KTH Biotechnology

Mining the transcriptome – methods and applications

Valtteri Wirta

Royal Institute of Technology,
School of Biotechnology
Stockholm, 2006

E-mail: vjwirta@kth.se

School of Biotechnology
Royal Institute of Technology
AlbaNova University Center
SE-106 91 Stockholm
Sweden

Printed at Universitetsservice US AB
Box 700 14
Stockholm

ISBN 91-7178-436-5

Valtteri Wirta (2006). Mining the transcriptome – methods and applications. Department of Gene Technology, School of Biotechnology, Royal Institute of Technology (KTH), Stockholm, Sweden

ISBN 91-7178-436-5

ABSTRACT

Regulation of gene expression occupies a central role in the control of the flow of genetic information from genes to proteins. Regulatory events on multiple levels ensure that the majority of the genes are expressed under controlled circumstances to yield temporally controlled, cell and tissue-specific expression patterns. The combined set of expressed RNA transcripts constitutes the transcriptome of a cell, and can be analysed on a large-scale using both sequencing and microarray-based methods.

The objective of this work has been to develop tools for analysis of the transcriptomes (*methods*), and to gain new insights into several aspects of the stem cell transcriptome (*applications*). During recent years expectations of stem cells as a resource for treatment of various disorders have emerged. The successful use of endogenously stimulated or *ex vivo* expanded stem cells in the clinic requires an understanding of mechanisms controlling their proliferation and self-renewal.

This thesis describes the development of tools that facilitate analysis of minute amounts of stem cells, including RNA amplification methods and generation of a cDNA array enriched for genes expressed in neural stem cells. The results demonstrate that the proposed amplification method faithfully preserves the transcript expression pattern. An analysis of the feasibility of a neurosphere assay (*in vitro* model system for study of neural stem cells) clearly shows that the culturing induces changes that need to be taken into account in design of future comparative studies. An expressed sequence tag analysis of neural stem cells and their *in vivo* microenvironment is also presented, providing an unbiased large-scale screening of the neural stem cell transcriptome. In addition, molecular mechanisms underlying the control of stem cell self-renewal are investigated. One study identifies the proto-oncogene Trp53 (p53) as a negative regulator of neural stem cell self-renewal, while a second study identifies genes involved in the maintenance of the hematopoietic stem cell phenotype.

To facilitate future analysis of neural stem cells, all microarray data generated is publicly available through the ArrayExpress microarray data repository, and the expressed sequence tag data is available through the GenBank.

Keywords: transcriptome, gene expression profiling, EST, microarray, RNA amplification, stem cells, neurosphere

LIST OF PUBLICATIONS

This thesis is based on the papers listed below, which will be referred to by their roman numerals

I. **Valtteri Wirta**, Anders Holmberg, Morten Lukacs, Peter Nilsson, Pierre Hilson, Mathias Uhlén, Rishikesh Bhalerao and Joakim Lundeberg. Assembly of a gene sequence tag microarray by reversible biotin-streptavidin capture for transcript analysis of Arabidopsis thaliana. BMC Biotechnology 5: 5, (2005).

II. Cecilia Williams*****, **Valtteri Wirta***, Konstantinos Meletis, Lilian Wikstrom, Leif Carlsson, Jonas Frisén and Joakim Lundeberg. Catalog of gene expression in adult neural stem cells and their in vivo microenvironment. Experimental Cell Research 312:1798-1812, (2006).

III. Maria Sievertzon, **Valtteri Wirta**, Alex Mercer, Konstantinos Meletis, Rikard Erlandsson, Lilian Wikstrom, Jonas Frisén and Joakim Lundeberg. Transcriptome analysis in primary neural stem cells using a tag cDNA amplification method. BMC Neuroscience 6:28, (2005).

IV. Maria Sievertzon, **Valtteri Wirta**, Alex Mercer, Jonas Frisén and Joakim Lundeberg. Epidermal growth factor (EGF) withdrawal masks gene expression differences in the study of pituitary adenylate cyclase-activating polypeptide (PACAP) activation of primary neural stem cell proliferation. BMC Neuroscience 6:55, (2005).

V. Konstantinos Meletis, **Valtteri Wirta**, Sanna-Maria Hede, Monica Nistér, Joakim Lundeberg and Jonas Frisén. p53 suppresses the self-renewal of adult neural stem cells. Development 133:363-369, (2006).

VI. Karin Richter*****, **Valtteri Wirta***, Lina Dahl, Sara Bruce, Joakim Lundeberg, Leif Carlsson, Cecilia Williams. Global gene expression analyses of hematopoietic stem cell-like cell lines with inducible Lhx2 expression. BMC Genomics 7:75, (2006).

***** These authors contributed equally to the work

All articles were printed with permission from the respective publisher.

Table of contents

Introduction

1. From genetic code to biological function

How is diversity generated, and what makes humans so different from other species? We share the same molecular building blocks – DNA, RNA and amino acids - as every other living organism on Earth, but obviously something distinguishes us from the rest.

For years it was thought that the number of genes might offer an explanation; the more genes the more advanced an organism would be, but today we know that this is not the case. Estimates of the number of human genes reached in the late 90's all the way up to 150,000 but sequencing of the genome has shown that we have no more than 24,000 protein-coding genes – only a few thousand more than the worm *Caenorhabditis elegans.*

The regulation of gene activity (gene expression levels) was already in 1975 identified as a potential explanation. King and Wilson compared gene sequences of humans and chimpanzees, and concluded: '*The intriguing result, ..., is that all the biochemical methods agree in showing that the genetic distance between humans and the chimpanzee is probably too small to account for their substantial organismal differences.*' and further that '*We suggest that evolutionary changes in anatomy and way of life are more often based on changes in the mechanisms controlling the expression of genes than on sequence changes in proteins. We therefore propose that regulatory mutations account for the major biological differences between humans and chimpanzees.*' (King, Science, 1975).

In addition to defining differences between species, regulation of gene expression also provides a fine-tuned control mechanism showing tissue-specific differences and controlling many biological processes in an organism. Everyday life wears out millions of skin cells daily; the intestine is constantly faced with different challenges, requiring ceaseless generation of new epithelial cells. Common for both these and most other tissues is the generation of new cells from undifferentiated tissue stem cells. These cells live a balanced life where the ratio of differentiation and self-renewal needs to be strictly controlled to ensure that all required cells are produced when needed throughout the life span of the organism. Gene expression plays a central regulatory role in controlling the balance between self-renewal and differentiation of these cells.

The central dogma of molecular biology – and a *dogma* it certainly has been – states that genetic information flows from genes, via RNA, to proteins (Figure 1A). Messenger RNA (mRNA) is generated in a process called transcription and is subsequently processed to yield a mature transcript. The maturation consists of several distinct steps, all of which are specifically regulated (Figure 1B). For many years it was assumed that the rate of RNA synthesis was the rate-limiting step indirectly also controlling the amount of protein synthesised. Today we know that the mRNA and protein levels do not always correlate. We further know that this is due to the extensive regulation of mRNA transcript processing and availability for translation – the mRNA can be sequestered in various cellular compartments and its degradation is regulated.

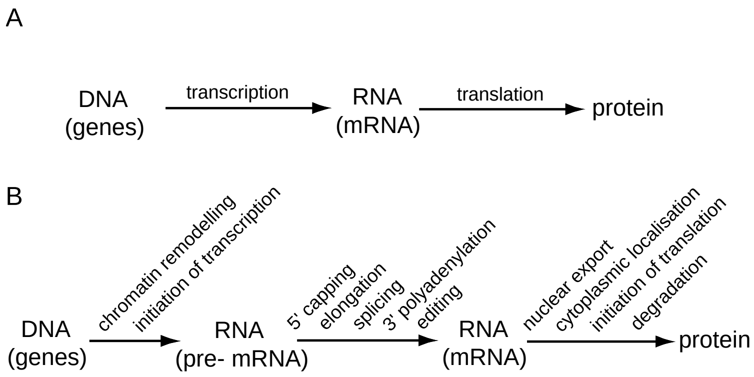

Figure 1. The flow of genetic information. A) The classical central dogma of molecular biology states that information in genes flows through an RNA intermediate to the proteins. B) Both transcription and translation are regulated at multiple steps.

Our view of RNA has dramatically changed during the last few years. Today we know that RNA actively functions as a regulator, a catalyser and a controller of several vital processes in the cell. These are functions that previously were attributed solely to proteins, but during recent years evidence for the role of RNA in these activities has emerged (Goodrich, Nat Rev Mol Cell Biol, 2006). The way the non-coding RNA (i.e. the type of RNA that does not encode proteins) functions can be summarised in three different ways: 1) binding through base pairing to target sequence, 2) folding on itself and catalysing a reaction (i.e. functioning as an enzyme), or 3) binding to a protein and modulating its activity. This is more complex than originally anticipated, but may in the end turn out to be more interesting and challenging.

What we need are methods to accurately quantify the levels of different types of RNA. This can be done in various different ways, including the use of microarrays to carry out a global analysis. Microarray-based gene expression analysis provides a snapshot of the expression levels of many (thousands) of the transcripts expressed in a cell.

This thesis presents six papers dealing with microarray-based gene expression analysis. Analysis of gene expression in (neural) stem cells is a central theme for most of the papers. The first paper describes a bead-based probe purification approach. The second paper describes an expressed sequence tag analysis of neural stem cells and generation of a ‘stem cell’ microarray. The third paper describes the use of a PCR-based amplification method for analysis of small amounts of RNA. The fourth and fifth papers analyse proliferation and self-renewal of adult neural stem cells isolated from the lateral ventricle wall of the mouse brain. In the last paper the role of Lhx2 in control of self-renewal of murine hematopoietic stem cell-like cells is discussed.

The *Introduction* provides a background summary of many aspects related to the papers forming the thesis. The work is focused on analysis of the protein-coding mRNA component of the transcriptome. However, to highlight the complexity of the RNA pool, the *Introduction* starts with an overview of several of the other RNA components in a cell (section 2). Next, a review of the current knowledge of transcription and transcript processing (section 3) is presented. This is followed by a discussion of alternatives to microarray-based methods for analysis of expression levels, starting with methods for analysis of individual transcripts (section 3.1), and followed by a description of tag-based methods for global analysis of gene expression levels (section 3.2). The *Introduction* ends with a review of microarray platforms (section 3.3) with focus on the spotted arrays, which are used in the papers included in this thesis.

2. RNA in a eukaryotic cell

2.1 Properties of ribonucleic acid

On both the chemical and structural level, RNA is similar to deoxyribonucleic acid (DNA) (Figure 2); both contain a pentose ring, a nitrogenous base and phosphate groups. The differences give rise to distinct properties that influence the way in which they are used by the cell. 1) Both use four nitrogenous bases that are attached to the 1' carbon of the pentose. DNA uses bases adenine (A), cytosine (C), guanine (G) and thymine (T), while RNA uses the first three bases and replaces thymine with uracil (U). In the DNA double helix the bases on the two opposite strands interact with each other through what has become known as Watson-Crick base pairing. Here, C forms three hydrogen bonds with G, and A forms two hydrogen bonds with T (Watson, Nature, 1953). 2) In contrast to DNA, the majority of the RNA in a cell is in single-stranded conformation and partially folds on itself (Littauer, Biochim Biophys Acta, 1959), influencing its stability and structure. 3) In RNA the 2' carbon of the pentose ring binds a hydroxyl (-OH) group. This renders the RNA molecule more flexible, but also more reactive (unstable) compared to DNA.

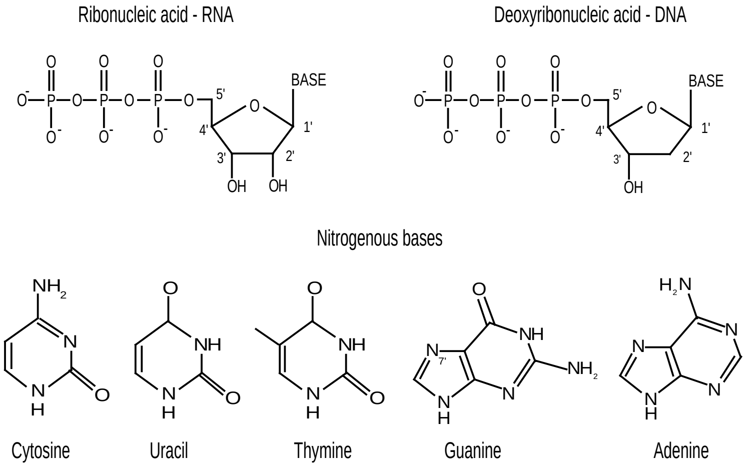

Figure 2. The structure of RNA and DNA and the five nitrogenous bases. RNA and DNA are shown as monomers with the tri-phosphate group attached to the 5' carbon. The bases are attached to the hydrogen-binding nitrogen atom shown at the bottom part of each base. The 7' position of the guanine becomes methylated in the 5' capping of an mRNA transcript (see later section for details).

4) The nitrogenous bases of RNA are frequently edited, increasing the complexity of the RNA molecule, and the diversity of reactions it can participate in. To date, more than 100 different RNA editing mechanisms are known, including both base substitutions (e.g. deamination of adenine to inosine), and base insertions and deletions ((Gott, C R Biol, 2003) and references therein). In many cases the effects of RNA editing are biologically important, as they introduce amino acid or reading frame changes, and introduce new open reading frames or introduce (or remove) stop codons. In addition the editing often has significant structural (e.g. pseudouridylation of all tRNAs) and functional effects (e.g. altered signalling properties for the 5-HT_{2C} serotonin receptor (Niswender, Ann N Y Acad Sci, 1998)) and may affect transcript splicing, transport (Zhang, Cell, 2001) and stability (Bass, Annu Rev Biochem, 2002). Interestingly, the base editing processes show tissue specific, developmental, hormonal and environmental regulation (Keegan, Nat Rev Genet, 2001).

2.2 Many flavours of RNA

The transcriptome is the combined set of all transcripts present in a cell at a certain time-point. mRNA, although widely studied and the focus of the present work, is only a minor component of the entire RNA population. Some years ago this population was considered to consist of highly abundant ribosomal RNA, transfer RNA and of rare protein-coding mRNA. During the last two decades evidence for the existence of other functionally and structurally important RNA molecules has been presented. Most intriguingly, research during the last few years has shown that only a fraction of the transcribed loci generate protein-coding transcripts and that almost the entire genome is transcribed (Carninci, Science, 2005).

This chapter summarises the different classes of RNA molecules (excluding mRNA, which will be discussed later), and briefly highlights their functional role. It is important to acknowledge that our understanding of RNA is still not complete, and that other functions will undoubtedly be discovered.

2.2.1 ribosomal RNA

The function of ribosomal RNA (rRNA) - the most abundant form of RNA in a cell – is to participate in and provide the structural framework for translation. The ribosome constitutes a factory site where amino acids carried by the tRNAs, through the ribosome's peptidyl transferase activity, are added to the nascent, growing polypeptide chain. The different steps in translation (initiation, decoding, peptidyl transferase reaction, translocation, and termination) are reviewed in detail in (Ramakrishnan, Cell, 2002).

The rRNA and the approximately 50-80 different ribosomal proteins assemble into two main subunits in all kingdoms of life (40S and 60S in eukaryotes). The structure of the rRNA provides a three-dimensional scaffold to which the proteins bind in a specific way, creating a highly ordered and spatially restricted unit. Three of the four rRNA transcripts are transcribed as one rRNA precursor in the nucleolus by RNA polymerase I (Pol I). The nucleolus is a non-membranous subcompartment of the nucleus, and the site where Pol I-dependent transcription and processing of rRNA takes place. Following transcription, the precursor is processed in at least ten distinct cleavage steps to generate the mature rRNA transcripts. The fourth rRNA transcript is transcribed by RNA polymerase III (Pol III). All rRNA transcripts are multiply encoded in the genome, providing sufficient amounts of transcription template to sustain a rapid cell growth (Prokopowich, Genome, 2003).

Determination of the ribosome structure (Ban, Science, 2000; Nissen, Science, 2000; Schluenzen, Cell, 2000; Schuwirth, Science, 2005; Wimberly, Nature, 2000; Yusupov, Science, 2001) showed that both riboproteins and rRNA are indispensable for the proper ribosome function (Wilson, Crit Rev Biochem Mol Biol, 2005). The structural characterisation also showed that the four rRNA bases participate to a varying degree in non-canonical base pairings; 30% of C, G and U and 62% of A are unbound or participate in non-canonical base pairing (i.e. base pairing that does not follow the Watson-Crick base pairing scheme) (Gutell, Proc Natl Acad Sci U S A, 1990; Noller, Science, 2005). However, even though the basic structural motifs for rRNA are the double-stranded helices, very few examples of contiguous Watson-Crick base pairing longer than 7 bp exist in the ribosome (Noller, Science, 2005). This makes it possible for the RNA to form a highly complex three-dimensional structure capable of complex interactions with a large number of proteins. These findings will probably also have important implications in other aspects of RNA biology.

2.2.2 transfer RNA

transfer RNA (tRNA) are small (~80 bases in length), heavily modified RNA molecules that each carry one single amino acid to the ribosome. tRNAs are highly abundant in a cell, for example during every yeast generation approximately 3-6 million tRNAs are produced. Each tRNA molecule contains four regions of intramolecular double helices formed by

Watson-Crick base pairing and three loops (D-, anticodon- and T-loop). The solving of the tRNA crystal structure in 1974 (Kim, Science, 1974; Robertus, Nature, 1974) showed that non-canonical base pairing, mediated by the hydroxyl group at the 2' carbon in the ribose, participates in creating the unique three-dimensional structure (Noller, Science, 2005). tRNAs are extensively modified before becoming fully mature: their 5' leader sequence is removed, the 3' trailer sequence is trimmed, the nucleotide CCA trimer is added to the 3' end, a large number of the bases are edited, and introns spliced. This processing requires more than 60 different proteins and includes several quality control steps. Recent work has also shown that several quality control steps ensure that only fully processed tRNAs are available to the ribosome and protein synthesis (Kadaba, Genes Dev, 2004), and that – surprisingly - retrograde transport of tRNA back into the nucleus takes place (Shaheen, Proc Natl Acad Sci U S A, 2005; Takano, Science, 2005).

2.2.3 small nucleolar RNA

A class of small nucleolar RNAs (snoRNA) is found in many eukaryotes and in archaea, but not in bacteria. To date more than 230 different snoRNAs are known (Griffiths-Jones, Nucleic Acids Res, 2005). snoRNAs are typically 60 to 300 bases long and are subdivided into two groups based on their secondary structure: C/D-box and H/ACA-box snoRNAs. The C/D-box snoRNAs bind to their target sequences (see below) through a 10-21 bp double helix and promote 2'-O-methylation at a position five bases upstream of the binding site. The H/ACA-box snoRNAs promote pseudouridylation through binding to the target sequences at two 3-10 bp regions and induce base editing at a position which is 15 bases upstream (Mattick, Hum Mol Genet, 2005).

snoRNAs are in most cases generated from introns of RNA polymerase II-transcribed mRNA through exonuclease activity, and were earlier considered as 'junk' RNA. In yeast and plants some snoRNAs are transcribed as polycistronic transcripts (see references in (Lau, Science, 2001)). Some snoRNAs are expressed in tissue-specific manner, which may potentially turn out to be of importance in the generation of specific response patterns for a variety of tissues (Cavaille, Proc Natl Acad Sci U S A, 2000; Runte, Hum Mol Genet, 2001). The primary targets of snoRNAs are the rRNA transcripts, snRNAs, and in some cases mRNA transcripts. It is also important to note that several of the snoRNAs are still considered orphan, i.e. they lack an identified target transcript.

2.2.4 small nuclear RNA

small nuclear RNA (snRNA) is a class of small eukaryotic RNA molecules transcribed by RNA polymerases II and III and found in the nucleus. snRNAs are involved in catalyzing the splicing reaction, and are important components of the splicesome (see chapter on splicing), providing the recognition of the splice sites at the exon-intron boundaries. One of the snRNAs (U1) has also been discovered to stimulate Pol II transcription through binding to the TFIIH subunit of the general transcription machinery (Kwek, Nat Struct Biol, 2002), providing evidence for an extensive interplay between the different steps of RNA transcription and processing (see later sections).

2.2.5 microRNA

The class of small RNA molecules that has been the 'rising star' of the decade is undoubtedly microRNAs (miRNA). miRNAs bind to mRNA transcripts, often at the 3' end, and exert post-transcriptional control of gene expression, either through translational repression or endonucleotic transcript cleavage. The initial evidence for RNA-mediated transcriptional repression came from studies in *C.elegans* by Fire *et al* in 1998, who, by injection of dsRNA, could promote transcriptional repression of specifically targeted genes (Fire, Nature, 1998). In principle, these small transcripts, which have later been found in all mammals, add a new level to regulation of gene expression (Lau, Science, 2001; Zamore, Cell, 2000); miRNAs establish tissue- and developmental-specific effects on the gene expression, without altering the transcription of the primary protein-coding mRNA transcript, *per se*.

The biogenesis of miRNAs starts from precursors containing a stem-and-loop region, which is cleaved by the RNase III endonuclease Drosha (RNASEN in *Homo sapiens*) in the nucleus. This results in a partially double-stranded, 65-75 bp pre-miRNA that is transported to the cytoplasm. Next Dicer, another RNase III endonuclease, generates the mature miRNA by cleavage of the pre-miRNA into a 21-22 bp fragment with two nucleotide (nt) overhangs at the 3' ends. Based on still partly unknown rules, one of the two strands is incorporated into the RNA-induced silencing complex ribonucleoprotein, which upon binding to the target mRNA regulates its translation and half-life.

To date 462, 367 and 228 miRNAs have been identified in humans, mice and rats, respectively (August 2006, release 8.2 of miRBase Sequence Database) (Griffiths-Jones, Nucleic Acids Res, 2006). These numbers are probably large underestimates as the current miRNA prediction algorithms are biased towards previously identified 'classical' miRNA structures and the frequency of false negatives is probably high (Mattick, Hum Mol Genet, 2005). Furthermore, many of the prediction algorithms look for sequences which are highly conserved across species, introducing further bias. To predict target mRNAs, an analysis carried out using four genomes and based on identification of regions in the 3' UTRs of mRNA transcripts complementary to the miRNA 2-to-7-nt seed region predicted that, in humans, there are >5,300 miRNA target transcripts (Lewis, Cell, 2005). The seed region is the stretch of nucleotides often found in the 5' end of the miRNA and which plays a key role in target recognition.

Recent research has shed light on many aspects of miRNA biology. Interestingly, families of miRNAs (based on the characteristic 5' seed of the miRNA) showed coordinated expression patterns in one skin lineage compared to another (Yi, Nat Genet, 2006), indicating a regulatory network that may efficiently target and post-transcriptionally regulate the expression of certain target transcripts. Recent work by Wu *et al* showed that at least some miRNAs promote deadenylation of the mRNA transcripts and thereby decrease the abundance of the messages (Wu, Proc Natl Acad Sci U S A, 2006). A striking example of this has been shown in zebrafish, where deadenylation of maternal transcripts in fertilized zygotes is mediated by the miR-430 family, which is expressed at the onset of the zygotic transcription. The miRNA-mediated deadenylation facilitates on global-scale the maternal-to-zygotic transition in early embryogenesis (Giraldez, Science, 2006). Another twist to the gene expression regulation by miRNAs is added by the findings showing that the adenosine deaminases (ADARs; convert adenosine to inosine) act on selected pre-miRNAs, and both change their target specificity through base editing and also promote rapid degradation of the miRNA (Yang, Nat Struct Mol Biol, 2006). There is also evidence supporting a pathogenic role for certain miRNAs; miR-372 and miRNA-373 have been shown to be oncogenic through suppressing an inhibitor of cyclin-dependent kinase activity, which in this case causes testicular germ cell tumours (Voorhoeve, Cell, 2006).

2.3 The life cycle of a eukaryotic messenger RNA transcript

RNA was described in 1961 as the '*unstable intermediate carrying information from genes to ribosomes for protein synthesis*' (Brenner, Nature, 1961). This statement still holds, but today we know that it is the messenger RNA (mRNA) that carries out this vital function. Transcription of an mRNA transcript is carried out by an RNA polymerase, but the maturation of a eukaryotic mRNA transcript is carried out in a multi-step process requiring hundreds of different proteins. The stages in the life cycle of an mRNA include transcription, addition of a protective 5' cap, polyadenylation, splicing, nuclear export, translation and, finally, degradation (Figure 3).

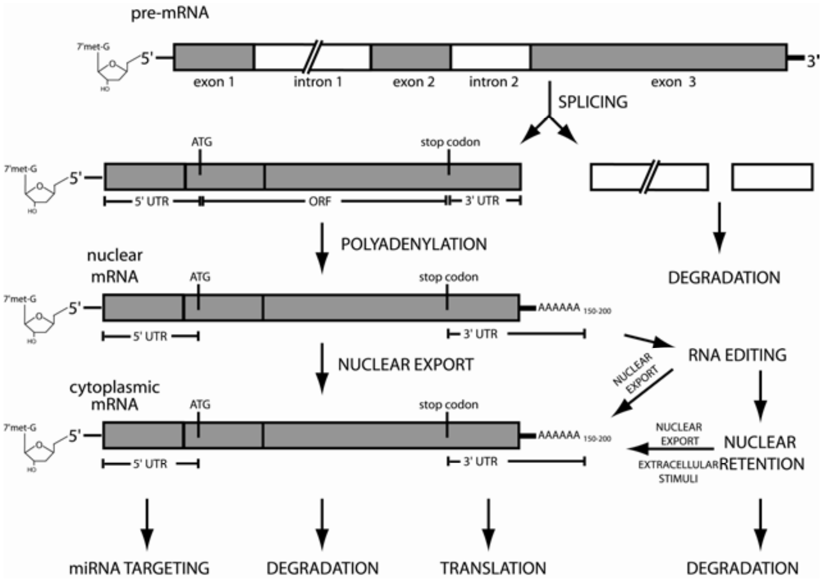

Figure 3. The life cycle of a eukaryotic mRNA transcript. ATG, starting methionine codon for translation; ORF, open reading frame; UTR, untranslated region.

Our understanding of the molecular steps controlling transcription and other mRNA maturation steps, as well as the spatial and temporal aspects, has dramatically changed during the last 5 to 10 years. The current model bears little resemblance to the knowledge that had accumulated by the mid 1980's (Kadonaga, Cell, 2004). Even when browsing through old undergraduate level molecular biology text books (late 90's) one realises that several fundamental concepts have been extensively revised. What was a simplistic picture of the different mRNA processing steps being separate and described as individual and compartmentalised reactions following each other, is today described as interactive steps dependent on each other and showing a shared use of resources and taking place in parallel and often at the same physical location.

In the following chapters eukaryotic transcription is reviewed. Only transcription carried out by RNA polymerase II (Pol II), which yields the protein-coding RNA transcripts, will be described.

2.3.1 Initiation of transcription

In eukaryotes transcription initiation is controlled by sequences both upstream and downstream of the first transcribed base (transcription start site, TSS). An extensive study of >500,000 TSSs in both human and mouse analysed these sequences (termed promoters) and categorised them into two main groups: 1) single dominant peak class promoters (termed SP) that often (~20%) contain a TATA-box and 2) broader promoters with multiple TSSs (termed BR) and that are CpG-rich and contain no TATA-boxes (Carninci, Nat Genet, 2006).

Interestingly, the study showed that the TATA-box containing SR type of promoter is less common, even though classically more extensively studied. The TATA-box is a cis-regulatory element consisting of a stretch of thymidine and adenosine nucleotides and located ~30 bp upstream of the TSS. SR promoters show high cross-species conservation and are commonly used to regulate tissue-specific gene expression. The second promoter group shows usage of multiple TSS and less evolutionary conservation, and is often related to ubiquitous gene expression. There are however exceptions to this; embryo and brain-specific genes show often use of BR type of promoters. Furthermore, CpG-rich promoters are susceptible to epigenetic alterations (e.g. methylation), which may be an important determinant for imprinting (i.e. the phenomenon whereby expression is turned off in the allele inherited from either mother or father). Interestingly, some CpG-rich promoters are bidirectional, providing a possible explanation for the large extent of sense and antisense transcription observed during recent years (reviewed in (Katayama, Science, 2005; Mattick, Hum Mol Genet, 2006)). The Carninci *et al.* study also provided, for the first time, evidence for large-scale use of exonic TSS, conserved between mouse and humans and generating a large abundance of truncated non-coding RNAs (Carninci, Nat Genet, 2006). Lastly, the TSSs indicate that transcription of approximately 58% of all protein-coding genes is initiated from two or more promoters, and that alternative methionines are used as translation initiating codons in 93% of cases with alternative promoter usage.

Initiation of transcription is mediated through binding by sequence-specific transcription factors (activators and repressors) to the promoter region of a gene and to other sequence elements in both proximal and distal positions relative to the TSS. These factors collectively recruit a partly pre-assembled Pol II complex consisting of several subunits and with a size of >500 kDa. One of the key components of Pol II, controlling transcription and association with other RNA processing machinery acting on the nascent transcript, is the carboxyl tail domain (CTD), which consists of 52 heptameric repeats. The overall efficiency of transcription is increased by the CTD, which at least partly functions to recruit other RNA processing enzymes to close proximity of the nascent transcript. This has been verified in experiments carried out with CTD-mutant Pol II variants and where reduced transcription efficiencies have been observed (Gerber, Nature, 1995; Lux, Nucleic Acids Res, 2005).

2.3.2 Capping of the 5’ end

Spatially, the phosphorylated CTD of the transcribing Pol II is located in close proximity to the nascent, 18-30 nt pre-mRNA transcript exiting the inner core of the polymerase, and it recruits the machinery required to process the 5’ cap structure of the pre-mRNA (Cho, Genes Dev, 1997; McCracken, Genes Dev, 1997). This setup facilitates early capping of the nascent pre-mRNA, providing immediate protection from 5’->3’ exonuclease-mediated transcript degradation. Capping generates an unusual 5’-5’ phosphodiester bond between the first nucleotide of the transcript and a 7’-methyl guanosine ribonucleotide (Shatkin, Cell, 1976) (Figures 2 and 3). In addition to providing protection from degradation, the cap structure is required for at least splicing, nuclear export and translation. The capping apparatus is conserved in all eukaryotes and contains three distinct enzymatic activities; 1) hydrolysis of the 5’ nucleotide to a diphosphate, 2) addition of a guanosine through a 5’-5’ bond, and 3) addition of a methyl group to the 7’ carbon of the guanosine base (Shuman, Prog Nucleic Acid Res Mol Biol, 2001).

2.3.3 Transcript elongation

During elongation Pol II extends the nascent transcript by approximately 30 nucleotides per second (Shilatifard, Annu Rev Biochem, 2003). This rate is however not achieved without encountering problems in the physical arrangement of the DNA template; in the nucleus the 3 billion bp of DNA are packed into chromatin. The smallest element of the chromatin structure, the nucleosome, is well characterised and consists of 146 bp of DNA wrapped 1.65 times around four pairs of histone molecules (Kornberg, Science, 1974; McDonald, Nature, 2005). The physical packing of DNA restricts transcription and has to be partly removed during elongation. Data suggests that an H2A-H2B dimer is removed from the nucleosome in front of the elongating Pol II, and that a histone pair is reattached afterwards (Belotserkovskaya, Science, 2003). Occasionally a different pair of histones is reattached, providing transcription-dependent remodelling of the chromatin.

2.3.4 Processing of the 3' end

The 3' end of a pre-mRNA is processed to yield a polyA-tail, typically 200-250 nt in mammals and 70-90 nt in yeast. The tail has several important functions; it is involved in export from the nucleus (transcripts lacking polyA-tail are commonly retained in the nucleus), translation and, most importantly, in control of the transcript turnover. The processing involves endonucleotic cleavage of the pre-mRNA, polyadenylation of the cleaved transcript and termination of the Pol II transcription. The cleavage is sequence specific and dependent on the presence of a polyadenylation signal, AAUAAA, and a downstream CA dinucleotide tag. The cleavage and polyadenylation specificity factor (CPSF) recognises and binds to the polyadenylation signal hexamer, recruits the cleavage stimulatory factor (CSF), two cleavage factors (CFI and CFII), and a poly(A) polymerase. After cleavage the polymerase extends the pre-mRNA with 200-250 adenosine nucleotides at the 3' end. The number of adenosines incorporated is determined by nuclear polyA-binding proteins that bind to the nascent polyA-tail through their RNA-binding domain. Pol II continues transcription of the 3' fragment of the cleavage reaction by >500 nt after the polyadenylation signal and then terminates, reviewed in (Rosonina, Genes Dev, 2006). The 3' fragment generated by the cleavage reaction is uncapped, and rapidly degraded in the nucleus.

It is important, however, to note that not all transcripts are polyadenylated. Detailed analysis of transcription from ten different human chromosomes using tiling arrays (microarrays that interrogate essentially every non-repetitive base of the genome) showed that only a fraction (19.4%) of all transcripts are polyadenylated ((Cheng, Science, 2005)). More research is needed to reveal the functional role and translational status, if any, of the transcripts without polyadenylation.

2.3.5 Splicing

In eukaryotes most pre-mRNA transcripts contain non-coding intron sequences that are removed in a splicing reaction (Berget, Proc Natl Acad Sci U S A, 1977; Chow, Cell, 1977). The length of the average human gene (27 kbp) is much more than the average size of the spliced, mature protein-coding mRNA (1.5 kb) (Lander, Nature, 2001). The number of introns is highly variable between different genes; olfactory receptors typically contain no introns, while the 2.4 Mbp dystrophin gene contains >75 introns. The majority of the genes contain a more modest number of introns, on average estimated to seven per gene (Lander, Nature, 2001).

Mechanism and regulation. Splicing involves two transesterification reactions that create a spliced mRNA and a lariat structure (Staley, Cell, 1998). A successful splicing reaction requires three stretches of nucleotides; a 5' splice site at the boundary between the 5' exon and the intron, a 3' splice site, and a branch site within the intron. The reaction is initiated with the 2'-OH group of the branch site adenosine attacking the phosphodiester bond of the 5' site, yielding a 3'-OH group at the 5' site. Next, this hydroxyl group attacks the phosphodiester bond of the 3' site, releasing the lariat (intron) and creating a normal phosphodiester bond between the 5' and 3' exons. In the nucleus these reactions are carried out by the splicesome, a multiunit riboprotein assembly containing five snRNA-protein complexes (snRNPs; one snRNA and multiple proteins in each) and additional 100+ proteins (Jurica, Mol Cell, 2003). The splicesome has a dynamic structure and involves several RNA:RNA, RNA:protein and protein:protein interactions; the key players are the U1 and U2 snRNPs that bind to the 5' and branch sites, and the U6 snRNP which is thought to be the catalytic component (Faustino, Genes Dev, 2003).

The transcriptome contains >10-fold excess of pseudo-splice sites (containing potential 5' and 3' splice sites and a branch site) that go unspliced. The *bona fide* splice sites are recognised and differentiated from the pseudo-sites using clusters of exonic splicing enhancers (ESE) and intronic splicing silencers (ISS) (Sun, Mol Cell Biol, 2000). ESE are commonly, but not exclusively, bound by serine and arginine rich proteins, while ISS are bound by heterogeneous nuclear ribonucleoproteins. Recent evidence however points to both classes of binders affecting splicing in both activating and repressing manner (Blanchette, Genes Dev, 2005), and these binders can be further covalently modified, providing an additional level of regulation (Shin, Nat Rev Mol Cell Biol, 2004).

Alternative splicing. In splicing, exons are always joined from 5' towards the 3' end of the pre-mRNA transcript. Under certain conditions one or more exons can be skipped, generating different patterns of exon joining – alternative splicing. In humans it has been estimated that 60-80% of the genes are alternatively spliced, that 30% show usage of alternative 3' exons, and that 80% of the splicing reactions change the protein sequence (Modrek, Nat Genet, 2002). Large numbers of exon combinations are possible for certain genes; the Down syndrome cell adhesion molecule gene in *Drosophila* and the neurexins and *CD44* genes in human can produce as many as 38,000, 3000 and 1,000 different isoforms, respectively (Schmucker, Cell, 2000; Zhu, Science, 2003). The extent of alternative splicing is also believed to be more important in certain tissues, such as the nervous system (Lee, Biol Psychiatry, 2003). Furthermore, dysregulated splicing has been associated with several diseases, e.g. certain types of cancer (Venables, Cancer Res, 2004).

Alternative splicing generates changes at the protein level through exon skipping, frameshift or downregulation of the transcript (through NMD, see later section). Typically, in the case of exon skipping, complete functional domains are affected, e.g. skipping of a transmembrane domain yielding a soluble protein (Xing, FEBS Lett, 2003). In more rare cases a new functional domain can be created; e.g. individually exons 2 and 4 encode a part of a nonfunctional transmembrane helix while skipping of exon 3 joins the exons 2 and 4 and generates a functional helix (Hiller, Genome Biol, 2005). Examples of alternative splicing modulating ligand specificity of growth factor receptors and adhesion molecules have also been observed (Lopez, Annu Rev Genet, 1998).

Regulation of alternative splicing is believed to be controlled through use of tissue-specific splicing factors binding to ESE and ISS elements, presumably binding to Pol II already at the transcription initiation phase, and through kinetics of the transcription. Experimental evidence for the model based on transcription kinetics has been generated by experiments where pausing during elongation, different classes of transcription activators, and more slowly elongating Pol II mutants have been used (de la Mata, Mol Cell, 2003; Eperon, Cell, 1988; Kadener, Proc Natl Acad Sci U S A, 2002; Nogues, J Biol Chem, 2002). In an *in vitro* study evidence for the kinetics of transcription affecting the choice of alternative splicing pattern was obtained; splicing of proximal exons is favoured initially during transcription, but over time more distal splicing takes over (Hicks, PLoS Biol, 2006). Supporting this, both experimental and computational approaches have concluded that the length of the flanking upstream intron influences the pattern of alternative splicing (Fox-Walsh, Proc Natl Acad Sci U S A, 2005).

In general, spliced transcripts are more abundant than unspliced transcripts. Data suggests that this is due to the increased protection from nuclear degradation provided by coating of the transcript with splicing factors and splicing more clearly directing the transcript to the mRNA processing pathway (Hicks, PLoS Biol, 2006). Most importantly, at every exon-exon junction the splicing machinery leaves a protein complex termed exon junction complex (EJC). This complex has been associated with mRNA processing steps such as nuclear export, translation (Matsumoto, Embo J, 1998), localisation (Hachet, Nature, 2004) and, crucially, transcript quality control and turnover (see chapter on RNA degradation).

2.3.6 Transcription factories

Where does Pol II-dependent transcription occur in the nucleus? The classical view suggests no distinct nuclear clustering of transcriptionally active sites. However, the identification of distinct transcriptionally active foci prepares the ground for a different view (Iborra, J Cell Sci, 1996; Jackson, Embo J, 1993; Wansink, J Cell Biol, 1993). In fact, each cell contains much fewer active sites of transcription than there are transcribed genes. Furthermore, co-transcription of genes 40 Mbp apart has been described (Osborne, Nat Genet, 2004). An alternative view of transcription localisation, based on the concept of 'transcription factories', has emerged to incorporate these distinct foci of transcription. Instead of Pol II moving along the chromatin, the chromatin moves through the transcription factory, which is 'semiattached' to the nuclear matrix (Jackson, Embo J, 1985). Energy released from NTP hydrolysis during polymerisation could be used to pull the chromatin through the factory (and there is certainly that amount of energy and force available (Yin, Science, 1995)). Enhancers and locus control elements can counteract this pulling and keep the gene longer in the factory, and hence increase the number of transcripts generated. This model is in agreement with results suggesting that large proportions of the genome are transcribed (Carninci, Science, 2005). To further facilitate pre-mRNA maturation, the various enzymatic complexes required for pre-mRNA processing may also be localised at the same factory foci, which would more or less immediately after transcription direct an mRNA molecule into the processing machinery (Osborne, Nat Genet, 2004). It is important to point out that many details of the transcription factory model are still unverified.

2.3.7 Nuclear mRNA export

The nucleus is surrounded by a phospholipid bilayer, effectively compartmentalising the nucleus from the cytoplasm. Movement of protein-bound mRNA (mRNP) inside the nucleus has recently been shown to occur through random diffusion, and it was further shown that adenosine triphosphate (the hallmark of active transport) is needed for reinitiation after mRNP particles stall in regions of dense chromatin (Vargas, Proc Natl Acad Sci U S A, 2005). Nuclear pore complex (NPC) is a large 50 MDa (yeast) to 125 MDa (human) assembly of proteins that facilitates transport of macromolecules through the membrane (Suntharalingam, Dev Cell, 2003). Particles smaller than 40 kDa diffuse through the NPC, while active transport is required for those of larger size. Components of the NPC also carry out quality control functions to ascertain that only completely spliced transcripts are exported (Galy, Cell, 2004).

Retention of a subset of polyadenylated transcripts in the nucleus is known to occur (Herman, Cell, 1976) and some of these transcripts are functional in the nucleus (e.g. *Xist* and its antisense transcript *Tsix* (Plath, Annu Rev Genet, 2002)). A recent study described the retention of a partly processed, adenosine-to-inosine edited transcript that, upon extracellular stimulation, was rapidly cleaved, exported to the cytoplasm and translated (Prasanth, Cell, 2005). Importantly, this study reveals an entirely new level of posttranscriptional mRNA regulation, which may be much more widely used than previously thought.

2.3.8 Translation

In translation the ribosome converts the genetic information, carried in an mRNA transcript, into a peptide by polymerising individual amino acids. Ribosomes are found 'free' in the cytoplasm, bound by the ER, in the mitochondria, and possibly in the nucleus (Iborra, Science, 2001; Iborra, J Cell Sci, 2004). The number of ribosomes is balanced in comparison to the number of transcripts; during rapid growth in yeast, 70-80% of transcripts are bound by at least one ribosome and approximately 85% of the ribosomes are engaged in translation (Arava, Proc Natl Acad Sci U S A, 2003; MacKay, Mol Cell

Proteomics, 2004). The number of ribosomes actively translating a given transcripts varies. For highly translated transcripts there can be one ribosome for every 30 bases, indicating an efficient – but also crowded – process. This is close to the maximal theoretical density, as the length of the RNA buried inside a ribosome is approximately 30 nt. At the other extreme, ribosome densities of 1/1000 nt are not uncommon (Arava, Proc Natl Acad Sci U S A, 2003; MacKay, Mol Cell Proteomics, 2004). Most transcripts are translated at a frequency somewhere between these extremes, indicating that the initiation of translation is the rate limiting step. For longer transcripts the density goes down, indicating less efficient initiation with growing transcript length (Arava, Nucleic Acids Res, 2005).

Control of the translation efficiency is carried out at both global and transcript-specific level. The former is achieved through control of the number of ribosomes and phosphorylation of translation initiation factors. Transcript-specific efficiency is modulated through binding of various RNA-binding translation factors, and through structural elements in the untranslated regions of some transcripts. For example, under stress or comparable conditions, a global translation downregulation takes place, while the translation efficiency for certain specific transcripts is highly increased. An example of this is the induced oncogenic signalling through Ras and Akt-signalling pathways in glioblastoma; during stimulation certain transcripts become much more frequently translated than normally. This regulatory effect on translation is in fact much larger than the upregulation on the transcriptional level (Rajasekhar, Mol Cell, 2003).

2.3.9 RNA degradation

Steady-state levels of mRNA transcripts can be achieved through use of stable transcripts and low transcriptional level, or through high transcriptional activity balanced by an efficient transcript degradation activity. However, only the latter allows for rapid adjustment to a change in the cell or its environment and this is the approach chosen by most organisms; half-lives for unstable transcripts in *E.coli,* yeast and mammals are typically only 1-2 min, 2-3 min, and 15 min, respectively (Herrick, Mol Cell Biol, 1990; Shyu, Genes Dev, 1989; Wang, Proc Natl Acad Sci U S A, 2002). Transcripts encoding transcription factors and other regulatory proteins are usually short lived, while transcripts for metabolic pathway enzymes are long lived. Also, members of the same functional class or macromolecular complex often have similar decay rates (Wang, Proc Natl Acad Sci U S A, 2002), and transcripts encoding orthologous genes typically have similar turnover patterns (McCarroll, Nat Genet, 2004). Collectively, these observations demonstrate that similar mechanisms are widely used to control transcript levels and that the half-life of a transcript reflects the function of the encoded protein.

The default state of an mRNA transcript is 'relative stability', and degradation is induced through either 5' decapping or 3' deadenylation. In both yeast and mammals, the first step is shortening of the polyA-tail (to 10 and 60 nt, respectively) (Chen, Mol Cell Biol, 1994). In yeast this is followed by removal of the cap, allowing for a 5'-3' exonucleatic degradation (Hsu, Mol Cell Biol, 1993; Muhlrad, Genes Dev, 1994). In mammals, a 3'-5' pathway, termed the exosome pathway, is the dominant mRNA degradation pathway ((Chen, Cell, 2001; Mukherjee, Embo J, 2002; Wang, Cell, 2001); reviewed in (Houseley, Nat Rev Mol Cell Biol, 2006; Meyer, Crit Rev Biochem Mol Biol, 2004)). The exosome is an evolutionarily conserved multiprotein complex found both in the nucleus and in the cytoplasm (Mitchell, Nat Struct Biol, 2000). The nuclear exosome is involved in maturation processes of for example snoRNAs and pre-rRNAs (Allmang, Embo J, 1999) and in degradation of incompletely processed mRNA transcripts. The cytoplasmic exosome's role is to survey and degrade mRNA transcripts. Exosome-mediated degradation can be controlled both through use of sequence-specific and general RNA-binding proteins (e.g. EJC) or through sequence elements typically located in the 3' untranslated region. As an example of sequence element-mediated decay, transcripts with 3' AU-rich elements (AREs) are often targeted for miRNA-dependent rapid degradation (Shaw, Cell, 1986; Yang, Genome Res, 2003). Transcripts that encode proteins which are required transiently, e.g. cytokines, growth factors and proto-oncogenes, often contain AREs (Houseley, Nat Rev Mol Cell Biol, 2006).

Different surveillance pathways exist to degrade partially or incorrectly processed transcripts, and transcripts with premature stop codons (PTC) that are often caused by alternative splicing; in fact, up to one third of all alternatively spliced transcripts may be targeted by rapid decay caused by introduction of premature termination codons (Hillman, Genome Biol, 2004). Surveillance pathways that are widely used include 1) nuclear, exosome-mediated surveillance to degrade partially or non-processed transcripts (Hilleren, Mol Cell, 2003; Moore, Cell, 2002), 2) cytoplasmic, exosome-mediated non-stop decay to identify transcripts that lack stop codons (Frischmeyer, Science, 2002; van Hoof, Science, 2002), and 3) cytoplasmic nonsense-mediated mRNA decay (NMD) that degrades transcripts with PTCs (He, Mol Cell, 2003; Maquat, Nat Rev Mol Cell Biol, 2004; Mendell, Nat Genet, 2004). The NMD pathway identifies transcripts with premature stops during a pioneer round of translation by searching for these codons upstream of exon-junction complexes (Ishigaki, Cell, 2001). NMD is initiated if a PTC is found >55 nt upstream of an exon-junction complex.

2.3.10 Balancing between degradation and translation

As already noted above, the life of a eukaryotic mRNA transcript is essentially a balance between destruction by various pathways in both the nucleus and in the cytoplasm, and translation (Figure 4). The initial default path for pre-mRNA transcripts is immediate degradation in the nucleus. However, various processing steps, such as splicing, increase the stability of the transcript, presumably through coating of the transcript with RNA-binding proteins. A mature mRNA transcript, on the other hand, is destined for translation, and the default state can be characterised as 'relative stability'. Again, multiple forces act together to shorten the half-lives of the transcripts through various mechanisms, including surveillance pathways, normal turnover, and localisation to defined cytoplasmic loci termed P-bodies. Collectively, these processes create a pool of mRNA transcripts that is balanced and regulated in terms of quantity, cellular localisation and availability for protein translation. Future gene expression studies need to take this into consideration.

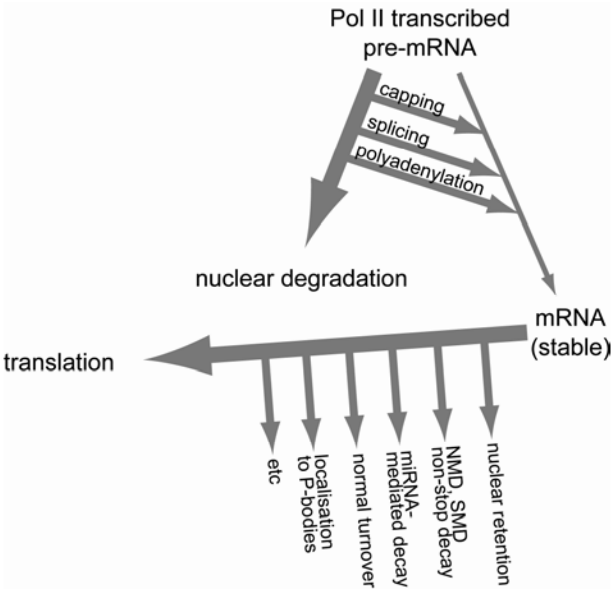

Figure 4. An mRNA transcript balances between degradation and translation. Abbreviations: mRNA, messenger RNA; NMD, non-stop mediated RNA decay; miRNA, micro RNA.

3. Tools for mining the transcriptome

Essentially every cell in an organism is, at any given time point, transcribing thousands of its genes in various quantities. As described in the previous section, the amount of a messenger RNA transcript is tightly regulated and there is an interest, both from basic science and clinical perspectives, in being able to accurately quantify the levels of different transcripts.

In this section both gene-by-gene methods and global methods for quantification of messenger RNA levels are described. The focus will be on the microarray technology, which is described last in the global methods subsection.

3.1 Gene-by-gene methods

Northern blot provides a quantification and size determination of a transcript in a complex mixture by first separating the transcripts by denaturing agarose gel electrophoresis, followed by a transfer to a membrane strip and hybridisation with a labelled probe (Alwine, Proc Natl Acad Sci U S A, 1977). However, the method is sensitive to RNA degradation and lacks a wide dynamic range. Labelling of the probes is commonly achieved using either radioisotopes or biotin. Typically DNA probes up to several hundred bp are used, but recently locked nucleic acid (a nucleic acid analogue) has been demonstrated to achieve a 10-fold improvement in sensitivity (Valoczi, Nucleic Acids Res, 2004).

Quantitative real-time reverse-transcription PCR (qRT-PCR) provides superior sensitivity for analysis of gene expression levels compared to other methods; analysis of even single cells is possible (reviewed in (Wong, Biotechniques, 2005)). First, a complex mixture of total RNA is converted to cDNA using reverse transcriptase with either random or gene-specific priming. Next, a 100-200 bp fragment is amplified and the accumulation of product measured after each cycle using a fluorophore that either specifically targets the amplicon or any double-stranded DNA. During the exponential phase of the amplification each PCR cycle doubles the amount of product and in $\log_2$ scale this corresponds to a linear increase. Extrapolation of the linear increase to the level of background provides an estimate of the initial starting amount of mRNA. Use of qRT-PCR has many advantages, making it the method-of-choice for high-accuracy - but low-throughput - gene expression analysis: 1) it offers a dynamic range of 7-8 log orders of magnitude (Morrison, Biotechniques, 1998), 2) it can achieve single-copy detection (Palmer, J Clin Microbiol, 2003), 3) it can be carried out in one step, 4) it has low coefficients of variation facilitating detection of small differences between samples (Gentle, Biotechniques, 2001), and 5) design of specific amplicons allows for discrimination between similar transcripts, such as gene family members.

3.2 Global methods

Global methods allow for a nearly-complete analysis of the transcriptome and are often associated with high costs and extensive data analysis. However, the methods have become popular; a query with the word ‘microarray’ in the NCBI’s PubMed yields >15,000 hits – an incredible increase since the first publications in early 1990’s. The race towards the $1,000 human genome has also provided the research community with two new ultra-high-throughput DNA sequencers. Roche and 454 Life Sciences market a bead-based pyrosequencing instrument producing tens of millions of bases of sequence every hour. Solexa, through its soon-to-be-released Clonal Single Molecule Array^{TM}, produce up to one billion bases of sequence per run. Various sequencing-based tag-counting methods are briefly outlined below and discussed in more detail elsewhere (Harbers, Nat Methods, 2005).

Expressed sequence tag (EST) sequencing generates random, 200-900 bp single-pass sequences of cDNA clones. The initial purpose of these sequences was to facilitate gene detection (Wilcox, Nucleic Acids Res, 1991), but they have also been used for estimation of gene expression levels. The main drawback is the low throughput (number of counts) caused by the high data generation costs (library generation and sequencing), low-quality of sequences and uncertainty of coverage of the transcript (i.e. whether it is full-length or not).

Serial analysis of gene expression (Velculescu, Science, 1995) was the first approach to provide large-scale absolute estimates of transcript frequencies, and relies on a biotinylated primer, streptavidin-coated beads and type IIs restriction endonucleases (that cleave outside the recognition site) to generate short tags from each transcript. The tags are concatemerised and sequenced using standard sequencing technology to derive a digital representation of the transcript frequencies. Originally SAGE was used to isolate approximately 14-bp 3’ tags, but the method has been later developed to also isolate 5’ tags and longer 26-bp tags.

Cap analysis of gene expression (Kodzius, Nat Methods, 2006) uses 5’ cap-trapping methods to selectively isolate full-length cDNAs and generates 20-bp tags from these. After isolation, the tags are ligated to yield ~700 bp concatemers, cloned into a vector and sequenced. The first-strand synthesis can be primed with random primers, allowing for analysis of polyA-negative transcripts. CAGE has recently been used for large-scale transcription start site mapping (Carninci, Nat Genet, 2006).

Massive parallel signature sequencing (Brenner, Nat Biotechnol, 2000). 3’ sequences of each transcript are isolated using biotinylated primer in the cDNA synthesis and cleavage with DpnII. Next, the 3’ signature sequences are ligated into specifically designed plasmid vectors containing 32-nt oligonucleotide tags (in total 16.8×10^6 different tags), and amplified using PCR. Use of a large number of tags provides a unique tag for each 3’ signature sequence, which is subsequently coupled to a 5-µm microbead. Each bead contains one type of capture tag complementary to one of the 32-nt oligonucleotide tags. Next the captured signature sequences are sequenced on beads to yield 16-20 nt signature tags, which are counted and mapped to the genome.

Gene identification signature (Ng, Nat Methods, 2005) relies on sequencing of concatenated 3’ and 5’ paired-end ditags and is perhaps more suitable for gene discovery projects than for gene expression profiling projects.

The number of tags generated with the different methods varies; using the EST approach counts of a few tens of thousands per library is achievable, while the MPSS generates approximately one million tags. With the remaining techniques counts up to a few hundred thousand are achieved. However, use of the ‘new’ sequencing techniques allow for generation of millions of tags, facilitating detection and reliable analysis of genes expressed at low levels. In fact, use of the 454 sequencing and GIS tag sequencing has been recently reported (Ng, Nucleic Acids Res, 2006).

The advantage of the sequencing-based tag counting methods is the possibility to map the obtained hits to the genome. For the CAGE technology, for example, this provides additional information, such as the identification and quantification of transcription start site usage. Use of these technologies is however restricted mainly to large-scale genome centres due to high sequencing costs.

3.3 Microarray-based methods

Microarrays are used to measure levels of mRNA transcripts, miRNAs, and proteins, but also to analyse characteristics of genomes (e.g. SNPs, gene copy number changes, and larger chromosomal gains and duplications). The fundamental underlying advantage of the technology is that a simultaneous, highly parallelised measurement of thousands of different targets is possible; in many cases allowing for analysis of all known protein-coding transcripts. The massive interest and large scientific expectations have generated an entire microarray industry providing various microarray-related products and services. Although many of the individual academic and industry-generated approaches are rather different, the unifying themes for the different microarray approaches are: 1) use of a solid support (e.g. glass, nylon filter, beads) to which probes are attached, 2) high density of probes facilitating large-scale analysis, 3) application of a complex sample to the array, 4) capture of target by its corresponding probe using base complementarity or antibody binding, and 5) detection system with a wide dynamic range, commonly based on fluorescent dyes. The array-based approaches are also substantially less expensive than alternative methods providing similar throughput (e.g. sequencing-based approaches, see previous section).

A typical gene expression level analysis starts with isolation of total RNA, followed by cDNA synthesis and labelling. Next the purified and labelled cDNA is applied onto a microarray containing thousands of immobilised probes, hybridised, washed and scanned. Arrays are bought from a commercial vendor (e.g. Affymetrix or Agilent Technologies), an academic microarray provider (e.g. KTH Microarray Center, http://www.ktharray.se), or produced in-house. A more detailed description of the different technical aspects of the sample preparation and the various microarray platforms is available in subsequent chapters.

3.3.1 Nomenclature

A defined nomenclature has been established to facilitate discussion of microarrays and to promote exchange of data. The most important distinction is made between the *probe* and the *target*, where probe refers to the DNA immobilised on the solid surface and target to the labelled sample hybridised onto the microarray. Probes are grouped into *blocks*, and probes in a block are generally more similar in terms of morphology and intensity to each other than to other probes on the arrays. The block structure is usually determined by the print-pin that deposited the DNA onto the array. A defined nomenclature for sharing of microarray data has also been generated, and is known as MAGE-ML (see chapter on Data sharing) (Spellman, Genome Biol, 2002).

3.3.2 Platforms

The main differences between the array platforms are the type of probe attached to the surface, the number of target samples (either one or two) that can be hybridised simultaneously on each array, and the principal expression measurement (ratio for two-channel arrays and absolute level estimate for single-channel experiments). The platforms differ also in target labelling and hybridisation, image analysis and initial low-level data analysis aspects. However, at the high-level data analysis phase (where biological inference is sought) the data analyses for the different array platforms converge, and the approaches and the interpretation of results generated are similar.

3.3.2.1 cDNA and other PCR-amplified probe arrays

The relatively low cost of cDNA array production, and the access to thousands of EST clones in the freezers in many laboratories, especially in the large-scale sequencing laboratories, and the commercial distribution of EST clone collections propelled the early development and popularity of the cDNA arrays in the late 1990's (DeRisi, Nat Genet, 1996; Schena, Science, 1995).

Manufacturing. The arrays are generated through PCR-amplification of clone collections using vector-specific primers, or through amplification of specific genomic regions (most commonly gene or promoter regions) using probe-specific primers. The double-stranded DNA amplicons are purified with one of the widely used purification techniques (ethanol precipitation or filter plates). Following a small-volume elution, the purified products are printed using specific instruments developed for microarray production. A print head with up to 48 pins is used to deposit the probes into a grid, generating one block for each print-tip. The total number of probe preparation steps is high, and therefore all steps are carried out in either 96- or 384-well format. To avoid plate handling errors, rigorous quality control steps, including complete, partial or random resequencing and agarose gel electrophoresis analysis of the purified clones is advantageous, but also labour intensive and costly.

Advantages and drawbacks. Advantages of the platform include: 1) low cost of arrays which allows for design of large experiments with extensive replication (note however that the initial probe preparation costs may be substantial), 2) the possibility of using two-colour detection, further facilitating use of complex design, 3) clone collections are widely available from multiple sources, 4) compatibility with most amplification protocols, and 5) established laboratory protocols. The drawbacks (many of which are shared with oligonucleotide and Affymetrix arrays, see below) include: 1) unspecific target-probe interaction due to the length of probes, 2) false negatives due to probes failing during preparation, 3) batch-to-batch variability in array production, 4) incomplete transcriptome coverage, 5) uncertainty over which region or isoform of a transcript is targeted with a given probe (the complete probe sequence is available only in few cases), 6) difficulties in maintaining high-quality probe collections (avoidance of evaporation, well-to-well contamination, plate rotation, etc), and 7) as the probes are double-stranded, measurement of both sense and antisense transcripts is confounded.

3.3.2.2 Oligonucleotide arrays

The concept of long (50 – 90 nt) spotted oligonucleotide arrays has recently been introduced for gene expression profiling (Hughes, Nat Biotechnol, 2001), offering higher specificity than is achievable using the cDNA array approach.

Manufacturing. Using publicly available genome sequences, oligonucleotides are designed *in silico* for each gene, and made as specific as possible, allowing monitoring of members of the same gene family. Furthermore, the melting temperatures of the oligonucleotides are taken into account to achieve uniform hybridisation conditions. The oligonucleotides are bought pre-synthesised, dissolved in appropriate printing buffer and printed using the same approach as the amplified cDNA clones.

Advantages and drawbacks. Use of these pre-synthesised oligonucleotides offers some advantages over the conventional cDNA arrays. 1) Arrays can be generated for any organism given that its genome sequence and gene predictions (or large-scale EST libraries) are available. Not surprisingly, oligonucleotide collections have been recently released for organisms such as grape, peach and tomato (Operon). 2) The probes are targeted to specific regions of genes, avoiding sequences that are shared between multiple genes, which also allows for some (limited) differentiation of splice variants of a gene. 3) Clone handling is reduced, minimising the risk for plate or clone handling errors. 4) Replacement plates are easy to obtain and are easily incorporated into an existing collection. 5) The probes are designed to have the same sense as the mRNA; hence they are complementary to the labelled cDNA generated from the mRNA, and a confounded measurement between sense and antisense strands is avoided. However, this is also a minor drawback, as the commonly used amplification approaches generate the opposite strand, making the use of these oligonucleotide arrays incompatible with the commonly used linear T7 amplification method (see section on Target amplification). In addition to many of the drawbacks listed for cDNA arrays (see points 2, 3, 4 and 6 in cDNA array section), the initial purchase investment for oligonucleotide collections is substantial.

3.3.2.3 In situ synthesised Affymetrix GeneChip arrays

These arrays are hybridised with only one sample and have gained popularity through their wide availability and ease of use.

Probe design. The probes are designed *in silico* and show 3' end bias to avoid non-unique transcript regions. Commonly 11-20 perfect match (PM) probes, together with their mismatch (MM) probes, are used to represent each transcript and collectively these are termed a *probe set*. The MM probes differ from their PM probes by one base in the central positions of the 25-mers; the base change destabilises the probe-to-target binding and is supposed to allow for estimation of non-specific binding. Depending on the data analysis approach, the intensities from the MM probes can be used to correct the signal from the PM probes (the value of this approach is however questionable (Irizarry, Nucleic Acids Res, 2003; Irizarry, Bioinformatics, 2006)).

Manufacturing. The oligonucleotides are synthesised directly on the array using photolithography chemistry (Fodor, Science, 1991; Lockhart, Nat Biotechnol, 1996; Pease, Proc Natl Acad Sci U S A, 1994). The synthesis is carried out nt-by-nt through synthesis in the 3'-5' direction. Briefly, a surface is coated with linkers containing a photosensitive group. A photolithographic mask directs light to pre-defined positions on the array and deprotects these through light-induced cleavage of the photosensitive group. Next, one of the four nucleotides is added, and allowed to couple. Extension with multiple nucleotides is avoided through use of nucleotides that are inhibited from multiple polymerisations by use of a protective photosensitive group. Next, a different mask is used and the process repeated with nucleotides added in predefined order to yield 25-mers.

Advantages and drawbacks. 1) The direct synthesis of probes on arrays avoids problems with incorrect plate handling and problematic spot morphology and ensures that the bath-to-batch variability is minimised. 2) Small feature sizes yield dense arrays. 3) Probes are single-stranded and hence nonconfounded measurements between overlapping transcripts are obtained. The drawbacks are 1) inflexibility in probe content is caused by the high cost of mask manufacture inhibiting frequent probe redesign, and hence several of the arrays are based on obsolete genome assemblies. 2) Sample preparation always includes linear amplification. 3) The arrays are expensive, rendering complex designs too expensive for common use.

3.3.3 Sample preparation

Analysis of complex tissues as such is of little value due to cellular heterogeneity. The bulk brain, for example, is a mixture of hundreds of different cell types, and unless the different cell types are specifically selected prior to mRNA extraction, the obtained gene expression profile will be a weighted average of the total gene expression in all the different cell types, with the most numerous cell type dominating. Hence, the homogeneity (purity) of the sample determines what is measured, and what biological conclusions can be drawn from the data.

Several approaches have been used to obtain homogeneous samples. 1) Experiments can be designed to include early sampling after induction of differentiation or treatment, which allows for monitoring of early events before secondary changes start to accumulate. 2) Selective markers (e.g. green-fluorescent protein expression) can be used in transfections to achieve high transfection rates. 3) Synchronised cell cultures, allowing for analysis of cell cycle phase-specific gene expression patterns can also be used (Spellman, Mol Biol Cell, 1998). To obtain samples with the highest degree of homogeneity, methods such as 4) fluorescence-activated cell sorting (FACS) or 5) laser-capture microdissection (LCM) can be used.

FACS is based on fluorescently-labelled antibodies binding to specific molecules on the surface of the desired cell, but not to other cells in the sample. Two cell types both expressing a marker, but in different quantities, can also be separated using FACS into 'marker-low' and 'marker-high' cell fractions. The cell-antibody conjugates are passed through a detector, one cell at a time, and when the desired cell is detected, an electric field is applied to direct the sample into a separate collection vessel. Improved separation is commonly achieved using multiple fluorophore-conjugated antibodies to identify a combination of different cell-surface molecules. Negative selection (absence of binding of a certain antibody), size selection and cell morphology can also be used as isolation criteria. The purification and isolation of the different stem, progenitor and mature cells of the hematopoietic lineage is a well-known example of successful application of the FACS technology to a complex heterogeneous biological sample (e.g. (Terskikh, Blood, 2003)). The main drawback of FACS technology is often the lack of antibodies for cell-surface proteins, or the lack of markers for a certain cell type.

LCM relies on recognition of a specific cell in a microscopic evaluation of the sample. Briefly, a single-cell or a thin layer of cells is attached to a solid support and the cells of interest identified. Next, a computer-controlled laser is used to excise these cells and isolate them into a collection vessel. The isolation technique relies on knowledge of the tissue structure and identification of specific cells, and is labour intensive, but has the potential to derive extremely pure samples.

3.3.4 Target preparation

Typically 10-20 µg of total RNA or 300-1000 ng of mRNA is required to label the target. This amount of material corresponds to approximately one to two million cells (assuming 10 pg of total RNA per cell), which is commonly obtained by *in vitro* cell culturing studies. Use of methods to obtain homogeneous samples compromises the yield, and hence a signal or target amplification method is required.

3.3.4.1 Signal amplification

These approaches include use of radioactive labelling with extended exposure times, dendrimer labelling, or tyramide signal amplification. Radioactive labelling is difficult to control; it easily gives rise to “bleeding” into adjacent features and is hazardous to work with. Amplification using the dendrimer technology relies on incorporation of a capture sequence into the cDNA and a post-hybridisation labelling where dendrimers that contain several fluorophores are directed using complementary binding towards the incorporated capture sequences in the hybridised target molecules (Stears, Physiol Genomics, 2000). These approaches facilitate analysis of nanogram amounts of target, at best, but are not compatible with amounts obtained from LCM or FACS.

3.3.4.2 Target amplification

To analyse low or sub-nanogram amounts of material, two fundamentally different target amplification methods are used: linear T7-based *in vitro* transcription (IVT) and PCR-based exponential amplification (Figure 5). Comprehensive literature reviews of target amplification approaches are available (Nygaard, Nucleic Acids Res, 2006; Sievertzon, PhD thesis, 2005).

Linear T7-based *in vitro* transcription amplification and modified variants (Baugh, Nucleic Acids Res, 2001; Eberwine, Proc Natl Acad Sci U S A, 1992; Van Gelder, Proc Natl Acad Sci U S A, 1990) are the most widely used amplification methods. IVT amplification typically yields approximately 300 to 500- fold amplification, but under optimised conditions up to 1000-fold amplification is achievable. Up to three rounds of amplification can be carried out for small amounts of starting material. Briefly, first-strand cDNA synthesis is primed with an oligo(dT) primer containing a T7 promoter sequence. Next, double-stranded cDNA is synthesised with a random primer, and is followed by a 3-12 hour isothermal IVT reaction at 37°C, during which the amplified RNA (aRNA) linearly accumulates. This approach was chosen at an early stage by Affymetrix as their target preparation approach – in fact, all samples used on Affymetrix arrays are subjected to IVT amplification.

Drawbacks include: 1) some array probes are incompatible due to target 3' bias introduced by the amplification. Synthesis of cDNA shortens the product towards the 3' end; unamplified material ranges from 200 bases to several kb, while the size range is 250 to 1800 bases and only 200 to 600 bases, respectively, for material amplified one or two rounds. Therefore, probes directed towards central or 5' regions of transcripts often lack corresponding target sequence in the aRNA population. A template-switching approach enriching for full-length transcripts has been developed (Wang, Nat Biotechnol, 2000). 2) Incompatibility with the oligonucleotide arrays occurs due to the amplified and labelled target and the probe having the same sense (strand orientation). To circumvent this, modified approaches incorporating the T7 promoter in the second-strand synthesis have been reported (Che, Lab Invest, 2004; Kaposi-Novak, Biotechniques, 2004; Marko, BMC Genomics, 2005; Rajeevan, Genomics, 2003; Schlingemann, Nucleic Acids Res, 2005). Another alternative is to use labelled aRNA (e.g. aminoallyl, biotin or labelled platinum conjugates) in the hybridisation, but the altered hybridisation conditions due to the reduced specificity of the RNA:DNA binding need to be addressed.

PCR-based amplification methods are more diverse, and are typically based on ligation of linker sequences to both ends of double-stranded cDNA, followed by a certain (often as limited as possible) number of PCR cycles to yield double-stranded DNA. These methods are generally assumed to introduce bias to the data due to transcript-length dependent or base composition (GC content) differences in amplification efficiencies. To circumvent this problem, approaches that restrict the length of the template and make it more uniform have been developed (Brady, Curr Biol, 1995; Brady, Methods Enzymol, 1993; Hertzberg, Plant J, 2001). The advantage of PCR-based methods over linear IVT methods is that much higher amplification is achievable; with PCR-based methods 10^8 to 10^9-fold amplifications can be obtained (Subkhankulova, Genome Biol, 2006). They are also faster, typically requiring only a few hours, and more cost-effective.

The performances of the various amplification methods have been analysed and found to yield good results for both IVT (e.g. (Zhu, Mol Genet Metab, 2006)) and for PCR-based approaches (e.g. (Goff, BMC Genomics, 2004; Iscove, Nat Biotechnol, 2002)). Subkhankulova *et al* compared PCR-based amplification approaches (global and 5' end template switch) with linear T7-based approaches for analysis of single cells and concluded that the PCR-based methods were more reliable than the linear transcription, and that template switching yielded fewer false positives, but also had a considerably lower absolute discovery rate (i.e. more false negatives) and more compressed ratios (Subkhankulova, Genome Biol, 2006). Similar results have been presented by Petalidis *et al* and Laurell *et al* (Laurell, J Biotech, 2006; Petalidis, Nucleic Acids Res, 2003).

3.3.4.3 Target labelling

The labelling approaches for two-channel array platforms differ from the approach chosen by Affymetrix for their one-channel system. Common for both platforms is the use of fluorescence, but the choice of dyes differs. The Affymetrix approach uses phycoerythrin, while the two-channel approaches commonly use cyanine, Cy or Alexa dyes.

For the cDNA and oligonucleotide arrays target labelling and incorporation of the fluorophore can be carried out, either directly or indirectly. In direct labelling the fluorophore is attached to the nitrogenous base of one of the nucleotides. Use of this approach is, however, expensive and often affected by incorporation difficulties; the reverse transcriptases have difficulties extending with the modified nucleotide, and they often yield shorter cDNA populations. Furthermore, the fluorophores also have different incorporation efficiencies that need to be addressed in the design of the experiment. Use of indirect labelling avoids most of these problems; during the cDNA synthesis only one type of modified nucleotide is used, avoiding differences in incorporation efficiencies. The modification is an attachment of an aminoallyl functional group through a linker to the base of one of the nucleotides. After cDNA synthesis a fluorophore with an ester group is chemically coupled to the aminoallyl group, which ensures that both dyes are incorporated at similar rates.

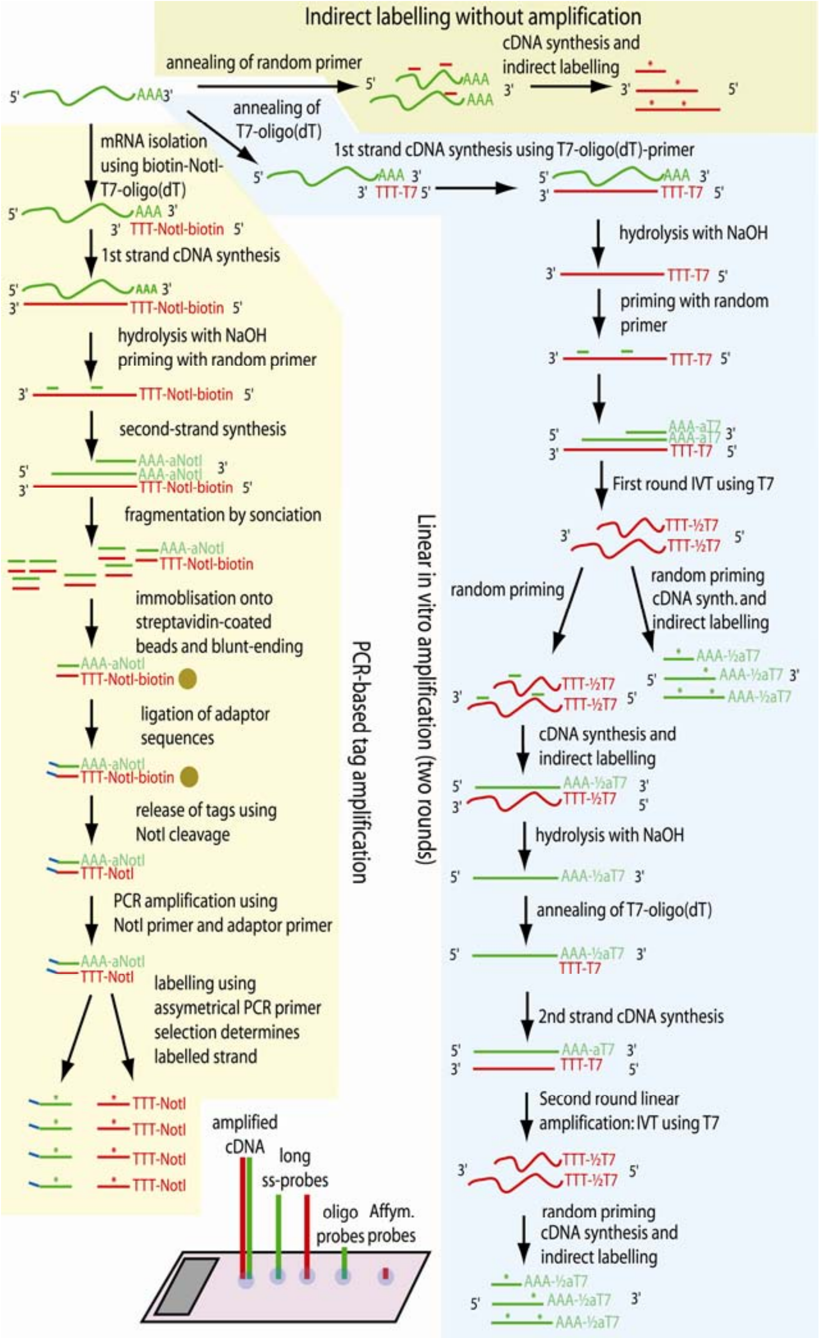

Figure 5. Schematic overview of the different amplification approaches. Colouring is used to distinguish the two strands. Straight lines represent cDNA, while curly lines represent RNA. The brown sphere represents a streptavidin-coated paramagnetic microbead. Labelled nucleic acid is indicated by an asterisk (*). The different types of array probes are schematically shown at the bottom.

For labelling of samples for hybridisation onto the Affymetrix arrays, the first step of the labelling is carried out during the IVT step generating the amplified RNA. Biotin-modified nucleotides are incorporated into the aRNA and dye coupling is carried out after hybridisation using phycoerythrin-streptavidin and biotinylated anti-streptavidin-antibody conjugates.

3.3.5 Hybridisation

Optimisation of hybridisation and wash conditions requires balancing between two opposing forces - hybridisation conditions that are too stringent need to be avoided as these give rise to low signals and increased noise in the downstream ratio estimates, while hybridisations that are too unspecific yield compressed ratios with little differential expression detectable. The main parameters that need to be considered for cDNA and oligonucleotide arrays are summarised below.

1) Probe length. Longer probes generate higher but more unspecific signals due to cross-reactivity with other than intended targets. 2) Hybridisation buffer. Most hybridisations are in either 3xSSC buffer or in 25-50% formamide buffer. The use of formamide increases specificity and allows for use of lower temperatures. Several commercial hybridisation buffers are also available, but these are often proprietary and the exact composition is unknown. 3) Hybridisation temperature. In general, higher temperature gives more specific hybridisation. The choice of hybridisation buffer also affects the hybridisation temperature; hybridisations in salt and formamide buffers are typically carried out at 65°C and 42°C, respectively. 4) Hybridisation duration. Most hybridisations continue for 16-42 hours, and increased hybridisation time (up to 66 hours) has been shown to increase the specificity (Sartor, Biotechniques, 2004). 5) Mixing. Several automated hybridisation stations achieve mixing of the target by pumping small volumes of the hybridisation buffer back and forth over the array surface or using a small-volume air bubble. Use of mixing during hybridisation provides an increased specificity and 2-3 fold increased sensitivity (Adey, Anal Chem, 2002; Schaupp, Biotechniques, 2005). 6) Wash stringency. The hybridisation step is followed by a wash step where unbound target is removed. The washing is typically carried out using multiple solutions with increasing stringency.

3.3.6 Scanning and image analysis

Spotted array hybridisations are typically scanned at 5 or 10-µm resolution one channel at time, generating two 25-100 Mb 16-bit images with a wide intensity range. The diameter of an array feature is typically 100 µm, yielding approximately 80-90 10-µm^2 pixels for each feature. To facilitate the image analysis and visualisation, the two images are overlaid to generate one 24-bit RGB pseudo-colour image with the red, green and yellow spots commonly associated with microarray data. The purpose of the image analysis step is to separate foreground and background intensities, to derive an estimate of the gene expression level for each feature and each channel, and to calculate various intensity and quality control parameters. The underlying concept of the various image analysis software (e.g. GenePix, Spot, Imagene) is essentially the same, with differences mainly on the user interface side.

The image analysis consists of three distinct steps: gridding, segmentation, and intensity extraction, which are reviewed in (Yang, Department of Statistics, University of California at Berkeley Technical Reports, 2000). *Gridding* identifies and assigns central coordinates using layout information provided by the user. The different software carry out this step automatically, commonly with some human intervention required to verify the correctness of the gridding. The *segmentation* step provides a distinction between foreground and background pixels for each feature. A number of methods have been developed to provide as sharp a distinction as possible, even allowing for identification of non-circular features. *Intensity extraction* is used to derive an estimate of expression level for each foreground feature and an estimate of its background. The background intensity was considered to represent the contribution of non-specific hybridisation to the slide surface. To represent the foreground and background intensities both mean and median values are commonly used. However, median values are often preferred as they are insensitive to strong outlier pixels (e.g. scratches, dust particles).

3.3.7 Experimental design

The aim of the experimental design is to make the experiment maximally informative given a certain amount of samples and resources, and to ensure that the questions of interest can be answered. The consequences of incorrect or bad design range from loss of statistical power and an increased number of false negatives to inability to answer the primary scientific question of the experiment. The number of arrays available is in most cases determined by financial resources, and because measurements between slides are more variable than measurements within slides, one of the most important issues is to determine how to allocate the different samples to a given set of arrays (hybridisation scheme). In addition to this, selection of the array platform and probe content of the array, the target preparation approach and what to replicate (biological samples, sampling, RNA extraction, RNA amplification, labelling or hybridisation), need to be considered. These issues should always be addressed, as significant costs are also associated with small-scale experiments.

Replication is carried out to control the three levels of variation in an experiment: biological variation (e.g. differences between animals), technical variation (e.g. differences caused by the RNA amplification), and measurement error (e.g. problems during hybridisation affecting parts of the array). Statistical testing can be carried out on any of these levels, but interpretations of the results differ. Is the purpose to analyse the difference between two mice (inference at the level of technical replicates), or is the purpose to generalise the results and draw conclusions at the level of a population (inference at the level of biological replicates)? It can be safely assumed that the purpose of most, if not all, experiments is to analyse differences at the population level, and hence biological replication is essential.

Hybridisation scheme. Once both the number of arrays and the number and types of samples have been determined, the next question is how to allocate the samples to the arrays. Several approaches have been proposed including direct designs, indirect reference designs, and loop designs (Churchill, Nat Genet, 2002; Glonek, Biostatistics, 2004; Kerr, Biostatistics, 2001; Yang, J Comput Biol, 2005; Yang, Nat Rev Genet, 2002). In considering the allocation of samples to arrays, it is important to identify the main scientific question of the study, and prioritise comparisons that directly attempt to answer this question. It is also important to use balanced designs so that treatments are not confounded with technical issues, such as dye assignments, batch of slides, day of hybridisation, etc. This is most easily achieved by using an equal number of technical replicates for each sample and by assigning an equal number of both dyes to each sample.

The difference between two samples, e.g. treated (T) and untreated (U), can be analysed directly or indirectly (Figure 6). A direct comparison between the two samples is the most straightforward, and requires a dye-swap approach (T labelled in Cy3 on slide 1 and in Cy5 on slide 2, and vice versa for U). Assuming the variance of one hybridisation is 1 (arbitrary unit of variance) in a direct design, use of two hybridisations reduces the variance for U vs. T comparison to ½. The two samples can also be compared using the reference design, where an unrelated sample (R) is used in each hybridisation (i.e. T vs. R and U vs. R). This is however associated with increased variance estimates; the variances are additive and yield a total variance of 2. The difference in efficiency for the two design approaches is therefore four-fold (½ vs. 2), to the advantage of the former. Use of a direct comparison is however seldom possible, as the number of samples in a study is in most cases >2. Despite the larger variance estimate associated with the reference design, it offers several advantages making it the most common type of design: 1) all comparisons are equally important, 2) the study can be extended by adding more samples (assuming more reference sample is available), 3) all samples are handled in the same way, 4) failed hybridisations do not affect the overall analysis and can easily be redone, and 5) analysis and interpretation of the results is straightforward.

An alternative to the reference design is the use of loop designs. Here samples are hybridised to arrays in a 'serial' manner. The main advantage of a loop design over a reference design is the improved variance estimate; in a reference design 50% of the data describes a sample that is not interesting for the primary scientific question and hence the variance components are inflated. In a loop design all data describe the samples of interest and the variance component is estimated more efficiently. No upper limit on number of samples that can be used in loop design exists, but for large loops (>10 samples) the data analysis becomes difficult, and the variance estimation between two samples become less efficient. The loop designs are sensitive to hybridisation failures, but use of combined loop and reference designs efficiently avoids this problem.

Studies have been carried out to calculate the most efficient study designs for different biological questions; Glonek et al proposed a number of 'admissible' designs for factorial and time-course experiments, where they calculated the optimal designs given a limited number of arrays and knowledge of the interactions (questions) most important in the study (Glonek, Biostatistics, 2004). Churchill (Churchill, Nat Genet, 2002) defined four rules for design of microarray experiments: 1) use biological replication, 2) make direct comparisons between samples whose contrasts are of most interest and use short paths to connect any samples that might be contrasted, 3) balance dyes and samples by dye swapping or looping, and 4) keep the goals of the experiment in mind. By following these rules most problems with bad designs are avoided, and efficient use of resources is likely. It is also important to acknowledge that no best design, suitable for all experiments, exists, and that the design issue should be considered individually for each experiment.

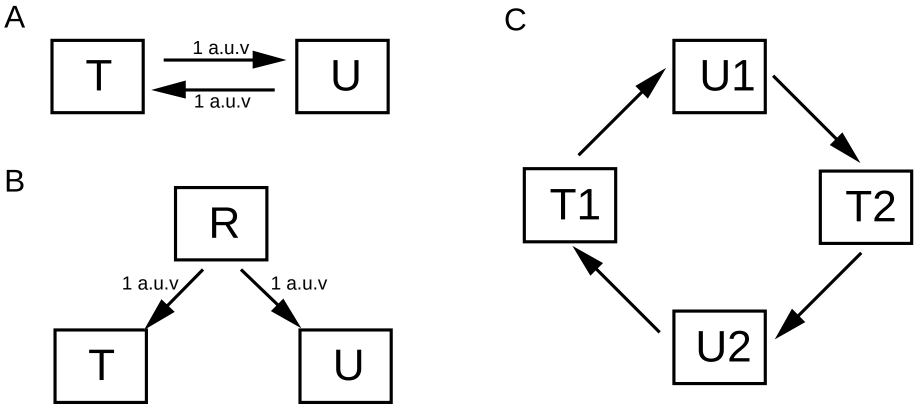

Figure 6. Examples of microarray experiment designs and hybridisation schemes. Each arrow represents one hybridisation, and the head of the arrow denotes Cy5 while the tail denotes Cy3. The samples are untreated (U), treated (T) and reference (R). The numbers above the arrows denote arbitrary units of variance (a.u.v).

3.3.8 Low-level data analysis

Microarray *low-level analysis* corrects for technical artefacts and other bias in the data, while *high-level analysis* seeks true biological differences between the samples. The main low-level analysis steps for two-channel microarray data are summarised below and a more detailed discussion is available in (Bengtsson, PhD thesis, 2004).

The ratio between two channels is the primary expression measurement used for downstream data analysis. The ratios are usually $\log_2$ transformed to yield a symmetrical distribution between up and down-regulated genes. In calculation of ratios, the signal measurement (mean or median of pixels constituting a feature) needs to be selected and an optional background subtraction carried out. Background subtraction is often associated with increased variance, especially for low-intensity features. Microarray data is often visualised in a ratio vs. intensity plot (MA-plot), which facilitates detection of systematic non-biological trends in the data (Figure 7). The next step is optional filtering of non-reliable features from further analysis. Typically these are features with low signal (close to background level, low signal-to-noise ratio, etc), or features that deviate from an expected feature in some aspect (e.g. morphologically through estimation of the feature's circularity). The drawback of filtering is that values are removed (generating missing values), which for some high-level data analysis steps requiring complete data sets introduces the need to estimate the missing values through imputation. Normalisation corrects for differences in the intensities between the channels (Figure 7). Equal amounts of target are used for labelling of both channels, and hence the differences in observed overall intensities are technical in nature and caused by either different amounts of mRNA in the samples, varying efficiencies of cDNA synthesis or dye incorporation, uneven hybridisations, or different scanning parameters or fluorophore properties. Several normalisation methods have been proposed, but the locally-weighted scatter plot smoothing (lowess) approach is most widely used (reviewed in (Smyth, Methods, 2003)). This carries out a local weighted linear regression on the $\log_2$-ratio values as a function of the $\log_2$ intensity, and subtracts the calculated best-fit average $\log_2$ ratio from the experimentally observed ratio for each data point. The approach also de-emphasises the contribution of outliers (differentially expressed genes). It is also important to ensure that the two primary requirements for the data are met before using the lowess normalisation; there should be a roughly equal proportion of up- and downregulated genes, and the majority of the genes should not be differentially expressed.

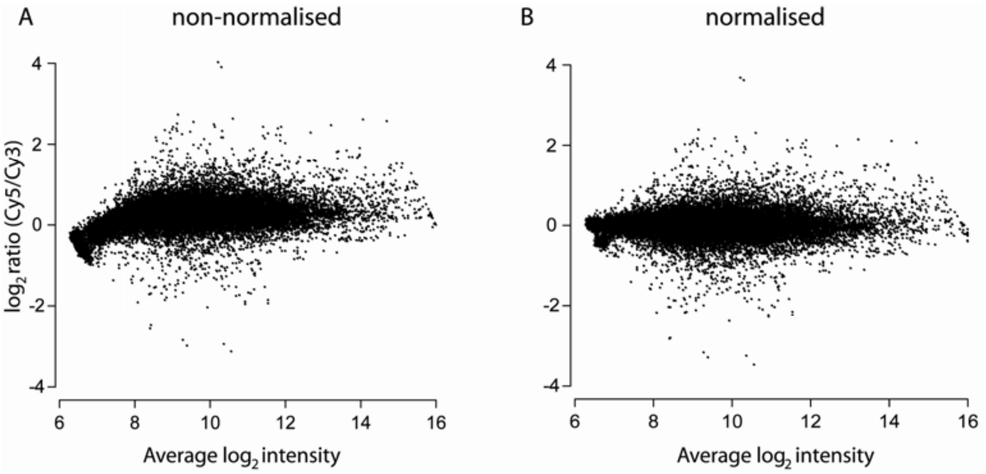

Figure 7. The effect of normalisation visualised using an MA-plot. The non-normalised data (left) shows a dip at low intensities towards negative M-values. After normalisation (right) this trend is removed.

3.3.9 High-level data analysis

Identification of differentially expressed genes between two or more samples is the purpose of the majority of microarray experiments. Initially, the selection of genes was carried out solely on the basis of the fold-change (ratio) value and an empirical cut-off (usually >2-fold change). However, this approach is sensitive to experimental noise – especially at the low-intensity range - and many of the genes selected are false positives. This approach is also unable to identify genes that show small changes, which are consistent over multiple samples. To circumvent these problems, selection based on statistical hypothesis testing is widely used by the microarray community. Many of the tests are based on the *t* test, which compares the means of two groups, considers the variance of the means and rejects the initial H_0 hypothesis of equal expression if the p-value associated with the t-statistics is small enough. A typical microarray experiment consists of thousands of independent tests (one for each probe). For some genes the variance will be very small simply by chance, and, as the variance component is in the denominator of the *t* test formula, a large t-statistics is obtained even if the actual fold-change value is small. To avoid this, moderated *t* tests are used, which add a small constant, estimated in one of several different ways, into the denominator of all the tests, and artificially increase the variance. The available moderated *t* tests differ in the way they determine this constant. The two most widely used moderated *t* tests are SAM (significance analysis of microarray) (Tusher, Proc Natl Acad Sci U S A, 2001) and the empirical Bayes moderated *t* test (Smyth, Stat Appl Genet Mol Biol, 2004).

Use of statistical testing to identify differentially expressed genes with normally employed p-value cut-offs may yield several false positives; with $p<0.01$ we expect one false positive (type I error) for every one hundred tests. Given that arrays contain up to 40,000 probes several hundred false positives are expected. This can be avoided by using family-wise error rate (FWER) controlling methods (e.g. the Bonferroni or Holm (Holm, Scandinavian Journal of Statistics, 1979) methods). However, by controlling the probability of false positives, these methods reduce statistical power and can generate many false negatives (type II error). These corrections also assume independency between the tests, which cannot be expected for gene expression data. Fortunately, instead of controlling the FWER, it is also possible to control the false-discovery rate (FDR; proportion of false positives), which offers a substantial increase in power (Benjamini, J Roy Stat Soc B Met, 1995). The price paid is a substantially increased number of false positives (FWER methods control the probability of obtaining one or more false positives). However, the global microarray-based gene expression measurements are often hypothesis-generating experiments, and hence a small percentage of false positives is acceptable. The FDR-methods are also able to deal with some dependency between the tests (Reiner, Bioinformatics, 2003).

Replication at multiple levels in the design of the experiment also needs to be considered. For example, replicated measurements on each array are more similar to each other than replicated amplifications of a small amount of RNA from two different mice, and hence these replicates need to be addressed separately. A common approach is to use averaging, but this is associated with loss of information. Linear model approaches that efficiently incorporate replication at different levels have been described (Smyth, Bioinformatics, 2005). Analysis of variance-based methods for identification of differentially expressed genes have also been described (Cui, Genome Biol, 2003; Kerr, J Comput Biol, 2000). These are also suitable for identification of confounding effects, e.g. effects such as day, dye, or batch of slides.

Questions commonly associated with statistical testing of differential expression are 1) ‘How many genes are differentially expressed?’ and 2) ‘Where should the p-value cut-off for statistical significance be drawn’? The answers, however, depend on the aim of the experiment and the resources available. Absolute interpretation of the p-values should be avoided, and emphasis should be put on the ranking of the genes. The latter is the most important output of the statistical testing and can be used to select an appropriate number (usually no more than ten to fifty genes) for downstream experimental confirmatory analysis, or for subsequent *in silico* enrichment analysis of theme and pathway terms.

Theme enrichment analysis. Depending on the purpose of the experiment and the magnitude in differential expression between the samples, lists of hundreds or even thousands of differentially expressed genes can be generated. These are often tedious to analyse as the roles of most genes are not directly intuitive. Use of functional themes, e.g. those defined by the Gene Ontology consortium (Ashburner, Nat Genet, 2000), and metabolic and signalling pathways (e.g. KEGG (Kanehisa, Nucleic Acids Res, 2000) and Biocarta) combined with statistical enrichment analysis is a valuable approach for reducing the level of manual analysis of the lists of genes (Hosack, Genome Biol, 2003). Common for all these approaches is the use of Fisher’s exact test to determine, for each theme or pathway, whether it is overrepresented in a given list of genes using a list of background frequencies (most often derived from the probe content of the arrays). Enrichment analyses are also affected by the multiple hypothesis testing problems, requiring use of FDR adjustment. Furthermore, the themes are often related to each other and cannot be considered independent. In fact, it is not uncommon to observe that several different themes are found enriched because they contain a shared core set of differentially expressed genes. Recently an alternative approach for theme and pathway analysis, based on gene set enrichment analysis, has been described (Subramanian, Proc Natl Acad Sci U S A, 2005). This approach circumvents the problem of lack of power with small gene set sizes and the drop of power associated with multiple hypotheses testing by use of a ranking-based scoring approach. Pathways for which many of the genes are present early in the ranked gene list (ranking either through fold-change, statistical significance or equivalent) are considered enriched and are given a high score.

Clustering, classification and dimension reduction tools are also widely used in analysis of microarray data. This is a diverse collection of tools, both in general these methods are often successful in revealing global trends in the data. Several reviews describing these methods are available elsewhere (e.g. (Azuaje, Brief Bioinform, 2003; Quackenbush, Nat Rev Genet, 2001; Slonim, Nat Genet, 2002)).

3.3.10 Microarray data sharing

Microarray experiments are associated with high expenses and the data can often be analysed in different ways answering multiple questions. Comparisons between data sets, and especially meta-analyses, are greatly improved if raw data is publicly available and properly described. To facilitate this, two widely used data storage and exchange repositories are available; ArrayExpress run by EBI (Parkinson, Nucleic Acids Res, 2005; Sarkans, Bioinformatics, 2005) and Gene Expression Omnibus (GEO) run by NCBI (Barrett, Nucleic Acids Res, 2005). Both accept submissions that fulfil the Minimum Information About a Microarray Experiment (MIAME) standards (Brazma, Nat Genet, 2001). The use of these databases has become widespread; ArrayExpress contains more than 1,500 experiments and over 45,000 hybridisations, while GEO is approaching 4,000 experiments and 90,000 hybridisations (July 2006). The purpose of the MIAME standard is to ensure that all essential information regarding the experiment underlying a publication is available, and that the interpretation of the results can be carried out properly. An increasing number of journals are also requiring the data to be publicly available in the repositories in order to publish the results. Also, a mark-up language has been developed to further facilitate data exchange and software integration (Spellman, Genome Biol, 2002).

3.3.11 Analysis software

The amount of data generated by even a simple microarray experiment is large and often contains several hundred megabytes or even gigabytes of data, which places certain requirements on the analysis software. To efficiently meet these requirements, both commercial (e.g. GeneSpring, Kensington Discovery Environment, and Pathway Expert], and academic open-source software solutions have been developed. The commercial software usually offers complete analysis functionalities starting from data preprocessing to identification of differentially expressed genes, enrichment and pathway analysis, clustering and various dimension reduction tools. The open-source software, on the other hand, is often more focused on a particular analysis step and is often written in R (a programming language and environment for statistical computing and graphics). In addition to many packages extending the functionality of R, the Bioconductor project (Gentleman, Genome Biol, 2004) provides a comprehensive collection of tools for all steps of microarray data analysis. TM4, a java-based open-source software suite is also available, providing an easy-to-use graphical interface (Saeed, Biotechniques, 2003). During the recent years open-source software has gained extensively in popularity, mainly due to the rapid availability of new R packages providing tools for analysis steps described in publications, a large user community improving existing functions, the possibility to modify and automate analysis steps, and the fact that the software is available at no cost at all.

4. Stem cells

The existence of stem cells has been known for years, but their occurrence in many tissues has not been reported until recently. For example, it was thought for years that neurons are generated only during the embryonic development, but pioneering experiments already in the 1960's demonstrated neurogenesis in the adult brain (Altman, Science, 1962). In recent years, the driving force for stem cell research has been both the desire for a deeper fundamental understanding of the biology itself, but also the high clinical expectations. Stem cells are expected to provide a source of cells for various cell-based therapies, including tissue repair and therapies for different degenerative diseases. Bone-marrow transplantations have been used for years to treat patients with leukaemia and other blood-related diseases, but the potential avenues for disease treatment are much wider, e.g. treatment of Parkinson's and other neural disorders (Lindvall, Nature, 2006), heart diseases (Srivastava, Nature, 2006) and many others.

The hallmarks of stem cells are that they 1) have the capacity for long-term self-renewal, and 2) have the potential to give rise to multiple differentiated cells. Stem cells divide asymmetrically to generate a copy of themselves (self-renewal) and a progenitor (precursor) cell. The progenitor cells often have reduced potential compared to the stem cells and divide to give rise to more differentiated progeny. Generally, the differentiation capacity is described as a stem cell's potential. *Totipotent* cells (e.g. fertilised oocytes) can give rise to any cell of an organism, *pluripotent* cells (e.g. embryonic stem (ES) cells) can give rise to all three embryonic germ layers, while *multipotent* (e.g. hematopoietic stem cells, HSC) can generate all cells of a certain tissue type.

Stem cells are generally divided into either ES cells or adult tissue stem cells. ES cells are derived from the inner cell mass of pre-implantation blastocysts. Murine ES cells have been in early use since the 1980's (Evans, Nature, 1981; Martin, Proc Natl Acad Sci U S A, 1981), while culturing of human ES cells was described in 1998 (Thomson, Science, 1998). Adult tissue stem cells are multipotent and give rise to progeny only of their restricted lineage. The primary role of tissue stem cells is to maintain the normal tissue homeostasis and carry out repair functions. *In vivo* stem cells reside in particular microenvironments, niches, that provide the stem cells with the cues required for control of self-renewal and differentiation (reviewed in (Moore, Science, 2006)). The niches for some tissue stem cells are characterised in detail (e.g. skin and small intestine). The molecular signalling in the microenvironments is often through short-distance signalling, for example through cell-to-cell contacts (e.g. through ephrin signalling (Holmberg, Cell, 2006)).

Neural stem cells reside in two neurogenic regions in the adult brain; the anterior wall of the lateral ventricle of the forebrain, from where they migrate to the olfactory bulb, and the dentate gyrus of the hippocampus (Gage, Science, 2000). The exact identities of these neural stem cells are still under debate. Cells with stem cell characteristics can be isolated from the adult brain using culturing conditions containing strong mitogens (basic fibroblast growth factor or epidermal growth factor) (Reynolds, Science, 1992). Under these conditions only cells with proliferative capacity form free-floating cell aggregates – neurospheres - that can subsequently be cultured for several passages (which fulfils the requirement for self-renewal) and that can be differentiated into all neural cell types (neurons, astrocytes, oligodendrocytes, i.e. differentiation into multiple cell types). Prolonged growth of neurospheres generates heterogeneous aggregates where only subsets of the cells retain the stem cell characteristics; hence the culturing conditions need to be strictly controlled. The neurosphere *in vitro* neural stem cell assay has recently been reviewed (Reynolds, Nat Methods, 2005).

During the last decade clear similarities between the stem cell and the cancer field have emerged. Both normal somatic stem cells and neoplastic cancer cells have the capacity for an extensive self-renewal. However, only a subset of the cells of a tumour have the capacity to reinitiate formation of seconday tumours, and these cells can therefore be considered as cancer stem cells (Bonnet, Nat Med, 1997; Pardal, Nat Rev Cancer, 2003;

Singh, Cancer Res, 2003). The similarity between cancer and normal stem cells suggests that the molecular mechanisms controlling the self-renewal of both types of stem cell types may be similar. Therefore, understanding the mechanisms controlling cancer stem cell self-renewal may improve our understanding of the self-renewal of normal tissue stem cells, which may in the end bring the clinical use of stem cells closer. The opposite is also true; understanding the self-renewal of normal somatic stem cells may turn out to be beneficial for treatment of cancers. Several key signalling pathways regulating normal stem cell self-renewal are known to also regulate proliferation of cancer cells, further supporting the existence of shared molecular mechanisms. Similarly, well-known tumour-suppressor and proto-oncogenes may play an important role in regulation of the the normal stem cell phenotype. Future research will probably reveal additional shared properties between the normal and cancer stem cells.

Present Investigation

The six papers forming this thesis are summarised in this section (Figure 8). These papers are centred on the themes presented in the Introduction. Common for all of the papers is the analysis of the protein-coding mRNA transcripts, which were reviewed in section 2.3. Another central theme of all but one of the papers (paper I) is the focus on stem cells (introduced in section 4), and especially on identification of mechanisms controlling their proliferation and self-renewal. Detailed understanding of the mechanisms underlying stem cell self-renewal is important and facilitates their future clinical use.

The stem cell transcriptome is analysed using two different global methods (reviewed in section 3.2); a sequencing-based approach (paper II) and a hybridisation-based microarray method (papers III – VI).

Section 5 describes the motivation and objectives for each of the papers, which are subsequently briefly described in sections 6.1 to 6.6.

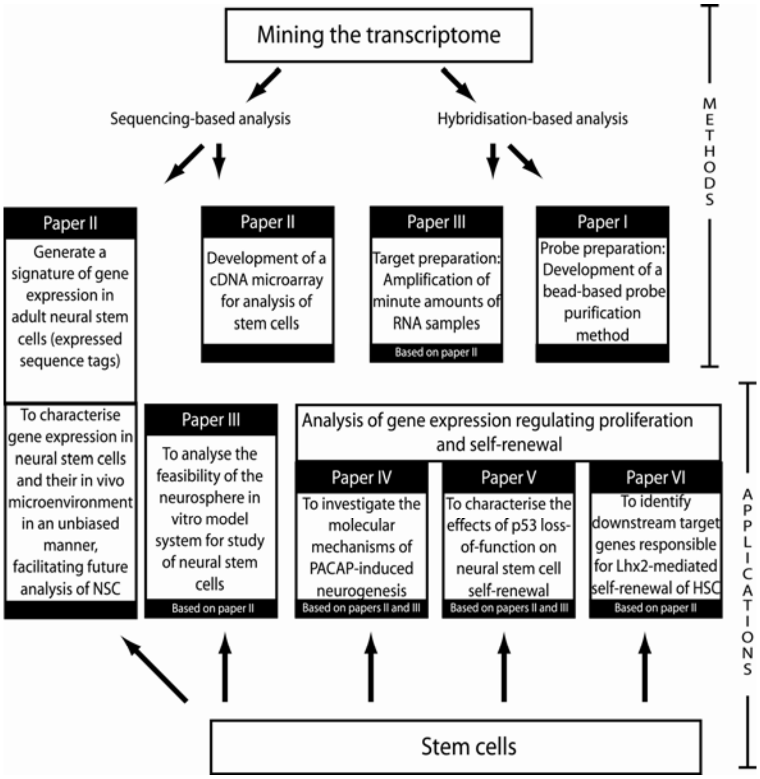
Mining the transcriptome
Sequencing-based analysis
Hybridisation-based analysis
METHODS
Paper II
Generate a signature of gene expression in adult neural stem cells (expressed sequence tags)
Paper II
Development of a cDNA microarray for analysis of stem cells
Paper III
Target preparation: Amplification of minute amounts of RNA samples
Based on paper II
Paper I
Probe preparation: Development of a bead-based probe purification method
To characterise gene expression in neural stem cells and their in vivo microenvironment in an unbiased manner, facilitating future analysis of NSC
Paper III
To analyse the feasibility of the neurosphere in vitro model system for study of neural stem cells
Based on paper II
Analysis of gene expression regulating proliferation and self-renewal
Paper IV
To investigate the molecular mechanisms of PACAP-induced neurogenesis
Based on papers II and III
Paper V
To characterise the effects of p53 loss-of-function on neural stem cell self-renewal
Based on papers II and III
Paper VI
To identify downstream target genes responsible for Lhx2-mediated self-renewal of HSC
Based on paper II
APPLICATIONS
Stem cells

Figure 8. Mining the transcriptome – methods and applications.

5. Objectives

5.1 Paper I

Amplicons need to be purified prior to printing to manufacture cDNA and other arrays based on PCR amplification of clones or genomic DNA. Size-exclusion and silica-based filter plates and ethanol precipitation have been widely used as purification approaches, but use of these is restricted to amplicons longer than approximately 300-500 bp.

In paper I we describe an alternative bead-based probe purification approach suitable for purification of short amplicons. The approach utilises the biotin-streptavidin interaction in a reusable format to specifically purify the biotinylated amplicon. We demonstrate the use of the technology for purification of 21,120 *Arabidopsis thaliana* gene sequence tags and show that the array performs as expected in an auxin time-point study.

5.2 Paper II

Several studies have analysed the neural stem cell transcriptome using microarray-based methods in the search for a stem cell gene expression signature - also known as a stemness signature (e.g. (Fortunel, Science, 2003; Ivanova, Science, 2002; Ramalho-Santos, Science, 2002)). The success of this approach is however dependent on the collection of probes available on the array in use.

The study presented in paper II uses expressed sequence tag (EST) analysis to characterise the gene expression in neural stem cells and their *in vivo* microenvironment (niche) in an unbiased manner. The obtained EST signature of neural stem cells is the largest publicly available, and provides a comprehensive resource for future stem cell research. We show a large overlap with several previously published stemness signatures, and provide several new known and unknown transcripts as candidates for further research. The obtained clones are also used to construct a cDNA microarray, enriched for genes expressed in neural stem cells and suitable for analysis of these and other stem cells.

5.3 Paper III

A neurosphere neural stem cell *in vitro* model can be generated by dissection of the wall of the lateral ventricle in the brain, and culturing the obtained cells under mitogen treatment. The magnitude of the variability introduced by the culturing is unknown, but needs to be investigated to facilitate future use of the neurosphere model system in comparative studies.

In paper III we analyse the variability between different isolation, passaging and culturing replicates, using the stem cell microarrays generated in paper II.

Extended growth of neurospheres generates heterogeneous aggregates containing thousands of cells and with only a minority of the cells retaining the capacity to reinitiate the formation of secondary spheres. Using frequent passaging and strict control of the culturing conditions, more homogeneous cell aggregates can be obtained (with up to 30% reinitiation capacity). However, this compromises the yield, requiring an amplification method for the target preparation.

In paper III we also investigate the performance of a 3’ tag-based amplification method for analysis of neurospheres.

5.4 Paper IV

The adult neural stem cells are a promising source of cells for treatment of various neural degenerative diseases. These cells could either be stimulated to proliferate and differentiate endogenously, or expanded in culture and transplanted back to the damaged site. Pituitary adenylate cyclase-activating polypeptide (PACAP) has previously been shown to endogenously stimulate neural stem cell proliferation, through a hereto unknown mechanism.

To provide an insight into the mechanisms responsible for PACAP-mediated increased neural stem cell proliferation, we used in paper IV the stem cell cDNA array to analyse the transcriptional effect of PACAP stimulation on *in vitro* cultured neural stem cells (neurospheres). The RNA amplification method and neurosphere culturing design described in paper III were used.

5.5 Paper V

The role of p53 (Trp53) in control of cell cycle and tumour progression has been established. Recent research has also shown a connection between stem cells and brain tumour formation. Hence, a role for p53 expression in control of neural stem cell self-renewal can be expected.

To shed light on the molecular mechanisms underlying p53-mediated control of stem cell self-renewal, we analyse in paper V the *in vivo* expression pattern of p53 in the lateral ventricle wall region (containing putative adult neural stem cells), and also analyse the effect of p53 loss-of-function on *in vitro* cultured neural stem cells.

5.6 Paper VI

A small number of hematopoietic stem cells (HSC) maintain the entire hematopoietic system throughout life. These are however difficult to study directly due to their low abundance and hence the molecular details controlling the HSC self-renewal are poorly understood. Expression of the LIM-homeobox gene *Lhx2* in murine hematopoietic or embryonic stem cells allows for generation of HSC-like cell lines, which both on the molecular and functional level closely resemble HSC.

In order to elucidate the molecular mechanisms responsible for Lhx2-mediated self-renewal of HSC we created cell lines with Lhx2 expression controlled by a tetracycline-responsive element, and analysed in paper VI the transcriptional changes induced by downregulation of Lhx2 expression using a time-point study.

6. Papers

6.1 Bead-based purification of microarray probes

In **paper I** we describe the use of a bead-based microarray probe purification method that is especially suitable for purification of short amplicons (100-500 bp). This method can be further modified to increase the capacity and facilitate the generation of single-stranded probes longer than 100 nucleotides (Klevebring and Wirta, unpublished results).

Microarray probes are typically purified using ethanol precipitation, size-exclusion filter plates or silica-based filter plates. These are low-cost (ethanol precipitation) and easily automated (filter plates) methods, but are less suitable for purification of short fragments. The purification method described in paper I is based on capture of biotinylated amplicons using paramagnetic, streptavidin-coated microbeads. The approach takes advantage of a recent finding showing that the biotin-streptavidin bond can be broken in a fully reversible fashion, without denaturation of the protein (Holmberg, Electrophoresis, 2005). Briefly, a PCR reaction is carried out with one of the primers biotinylated. Next, the biotinylated product is bound to the microbeads in high-salt conditions and unbound product removed through washing. Elution is achieved by breaking the streptavidin-biotin bond using deionised water; the immobilised products kept in suspension are heated in deionised water to 80°C (1°C/2 s) for 1 second and cooled to room temperature (1°C/2 s). Efficient elution is achieved through a combination of elevated temperature and appropriate temperature ramping. After purification the beads are regenerated (washed and transferred to appropriate storage buffer), which is a key feature of the described purification approach. The ability to use the beads many times dramatically reduces the cost per probe. The entire purification protocol is also suitable for automation on instruments with the capacity to carry out magnetic separations and that are equipped with a peltier thermal element.

We investigate parameters such as bead capacity, binding time, bead regeneration and multiple uses, and demonstrate the use of the technology for purification of 21,120 *Arabidopsis thaliana* gene sequence tags (GST). The GST are 150 to 500 bp (median 220 bp) gene fragments designed to be specific (which is important for organisms such as *A.thaliana* that have undergone genome duplications) and amplified using a two-step approach (first round using gene-specific primers and the second using primers targeted towards 5' capture sequences introduced in the gene-specific primers) (Hilson, Genome Res, 2004). We also demonstrate the use of the GST arrays in a proof-of-principle study where we analyse changes in the transcriptome in a time-point study where *A.thaliana* seedlings are treated with the well-known plant hormone indole-3-acetic acid (auxin). Out of the identified 120 upregulated genes 17 are known auxin target genes, providing validation for the results.

6.2 Neural stem cell expressed sequence tag analysis

In **paper II** we analysed the transcriptomes of neural stem cells and their *in vivo* microenvironment using expressed sequence tag (EST) analysis. Use of EST counts to estimate gene expression levels is approximate and provides a reliable expression estimate for moderately to highly expressed genes. However, on the global level the unsubtracted EST profile reflects the transcriptional activity of the sample and provides an unbiased detection of gene expression.

The large-scale sequencing was primarily carried out on neurospheres (neural stem cell *in vitro* model, NS) and on the neural stem cell *in vivo* microenvironment (lateral ventricle wall, LVW). We used small, early passage neurospheres with a high proportion of cells retaining the capability to reinitiate neurosphere formation to control for sample heterogeneity, and to avoid the influence of culture adaptations. The neurosphere library was normalised to increase the probability of detecting genes expressed at low levels. For analysis of the microenvironment we chose an inclusive strategy, i.e. the library was created from unfractionated tissue, which ensures that most cell types in this region, including the neural stem cells and the progenitor cells, are included. We also sequenced a hematopoietic stem cell line (BM-HPC library) to characterise the gene expression in a well-characterised stem cell-like population.

In total, 50,792 high-quality sequences were generated; 25,501 from the NS library, 14,884 from the LVW library and 10,407 from the BM-HPC library. The sequencing of the neurosphere library in this study is the largest single-library EST effort in the public databases for characterisation of stem cells. Analysis of normalised tag counts shows that the BM-HPC library had the largest within-library redundancy, in the LVW library more genes but with less redundancy were detected, and the neurosphere library had the largest number of detected genes, as would be expected for a normalised library. We carried out several between-library comparisons and identified 1,065 transcripts expressed in all three libraries. In a comparison with a previously published differentiated neurosphere library (Sharov, PLoS Biol, 2003), 639 transcripts were found expressed only in the three stem cell libraries (NS, LVW, BM-HPC). These transcripts, representing the undifferentiated phenotype included several genes involved in 'cell proliferation' and 'cell death'. We also compared the gene expression in the NS library to previously published 'stemness'" signatures. The majority (78% and 62% for (Fortunel, Science, 2003; Sharov, PLoS Biol, 2003), respectively) of the transcripts included in these signatures were found expressed in the NS library. Additional analyses carried out in the study include identification of overrepresented transcripts in the different libraries and analysis of rare transcripts, as well as comparisons with public large-scale *in situ* hybridisation efforts and with immunohistochemically stained tissue microarrays in the Human protein atlas (www.proteinatlas.org).

We have also used the clones derived within the framework of this study to generate a cDNA microarray suitable for neural stem cell-related studies. The current version also contains clones derived from an EST analysis of ES cells (Wirta, unpublished results).

6.3 Feasibility of a PCR-based tag-amplification method for analysis of neural stem cells

In **paper III** we evaluated the feasibility of a PCR-based 3' tag amplification method (Figure 5) for analysis of *in vitro* cultured neural stem cells (neurospheres), and investigated culturing-induced changes in the neural stem cell transcriptomes.

The employed amplification approach is based on isolation of 3' tags of mRNA transcripts through use of a biotinylated primer during first-strand cDNA synthesis and streptavidin-coated paramagnetic microbeads. Briefly, following second-strand synthesis, the obtained cDNA population is randomly fragmented using sonication. 3' fragments are selectively isolated using the biotin molecule incorporated during the cDNA synthesis and streptavidin-coated beads. Next, a common linker is ligated to the 5' end of the bound tags, followed by enzymatic tag release from the beads. Amplification is achieved using PCR with primers complementary to the 5' linker and to a capture sequence included in the primer used for the first-strand synthesis. The presented amplification approach generates a uniform fragment pool ranging from 100 bp up to approximately 600 bp, circumventing size-dependent bias commonly associated with PCR-based transcriptome amplification.

Culturing of microdissected brain regions, e.g. lateral ventricle wall (LVW) region or dentate gyrus of the hippocampus, is a commonly used *in vitro* enrichment approach for selection of cells with neural stem cell properties. The obtained cell aggregates – the neurospheres – are complex structures consisting of many different types of cells. Extended culturing and passaging of neurospheres has been associated with gain of altered properties (Morshead, Nat Med, 2002). Use of controlled culturing and frequent passaging to keep the sizes of the cell aggregates small facilitates generation of neurospheres with high (up to 30%) capacity for secondary neurosphere initiation.

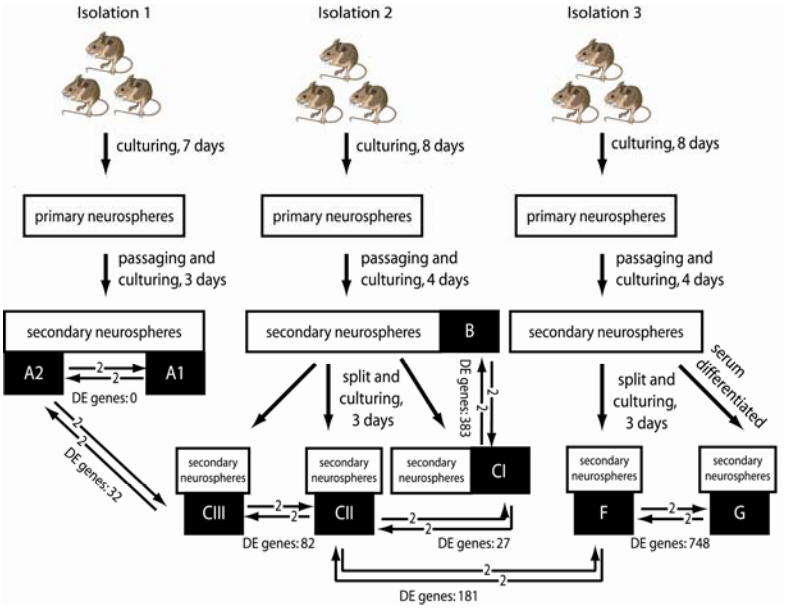

Figure 9. Design of the experiment. White and black squares are samples and amplifications, respectively. The number of differentially expressed genes is shown. The number in each hybridisation arrow indicates replicated hybridisations.

To facilitate the use of neurospheres for future studies, we investigated culturing-induced changes at the transcriptional level by analysis of neurospheres from three different LVW isolations, different passages and different splitted aliquots (Figure 9).

The results show reliable performance of the described amplification protocol; no genes were found differentially expressed in the technical replicates (A1 vs. A2), indicating that methodological noise can be considered minor. Analysis of the isolation, passaging and culturing replicates showed some variability between parallel neurosphere culturing replicates, and a higher number of differentially expressed genes between neurospheres from different passages and different isolations. Many of the genes identified as differentially expressed between the passaging and isolation replicates have large fold-change values, while the differences between culturing replicates are in general of smaller magnitude. Also, the genes found differentially expressed in the two culturing replicate comparisons showed only minor overlap, indicating that these are random, non-systematic changes.

In conclusion, the data demonstrate that a 3-4 day culturing, *per se,* is sufficient to induce changes, but careful experimental planning and use of neurospheres from the same isolation and same passage allows for use of neurospheres in comparative studies. Use of biological replicates is advantageous as the culturing-induced changes are small and non-systematic, and can hence be easily averaged out through use of biological replicates.

6.4 The effect of PACAP on neural stem cell proliferation

Pituitary adenylate cyclase-activating polypeptide (PACAP) binds to the G-protein coupled PACAP receptor 1, and through intracellular signalling cascades induces neurogenesis *in vivo*, and proliferation of neural stem cells both *in vivo* and *in vitro* (Mercer, J Neurosci Res, 2004). To shed light on the molecular mechanisms of the PACAP-induced proliferation, in **paper IV** we carried out a PACAP stimulation study of neurospheres.

The experiment was carried out using microarrays created from the three EST libraries analysed in paper II. These arrays are enriched for genes expressed in neurospheres and hence offer a relevant collection of probes for study of neurosphere proliferation. To facilitate identification of PACAP-specific effects, the design of the study included technical controls to account for variability caused by target amplification and hybridisation, and two control treatments inducing either proliferation or differentiation. The design of the experiment was further based on the findings of paper III; all secondary neurospheres used originated from one isolation of primary neurospheres, and all treatments were tested using samples cultured in parallel and from the same passage. Each treatment was replicated and the subsequent target amplification carried out using the 3’ end tag amplification method, also evaluated in paper III. The treatments were allocated into two treatment replicate groups to allow an analysis with replicates included at several different levels (replicated features on the array, replicated dye-swap hybridisations, replicated treatments).

In line with the findings of paper III, we again found essentially no significant variability between technical amplification replicates. For each treatment, the replicates had high overall M-value ($\log_2$ of fold-change) correlations (0.85 – 0.88) and 60-70% of the genes identified as differentially expressed were shared. Analysis of the M-values of the non-shared differentially expressed genes indicated lack of statistical power and suggests that the true overlap is probably even higher (see paper for details). In a combined analysis of the differentially expressed genes from the PACAP treatment, the proliferation control and the differentiation control treatment, a high proportion of shared transcripts was observed. This indicates a surprisingly similar effect on gene expression levels. A detailed analysis of the non-overlapping genes suggests that they also have the similar M-value trends, and are close to reaching statistical significance (i.e. they are probably false negatives).

Collectively these analyses demonstrate that the three treatments (PACAP, differentiation control and proliferation control) induce similar gene expression changes. A likely explanation is that the removal of the growth factor (EGF) from the neurosphere culture medium, coinciding with the treatment initiation, masks the specific gene expression changes caused by the different stimuli. In retrospect, this is not surprising given that EGF is a strong mitogen. Future studies need to take these results into account when designing neurosphere treatment experiments, possibly necessitating the development of mitogen-free neurosphere culturing methods.

6.5 The role of p53 in control of self-renewal of adult neural stem cells

Several proto-oncogenes and tumour suppressors are known to control self-renewal of normal tissue stem cells, indicating that the molecular regulation of tissue and cancer stem cells is similar, and that tumour formation can be viewed as an excessive stem cell expansion. p53 (official gene symbol Trp53) is the key player in tumour development, but its role in normal tissue stem cells has not been addressed previously. In **paper V** we investigate the effects of p53 loss-of-function and how it affects adult neural stem cells, and provide evidence for its role in negative regulation of neural stem cell self-renewal.

Using immunohistochemistry we observed substantially higher expression of p53 in the neural stem cell lineage (especially in stem cells and progenitors) than in other cells in the adult brain. Labelling studies using a nucleotide analogue identified an increase of proliferating cells in the lateral ventricle wall of *p53*$^{-/-}$ mice compared to their wild-type littermates. Also, the number and proportion of neurosphere generating cells was increased in the LVW of *p53*$^{-/-}$ mice, as was the size of the generated neurospheres. Next, characterisation of small neurospheres (4-20 cells; to avoid generation of secondary effects caused by neurosphere heterogeneity) revealed that *p53*$^{-/-}$ neurospheres had 1) an increased proliferation activity, 2) a decreased apoptosis activity, and 3) that both wild-type and *p53*$^{-/-}$ neurospheres have the potential to differentiate into all neural cell types. Collectively, these results indicate that the increased self-renewal in *p53*$^{-/-}$ is through a combination of increased proliferation and survival (i.e. decreased apoptosis).

To identify the molecular program leading to increased self-renewal, we compared *p53*$^{-/-}$ and wild-type neurospheres using the 'stem cell' microarray generated using clones from paper II. In total, we identified 325 genes that were differentially expressed. Functional classification using Gene Ontology revealed that many of these genes are implicated in control of cell proliferation. *p21* (official gene symbol *Cdkn1a*), a direct target gene of p53, was the most strikingly dysregulated gene, which was verified using qRT-PCR. Western blot analysis furthermore showed a reduction also at the protein level. p21 is a well-known negative regulator of the cell cycle through its inhibition of G1-to-S and G2-to-mitosis transitions. p21 is also known to regulate self-renewal in hematopoietic stem cells (Cheng, Nat Med, 2000), and *p21*$^{-/-}$ has recently been associated with an increased proliferation in the LVW and in neurospheres (Kippin, Genes Dev, 2005).

Taken together, these data indicate that in the absence of p53, several cell cycle regulators are dysregulated, including p21. The data implicate p53 as a suppressor of tissue stem cell self-renewal.

6.6 Lhx2-mediated self-renewal of hematopoietic stem cells

In **paper VI** we investigate the downstream effector genes of Lhx2, which promote stem cell self-renewal.

Throughout the life span of an organism, hematopoietic stem cells (HSC) divide asymmetrically to give rise to several lineages, effectively maintaining the entire hematopoietic system. During embryogenesis, expansion of the hematopoietic system takes place in the liver, indicating that the fetal liver microenvironment may promote HSC expansion and self-renewal. Development of the liver and expansion of the hematopoietic system are also temporally connected, which suggests that the mechanisms may be overlapping. It has further been shown that the LIM-homeodomain transcription factors, especially Lhx2, are important for proper liver development; *Lhx2*$^{-/-}$ embryos develop a small and disorganised liver and lethal anaemia. In **paper VI** we generate HSC-like cells by expression of the *Lhx2* gene in embryonic stem cells. Similar cell-lines generated previously from bone-marrow cells can long-term engraft stem cell-deficient lethally-irradiated mice, indicating that they have stem cell capabilities (long-term self-renewal and differentiation to multiple cell types). The expression of Lhx2 can be turned off, facilitating the identification Lhx2 downstream target genes controlling the self-renewal of these cells.

Using a time-point study we identified 267 genes (141 downregulated and 126 upregulated) that were differentially expressed at 36, 72 and 96 hours after downregulation of Lhx2 expression. These are genes that are putatively involved in HSC self-renewal, differentiation and organ development. The array data confirmed the downregulation of the *Lhx2*, as it was the gene with the largest decrease in expression. Functional analysis using Gene Ontology terms revealed significant enrichment of themes like ‘regulation of signal transduction’, ‘organogenesis’ and ‘cell death’ among the downregulated genes. To verify the array data biologically interesting genes from both the upregulated (n=10) and downregulated (n=10) genes were chosen and analysed using quantitative real-time PCR. The data for all genes included in the verification was in agreement with the microarray data. We next selected 13 genes for *in situ* hybridisations on tissues that express Lhx2 during embryogenesis (olfactory epithelium, hair follicles, cerebral cortex of the forebrain and liver lobes). For six of the genes (*Nuak1*, *Tmem2*, *Etv5*, *Enc1*, *Csrp2* and *Tgfb1il*) we observed an overlapping expression pattern (with Lhx2) in at least two of the tissues. These results suggest that the mechanism whereby Lhx2 immortalises HSC may partly overlap with the function of Lhx2 in the development of a variety of organs.

Interestingly, Rhee *et al* have recently reported that Lhx2 maintains the stem cell character of hair follicles (Rhee, Science, 2006). Another recent study revealed that Etv5 is required for transcriptional control of the spermatogonial stem cell niche (Chen, Nature, 2005). Taken together, these two findings provide additional validation of the approach selected in our study and the obtained results, and suggest that additional detailed studies of genes identified in our study may turn out to be fruitful.

Future perspectives

The magnitude and complexity of the regulatory role played by RNA has so far remained to a large extent hidden, and only recently have we started to understand the scope of it. Future studies will undoubtedly shed further light on various, perhaps even hitherto unknown, regulatory functions exerted by the RNA.

The transcriptome contains a diverse collection of various mRNA transcripts, of which many have the potential to encode a protein. An extensive network of regulatory processes ensures that only a subset of the transcribed sequences end up being available for translation by the ribosome. The cells invest a large amount of energy into regulation of gene expression; large-scale synthesis of transcripts is balanced by a massive and rapid degradation, often already in the nucleus. This allows cells to rapidly adjust to a changing cellular environment.

The analysis methods used today for transcriptional profiling do not yet fully take the complexity of RNA into account, suggesting that future technological improvements may yield major benefits. Many of the currently available methods are in fact designed only for analysis of protein-coding RNA molecules. The extent of transcription from both sense and anti-sense strands described during the last few years requires development of microarray platforms for simultaneous, non-confounded detection of both strands separately. Furthermore, the extent of alternative splicing and the exon-initiated transcription (Carninci, Nat Genet, 2006) necessitates development of gene expression tools capable of separating the different isoforms and the transcripts with coding or non-coding potential. Our understanding of gene regulation will also be greatly improved if methods are developed for distinguishing RNA accessible for transcription from the non-accessible RNA. Further, by measuring ribosome-bound RNA levels using tools already available today, a better correlation with protein levels may be obtained.

Improvements in sample preparation approaches will also be useful. One improvement foreseeable in the near future is the identification of new cell-specific surface markers and production of antibodies targeting these, which will greatly facilitate fluorescence-activated cell sorting and enrichment of pure cell populations. Lack of markers has been an obstacle for the stem cell field, and a panel of markers distinguishing stem cells *in vivo* would be of great benefit.

The high cost associated with the genome-wide gene expression analysis methods means that generated data should be used as efficiently as possible. Several efforts to facilitate data exchange have been carried out (Brazma, Nat Genet, 2001; Spellman, Genome Biol, 2002) and these standards have been widely adopted by most journals. In the current era of ‘systems biology’ several research fields are fusing, necessitating large-scale data exchange, sometimes between different platforms (perhaps even measuring different macromolecules). The format of this exchange needs to be resolved, preferentially sooner than later, in order to facilitate integrative data analysis and to provide guidelines for design of appropriate databases and data exchange tools.

The microarray technology has been available for gene expression analysis for approximately ten years. During this era several improvements and discoveries have been made and several supporting tools have been developed: 1) the number of probes has increased from a few thousand to ~50,000 for in-house produced arrays and to ~6.5M for Affymetrix arrays, 2) a new generation of open-source data analysis tools has been specifically developed for microarray data analysis (e.g. Bioconductor and TM4), 3) genome sequences for many organisms are available facilitating *in silico*-based probe design, 4) computer capacity has increased dramatically, allowing more complex analyses, and 5) a new view of gene expression that clearly deviates from the “one gene – one transcript – one protein”-dogma has emerged. What impact the more recent technological developments (e.g. the new generation of DNA sequencers based on sequencing-by-synthesis chemistry) will have on gene expression analysis remains to be seen, but it will undoubtedly leave at least a few marks.

Abbreviations

A.thaliana	*Arabidopsis thaliana*
ARE	AU-rich element
aRNA	amplified RNA
bp	base pairs
C.elegans	*Caenorhabditis elegans*
CAGE	cap analysis of gene expression
CTD	carboxy tail domain
DMSO	dimethyl sulphoxide
DNA	deoxyribonucleic acid
dsDNA	double-stranded DNA
EJC	exon junction complex
ESE	exonic splicing enhancer
EST	expressed sequence tag
FDR	false-discovery rate
FWER	family-wise error rate
GIS	gene identification signature
GST	gene sequence tag
H.sapiens	*Homo sapiens*
ISS	intronic splicing silencer
kb	kilo bases
kbp	kilo base pairs
LVW	lateral ventricle wall
Mbp	million base pairs
MDa	million Dalton
miRNA	microRNA
MM	mismatch probe
MPSS	massive parallel signature sequencing
mRNA	messenger RNA
mRNP	messenger ribonucleoprotein complex
NMD	nonsense-mediated decay
NPC	nuclear pore complex
NS	neurosphere cDNA library
NSC	neural stem cell
nt	nucleotides
NTP	ribonucleotide triphosphate
PABP	polyA-binding protein
PACAP	pituitary adenylate cyclase-activating polypeptide
PCR	polymerase chain reaction
PM	perfect match probe
Pol I, II, III	RNA polymerase I, II, III
PTC	premature termination codon
qRT-PCR	Quantitive real-time reverse-transcription PCR
RNA	ribonucleic acid
rRNA	ribosomal RNA
SAGE	serial analysis of gene expression
SDS	sodium dodecyl sulphate
SMD	Staufen-1 mediated RNA decay
snoRNA	small nucleolar RNA
snRNA	small nuclear
snRNP	small nuclear RNA protein complex
SSC	sodium chloride / sodium citrate buffer
tRNA	transfer RNA
TSS	transcription start site
UTR	untranslated region

Acknowledgements

Joakim, min handledare, för ett starkt kontinuerligt stöd under fem år. Det är sällan du har sagt nej till mina önskemål och du har hela tiden sett till att jag har tillräckligt många projekt att jobba med. Tack också för att du höll en plats åt mig i din grupp under mitt lumpenår. Ser framemot många nya spännande samarbetsprojekt i framtiden.

Mathias och alla andra som hjälpt till att skapa en mycket trevlig atmosfär på vår avdelning. Det är nog få unnat att kunna säga att det är roligt att gå till jobbet (nästan) varje dag. Jag kan.

Arraygruppen, alla nuvarande och före detta medlemmar. Ett extra tack till Annelie som alltid prompt leverar nya slides och till alla tekniker (speciellt Anna) som hjälpt till under åren.

Jonas och Dinos, det har varit ett stort nöje att få jobba med er. Det blev en hel hög med pek till slut. Stort tack även till alla andra samarbetspartners både på KTH, KI, Umeå universitet och Neuronova.

Stort tack också till alla er som under årens lopp inspirerat mig till en forskarkarriär: Terho Lehtimäki, som under tre somrar i slutet på 90-talet visade vad effektivitet och positiv attityd till forskning är; Stefan Nordlund, som hjälpte till med flera saker under mina år på SU och som lärde mig vad *Rhodospirillum rubrum* är; Edvard Smith, för att jag fick göra ett forskarskoleprojekt i din grupp på Novum och för att du erbjöd mig en doktorandplats.

Ett stort tack till rumskompisarna Malin, Jorge, Erik, Bill och Bull som bidragit till att mina fem år på KTH gått så fort och att det har varit roligt.

Forskning är inte gratis och utan ekonomiskt stöd från Knut och Alice Wallenbergs stiftelse och Stiftelsen för Strategisk Forskning hade det inte gått.

Alla som hjälpt till med avhandlingen. Carolyn, för att du hjälpt till med språket. Jag är otroligt tacksam för all den tid du lagt ner, inte bara på avhandlingen, men även på flera av peken. Albin, Joakim, Peter, Johan och Sara för att ni gett en massa värdefulla kommentarer. Texten blev kanske lite väl lång, men ni tog er igenom den ändå. Tack till alla ni som sett till att det har blivit även en massa annat än bara skrivande denna sommar: Gösta och Barbro Bruce för att ni tog hand om mig en vecka i Blekinge, Kaarle, Mari och Sara för att Greklandsveckan blev en lyckad kombination av semester och skrivande, och Carsten och Debbie för att ni såg till att det blev en liten rundtur ner till Tyskland.

Bubbarna Marcus, Johan och Daniel för arrayandet, analyserandet och lunchandet, det hade inte varit lika roligt utan er.

Alla mina andra kompisar från världen utanför KTH, både i Finland och Sverige. Hoppas vi syns på festen, annars får vi ta en öl efteråt.

Mina föräldrar, Jaana och Ole, för att ni alltid hjälper till när det behövs och kommer med kloka råd om strategiska val i livet, och i karriären.

Kaarle, för att du är helt enkelt den bästa lillebror och kompis man kan ha.

Sara, för att du har haft tålamod med mig under författandet och för att du hjälpt till med avhandlingen på ett fantastiskt sätt. Jag älskar dig.

References

Adey NB, Lei M, Howard MT, Jensen JD, Mayo DA, Butel DL, Coffin SC*, et al.* Gains in sensitivity with a device that mixes microarray hybridization solution in a 25-microm-thick chamber. *Anal Chem* 2002;74(24):6413-7.

Allmang C, Kufel J, Chanfreau G, Mitchell P, Petfalski E and Tollervey D. Functions of the exosome in rRNA, snoRNA and snRNA synthesis. *Embo J* 1999;18(19):5399-410.

Altman J. Are new neurons formed in the brains of adult mammals? *Science* 1962;135:1127-8.

Alwine JC, Kemp DJ and Stark GR. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci U S A* 1977;74(12):5350-4.

Arava Y, Boas FE, Brown PO and Herschlag D. Dissecting eukaryotic translation and its control by ribosome density mapping. *Nucleic Acids Res* 2005;33(8):2421-32.

Arava Y, Wang Y, Storey JD, Liu CL, Brown PO and Herschlag D. Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* 2003;100(7):3889-94.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP*, et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25(1):25-9.

Azuaje F. Clustering-based approaches to discovering and visualising microarray data patterns. *Brief Bioinform* 2003;4(1):31-42.

Ban N, Nissen P, Hansen J, Moore PB and Steitz TA. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* 2000;289(5481):905-20.

Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D*, et al.* NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic Acids Res* 2005;33(Database issue):D562-6.

Bass BL. RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem* 2002;71:817-46.

Baugh LR, Hill AA, Brown EL and Hunter CP. Quantitative analysis of mRNA amplification by in vitro transcription. *Nucleic Acids Res* 2001;29(5):E29.

Belotserkovskaya R, Oh S, Bondarenko VA, Orphanides G, Studitsky VM and Reinberg D. FACT facilitates transcription-dependent nucleosome alteration. *Science* 2003;301(5636):1090-3.

Bengtsson H. Low-level analysis of microarray data. PhD thesis, Lund University 2004.

Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met* 1995;57(1):289-300.

Berget SM, Moore C and Sharp PA. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A* 1977;74(8):3171-5.

Blanchette M, Green RE, Brenner SE and Rio DC. Global analysis of positive and negative pre-mRNA splicing regulators in Drosophila. *Genes Dev* 2005;19(11):1306-14.

Bonnet D and Dick JE. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat Med* 1997;3(7):730-7.

Brady G, Billia F, Knox J, Hoang T, Kirsch IR, Voura EB, Hawley RG*, et al.* Analysis of gene expression in a complex differentiation hierarchy by global amplification of cDNA from single cells. *Curr Biol* 1995;5(8):909-22.

Brady G and Iscove NN. Construction of cDNA libraries from single cells. *Methods Enzymol* 1993;225:611-23.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J*, et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001;29(4):365-71.

Brenner S, Jabob F and Meselson M. An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis. *Nature* 1961;190(4776):567-656.

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S*, et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 2000;18(6):630-4.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R*, et al.* The transcriptional landscape of the mammalian genome. *Science* 2005;309(5740):1559-63.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA*, et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 2006;38(6):626-35.

Cavaille J, Buiting K, Kiefmann M, Lalande M, Brannan CI, Horsthemke B, Bachellerie JP*, et al.* Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc Natl Acad Sci U S A* 2000;97(26):14311-6.

Che S and Ginsberg SD. Amplification of RNA transcripts using terminal continuation. *Lab Invest* 2004;84(1):131-7.

Chen C, Ouyang W, Grigura V, Zhou Q, Carnes K, Lim H, Zhao GQ*, et al.* ERM is required for transcriptional control of the spermatogonial stem cell niche. *Nature* 2005;436(7053):1030-4.

Chen CY, Chen TM and Shyu AB. Interplay of two functionally and structurally distinct domains of the c-fos AU-rich element specifies its mRNA-destabilizing function. *Mol Cell Biol* 1994;14(1):416-26.

Chen CY, Gherzi R, Ong SE, Chan EL, Raijmakers R, Pruijn GJ, Stoecklin G*, et al.* AU binding proteins recruit the exosome to degrade ARE-containing mRNAs. *Cell* 2001;107(4):451-64.

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J*, et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 2005;308(5725):1149-54.

Cheng T, Rodrigues N, Dombkowski D, Stier S and Scadden DT. Stem cell repopulation efficiency but not pool size is governed by p27(kip1). *Nat Med* 2000;6(11):1235-40.

Cho EJ, Takagi T, Moore CR and Buratowski S. mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain. *Genes Dev* 1997;11(24):3319-26.

Chow LT, Gelinas RE, Broker TR and Roberts RJ. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* 1977;12(1):1-8.

Churchill GA. Fundamentals of experimental design for cDNA microarrays. *Nat Genet* 2002;32 Suppl:490-5.

Cui X and Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol* 2003;4(4):210.

de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein M, Pelisch F, Cramer P*, et al.* A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell* 2003;12(2):525-32.

DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y*, et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996;14(4):457-60.

Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M*, et al.* Analysis of gene expression in single live neurons. *Proc Natl Acad Sci U S A* 1992;89(7):3010-4.

Eperon LP, Graham IR, Griffiths AD and Eperon IC. Effects of RNA secondary structure on alternative splicing of pre-mRNA: is folding limited to a region behind the transcribing RNA polymerase? *Cell* 1988;54(3):393-401.

Evans MJ and Kaufman MH. Establishment in culture of pluripotential cells from mouse embryos. *Nature* 1981;292(5819):154-6.

Faustino NA and Cooper TA. Pre-mRNA splicing and human disease. *Genes Dev* 2003;17(4):419-37.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE and Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* 1998;391(6669):806-11.

Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT and Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991;251(4995):767-73.

Fortunel NO, Otu HH, Ng HH, Chen J, Mu X, Chevassut T, Li X*, et al.* Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature". *Science* 2003;302(5644):393; author reply

Fox-Walsh KL, Dou Y, Lam BJ, Hung SP, Baldi PF and Hertel KJ. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc Natl Acad Sci U S A* 2005;102(45):16176-81.

Frischmeyer PA, van Hoof A, O'Donnell K, Guerrerio AL, Parker R and Dietz HC. An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. *Science* 2002;295(5563):2258-61.

Gage FH. Mammalian neural stem cells. *Science* 2000;287(5457):1433-8.

Galy V, Gadal O, Fromont-Racine M, Romano A, Jacquier A and Nehrbass U. Nuclear retention of unspliced mRNAs in yeast is mediated by perinuclear Mlp1. *Cell* 2004;116(1):63-73.

Gentle A, Anastasopoulos F and McBrien NA. High-resolution semi-quantitative real-time PCR without the use of a standard curve. *Biotechniques* 2001;31(3):502, 4-6, 8.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B*, et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004;5(10):R80.

Gerber HP, Hagmann M, Seipel K, Georgiev O, West MA, Litingtung Y, Schaffner W*, et al.* RNA polymerase II C-terminal domain required for enhancer-driven transcription. *Nature* 1995;374(6523):660-2.

Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Van Dongen S, Inoue K, Enright AJ*, et al.* Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science* 2006;312(5770):75-9.

Glonek GF and Solomon PJ. Factorial and time course designs for cDNA microarray experiments. *Biostatistics* 2004;5(1):89-111.

Goff LA, Bowers J, Schwalm J, Howerton K, Getts RC and Hart RP. Evaluation of sense-strand mRNA amplification by comparative quantitative PCR. *BMC Genomics* 2004;5(1):76.

Goodrich JA and Kugel JF. Non-coding-RNA regulators of RNA polymerase II transcription. *Nat Rev Mol Cell Biol* 2006;7:612-6.

Gott JM. Expanding genome capacity via RNA editing. *C R Biol* 2003;326(10-11):901-8.

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A and Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 2006;34(Database issue):D140-4.

Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR and Bateman A. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* 2005;33(Database issue):D121-4.

Gutell RR and Woese CR. Higher order structural elements in ribosomal RNAs: pseudo-knots and the use of noncanonical pairs. *Proc Natl Acad Sci U S A* 1990;87(2):663-7.

Hachet O and Ephrussi A. Splicing of oskar RNA in the nucleus is coupled to its cytoplasmic localization. *Nature* 2004;428(6986):959-63.

Harbers M and Carninci P. Tag-based approaches for transcriptome research and genome annotation. *Nat Methods* 2005;2(7):495-502.

He F, Li X, Spatrick P, Casillo R, Dong S and Jacobson A. Genome-wide analysis of mRNAs regulated by the nonsense-mediated and 5' to 3' mRNA decay pathways in yeast. *Mol Cell* 2003;12(6):1439-52.

Herman RC, Williams JG and Penman S. Message and non-message sequences adjacent to poly(A) in steady state heterogeneous nuclear RNA of HeLa cells. *Cell* 1976;7(3):429-37.

Herrick D, Parker R and Jacobson A. Identification and comparison of stable and unstable mRNAs in Saccharomyces cerevisiae. *Mol Cell Biol* 1990;10(5):2269-84.

Hertzberg M, Sievertzon M, Aspeborg H, Nilsson P, Sandberg G and Lundeberg J. cDNA microarray analysis of small plant tissue samples using a cDNA tag target amplification protocol. *Plant J* 2001;25(5):585-91.

Hicks MJ, Yang CR, Kotlajich MV and Hertel KJ. Linking splicing to Pol II transcription stabilizes pre-mRNAs and influences splicing patterns. *PLoS Biol* 2006;4(6):e147.

Hiller M, Huse K, Platzer M and Backofen R. Creation and disruption of protein features by alternative splicing -- a novel mechanism to modulate function. *Genome Biol* 2005;6(7):R58.

Hilleren PJ and Parker R. Cytoplasmic degradation of splice-defective pre-mRNAs and intermediates. *Mol Cell* 2003;12(6):1453-65.

Hillman RT, Green RE and Brenner SE. An unappreciated role for RNA surveillance. *Genome Biol* 2004;5(2):R8.

Hilson P, Allemeersch J, Altmann T, Aubourg S, Avon A, Beynon J, Bhalerao RP*, et al.* Versatile gene-specific sequence tags for Arabidopsis functional genomics: transcript profiling and reverse genetics applications. *Genome Res* 2004;14(10B):2176-89.

Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979;6:65-70.

Holmberg A, Blomstergren A, Nord O, Lukacs M, Lundeberg J and Uhlen M. The biotin-streptavidin interaction can be reversibly broken using water at elevated temperatures. *Electrophoresis* 2005;26(3):501-10.

Holmberg J, Genander M, Halford MM, Anneren C, Sondell M, Chumley MJ, Silvany RE*, et al.* EphB receptors coordinate migration and proliferation in the intestinal stem cell niche. *Cell* 2006;125(6):1151-63.

Hosack DA, Dennis G, Jr., Sherman BT, Lane HC and Lempicki RA. Identifying biological themes within lists of genes with EASE. *Genome Biol* 2003;4(10):R70.

Houseley J, LaCava J and Tollervey D. RNA-quality control by the exosome. *Nat Rev Mol Cell Biol* 2006;7(7):529-39.

Hsu CL and Stevens A. Yeast cells lacking 5'-->3' exoribonuclease 1 contain mRNA species that are poly(A) deficient and partially lack the 5' cap structure. *Mol Cell Biol* 1993;13(8):4826-35.

Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM*, et al.* Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* 2001;19(4):342-7.

Iborra FJ, Jackson DA and Cook PR. Coupled transcription and translation within nuclei of mammalian cells. *Science* 2001;293(5532):1139-42.

Iborra FJ, Jackson DA and Cook PR. The case for nuclear translation. *J Cell Sci* 2004;117(Pt 24):5713-20.

Iborra FJ, Pombo A, Jackson DA and Cook PR. Active RNA polymerases are localized within discrete transcription "factories' in human nuclei. *J Cell Sci* 1996;109:1427-36.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B and Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 2003;31(4):e15.

Irizarry RA, Wu Z and Jaffee HA. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics* 2006;22(7):789-94.

Iscove NN, Barbara M, Gu M, Gibson M, Modi C and Winegarden N. Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. *Nat Biotechnol* 2002;20(9):940-3.

Ishigaki Y, Li X, Serin G and Maquat LE. Evidence for a pioneer round of mRNA translation: mRNAs subject to nonsense-mediated decay in mammalian cells are bound by CBP80 and CBP20. *Cell* 2001;106(5):607-17.

Ivanova NB, Dimos JT, Schaniel C, Hackney JA, Moore KA and Lemischka IR. A stem cell molecular signature. *Science* 2002;298(5593):601-4.

Jackson DA and Cook PR. Transcription occurs at a nucleoskeleton. *Embo J* 1985;4(4):919-25.

Jackson DA, Hassan AB, Errington RJ and Cook PR. Visualization of focal sites of transcription within human nuclei. *Embo J* 1993;12(3):1059-65.

Jurica MS and Moore MJ. Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell* 2003;12(1):5-14.

Kadaba S, Krueger A, Trice T, Krecic AM, Hinnebusch AG and Anderson J. Nuclear surveillance and degradation of hypomodified initiator tRNAMet in S. cerevisiae. *Genes Dev* 2004;18(11):1227-40.

Kadener S, Fededa JP, Rosbash M and Kornblihtt AR. Regulation of alternative splicing by a transcriptional enhancer through RNA pol II elongation. *Proc Natl Acad Sci U S A* 2002;99(12):8185-90.

Kadonaga JT. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* 2004;116(2):247-57.

Kanehisa M and Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28(1):27-30.

Kaposi-Novak P, Lee JS, Mikaelyan A, Patel V and Thorgeirsson SS. Oligonucleotide microarray analysis of aminoallyl-labeled cDNA targets from linear RNA amplification. *Biotechniques* 2004;37(4):580, 2-6, 8.

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H*, et al.* Antisense transcription in the mammalian transcriptome. *Science* 2005;309(5740):1564-6.

Keegan LP, Gallo A and O'Connell MA. The many roles of an RNA editor. *Nat Rev Genet* 2001;2(11):869-78.

Kerr MK and Churchill GA. Experimental design for gene expression microarrays. *Biostatistics* 2001;2(2):183-201.

Kerr MK, Martin M and Churchill GA. Analysis of variance for gene expression microarray data. *J Comput Biol* 2000;7(6):819-37.

Kim SH, Suddath FL, Quigley GJ, McPherson A, Sussman JL, Wang AH, Seeman NC*, et al.* Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science* 1974;185(149):435-40.

King MC and Wilson AC. Evolution at two levels in humans and chimpanzees. *Science* 1975;188(4184):107-16.

Kippin TE, Martens DJ and van der Kooy D. p21 loss compromises the relative quiescence of forebrain stem cell proliferation leading to exhaustion of their proliferation capacity. *Genes Dev* 2005;19(6):756-67.

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D*, et al.* CAGE: cap analysis of gene expression. *Nat Methods* 2006;3(3):211-22.

Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. *Science* 1974;184(139):868-71.

Kwek KY, Murphy S, Furger A, Thomas B, O'Gorman W, Kimura H, Proudfoot NJ*, et al.* U1 snRNA associates with TFIIH and regulates transcriptional initiation. *Nat Struct Biol* 2002;9(11):800-5.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K*, et al.* Initial sequencing and analysis of the human genome. *Nature* 2001;409(6822):860-921.

Lau NC, Lim LP, Weinstein EG and Bartel DP. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* 2001;294(5543):858-62.

Laurell C, Wirta, V., Nilsson, P., Lundeberg, J. Comparative analysis of a 3' end tag PCR and a linear RNA amplification approach for microarray analysis. *J Biotech* 2006;In press.

Lee CJ and Irizarry K. Alternative splicing in the nervous system: an emerging source of diversity and regulation. *Biol Psychiatry* 2003;54(8):771-6.

Lewis BP, Burge CB and Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 2005;120(1):15-20.

Lindvall O and Kokaia Z. Stem cells for the treatment of neurological disorders. *Nature* 2006;441(7097):1094-6.

Littauer UZ and Eisenberg H. Ribonucleic acid from Escherichia coli; preparation, characterization and physical properties. *Biochim Biophys Acta* 1959;32:320-37.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M*, et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996;14(13):1675-80.

Lopez AJ. Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu Rev Genet* 1998;32:279-305.

Lux C, Albiez H, Chapman RD, Heidinger M, Meininghaus M, Brack-Werner R, Lang A*, et al.* Transition from initiation to promoter proximal pausing requires the CTD of RNA polymerase II. *Nucleic Acids Res* 2005;33(16):5139-44.

MacKay VL, Li X, Flory MR, Turcott E, Law GL, Serikawa KA, Xu XL*, et al.* Gene expression analyzed by high-resolution state array analysis and quantitative proteomics: response of yeast to mating pheromone. *Mol Cell Proteomics* 2004;3(5):478-89.

Maquat LE. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat Rev Mol Cell Biol* 2004;5(2):89-99.

Marko NF, Frank B, Quackenbush J and Lee NH. A robust method for the amplification of RNA in the sense orientation. *BMC Genomics* 2005;6(1):27.

Martin GR. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc Natl Acad Sci U S A* 1981;78(12):7634-8.

Matsumoto K, Wassarman KM and Wolffe AP. Nuclear history of a pre-mRNA determines the translational activity of cytoplasmic mRNA. *Embo J* 1998;17(7):2107-21.

Mattick JS and Makunin IV. Small regulatory RNAs in mammals. *Hum Mol Genet* 2005;14 Spec No 1:R121-32.

Mattick JS and Makunin IV. Non-coding RNA. *Hum Mol Genet* 2006;15 Spec No 1:R17-29.

McCarroll SA, Murphy CT, Zou S, Pletcher SD, Chin CS, Jan YN, Kenyon C*, et al.* Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat Genet* 2004;36(2):197-204.

McCracken S, Fong N, Rosonina E, Yankulov K, Brothers G, Siderovski D, Hessel A*, et al.* 5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. *Genes Dev* 1997;11(24):3306-18.

McDonald D. Milestones: Gene Expression (1973-1974) The nucleosome hypothesis - An alternative string theory. *Nature* 2005:9.

Mendell JT, Sharifi NA, Meyers JL, Martinez-Murillo F and Dietz HC. Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nat Genet* 2004;36(10):1073-8.

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

Meyer S, Temme C and Wahle E. Messenger RNA turnover in eukaryotes: pathways and enzymes. *Crit Rev Biochem Mol Biol* 2004;39(4):197-216.

Mitchell P and Tollervey D. Musing on the structural organization of the exosome complex. *Nat Struct Biol* 2000;7(10):843-6.

Modrek B and Lee C. A genomic view of alternative splicing. *Nat Genet* 2002;30(1):13-9.

Moore KA and Lemischka IR. Stem cells and their niches. *Science* 2006;311(5769):1880-5.

Moore MJ. Nuclear RNA turnover. *Cell* 2002;108(4):431-4.

Morrison TB, Weis JJ and Wittwer CT. Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification. *Biotechniques* 1998;24(6):954-8, 60, 62.

Morshead CM, Benveniste P, Iscove NN and van der Kooy D. Hematopoietic competence is a rare property of neural stem cells that may depend on genetic and epigenetic alterations. *Nat Med* 2002;8(3):268-73.

Muhlrad D, Decker CJ and Parker R. Deadenylation of the unstable mRNA encoded by the yeast MFA2 gene leads to decapping followed by 5'-->3' digestion of the transcript. *Genes Dev* 1994;8(7):855-66.

Mukherjee D, Gao M, O'Connor JP, Raijmakers R, Pruijn G, Lutz CS and Wilusz J. The mammalian exosome mediates the efficient degradation of mRNAs that contain AU-rich elements. *Embo J* 2002;21(1-2):165-74.

Ng P, Tan JJ, Ooi HS, Lee YL, Chiu KP, Fullwood MJ, Srinivasan KG*, et al.* Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Res* 2006;34(12):e84.

Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, Gupta S*, et al*. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Methods* 2005;2(2):105-11.

Nissen P, Hansen J, Ban N, Moore PB and Steitz TA. The structural basis of ribosome activity in peptide bond synthesis. *Science* 2000;289(5481):920-30.

Niswender CM, Sanders-Bush E and Emeson RB. Identification and characterization of RNA editing events within the 5-HT2C receptor. *Ann N Y Acad Sci* 1998;861:38-48.

Nogues G, Kadener S, Cramer P, Bentley D and Kornblihtt AR. Transcriptional activators differ in their abilities to control alternative splicing. *J Biol Chem* 2002;277(45):43110-4.

Noller HF. RNA structure: reading the ribosome. *Science* 2005;309(5740):1508-14.

Nygaard V and Hovig E. Options available for profiling small samples: a review of sample amplification technology when combined with microarray profiling. *Nucleic Acids Res* 2006;34(3):996-1014.

Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, Goyenechea B*, et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet* 2004;36(10):1065-71.

Palmer S, Wiegand AP, Maldarelli F, Bazmi H, Mican JM, Polis M, Dewar RL*, et al.* New real-time reverse transcriptase-initiated PCR assay with single-copy sensitivity for human immunodeficiency virus type 1 RNA in plasma. *J Clin Microbiol* 2003;41(10):4531-6.

Pardal R, Clarke MF and Morrison SJ. Applying the principles of stem-cell biology to cancer. *Nat Rev Cancer* 2003;3(12):895-902.

Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A*, et al.* ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 2005;33(Database issue):D553-5.

Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP and Fodor SP. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci U S A* 1994;91(11):5022-6.

Petalidis L, Bhattacharyya S, Morris GA, Collins VP, Freeman TC and Lyons PA. Global amplification of mRNA by template-switching PCR: linearity and application to microarray analysis. *Nucleic Acids Res* 2003;31(22):e142.

Plath K, Mlynarczyk-Evans S, Nusinow DA and Panning B. Xist RNA and the mechanism of X chromosome inactivation. *Annu Rev Genet* 2002;36:233-78.

Prasanth KV, Prasanth SG, Xuan Z, Hearn S, Freier SM, Bennett CF, Zhang MQ*, et al.* Regulating gene expression through RNA nuclear retention. *Cell* 2005;123(2):249-63.

Prokopowich CD, Gregory TR and Crease TJ. The correlation between rDNA copy number and genome size in eukaryotes. *Genome* 2003;46(1):48-50.

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

Rajasekhar VK, Viale A, Socci ND, Wiedmann M, Hu X and Holland EC. Oncogenic Ras and Akt signaling contribute to glioblastoma formation by differential recruitment of existing mRNAs to polysomes. *Mol Cell* 2003;12(4):889-901.

Rajeevan MS, Dimulescu IM, Vernon SD, Verma M and Unger ER. Global amplification of sense RNA: a novel method to replicate and archive mRNA for gene expression analysis. *Genomics* 2003;82(4):491-7.

Ramakrishnan V. Ribosome structure and the mechanism of translation. *Cell* 2002;108(4):557-72.

Ramalho-Santos M, Yoon S, Matsuzaki Y, Mulligan RC and Melton DA. "Stemness": transcriptional profiling of embryonic and adult stem cells. *Science* 2002;298(5593):597-600.

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

Reynolds BA and Rietze RL. Neural stem cells and neurospheres--re-evaluating the relationship. *Nat Methods* 2005;2(5):333-6.

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

Rhee H, Polak L and Fuchs E. Lhx2 maintains stem cell character in hair follicles. *Science* 2006;312(5782):1946-9.

Robertus JD, Ladner JE, Finch JT, Rhodes D, Brown RS, Clark BF and Klug A. Structure of yeast phenylalanine tRNA at 3 A resolution. *Nature* 1974;250(467):546-51.

Rosonina E, Kaneko S and Manley JL. Terminating the transcript: breaking up is hard to do. *Genes Dev* 2006;20(9):1050-6.

Runte M, Huttenhofer A, Gross S, Kiefmann M, Horsthemke B and Buiting K. The IC-SNURF-SNRPN transcript serves as a host for multiple small nucleolar RNA species and as an antisense RNA for UBE3A. *Hum Mol Genet* 2001;10(23):2687-700.

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J*, et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 2003;34(2):374-8.

Sarkans U, Parkinson H, Lara GG, Oezcimen A, Sharma A, Abeygunawardena N, Contrino S*, et al.* The ArrayExpress gene expression database: a software engineering and implementation perspective. *Bioinformatics* 2005;21(8):1495-501.

Sartor M, Schwanekamp J, Halbleib D, Mohamed I, Karyala S, Medvedovic M and Tomlinson CR. Microarray results improve significantly as hybridization approaches equilibrium. *Biotechniques* 2004;36(5):790-6.

Schaupp CJ, Jiang G, Myers TG and Wilson MA. Active mixing during hybridization improves the accuracy and reproducibility of microarray results. *Biotechniques* 2005;38(1):117-9.

Schena M, Shalon D, Davis RW and Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;270(5235):467-70.

Schlingemann J, Thuerigen O, Ittrich C, Toedt G, Kramer H, Hahn M and Lichter P. Effective transcriptome amplification for expression profiling on sense-oriented oligonucleotide microarrays. *Nucleic Acids Res* 2005;33(3):e29.

Schluenzen F, Tocilj A, Zarivach R, Harms J, Gluehmann M, Janell D, Bashan A*, et al.* Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell* 2000;102(5):615-23.

Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE*, et al.* Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* 2000;101(6):671-84.

Schuwirth BS, Borovinskaya MA, Hau CW, Zhang W, Vila-Sanjurjo A, Holton JM and Cate JH. Structures of the bacterial ribosome at 3.5 A resolution. *Science* 2005;310(5749):827-34.

Shaheen HH and Hopper AK. Retrograde movement of tRNAs from the cytoplasm to the nucleus in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* 2005;102(32):11290-5.

Sharov AA, Piao Y, Matoba R, Dudekula DB, Qian Y, VanBuren V, Falco G*, et al.* Transcriptome analysis of mouse stem cells and early embryos. *PLoS Biol* 2003;1(3):E74.

Shatkin AJ. Capping of eucaryotic mRNAs. *Cell* 1976;9(4):645-53.

Shaw G and Kamen R. A conserved AU sequence from the 3' untranslated region of GM-CSF mRNA mediates selective mRNA degradation. *Cell* 1986;46(5):659-67.

Shilatifard A, Conaway RC and Conaway JW. The RNA polymerase II elongation complex. *Annu Rev Biochem* 2003;72:693-715.

Shin C and Manley JL. Cell signalling and the control of pre-mRNA splicing. *Nat Rev Mol Cell Biol* 2004;5(9):727-38.

Shuman S. Structure, mechanism, and evolution of the mRNA capping apparatus. *Prog Nucleic Acid Res Mol Biol* 2001;66:1-40.

Shyu AB, Greenberg ME and Belasco JG. The c-fos transcript is targeted for rapid decay by two distinct mRNA degradation pathways. *Genes Dev* 1989;3(1):60-72.

Sievertzon M. Transcript profiling of small tissue samples using microarray technology. PhD thesis, Royal Insitute of Technology 2005.

Singh SK, Clarke ID, Terasaki M, Bonn VE, Hawkins C, Squire J and Dirks PB. Identification of a cancer stem cell in human brain tumors. *Cancer Res* 2003;63(18):5821-8.

Slonim DK. From patterns to pathways: gene expression data analysis comes of age. *Nat Genet* 2002;32 Suppl:502-8.

Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004;3(1):Article3.

Smyth GK, Michaud J and Scott HS. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* 2005;21(9):2067-75.

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

Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D*, et al.* Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol* 2002;3(9):RESEARCH0046.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO*, et al.* Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 1998;9(12):3273-97.

Srivastava D and Ivey KN. Potential of stem-cell-based therapies for heart disease. *Nature* 2006;441(7097):1097-9.

Staley JP and Guthrie C. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* 1998;92(3):315-26.

Stears RL, Getts RC and Gullans SR. A novel, sensitive detection system for high-density microarrays using dendrimer technology. *Physiol Genomics* 2000;3(2):93-9.

Subkhankulova T and Livesey FJ. Comparative evaluation of linear and exponential amplification techniques for expression profiling at the single-cell level. *Genome Biol* 2006;7(3):R18.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102(43):15545-50.

Sun H and Chasin LA. Multiple splicing defects in an intronic false exon. *Mol Cell Biol* 2000;20(17):6414-25.

Suntharalingam M and Wente SR. Peering through the pore: nuclear pore complex structure, assembly, and function. *Dev Cell* 2003;4(6):775-89.

Takano A, Endo T and Yoshihisa T. tRNA actively shuttles between the nucleus and cytosol in yeast. *Science* 2005;309(5731):140-2.

Terskikh AV, Miyamoto T, Chang C, Diatchenko L and Weissman IL. Gene expression analysis of purified hematopoietic stem cells and committed progenitors. *Blood* 2003;102(1):94-101.

Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, Marshall VS and Jones JM. Embryonic stem cell lines derived from human blastocysts. *Science* 1998;282(5391):1145-7.

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

Valoczi A, Hornyik C, Varga N, Burgyan J, Kauppinen S and Havelda Z. Sensitive and specific detection of microRNAs by northern blot analysis using LNA-modified oligonucleotide probes. *Nucleic Acids Res* 2004;32(22):e175.

Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD and Eberwine JH. Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc Natl Acad Sci U S A* 1990;87(5):1663-7.

van Hoof A, Frischmeyer PA, Dietz HC and Parker R. Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. *Science* 2002;295(5563):2262-4.

Wang E, Miller LD, Ohnmacht GA, Liu ET and Marincola FM. High-fidelity mRNA amplification for gene profiling. *Nat Biotechnol* 2000;18(4):457-9.

Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D and Brown PO. Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci U S A* 2002;99(9):5860-5.

Wang Z and Kiledjian M. Functional link between the mammalian exosome and mRNA decapping. *Cell* 2001;107(6):751-62.

Wansink DG, Schul W, van der Kraan I, van Steensel B, van Driel R and de Jong L. Fluorescent labeling of nascent RNA reveals transcription by RNA polymerase II in domains scattered throughout the nucleus. *J Cell Biol* 1993;122(2):283-93.

Vargas DY, Raj A, Marras SA, Kramer FR and Tyagi S. Mechanism of mRNA transport in the nucleus. *Proc Natl Acad Sci U S A* 2005;102(47):17008-13.

Watson JD and Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 1953;171(4356):737-8.

Velculescu VE, Zhang L, Vogelstein B and Kinzler KW. Serial analysis of gene expression. *Science* 1995;270(5235):484-7.

Venables JP. Aberrant and alternative splicing in cancer. *Cancer Res* 2004;64(21):7647-54.

Wilcox AS, Khan AS, Hopkins JA and Sikela JM. 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* 1991;19(8):1837-43.

Wilson DN and Nierhaus KH. Ribosomal proteins in the spotlight. *Crit Rev Biochem Mol Biol* 2005;40(5):243-67.

Wimberly BT, Brodersen DE, Clemons WM, Jr., Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T*, et al.* Structure of the 30S ribosomal subunit. *Nature* 2000;407(6802):327-39.

Wong ML and Medrano JF. Real-time PCR for mRNA quantitation. *Biotechniques* 2005;39(1):75-85.

Voorhoeve PM, le Sage C, Schrier M, Gillis AJ, Stoop H, Nagel R, Liu YP*, et al.* A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors. *Cell* 2006;124(6):1169-81.

Wu L, Fan J and Belasco JG. MicroRNAs direct rapid deadenylation of mRNA. *Proc Natl Acad Sci U S A* 2006;103(11):4034-9.

Xing Y, Xu Q and Lee C. Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. *FEBS Lett* 2003;555(3):572-8.

Yang CW, Hsiao CF and Chou CK. Evaluation of experimental designs for two-color cDNA microarrays. *J Comput Biol* 2005;12(9):1202-20.

Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M and Darnell JE, Jr. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res* 2003;13(8):1863-72.

Yang HY, Buckley M, Dudoit S and Speed T (2000). TechReport 584: Comparison of methods for image analysis on c{DNA} microarray data. Department of Statistics, University of California at Berkeley Technical Reports.

Yang W, Chendrimada TP, Wang Q, Higuchi M, Seeburg PH, Shiekhattar R and Nishikura K. Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nat Struct Mol Biol* 2006;13(1):13-21.

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

Yi R, O'Carroll D, Pasolli HA, Zhang Z, Dietrich FS, Tarakhovsky A and Fuchs E. Morphogenesis in skin is governed by discrete sets of differentially expressed microRNAs. *Nat Genet* 2006;38(3):356-62.

Yin H, Wang MD, Svoboda K, Landick R, Block SM and Gelles J. Transcription against an applied force. *Science* 1995;270(5242):1653-7.

Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH and Noller HF. Crystal structure of the ribosome at 5.5 A resolution. *Science* 2001;292(5518):883-96.

Zamore PD, Tuschl T, Sharp PA and Bartel DP. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* 2000;101(1):25-33.

Zhang Z and Carmichael GG. The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. *Cell* 2001;106(4):465-75.

Zhu B, Xu F and Baba Y. An evaluation of linear RNA amplification in cDNA microarray gene expression analysis. *Mol Genet Metab* 2006;87(1):71-9.

Zhu J, Shendure J, Mitra RD and Church GM. Single molecule profiling of alternative pre-mRNA splicing. *Science* 2003;301(5634):836-8.