Transcriptional profiling of human embryonic stem cells and their functional derivatives

Jane Synnergren

DOCTORAL DISSERTATION

To be defended 28th of October 2010

Department of Clinical Chemistry and Transfusion Medicine
Institute of Biomedicine at Sahlgrenska Academy
University of Gothenburg, Sweden

FACULTY OPPONENT

Professor Mahendra Rao
Buck Institute for Age research
Novato, CA

To my lovely family

Tommy, Sara, and Robin

-you gave me inspiration, comfort, and mental relaxation whenever best needed

"Science is organized knowledge,
wisdom is organized life."

[Immanuel Kant , 1724-1804]

"It is easy to lie with statistics.
It is hard to tell the truth without statistics."

[Andrejs Dunkels, 1939-1938]

"….technology tends to overwhelm common sense."

[David. A. Freedman, 1938-2008]

Abstract

Human embryonic stem cells (hESCs) represent populations of pluripotent, undifferentiated cells with unlimited replication capacity, and with the ability to differentiate into any functional cell type in the human body. Based on these properties, hESCs and their derivatives provide unique model systems for basic research on embryonic development. Also, industrial *in vitro* applications of hESCs are now beginning to find their way into the fields of drug discovery and toxicology. Moreover, hESC-derivatives are anticipated to be promising resources for future cell replacement therapies. However, in order to fully utilize the potential of hESCs it is necessary to increase our knowledge about the processes that govern the differentiation of these cells. At present, some of the major challenges in stem cell research are heterogeneous cell populations, insufficient yield of the differentiated cell types and immature derivatives with limited functionality. To address these problems, a better understanding of the regulatory mechanisms that control the lineage commitment is needed. The aim of this thesis has been to increase the knowledge of the global transcriptional programs which are activated when cells differentiate along specific pathways, and to identify key genes that show differential expression at specific stages of differentiation. The results indicate that hESCs express a unique set of housekeeping genes that are stably expressed in this specific cell type and in their derivatives, which highlights the importance of proper validation of reference genes for usage in hESCs. Furthermore, an extensive characterization of hESCs and differentiated progenies of the cardiac and hepatic lineages has been conducted, and sets of differentially expressed genes were identified. Two different protocols, which mediate definitive and primitive endoderm respectively, were studied, and important discrepancies between these two cell types were identified. Moreover, the global expression profile of hESC-derived cardiomyocyte clusters were thoroughly investigated and compared to that of foetal and adult heart. To further study regulatory mechanisms of importance during stem cell differentiation, the global expression of microRNAs (miRNAs) was also investigated. Putative target genes of differentially expressed miRNAs were identified using computational predictions, and their mRNA expression was analysed. Notably, an interesting correlation between the miRNA and mRNA expression was observed, which supports the general notion that miRNAs bind to and degrade their target mRNAs, and thus act as fine-tuning regulators of gene expression. Taken together, the results described in this thesis provide important information for further studies on regulatory mechanisms that control the differentiation of hESCs into functional cell types such as cardiomyocytes and hepatocytes.

List of publications

This thesis is based on the following papers, referred to in the text by their roman numerals:

I. **Jane Synnergren**, Theresa L. Giesler, Sudeshna Adak, Reti Tandon, Karin Noaksson, Anders Lindahl, Patric Nilsson, Deidre Nelson, Björn Olsson, Mikael C.O. Englund, Stewart Abbot, Peter Sartipy (2007). Differentiating human embryonic stem cells express a unique housekeeping gene signature. *Stem Cells*, 25(2): 473-480.

II. **Jane Synnergren**, Karolina Åkesson, Kerstin Dahlenborg, Hilmar Vidarsson, Caroline Améen, Daniella Steel, Anders Lindahl, Björn Olsson, Peter Sartipy (2008). Molecular signature of cardiomyocyte clusters derived from human embryonic stem cells. *Stem Cells*, 26(7): 1831-1840.

III. **Jane Synnergren**, Nico Heins, Gabriella Brolén, Gustav Eriksson, Anders Lindahl, Johan Hyllner, Björn Olsson, Peter Sartipy, Petter Björquist (2010). Transcriptional profiling of human embryonic stem cells differentiating to definitive and primitive endoderm and further towards the hepatic lineage. *Stem Cells Dev*. Jul;19(7): 961-78.

IV. **Jane Synnergren**, Caroline Améen, Anders Lindahl, Björn Olsson, Peter Sartipy. Expression of microRNAs and their target mRNAs in human stem cell derived cardiomyocyte clusters and in heart tissue. Accepted for publication in *Physiol Genomics,* 2010 Sep 14. [Epub ahead of print]

Abbreviations

AH	adult heart
cDNA	complementary DNA
CM	cardiomyocyte
CMC	cardiomyocyte clusters
cRNA	complementary RNA
CV	coefficient of variation
DE	definitive endoderm
DNA	deoxyribonucleic acid
EB	embryoid body
ECM	extracellular matrices
END-2	endoderm-like cell line
ESC	embryonic stem cells
EST	expressed sequence tag
FC	fold change
FDR	false discovery rate
FH	foetal heart
GO	gene ontology
GSA	gene set analysis
HD	high density
hESC	human embryonic stem cells
HGF	hepatocyte growth factor
HKG	housekeeping gene
ICM	inner cell mass
IGA	individual gene analysis
iPS	induced pluripotent stem
miRNA	microRNA
MPSS	massively parallel signature sequencing
mRNA	messenger RNA
PCA	principle component analysis
PCR	polymerase chain reaction
PIN	protein interaction network
PrE	primitive endoderm
RISC	RNA-induced silencing complex
RMA	robust multichip average
RNA	ribonucleic acid
RNAP	RNA polymerase
RT-PCR	reverse transcription polymerase chain reaction

SAGE	serial analysis of gene expression
SAM	significance analysis of microarray data
SCID	severe combined immunodeficiency
SOM	self organising maps
STRING	search tool for the retrieval of interacting genes
tRNA	transport RNA
UD	undifferentiated

Gene symbols

ACTB	beta actin
AFP	alpha fetoprotein
ALB	albumin
CAV2	caveolin 2
CD44	CD44 molecule (gene)
CDH17	cadherin 17
CEBPA	CCAAT/enhancer binding protein, alfa
CER1	cerberus 1
CLIC5	chloride intracellular channel 5
COL8A1	collagen, type VIII, alpha 1
CXCR4	chemokine (C-X-C motif) receptor 4
DPP4	dipeptidyl-peptidase 4
EMP1	epithelial membrane protein 1
EPAS1	endothelial PAS domain protein 1
FBXL12	F-box and leucine-rich repeat protein 12
FHOD3	formin homology 2 domain containing 3
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
GATA4	GATA binding protein 4
GSC	goosecoid
HPRT	hypoxanthine guanine phosphoribosyl transferase
ITGB3	integrin, beta 3
KRT7	keratin 7
LONRF2	LON peptidase N-terminal domain and ring finger 2
MEF	mouse embryonic fibroblasts
MEF2C	myocyte enhancer factor 2C
MET	met proto-oncogene (hepatocyte growth factor receptor)
MIXL1	Mix1 homeobox-like 1

MSRB3	methionine sulfoxide reductase B3
MYH6	myosin, heavy chain 6, cardiac muscle, alpha
MYH7	myosin, heavy chain 7, cardiac muscle, beta
NANOG	Nanog homeobox
NFAT	nuclear factor of activated T-cells 5, tonicity-responsive
NKX2.5	NK2 transcription factor related, locus 5
NPPA	natriuretic peptide precursor A
NTN4	netrin 4
OCT4	POU class 5 homeobox 1 (POU5F1)
PLD1	phospholipase D1, phosphatidylcholine-specific
PLN	phospholamban
RBM24	RNA binding motif protein 24
RNF7	ring finger protein 7
RUNX1	runt-related transcription factor 1
SERPINA7	serpin peptidase inhibitor, cladeA (alpha-1antiproteinase, antitrypsin) member 7
SOX17	SRY (sex determining region Y)-box 17
SOX2	SRY (sex determining region Y)-box 2
TCEA3	transcription elongation factor A, 3
TF	transferrin
TM4SF1	transmembrane 4 L six family member 1
TNNT2	troponin T type 2 (cardiac)
TUBB	beta tubulin
UBD	ubiquitin D
α-MHC	alpha-myosin heavy chain

Table of Contents

Introduction

Stem cells are generic cells that can develop into many different types of cells. As such they can serve as an important repair system for the organism and they have therefore received a lot of interest from scientists during the last decades. In general, there are two main types of stem cells: *embryonic* and *adult* stem cells, and these two types have different characteristics and different potential [1]. In 1998, the first success of culturing human embryonic stem cells (hESCs) *in vitro* over multiple passages was reported [2] and since then, hESCs have attracted incredible attention as they offer great possibilities within many medical fields. Recently, researchers have also been able to successfully re-program differentiated somatic cells into an induced pluripotent state (i.e., iPS-cells) that in the future potentially will allow for the creation of patient- and disease-specific stem cells [3]. In basic research, stem cells can provide a human model system, important for studying fundamental processes during embryonic development [4]. They can provide tools for development of new drugs, and they offer great possibilities in regenerative medicine and for curing various diseases [5-7]. However, there are many obstacles to overcome before the potential of these cells can be fully realised. One of the most important issues is to increase the understanding about the gene regulatory mechanisms that control the differentiation of hESCs. Therefore, this thesis will focus on analyses of global gene expression during differentiation of hESCs towards the cardiomyocyte (CM) and hepatocyte lineages, with the aim to extend our knowledge of the transcriptional programs that are activated during these differentiation processes.

Definition of stem cells

Stem cells have two key characteristics, they can self-replicate for an indefinite period of time and they can differentiate into many specialised cell types [2]. The two main types of stem cells, *adult stem cells* and *embryonic stem cells,* have different origins and characteristics (further described below). Various types of cells have diverse degrees of differentiation potential. By definition, a *totipotent* cell can specialise into any cell type in an organism including the extraembryonic tissues, and a *pluripotent* cell can differentiate into any of the three germ layers mesoderm, endoderm and ectoderm, a *multipotent* cell can specialise into several cell types (usually present within one tissue/organ). Finally, *unipotent* cells can only specialise into one mature cell type. Adult stem cells are undifferentiated (unspecialised) cells that are present in a differentiated (specialised) tissue. They can self-renew for the lifetime of the organism and they are multipotent, i.e., can differentiate into any of the specialised cell types of the tissue from which they originate [1]. Embryonic stem cells (ESCs) are present in the inner cell mass (ICM) of the blastocyst only for a short time during the earliest stages of the development of the embryo. The ESCs can proliferate and they can differentiate into all different cell types in the organism, and are therefore referred to as pluripotent cells.

Human embryonic stem cells

Human ESCs represent populations of pluripotent, undifferentiated cells with unlimited replication capacity, and with the ability to differentiate into the three germ layers (ectoderm, endoderm and mesoderm) and further towards all the different types of cells in the human body [2]. These cell populations grow as compact colonies of undifferentiated cells on mouse [2, 8] or human [9] feeders (Figure 1). They can also be cultured in feeder-free conditions using matrix and conditioned medium [10]. Recent reports also demonstrate defined culture conditions for hESCs [11-14]. Importantly, hESCs can be maintained *in vitro* in their pluripotent state or they can be coaxed to differentiate along specific pathways to form a variety of specialised cell types [15].

Figure 1. Human ESCs on three different feeder systems.
Shown to the left is a mouse feeder system, in the middle is a human feeder system, and to the right is a feeder-free culture system. The illustration is a courtesy from Cellartis AB.

The potential of human embryonic stem cells

Due to the characteristics of hESCs, these cells are extremely promising in a wide range of applications. They constitute a model system for studying basic developmental processes and the formation of different tissues and organs, which, for ethical reasons, otherwise cannot be done in humans. Moreover, they provide platforms for various *in vitro* applications (e.g., in drug discovery), models for studying various diseases, and in the future, hESCs and their differentiated progenies are promising resources for cell replacement therapies [4, 5, 16].

Derivation of human embryonic stem cells

Human ESCs are derived from a 4-6 days old fertilized egg at the blastocyst stage. The blastocyst possesses three different structures; the ICM, which later forms the embryo by transformation through the three germ layers, the cavity known as the blastocoele, and an outer layer of cells called the trophoblast, which surrounds the blastocoele and later forms the placenta [1]. At this stage the ICM is isolated by the use of microsurgery or enzymatic dispersion of the trophoblast (Figure 2). The isolated ICM is then plated into a culture dish coated with e.g., mouse or human fibroblasts or matrigel, to which the cells attach and grow in specific media. The presence of a feeder layer is essential for ESCs since they provide signals necessary for sustaining the pluripotent phenotype. When the cells attach to the feeders they start to proliferate and the colony spreads over the surface.

To keep them in an undifferentiated state, the cells need to be passaged (dissociated and re-plated) before they start to form 3D structures. The passaging can be performed either mechanically or enzymatically. The dissociated pieces of cell colonies are re-plated on new feeders where they grow as individual colonies with preserved undifferentiated morphology, and this process is repeated every 4-5 days (Figure 2).

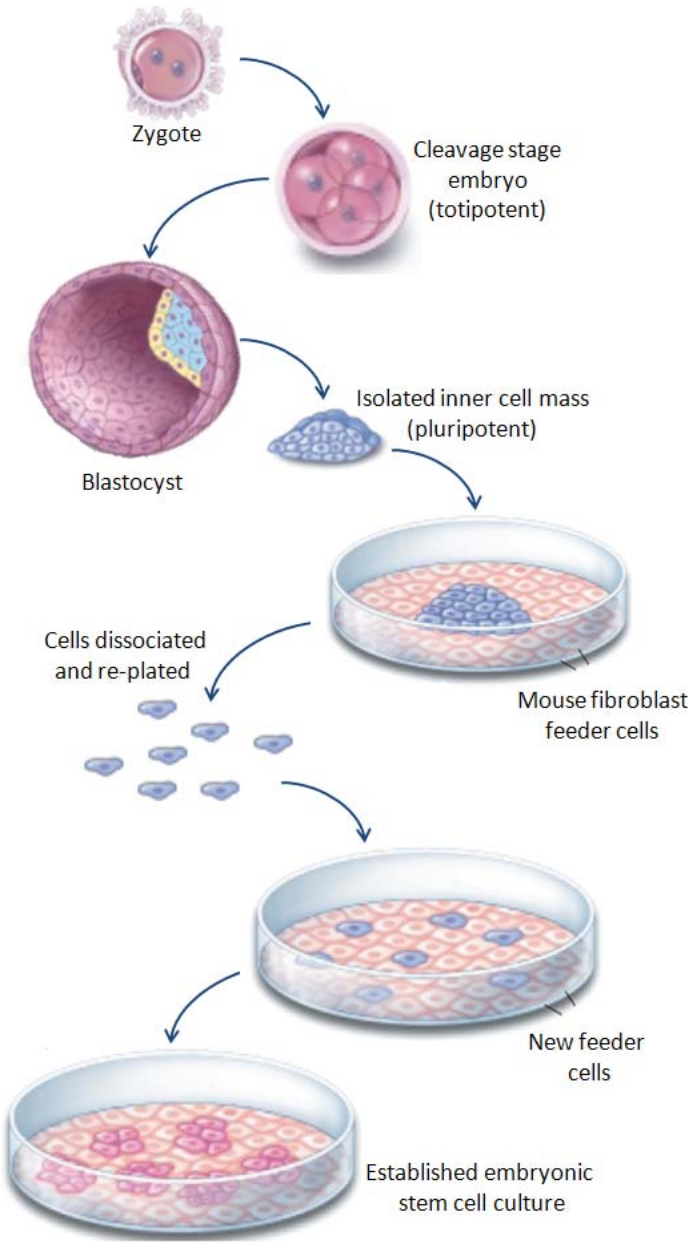

Figure 2. Derivation of a human embryonic stem cell line.
Surplus *in vitro* fertilized eggs at day 4-6 after fertilization are used to establish a cell line. The ICM is isolated and placed on a coated culture dish. When the cells attach to the surface they start to proliferate. To keep the cells in an undifferentiated state they must be regularly passaged, and placed on new dishes to prevent the formation of 3D structures. Cells at the cleavage stage embryo are totipotent and cells in the isolated ICM are pluripotent. Illustration is reproduced from [1], with permission from Therese Winslow.

Characterisation of human embryonic stem cells

To establish the identity of hESCs and their functional derivatives, the cells need to be extensively characterised. This includes morphological inspection, analysis of telomerase activity, karyotyping, investigation of pluripotency, expression of unique cell-surface antigens and tissue-specific enzymatic activity, as well as expression of typical marker genes [17]. It has been demonstrated that high telomerase activity in ESCs correlates well with their ability to proliferate indefinitely in culture [17]. Moreover, analysis of the nuclear chromosomal karyotype provides means to assess the genetic stability of established hESC lines, which may be affected if hESCs are maintained in culture for extended periods of time [18]. Their ability to differentiate to various cell types is analysed both *in vitro*

and *in vivo*. The pluripotency *in vitro* is typically assessed by formation of embryoid bodies (EBs) [19] which initiate spontaneous differentiation. Antibody techniques are then used to stain the cells for typical markers, representative of all three germ layers. To assess the pluripotency *in vivo,* the hESCs are injected under the kidney capsule of SCID (Severe Combined Immunodeficiency) mice to let form teratomas, and these teratomas are then analysed to confirm that all three germ layers are represented in the tumours. Global transcriptional profiling provides a powerful characterization method as one can define a transcriptional fingerprint for hESCs and their differentiated progenies, and identify novel markers. The focus for this thesis project has been to characterise hESCs and their functional derivatives, by performing global gene expression profiling using microarrays.

Differentiation of human embryonic stem cells

As demonstrated by several investigators [19-21], hESCs are pluripotent and can efficiently differentiate into all the three germ layers mesoderm, endoderm, and ectoderm, and further into various functional cell types (Figure 3). However, these are extremely complicated processes that are dependent on many different parameters such as timing, concentrations and combinations of growth factors, as well as other cell culture conditions. Currently, a major goal for hESC research is to learn how to control the differentiation into specific functional cells, which is required for the future use of these cells in drug development, in screening studies for toxins, and in therapeutic applications. In recent years, significant progress towards the understanding of cellular differentiation has been fuelled, in part, by studying gene expression using microarrays [22-27] and this thesis project has contributed to this progress. The ectoderm germ layer and its derivatives is the most studied of these three, and has hence not been further investigated in this project. Instead, we have in detail explored the differentiation through the mesoderm and endoderm germ layers, and investigated the derivatives thereof, such as cardiomyocytes and hepatocytes.

Gene transcription and protein translation

Gene expression is the process by which information from a gene is copied from the gene to an mRNA sequence, which is then used in the synthesis of a functional gene product, a protein. The genetic code is mediated by the gene expression, and the process from transcription of a gene to a functional protein involves several steps, such as transcription of the gene in the nucleus, and transport of the mRNA to the cytoplasm where translation to a protein is carried out aided by tRNAs (Figure 4). The properties of the expression products give rise to the phenotype of an organism. By means of gene regulation, the cell has control over its structure and function, and this is the basis for cellular differentiation, morphogenesis and the versatility and adaptability of any organism [28].Transcriptional regulation is also essential for evolutionary changes, since control of the timing, location, and amount of gene expression often have profound effects on the functions of the gene in a cell [28]. When conducting gene expression studies, it is important to understand the basic concepts behind these processes for a proper interpretation of the data.

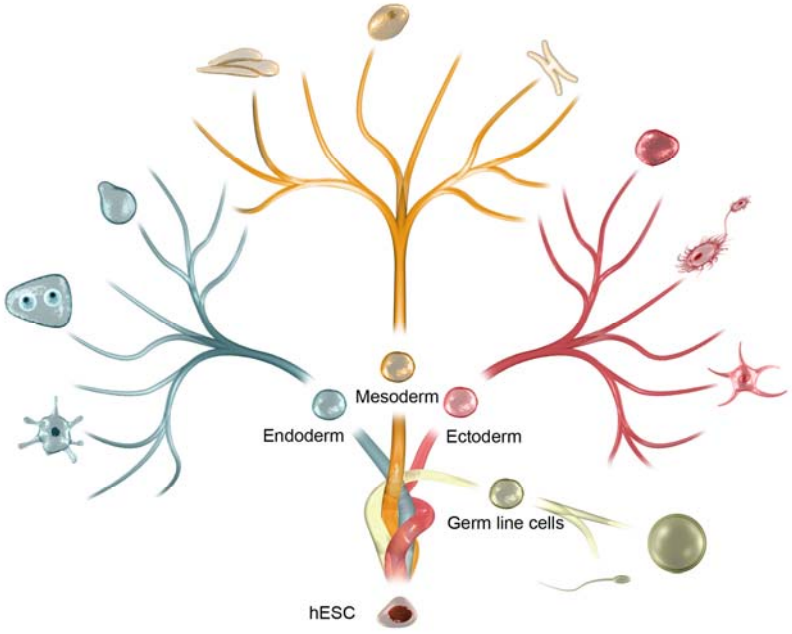

Figure 3. Differentiation of human embryonic stem cells.
The pluripotent stem cells differentiate through the three germ layers mesoderm, endoderm, and ectoderm, and further into specialised cell types. The Illustration reproduced from Jensen et al. [105], with permission from John Wiley and Sons.

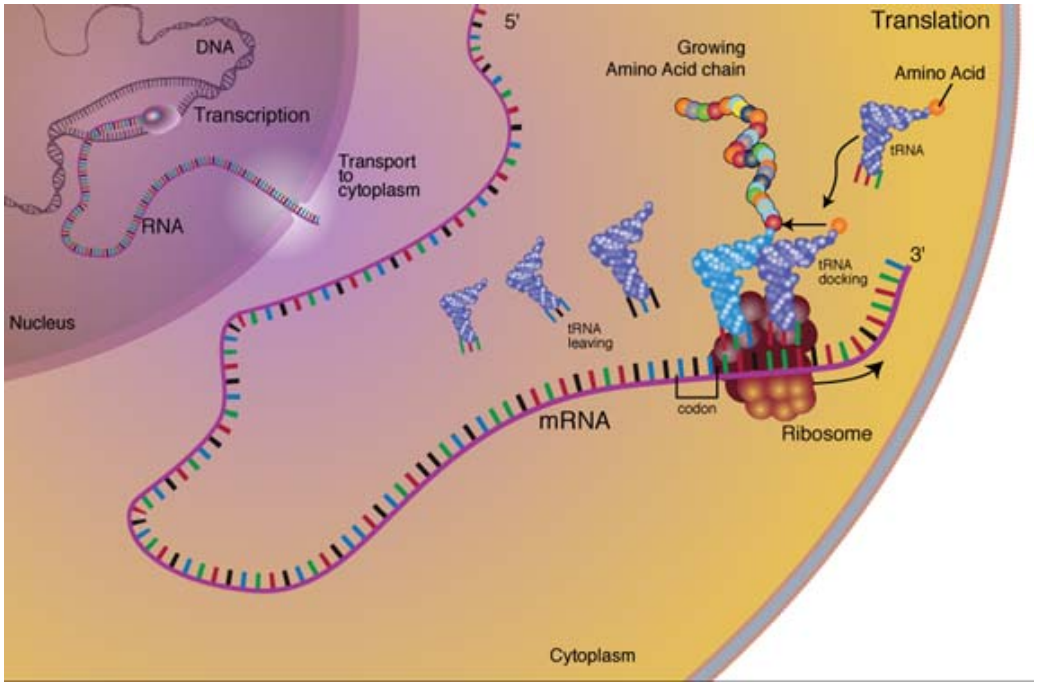

Figure 4. Overview of the transcription and translation processes in a cell.
Messenger RNAs (mRNA) are transcribed from a gene and transported from the nucleus to the cytoplasm, where the translation to a protein is carried out by ribosomes. The amino acids, which are syntesised to a polypetide, are transported to the ribosome by tRNAs. Illustration reproduced from Talking Glossary of Genetics.

Transcriptional regulation

The transcription of genes involves intricate dynamic dependencies which makes it challenging to study. Several mechanisms have been shown to be critical for the initiation of transcription, the rate of transcription, and the subsequent processing of the mRNA. These regulatory mechanisms control when the transcription occurs and the amount of mRNA produced [28]. The transcription of a gene is carried out by RNA polymerase and the process is regulated by several components [28] (Figure 5).

- *Specificity factors* control the ability of the RNA polymerase to bind to a specific promoter or set of promoters.
- *Repressors* bind to non-coding regions, close to or overlapping with the promoter for a gene, and impede the RNA polymerase's progress along the DNA strand, thus hampering the transcription of the gene.
- *General transcription factors* aid in positioning the RNA polymerase at the start of a protein coding sequence.
- *Activators* enhance the interaction between the RNA polymerase and the specific promoter.
- *Enhancers* are sites on the DNA helix that are bound to by activators in order to loop the DNA and bring a specific promoter to the initiation complex.

Splicing of mRNA

Splicing is a modification of an RNA molecule post-transcription, in which introns are removed and exons are joined together (Figure 6). Hence, after transcription of a gene the pre-mRNA is spliced to mRNA, typically in a series of reactions. This is necessary before the mRNA can leave the nucleus and be transported to the cytoplasm, where it is translated to a protein. The presence of introns in the genome is restricted only to the eukaryotic organisms. Splicing is performed mainly by sets of small nuclear RNAs that together with sets of proteins form the spliceosome, which is responsible for the splicing in the cell [28]. RNA splicing allows for packing of more information into every gene as the transcripts from one single gene can be spliced in various ways to produce different mRNAs, depending on the cell type in which the gene is being expressed or the stage of the development of the organism [28]. As a consequence, different proteins can be produced by the same gene and it is estimated that 60% of the human genes undergo such alternative splicing [28]. Thus, RNA splicing increases the already enourmous coding potential of eukaryotic genomes, at the same time as it complicates the studies of gene transcription. This is because the complexity increases dramatically when there, as in many cases, are several different transcripts transcribed by one single gene.

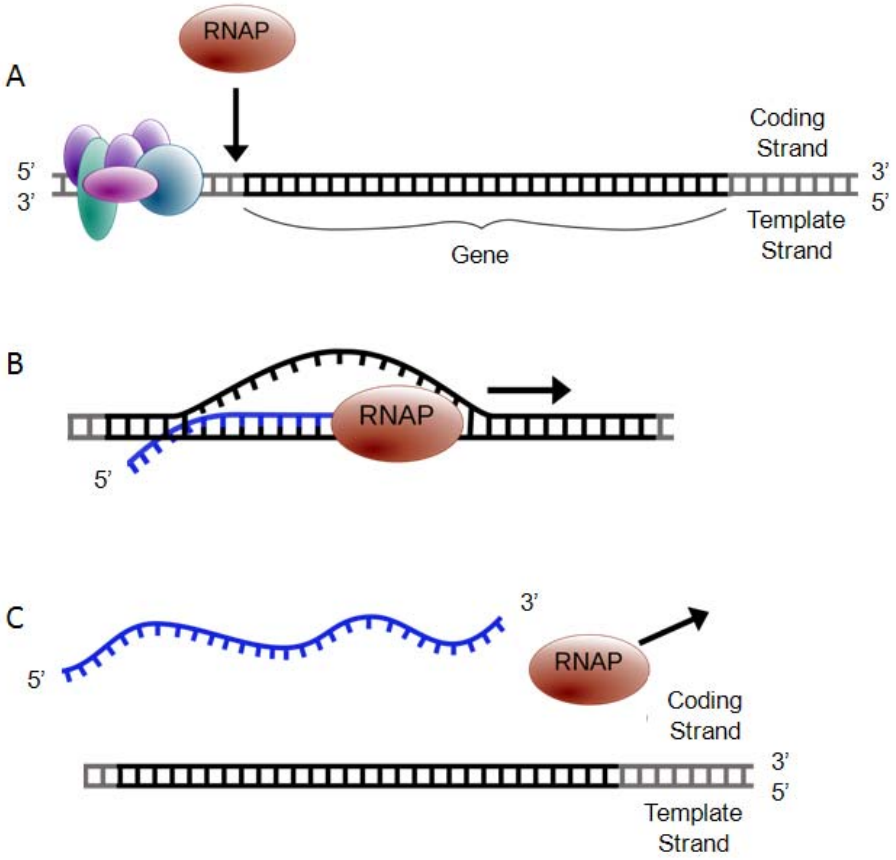

Figure 5. Transcription of a gene.
A: The initiation of transcription is guided by attachment of a collection of proteins, transcription factors, that bind to the promoter and mediates the binding of the RNA polymerase (RNAP). B: RNAP traverses the template strand and uses base pairing complementary with the template strand to create an RNA copy (blue). C: At the termination of transcription, the RNAP is released from the template strand and a tail of adenines is added to the mRNA sequence at the 3' end, in a process called polyadenylation. The illustration was modified and re-produced from Wikipedia Commons.

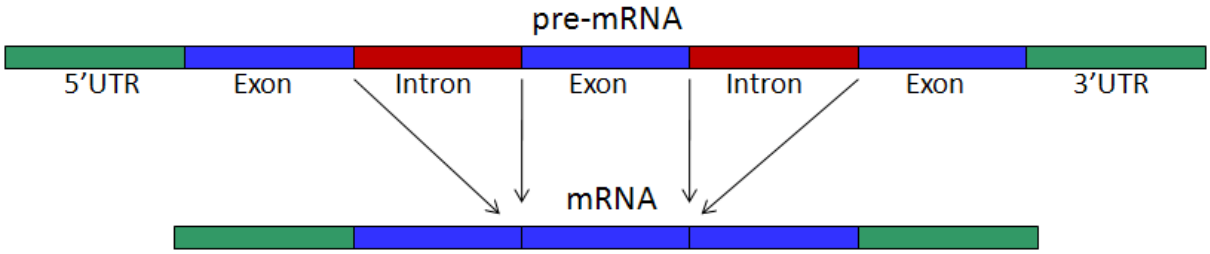

Figure 6. Splicing of pre-mRNA.
The introns are removed before formation of the mRNA sequence. Different sets of exons can be selected to form the mRNA which means that one pre-mRNA can give rise to several variants of mRNA sequences.

Translation to protein

After splicing the pre-mRNA into mRNA, the transcript is transported from the nucleus to the cytoplasm, where the translation occurs by means of ribosomes, which bind to the mRNA sequence. The same mRNA sequence can be translated many times, and therefore, the period of time that a mature mRNA molecule persists in the cell influences the amount of protein that is produced. The lifetime of mRNAs differs considerably and is dependent on the nucleotide sequence of the mRNA itself, as well as the type of cell in which the mRNA is produced. The typical lifetime for mRNA molecules in eukaryotic cells ranges from 30 minutes up to 10 hours [28]. One nucleotide cannot directly be translated to an amino acid since there are only four types of nucleotides in the mRNA, and 20 different types of amino acids that build up a protein. Therefore the information is translated into amino acid sequences by means of the *genetic code*. The sequence of nucleotides in the mRNA is read in groups of three, denoted *codons*, which increases the number of unique combinations [28]. Each codon specifies one amino acid, and small transfer molecules known as tRNAs match the amino acids to the correct codon.

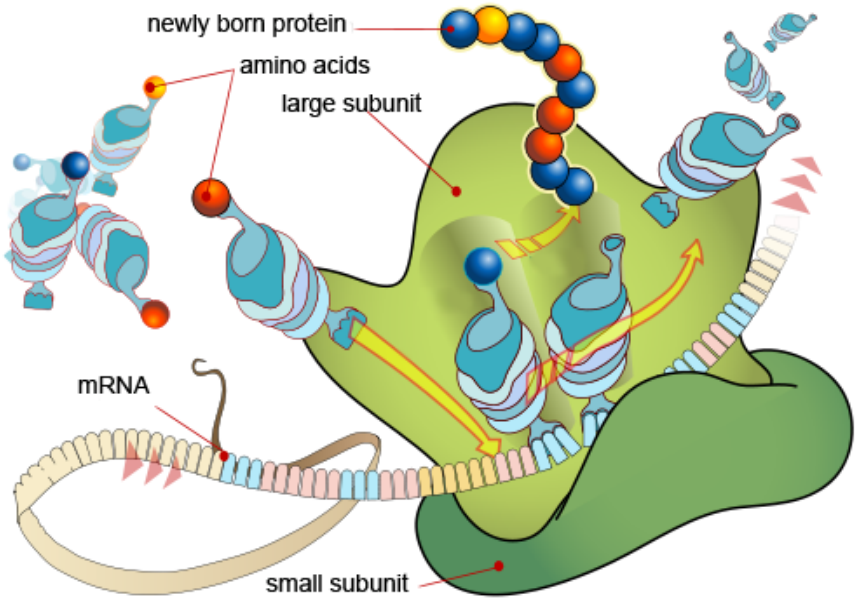

Figure 7. Translation of the mRNA into a protein takes place in ribosomes.
Amino acids are transported by means of tRNAs to the ribosome, where they are bound to each other in a polypeptide that forms the new protein. The order in which the amino acids are bound together is determined by the order of the nucleotides in the mRNA sequence. Illustration reproduced with permission from Mariana Ruiz Villarreal.

The genetic code is partly redundant, since several codons can specify a single amino acid. Depending on where in the sequence the de-coding begins, each mRNA sequence can be translated in three different, non-overlapping, reading frames but only one of these is the correct one [28]. The translation of an mRNA begins with a specific start codon (AUG) and is then performed in the direction 5' cap to 3' end. The translation of the codons and the synthesising of the amino acids into a polypeptide that forms the protein are performed by the ribosomes (Figure 7). The specific amino acids that are chained together into a polypeptide are carried to the ribosome by tRNAs. Once protein synthesis has been initiated, each new amino acid is added to the elongating chain in a cycle of reactions. The end of a protein coding mRNA is indicated by the presence of one of three stop codons (UAA, UAG, UGA), which signals to the ribosome to stop the translation. After the protein is synthesized, important post-translational modifications are carried out which extends the range of functions of the protein, by attaching to it other biochemical functional groups [28].

Housekeeping genes

Housekeeping genes (HKGs) are genes that are involved in basic functions needed for the sustenance of the cell, and are assumed to be constitutively expressed in different cell types and under various conditions [29]. They have therefore been used as endogenous controls in normalisation of gene expression data, which aims to reduce non-biological variation [30]. However, with the advent of genome-wide expression profiling, the mRNA levels of many HKGs were observed to vary extensively between different cell types [31]. Therefore, researchers instead turned to various statistical methods for normalising large scale gene expression data [32, 33]. These methods are based on the assumption that most of the measured genes remain unchanged, which is usually correct in large scale genome-wide studies [32]. However, smaller experiments, where focused arrays or quantitative real-time PCR are used still require carefully selected and validated HKGs for normalisation, to adequately correct for inter-sample variation [34, 35]. In general, investigators have also used the traditional HKGs (e.g., GAPDH, ACTB, TUBB) in studies of hESCs [36, 37]. However, it is well known that the expression of several of these genes varies considerably in adult tissues, and their suitability as reference genes in hESCs requires further investigation.

MicroRNAs

An additional level of cellular regulation involves a family of tiny molecules, known as microRNAs (miRNAs). These are 19–25 nucleotide non-coding RNAs that bind to the 3' untranslated region of target mRNAs through imperfect matching. In mammalian genomes, miRNAs are predicted to regulate the expression of approximately 30% of the protein-coding genes [38]. Knowledge about the biological functions of most miRNAs identified thus far is still lacking, but it has been shown that they play important roles in embryo development, determination of cell fate, cell proliferation, and cell differentiation [39, 40].

Processing of miRNAs

MicroRNAs are derived from approximately 70 nucleotide long precursors, encoded by introns or intergenic regions, and are expressed in most organisms ranging from plants to humans. Figure 8 outlines schematically the different steps in the generation of mature miRNAs. The primary miRNAs are processed and cleaved in the cell nucleus by an enzyme called Drosha, which works in concert with the RNA binding protein Pasha. Subsequently, these pre-miRNAs are transported to the cytoplasm by exportin-5. In the cytoplasm, further cleavage is performed by Dicer. One of the remnant single strands (the so called "guide strand") is selected by an Argonaute protein and is integrated into the RNA-induced silencing complex (RISC).

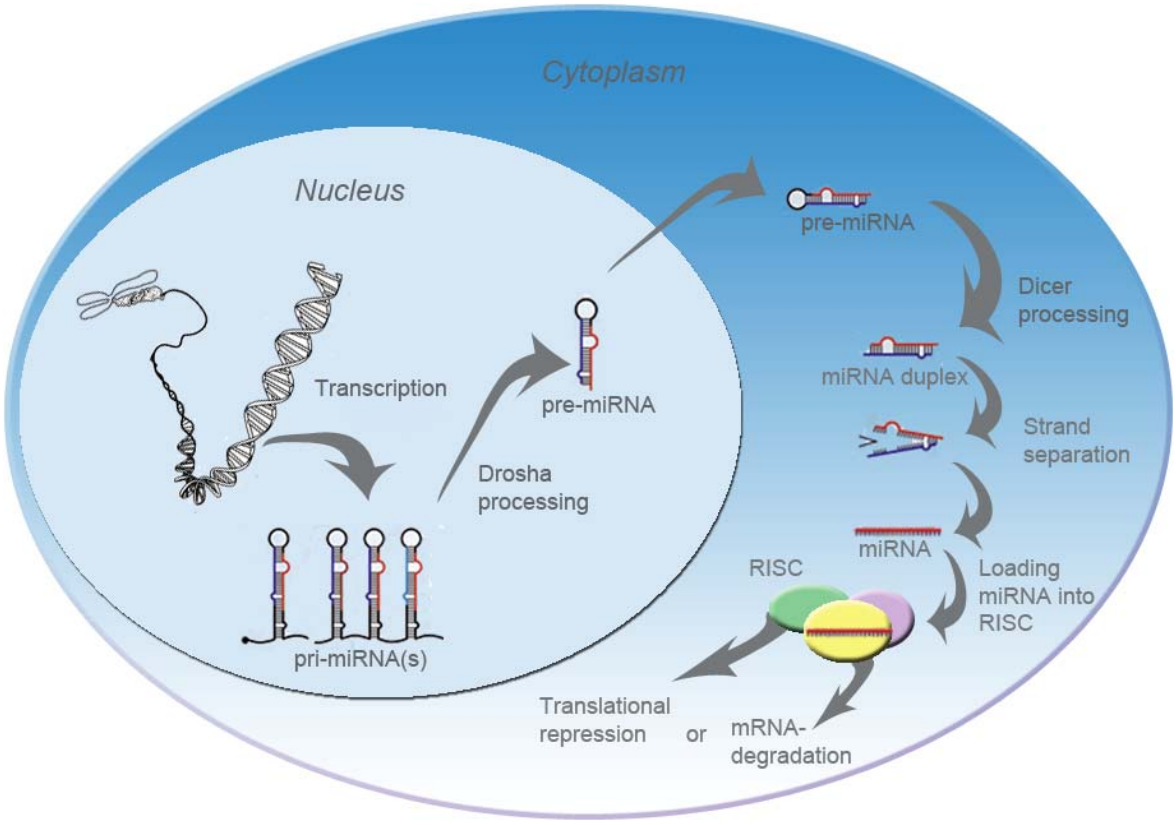

Figure 8. MiRNA processing from transcription to mature miRNA.
The primary miRNA (pri-miRNA) is processed and cleaved into pre-miRNAs in the cell nucleus by the enzyme Drosha. These pre-miRNAs are then transported to the cytoplasm by exportin-5. In the cytoplasm, further cleavage is performed by an enzyme called Dicer. One of the remaining single strands (the *guide strand*) is selected by an Argonaute protein and is then integrated into the RISC complex.

Functions of miRNAs

Many miRNAs appear to be expressed at different levels in various tissues, and the maturation and function of the tissues seem to be influenced by their presence. Interestingly, results from recent studies have indicated important roles for miRNAs in the control of diverse aspects of heart formation and cardiac function [41, 42]. It is also known that miRNAs are involved in various types of cancer by targeting tumour suppressing genes [43, 44]. MicroRNAs bind to their target mRNAs and negatively regulate their expression, either by repression of translation or by degradation of the mRNA [38]. Increased expression levels of miRNAs can also result in upregulation of previously suppressed target genes either directly, by decreasing the expression of inhibitory proteins and/or transcription factors, or indirectly, by inhibiting the expression levels of inhibitory miRNAs [45]. Depending on the state of the cell, miRNAs have also been observed to affect the translation of target mRNAs by regulation of their stability [45, 46]. Moreover, it has been shown that combinatorial regulation by miRNAs is common, which enables complex regulatory programs that are exceptionally challenging to dissect [47].

Global transcriptional profiling techniques

There are several high throughput techniques for measuring gene expression at large scale, such as expressed sequence tags (EST)-enumeration, Serial Analysis of Gene Expression (SAGE), Massively Parallel Signature Sequencing (MPSS) and different types of microarrays (described in more detail below). In EST-enumeration the expression levels are assessed by counting the number of ESTs for a particular gene, in a random selection of transcripts from a cDNA library derived from the sample. The ESTs are clustered into groups of sequences originating from the same transcript, and a longer consensus sequence is defined, which is then aligned to the genome to find the matching gene sequence. Both SAGE and MPSS are sequencing based techniques that use tags to identify and count the mRNAs, but the biochemical manipulation and the sequencing approaches differ substantially between these techniques. Both methods are based on the principle that a short sequence tag contains sufficient information to uniquely identify a transcript, provided that the tag is obtained from a specific position within each transcript.

In SAGE, short tags, usually 9-10 base pairs in length are extracted from each mRNA, at a defined position. These tags are then linked together to form long serial molecules that can be cloned and sequenced. The quantification is performed by counting the number of times a specific tag is observed in the sequenced molecule. Finally, the tags are matched to the corresponding genes. In MPSS, the extracted signatures are longer, 17-20 base pairs. Everyone of these signatures is cloned into a vector, which is labelled with a unique 32 base pair oligonucleotide tag. The tag is then attached to one of millions of microbeads, by hybridization of the tag to a complementary sequence on the bead. The signatures on the microbeads are then sequenced and matched to the corresponding genes, and subsequently quantified by counting the number of beads.

The longer tag sequences, used in MPSS, provide higher specificity compared to SAGE. Another advantage of MPSS is the larger library size. One disadvantage that applies to both SAGE and MPSS is the loss of certain transcripts due to lack of restriction enzyme recognition sites, and ambiguity in tag annotation. Compared to microarray techniques, sequencing techniques, which are not based on hybridizations, give on the other hand a more exact quantitative value. This is because the number of transcripts is counted directly, instead of quantifying spot intensities which are prone to background noise. Another advantage is that the mRNA sequences do not need to be known beforehand, and therefore also previously unknown transcripts can be detected. Nevertheless, microarray experiments are much cheaper to perform and are therefore usually used in large scale experiments.

Microarray technology

The microarray technology was introduced in the early 1990s, and during the last two decades the precision of the technology has increased considerably and, at the same time, the cost has decreased. Microarrays render the possibility to monitor the expression of thousands of genes simultaneously. Investigators are using the microarray technology to try to understand fundamental aspects of growth and development as well as to explore the pathogenesis of many human diseases. By monitoring the cells at various time points during a biological process or at specific biological conditions, one obtains snapshots of the global transcriptional profile at different stages. The principle behind the microarray technology is base pairing of DNA/RNA. When two complementary sequences come together, such as the immobilized probe on the array and the mobile target in the sample, they will lock together (hybridise). The microarray consists of a surface on which millions of probes are immobilised. The surface is divided into features (locations) and each feature on the microarray has a superfluous number of probes that correspond to a specific transcript.

When labelled target transcripts are hybridised onto the microarray, these bind complementary to their probes (Figure 9). The general procedure for performing a microarray experiment (which varies somewhat depending on the type of system) includes a series of steps [48]. Initially, the RNA is reverse transcribed, usually to cDNA, and labelled with a fluorophore, and then the solution is hybridised onto the array. After the hybridisation, the arrays are thoroughly washed, rinsed, and dried to remove non-hybridised transcripts from the surface. They are subsequently scanned to measure the fluorescence intensity for each feature on the array and these intensities are then translated into expression values. The feature intensities are directly proportional to the number of transcripts corresponding to each gene, and thus to the expression level of the gene.

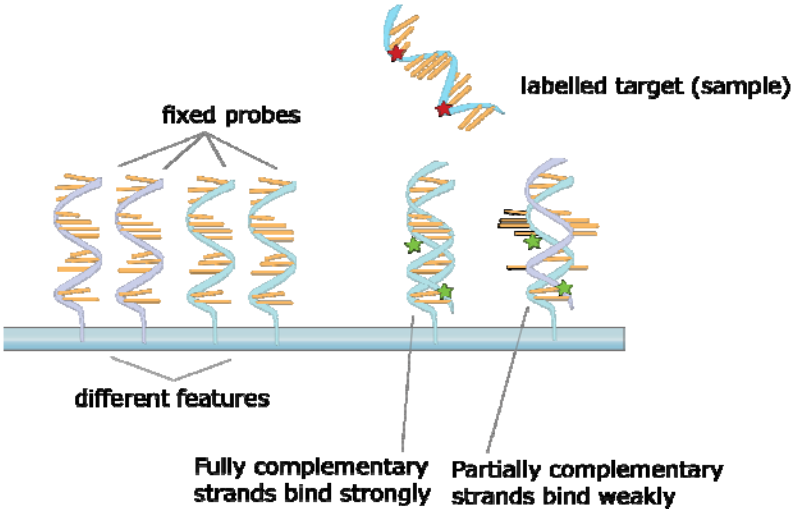

Figure 9. Schematic picture showing hybridisation of targets onto the microarray.
Labelled targets are hybridised on the array by the principle of base pairing. The array consists of different features (locations) that represent different genes. Each feature has a superfluous number of identical probes immobilised. Only fully complementary strands bind strongly during the hybridisation. Weakly bound targets are removed during the washing of the microarrays. Illustration reproduced from Wikipedia Commons.

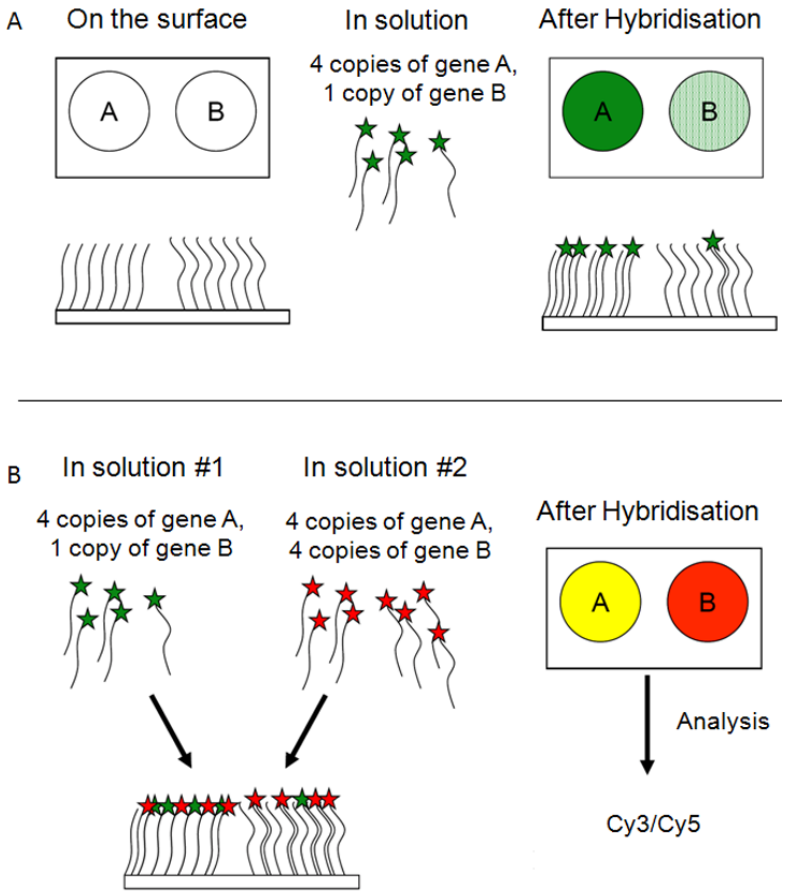

Figure 10.
Overview of one- and two-channel hybridisation.
Round-shaped features contain superfluous identical probes that hybridise with labelled targets from the samples. The shape of the features may vary between different microarray platforms. The intensity of the colour is proportional to the number of probes that are hybridised to that feature. Panel A shows the one-channel system and panel B the two-channel system. Yellow colour means equal amounts of red and green labelled targets.

Different types of microarrays

There are several different types of microarrays and the broadest distinction is whether the probes are spatially arranged on a slide made of glass, silicon or plastic or, if they are coded on microscopic polystyrene beads. They can be fabricated using different techniques, where the most common ones are robotic printing of the features on the array or synthesis of the probes *in situ* using techniques such as photolithography. Moreover, the arrays vary in the way the signals are detected, and they are designed for hybridisation of either one or two samples on the same array (one- or two-channel arrays). On one-channel arrays (also called oligonucleotide arrays) only one sample can be hybridised on each array, and the intensity levels are measured rather than the ratio between two intensities (Figure 10A). Therefore, comparison of two conditions requires two separate single-dye hybridisations. On two-channel arrays two samples are labelled with two different fluorophores, typically Cy3 and Cy5, which have different fluorescence emission wavelengths. The two Cy-labelled cDNA samples are mixed and hybridised to a single microarray (Figure 10B). Since the fluorophores have different excitation wavelengths it is possible to split the two signals during the scanning and calculate the intensities of each fluorophore, and use this in ratio-based analysis to identify up- and downregulated genes. One benefit of one-channel arrays is that the data is more easily compared to data from different experiments, as long as batch effects have been accounted for. However, using the one-channel system may require twice as many microarrays to compare samples within an experiment than with the two-channel system. Depending on which system is used, the experimental design, and the generated data, the subsequent data analysis may differ.

CodeLink microarrays

CodeLink™ Human Whole Genome Bioarrays are one-channel arrays that use 30-mer probes, which mainly target transcripts selected from the NCBI UniGene, RefSeq, and dbEST databases [49]. These arrays are based on polyacrylamide substrate which is photocross-linked to a glass slide and which has specific functional groups to which the 5' end of an oligonucleotide is attached via a hexylamine linker [49]. This 3D hydrophilic polymer matrix surface facilitates probe-target hybridisation [50] and yields improvements in spot density [49]. CodeLink Bioarrays have demonstrated high sensitivity for low expressed targets, low variability between arrays, and high specificity in distinguishing between highly homologous sequences [49, 51, 52].

Affymetrix microarrays

The Affymetrix platform is the most widely used commercial platform, providing a whole range of different types of arrays and covering various species. Affymetrix arrays are *in situ* synthesized, applying the photolithography technology to synthesise thousands to millions of 25-mer cDNA oligonucleotides in parallel [53]. By using light-sensitive masking agents, a sequence is "built", one nucleotide at a time, across the entire array. Typical for Affymetrix arrays are the multiple probe pairs for each transcript [54] (Figure 11). One

single probe pair consists of a perfect match sequence and a corresponding mismatch sequence, with a mismatch at the 13th nucleotide, designed to measure the amount of non-specific binding [54]. Each transcript is represented by 11-20 probe pairs, referred to as a probe set, and these probe pairs target the transcripts at the 3' end. A new type of Affymetrix array which recently has entered the market is the Whole Transcript arrays, including both Gene ST 1.0 and Exon ST 1.0 arrays [55]. The characteristic of these arrays is that they have an increased number of probes targeting exons along the whole transcript and not only in the 3' end [55]. The Gene ST 1.0 array has 1-2 probes per exon and the more comprehensive Exon ST 1.0 has four probes per exon. The main differences between GeneChip 133 Plus 2.0 and the newer Gene ST 1.0 are the following [55].

- cDNA instead of cRNA is hybridised to the arrays, which results in a more specific binding
- Random priming is applied instead of poly dT, thus querying target exons along the whole transcript instead of only in the 3' end
- Gene ST 1.0 covers a more restricted set of only well annotated transcripts from RefSeq, Ensembl, and GeneBank.

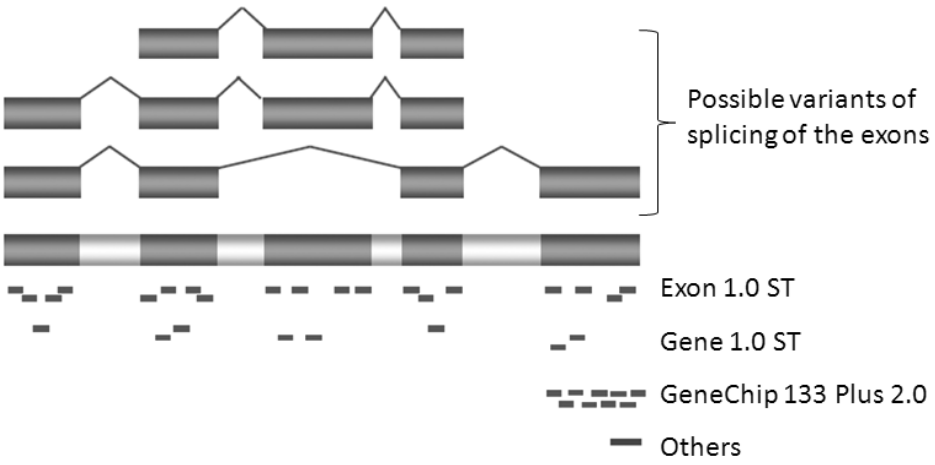

Figure 11. Distribution of the probes along the transcripts.
For the Exon 1.0 ST and the Gene 1.0 ST arrays the distribution of probes is querying the whole length of a transcript instead of only in the 3' end as for GeneChip 133 Plus 2.0, as well as for other types of arrays. This increases the sensitivity and specificity of the microarrays and also makes it possible to detect different splicing variants of a transcript.

Reliability and reproducibility of microarray data

The microarray technology has had tremendous impact on gene expression analysis during the last decade. However, publications of studies with dissimilar or even contradictory results have raised concerns regarding the reliability of this technology [56-60]. For example, several global gene expression studies of stem cells have shown poor overlap [61-63]. To address these and other concerns, such as performance and data analysis issues, the MicroArray Quality Control project was initiated by the US Food and Drug Administration. Using an impressive number of laboratories, this comprehensive study

showed both intra-platform consistencies across laboratories and a high level of inter-platform concordance in terms of genes identified as differentially expressed [59, 60]. Nevertheless, there are several issues to be aware of when using this technology, which can introduce substantial biases in the final results. Examples of such issues to consider are:

- Cross-hybridisation: There is a risk that some mRNAs may cross-hybridise to probes on the array that are supposed to detect other mRNAs.
- Fold change compression: Due to various technical limitations, such as limited dynamic range and signal saturation, a certain level of fold change (FC) compression is expected for microarray data compared to e.g., RT-PCR data [64, 65].
- Poor sensitivity for low expressed transcripts: Problems with relatively poor sensitivity in detecting small FCs have been reported for several microarray platforms [64].
- Cross-platform inconsistency: Inconsistent probe annotations across platforms, which leads to difficulties to ascertain that probes on various platforms aimed at the same gene do in fact quantify the same mRNA transcript [58].
- Dye-biases: In two-channel systems the fluorescent dyes usually have different dynamic ranges and quantum yields, which is partially adjusted for by appropriate normalisation but may not be completely eliminated.
- Non-biological variations: There is always a risk that variations may be introduced during the experimental procedure (e.g., different persons performing the experiment, minor variations in temperature or duration for the reverse transcription and hybridisation) [66] and these sometimes add substantial noise to the system. However, this source of variation is not unique to microarray experiments but is also an issue in other reverse transcription reactions [64].

Bioinformatics

The work in this thesis has a strong focus on bioinformatics, which is the application of statistics and computer science to the field of molecular biology. Bioinformatics has arisen from the needs of biologists to interpret the vast amounts of data that constantly are generated in e.g., genomics, proteomics, and functional genomics research. The primary goal of bioinformatics is to increase the understanding of biological processes by development and application of computational techniques. However, dealing with bioinformatics is challenging and in biology there are no rules without exception, and biological processes are extremely complex with a vast number of interacting components that are dependent in various ways. Yet another challenge is that most of the data is fragmented, incomplete, and noisy. There is therefore also a need for bioinformatic tools that allow researchers to compare carefully the relationship between new data and data that has been validated by experiments [67]. Large scale gene expression experiments generate enormous datasets that are computationally demanding to analyse. Today, there are a lot of tools and software available, both commercially and open source, for solving various bioinformatic problems, such as identification of differentially expressed genes, clustering of data, and identification of interaction networks.

Scientific aim

The overall aim of this thesis was to increase the understanding of the transcriptional programs that are active during hESC differentiation towards the cardiac and hepatic lineages, and contribute with knowledge that may assist future studies of regulatory mechanisms that control hESC differentiation. Such knowledge can be genes that are differentially expressed in various stages during the differentiation and thus might be candidate genes in regulatory mechanisms.

Specific aims

- To investigate the stability of commonly used HKGs in differentiating hESCs and identify a novel set of HKGs that show stable expression in hESCs and derivatives thereof (Paper I).

- To analyse the global gene expression patterns and identify differentially expressed genes and induced pathways in hESC-derived cardiomyocyte clusters (Paper II).

- To analyse the global gene expression patterns and identify differentially expressed genes in hESCs that differentiate towards endoderm and further into hepatocyte-like cells (Paper III).

- To investigate the correlation between miRNA and mRNA expression in hESC-derived cardiomyocyte clusters and in foetal and adult heart tissue, and identify miRNAs that are differentially expressed in both hESC-derived cardiomyocyte clusters and in heart tissue samples (Paper IV).

Gene expression data

The biological materials that have been subjects of investigation are derived from hESCs and differentiated derivatives thereof (Cellartis AB, Göteborg, www.cellartis.se). Details regarding the preparation of the cell material used in each study can be found in Paper I-IV.

Microarray experiments

A number of microarray experiments have been conducted during this thesis project, to generate several extensive gene expression datasets from hESCs and their derivatives. RNA was extracted from the collected cell material using standard methods, and subsequently analysed with microarrays. Three different types of microarrays have been used in the project.

- CodeLink Human Whole Genome Bioarrays (GE Healthcare, Piscataway, NJ)
- GeneChip Human, HGU 133 Plus 2.0 (Affymetrix, Santa Clara, CA)
- Gene ST 1.0 arrays (Affymetrix, Santa Clara, CA)

All three types are one-channel arrays, which mean that the generated datasets consist of relative expression values rather than ratios between two samples (as for two-channel arrays). However, since different microarray systems have been used, the data is not directly comparable across the experiments.

Microarray experiment in Paper I

The first study, described in Paper I, was designed to investigate the stability of commonly used HKGs in data from hESCs, to validate their usability as reference genes in subsequent studies in this thesis project. Subsequently, we also aimed to define a novel set of genes that showed stable expression in hESCs and their differentiated progenies. For this purpose, the CodeLink Human Whole Genome Bioarrays, targeting approximately 57,000 transcripts and ESTs, was applied to generate gene expression data. The CodeLink arrays have shown particularly high sensitivity for low expressed transcripts [51]. The experimental design in this study (Figure 12) included a high density (HD) protocol, which is a spontaneous differentiation protocol where the hESCs were maintained on mouse embryonic fibroblasts (MEF), and harvested at day 5, 11 and 25 after passage for subsequent RNA extraction. In the second protocol, the hESC cultures are transferred from MEF to suspension for EB formation. At day 11, after six days in suspension, the EBs were plated onto gelatin-coated culture dishes to allow for further differentiation. At day 25, i.e., 14 days after plating of the EBs, the cells were harvested for RNA extraction. This experimental set-up was repeated for the three hESC lines SA001, SA002 and SA002.5 (Cellartis AB, Göteborg) and run in triplicates. Total RNA was extracted from all samples using Qiagen RNeasy Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. DNase treatment was performed on-column using Qiagen RNase-free DNase Kit (Qiagen). Total RNA was used to generate cRNA,

which was then assessed for quality before being hybridised onto the microarrays. The arrays were then washed and scanned and the expression values were extracted. Bad quality spots were filtered and the data was median normalised and $\log_2$ transformed before subsequent data analysis.

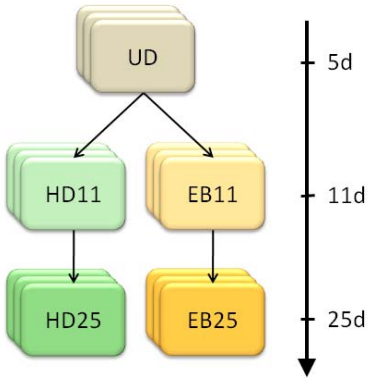

Figure 12. Experimental design for microarray experiment described in Paper I.
Three different time points (5, 11 and 25 days) and two differentiation protocols (HD and EB) were included in the experiment which was repeated in three different cell lines.

Microarray experiment in Paper II

The purpose of the study described in Paper II was to characterise hESC-derived cardiomyocyte clusters (CMCs) at the gene expression level and globally investigate their transcriptional patterns. This required only a rather simple design with no more than two groups to compare, undifferentiated (UD) hESCs and hESC-derived CMCs. The material consisted of one pooled sample of UD hESCs and two different biological replicates of pooled hESC-derived CMCs, harvested at a number of time points up to 22 days after initiation of differentiation (Figure 13). The hESC line SA002 was used in this experiment. Due to technical issues, two separate sets of microarray experiments were conducted. In the first, one-cycle amplified RNA was used, while in the second set of experiments two-cycle amplified RNA was used due to the limited amount of available RNA for some of the samples. Even though no obvious differences between the two datasets could be observed, all subsequent calculations between samples were conducted within each experiment separately. The quality of the RNA and cRNA, labelled by *in vitro* transcription, was tested and the fragmented cRNA was then hybridised to the microarrays.

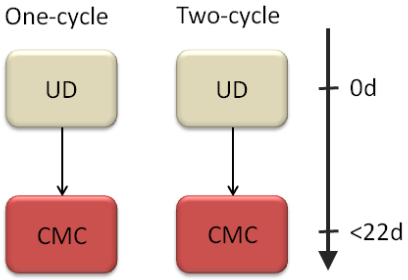

Figure 13. Experimental design for microarray experiment described in Paper II.
Two different groups (UD and CMC) were included in the experiment which was repeated two times using one-cycle and two-cycle amplification, respectively. Cell line SA002 was used in the experiment.

Each sample was hybridised to duplicate arrays from the Affymetrix microarray platform, GeneChip 133 Plus 2.0 (Affymetrix, Santa Clara, CA), targeting approximately 54,000 transcripts. The main reason for switching to the Affymetrix platform was the availability of standardised procedures for data analysis. Extraction of expression values and scaling of data were performed using the MAS5 algorithm and transcripts flagged as 'Absent' on all arrays were filtered and the data was $\log_2$ transformed before the data analysis.

Microarray experiment in Paper III

Paper III describes a comparison between hESCs differentiated through the endoderm, either definitive endoderm (DE) or primitive endoderm (PrE), as well as a global transcriptional characterisation of endoderm, hepatocyte progenitors, and hepatocyte-like cells. A comprehensive experimental design was applied in this work including three cell lines (SA002, SA167, and SA461) and four time points, as well as two separate differentiation protocols (Figure 14). The hepatocellular carcinoma cell line (HepG2) was included as a reference sample in the experiment. Similarly as in Paper II, the human GeneChip 133 Plus 2.0 microarray from Affymetrix was used, and each sample was cultured and harvested in biological duplicates. The RNA was extracted and assessed for quality before generation of cRNA, and subsequently hybridised to the arrays using similar procedure as in Paper II. The raw data was extracted and normalised using MAS5 and filtered and $\log_2$ transformed before subsequent data analysis.

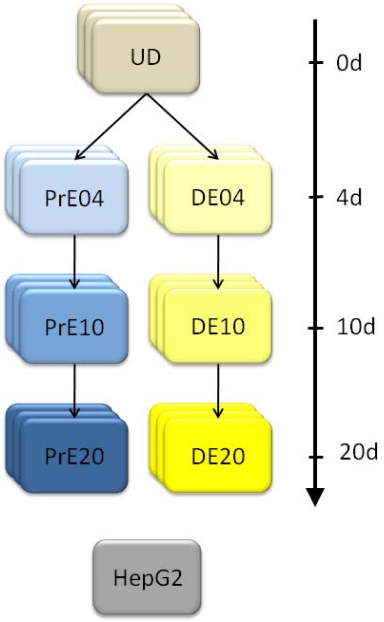

Figure 14. Experimental design for the microarray experiment described in Paper III.
Four time points (UD, 4 days, 10 days, and 20 days) and two differentiation protocols (PrE and DE) were included in the experiment, which was repeated for three different cell lines (SA002, SA167, and SA461). HepG2 was included as a reference sample in the study.

Microarray experiment in Paper IV

To further our understanding of the regulatory mechanisms of transcription and translation, the study described in Paper IV investigated the putative correlation between mRNA and miRNA expression. Thus, mRNA and miRNA microarray experiments were designed in which matched samples from hESCs and hESC-derived CMCs were collected for global mRNA and miRNA profiling.

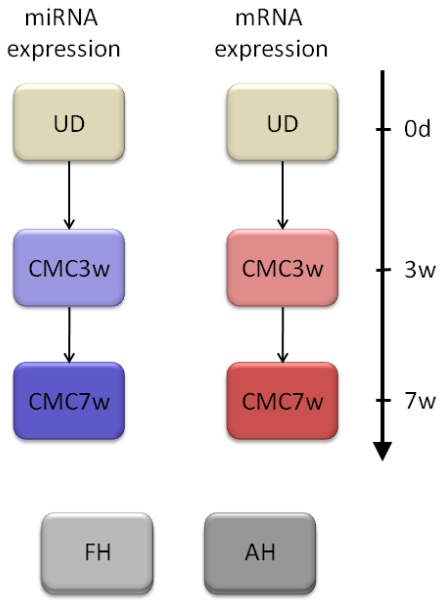

Figure 15. Experimental design for the microarray experiment described in Paper IV. Three time points (UD, CMC3 weeks, and CMC7 weeks) were analysed and foetal heart (FH) and adult heart (AH) were included as reference samples. Both miRNA (blue) and mRNA (red) expression were analysed in parallel.

Cell line SA002 was used in this experiment. Total RNA was extracted using the Ambion miRVana miRNA isolation kit (Ambion, www.ambion.com) which preserves small molecules. The RNA was split into two aliquots, and microarray experiments were conducted in parallel to measure both miRNA and mRNA expression of paired samples. As illustrated in Figure 15, the material consisted of samples of UD hESCs and hESC-derived CMCs, cultured for 3 (CMC3w) and 7 weeks (CMC7w) after onset of differentiation. Each sample collection was repeated three times to generate biological replicates. In addition, triplicate samples from foetal heart (FH) and adult heart (AH) (Yorkshire Bioscience, www.york-bio.com) were included as reference material.

The miRNA expression was measured using the miRCURY™ LNA array version 11.0 from Exiqon (www.exiqon.com), following the manufacturer's instructions. After hybridisation, the microarray slides were washed and scanned and the image analysis was carried out using the ImaGene 8.0 software (BioDiscovery, www.biodiscovery.com). The quantified signals were background corrected and normalised using the global Lowess regression algorithm. For investigation of the mRNA expression, the Whole Transcript Gene ST 1.0 arrays (Affymetrix) were used. Expression signals were extracted and normalised by means of the Expression Console™ (Affymetrix) applying the Robust Multichip Average (RMA) normalisation method that by default outputs $\log_2$ transformed values.

Bioinformatic and statistical analysis

Analysis of microarray data

The raw data from microarray experiments need to be pre-processed in several steps, before conducting any high level data analysis. Depending on the array type and the platform, these pre-processing steps vary, but basically involve subtraction of background and normalisation for removal of non-biological variations. The data are also typically $\log_2$-transformed to achieve roughly normally distributed data, and potential outliers are excluded before performing the high level analysis. Due to the large amounts of data generated in microarray experiments, advanced bioinformatic algorithms (described below) are required for efficient interpretation of the data into valuable biological information. In the area of gene expression analysis there are e.g., algorithms for:

- identification of differentially expressed genes
- clustering of gene expression data
- pathway analysis
- derivation of protein interaction networks
- functional annotation of regulated genes

The majority of the work in this project was carried out by using the free R software environment (http://www.r-project.org). This software is particularly useful for analysis of microarray data as it has packages for normalisation/standardisation and statistical computing, as well as graphics. R can be used as a powerful standalone programming language, but the most prominent advantages are indeed all the implemented functions that are freely available and ready to use, and which make the R environment both flexible and extendible.

Identification of differentially expressed genes

For the identification of differentially expressed genes, two different methods have mainly been applied, both available in R. These are the Significance Analysis of Microarray Data (SAM) [68] which is included in the Siggenes package (http://www.bioconductor.org), and the Fold Change method (FC). SAM is a statistical method for identification of differentially expressed genes, which controls for the false discovery rate (FDR). Briefly, the algorithm assigns a score to each gene based on differences in expression between conditions, relative to the standard deviation of repeated measurements. The FDR is determined by using permutations of the repeated measurements to estimate the percentage of genes identified by chance. The FC method calculates the ratio between two samples, but provides no statistics regarding the significance of the results. The characteristics of the dataset and the experimental design decide whether SAM or FC is the most appropriate method to use.

Clustering of gene expression data

To reduce the dimensionality and facilitate interpretation of microarray data one can apply different clustering techniques, such as hierarchical clustering, K-means [69], principle component analysis (PCA) [70] or self-organising maps (SOMs) [71], to group transcripts with similar transcriptional profiles. The purpose of clustering is to identify co-regulated and functionally related genes in large datasets. In this project, the agglomerative hierarchical clustering approach has mainly been used, which starts with clusters containing a single item, and iteratively links and merges the two closest clusters together based on a distance measure. After each step, all the distances between the newly formed clusters are re-calculated. The output is a relationship tree (dendrogram) where the branches represent similarity.

Pathway analysis

There are two main approaches for identification of pathways, which are differentially expressed across various experimental conditions. These are Individual Gene Analysis (IGA) methods and Gene Set Analysis (GSA) methods [72]. IGA is the most widely used approach and evaluates the significance of individual genes between two groups of compared samples. Methods using this approach typically yield a list of differentially expressed genes from a cut-off threshold, and evaluate this list for the enrichment of genes participating in specific pathways from a pathway database. A limitation with IGA approaches is that the final result is considerably affected by the selected threshold, which is often arbitrarily chosen. Notably, many genes with moderate, but biologically meaningful, expression differences are discarded by a strict cut-off threshold, which implies a reduction in statistical power. The GSA approach directly scores pre-defined pathways or gene sets based on differential expression, and specifically aims to identify pathways with subtle but coordinated expression changes that cannot be detected by IGA methods [72, 73]. It is based on the principle that even weak expression changes for groups of related genes can have important effects. From a biological perspective, GSA methods are promising because functionally related genes often display coordinated expression [72]. In this thesis, methods that apply the IGA approach have mainly been used, where the lists of differentially expressed genes from various experiments have been analysed for enrichment of genes that are recognised as interacting components in known cellular pathways, represented in the KEGG (http://www.genome.jp/kegg) or the BioCarta (http://www.biocarta.com) pathway databases. Various bioinformatic resources such as WebGestalt [74] and DAVID [75] have been utilised to perform these analyses.

Protein interaction networks

Protein–protein interactions are of central importance for virtually every biological process in a living cell. Typically, signal transduction, where mechanical/chemical stimuli to a cell are converted into specific cellular responses, plays a fundamental role in many biological processes and in many diseases. To investigate the putative interactions among proteins from the significantly up- or downregulated genes identified from a biological experiment, protein interaction networks can be computationally generated by combining the experimental data with information from interaction databases with predicted data.

Several tools to aid derivation of protein interaction networks are available and currently the most comprehensive and freely available one, and the one that has been applied in the analyses described in this thesis, is STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) (http://string-db.org/), which is a database and web resource for experimentally determined and predicted protein-protein interactions [76, 77]. STRING includes both physical and functional interactions, and it weights and integrates information from numerous sources, including experimental repositories, computational prediction methods and public text collections [77]. Thus, STRING is acting as a meta-database that maps all interaction evidence into a common set, which is then graphically visualised in a protein interaction network.

Functional annotation of differentially expressed genes

To further explore the functional properties of a group of differentially expressed genes, one can use information from Gene Ontology (GO) [78] and assess the enrichment of GO annotations (terms describing the genes or gene products). GO consists of three categories of annotation terms, Biological Processes (19,289), Molecular Functions (8,761) and Cellular Components (2,750). The figures in the parentheses represent the number of annotation terms in each category as per Aug 2010. By comparing with a reference group, overrepresentation of annotations among sets of genes can be calculated by dividing the observed number of genes holding a specific annotation with the expected number of genes with that annotation. All genes represented on the arrays are commonly used as the reference group in these calculations. There are many tools available for performing GO annotation enrichment analysis, and in this thesis we have used FatiGO [79], WebGestalt [74], and DAVID [75] to understand more about the biological properties of the differentially expressed genes.

Results in summary

Paper I: Differentiating human embryonic stem cells express a unique housekeeping gene signature

As an initial part of this thesis project, the stability of commonly used endogenous controls, used for normalisation of gene expression levels in various somatic tissues, was investigated. Data from three hESC lines (SA001, SA002, and SA002.5) and two differentiation protocols were generated. Typically, investigators have used the traditional HKGs (e.g., GAPDH, TUBB, ACTB) as controls also in studies of hESCs [37, 80]. However, it is well known that the expression of several of these genes varies considerably in adult tissues, and their suitability as control genes in hESCs had not yet been thoroughly investigated. It had already been shown that the RNA levels of HPRT and β-tubulin varied substantially in differentiating mouse ESCs [81]. This prompted us to investigate the stability of commonly used HKGs in differentiating hESCs. We applied the CodeLink Human Whole Genome Bioarrays (GE Healthcare) to generate global gene expression data from three different cell lines as described in Paper I. We investigated the stability of a group of 56 commonly used HKGs in this novel dataset and notably, only four of these HKGs showed stability in our data. Therefore, a novel set of genes that were stably expressed in this dataset was identified. Based on a stability threshold of coefficient of variation (CV) < 20% we identified 292 putative reference genes in our dataset. The threshold was defined based on results from a technical study [51] where various array platforms, including CodeLink, were compared and evaluated. This novel set of 292 stably expressed genes was further subdivided into three groups; genes with high, medium, and low expression. We also validated our resulting list of stably expressed genes in eight other independent hESC lines from two other studies [21, 82]. Although these data had been generated under considerably different conditions (e.g., different laboratories, cell lines, culture conditions, and array platforms) we observed interesting overlaps with our results. The intersection between all three studies contained a total of six stably expressed genes shown in Table 3 in Paper I. Among these genes were RNF7 and FBXL12, which are both involved in cell-cycle progression and development.

Paper II: Molecular signature of cardiomyocyte clusters derived from human embryonic stem cells

The next step in the project was to perform a detailed characterisation of hESC-derived functional cell types, such as CMs and hepatocytes, and explore the transcriptional program that is activated during differentiation. In Paper II, selected colonies of hESC-derived contracting clusters of CMs were manually dissected, and pooled for subsequent microarray analysis. These samples were compared with samples from UD hESCs. Cell line SA002, which forms cells of the cardiac lineage relatively efficiently, was used for this study. We generated beating CMCs from hESCs, and compared the gene expression profiles of these clusters with profiles of UD hESCs. Using the SAM algorithm [68], we identified 530 genes that were specifically upregulated in the CMCs and 40 genes that were downregulated. Among the upregulated genes, there are several that have been used

before to characterise hESC-derived CMs (e.g., MYH6, MYH7, PLN, TNNT2, NPPA, GATA4, and MEF2C). The functional properties of the upregulated genes in the hESC-derived CMCs were further investigated, using available Gene Ontology (GO) annotations. Among the enriched GO annotations were 'muscle contraction', 'development of mesoderm and muscle', 'cellular differentiation', 'calcium ion binding', and 'tropomyosin binding'. In addition, we identified possible protein interactions among the products of the upregulated genes in the hESC-derived CMCs. Interestingly, substantially more interactions were identified among these gene products compared to randomly generated sets of proteins. Moreover, several induced cellular pathways were identified, that may be important for cardiogenic induction of hESCs as well as for sustaining the CM phenotype. Taken together, these results provide valuable information about the molecular programs that are active in hESC-derived CMCs.

Paper III: Transcriptional profiling of human embryonic stem cells differentiating to definitive and primitive endoderm and further towards the hepatic lineage

Using a similar approach as in Paper II, the transcriptional program that controls the endoderm induction and further differentiation to hepatocyte-like cells was investigated. The endoderm lineage can be subdivided into the DE which further develops into liver, pancreas and lung, and the PrE which develops into the yolk sack, where it forms the placenta. In Paper III, we analysed global gene expression data from DE and PrE differentiation and compared the transcriptional patterns in these two cell lineages with UD cells, as well as with control samples from the well characterised hepatocellular carcinoma cell line HepG2. Three different stem cell lines were included in the study (SA002, SA167, and SA461, Cellartis AB, Göteborg) and for each cell line the experiment was repeated twice. Using two differentiation protocols, the DE-protocol and the intrinsic (I)-protocol that mainly mediates PrE differentiation, we identified differences and similarities between these two endodermal subtypes. We also thoroughly characterised the DE-derivatives, by identifying up- and downregulated genes at each of the three differentiation time points 4 days (DE), 10 days (DE-Prog), and 20 days (DE-Hep). In total, we identified 167, 439, and 921 transcripts which were significantly upregulated in DE, DE-Prog, and DE-Hep, respectively, when compared to UD samples. Interestingly, none of these transcripts were significantly enriched in the PrE derivatives. Well-known markers for DE, such as SOX17, CXCR4, CER1 and GSC, showed a distinct peak of expression in the DE time point in all the three investigated cell lines. AFP was highly expressed in the samples from PrE at 4 days and expressed at low levels in the corresponding DE samples. The opposite pattern was observed for AFP at the 10 day time point, for which the expression drastically had increased in the DE samples and decreased in the corresponding PrE samples. At the final time point, when using the DE differentiation protocol, several genes expressed in mature hepatocytes, e.g., ALB, DPP4, SERPINA7, TF, TM4SF1 and UBD [83-87, 88], showed increased mRNA levels. Notably CD44, known to be expressed in hepatocyte progenitors [87, 89, 90], also showed high expression at 20 days, which indicates that the DE-Hep used in this study have an

immature phenotype, and/or contain a fraction of hepatocyte progenitors. Interestingly, ALB, which is a well-known marker for mature hepatocytes [91-94], showed 3-11 times higher expression in DE-Hep than in the corresponding PrE-derivatives, and the expression of ALB was about 1,000-fold higher in the DE-Hep than in the UD samples. Paper III describes, for the first time, transcriptional differences at the global scale between DE-differentiation and PrE-differentiation. Our results also provide important contributions to the characterisation of hESC-derived hepatocyte-like cells at the global gene expression level.

Paper IV: Expression of microRNAs and their target mRNAs in human stem cell derived cardiomyocyte clusters and in heart tissue

In Paper IV, an additional level of gene regulation was explored, by characterising hESC-derived CMCs with respect to their miRNA expression. Global microarrays were employed to measure the expression of both miRNA and mRNA in parallel in samples of CMC, harvested at two different time points, 3 weeks and 7 weeks after onset of differentiation, as well as in UD cells and in foetal and adult heart tissue samples. The SAM statistical algorithm was applied to identify differentially expressed miRNAs and mRNAs in these datasets, by using a control sample of UD cells. Notably there were more than twice as many up- than downregulated miRNAs in the samples of CMCs, indicating the importance of increased expression of specific miRNAs during cardiac development. Furthermore, we also identified more differentially expressed miRNAs (both up- and downregulated) in the CMC samples than in the foetal and adult heart tissue samples. To define a set of miRNAs of putative importance in cardiac-like cells, differentially expressed miRNAs in samples from CMC and in samples from foetal and adult heart were compared, and an overlap of regulated miRNAs in all four samples was identified. Moreover, possible correlations between differentially expressed miRNAs and mRNAs were investigated, by first conducting computational predictions for the differentially expressed miRNAs, and then determining putative concordance in miRNA expression and mRNA levels of the predicted target genes. Interestingly, a correlation between the global miRNA expression and corresponding target mRNA expression was observed. Using three different sources of cardiac tissue-like samples, a clear similarity between *in vitro* hESC-derived CMCs and their *in vivo* counterparts was identified. To further explore the biology of the predicted target genes of the differentially expressed miRNAs, enrichment of GO annotations was determined and strikingly, several of the over-represented GO terms relate to cardiac function and cardiac development. A number of induced cellular pathways were also identified among the predicted target genes, and several of these have been demonstrated to be important in cardiac development or functions e.g., 'NFAT and Hypertrophy of the heart', 'Wnt signaling pathway' and 'Calcium signaling pathway'. In conclusion, results from Paper IV provide an excellent starting point for further studies regarding the functional properties of the differentially expressed miRNAs in the context of cardiogenesis and regeneration of cardiac tissue.

Discussion and implication of results

This section provides a more general discussion of how the results from this thesis project compare to results from similar studies conducted by other investigators, and possible implications of our results.

The importance of validation of reference genes in human embryonic stem cells and their derivatives (Paper I)

Global gene expression analysis has become a widely used tool for assessing the molecular state of various cells and tissues. Generally, investigators report on genes that are significantly up- and downregulated, in relation to a control or basal state. Much less reported on is the identification of genes that remain constant during different experimental conditions. However, these genes can provide important information on the basal activities and states of the cells. Moreover, stably expressed genes (i.e., HKGs) represent reference genes that can be used for calibration of gene expression data across various samples. However, previous studies have shown that the expression patterns of commonly used reference genes can vary extensively [31, 81, 95, 96]. This suggests that the use of HKGs as reference genes for normalisation without appropriate validation might lead to systematic errors in the calculation of FC in gene expression levels [31]. In Paper I, we analysed the stability of HKGs commonly used as reference genes in somatic cells, and reported on their variability in gene expression data from hESCs and their derivatives. As a result, we proposed a novel set of candidate HKGs that showed stable expression in differentiating hESCs. Notably, despite several reports [31, 81, 95-97] about the variability of common reference genes such as GAPDH and ACTB, these are still frequently used as default reference genes in studies of stem cells and their differentiated derivatives [98, 99]. Typically, these genes are used as calibrators, without proper validation of their stability, which may introduce errors and compromise the interpretation of the results. More research is urgently needed to extend our knowledge about reliable reference genes in different cell types, particularly in stem cells. Importantly, our proposed set of putative HKGs needs further refinement, and the stability of these genes needs to be analysed in additional gene expression datasets from hESCs. Even though relatively few studies have been performed to identify HKGs also in other tissues, one can still conclude that it is unlikely that a standard set of reference genes can be identified, which will show stable expression in all cell types under all experimental conditions. For example human myocardium has been analysed for stably expressed genes and nine putative reference genes have been proposed [100], but none of them were selected as stably expressed in our dataset from hESCs. Ultimately, as more global gene expression datasets are being generated, a set of genes that show stability in a wide range of stem cell lines and differentiated lineages can hopefully be identified. Such a set of genes would be extremely useful as standard reference genes in various stem cell experiments, and would be likely to reduce the risk of introducing systematic biases due to instability of the reference genes.

Considerable overlap of gene expression patterns in hESC-derived cardiomyocyte studies (Paper II)

Despite the substantial progress made by different investigators during recent years, the knowledge of the molecular signature of hESC-derived CMs and the factors that induce cardiogenesis during embryonic development still remains limited. A large proportion of the work in this thesis project has been focused on understanding the expression patterns in hESCs that differentiate towards the CM lineage. In Paper II, the global transcriptional profile of hESC-derived CM clusters was compared to that of hESCs. Up- and downregulated genes were identified and thoroughly analysed. Direct comparisons of results between different microarray studies are sometimes difficult to make, since the experiments performed often have major differences in differentiation models, microarray platforms, cell lines used, and experimental set-ups. This partly may explain the observed problems with poor overlap between published results from different stem cell studies [61-63]. When comparing results from multiple microarray experiments, one should preferably re-analyse the data from the raw data files from each experiment, using a consistent data mining approach for all the datasets. However, as an alternative, we compared published lists of significantly enriched genes from similar studies for overlap of differentially expressed genes during CM differentiation. Interestingly, when placing our results in a wider context and comparing with data from other studies, a substantial overlap was observed. Importantly, in addition to the above mentioned challenges when comparing microarray data from different experiments, the final cell populations that have been analysed in these studies differ in their composition [98] [101]. Nevertheless, in contrast to previous findings [61-63], we identified notable similarities across our data and results from the other three global expression studies that so far have been published on hESC-derived CM-like cells [98, 101, 102]. Importantly, this strengthens the reliability of the microarray technology and verifies that hESC-derived CMs express a uniform transcriptional profile, despite different cell lines and major differences in how these cells are derived. Comparing with results from the study performed by Beqqali et al. [102], where hESC-derived CMs were generated by co-culture with END-2 cells [103], 15 genes were reported as enriched in their hESC-derived CMs and in foetal heart tissue. Notably, eight (53%) of these genes are also upregulated in our hESC-derived CMs (e.g., TNNT2, PLN, and MYL7).

Another study published on hESC-derived CMs [98] report on analyses made on material from hESCs, hESC-derived beating EBs, hESC-derived CMs which were Percoll purified to 40-45% CMs, and purified CMs from foetal heart (FH) tissue samples. Notably, their study focused on transitions from one stage to the next one, and consequently they compared hESCs-EBs, EBs-CMs, and CMs-FH. In our work we compared hESCs with hESC-derived CM clusters and the corresponding direct comparison of CMs and hESCs was not done by Cao et al. [98] which hampers the comparison of our results. Nevertheless, we found that 33% of our upregulated genes in the CM clusters were in their study identified as enriched already at the EB stage. Six of our genes that were enriched in the CM clusters (CLIC5, RUNX1, COL8A1, LONRF2, MSRB3, CAV2) were upregulated in

CMs compared to EBs, and five of our upregulated genes in the CM clusters (EPAS1, ITGB3, PLD1, MSRB3, EMP1) were significantly enriched in FH compared to the CM sample. A similar comparison was made regarding the repressed genes across these two studies and 17 (43%) of the 40 significantly downregulated genes in our data were already repressed at the EB stage in [98] and one gene was among the genes that were significantly downregulated between FH and CMs. Again, no comparison was made between CMs and hESCs [98] regarding downregulated genes, but such a comparison is anticipated to generate a higher overlap with our list of genes that were downregulated in CM clusters.

The most recent work on global gene expression of hESC-derived CMs used a transgenic cell line with a construct comprising the CM-restricted alpha-myosin heavy chain (α-MHC) promoter [101]. They applied antibiotic selection to purify their population of hESC-derived CMs and achieved a 99% pure population [101]. Foetal and adult heart tissue were used as reference samples but notably, these samples were not purified but contained a mixture of the cell types present in heart tissue. Despite substantial differences, such as different cell lines, differentiation protocols, purity of CMs, sampling day etc, a prominent overlap was observed between our data and the data from Xu and colleagues [101]. In total 147 (27%) of our 540 upregulated genes were also identified as significantly upregulated in their population of CMs, when compared to UD and EB samples. Remarkably, 115 (78%) of these 147 genes also show upregulation in the FH and AH samples in data from Xu et al. Strikingly, a subset of 57 genes that show upregulation in our hESC-derived CM clusters is also overlapping with the upregulated genes both in Cao et al. [98] and in Xu et al [101]. All of these 57 genes also show significant upregulation in FH and AH. Furthermore, three (RBM24, TCEA3, and FHOD3) of the four candidate novel cardiac markers, which by Xu and co-workers were validated by *in situ* hybridisation during early mouse development, were indeed significantly upregulated in our study of hESC-derived CM clusters. The fourth one (C15orf52) was not present on the arrays we used in our study. Interestingly, TCEA3 is also among the 57 genes that overlapped across all three studies [98, 101, 104]. Taken together, this suggests that there are substantial similarities between the CM cell populations obtained from hESCs, independent of differentiation protocols and cell lines used.

Transcriptional patterns in hESC-derived hepatocyte-like cells differentiated through definitive endoderm (Paper III)

Differentiated hepatocyte-like cells represent promising tools for target evaluation, studies of metabolism and safety assessment of new drug candidates [7, 105]. They also have the potential to serve as an inexhaustible cell source for hepatocyte transplantation [106]. However, much more research is needed about how to efficiently direct hESCs to DE, hepatic progenitors, and hepatocyte-like cells in order to optimise the properties of the resulting functional cells. Even though several investigators have reported on the capacity of hESCs to differentiate into various specific cell types, the generation of fully functional hepatocytes from hESCs has proven to be particularly challenging. The first report on

differentiation of hESC to hepatocytes-like cells was published in 2003 [107], but even if these cells expressed genes such as ALB and AFP they were of PrE, which lack expression of specific DE markers. It was not until 2005 that protocols for DE differentiation were published [108]. Shortly after, investigators also started to report on further differentiation of hESCs into hepatocyte-like cells [83, 85, 86, 93, 109, 110] with various efficiencies. However, global transcriptional studies on hESC-derived hepatocytes are still rare, most likely due to the difficulty of developing efficient differentiation protocols. To the best of our knowledge, besides our Paper III, global transcriptional profiling of hESC-derived hepatocytes has only been reported once [84], in which AFP^+ cells were selected for gene expression analysis. However, the data analysis approach used in that study differs from ours, since instead of generating lists of up- vs. downregulated genes during hepatocyte differentiation, they applied Gene Set Enrichment Analysis [73] (described above) to identify affected sets of genes during hepatic specification, e.g., molecular pathways etc. An overlap comparison of lists of enriched genes in different stages is therefore not possible. However, when investigating the expression of the 30 marker genes, which specifically were reported as enriched in their AFP^+ population, we observed important similarities with our results. Examples of genes that are enriched in the early DE stage in both studies are typical DE-markers such as SOX17, FOXA2 and MIXL1. Moreover, AFP, which is expressed during hepatocyte specification but also at very early stages of PrE, is upregulated accordingly in our data. Additionally, CDH17 and KRT7, which are reported as expressed during hepatocyte specification [84] are induced in our hepatocyte-like cells, which might indicate that this population also contains more immature cells. Also other genes such as ALB, NTN4, MET, and CEBPA, which are known to be expressed in mature hepatocytes, show induced expression patterns in both studies.

Of special note from our results are the less reported genes TM4SF1 and UBD, which demonstrate highly interesting expression in our experiments. Their expression patterns were consistent across all three cell lines, indicating a putative importance of controlled expression during hepatocyte differentiation. Both these genes demonstrate increasing expression levels during the differentiation towards hepatocytes, with the peak expression in the most mature samples. Interestingly, and consistent with our observations, these genes have also previously been reported as reliable hepatocyte markers [88]. Another observation from our results is that the global transcriptional activity increases dramatically as the cells differentiate, with a larger number of differentially expressed genes in the more mature samples. In Paper III, our investigations focus on upregulated genes, but downregulation may also be equally important for stem cell differentiation and needs further investigation.

Despite the challenges of efficiently generating hepatocytes from hESCs, important steps towards cell-based therapeutics and toxicology testing using hESC-derived hepatocytes have recently been made. Reports have been published that address major obstacles in this field, describing significant improvements of the differentiation protocols. Several of

these reports have emphasised the importance of Activin A for differentiation of DE [83, 93, 108, 109], and Hay et al. [93] also reported on the importance of Wnt3a signalling for generation of functional hepatic endoderm. To be able to further optimise the culturing conditions for induction of hepatic differentiation of hESCs, the effects of various extracellular matrices (ECMs) and growth factors need to be better understood. Accordingly, Ishii and colleagues [111] studied the effect of different culturing conditions and concluded that adding Activin A and HGF to the medium for cells cultured on Matrigel was useful conditions for efficient differentiation of hESCs to DE and further on towards the hepatic lineage.

One important application for hESC-derived hepatocytes is to provide a tool for pharmacology and toxicology studies. Though, this requires that the cells demonstrate metabolic functionality similar to what is observed in primary hepatocytes. Although some functionality is still missing, significant progress of the metabolic activities in hESC-derived hepatocytes was recently reported [112]. It may be questioned if complete hepatocytes maturation is feasible in the simple 2D culturing systems that are used in most of the current published reports. Perhaps more sophisticated 3D systems are required to increase the enzymatic activity and achieve metabolic systems comparable to primary hepatocytes [105]. Therefore, extensive studies of more advanced systems are on-going in order to improve hepatic culture conditions [113]. The metabolic functions of the differentiated cells need to be further improved to fulfil the requirement as a powerful tool for safety testing and drug development. Another issue for future clinical applications is the need of xeno-free culturing systems. Despite the fact that diverse culturing systems for generation of hESC-derived hepatocytes have been published, most of these differentiation approaches are still based on culture media containing serum, complex matrixes, or MEF, which is not compatible with the long term goal of producing hepatocytes for future clinical applications. In this regard, Touboul and co-workers [114] have reported on the generation of functional hepatocytes from hESCs using fully defined culturing conditions.

Despite important improvements of differentiation protocols, more studies addressing global transcriptional profiling of hESC-derived hepatocytes are needed and could contribute to unravel critical regulatory mechanisms that may be important for achieving efficient hepatocyte differentiation. The regulatory effects of miRNA expression during hepatic differentiation also need to be investigated, and interestingly, a distinct global miRNA expression profile has recently been reported in definitive endoderm [115].

MicroRNAs as important regulators in lineage specification and during cardiomyocyte differentiation (Paper IV)

MicroRNAs are small non-coding molecules, which have been shown to be involved in cell fate decisions of pluripotent ESCs, by controlling the activation and repression of lineage-specific genes. As a result, miRNAs may provide new means of altering stem cell

fate and differentiation processes. These tiny molecules act as post-transcriptional regulators and notably, several miRNAs have been shown to play key roles during heart development and in cardiac function [39, 42, 116-119]. It was recently demonstrated that miR-1 reinforces the expression of one of the earliest cardiac markers, NKX2.5, in both murine and human ESC lines and that it increases the fraction of contracting CMs compared to control samples [41]. Another group of miRNAs that are expressed during the stem cell state and progressively declines during differentiation are the nearly identical miRNAs miR-302a-d, collectively referred to as miR-302 [120]. By controlling the germ layer specification and promoting mesendodermal fate specification while inhibiting neuroectoderm formation, it is suggested that miR-302 has a crucial role in embryogenesis [120]. In Paper IV, we have analysed hESC-derived CMs and heart tissue samples with regard to their miRNA expression, and identified significantly up- and downregulated miRNAs. In line with previous reports, our data confirms that several variants of miR-302 are highly expressed in UD cells, and a substantial downregulation is also observed in differentiated progenies as well as in foetal and adult heart tissue. Moreover, we identified miR-208a/b and miR-499 as significantly induced in all cardiac-like samples. Interestingly, these miRNAs have also recently been reported as enriched in cardiac tissue by others [121-123]. Both miR-499 and miR-1 have been suggested to regulate the proliferation of human CM progenitors and their further differentiation into CMs [123]. In addition, miR-499 has also been proposed as a marker for acute myocardial infarction in humans [121]. Moreover, miR-208 has been suggested as a marker for myocardial injury in rat [122], and as a regulator of cardiac hypertrophy in mice [124].

Together with our results, these reports emphasise the importance of miRNAs for cardiac development and as potentially useful markers in clinical applications. In the future, some of these miRNAs may also serve as prospective drug targets in various cardiac injuries. There is accumulating evidence supporting the importance of miRNAs for hESC self-renewal, pluripotency, and differentiation. Determining miRNAs that are associated with re-programming will yield significant insight into the specific miRNA expression patterns that are required for pluripotency. To further investigate which miRNAs that are associated with re-programming, investigators have now started to characterise the miRNA expression during the re-programming of iPS cells [125]. Interestingly, results show that miR-302, which is significantly enriched in our UD samples, is upregulated in both hESCs and in iPS cells [125]. Even though much effort has been focused on identifying miRNAs that are important during cardiac development, information about the correlation between global miRNA and mRNA expression is lacking. In Paper IV, we partly addressed this issue and investigated in parallel the miRNA and mRNA expression in CMCs derived from hESCs and in foetal and adult heart tissues, and observed a correlation between the differentially expressed miRNAs and their putative target genes. As expected, a negative correlation between miRNA expression and mRNA expression was dominating, although a positive correlation was observed for a fraction of the predicted target genes. A possible explanation for this may be an indirect regulation of transcription factors.

Limitations of this work

Although much effort has been invested to design robust experimental set-ups and produce extensive datasets that are then rigorously analysed, there are limitations in this thesis work that should be pointed out.

The different cell populations analysed in this thesis project were isolated and prepared without applying specific purification steps, and therefore may contain a mixture of cells and not only the specific cell type of interest. There is therefore an inherent risk that the observed expression patterns origin from other cells of unknown origin that may be present in the sampled populations. There is also a risk that signals from low expressed genes are buried in the random noise present in heterogeneous populations. To account for this, e.g., in the experiments on the hESC-derived CMCs, specific care was taken to only harvest the beating areas with a minimum of surrounding non-contracting cells.

Because of the high costs associated with microarrays, a general limitation in most global gene expression experiments is that relatively few repetitions are conducted. This is also a weakness in the experiments described in this thesis, where a limited number of replicates have been carried out, partly due to the high costs and partly due to limited amounts of available cell material. As compensation, the replicated arrays within the experimental groups have been carefully compared to assess low variability. Also, despite the low numbers of replicates, we have whenever possible used statistical methods with p-value < 0.05 for selection of differentially expressed genes.

As this work has been carried out in collaboration with Cellartis AB, a biotech company that is specialised in stem cell technology, cell lines derived at their lab have exclusively been used in the experiments. It would have been advantageous to also include cell lines from other labs in the experiments. However, we have instead carefully compared our results with similar studies performed on other cell lines, and investigated the overlap between their results and ours.

In Paper III, we used FC as a method for selection of differentially expressed genes, and therefore these results lack statistics such as p-values. The reason for using FC instead of e.g., SAM was that large variations in magnitude of regulation within the groups of up- and downregulated genes were observed, meaning that some genes had a FC>3 in one cell line while the same group of genes had a FC>10 in another cell line. Genes with high variability within the groups of samples are not selected as significant by SAM. However, we still believe that genes with a consistent pattern of up- or downregulation are interesting to report, even though the magnitude of regulation varies between the cell lines. Observed variations were most pronounced at the first differentiation time point, DE04, indicating that the different cell lines are less synchronised during the early stages of DE differentiation. However, for the last time point, DE20, this type of variation flattens out and consequently, the SAM and FC methods yield highly overlapping results.

Induced pluripotent stem cells and future perspectives

The most recent progress in the field of human pluripotent stem cells is the successful re-programming of differentiated somatic cells into induced pluripotent (iPS) cells [3]. Human iPS-cells are generated from somatic cells by over-expression of specific factors and these cells share many characteristics with hESCs, including multi-lineage differentiation potential and infinitive proliferation capabilities *in vitro*. By using this technique, genetically identical somatic cells can be generated as model systems for basic studies of human disease [3]. The differentiation potential of human iPS-cells has already been demonstrated by several investigators with successful differentiation of iPS-cells to e.g., neurons [126], CMs [127, 128] and hepatocytes [129, 130, 131] representing cell types from the ectoderm, mesoderm and endoderm germ layers respectively. However, extensive characterization of the iPS-cells and their differentiated progenies at the transcriptional level is still limited, but is nevertheless necessary to be able to assess their similarity to hESCs.

With regard to cardiac differentiation, iPS-derived CMs have been evaluated in terms of differentiation efficiency, contraction rates, CM marker genes, proliferation, and electrophysiology. [127, 128]. The results suggest that iPS cells are a viable alternative to ESCs for a variety of research applications and importantly, also, potentially, as an autologous cell source for cardiac repair. Successful differentiation of iPS-cells into hepatocyte-like cells has also been shown by Song and co-workers [129]. This pioneering study reported that iPS-derived hepatocyte-like cells possessed similar functionality to hESC-derived hepatocytes. Even though much more research is needed, e.g., regarding the effect of the transduction on the iPS cells, the progress recently reported provides important steps towards cell-based therapies. In the future, iPS-derivatives hold great potential in clinical applications as this technique avoid problematic immune system reactions, which is an issue not only in cell-based therapies but also in organ transplantations. Rapid progress of further development of iPS-cells is expected in the very near future.

Concluding remarks

Human ESCs have a tremendous potential in many different applications such as drug development, regenerative medicine, and as a model system in basic research. However, to fully realise the potential of these cells we need to better understand the regulatory mechanisms that control the differentiation of hESCs into various functional cell types. The aim of this thesis was to increase the understanding of the transcriptional programs that are active during stem cell differentiation, with particular focus on CM and hepatocyte differentiation. To meet this aim, global transcriptional profiling of hESCs and differentiated progenies of the cardiac and hepatic lineages was performed, to generate comprehensive datasets, especially useful for characterisation purposes, identification of differentially expressed genes, and investigations of gene regulation. When analysing gene expression data it is common to use endogenous controls, to calibrate and normalise the data. Therefore, as an initial step, the stability of commonly used reference genes in hESCs and their derivatives was investigated, to evaluate their usefulness as calibrators in subsequent studies. The results from that work demonstrated unacceptably high variability for most of the commonly used reference genes, which prompted us to identify a novel set of HKGs that are more reliable to use as reference genes in data from hESCs. Microarrays have previously been demonstrated a useful tool for identification and evaluation of HKGs, and many adult and foetal tissues have been analysed for identification of stably expressed genes [31, 132-134], but none of these studies have included hESCs.

In the next two parts of the project we analysed global gene expression patterns in hESC-derived CMs and hepatocytes, and identified up- and downregulated genes in different stages during the differentiation processes. Interesting transcriptional patterns were revealed and our results both confirmed the expression of known marker genes, as well as identified large sets of novel genes that demonstrated significant up- or downregulation in specific stages during the differentiation. The expected expression patterns of known marker genes confirm the efficiency of the differentiation protocols. Nevertheless, the most interesting results are the sets of novel genes that have not previously been associated with differentiation or developmental processes in these particular cell types. It is hypothesised that many of these genes have the potential to serve as novel markers and provide important information for optimisation of the differentiation protocols, and further validation and investigation of their functional properties is therefore suggested as future work.

To further investigate possible regulatory mechanisms that may control gene expression in hESCs and their differentiated progenies, the final part of the thesis project focused on investigation of miRNA expression in hESCs that differentiate towards the CM lineage. By exploring the miRNA and mRNA expression in parallel, interesting correlations between these interacting molecules were revealed. Significantly differentially expressed

miRNAs in hESC-derived CM clusters were identified, providing a novel level of characterisation of this cell population. The results show that specific miRNAs may serve as important complementary markers for this specific cell type. A similar characterisation of the hESC-derived hepatocytes would likely reveal interesting miRNA patterns also in this cell type and is therefore also suggested as future work. Taken together, this thesis utilise data from microarray experiments from hESCs and their differentiated progenies, and provides a transcriptional characterisation of various stages during the differentiation processes. The results presented here increase our knowledge of the global transcription machinery that is activated during differentiation, and provide a foundation for further dissection of the molecular mechanisms that drives the heart and liver specification from human pluripotent stem cells.

References

1. Stem Cells: Scientific Progress and Future Research Directions: National Institute of Health, http://stemcells.nih.gov/info/2001report/2001report.htm 2001.
2. Thomson JA, Itskovitz-Eldor J, Shapiro SS, et al. Embryonic stem cell lines derived from human blastocysts. *Science.* 1998;282:1145-1147.
3. Takahashi K, Tanabe K, Ohnuki M, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell.* 2007;131:861-872.
4. McKay R. Stem cells--hype and hope. *Nature.* 2000;406:361-364.
5. McNeish J. Embryonic stem cells in drug discovery. *Nat Rev Drug Discov.* 2004;3:70-80.
6. Sartipy P, Björquist P, Strehl R, et al. Pluripotent human stem cells as novel tools in drug discovery and toxicity testing. *IDrugs.* 2006;9:702-705.
7. Sartipy P, Björquist P, Strehl R, et al. The application of human embryonic stem cell technologies to drug discovery. *Drug Discov Today.* 2007;12:688-699.
8. Reubinoff BE, Pera MF, Fong CY, et al. Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro. *Nat Biotechnol.* 2000;18:399-404.
9. Richards M, Fong CY, Chan WK, et al. Human feeders support prolonged undifferentiated growth of human inner cell masses and embryonic stem cells. *Nat Biotechnol.* 2002;20:933-936.
10. Xu C, Inokuma MS, Denham J, et al. Feeder-free growth of undifferentiated human embryonic stem cells. *Nat Biotechnol.* 2001;19:971-974.
11. Nagaoka M, Si-Tayeb K, Akaike T, et al. Culture of human pluripotent stem cells using completely defined conditions on a recombinant E-cadherin substratum. *BMC Dev Biol.* 2010;10:60.
12. Rodin S, Domogatskaya A, Ström S, et al. Long-term self-renewal of human pluripotent stem cells on human recombinant laminin-511. *Nat Biotechnol.* 2010; 28:611-615.
13. Villa-Diaz LG, Nandivada H, Ding J, et al. Synthetic polymer coatings for long-term growth of human embryonic stem cells. *Nat Biotechnol.* 2010;28:581-583.
14. Melkoumian Z, Weber JL, Weber DM, et al. Synthetic peptide-acrylate surfaces for long-term self-renewal and cardiomyocyte differentiation of human embryonic stem cells. *Nat Biotechnol.* 2010;28:606-610.
15. Keller G. Embryonic stem cell differentiation: emergence of a new era in biology and medicine. *Genes Dev.* 2005;19:1129-1155.
16. Beqqali A, van Eldik W, Mummery C, et al. Human stem cells as a model for cardiac differentiation and disease. Cell Mol Life Sci. 2009;66:800-813.
17. Heins N, Englund MC, Sjöblom C, et al. Derivation, characterization, and differentiation of human embryonic stem cells. *Stem Cells.* 2004;22:367-376.
18. Draper JS, Smith K, Gokhale P, et al. Recurrent gain of chromosomes 17q and 12 in cultured human embryonic stem cells. *Nat Biotechnol.* 2004;22:53-54.
19. Itskovitz-Eldor J, Schuldiner M, Karsenti D, et al. Differentiation of human embryonic stem cells into embryoid bodies compromising the three embryonic germ layers. *Mol Med.* 2000;6:88-95.
20. Kehat I, Kenyagin-Karsenti D, Snir M, et al. Human embryonic stem cells can differentiate into myocytes with structural and functional properties of cardiomyocytes. *J Clin Invest.* 2001;108:407-414.
21. Xu RH, Chen X, Li DS, et al. BMP4 initiates human embryonic stem cell differentiation to trophoblast. *Nat Biotechnol.* 2002;20:1261-1264.

22. Abeyta MJ, Clark AT, Rodriguez RT, et al. Unique gene expression signatures of independently-derived human embryonic stem cell lines. *Hum Mol Genet.* 2004;13:601-608.
23. Bhattacharya B, Miura T, Brandenberger R, et al. Gene expression in human embryonic stem cell lines: unique molecular signature. *Blood.* 2004;103:2956-2964.
24. Boyer LA, Lee TI, Cole MF, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell.* 2005;122:947-956.
25. Brandenberger R, Khrebtukova I, Thies RS, et al. MPSS profiling of human embryonic stem cells. *BMC Dev Biol.* 2004;4:10.
26. Sato N, Sanjuan IM, Heke M, et al. Molecular signature of human embryonic stem cells and its comparison with the mouse. *Dev Biol.* 2003;260:404-413.
27. Sperger JM, Chen X, Draper JS, et al. Gene expression patterns in human embryonic stem cells and human pluripotent germ cell tumors. *Proc Natl Acad Sci U S A.* 2003;100:13350-13355.
28. Alberts B, Bray D, Hopkin K, et al. *Essential Cell Biology*: Garland Science, Taylor & Francis Group, New York; 2004.
29. Watson JD, Hopkins, N. H. et al. . *Molecular Biology of the Gene* Menlo Park CA: Benjamin-Cummins; 1987.
30. Eisenberg E, Levanon EY. Human housekeeping genes are compact. *Trends Genet.* 2003;19:362-365.
31. Lee PD, Sladek R, Greenwood CM, et al. Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Res.* 2002;12:292-297.
32. Speed T. *Statistical analysis of gene expression microarray data.* Boca Raton: Chapman & Hall/CRC CRC press LRC; 2003.
33. Yang YH, Dudoit S, Luu P, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 2002;30:e15.
34. Abruzzo LV, Lee KY, Fuller A, et al. Validation of oligonucleotide microarray data using microfluidic low-density arrays: a new statistical method to normalize real-time RT-PCR data. *Biotechniques.* 2005;38:785-792.
35. Hoerndli FJ, Toigo M, Schild A, et al. Reference genes identified in SH-SY5Y cells using custom-made gene arrays with validation by quantitative polymerase chain reaction. *Anal Biochem.* 2004;335:30-41.
36. Cai J, Chen J, Liu Y, et al. Assessing self-renewal and differentiation in hESC lines. *Stem Cells.* 2005.
37. Yang AX, Mejido J, Luo Y, et al. Development of a focused microarray to assess human embryonic stem cell differentiation. *Stem Cells Dev.* 2005;14:270-284.
38. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell.* 2004;116:281-297.
39. Sartipy P, Olsson B, Hyllner J, et al. Regulation of 'stemness' and stem cell differentiation by microRNAs. *IDrugs.* 2009;12:492-496.
40. Wang Y, Keys DN, Au-Young JK, et al. MicroRNAs in embryonic stem cells. *J Cell Physiol.* 2009;218:251-255.
41. Ivey KN, Muth A, Arnold J, et al. MicroRNA regulation of cell lineages in mouse and human embryonic stem cells. *Cell Stem Cell.* 2008;2:219-229.
42. van Rooij E, Olson EN. MicroRNAs: powerful new regulators of heart disease and provocative therapeutic targets. *J Clin Invest.* 2007;117:2369-2376.

43. Lu J, Getz G, Miska EA, et al. MicroRNA expression profiles classify human cancers. *Nature.* 2005;435:834-838.
44. Zhu S, Wu H, Wu F, et al. MicroRNA-21 targets tumor suppressor genes in invasion and metastasis. *Cell Res.* 2008;18:350-359.
45. Gregory PA, Bert AG, Paterson EL, et al. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat Cell Biol.* 2008;10:593-601.
46. Vasudevan S, Tong Y, Steitz JA. Switching from repression to activation: microRNAs can up-regulate translation. *Science.* 2007;318:1931-1934.
47. Zhou Y, Ferguson J, Chang JT, et al. Inter- and intra-combinatorial regulation by transcription factors and microRNAs. *BMC Genomics.* 2007;8:396.
48. Forster T, Roy D, Ghazal P. Experiments using microarray technology: limitations and standard operating procedures. *J Endocrinol.* 2003;178:195-204.
49. Ramakrishnan R, Dorris D, Lublinsky A, et al. An assessment of Motorola CodeLink microarray performance for gene expression profiling applications. *Nucleic Acids Res.* 2002;30:e30.
50. Weckx S, Carlon E, DeVuyst L, et al. Thermodynamic behavior of short oligonucleotides in microarray hybridizations can be described using Gibbs free energy in a nearest-neighbor model. *J Phys Chem B.* 2007;111:13583-13590.
51. Shippy R, Sendera TJ, Lockner R, et al. Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations. *BMC Genomics.* 2004;5:61.
52. Miller RM, Callahan LM, Casaceli C, et al. Dysregulation of gene expression in the 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine-lesioned mouse substantia nigra. *J Neurosci.* 2004;24:7445-7454.
53. Pease AC, Solas D, Sullivan EJ, et al. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci U S A.* 1994;91:5022-5026.
54. Han ES, Wu Y, McCarter R, et al. Reproducibility, sources of variability, pooling, and sample size: important considerations for the design of high-density oligonucleotide array experiments. *J Gerontol A Biol Sci Med Sci.* 2004;59:306-315.
55. Pradervand S, Paillusson A, Thomas J, et al. Affymetrix Whole-Transcript Human Gene 1.0 ST array is highly concordant with standard 3' expression arrays. *Biotechniques.* 2008;44:759-762.
56. Tan PK, Downey TJ, Spitznagel EL, Jr., et al. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* 2003;31:5676-5684.
57. Kuo WP, Jenssen TK, Butte AJ, et al. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics.* 2002;18:405-412.
58. Draghici S, Khatri P, Eklund AC, et al. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.* 2006;22:101-109.
59. Chen JJ, Hsueh HM, Delongchamp RR, et al. Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. *BMC Bioinformatics.* 2007;8:412.
60. Shi L, Reid LH, Jones WD, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol.* 2006;24:1151-1161.
61. Ivanova NB, Dimos JT, Schaniel C, et al. A stem cell molecular signature. *Science.* 2002;298:601-604.

62. Ramalho-Santos M, Yoon S, Matsuzaki Y, et al. "Stemness": transcriptional profiling of embryonic and adult stem cells. *Science.* 2002;298:597-600.
63. Fortunel NO, Otu HH, Ng HH, et al. Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature". *Science.* 2003;302:393.
64. Wang Y, Barbacioru C, Hyland F, et al. Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays. *BMC Genomics.* 2006;7:59.
65. Yuen T, Wurmbach E, Pfeffer RL, et al. Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.* 2002;30:e48.
66. Frantz S. An array of problems. Nat Rev Drug Discov. 2005;4:362-363.
67. Cohen J. Bioinformatics—An Introduction for Computer Scientists. *ACM Computing Surveys.* 2004;36:122-158.
68. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A.* 2001;98:5116-5121.
69. Świniarski R, Cios KJ, Pedrycz W. *Data mining methods for knowledge discovery*: Kluwer Academic; 1998.
70. Jolliffe IT. *Principal Component Analysis*: Springer-Verlag; 1986.
71. Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A.* 1999;96:2907-2912.
72. Nam D, Kim SY. Gene-set approach for expression pattern analysis. *Brief Bioinform.* 2008;9:189-197.
73. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet.* 2003;34:267-273.
74. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.* 2005;33:W741-748.
75. Huang da W, Sherman BT, Tan Q, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* 2007;35:W169-175.
76. von Mering C, Jensen LJ, Kuhn M, et al. STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* 2007;35:D358-362.
77. Jensen LJ, Kuhn M, Stark M, et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* 2009;37:D412-416.
78. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000;25:25-29.
79. Al-Shahrour F, Diaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics.* 2004;20:578-580.
80. Cai J, Chen J, Liu Y, et al. Assessing self-renewal and differentiation in human embryonic stem cell lines. *Stem Cells.* 2006;24:516-530.
81. Murphy CL, Polak JM. Differentiating embryonic stem cells: GAPDH, but neither HPRT nor beta-tubulin is suitable as an internal standard for measuring RNA levels. *Tissue Eng.* 2002;8:551-559.
82. Skottman H, Mikkola M, Lundin K, et al. Gene expression signatures of seven individual human embryonic stem cell lines. *Stem Cells.* 2005;23:1343-1356.

83. Agarwal S, Holton KL, Lanza R. Efficient differentiation of functional hepatocytes from human embryonic stem cells. *Stem Cells.* 2008;26:1117-1127.
84. Chiao E, Elazar M, Xing Y, et al. Isolation and transcriptional profiling of purified hepatic cells derived from human embryonic stem cells. *Stem Cells.* 2008;26:2032-2041.
85. Duan Y, Catana A, Meng Y, et al. Differentiation and enrichment of hepatocyte-like cells from human embryonic stem cells in vitro and in vivo. *Stem Cells.* 2007;25:3058-3068.
86. Hay DC, Zhao D, Ross A, et al. Direct differentiation of human embryonic stem cells to hepatocyte-like cells exhibiting functional activities. *Cloning Stem Cells.* 2007;9:51-62.
87. Kon J, Ooe H, Oshima H, et al. Expression of CD44 in rat hepatic progenitor cells. J *Hepatol.* 2006;45:90-98.
88. D´Amour K, Baetge, Emmanuel. Hepatocytes lineage cells. Vol U.S.Patent # WO2007/127454 A2. U.S.A; 2007.
89. Dan YY, Riehle KJ, Lazaro C, et al. Isolation of multipotent progenitor cells from human fetal liver capable of differentiating into liver and mesenchymal lineages. *Proc Natl Acad Sci U S A.* 2006;103:9912-9917.
90. Schmelzer E, Zhang L, Bruce A, et al. Human hepatic stem cells from fetal and postnatal donors. *J Exp Med.* 2007;204:1973-1987.
91. Duncan SA. Mechanisms controlling early development of the liver. *Mech Dev.* 2003;120:19-33.
92. Gouon-Evans V, Boussemart L, Gadue P, et al. BMP-4 is required for hepatic specification of mouse embryonic stem cell-derived definitive endoderm. *Nat Biotechnol.* 2006;24:1402-1411.
93. Hay DC, Fletcher J, Payne C, et al. Highly efficient differentiation of hESCs to functional hepatic endoderm requires ActivinA and Wnt3a signaling. *Proc Natl Acad Sci U S A.* 2008;105:12301-12306.
94. Lavon N, Benvenisty N. Study of hepatocyte differentiation using embryonic stem cells. *J Cell Biochem.* 2005;96:1193-1202.
95. Suzuki T, Higgins PJ, Crawford DR. Control selection for RNA quantitation. *Biotechniques.* 2000;29:332-337.
96. Wu YY, Rees JL. Variation in epidermal housekeeping gene expression in different pathological states. *Acta Derm Venereol.* 2000;80:2-3.
97. Synnergren J, Giesler TL, Adak S, et al. Differentiating human embryonic stem cells express a unique housekeeping gene signature. *Stem Cells.* 2007;25:473-480.
98. Cao F, Wagner RA, Wilson KD, et al. Transcriptional and functional profiling of human embryonic stem cell-derived cardiomyocytes. *PLoS ONE.* 2008;3:e3474.
99. Stock P, Staege MS, Muller LP, et al. Hepatocytes derived from adult stem cells. *Transplant Proc.* 2008;40:620-623.
100. Pilbrow AP, Ellmers LJ, Black MA, et al. Genomic selection of reference genes for real-time PCR in human myocardium. *BMC Med Genomics.* 2008;1:64.
101. Xu XQ, Soo SY, Sun W, et al. Global expression profile of highly enriched cardiomyocytes derived from human embryonic stem cells. *Stem Cells.* 2009;27:2163-2174.
102. Beqqali A, Kloots J, Ward-van Oostwaard D, et al. Genome-wide transcriptional profiling of human embryonic stem cells differentiating to cardiomyocytes. *Stem Cells.* 2006;24:1956-1967.
103. Passier R, Oostwaard DW, Snapper J, et al. Increased cardiomyocyte differentiation from human embryonic stem cells in serum-free cultures. *Stem Cells.* 2005;23:772-780.

104. Synnergren J, Åkesson K, Dahlenborg K, et al. Molecular signature of cardiomyocyte clusters derived from human embryonic stem cells. *Stem Cells.* 2008;26:1831-1840.
105. Jensen J, Hyllner J, Björquist P. Human embryonic stem cell technologies and drug discovery. *J Cell Physiol.* 2009;219:513-519.
106. Haridass D, Narain N, Ott M. Hepatocyte transplantation: waiting for stem cells. *Curr Opin Organ Transplant.* 2008;13:627-632.
107. Rambhatla L, Chiu CP, Kundu P, et al. Generation of hepatocyte-like cells from human embryonic stem cells. *Cell Transplant.* 2003;12:1-11.
108. D'Amour KA, Agulnick AD, Eliazer S, et al. Efficient differentiation of human embryonic stem cells to definitive endoderm. *Nat Biotechnol.* 2005;23:1534-1541.
109. Cai J, Zhao Y, Liu Y, et al. Directed differentiation of human embryonic stem cells into functional hepatic cells. *Hepatology.* 2007;45:1229-1239.
110. Hay DC, Zhao D, Fletcher J, et al. Efficient differentiation of hepatocytes from human embryonic stem cells exhibiting markers recapitulating liver development in vivo. *Stem Cells.* 2008;26:894-902.
111. Ishii T, Fukumitsu K, Yasuchika K, et al. Effects of extracellular matrixes and growth factors on the hepatic differentiation of human embryonic stem cells. *Am J Physiol Gastrointest Liver Physiol.* 2008;295:G313-321.
112. Duan Y, Ma X, Zou W, et al. Differentiation and characterization of metabolically functioning hepatocytes from human embryonic stem cells. *Stem Cells.*28:674-686.
113. Hewitt NJ, Lechon MJ, Houston JB, et al. Primary hepatocytes: current understanding of the regulation of metabolic enzymes and transporter proteins, and pharmaceutical practice for the use of hepatocytes in metabolism, enzyme induction, transporter, clearance, and hepatotoxicity studies. *Drug Metab Rev.* 2007;39:159-234.
114. Touboul T, Hannan NR, Corbineau S, et al. Generation of functional hepatocytes from human embryonic stem cells under chemically defined conditions that recapitulate liver development. *Hepatology.* 2010;51:1754-1765.
115. Hinton A, Afrikanova I, Wilson M, et al. A distinct microRNA signature for definitive endoderm derived from human embryonic stem cells. *Stem Cells Dev.* 2010;19:797-807.
116. Divakaran V, Mann DL. The emerging role of microRNAs in cardiac remodeling and heart failure. *Circ Res.* 2008;103:1072-1083.
117. Thum T, Catalucci D, Bauersachs J. MicroRNAs: novel regulators in cardiac development and disease. *Cardiovasc Res.* 2008;79:562-570.
118. Zhang C. MicroRNAs: role in cardiovascular biology and disease. *Clin Sci (Lond).* 2008;114:699-706.
119. Cordes KR, Srivastava D. MicroRNA regulation of cardiovascular development. *Circ Res.* 2009;104:724-732.
120. Rosa A, Spagnoli FM, Brivanlou AH. The miR-430/427/302 family controls mesendodermal fate specification via species-specific target selection. *Dev Cell.* 2009;16:517-527.
121. Adachi T, Nakanishi M, Otsuka Y, et al. Plasma microRNA 499 as a biomarker of acute myocardial infarction. *Clin Chem.* 2010;56:1183-1185.
122. Ji X, Takahashi R, Hiura Y, et al. Plasma miR-208 as a biomarker of myocardial injury. *Clin Chem.* 2009;55:1944-1949.
123. Sluijter JP, van Mil A, van Vliet P, et al. MicroRNA-1 and -499 regulate differentiation and proliferation in human-derived cardiomyocyte progenitor cells. *Arterioscler Thromb Vasc Biol.* 2010;30:859-868.

124. Callis TE, Pandya K, Seok HY, et al. MicroRNA-208a is a regulator of cardiac hypertrophy and conduction in mice. *J Clin Invest.* 2009;119:2772-2786.
125. Wilson KD, Venkatasubrahmanyam S, Jia F, et al. MicroRNA profiling of human-induced pluripotent stem cells. *Stem Cells Dev.* 2009;18:749-758.
126. Swistowski A, Peng J, Liu Q, et al. Efficient Generation of Functional Dopaminergic Neurons from Human Induced pluripotent Stem Cells under Defined Conditions. *Stem Cells.* 2010 Aug 16. [Epub ahead of print]
127. Gai H, Leung EL, Costantino PD, et al. Generation and characterization of functional cardiomyocytes using induced pluripotent stem cells derived from human fibroblasts. *Cell Biol Int.* 2009;33:1184-1193.
128. Zhang J, Wilson GF, Soerens AG, et al. Functional cardiomyocytes derived from human induced pluripotent stem cells. *Circ Res.* 2009;104:e30-41.
129. Song Z, Cai J, Liu Y, et al. Efficient generation of hepatocyte-like cells from human induced pluripotent stem cells. *Cell Res.* 2009;19:1233-1242.
130. Si-Tayeb K, Noto FK, Nagaoka M, et al. Highly efficient generation of human hepatocyte-like cells from induced pluripotent stem cells. *Hepatology.* 2010;51:297-305.
131. Sullivan GJ, Hay DC, Park IH, et al. Generation of functional human hepatic endoderm from human induced pluripotent stem cells. *Hepatology.* 2010;51:329-335.
132. Buckingham M, Meilhac S, Zaffran S. Building the mammalian heart from two sources of myocardial cells. *Nature reviews.* 2005;6:826-835.
133. Hsiao LL, Dangond F, Yoshida T, et al. A compendium of gene expression in normal human tissues. *Physiol Genomics.* 2001;7:97-104.
134. Warrington JA, Nair A, Mahadevappa M, et al. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol Genomics.* 2000;2:143-147.

Acknowledgement

I wish to express my sincere gratitude to all my friends and colleagues who have contributed to this thesis in different ways. Special thanks to:

AndersLindahl, my supervisor, for giving me the opportunity to be a PhD student in your research group even thou my background was merely ones and zeros, and despite the fact that I knew absolutely nothing about stem cells from the start. Nevertheless, I have always felt that you trusted and believed in me.

Björn Olsson, my supervisor at the University of Skövde, for all your excellent advices and support and for being such a good mentor. Your constructive feedback has been inclusive and valuable and contained comments on my work, encouragement, inspiration and even English lesions, which has been highly appreciated. You have also been an excellent manager for the bioinformatic research group at the University of Skövde.

Peter Sartipy, my supervisor at Cellartis AB, for being such an extraordinary person. You have been encouraging, inspiring, understanding, and supporting at the same time as you have critically scrutinized and corrected my work. In addition, you have been a fantastic colleague and friend during the entire thesis project and I hope that I will get the opportunity to work together with you also in future projects.

All colleagues at Anders Lindahl's Research Group, Sahlgrenska Akademy
I would particularly thank **Julia, Marianne,** and **Johan,** for interesting discussions and collaborations in various projects, and to **Ulla,** for help with administration tasks.

All colleagues at Högskolan, Skövde.
Former and present members of the bioinformatic group **Jonas, Angelica, Zelmina, Simon, Dan, Benjamin, Jessica** and **Sanja,** thanks for good discussions at our meetings and for nice lunches at ThaiThai. You have made my office time in Skövde to a pleasure. Special thanks to Jonas, who is a co-author on several of my papers, and to Angelica for taking good care of the research group, while Björn was on paternity leave.

All colleagues at the **System biology research center**.
Andreas and **Patric,** thanks for successful work with applications.
Jenny and **Lina,** thanks for your nice company at the office and **Eva, Jasmine, Karin** and **Afrouz,** for enjoyable lunch breaks.

All colleagues in the **Information fusion program**, no one mentioned, no one forgotten, our program meetings have given me a wider perspective of my work.

All colleagues at Cellartis AB:
Johan Hyllner, for your outstanding positive attitude that you readily share with everybody. Thanks for giving me the opportunity to collaborate with Cellartis, it has been a real pleasure.

Former and present members in the **Mesoderm research group**:
Hilmar, the party guy, hope you have a good time in Island, Cellartis will never be the same without you. **Marcus,** the transfection guy, thanks for your friendly attitude. **Yalda**, for being such a nice person, hope you enjoy your maternity leave. **Kerstin,** for being such a funny person, I do enjoy your friendship. **Sofie**, for nice meeting and lunch chats. **Karolina**, you were the first one to show me the beating cells, I was fascinated, and I still am. Special thanks to **Caroline** and **Daniella**, for good scientific discussions and for help with proofreading my thesis. Caroline has also done lot of experimental work (with assistance from other group members) as well as being a superb officemate.

Members in the **Endoderm research group:**
Petter Björquist, you are a fantastic person and a fantastic project leader, I admire you! Thanks for having me in your group. **Gabriella,** for your lovely enthusiasm, **Gustav** for your helpfulness, and **Nico**, for invaluable scientific input. **Karin Norrman** and **Anna,** for your positive attitude, **Barbara** and **Jenny L,** for nice time as officemates. **Janne,** for your pleasant company at all conferences. **Josefina,** for your nice company at the trip to Brussels. **Marie R**, for giving me a nice introduction at Cellatis. **Sara, Carina S, Susanna, Maria U, and Tina,** for nice chats at meetings and lunch breaks.

Mikael E, thanks for taking good care of me when I first started at Cellartis, and for sharing your office with me. **Anders A,** for your stand-up comedy skills that you have practiced at various Cellartis parties, you have a lot of laughs on you responsibility. **Camilla K,** for taking your time to proofread my thesis. **Karin Noaksson,** for your pedagogic explanations about stem cells during my introduction at Cellartis. **Katarina A,** for demonstrating how the work in the lab is carried out. **Gunilla,** for demonstration of the characterisation program. **Katarina E**, for showing me how animal testing can be carried out in a respectful way. **Angelica N,** for nice chats about wedding dresses. **Maria F**, for nice chats during lunches and bus trips. **Jenny G,** for nice chats about anything. **Mia E**, for your patience with the tox-dataset, shortly I will take a look at that. **Fredrik**, for being a good officemate. **Raymond**, for system support and good scientific discussions. **Catharina E**, for enjoyable company at conferences.

To all persons in the **production team**, for delivery of cells to my experiments.
To all persons at the **QC,** for confirming the quality of the cells.
To all persons at the **administration**, for taking care of paper work and patent issues.
To all others which I have not mentioned by your names, but who have been equally important to me.

Friends and relatives
To **Pontus,** my fellow student during my undergraduate studies, without your enjoyable company and your computer support at my first years of studies I would never have made the rest. To **Ammi** and **Hanna**, my friends and colleagues during my entire university time, you turned every coffee and lunch break into funny memories.

My friends from our girl club meetings, **Gudrun, Margareta, Berit** and **Helena,** who always have believed that I am smarter than I actually am. You have given me a lot of useful self-confidence.

To **Elisabeth, Per, Anette, Jan J, Lena** and **Jan H**, for your fantastic friendship and company during boat-trips, holidays, and weekends, which have given me lots of laughs and relaxing breaks whenever needed the best.

To **Ulrika,** for your helpfulness, encouragement, consideration, and for taking your time to proofread my thesis.

To my parents **Inga** and **Gillis,** for raising me to be independent and with believe in myself, and to my sisters **Anne, Kate, Maud** and my brother **Tony** and their families, for always being interested in my work and for their effort in trying to understand what I am actually doing. To my family in law, **Christina** and **Håkan,** for their endless support, to **Camilla**, for always being interested and encouraging, and to **Staffan,** for nice chats.

Most of all I would like to thank my outstanding family **Tommy, Sara** and **Robin,** for unconditional support and encouragement, and for your endless patience when I have been working evenings, weekends, and holidays. This thesis is dedicated to you!