.

Wolfinger, R. D., G. Gibson, et al. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol* 8(6): 625-37.

Wurmser, A. E., K. Nakashima, et al. (2004). Cell fusion-independent differentiation of neural stem cells to the endothelial lineage. *Nature* 430(6997): 350-6.

Xiang, C. C., M. Chen, et al. (2003). Probe generation directly from small numbers of cells for DNA microarray studies. *Biotechniques* 34(2): 386-8, 390, 392-3.

Xiang, C. C., O. A. Kozhich, et al. (2002). Amine-modified random primers to label probes for DNA microarrays. *Nat Biotechnol* 20(7): 738-42.

Xie, Y., K. S. Jeong, et al. (2004). A case study on choosing normalization methods and test statistics for two-channel microarray data. *Comparative and Functional Genomics* 5: 432-444.

Yamamoto, M., T. Wakatsuki, et al. (2001). Use of serial analysis of gene expression (SAGE) technology. *J Immunol Methods* 250(1-2): 45-66.

Yang, G. P., D. T. Ross, et al. (1999). Combining SSH and cDNA microarrays for rapid identification of differentially expressed genes. *Nucleic Acids Res* 27(6): 1517-23.

Yang, I. V., E. Chen, et al. (2002). Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol* 3(11): research0062.

Yang, M. C., Q. G. Ruan, et al. (2001). A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. *Physiol Genomics* 7(1): 45-53.

Yang, Y. H., M. J. Buckley, et al. (2002). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics* 11: 108-136.

Yang, Y. H., S. Dudoit, et al. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Re*s 30(4): e15.

Yang, Y. H., S. Dudoit, et al. (2001). Normalization for cDNA microarray data. Microarrays: Optical Technologies and Informatics. M. L. Bittner, Y. Chen, A. N. Dorsel and E. R. Dougherty. Volume 4266 of Proceedings of SPIE.

Yang, Y. H. and T. Speed (2002). Design issues for cDNA microarray experiments. *Nat Rev Genet* 3(8): 579-88.

Ye, S. Q., L. Q. Zhang, et al. (2000). miniSAGE: gene expression profiling using serial analysis of gene expression from 1 microg total RNA. *Anal Biochem* 287(1): 144-52.

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

Zeeberg, B. R., W. Feng, et al. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* 4(4): R28.

Zeng, J., R. A. Gorski, et al. (1994). Differential cDNA cloning by enzymatic degrading subtraction (EDS). *Nucleic Acids Res* 22(21): 4381-5.

Zhang, L., W. Zhou, et al. (1997). Gene expression profiles in normal and cancer cells. *Science* 276(5316): 1268-72.

Zhao, H., T. Hastie, et al. (2002). Optimization and evaluation of T7 based RNA linear amplification protocols for cDNA microarray analysis. *BMC Genomics* 3(1): 31.

Zhao, N., H. Hashida, et al. (1995). High-density cDNA filter analysis: a novel approach for large-scale, quantitative analysis of gene expression. *Gene* 156(2): 207-13.

Zhu, Y. Y., E. M. Machleder, et al. (2001). Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* 30(4): 892-7.