Subtracted Approaches to Gene Expression Analysis in Atherosclerosis

Stina Boräng

Royal Institute of Technology
Department of Biotechnology

Stockholm  2004
© Stina Boräng

Department of Biotechnology
Royal Institute of Technology
Alba Nova University Center
SE-106 91 Stockholm
Sweden

Printed 2003 at
Universitetsservice US AB
Box 700 14
SE-100 44 Stockholm
Sweden
ISBN 91-7283-653-9
Abstract

Gene expression analysis has evolved as an extensive tool for elucidation of various biological and molecular events occurring in different organisms. A variety of techniques and software tools have been developed to enable easier and more rapid means of exploring the genetic information. A more effective approach than exploring the whole content of genes expressed under certain conditions is to study fingerprint assays or to use subtracted cDNA libraries to identify only differentially expressed genes.

The objective for the work in this thesis has been to explore differentially expressed genes in atherosclerosis. This was done by applying and modifying a protocol for the subtractive approach RDA (Representational Difference Analysis) in different model systems.

Initially, the molecular effects of an anti-atherosclerotic drug candidate were elucidated. In addition, two alternative approaches to identify differentially expressed genes obtained after iterative rounds of RDA subtraction cycles were evaluated. This revealed that in most cases, the shotgun approach in which the obtained gene fragments are cloned without any prior selection has clear advantages compared to the more commonly used selection strategy, whereby distinct bands are excised after gel electrophoresis.

A key process in the atherosclerotic plaque initiation is the phenotypic change of macrophages into foam cells, which can be triggered in a model system by using macrophages exposed to oxidised LDL. To investigate the genes expressed in this process, the RDA technique was combined with microarray analysis, which allows for selectivity and sensitivity through RDA, as well as rapid high-throughput analysis using microarrays. The combination of these techniques enables significant differences in gene expression to be detected, even for weakly expressed genes and the results to be reliably validated in a high throughput manner.

Finally, investigation of the focal nature of atherosclerotic lesions and gene expression profiling were studied using in vivo aortic tissues from ApoE-/- and LDLR -/- mice. The study was based on a comparison between localisations that are likely, and others that are unlikely, to develop atherosclerotic plaques, and the RDA technique was employed to explore differential gene expression.

© Stina Boräng, 2003

Keywords: Representational Difference Analysis, atherosclerosis, gene expression profiling
I am not afraid of storms,
for I am learning how to sail my ship

Louisa May Alcott
LIST OF PUBLICATIONS

This thesis is based on the following manuscripts, which in the text will be referred to by their roman numerals:


Table of contents

Populärvetenskaplig sammanfattning .......................................................... 11
INTRODUCTION
1 Genome discovery .................................................................................. 16
   1.1 The Human Genome Project ....................................................... 17
2 Global analysis of gene expression ...................................................... 19
   2.1 Expressed sequence tags ............................................................ 20
   2.2 Serial analysis of gene expression ............................................ 22
   2.3 DNA microarrays .......................................................................... 24
      2.3.1 Spotted arrays (cDNA) ....................................................... 25
         2.3.1.1 Array fabrication ....................................................... 27
         2.3.1.2 Target preparation and hybridisation ....................... 27
         2.3.1.3. Data analysis ........................................................... 27
3 Selective analysis of differential gene expression ..................................... 29
   3.1 Differential display and RNA arbitrarily primed PCR ................... 29
   3.2 Suppression subtractive hybridisation ......................................... 32
   3.3 Representational difference analysis ........................................... 34
4 Tools for gene expression sequence tag analysis .................................... 37
   4.1 Preprocessing of sequences ....................................................... 37
   4.2 Assembly ...................................................................................... 37
   4.3 Annotation .................................................................................... 38
5 Tools for microarray analysis .................................................................. 39
   5.1 Image analysis ............................................................................. 39
   5.2 Normalization ............................................................................... 40
   5.3 Selection of differentially expressed genes ..................................... 40
PRESENT INVESTIGATION
6 Pathogenesis of atherosclerosis ........................................................... 42
7 Differential gene expression in atherosclerosis ..................................... 44
   7.1 Treatment with a therapeutic drug candidate (Paper I) ............... 44
   7.2 Foam cell formation in atherosclerotic lesions (Papers II and III) 46
   7.3 Focal localisation of atherosclerotic plaques (Paper IV) .......... 48
8 Signature Tag RDA .............................................................................. 51
9 Concluding remarks .............................................................................. 53
Acknowledgements
References
Original Papers I-IV
Populärvetenskaplig sammanfattning


Figur 1. Schematisk bild över hur DNA transkriberas till mRNA som vid ribosomer styr sammanfogningen av aminosyror till proteiner.

Ordet cell betyder egentligen rum, vilket kan liknas vid att cellen har en cellvägg som omsluter en rad olika organeller. Organellerna har olika funktioner i cellen, precis som möblerna i ett rum där en del har till uppgift att vara bekväma, andra snygga osv. Den viktigaste organellen är cellkärnan som fungerar som sambandscentral för alla de processer som sker inuti cellen. I cellkärnan finns organismens arvsanlag, dvs dess gener, vilka är uppbyggda av DNA. Det finns även organismer som inte har någon cellkärna, som bakterier där allt DNA istället ligger fritt inuti cellen. 1943 kom Oswald Avery på att en cells DNA är det som innehåller all genetisk information och tio år senare upptäckte Francis Crick och James Watson strukturen för DNA, vilket de också fick Nobelpriset för.

En DNA-molekyl är uppbyggd av fyra olika beståndsdelar som kallas för nukleotider. Dessa brukar förkortas A, C, G respektive T och den inbördes ordningen (sekvensen) av dessa kodar för vad genen har för funktion. Ordningen på nukleotiderna bestämmer nämligen vilket protein som ska

Arbetet som ligger till grund för denna avhandling bygger till stor del på en molekyl som kallas mRNA. Denna bildas som ett mellanled i syntesen av proteiner från DNA (se Figur 1). Genom att studera mRNA kan man i många hänseenden lättare identifiera en gen än om man skulle detektera och följa det färdiga proteinet som bildas av genen. mRNA är emellertid en något instabil molekyl och man omvandlar den därför ofta till en komplementär DNA-molekyl (cDNA), eftersom DNA är stabilare och lättare att arbeta med än RNA. Alla gener i en cell är inte aktiva hela tiden utan väntar på en signal för att aktiveras. Det är alltså bara de gener som är aktiva under rådande omständigheter som omvandlas till mRNA och vidare till proteiner. Det finns idag en mängd olika metoder för att ta reda på vilka gener som är aktiva (gene expression) och avhandlingen beskriver en rad sådana. Man kan jämföra celler för att studera vilka gener som är aktiva på olika ställen inom samma organism eller mellan individer som behandlats olika. Det finns även ett antal metoder som koncentrerar sig enbart på att hitta skillnader mellan olika celler. Ett exempel på en sådan metod förkortas RDA och just denna metod har använts i alla artiklar som denna avhandling bygger på.

Samtliga fyra artiklar handlar om att hitta skillnader i genexpression i modellsystem som involverar gener påverkade av åderförkalkning. I den första artikeln jämförs celler som behandlats med en ny medicinkandidat mot åderförkalkning med obehandlade celler för att försöka ta reda på exakt hur

---

Subtracted Approaches to Gene Expression Analysis in Atherosclerosis
medicinkandidaten reglerar de cellulära mekanismerna, dvs vilka gener som aktiveras eller inaktiveras. I artikel två och tre jämförs celler som utvecklats olika långt i åderförkalkningsprocessen. Härigenom kan man få ledtrådar om vilka gener som påverkar denna utveckling och därigenom möjligen förstå hur man skulle kunna hindra förloppet av åderförkalkning. Denna vanliga och allvarliga sjukdom har en tendens att uppstå främst där blodkärlen böjer eller delar på sig. För att försöka ta reda på varför det är så genomfördes det sista arbetet där celler från blodkärlens krökar och förgreningar jämfördes med raka blodkärl för att studera skillnader i genexpression.
Introduction
1 Genome discovery

Francis Collins, the director of the National Human Genome Research Institute (NHGRI), perhaps best described the essential properties of the complex human genome in 2001:

“It’s a history book – a narrative of the journey of our species through time. It’s a shop manual, with an incredibly detailed blueprint for building every human cell. And it’s a transformative textbook of medicine, with insights that will give health care providers immense new powers to treat, prevent and cure disease.”

In other words, the genome of any organism contains all the information required to understand its physiological nature, development and evolutionary history. Today, several genomes of different organisms have been determined, and the task for researchers now is to decipher their transcribed genes, or transcriptome, and draw correlations with the complex corresponding protein network, the proteome. Exploration and elucidation of these intricate features of every cell is most commonly known nowadays as functional genomics.

Technologies used in molecular biology have constantly evolved and improved. During the past 25 years, two landmark technologies within the field have been developed. First, two independent methods for DNA sequencing were invented, one by Allan Maxam and Walter Gilbert, the other by Fred Sanger and coworkers (Maxam and Gilbert 1977; Sanger et al. 1977). Second, Kary Mullis devised the polymerase chain reaction (PCR) technique, enabling the rapid multiplication of DNA fragments (Mullis et al. 1986). Like DNA sequencing, PCR has completely revolutionised molecular genetics, enabling a whole new approach for the study and analysis of genes.

An intense collective effort is underway to map the genome of various organisms, for a number of reasons. For instance, detection of genes (or regions of genes) that are seldom affected by mutations or other changes (conserved domains) may provide new insights into gene function. On the other hand, single-base sequence variations occurring at specific locations in the genome, single nucleotide polymorphisms (SNPs), occur throughout the whole genome
(Brookes 1999), and genotyping, i.e. comparison of the SNPs between individuals within the same species, can increase our knowledge, for instance, genetically related diseases. Sequencing of the human genome along with other organisms, from yeast to chimpanzees, has given rise to a growing biological research field called comparative genomics. Although all organisms appear to be different and behave in various ways, all of their genomes are composed of DNA. Comparative genomics offers researchers possibilities to pinpoint the signals that control gene function, and thus new approaches for treating diseases. In addition, the identification of regions of similarity and differences among species may facilitate understanding of the structure and function of genes, and to address questions such as why chimpanzees do not suffer from some of the severe diseases that affect humans, such as HIV, although human and chimp DNA sequences are estimated to be 98.8% identical. Data on the genomes of more than 800 organisms, representing both completely sequenced organisms and organisms for which sequencing is still in progress, can be found at http://www.ncbi.nlm.nih.gov.

1.1 The Human Genome Project

The human genome includes approximately three billion base pairs, packaged in the 23 pairs of chromosomes. Fifty years after James Watson and Francis Crick proposed a double helical structure of DNA in 1953 (Watson and Crick 1953), the complete sequence of the human genome is now available. To obtain this sequence, an international, collaborative research effort was initiated in 1990, called the Human Genome Project (HGP). The goal of this effort was to create a public database containing genetic information as a resource for scientific discovery within a time limit of 15 years. In February 2001, HGP published a draft version of the human genome sequence (Lander et al. 2001). Simultaneously, another research group led by Craig Venter of Celera Genomics published their less widely available draft version of the human genome sequence (Venter et al. 2001).

More than two years ahead of schedule and for much lower cost than originally estimated, HGP announced that the sequence was finished, and published it,
in April 2003. The HGP defines a finished sequence as being highly accurate (with fewer than one error per 10,000 letters) and highly contiguous (the only remaining gaps corresponding to regions where the sequences cannot be reliably resolved with current technology). From the sequence, the human genome is estimated to contain between 30,000 and 40,000 genes, in contrast to earlier estimates ranging from 50,000 to 140,000 genes (Liang et al. 2000). About 99% of the gene-containing parts of the human genome are covered, with an accuracy of 99.99%. The major findings from HGP’s enormous efforts are that genes account for a relatively small amount of the human genome and the architecture of human proteins is very complex compared to that of other species. Since the genomic sequence for eukaryotes only contains a small portion of coding regions (around 2% in the human genome (Lander et al. 2001)), intensive efforts are required to identify and sequence regions with altered expression levels using genomic sequencing. The full sequence of the human genome provides a huge source of information about the structure, organisation and function of the human genes. As Nobel laureate James D. Watson stated:

“The completion of the Human Genome Project is a truly momentous occasion for every human being around the globe.”
2 Global analysis of gene expression

Obviously, even before the human genome was completed, many research groups strove to identify genes involved in a wide range of cellular processes. In many cases, exploring genes that are only active under circumstances of specific interest would immensely simplify this procedure. This can be done by investigating the genetic information available at the RNA rather than the DNA level. Through splicing events, all introns and occasionally exons are removed from the transcript. The processed mRNA is then translated into proteins. The function of the proteins produced can be dramatically altered through the production of different splice variants of the transcribed genes (Graveley 2001).

The analysis of gene expression in specific tissues and physiological processes has rapidly developed over the last twenty years, and it is now potentially possible to identify all of the genes expressed in a specific tissue. The introduction of PCR together with other technological improvements (such as microarrays) has simplified the discovery of differentially expressed genes.

The fundamental principle for monitoring gene expression in tissues involves extraction of total RNA followed by isolation of the mRNA fraction and reverse transcription into cDNA, that can be cloned into bacterial plasmids, resulting in a cDNA library, and the whole spectrum of mRNAs are represented. Collectively, all the cDNAs in a library represent the frequency of expression of different genes, i.e. the most abundantly expressed genes will generate the most abundant copies of individual clones. It has been estimated that 20 % of the total mRNA population in the human genome are represented by less than 100 different transcripts in a cell (Gibson and Muse 2002). To identify less abundant genes, libraries can be normalized (Soares et al. 1994; Bonaldo et al. 1996). That is, the most abundant genes can be subtracted and discarded through reassociation kinetics, theoretically leaving only the weakly expressed transcripts or single copies of each gene, depending on the normalization strategy chosen.
2.1 Expressed sequence tags

One of the most widely used approaches for gene identification and gene expression profiles in various tissues, cell types, or developmental stages, is to generate expressed sequence tags (ESTs) (Adams et al. 1991). Briefly, an EST is part of a sequence from a cDNA clone that corresponds to an mRNA. When constructing an EST library, it is often beneficial to utilize the polyA region of the mRNA. Usually, a restriction site (typically NotI) is introduced together with a polyT oligonucleotide to prime the first strand synthesis of cDNA. However, extension of the first strand of cDNA is often incomplete due to the presence if inhibitory secondary structures, so fragments of various lengths are produced. Different approaches to generate full-length clones are available, (see, for instance (Carninci and Hayashizaki 1999; Das et al. 2001; Suzuki and Sugano 2001)). Following second strand synthesis, adaptors (usually with an EcoRI overhang) are ligated to both ends. Enzyme restriction with NotI then enables directional cloning into a suitable phage or plasmid. In most cases, sequencing of the clones is performed from the 5’-end to avoid problems reading through the polyadenylated 3’-region. To circumvent directed cloning of the 3’-end regions, random hexamers can be utilized in the cDNA synthesis (Dudley et al. 1978; Dias Neto et al. 2000), although this approach is not widely employed. In theory, the complete transcribed region, except for the outermost ends, will then be represented.

By choosing clones for sequencing in a randomised manner, it is possible to construct a profile of the transcriptional activity. Generally, an EST library is sequenced until the yield of novel clones is reduced to less than 10-20 %. The possibility of using normalized libraries could be advantageous in this respect. Once the clones have been sequenced, they are made available through the IMAGE Consortium (http://image.llnl.gov), and the sequences are deposited in electronic databases. The National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) maintains a repository of ESTs. Table 1 displays the number of public EST entries at this date (November, 2003) for a selection of organisms (http://www.ncbi.nlm.nih.gov/dbEST).
Table 1. Number of public EST entries for a selection of organisms.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Number of public EST entries (<a href="http://www.ncbi.nlm.nih.gov/dbEST">http://www.ncbi.nlm.nih.gov/dbEST</a>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens (human)</td>
<td>5,427,257</td>
</tr>
<tr>
<td>Mus musculus + domesticus (mouse)</td>
<td>3,944,197</td>
</tr>
<tr>
<td>Rattus sp. (rat)</td>
<td>538,251</td>
</tr>
<tr>
<td>Triticum aestivum (wheat)</td>
<td>503,828</td>
</tr>
<tr>
<td>Ciona intestinalis</td>
<td>492,488</td>
</tr>
<tr>
<td>Gallus gallus (chicken)</td>
<td>451,573</td>
</tr>
<tr>
<td>Zea mays (maize)</td>
<td>384,103</td>
</tr>
<tr>
<td>Danio rerio (zebrafish)</td>
<td>362,483</td>
</tr>
<tr>
<td>Hordeum vulgare + subsp. vulgare (barley)</td>
<td>348,282</td>
</tr>
<tr>
<td>Xenopus laevis (African clawed frog)</td>
<td>344,747</td>
</tr>
<tr>
<td>Glycine max (soybean)</td>
<td>344,524</td>
</tr>
<tr>
<td>Bos taurus (cattle)</td>
<td>331,139</td>
</tr>
<tr>
<td>Drosophila melanogaster (fruit fly)</td>
<td>261,414</td>
</tr>
<tr>
<td>Oryza sativa (rice)</td>
<td>260,901</td>
</tr>
<tr>
<td>Saccharum officinarum</td>
<td>246,301</td>
</tr>
<tr>
<td>Silurana tropicalis</td>
<td>215,211</td>
</tr>
<tr>
<td>Caenorhabditis elegans (nematode)</td>
<td>215,200</td>
</tr>
<tr>
<td>Arabidopsis thaliana (thale cress)</td>
<td>190,732</td>
</tr>
<tr>
<td>Medicago truncatula (barrel medic)</td>
<td>187,763</td>
</tr>
<tr>
<td>Sus scrofa (pig)</td>
<td>171,920</td>
</tr>
<tr>
<td>Sorghum bicolor (sorghum)</td>
<td>161,766</td>
</tr>
<tr>
<td>Dictyostelium discoideum</td>
<td>155,032</td>
</tr>
<tr>
<td>Chlamydomonas reinhardtii</td>
<td>154,600</td>
</tr>
<tr>
<td>Lycopersicon esculentum (tomato)</td>
<td>150,228</td>
</tr>
<tr>
<td>Schistosoma mansoni (blood fluke)</td>
<td>139,064</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>135,542</td>
</tr>
<tr>
<td>Anopheles gambiae (African malaria mosquito)</td>
<td>130,731</td>
</tr>
<tr>
<td>Pinus taeda (lobolly pine)</td>
<td>110,622</td>
</tr>
<tr>
<td>Oryzias latipes (Japanese medaka)</td>
<td>103,098</td>
</tr>
<tr>
<td>Oncorhynchus mykiss (rainbow trout)</td>
<td>102,224</td>
</tr>
</tbody>
</table>

Genome projects including various model organisms have taken advantage of EST studies because of its suitability for the discovery of new genes, physical mapping of genomes, and identification of coding regions in genomic sequences (Adams et al. 1991). Also, comparative EST analysis provides a valuable resource for various biological research fields. For example, it allows evaluation of gene expression patterns in response to different biological signals, thereby enhancing the understanding of cellular biology and physiology (Lee et al. 1995).
2.2 Serial analysis of gene expression

In 1995 a new, more rapid means of tag sequencing (Serial Analysis of Gene Expression, SAGE) was described with the potential to significantly increase throughput capacity (Velculescu et al. 1995). In contrast to the EST approach, SAGE allows the sequencing of multiple tags within a single clone, thereby reducing both time and sequencing costs. The major application of SAGE was expected to be comparison of gene expression patterns in different developmental and disease states, but today it is also used in a variety of applications to study functional genomics in different organisms.

Figure 2 shows the principle of SAGE, in which cDNA is reversely transcribed from mRNA using a biotinylated oligo-dT primer. The cDNA is then cleaved with a restriction endonuclease (anchoring enzyme) and the 3'-terminal cDNA fragments are bound to streptavidin-coated beads. The captured cDNA is divided into two aliquots and ligated to one of two oligonucleotide linkers containing a recognition site for a tagging enzyme, which belongs to the class IIS restriction endonucleases and hence cleaves DNA at a specific distance 3’ to the recognition site. Cleavage with the tagging enzyme will yield short (~9-14 bp) tags of cDNA that can be ligated to each other, forming ditags. Ligated tags are used as templates for PCR amplification with linker-specific primers, and the PCR products are cleaved by the anchoring enzyme, concatemerized into long continuous stretches of DNA, cloned into a plasmid vector then sequenced. This allows for high-throughput sequencing of up to 50 tags per sequence run.

SAGE analysis has a number of unique advantages over other techniques for global gene expression analysis (Velculescu et al. 1997). The rapid, high-throughput sequencing and analysis of tags generates reliable expression profiles and enables the discovery of rare and novel gene transcripts since SAGE theoretically generates a tag for every cellular mRNA. Among the negative aspects of the method are limitations due to the short tag length generated. This problem may lead to failure of a tag to match and uniquely identify sequences in SAGE reference databases, especially if it is situated in a
Figure 2. Schematic of the SAGE procedure.
conserved region (Ishii et al. 2000; Kannbley et al. 2003). Other disadvantages are that some of the cDNAs may lack the restriction site used to construct the SAGE library, and the technique provides no information about splice variants. Although SAGE has been widely adopted for global analysis of gene expression, its applicability is limited by the large amount of RNA required. Many research groups have tried to overcome this problem by using either PCR amplification of starting cDNA materials, as in SAGE-lite (Peters et al. 1999) and PCR-SAGE (Neilson et al. 2000), or PCR reamplification of SAGE ditags, as in microSAGE (Datson et al. 1999) and SADE (Virlon et al. 1999). These methods all include an additional amplification step, which may introduce bias in quantitative analysis of gene expression. In the year 2000, miniSAGE was introduced as a modified SAGE protocol that does not require any additional amplification (Ye et al. 2000). This technique allows for gene expression profiling using only 1 µg total RNA, and is still considered to require less starting material than any other SAGE technique. Despite its drawbacks, SAGE is a general and powerful technique allowing not only global gene expression profiling of various eukaryotic organisms, but also the identification of genes that are exclusively expressed under various cellular conditions (Yamamoto et al. 2001).

2.3 DNA microarrays

Since the DNA microarray technique was first described in 1995 (Schena et al. 1995) it has been extensively employed for large-scale analysis of gene expression in the field of functional genomics. This high-capacity system can be used to measure the relative quantities of specific mRNAs representing tens of thousands of genes in two or more tissue samples in a single experiment. Three types of microarray can be constructed, depending on the source of the immobilized DNA. In this thesis, only the basic methodology concerning spotted microarrays (cDNA or longmers) will be discussed, although both genomic DNA (Forozan et al. 1997) and in situ synthesised oligonucleotides are being used as probes attached to microarray surfaces. An important example of the latter is the Affymetrix platform (Lockhart et
al. 1996), where oligonucleotides are synthesised directly on the surface by photolithography and solid-phase chemistry (Fodor et al. 1993) to produce probes, 20-25 oligonucleotides in length. Multiple probe pairs (one perfect-match oligonucleotide is paired with a mismatch oligonucleotide) for different regions of each gene are designed, allowing for mean values of signal intensities to be calculated.

2.3.1 Spotted arrays (cDNA)

The use of cDNA microarrays to examine numerous genes in parallel is currently one of the most common approaches for gene expression profiling. cDNA fragments are first amplified and sequenced before being spotted onto a suitable surface, generally glass microscope slides, at high density. Relative expression levels of genes represented in the array can be analysed by comparing fluorescence intensities between two fluorescently labelled samples hybridised to the array. Normalization of the collected data is essential for adjustment of fluorescence intensity ratios, and should not be neglected. Various authors, including Priti Hegde and colleagues (Hegde et al. 2000), have optimised several steps involved in the process of constructing reliable and reproducible array platforms. The general procedure of cDNA microarray experiments is outlined in Figure 3. Recently, a single long probe (longmer, consisting of 50-70 oligonucleotides) for each gene was introduced as an alternative to the well-established methods using cDNA or in situ synthesized oligonucleotides attached to the arrays (Shoemaker and Linsley 2002). Longmer arrays can be fabricated and analysed in roughly the same standard manner as for spotted cDNA arrays. In addition, the longmer strategy saves time since no amplification of cDNA clones is required, and comparison with in situ synthesized 25-mer probes has produced promising results (Barczak et al. 2003).
Subtracted Approaches to Gene Expression Analysis in Atherosclerosis

Figure 3. cDNA microarray flowchart.
2.3.1.1 Array fabrication
When creating an array, either cDNAs representing all known genes for the organism studied or a subset of clones representing only the genes of interest to the particular study can be used. In either case, amplified PCR products are spotted onto a slide by a high-speed robotic system, and the array is further processed to attach the DNA sequences to the surface and denature them. Concurrently, both positive and negative control gene sequences should be printed onto the slide to validate the data generated. To avoid intra-slide variations, cDNA clones are generally spotted with duplicates (or triplicates) spread over the slide surface. Even the slightest changes in the microenvironment, such as modifications of the slide surface, spotting buffer, temperature, and relative humidity, may affect the quality of the spotted gene fragments and hybridisation strength (Lander 1999).

2.3.1.2 Target preparation and hybridisation
RNA from two different sources is used for reverse transcription into single stranded cDNA in the presence of nucleotides labelled with two different fluorescent dyes (typically Cy3 and Cy5), one for each sample. The labelled reaction products are purified, mixed, and hybridised to the array surface, allowing the differentially labelled cDNAs to bind the corresponding nucleic acid molecules spotted onto the surface in a competitive manner. However, prior to hybridisation of fluorescently labelled cDNAs to the array, depending on the slide used, its surface may need blocking or inactivation of active molecules coated on the surface to reduce background signals.

2.3.1.3 Data analysis
High-resolution confocal fluorescence scanning of the array provides data on the relative signal intensities and ratios between the samples for the genes represented on the microarray. This allows relative expression levels to be estimated, and differentially expressed genes to be determined. A number of factors, such as RNA quantity and quality, labelling efficiency, and detection of intensity signals from the different laser wavelengths, all affect the ratios obtained. Therefore, normalization of the data obtained is of utmost importance and a number of different strategies have been developed and utilized for this purpose (Schuchhardt et al. 2000; Kerr and Churchill 2001; Yang et al. 2002).
Subtracted Approaches to Gene Expression Analysis in Atherosclerosis
3 Selective analysis of differential gene expression

The identification of differentially expressed mRNAs has been used to help understand not only gene function, but also the underlying molecular mechanisms of particular biological systems. A more effective approach than exploring the whole content of genes expressed under certain conditions is to study fingerprint assays or to use subtracted cDNA libraries to identify only differentially expressed genes. This can heavily reduce the time, money and effort involved. To do this, a variety of selective techniques have been developed, and some of the most frequently used techniques are described below.

3.1 Differential display and RNA arbitrarily primed PCR

To meet the needs for isolating and identifying genes that are differentially expressed in various cells and conditions, Peng Liang and Arthur Pardee developed a new technology called differential display (DD) in 1991 (Liang and Pardee 1992). Briefly, the method is based on primer-sets in which the first primer is anchored to mRNA in the polyadenylated region of the 3’-end, while the other is anchored with arbitrary spacing upstream from the first. This yields a subpopulation of mRNAs, which can be reversibly transcribed into cDNA, amplified, and resolved on a polyacrylamide gel. Differentially expressed genes between two or more samples can then be detected in parallel, and further explored. The principles of DD are schematically illustrated in Figure 4A.
Subtracted Approaches to Gene Expression Analysis in Atherosclerosis

Figure 4. Overview of two fingerprint assays, (A) differential display (B) RNA arbitrarily primed PCR.
Using an oligo-dT primer with two additional nucleotides at the 3'-end (5'-oligo dT-VN-3', where V = A, C or G; and N = A, C, G or T) will generate a subpopulation corresponding to 1/12 of the total mRNA population. Such primers permit the initiation of reverse transcription of only this arbitrary set of mRNAs. As a 5' primer, a short oligonucleotide with an arbitrary base sequence is used. After amplification, this will yield products of various sequence lengths corresponding to different mRNAs. The gene fragments from each sample can then be visualized and separated by gel electrophoresis. Differentially expressed genes can then be further explored by excision, cloning, and sequencing of bands that differ between samples. Using other sets of primers will obviously generate a different subpopulation of gene fragments. Therefore, repeated experiments with other primers are required to cover the complete gene expression profile.

Over the years, criticism over the high number of false positives produced by DD, for example Dana Crawford et al (Crawford et al. 2002) has been raised, who also discuss the statistically predicted need for 240 different sets of primers to cover all mRNAs in a cell. However, intense efforts have been made by a number of research groups to resolve the problems and to refine and improve DD from a technological perspective (Liang 2002; Stein and Liang 2002).

An additional technique, the closely related RNA arbitrarily primed PCR (RAP-PCR) protocol, was developed by John Welsh and colleagues (Welsh et al. 1992) (Figure 4B). The only real difference in this approach compared to DD, is that primers with arbitrary oligonucleotides are used for both first and second strand cDNA synthesis, enabling a subpopulation of gene fragments spread throughout the genes, rather than just the 3'-ends, to be generated. This also enables analysis of non-RNA species.
3.2 Suppression subtractive hybridisation

A number of techniques have been developed to study differential gene expression between two different sources. One such method is to generate subtracted cDNA libraries. Suppression subtractive hybridisation (SSH), published in 1996 by Luda Diatchenko et al (Diatchenko et al. 1996), is a PCR-based cDNA subtraction method that combines normalization of fragments with both high and low abundance, and subtraction of gene fragments present in the two cDNA populations. A schematic view of the SSH procedure is outlined in Figure 5.

One cDNA population containing differentially expressed gene fragments of interest is termed the “tester” population, while the other is termed the “driver”. First, both tester and driver cDNAs are digested by restriction enzymes, and the tester is subdivided into two batches that are equal in all respects. Different sets of linkers containing long, inverted terminal repeats (Lukyanov et al. 1995) are ligated onto the cDNAs in the two batches, resulting in two tester populations. Tester and excess driver are mixed, heat-denatured, and annealed in a first hybridisation, leading to normalization, i.e. equalization of the abundance of the cDNAs in the tester. A second hybridisation step is then performed with a mixture of both tester populations as well as new driver to allow for the possibility that ds DNA may be formed, originating from both tester populations. These fragments with different linkers at the 3’- and 5’-ends are favoured in the following PCR amplification by using a pair of primers corresponding to the outer part of the two linkers. Hybridisation products consisting of tester/tester duplexes with the same linker (i.e. the same long, inverted terminal repeats) at the ends, will form stable hairpin-like structures after each denaturation-annealing PCR step, preventing linker-specific primers from annealing. In this manner, only rare target fragments are enriched, although the number of false positives obtained using the SSH method is relatively high.
SUPPRESSION SUBTRACTIVE HYBRIDISATION (SSH)

First hybridisation
Sample 1 the TESTER with linker 1
Sample 1 the TESTER with linker 2
Sample 2 the DRIVER

Second hybridisation
Mix samples
Add new DRIVER
One "new" product

Fill in ends and add primers for amplification
and

(No amplification) Exponential amplification

Figure 5. Scheme of the SSH method.
3.3 Representational difference analysis

Representational difference analysis (RDA) is a PCR-coupled subtractive enrichment procedure originally developed for detecting differences between two genomes (Lisitsyn and Wigler 1993). To avoid the complexity of using genomic DNA as starting material, Michael Hubank and colleagues (Hubank and Schatz 1994) adopted the protocol for use with differentially expressed genes. The method relies on restriction digestion of cDNA and ligation of adapters to a PCR-amplified subset of all gene sequences, and is outlined in Figure 6.

Endonuclease restriction of the cDNA, followed by linker ligation and PCR amplification, will generate a representation of the transcribed genes originating from the mRNA population. The cDNA representation from which unique sequences are sought is designated the “tester”, and the cDNA representation that is used to subtract sequences common to the two populations is designated the “driver”. To enrich differentially expressed genes, the linkers are removed by restriction digestion and another set of linkers is ligated to the tester fragments. The tester and driver are subjected to cross-hybridisation with excess amount of driver to “drive out” fragments that are also present in the tester, leaving three variants of hybridisation products: driver-specific fragments with no linkers, fragments common to both tester and driver with just one linker, and tester-specific fragments containing linkers in both 5’- and 3’-ends allowing for exponential amplification with linker-specific primers. The resulting pool of gene fragments after PCR amplification is denoted the first difference product (DP). Further rounds of linker removal and ligation of new sets of linkers, cross-hybridisations with more stringent ratios, and PCR amplifications are required to enrich differentially expressed genes with a low background of false positives. Both upregulated and downregulated genes in a model system can be identified in two parallel experiments by interchanging the tester and driver sources.
Figure 6. Schematic diagram of cDNA RDA.
In contrast to other methods, RDA makes it possible to rapidly reduce the number of genes represented to a few of the most potentially interesting by eliminating fragments present in roughly equal proportions in the two populations and leaving only those that are differentially expressed. Thus, reductions are achieved in the number of sequencing and data analysis steps, and consequent reductions in the time involved. However, there are some disadvantages. There is a high likelihood that two restriction enzyme restriction sites will be present in a mRNA of average length, and thus the intervening gene sequence will be amplified (although it should be noted that each of the sequences will not be equivalently amplified, so there will be sequence bias during the multiple rounds of PCR amplification, resulting in loss of some PCR products). Furthermore, some mRNAs may only harbour on site for the restriction enzyme selected, resulting in loss of that particular gene fragment. One way to overcome this problem would be to repeat the RDA protocol on the same cDNA populations using a different restriction enzyme, and comparing the results. RDA is not the method of choice when many differences are expected between two samples. Under such circumstances, RDA will probably enrich the fragments that are most efficiently amplified, and not necessarily the differentially expressed fragments of interest (Hubank and Schatz 1999).

In recent years, many research groups have optimised and modified the RDA protocol in diverse ways (most of which are not discussed in this thesis). For example, Jacob Odeberg and colleagues (Odeberg et al. 2000) further developed the RDA protocol taking advantage of solid-phase technology to simplify removal of digested linkers and uncleaved fragments, and to enable low amounts of starting material to be used. This modified protocol was used as a basis for the work presented in this thesis.
4 Tools for gene expression
sequence tag analysis

Regardless of the strategy chosen for gene expression profiling, it will generate a massive amount of sequence data. To facilitate management and analysis of the data obtained, powerful computational resources are required. The process for data analysis of gene sequences obtained from EST sequencing efforts or shotgun sequencing of subtracted approaches, can be broadly divided into three steps: pre-processing of the sequences, assembly, and annotation. For each of the three stages, a wide range of software tools have been developed, but here, only the main principles are described.

4.1 Preprocessing of sequences

Raw data from sequencing instruments needs to be passed through several processes before being entered into a subsequent assembly program. These processes include screening for the vector sequence, quality evaluation, and conversion of data formats. Manual editing of each sequence in a graphical user interface can have a powerful impact on the resulting sequences, although it is very time consuming, especially for large-scale sequencing projects. Instead, batches of sequences can be passed through the system in an automatic way. Neglecting the pre-processing step will generate sequences with poor quality, possibly leading to incorrect annotation of genes.

4.2 Assembly

In any sequencing project, the goal is to assemble all sequences with homology greater than a suitable threshold value into one cluster. This process involves comparison of sequences, finding overlapping regions, and integrating those satisfying pre-set computational criteria. At least two different approaches to achieve this have been developed. In the first, every new sequence entering
the assembly program is compared with sequences that have already been integrated. If a sufficient match value is reached, the sequences are merged into one contig and a consensus sequence is created, i.e. a single representative sequence for the cluster. In the second, all sequences are compared to each other simultaneously, and the best matches of sequences will be joined first. The second approach is generally preferred, although it demands much higher computational power.

Since sequences originating from the same gene family may be difficult to distinguish from each other, problems in assembly may also arise. In addition, different splice variants as well as chimeric clones, in which two or more gene fragments are brought together before ligation into a vector, may cause problems in the assembling procedure.

4.3 Annotation

When consensus sequences for all gene fragments obtained in a project have been determined, the next step is to find out which genes they represent. Numerous software tools have been developed for this purpose, one of the most common being BLAST (Basic Local Alignment Search Tool) (Altschul et al. 1990) for sequence-to-sequence comparisons against a suitable database. Instead of using consensus sequences, all sequences can be individually used for homology searches, and thus expression profiles can be built up.
5 Tools for microarray analysis

The microarray technology has rapidly evolved, and is now employed in an almost standardised manner, with all reagents, printing robots and scanners commercially available. Analysis of the data obtained is, however, constantly being refined. The analysis of spotted microarray data can be broadly divided into three steps: image analysis, normalization, and selection of differentially expressed genes.

5.1 Image analysis

To evaluate the data obtained from microarray experiments, the array is physically scanned to create a digital image of the red and green fluorescence emissions (Cy5 and Cy3 respectively) from the array. Overlaying the output images of the Cy5 and Cy3 channels reveals physical information, such as spot morphology, hybridisation uniformity, and background artefacts such as dust particles. In addition, overlay images provide rough estimations of differentially expressed genes. After scanning, each spot must be located and linked to a clone ID. The primary purpose of the image analysis step is to calculate a foreground and a background intensity value for each spot, enabling adjustments for local variations in the array. Furthermore, the intensity values can flag for unreliable spots. The oldest method for spot intensity value determination is the histogram method (Chen et al. 1997), where a histogram is formed from the intensities of the pixels within a mask covering the spotted surface. Pixels are defined as foreground if their value is greater than a pre-set threshold, otherwise they are defined as background. Other strategies rely on finding spots as joined groups of foreground pixels, by fitting a circle of constant diameter to all spots in the array, or by allowing the circle’s diameter to change for each spot. Using these methods, the background values must be determined separately. One way of doing this is to consider all pixels outside the spots, but inside the bounding box, as local background. Once the intensity values have been estimated, the most common procedure is to subtract the background intensity from the foreground for each spot.
5.2 Normalization

The purpose of normalization is to adjust the individual hybridisation intensities so that relevant biological information can be obtained. Most normalization algorithms can be applied either to the whole array (globally) or to a subset of genes represented in the array (locally). Common factors to introduce red-green bias in a spotted microarray experiment are those related to labelling efficiencies and scanning properties. In addition, variations in different positions of the spotted area may occur, or even between different slides. Hence, normalization of the data obtained must be performed prior to any calculations of relative expression levels for the genes analysed, enabling further explorations of biologically relevant expression patterns. There are many approaches for normalization of spotted microarray data, some of those reviewed by John Quackenbush (Quackenbush 2002) and Gordon Smyth and colleagues (Smyth et al. 2003).

5.3 Selection of differentially expressed genes

The data obtained from microarray experiments are often used to screen for differentially expressed genes between one or more sample pairs. The general procedure is to choose a statistical method for ranking genes from high to low evidence of differential expression, followed by choosing a cut-off value above which the genes are determined to be significantly expressed (as reviewed in (Smyth et al. 2003)). When differentially expressed genes have been identified, a common approach is to group genes with similar expression profiles into clusters (Eisen et al. 1998), potentially revealing co-regulated genes with correlations not detected otherwise.
Present investigation
6 Pathogenesis of atherosclerosis

Atherosclerosis is an inflammatory disease that is believed to be the principal cause of death in modern society. Large and medium-sized blood vessels are mainly affected, and the atherosclerotic lesions tend to occur in regions with turbulent blood flow, such as branches, bifurcations and curved sections (Davies 1997; Ross 1999). One of many risk factors associated with atherosclerosis is elevated cholesterol levels in the vascular system. Under normal circumstances, cholesterol and its derivates function as membrane lipids or are stored as lipid droplets in cells for later use. The major carriers of blood cholesterol, low density lipoproteins (LDLs), constantly circulating in the vascular system and are associated with the buildup of cholesterol in
atherosclerotic plaques. Endothelial cells that line arteries transport LDLs into vessel walls (Kruth 2001), in both atherosclerotic and non-atherosclerotic states. Small and dense LDL particles have a high affinity for matrix proteins, like collagens, that are expressed following injury to endothelial cells (Heeneman et al. 2003) and are therefore easily trapped upon entering the arterial wall, causing diffuse arterial intimal thickening and progression into fatty streak lesions. In response, monocytes migrate and adhere to the surface of the thickened area, then transmigrate through the cells (Figure 7).

The monocytes are differentiated into macrophages, which express scavenger receptors on their surface, attract oxidised LDL (oxLDL) and are further transformed into lipid-rich foam cells. Foam cells also originate from vascular smooth muscle cells that have undergone phenotypic conversion into macrophage-like cells, thus mimicking their progression and transformation (Ricciarelli et al. 2000). Foam cells are the major components of the atherosclerotic plaques and, regardless of their origin, they tend to die as a result of apoptosis, because of the intracellular accumulation of free cholesterol (Kellner-Weibel et al. 1998). Macrophage-derived foam cells may also escape from the lesions into the peripheral circulation and die elsewhere through apoptosis (Takahashi et al. 2002). As they mature, atherosclerotic plaques may protrude into the lumen, narrowing the lumen of the artery. This may lead to ischaemic symptoms, although the most severe consequences are plaque rupture and thrombosis as a result of superficial erosion of the endothelium or uneven thinning and rupture of the lesion (Lee and Libby 1997; Rosenfeld 2000).
7 Differential gene expression in atherosclerosis

7.1 Treatment with a therapeutic drug candidate (Paper I)

In attempts to develop treatments for severe diseases, the pharmaceutical industries worldwide are constantly aiming to discover new drugs. The molecular mechanisms involved in the co-regulation of multiple genes that affect responses to these drugs are generally poorly understood. A possible way to elucidate the complex molecular interactions involved is via differential gene expression profiling. cDNA tag sequencing methods are preferable to microarray-based gene expression analysis for this purpose, since they provide absolute estimates of gene expression frequencies. Even better are techniques that focus on key fractions of genes being expressed, using “selective” techniques.

In Paper I, we describe how a solid-phase RDA technique can be used to elucidate the molecular effects of N,N’-Diacetyl-L-cystine (DiNAC) (Sarnstrand et al. 1999), an anti-atherosclerotic drug candidate. As a test system for this purpose, a monocytic cell line (THP-1) (Auwerx 1991) was used. The THP-1 cell line can be activated by various stimulants to differentiate into phenotypes mimicking macrophages in atherosclerosis. Here, THP-1 cells were activated with lipopolysaccharide (LPS), and compared with identical LPS-activated cells exposed to DiNAC.

Total RNA was extracted from the cells, mRNA was isolated using oligo-dT paramagnetic beads and cDNA was synthesised. The double-stranded cDNA, from both non-treated and drug-treated cells, was digested and used as starting material for multiple PCR-reactions to obtain approximately 150 mg DNA serving as representations in the RDA analyses. Three consecutive rounds of subtractive hybridisation were performed to obtain three difference products for both drug-treated and non-treated cells.
In this study we evaluated two alternative approaches to identify differentially expressed genes obtained after iterative rounds of RDA subtraction cycles. Previously, the most commonly used approach to select and isolate RDA fragments was to electrophoretically separate, excise, purify and clone them, assuming fragments generated in this way to be a representative set of the genes that are differentially expressed in the sets of cells or tissues under examination. Here, we used two different procedures to identify genes for which expression levels differed between the two materials. The first was the commonly used selection strategy, whereby we excised both distinct bands and band-patterned smears (size selection strategy), and the second was a shotgun approach in which the entire contents of the third set of differential products were cloned without any prior selection. A high number of different contigs (150 out of 197) were obtained from the size-selected fragments, demonstrating that the gene fragments in these products display a high degree of diversity. The analysis of the shotgun approach resulted in 54 out of 309 different contigs. These results suggest that the separation a complex mixture of fragments in the electrophoresis step may be inadequate to give a true reflection of quantitative differences between the test materials, and conclusions based on such separations may be somewhat misleading.

The obtained sequences were compared by BLAST (Altschul et al. 1990) to the nucleotide sequences included in UniGene (build 89) and the Expressed Gene Anatomy Database (EGAD). To verify that the obtained gene frequencies reflected genuine quantitative differences, real-time PCR was performed on a selection of gene fragments using the cDNA representations as templates.

The quality of the overall results of an RDA experiment is obviously dependent on the cloning strategy chosen to obtain the difference products, and our results suggested that the shotgun procedure has clear advantages. Hence, shotgun cloning approaches were adopted in the studies reported in all the following papers.
7.2 Foam cell formation in atherosclerotic lesions (Papers II and III)

Transcript profiling represents an important first step in understanding the diversity of cellular roles and mechanisms of genes. Several different methodologies have been developed for this purpose recently. Here, we used an RDA technique to analyse the early gene expression in macrophages accompanying the phenotypic changes into foam cells upon exposure to oxLDL. We have shown that shotgun RDA and large-scale DNA sequencing can be an attractive approach to monitor differential expression and that analysis of difference products can be analysed, to a certain extent, with high-throughput microarray techniques.

The monocytic THP-1 cell line was used again (Papers II and III) as a model system to study differential gene expression. However, this time the cells were stimulated with phorbol 12-myristate 13-acetate (PMA) to establish a macrophage phenotype, and then with oxLDL to trigger macrophage differentiation into foam cells. Total RNA was extracted from both oxLDL-treated and non-treated cells, mRNA was isolated using oligo-dT paramagnetic beads, and then cDNA was synthesised. A solid-phase RDA protocol was applied to the two different materials (treated and non-treated cells) with three consecutive rounds of hybridisation. The six sets of difference products were shotgun cloned and approximately 300 clones per differential product were randomly chosen and sequenced. The obtained sequences were compared by BLAST (Altschul et al. 1990) to the nucleotide sequences included in UniGene and EGAD. In parallel, a non-redundant set of clones from each data set was printed in triplicate onto amino-silane coated glass slides together with positive and negative control genes from the human and Arabidopsis thaliana genomes. Labelled targets (differential products) for hybridisation were generated by PCR in the presence of Cy3- or Cy5- labelled dCTPs. Scanning was performed using a confocal laser scanner and images thus obtained were analysed with GenePixPro 3.0 software.
The results revealed that around 70% of the assembled contigs of the third difference products comprised unique sequences (singletons). Also, approximately 20% of all gene fragments in the final difference products represented novel transcripts that had not been detected in the previous rounds of subtractive hybridisations. The substantial number of different gene fragments present after three rounds of subtractive enrichment demonstrates the complexity of gene regulatory events, as well as the RDA technique’s ability to detect rare transcripts. However, the relative expression levels derived using RDA may be misleading if an insufficient number of clones is sequenced and analysed. One way to obtain more exact estimates of expression rates could be to combine the RDA technique with large-scale microarray analysis. Microarray technology is a powerful tool, enabling expression profiles to be determined for thousands of genes simultaneously, although the detector’s sensitivity limits the ability to detect differences in the abundance of weakly expressed transcripts. It also requires prefabricated arrays harbouring spots representing all genes of interest, unless an array representing the total transcriptome of the organism is used. Until such microarrays are available/used for the organism under study, transcript-profiling methods that allow gene discovery (such as RDA) will yield information that would otherwise be missed.

The performance of the microarray assay in this study demonstrated both high specificity (no cross-hybridisation to negative control genes) and very high sensitivity, since 97% of the microarray elements repeatedly gave signals above the intensity threshold we set (local background plus two standard deviations). Also, as expected, the majority of the spots were red or green, indicating that they were differentially expressed, even though the expression levels of barely 32% of all replicates exceeded the minimum twofold expression ratio nominally required to accept expression as being differential.

The biological data derived in this study include information on genes that play crucial roles in cell cycle control and proliferation, inflammatory responses, several pathways that had not previously been implicated in atherosclerosis, and the peroxisome proliferator-activated receptor...
(PPARgamma) pathway, which has previously been implicated in the initiation and progression of atherosclerosis. Accumulating data suggest that PPARgamma plays a central role in the macrophage response to high extracellular concentrations of oxLDL (Tontonoz et al. 1998; Kersten et al. 2000). Several previously known PPARgamma target genes, e.g. the gene encoding adipophilin (Pelton et al. 1999), were identified and their up-regulation in the oxLDL-treated cells was confirmed. This was also the case for the class B scavenger receptor CD36, which is considered to play a critical role in atherosclerosis foam cell formation by mediating the uptake of ligands like oxidized lipoproteins (Tontonoz et al. 1998), apoptotic cells, and collagens.

In conclusion, we show that random sequencing of the difference products generated an accurate transcript profile and that regulations of the obtained gene fragments can be confirmed on a large-scale microarray analysis. The combination of these techniques enables significant differences in gene expression to be detected, even for weakly expressed genes and the results to be reliably validated in a high throughput manner.

7.3 Focal localisation of atherosclerotic plaques (Paper IV)

When exposed to sustained haemodynamic forces as a result of rapid (and, especially, turbulent) blood flow, changes occur in the vascular endothelium in terms of both structure and function. These changes in the vessel walls have great impact on the initiation and progression of atherosclerosis. It has been known for quite some time that regions with turbulent blood flow are more likely to develop atherosclerotic plaques than regions with more uniform blood flow (Figure 8).

In the studies reported in Paper IV, we used a solid-phase RDA protocol to investigate the focal nature of atherosclerotic lesions and gene expression profiling in vivo. The investigations were based on a comparison between
localisations that are likely, and others that are unlikely, to develop atherosclerotic plaques in the aorta in ApoE-/- and LDLR -/- mice. The aorta of each of these mice was cleaned of adipose tissue and dissected into plaque-prone localisations (the aortic arch and proximal part of the abdominal aorta) and plaque-resistant localisations (the descending thoracic aorta and distal part of the abdominal aorta). The tissues were snap-frozen in liquid nitrogen, total RNA was extracted and cDNA was synthesised using just 6 mg total RNA. The double-stranded cDNA from the two materials was used as starting material in multiple PCR-reactions to obtain a total of approximately 500 µg DNA for each material, serving as representations in the RDA protocol.

These representations, together with the first and second difference products generated by the RDA technique, were shotgun cloned and more than 400 clones from each dataset were sequenced. Each sequence was manually edited and clustered into contigs using the Staden software package (Staden 1996). This revealed that the number of clusters successively increased during the RDA procedure, showing enrichment of differentially expressed gene fragments (Table 2). Almost 2800 gene fragments potentially involved in the development of atherosclerotic lesions were compared by BLAST (using the E-value < 10^{-20}) to the representative nucleotide sequences included in UniGene (build 100), 52 % of which represented novel transcripts. To independently confirm the differential expression identified by RDA, a small subset of clones was selected for confirmation with real-time PCR using the cDNA representations as template. The results confirmed eleven out of twelve transcripts to be differentially expressed, showing the sensitivity and reliability of the RDA technique.
The expression levels of several of the obtained differential transcripts appear to be modulated by shear stress in the arteries. Such mechanotransduction preferentially occurs at specialised invaginated microdomains in the endothelial membrane, called caveolae. The function of caveolae has been debated, but it now seems clear that they are stable membrane domains that are kept in place by the actin cytoskeleton (van Deurs et al. 2003). Caveolae are important in the organisation of cell surface receptors and the regulation of various signal transduction systems, such as the system regulating cholesterol uptake. In this study we found increased expression of caveolin, the major structural element of caveolae, as well as coflin, an actin-binding gene, in the vessel localisations thought to be especially susceptible to plaque formation. Another up-regulated membrane protein, co-localised with caveolin, is CD36, which was also detected in the studies reported in Paper III.

Table 2. The distribution of clustered sequences, showing the enrichment of commonly expressed gene fragments. Data from upregulated genes in plaque prone regions of the mouse aorta (Paper IV).

![Cluster and Singleton Distribution](image)
As discussed above, RDA is a powerful technique for differential gene expression profiling, although it has several disadvantages. Since RDA relies on endonuclease restrictions of a pool of unknown cDNAs at specific restriction enzyme sites, fragments that only harbour one such site may be lost. Also, the majority of sequences in many databases represent the 3’- or 5’-ends of the cDNAs, while RDA generates fragments that are scattered throughout the cDNA region, except for the parts closest to their ends. Thus, identifying the obtained gene sequences may be problematic.

To address this problem we are developing a method (“Signature Tag RDA”) for identifying differentially expressed genes based solely on the 3’-ends of cDNAs. In order to do this we have combined RDA with a strategy developed for the amplification of cDNA tags (“signature tags”), in which the cDNAs are randomly fragmented into short tags of similar length, and the 3’-end (“signature tag”) population is then isolated and amplified by PCR amplification (Sievertzon et al. 2003). The strategy for the non-biased PCR amplification of 3’-end signature tags is outlined in Figure 9.

To study differential gene expression using this approach, the Lateral Hypothalamic Area (LHA) of very overweight and slightly overweight rats has been used as a model system. First, mRNA was isolated from LHA tissue with a designed 5’-biotinylated oligo-dT primer containing enzymatic restriction sites needed in subsequent steps of the protocol. cDNA synthesis was then performed, followed by random fragmentation of the cDNAs through sonication into 100-600 bp fragments. Biotinylated 3’-end signature tags from the fragmented cDNA population were isolated onto paramagnetic streptavidin-coated beads and the non-biotinylated fragments were removed. The ends of the immobilised signature tags were repaired and blunt end adaptors containing PCR primer sites and enzyme restriction sites suitable for RDA were ligated onto the 3’-end signature tags. The signature tags were released from the magnetic beads through NotI restriction digestion and then subjected to nested PCR amplification using primers designed in-house. The obtained pools of
**Figure 9.** The principle of Signature Tag RDA, based on identification of differentially expressed genes utilizing 3'-end signature tags.
amplified cDNA (for very overweight and slightly overweight rats) then served as representations for the RDA technique, which should theoretically represent the original transcripts expressed. Having obtained such representations the RDA technique can be used as previously described, except that all primers and linkers have to be designed and appropriate adjustments have to be made to the PCR conditions.

Using the signature tag strategy avoids the problems with RDA earlier discussed. This strategy relies on random fragmentation of cDNA populations followed by ligation of relevant adaptors suitable for RDA. Hence, the risk of losing fragments that only harbour one specific restriction site is avoided. Focusing entirely on the 3'-ends of the transcripts represented provides another huge advantage. Furthermore, fragmentation of the cDNA populations minimizes the risk of biased amplification due to the parallel amplification of templates of several different sizes.

9 Concluding remarks

The work in this thesis describes further developments of especially the representational difference analysis (RDA) technique for selective differential gene expression. This can be used independently as a tool for gene expression profiling, but has recently also been combined with global microarray analysis (Andersson et al. 2002), which indicates that combination of technologies can be an important complement for future efforts to identify differentially expressed genes.
Acknowledgements

Först och främst ett stort tack till alla som jag kanske glömmer räkna upp här (hemska tanke…).

Sen till övriga:

Joakim, för att du tagit dig an mig. För din exakta gräns mellan skvaller och hemlighet, och för att du mer än jag trott på att detta varit möjligt.

Mathias, för att du skapat och gjort denna grupp till vad den är.

Sophia, Per-Åke och Stefan, för att ni alltid tagit er tid med mina frågor när jag inte har haft någon annan att vända mig till.

Jacob som lärt mig mycket om mycket.

Anna för att jag alltid fått låna din hjärna när min egen totalhavererat.

Nina och Malin för alla ingående samtal om magar och barnen som till slut kommer ut ur dem.


Anders Thelin för all din hjälp och för att du lärt mig uttala ”anrika”.

Tove, min labbkompis, för all tid vi delat på resor och allt jobbsnack som man ju faktiskt måste ha ibland.

Lotta för din glada personlighet och att jag lyckats få in dig på fågelspåret.

PerU och Valtteri, för att ni alltid tagit er tid till alla mina frågor och faktiskt gjort det ni kunnat för att lösa mina problem så fort det nu varit möjligt. En inte helt vanlig egenskap, tack.
Till Karin, Jenny, Anna och Ingrid: Vi får grilla nästa sommar istället.

Annica och Anna W för enormt stöd de stunder jag behövt det.


Karin och Petra för alla våra fantastiska middagar.


Cloetta, för att ni tillverkar kexchoklad.

Anders för barnpassning och för att du är den bästa storebror man kan ha.

Föräldrar och svärföräldrar för att ni är bäst! Jag älskar er allihop.

Och så förstås min egen lilla familj. Peter för att du vill dela ditt liv med mig med allt vad det innebär. Mina älskade ungar, Andreas och förstås lilla knyten i magen som fortfarande har vett att stanna där inne fast mamma har varit lite stressad…
Subtracted Approaches to Gene Expression Analysis in Atherosclerosis

References


Subtracted Approaches to Gene Expression Analysis in Atherosclerosis


