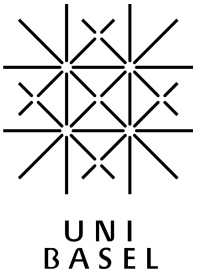
The relative contributions of transcriptional and post-transcriptional regulation to steady-state messenger RNA levels



Inauguraldissertation

zur Erlangung der Würde eines Doktors der Philosophie vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät der Universität Basel

von

Sylvia Tippmann

aus Chemnitz, Deutschland

Basel, 2013

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Dirk Schuebeler und Prof. Peter F. Stadler

Basel, den 22. Mai 2012

Dekan: Prof. Dr. M. Spiess

"You can not explain the whole world but at least have fun while you try to understand part of it." [Ryan J. Taft, 2006]

Acknowledgements

First of all I would like to thank my two PhD supervisors Dirk Schübeler and Michael Stadler for giving me the opportunity to learn and research under their guidance. Dirk, while giving me the freedom to choose my own project, always kept the big picture. He explained biological processes to me when necessary and was never tiered finding new allegories to illustrate them [*"RNAP II and the truck on the highway"*]. Michael always took the time to solve theoretical challenges with me, no matter if for 5 minutes or two hours and in every discussion I learned something new. Moreover, I would like to acknowledge Peter Stadler for being very supportive during the whole project, giving me really fast and sharp advices, even from remote locations.

I also wish to mention all the other people that made a difference during the time of my PhD: all the past and current members of the Schübeler Lab, which made researching at the FMI enjoyable, as well as Lukas and Dimos for smart and critical comments on almost everything. Last but not least I would like to thank my parents and all my friends inside and outside of the research bubble for taking care of me - in general.

Contents

List of Figures

LIST OF FIGURES

Chapter 1

Summary

The regulation of gene expression in eukaryotes is a complex process balancing two opposing schemes into one regulatory network. Stable maintenance of gene expression patterns is as important as quick adaption to intrinsic and extrinsic stimuli. Over the past years it has emerged that gene regulation is a multistep process occurring at many levels. On the level of DNA and chromatin it is determined how efficiently a gene is transcribed by RNA polymerase II (RNAP II) in the first place. Influenced by many processing steps, which are mediated amongst others by RNA binding proteins (RBPs), only a fraction of a respective gene arrives to the cytoplasm, where more regulatory processes alter the lifetime of messenger RNA (mRNA), during which it is available for translation into protein. Due to the local separation of nucleus and cytoplasm in eukaryotes it is intuitive to imagine a stepwise process, which can be split up in transcriptional regulation in the first place and subsequent post-transcriptional regulation.

At the beginning of my PhD high resolution genome-wide data of chromatin modifications [Barski et al., 2007; Mikkelsen et al., 2007] and transcription [Mortazavi et al., 2008; Wang et al., 2009] became available, which allowed a global correlation of mRNA expression with chromatin features. Also supported through RNA sequencing data, more small regulatory RNAs were discovered and their expression linked to specific cell types [Carninci, 2009; Core et al., 2008; Seila et al., 2008; Wang et al., 2009]. Both, histone marks influencing the chromatin environment and post-transcriptional processes operating on RNA level, have a contribution to the final mRNA concentration per gene in a cell. It was still largely unknown if these processes are separable and how much each process contributes to the final mRNA expression.

1. SUMMARY

Therefore we set out to define the relative contributions of transcriptional and post-transcriptional regulation which shape the mRNA profile in a cell. To this end we obtained all necessary data from murine embryonic stem cells, which are differentiated into neurons in cell culture. Modifications at histone H3 (di-methylation of lysine 4 at histone tail H3 (H3K4me2), tri-methylation of lysine 27 at histone tail H3 (H3K27me3) and tri-methylation of lysine 36 at histone tail H3 (H3K36me3)) and RNAP II occupancy were derived by chromatin immunoprecipitation (ChIP) followed by deep sequencing to predict transcription rate. In addition we measure mRNA decay rates of protein coding genes both, by transcription arrest and pulse labeling and infer expression profiles of micro RNAs (miRNAs) during neuronal differentiation by small RNA sequencing.

Our integrative analysis in ESC revealed that chromatin marks are very good predictors of steady-state mRNA level. Especially, H3K36me3, which is a co-transcriptional histone mark, is highly correlated with mRNA abundance when integrated over the whole gene body. This is in contrast to two other studies [Cheng and Gerstein, 2011; Karlic $\mathrm{et\ al.}$, 2010], which also use histone marks to predict mRNA expression, however because their analysis is restricted to regions around the TSS, they do not use the full predictive power of the H3K36me3. Here we show that with H3K36me3, additional two promoter proximal histone marks and RNAP II occupancy, we can explain most of the variance in mRNA levels ($\sim$85%). Based on this result we went on to ask which regulatory mechanism could explain the additional variance in transcript levels, and investigated the contribution of mRNA decay to steady-state levels in general and in particular focus on miRNA-mediated degradation of transcripts.

This analysis, integrating mRNA half-life of each transcript in a model together with transcription-relevant measures, shows, that degradation has a minor quantitative impact on mRNA levels ($<$2%). This is in accordance with two recent publications in murine fibroblast and dendritic cells [Rabani $\mathrm{et\ al.}$, 2011; Schwanhäusser $\mathrm{et\ al.}$, 2011], which show, by measuring mRNA transcription rate and modeling RNA decay, a similar ratio of transcriptional and post-transcriptional regulation to quantify mRNA levels. Furthermore, we were interested in the quantitative contribution of mRNA degradation, which is mediated by miRNAs specifically. To this end we established weighted miRNA-target connections by

combining a posterior probability score [Gaidatzis et al., 2007] of interaction with experimentally inferred miRNA expression data. On a subset of likely miRNA target genes we can see a small effect of miRNA-mediated post-transcriptional decay, however on a genome-wide level the quantitative contribution of this regulatory layer is too small to be detectable.

Together, our findings establish a chromatin-based quantitative model for the contribution of transcriptional and post-transcriptional regulatory processes to steady-state levels of messenger RNA and support the recent notion that the lion share of mRNA expression regulation is happening at the level of transcription [Rabani et al., 2011; Schwanhäusser et al., 2011].

1. SUMMARY

Chapter 2

Introduction

Every multicellular organism originates from a single fertilised egg. During metazoan development this single cell divides and gives rise to many specialised cell types with different phenotypes and functions. While the genetic information of these cells is a constant, their set of expressed genes is subject to major changes throughout differentiation. This process requires the coordinated regulation of gene expression, which is a complex, multi-layered process in eukaryotes. Gene expression regulation describes the whole processes, that cells use to regulate the way that information in genes is turned into gene products. At any step the gene's expression may be modulated, from transcription of the DNA to RNA, during splicing, export to the cytoplasm, before, during and after translation. While there is a fairly good understanding of the mechanistic details of each of the regulatory processes, the interaction between them has not been studied until recently.

In bacteria regulatory pathways from DNA over RNA to protein are often directly coupled due to the lack of a compartmentalisation [Montero Llopis et al., 2010]. Coupling, in a non-direct way, might also occur in eurkaryotes [Dahan et al., 2011], however we can distinguish processes that happen in the nucleus from cytoplasmic events.

The following paragraphs will summarize current knowledge on transcriptional regulation of RNA synthesis and post-transcriptional down-regulation of mRNA. I will also introduce quantitaive measures that provide potential readouts of these regulatory layers, relating to my PhD thesis project.

2.1 Transcriptional Gene Regulation

Throughout evolution, complexity of organisms scales with genome size. Paradoxically, the number of genes does not match up with this increase in size and complexity, a phenomenon termed the c-value enigma [Gregory, 2001]. The mouse genome for example is 240 times bigger than budding yeast *saccharomyces cerevisiae*, however it encodes only 4 times more proteins (23,000 genes in mouse [Waterston et al., 2002] vs 5,500 in yeast [Kellis et al., 2003]). This raises the question, how is complexity achieved? One explanation is the number of transcription factor (TF) genes, which increases exponentially (exponent=1.26) with the number of total protein coding genes in an eukaryotic organism [van Nimwegen, 2003]. TFs are the most prominent and best studied mediators for gene expression regulation [Vaquerizas et al., 2009]. Their recognition motifs are on average 6-8 bp in length, in prokaryotes as well as in and eukaryotes, and in many cases the binding motifs are degenerate [Wray et al., 2003]. In large vertebrate genomes however, where only a small portion encodes proteins or regulatory RNAs [Waterston et al., 2002] this poses a major challenge: in the mouse genome for example, assuming a random sequence distribution, any potential 6-mer motif could bind more than 732,400 times. From ChIP-sequencing experiments we know that the actual number of sites bound by a TF in a cell is considerably smaller. Combinatorial regulation of transcription factors [Bilu and Barkai, 2005] could possibly confer specificity of TF binding, but further structuring of large genomes is required to guide the TFs to their respective target sites and thereby reduce random binding.
Chromatin modifying mechanisms co-evolved with genome size: although the use of chromosomal architectural proteins variants is conserved back to eubacteria, in the transition from pro- to eukaryotes, mechanisms for 'writing' chemical modifications, that constitute persistent signals, onto chromatin appeared [Prohaska et al., 2010].
In the following sections I will give an overview of cellular processes that contribute to transcriptional regulation on the level of chromatin in eukayotes.

2.1.1 DNA, Histones and Chromatin

Roughly two meters of DNA are in the nucleus of every mammalian cell. For obvious packaging but also regulatory purposes the DNA is highly compacted, where the chromosome represents the highest compaction form. The chromosome is composed of a highly folded

30 nm chromatin fibre of packaged nucleosomes. Nucleosome structure, so called 'beads on a string', consists of DNA wrapped around histones thereby achieving a high initial condensation (Figure 2.1). Nucleosomes consist of approximately 150 bp of DNA wrapped around

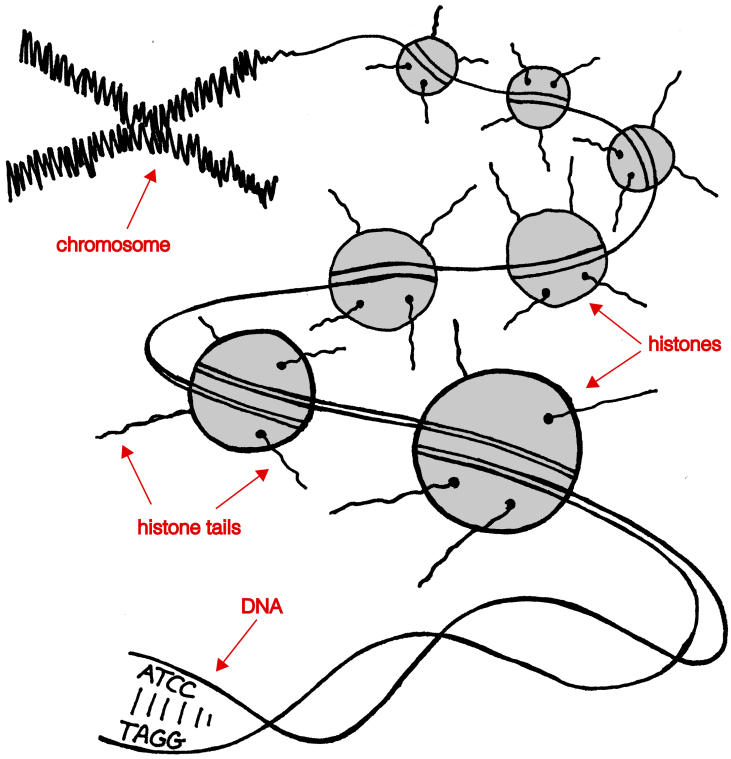


Figure 2.1: Model of chromatin compaction of DNA in the nucleus depicts the DNA double-strand, which is wrapped around histone proteins to form the nucleosome. This 'beads on a string' structure of nucleosomes is then further compacted into a 30nm fiber with the help of linker histone H1 and even more condensed on further scaffold proteins leading to a chromosome structure.

a protein octamer of four core histones, H3, H4, H2A and H2B [Kornberg and Thomas, 1974]. Together with so called linker histones (H1), this results in a more than 50-fold compaction of the genome in the nucleus of the cell, termed chromatin (Figure 2.1). In addition to packaging this chromatin conformation also allows to make DNA more or less accessible for TFs to bind. The tight structure of DNA wrapped around histones is in itself rather in-accessible [Lam et al., 2008; Struhl, 1999; Workman and Kingston, 1998].

Alteration of core histone stoichiometry in yeast leads to constitutive activation of many

inducible genes (Han and Grundstein 1988). This provides support for the repression of basal transcription through chromatin assembly. Chemical modifications on histone tails lead to recruitment of non-histone proteins or directly influence the electric charge of the chromatin and can thereby pull nucleosomes closer together or push them apart (Figure 2.1). This regulates the access of TFs and ultimately the transcription machinery to the DNA and renders the chromatin either permissive or repressive for transcription.

2.1.2 DNA Methylation

In theory, a methyl group can be added to any of the 4 nucleic acids, making it a methyl-A, methyl-G, methyl-T or methyl-C. However, in eukaryotes DNA methylation is exclusively found at cytosine residues. Not all eukaryotes methylate their genomes, for example yeast and the roundworm *C. elegans* contain no methylated cytosines at all [Antequera et al., 1984; Simpson et al., 1986], while all vertebrates seem to display genome-wide DNA methylation which, in mammals, mostly occurs in the context of CpG dinucleotides [Suzuki and Bird, 2008]. Genome-wide studies revealed a bimodal distribution of CpG methylation, with most of the genome being highly methylated (that is 80-100% methylation) and a few regions largely devoid of methylation, which correspond to relative local enrichments of CpG dinucleotides, called CpG islands [Bird, 1986]. CpG islands mainly co-localize with promoters, the transcription regulatory unit of a gene. Recently, however our laboratory identified novel regions, which are not CpG islands but which nevertheless have low methylation levels, termed low methylated regions (LMRs) [Stadler et al., 2011].

DNA Methylation and Transcription

Early studies in mouse and cancer cells lines connected DNA methylation with X-inactivation, imprinting and transposon silencing and led to the common theme that DNA methylation functions to maintain a repressed chromatin state and silence promoter activity [Bird and Wolffe, 1999; Suzuki and Bird, 2008]. Although it was not initially appreciated that DNA methylation could be a transient mark, large-scale studies revealed that many promoters and LMRs vary in their DNA methylation according to cell type [Bibikova et al., 2006; Eckhardt et al., 2006; Mohn et al., 2008; Rakyan et al., 2004; Rollins et al., 2006; Stadler et al., 2011; Weber et al., 2007]. The results showed that the majority of the analyzed

regions do not show a continuum of CpG methylation levels. Instead they were either hypomethylated (less than 30% of CpG sites) or hypermethylated (more than 70% of CpG sites), suggesting two alternative states: silent and methylated or active and unmethylated. The effect of DNA methylation on gene transcription seems to depend on the CpG content of the promoter. Single gene studies suggested that methylated, CpG-poor promoters can repress transcription [Boyes and Bird, 1992; Schübeler et al., 2000]. Genome-wide measurements of DNA methylation showed that some CpG-poor promoters are methylated, even when the corresponding gene is actively transcribed [Ball et al., 2009; Meissner et al., 2008; Weber et al., 2007]. In contrast to CpG-poor promoters, DNA methylation at promoters with high CpG content, is clearly anti-correlated with transcription of the associated gene [Weber et al., 2007]. Two models have been proposed, for the mechanism, by which the transcriptional inhibition occurs [Appanah et al., 2007; Schübeler et al., 2000], however both act at the level of transcription initiation: One model postulates that DNA methylation inhibits the binding of methylation-sensitive TFs, the second model is more indirect where proteins specifically binding to methylated CpGs recruit co-factors, which in turn repress transcription. For most known methyl-CpG-binding domain proteins (MBDs) an interaction with factors that set up a repressive chromatin environment has been reported. A variety of such MBDs are known and for most of these proteins, it has been reported that they interact with factors that set up a repressive chromatin environment such as HDACs and the NURD complex [Clouaire and Stancheva, 2008]. However, not only promoter proximal DNA methylation has an influence on gene expression: A recent study reported regions with intermediate CpG content, that have low methylation levels and are cell-type specific. These loci are likely to be distal regulatory regions and are occupied by cell type specific TFs [Stadler et al., 2011].

2.1.3 Histone Modifications

Histones consist of a globular center and flexible arms, protruding from the center, called 'histone tails', which have many basic, or positively charged, amino acids (Figure 2.2). It was found that removal of histone tails from the nucleosome with the protease trypsin facilitates binding of TATA binding protein (TBP) [Godde et al., 1995] and other TFs [Lee et al., 1993] and causes specific effects on gene expression [Kayne et al., 1988]. This lead to the conclusion that the N-terminal tails of the core histones have an important role in

regulating TF access to the DNA [Godde et al., 1995]. Importantly N-terminal tails of histones are targets for enzymes that modify chromatin structure. Modifications on histones take place on the N-terminal tails, mostly of histone H3 and H4, which stick out from the nucleosome core. They contain more than 60 sites which are subject to post-translational modificationss (PTMs) such as acetylation, methylation, ubiquitination, phosphorylation, sumoylation and others [Kouzarides, 2007] (methylations and acetylations of N-terminal tails illustrated in Figure 2.2). Later studies revealed that PTMs are highly dynamic and have a regulatory role [Brownell et al., 1996; Rea et al., 2000; Taunton et al., 1996].

Modifications associated with active transcription, such as acetylation of histone 3 and histone 4 or di- and tri-metylation of H3K4, are termed euchromatic modifications, whereas modifications localized to inactive genes, such as H3K9 methylation and H3K27 methylation are referred to as heterochromatic modifications (reviewed in Li et al. [2007]). The concept is sketched in Figure 2.3. In the following subsections I will briefly discuss active (euchromatin) and repressive (heterochromatin) histone modifications and especially highlight the histone lysine methylations, H3K27, H3K4 and H3K36, which will be important for my thesis.

Histone Acetylation

Histone acetylation, similarly to the removal of histone tails, alters the constraints on the wrapping of DNA on the nucleosome [Bauer et al., 1994] and reduces the stability with which these flexible domains bind to DNA [Cary et al., 1982]. Histone acetylation neutralises the charge of nucleosomes, thereby destabilizes nucleosomes, increases DNA accessibility and leads to non-histone protein binding to DNA *in vitro* [Imbalzano et al., 1994; Lee et al., 1993; Vettese-Dadey et al., 1996]. Since transcriptional co-activators in yeast and human have the capacity to acetylate histones [Brownell et al., 1996], an attractive hypothesis is that targeted histone acetylation followed by the disruption of chromatin will have a major causal role in gene regulation [Wolffe and Pruss, 1996]. Acetylated lysines on histones H2B, H3 and H4 are largely overlapping and highly correlated with active transcribed regions in yeast [Pokholok et al., 2005], fly [Schubeler et al., 2004] and human [Wang et al., 2008]. With the exception of H4K16 acetylation, which directly interferes with higher order chromatin structure [Shogren-Knaak et al., 2006], acetylation of individual lysines conveys little

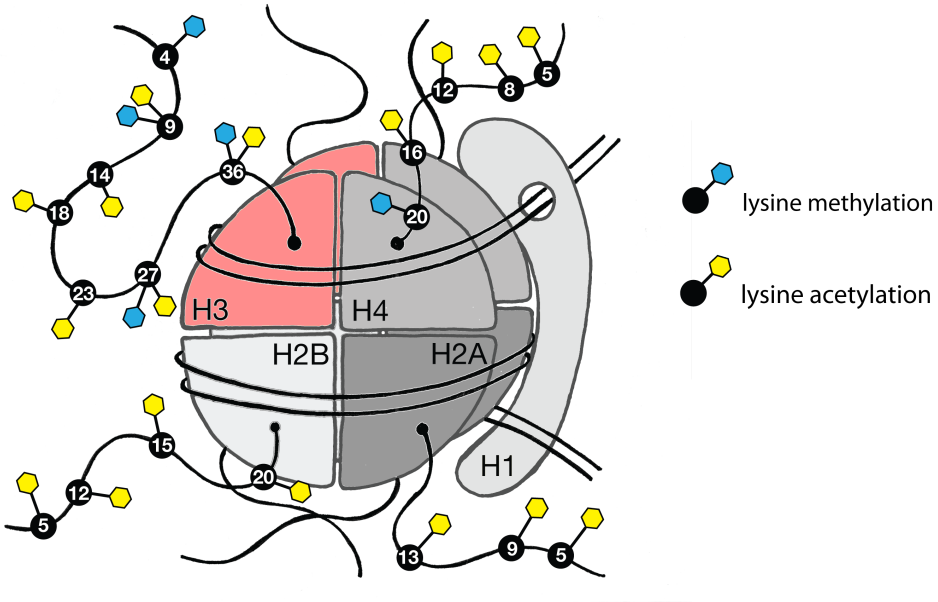


Figure 2.2: A nucleosome is composed of a protein octamer consisting of the four core histones, H3, H4, H2A and H2B and the double stranded DNA. C- and N-terminal histone tails of the core histones can be modified, here only lysine modifications, methylations and acetylations, are depicted. The linker histone H1 aids in compaction of the chromatin.

specificity. It is rather the cumulative effect of acetyl groups at multiple lysines which is important for regulating DNA accessibility.

Histone Methylation

In contrast to acetylation, histone methylation is often catalyzed by a specific enzyme at a specific site and results in unique functions. Methylation of histones can either occur at lysine or arginine residues. The same residue can exist in mono- (me1), di- (me2) or tri-methylation state (me3) state which adds another level of regulatory potential. Several lysines display diverging functions and localization in the genome depending on their methylation state (Barski et al., 2007; Peters and Schubeler, 2005).
ChIP experiments showed that active genes are methylated at lysine 4 and 79 of histone H3 (H3K4and H3K79) and lysine 36 of histone H3 (detailed introduction to H3K36me, in section 2.1.5) [Barski et al., 2007; Pokholok et al., 2005; Schubeler et al., 2004], therefore these modifications are thought to have a role in transcription. H3K36me and H3K79me display

a broader distribution within the gene body, while H3K4 methylation states show a distinct promoter proximal profile: K4me3 peaks at start sites, K4me2 and K4me3 downstream of the transcription start site (TSS) [Li et al., 2007; Pokholok et al., 2005]. H3K4 methylation can be bound by chromatin remodelling complexes and different histone acetyltransferases, creating accessible chromatin and may thereby directly contribute to transcription initiation [Santos-Rosa et al., 2003; Taverna et al., 2006]. Although H3K4me3 can be directly bound by the general transcription factor TFIID and thereby might facilitate transcription [Vermeulen et al., 2007], it is not exclusively located at transcribed regions in mammals. Recent data indicates that in contrast to invertebrates H3K4me2/3 are not exclusively marking actively transcribed regions, depending on the CpG content of the promoter this mark correlates with low or high levels of RNAP II [Bernstein et al., 2006; Guenther et al., 2007; Mikkelsen et al., 2007; Roh et al., 2006; Weber et al., 2007].
An additional mark shown to be enriched at transcribed genes is H3K79 methylation [Schubeler et al., 2004]. All three methylation variants of H3K79 are catalyzed by DOT1, the only lysine histone methyl-transferase (HMT) that does not contain a SET domain [van Leeuwen et al., 2002]. The role of this modification in regulation of transcription, however, remains still unclear.
In yeast, a second HMT, named SET2, mediates H3K36 methylation, another mark associated with transcription. Upon methylation of H3K36, the histone deacetylase complex Rpd3 removes acetylation [Keogh et al., 2005], which has been suggested to be involved in preventing spurious transcription [Carrozza et al., 2005]. In Section 2.1.5 I will introduce H3K36me3 and it's role in transcription in more detail.

Inactive loci display a different set of methylation marks mainly consisting of methylation of H3K9, H4K20, and H3K27. H3K27 di- and tri-methylation predominantly localizes to CpG-rich regions and is excluded from regions carrying H3K9 methylation. H3K27me3 is known as a mechanistic intermediate during transcriptional repression by polycomb-group (PcG) proteins. Polycomb-mediated repression is carried out by the two polycomb-repressive-complexes PRC2 and PRC1. While PRC2 sets the H3K27me3 mark, PRC1 is thought to be the reader protein, which in turn ubiquitinates lysine 119 at histone H2A [Simon and Kingston, 2009]. The two PRC complexes are thought to mediate repression by inhibiting chromatin remodeling, blocking transcription and/or by mediating chromatin compaction [Margueron et al., 2008].

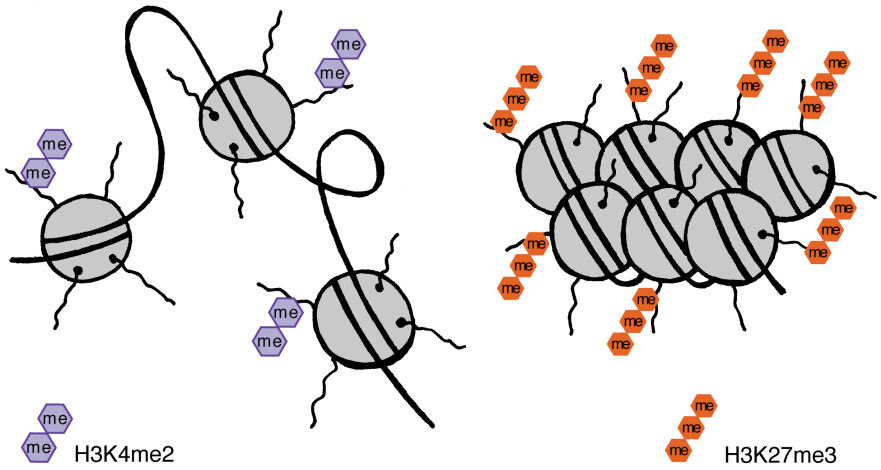


Figure 2.3: The figure demonstrates two different states of chromatin: active, accessible chromatin (left) and inactive, 'closed' chromatin (right). Each state is accompanied by characteristic modifications of histone tails. Here only two representative histone marks are depicted: H3K4me2 at accessible chromatin and H3K27me3 at closed chromatin.

PcG proteins and H3K27me3 occupy many inactive promoters of key developmental regulators in embryonic stem (ES) cells, suggesting that they maintain pluripotency and cellular identity in these cells [Boyer et al., 2006]. Also in later steps of differentiation PcG proteins were shown to play an important role [Ezhkova et al., 2009; Mohn et al., 2008].

2.1.4 RNA Polymerase II and Transcription

The first step, at which the expression level of genes is regulated in eukaryotes, is RNA transcription in the nucleus of the cell. RNAP II is the enzyme that transcribes all genes encoding mRNA as well as as some structural or regulatory RNAs. A feature which distinguishes RNAP II from the other two eukaryotic RNA polymerases is the extended carboxyl-terminal domain (CTD) of the largest RNAP II subunit Rpb1. The 52 copies of the CTD are subject to modifications during the transcription cycle. While serine 5 phosphotylation of the CTD is indicative of pausing the serine 2 phosphorylated form is characteristic for elongating polymerase [Phatnani and Greenleaf, 2006]. The phosphorylation affects the CTD's conformation and ability to associate with factors involved in elongation, RNA processing and termination of transcription (reviewed in Saunders et al. [2006]).

Initiation of Transcription

Before transcription initiates RNAP II is positioned at the core promoter by a combination of the general transcription factors (GTFs) TFIID, TFIIA and TFIIB to form the pre-initiation complex (PIC) [Thomas and Chiang, 2006]. TFIIH then melts 10-15bp of the DNA in order to position the single stranded template of RNAP II to initiate RNA synthesis. RNAP II CTD gets phosphorylated at serine 5 during the first 30 bp, before elongation starts. The phosphorylated CTD then recruits factors important for productive elongation and mRNA processing [Buratowski, 2003] to the transcription machinery.
Even though this appears straight forward, the rate of transcription is subject to regulation at each of these steps: A study using model fitting based on photo-bleeching and live imaging in a human cell line, predicted that only 13% of RNAP II, which interacts with the promoter, are delivered to the initiation step and only 8.6% of these RNAP II engage in productive elongation [Darzacq et al., 2007]. In total this means that on average only one RNAP II out of 90 interaction events produces a mature mRNA molecule, suggesting a tight transcriptional regulation.

The packaging of DNA into chromatin contributes largely to this tight regulation, from activator binding over PIC formation to productive elongation. A prominent example is the PHO5 promoter in yeast, which contains one exposed binding site for the TF Pho4 located in the linker DNA between two nucleosomes, while additional binding sites are buried within nucleosomes [Adkins et al., 2004; Almer and Hörz, 1986; Boeger et al., 2004]. During induction, Pho4 binds to the accessible site first, recruits proteins which modify histones and remodel nucleosomes, and thereby expose the secondary binding sites to the TF.
Since the chromatin conformation of DNA is already repressive in itself, regions of active transcription need to be relieved of condensation. Indeed in yeast it has been shown, that highly transcribed genes have a lower nucleosome occupancy than intergenic regions, with pronounced nucleosome depletion in promoter regions [Pokholok et al., 2005]. But not only at the initiation step, chromatin needs to be de-condensed, also during transcription elongation the barrier posed by nucleosomes in the coding regions, has to be overcome, either by completely dis- and reassembling nucleosomes or by modifying histone tails.

Elongation of Transcription

Recent studies have challenged that transcription is predominantly regulated at the level of RNAP II binding and initiation and it is now apparent that regulation at the elongation step is equally important [Min et al. [2011], also reviewed in Saunders et al. [2006]]. Elongation is divided into three distinct stages: promoter escape, promoter proximal pausing and productive elongation. Each of these stages involves a different behavior and stability of the transcription complex and a specific manipulation of the chromatin environment.

Promoter escape begins after the assembly of the PIC and with the onset of transcription initiation, from this point the transcription complex is termed the initially transcribing complex (ITC). If RNAP II is subjected to other challenges, the ITC can still abort the nascent RNA until about 23 bp downstream of the promoter [Pal and Luse, 2003]. Promoter escape is considered complete and the ITC becomes an early elongation complex (EEC) when the Rpb7 subunit of RNAP II stably associates with the nascent RNA [Ujvári and Luse, 2006]. The nascent RNA can also bind the CTD, which might affect transcription elongation [Kaneko and Manley, 2005].

Another step, other than RNAP II recruitment or transcription initiation, is rate limiting and a target of regulation: promoter proximal pausing. This is an event in which the forward movement of elongation competent transcription complexes is temporarily blocked owing to template sequence, regulatory factors or both. High-resolution analysis showed that the pausing occurs at several sites from +20 to +40 [Giardina et al., 1992; Rasmussen and Lis, 1993]. Pausing can provide a checkpoint to assess whether the RNAP II is correctly prepared for productive elongation, and allows rapid regulation of gene expression. Capping enzyme associates with the Ser5-phosphorylated CTD of RNAP II [Wen and Shatkin, 1999], and the nascent RNA becomes capped during elongation through the pause site [Rasmussen and Lis, 1993]. The phosphorylated CTD stimulates capping enzyme activity *in vitro* [Wen and Shatkin, 1999]. Promoter proximal pausing might facilitate correct capping, and a correctly capped nascent RNA might be a prerequisite for escape from the pause [Pei et al., 2003].

Several factors are required for the efficient release of paused RNAP II into productive elongation, after which RNAP II proceeds through the remainder of the gene. This is proposed to occur by the action of the positive transcription-elongation factor-b (P-TEFb) complex. P-TEFb phosphorylates factors facilitating the paused state, DSIF, NELF and Ser2 of the

RNAP II CTD [Yamada et al., 2006]. Upon transition to productive elongation DSIF remains associated but NELF leaves the elongation complex [Wu et al., 2005].

Termination of Transcription

Finally, termination of transcription requires the dissociation of RNAP II and the transcription complex from the template. This may occour either through a conformational change in RNAP II following transcription of the poly(A) site [Zhang and Gilmour, 2006] or by an RNA exonuclease mediated degradation of mRNA, that is still associated to RNAP II and thereby stimulates its termination ('torpedo model' Luo and Bentley [2004]).

2.1.5 Readout of Transcription: H3K36me3

The presence of elongating RNAP II is the sign of active transcription of genes, however, by common methods, such as ChIP, the moving enzyme is hardly detectable along the gene body. A more stable readout for transcription would therefore be a histone modification, which is set co-transcriptionally: H3K36me3.
In yeast all three H3 lysine 36 methylation marks, mono-, di- and tri-methylation are mediated by the non-essential SET domain-containing (Set2) protein. It associates with the large subunit of RNAP II (Rpb1) in its hyperphosphorylated form during transcriptional elongation and deposits the trimethyl group onto H3K36 [Kizer et al., 2005; Li et al., 2003, 2002; Xiao et al., 2003]. In addition the RNAP II, CTD kinase 1 (Ctk1) and the elongation factor Spt6 regulate the levels of H3K36 tri- but not di-methylation [Lin et al., 2010; Youdell et al., 2008].
In metazoa the lysine 36 methytransferases are essential and specific for each level of methylation. H3K36 mono- and di-methylation are set by nuclear receptor binding SET domain protein 1 (NSD1) in human [Lucio-Eterovic et al., 2010], shown through enzymatic essays [Li et al., 2009] and structural data [Qiao et al., 2011]. Maternal effect sterile 4 (MES-4) is the NSD1 orthologue in fly [Bell et al., 2007] and worm [Bender et al., 2006], and although it exclusively sets mono- and di-methyl groups it indirectly regulates the H3K36 tri-methylation by adjusting the availability of substrates to the tri-methylating enzymes. In worm and fly the tri-methylating enzymes are termed histone-methyltransferase-like 1

(MET-1) [Andersen and Horvitz, 2007] and Set2 [Bell et al., 2007] respectively. The human orthologue SET domain-containing 2 (SETD2) (aka HYPB or KMT3A) indeed requires the NSD1 mediated substrate of H3K36me2 [Edmunds et al., 2008] to set tri-methylation. It was shown that even with normal levels of H3K36me2 a depletion of SETD2 results in reduced H3K36me3 levels [Yuan et al., 2009].
Similarly to yeast Set2, human SETD2 interacts with RNAP II during elongation to target H3K36 [Sun, 2005; Yuan et al., 2009]. This interaction is also regulated by the phosphorylated residues in the CTD of Rpb1. During elongation heterogeneous nuclear RNAs (hnRNAs), including precursors and mature mRNA, associate with specific proteins to form heterogenous ribonucleoprotein (hnRNP) complexes. Knockdown analyses of one of those proteins, heterogenous ribonucleoprotein L (hnRNPL), revealed decreased levels of H3K36 tri- but not mono- or di-mathylation [Yuan et al., 2009], indicating that hnRNPL interacts with SETD2 during active transcription.
It was shown in single gene experiments [Bannister et al., 2005; Edmunds et al., 2008; Vakoc et al., 2006] as well as genome-wide studies [Barski et al., 2007; Bell et al., 2007; Mikkelsen et al., 2007; Pokholok et al., 2005] that H3K36me3 levels are correlated with the expression of active genes. In metazoan and yeast H3K36me3 has a characteristic distribution pattern increasing towards the 3' ends of transcription units [Barski et al., 2007; Bell et al., 2007; Pokholok et al., 2005] . In chicken, there is a shift from mono- to trimethylation of H3K36 from the promoters to the 3' ends of active genes [Bannister et al., 2005]. Consistent with a role for H3K36me in transcription, data from yeast denote that H3K36me prevents cryptic initiation via recruiting a histone deacetlyase to the body of genes, which in turn presumably leads to a less accessible chromatin structure (Carrozza et al., 2005).

Several large-scale bioinformatic studies have analysed both the positions of nucleosomes and their modification status within the genomes of humans, *C. elegans*, *D. melanogaster* and mice [Kolasinska-Zwierz et al., 2009; Schwartz et al., 2009; Spies et al., 2009].
In each case, nucleosomes were enriched specifically at exonic sequences. Although the increased deposition of nucleosomes at exons guarantees a bias in histone modifications within exons relative to those within introns, it is also clear that a subset of modifictions is specifically enriched here. This is particularly true for H3K36me3 but also includes methylation at H3K79, H4K20 and H2BK5 [Schwartz et al., 2009]. Each analysis also found

that the H3K36me3 bias is more pronounced within exons further downstream of the transcription start site. However it is subject to debate whether there is a causal relationship between the histone modification and the exonic position and if yes, which is cause and which is consequence [Kim et al., 2011; Schwartz et al., 2009]

Bioinformatic Aspects of H3K36me3 as a Readout

For our study we use H3K36me3 as a readout of transcription. To this end, chromatin is isolated and fragmented. DNA fragments which are associated with histones carrying H3K36me3 are enriched by ChIP and analyzed by deep-sequencing on an Illumina GA II. The raw data obtained by deep sequencing are ∼80 million sequence strings ('reads') of size 36nt. To obtain a quantitatively meaningful H3K36me3 level per gene, some processing steps need to be considered. We initially, filtered low-complexity reads based on their dinucleotide entropy, which is calculated by:

$$H = \sum_i f_i log(f_i),$$

where f_i is the frequency of dinucleotide i in the read and the $\sum$ is over all dinucleotides (i from 1 to 16). Reads were filtered out if H was less than half the dinucleotide entropy of the genome, typically removing less than 0.5% of the reads in the given sample. In order to assign H3K36me3 enrichments to genes, the reads have to be mapped to their respective position in the genome. A read can possibly map to each position in the mouse genome, which is 3*10^9 bases in size. A brute-force approach to the mapping problem would therefore in the worst case require 80 million times 3*10^9 pairwise comparisons, which even with the fast development of computational hardware, would be too time intensive. To overcome this limitation the concept of suffix trees is applied, which was introduced in the 70s by Weiner [Weiner, 1973] and later speed-up by Ukkonen [Giegerich, 1997]. The genome is decomposed into a 'tree structure' for once and subsequently each read mapping event runs in the time of the read length. In addition this allows to even map reads, which have mismatches (e.g. due to sequencing errors) to their locations in the genome. Alignments to the mouse genome allowing two mismatches per read were performed by the software bowtie, which implements this algorithm [Langmead et al., 2009]. Due to repeat elements and pseudogenes a read can possibly map to multiple locations in the genome, which all

have the same probability to be the origin of this read. We therefore allow a read to map up to 100 times not to restrict our analysis to uniquely mapping reads. In addition to track genomically untemplated hits (e.g., exon-exon junctions), the reads were also mapped to an annotation database containing known mouse sequences. To account for the multiple assignment of reads each alignment was weighted by the inverse of the number of hits for this read. All further quantifications were based on weighted alignments. To quantify the level of H3K36me3 per gene we had to ensure that that the signal was not blurred by antisense transcripts or overlapping genes with a shifted TSS. For illustration of this problem assume we want to relate the H3K36me3 level of a region in the genome with the mRNA level of a gene transcribed from this region (from one specific strand). The ChIP data inherently lacks information about the strand because the IP is done on double stranded DNA bound to histones, however the RNA sequencing data is specific for one strand. To exclude that we do not associate transcript abundance with H3K36me3 signal from an overlapping gene location, we stringently excluded based on annotation all mRNA transcripts, which either overlap with another transcript on the complementary strand or with an overlapping transcript on the same strand but shifted TSS. In addition we had to consider that there may be several annotated transcript variants of a gene due to alternative splicing, therefore we selected the transcript version of median length to be the 'representative' transcript of this gene. These filtering steps left us with around 10.000 genes, distant enough to other transcripts to be safely quantified as separate entities. For those transcripts H3K36me3 reads were summed up over the whole gene body and divided by the length of the gene, to yield a H3K36me3-density per gene. This density was later logarithmically transformed for use in the linear regression.

2.2 Co- and Post-Transcriptional Gene Regulation

For many years, it has been assumed that transcriptional regulation of genes is the major source of differential gene expression. However, it becomes more and more evident, that transcriptional regulation can only partly explain why and at what level proteins are expressed. Accordingly, quantitative mRNA expression studies are insufficient to predict protein levels [Gygi et al., 1999].

As co-transcriptional gene regulation I will refer to all mechanisms targeting the transcript once RNA polymerase has started to transcribe until it releases the mRNA. Following this scheme, post-transcriptional control of gene expression begins with transcription termination in the nucleus and extends over mRNA export to all effects, which alter mRNA abundance in the cytoplasm of the cell before translation into protein.

2.2.1 Co-transcriptional RNA processing

Co-transcriptionally, several processing steps have to take place to transform the pre-mRNA into mature mRNA: capping, splicing and poly-adenylation (Figure 2.4). Consequently the complexes that mediate this mRNA processing have to be tightly linked in space and time to the transcription machinery [Proudfoot et al., 2002], which in turn makes them equally dependent in chromatin.

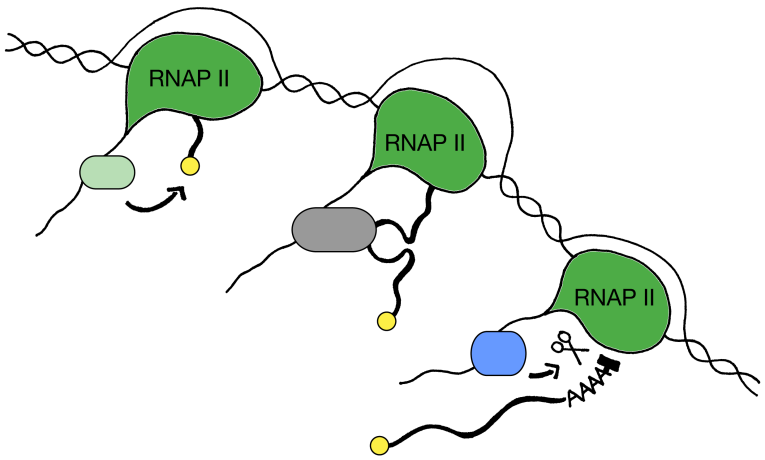


Figure 2.4: RNA processing happens co-transcriptional. The CTD of the RNAP II serves as a scaffold for modifying enzymes which aid in capping, RNA splicing and poly-adenylation.

5' End Processing: Capping

The first RNA processing event to occur on the nascent transcript is 5' end capping, which happens within the first 40 nucleotides. Three enzymes, a triphosphatase, a guanyl transferase, and a methyl transferase, all act in concert to add a cap to the 5' terminus of the primary transcript [Shuman, 2001]. The first two activities are present on a single polypeptide in mammals which gets recruited to the RNAP II initiation complex once the CTD has become activated by Ser5 phosphorylation. Through direct association with CTD Ser5P, the capping enzyme acts on nascent transcripts as soon as they emerge from the elongating RNAP II. Capping may well be a key component of the switch that pushes RNAP II from abortive early elongation into fully processive elongation across the body of the gene. Furthermore the 5' cap allows the mature mRNA to circularize, thereby confering stability and protecting from degradation [Rasmussen and Lis, 1993].

Transcript Splicing

In eukaryotes most pre-mRNA is composed of protein-encoding exons and large noncoding intervening sequences, or introns. In the splicing process introns are removed and exons are joined together to form the mature mRNA, used in translation to produce the correct protein. Selective inclusion of different coding sequences (alternative splicing) results in the production of different protein isoforms. For many eukaryotic introns, with exception of self-splicing introns, splicing is catalyzed by the spliceosome. It consists of the U1, U2, U4, U5 and U6 small nuclear RNPs (snRNPs) in conjunction with a large number of additional proteins (reviewed in Stark and Lührmann [2006]). A series of RNA–RNA, RNA–protein, and protein–protein interactions within the spliceosome is needed to remove intronic regions and subsequently join exons, producing a mature transcript (reviewed in Collins and Guthrie [2000]). Intron identification relies on specific sequences defining the 5' and 3' splice site. In mammals, many genes contain multiple introns that are up to hundreds of thousands of nucleotides in length [Waterston et al., 2002]. The presence of potential splice sites in eukaryotes is not necessarily leading to selection of these sites by the spliceosome. Trans-acting regulatory factors bound by pre-mRNA regulatory elements enhance or repress the recruitment of snRNP to splice sites. These multiple factors together determine the actual splice site *in vivo*. In mouse more than 50% of the transcripts are subject to alternative

splicing, represents an important source of flexibility in gene expression.
As part of the large splicing complex, there are a number of proteins, which leave a mark on spliced mRNAs and thereby direct localization, translation and decay of the mature mRNA. The most studied eukaryotic splice-dependent mark is the exon junction complex (EJC). EJCs are stably deposited ∼20 nucleotides upstream of exon-exon junctions [Le Hir et al., 2000]. They play a role in non-sense mediated decay (NMD) and directly enhance translation initiation by promoting the pioneer round of translation [Moore and Proudfoot, 2009]. In addition the THO/TREX complex associates with spliced mRNAs at the 5'-most exon and promotes rapid export to the cytoplasm [Valencia et al., 2008]. Finally, a number of DEAD-box proteins have recently been found to associate with mRNAs in a splice-dependent manner. These proteins seem to influence many aspects of mRNA metabolism [Rosner and Rinkevich, 2007]. All these evidences show that spliced mRNAs carry numerous protein marks related to their splicing history, which has important downstream effects.

3' End Processing: PolyA Addition

PolyA addition, or polyadenylation, occurs during the completion of the transcriptional process, following transcription of the poly(A) site and cleavage of the transcript. 3' cleavage and polyadenylation of pre-mRNA are dictated by polyA signals that define the end of the mRNA. These signals are recognized by a substantial cleavage/polyadenylation protein complex (polyA complex) that is recruited to the Serine 2 phosphorylated form (Ser2P) of the CTD through direct CTD-interacting domains (CIDs) as well as RNA binding domains (RBDs) that specifically recognize the pre-mRNA polyA signals. Specific CIDs and RBDs have been identified on individual polyA complex subunits [Proudfoot, 2004].
Polyadenylation, the final stage in pre-mRNA cotranscriptional processing, is a critical control point in preventing aberrant gene expression. When 3' processing is either inefficient or compromised by gene mutation, the nuclear exosome is recruited to rapidly degrade the unwanted transcript. Finally, polyadenylation facilitates mRNA release from the transcription site and its ultimate export through the nuclear pore complex (NPC) to cytoplasmic translation. Like the 5'-cap structure, the 3'-polyA tail is important for mRNA stability in the cytoplasm.

2.2.2 Export of mRNA into the cytoplasm

Before an mRNA is exported into the cytoplasm it has to pass several mRNA quality control steps. Splicing defective mRNAs as well as transcripts with aberrant 3'-ends are retained at the site of transcription and directly degraded by the exosome in the nucleus. Once an mRNA has passed the nuclear surveillance system, mRNA export factors, which have been deposited on the mRNA during processing, interact with nuclear pore proteins and mediate the transport of the mature mRNA into the cytoplasm [Hocine et al., 2010].
In mammals only about 5% of the total mass of RNA synthesized ever leaves the nucleus. In section 2.2.1 the extensive mRNA processing, including splicing, capping, polyadenylation and quality control was discussed. A large fraction of the transcripts that does not pass these steps or is otherwise damaged, is immedately degraded. The export of the mature mRNA transcript is delayed until all processing has been completed.
One of the few well described examples of regulated nuclear export that of the human immunodeficiency virus (HIV). The viral RNA directs the formation of double stranded DNA and its insertion into the host genome, where it gets transcribed by the host cell's RNAP II. In order to produce progeny virus complete unspliced, intron containing, transcripts need to be exported to the cytoplasm to be packaged into newly synthesized viral capsids. To overcome the host cell's normal block to export unspliced mRNA, HIV encodes a protein REV, which, once translated, binds to the pre-mRNA of the virus in the nucleus and shuttles it though the nuclear pore by interacting with the export receptor exportin 1.
A key mediator of nuclear mRNA export is the THO/TREX complex, mentioned in section 2.2.1. Consisting of the pentameric THO complex, which functions in transcription elongation, and the mRNA export factors REF/Aly and UAP56, it associates with the 5'-most exon of spliced mRNAs. UAP56 functions in spliceosome assembly [Iglesias and Stutz, 2008; Köhler and Hurt, 2007], while REF/Aly bridges the mRNA to the export receptor NXF1/ TAP. In mammals, REF/Aly and UAP56 appear to be recruited as a consequence of splicing: when uncoupled from transcription *in vitro*, THO/TREX complex recruitment is strongly 5' cap and splicing dependent [Cheng et al., 2006; Masuda et al., 2005]. REF/Aly binding can potentially increase the speed and efficiency of the export process [Valencia et al., 2008] but is not essential for export in metazoans [Gatfield and Izaurralde, 2002]. In addition is was proposed that the positioning of the THO/TREX complex at the 5'-end of spliced mRNAs influences direction of export, so that mRNAs exit the nuclear pore with

the 5'-end first to directly become engaged in translation [Valencia et al., 2008]. In addition to the THO/TREX complex, serine/arginine-rich (SR) and SR-like proteins can also function as mRNA export adaptors [Huang and Steitz, 2005]. These proteins are initially recruited to pre-mRNAs for splicing in a hyperphosphorylated state, and become partially dephosphorylated during the splicing reaction. Thus, it has been suggested that the export competence of the spliced messenger ribonucleoproteins (mRNP) is signaled by the phosphorylation status of the bound SR proteins [Huang and Steitz, 2005].

2.2.3 Determinants of mRNA half-life

In procaryotes the rapid synthesis and degradation of mRNA is essential for their capacity to adapt quickly to the environment. Transcripts in bacteria like *E.coli* live in the cytoplasm on average less then 5 minutes [Bernstein et al., 2004]. In eukaryotes, the dynamic range of transcript half-life is much bigger: housekeeping transcripts, from the β-globin gene for example, can be present for more than 10 hrs [Sharova et al., 2009] while TF-mRNAs are degraded relatively fast [Yang et al., 2003].

As described in section 2.2.1 most mRNAs acquire a 5' cap structure and a 3' polyA tail during co-transcriptional processing in the nucleus. A so called cap-binding complex induces the circularization of the transcript, which both, facilitates translation and protects it from degradation.

There are two general ways a transcript can be degraded: from the 3' or from the 5' end. From the 3' end the polyA tail gets shortened as soon as the transcript is exported to the cytoplasm. PolyA shortening is like a timer that counts down lifetime. When the polyA tail reaches a critical length, in mammals $\sim$25 nt, two pathways of degradation diverge: (I) Either exonucleases continue to shorten the transcript from the 3' end into the coding region or (II) the 5' cap is removed (decapping) and the exposed mRNA is rapidly degraded from the 5' end by the exonuclease Xrn1. Most eukaryotic RNA is actually degraded by both mechanism.

Usually, specific sequence properties of each transcript determine how fast the degradation occurs and thereby how long the mRNA is available in the cytoplasm to be subject to translation into protein. Especially 3' un-translated region (UTR) sequences often carry binding sites for proteins, which specifically enhance or slow down the rate of polyA shortening, decapping or 3'-5' degradation. At the same time translation itself regulates the stability of the respective mRNA: if ribosome and translation machinery are bound, degrading enzymes

are less likely to access and act on this transcript. (reviewed in Parker and Song [2004])
Apart from the two general ways of degradation there are cases where specific nucleases cleave the mRNA internally, which leads to rapid degradation. Transcripts which are degraded in this way, usually carry specific sequences in their 3' UTR, which serve as recognition sites for endonucleases to bind.

Sequence-specific mRNA repression

Post-transcriptional regulation is mediated by RBPs or small RNAs, so called, trans-acting factors, which bind to specific cis elements in UTRs of an mRNA. This binding can then influence mRNA degradation, sequestration, localization and translation. Most regulatory sequences bound by trans-acting factors, are located within the 3' UTR of an mRNA [Merritt et al., 2008; Stark et al., 2005].

Trans-acting factors and cis-acting elements

Although the 3' UTR in a long linear RNA molecule is quite distant from the cap, the closed loop structure, discussed above, brings both of these features into close proximity and thereby allows the 3' UTR to impact on translation initiation. Numerous cis elements located in the the 3' UTR have been described, however, only few reports, find regulatory sequences in the 5' UTR. For instance AU rich element (ARE) are found in mRNAs encoding for cytokines, interleukins and proto-oncogenes [Caput et al., 1986; Shaw and Kamen, 1986]. Several ARE binding proteins (ARE-BPs) have been identified, which tightly regulate the turnover of transcripts they bind to: While the CCCH tandem zing-finger protein tristetraprolin (TTP) promotes mRNA degradation [Lykke-Andersen and Wagner, 2005], the ELAV protein family member HuR, another ARE-BP, has a stabilizing effect on its target transcript [Fan and Steitz, 1998].
Proteins that bind to 3' UTR elements can influence the stability of the transcript in several ways. They can regulate mRNA transport within the cytoplasm or assemble repressive complexes which sequester the mRNA away from the translation machinery. Moreover, trans-acting factors may recruit mRNA decay enzymes, thereby inducing degradation. Besides RBPs, another group of important trans-acting factors are small regulatory RNAs, like

piRNAs and miRNAs.
In the following section I will describe miRNAs in more detail as they will be the most relevant trans-acting factor for my PhD thesis.

2.2.4 Transcript decay by MicroRNAs

miRNAs were first discovered in 2001 in *C.elegans* [Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001]. Since then, this species of small RNA became recognized as key regulators in gene expression, influencing a wide range of biological processes, post-transcriptionally, including cell proliferation, differentiation, metabolism and development (reviewed in Krol et al. [2010]).
Like mRNAs, miRNAs are initially transcribed by RNAP II in the nucleus, where they form pri-miRNA precursors, folded into a so called 'hairpin' structure. These precursors are processed by the endoribonuclease Drosha, yielding shorter 'hairipins', termed pre-miRNAs, which are subsequently exported to the cytoplasm by the export factor exportin 5. In the cytoplasm a second processing enzyme, Dicer, cuts the loop of the folded pre-miRNA and leaves a 22nt long double-stranded RNA. From this double-strand one, the mature miRNAs, is incorporated together with several RNA binding proteins into the miRNA induced silencing complex (miRISC). The miRISC locates its targets via basepairing between the loaded miRNA and the target 3' UTR and thereby represses mRNA expression. Key components of the miRISC, and crucial for target mRNA repression are the Argonaute and GW182 proteins, which interact with other proteins to affect translation initiation or recruit mRNA decay enzymes (reviewed in Krol et al. [2010]). Initially, it was believed that in animals miRNAs would affect gene expression mainly via translation inhibition, because the complementary region between the miRNA and its target mRNA is very short (6-8 nt), in contrast to the almost full complementarity in plants [Llave et al., 2002; Rhoades et al., 2002]. While a lot of progress was made understanding the biogenesis and function of miRNAs the actual mechanism that miRNAs use to regulate gene expression is subject to a controversy (reviewed in Huntzinger and Izaurralde [2011]). There are in principle two different views: (I) miRNAs function on the level of actual mRNA degradation or (II) they only inhibit translation of the target but leave the transcript intact (Figure 2.5). The latter mechanism, translational repression, has been suggested to occur in four different ways: inhibition of translation initiation, inhibition of translation elongation, premature termination

of translation, and co-translational protein degradation.

The first studies on miRNA-mediated repression mechanism in *C.elegans* suggested that

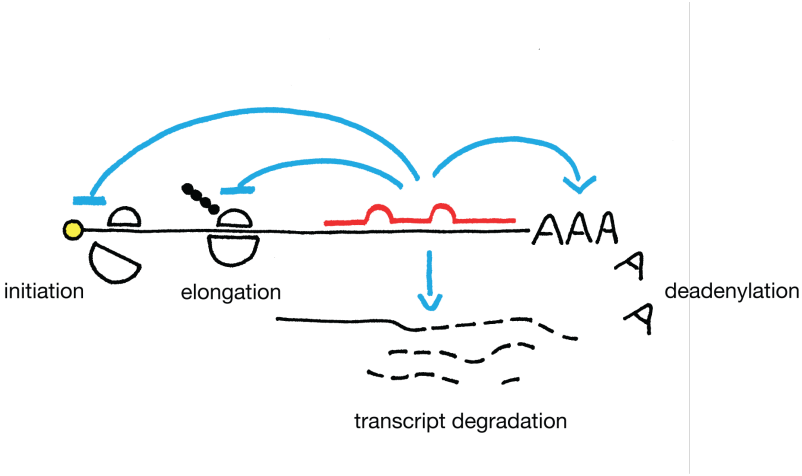


Figure 2.5: Proposed models of miRNA mediated gene expression silencing. A miRNA might act on different stages of gene expression: it might prevent transcription initiation or elongation or act to degrade the target mRNA by deadenylation and subsequent decay mechanisms.

the repression happens post-initiation, because protein expression of target mRNAs was inhibited while RNA could still be detected on polysomes [Maroney et al., 2006; Olsen and Ambros, 1999; Seggerson et al., 2002]. This could be either because ribosomes drop off from the transcript prematurely [Petersen et al., 2006] or because proteins are degraded co-translationally [Nottrott et al., 2006].

Contrasting studies showed the absence of miRNA targets from the polysomal fraction [Pillai et al., 2005] and concluded that translation is repressed already at the initiation step. This theory was supported by the observation that miRNA mediated silencing could be avoided if translation was driven by an internal ribosome entry site (IRES) [Iwasaki et al., 2009; Mathonnet et al., 2007]. The transcription initiation complex eIF4F, which binds polyA tail and cap, was actually observed to be affected because adding purified eIF4F continuously abrogated silencing [Ding and Grosshans, 2009].

This last finding may not be in conflict with the option that miRNA target supression acts on the level of transcript degradation. Because of the imperfect pairing of the miRNA with its target endonucleolytic cleavage it is unlikely, however the miRNA can direct its target to the cellular 5'-3' miRNA decay pathway, where the circular conformation of the transcript is broken up and progressive deandenylation takes place. The degradation theory is supported

by numerous evidences from specific miRNA-target pairs as well as transcriptome studies. The depletion of a miRNA lead to increased abundance of mRNAs with complementary target sites [Baek et al., 2008; Krützfeldt et al., 2005] and conversely, the introduction of a miRNA into the cell resulted in decreased levels of potential targets [Baek et al., 2008; Guo et al., 2010; Hendrickson et al., 2009; Lim et al., 2005]. In addition the depletion of any essential proteins of the miRNA biogenesis pathway had the same effect as deleting mature miRNAs: target mRNAs accumulated [Behm-Ansmant et al., 2006; Eulalio et al., 2009, 2007; Giraldez et al., 2006; Rehwinkel et al., 2005]. Without interfering with a cell, expression profiles of differentiating cells show anticorrelation of miRNA expression and target [Farh et al., 2005; Stark et al., 2005].

Studies employing quantitative mass spectrometry agreed that miRNAs have only a minor effect on protein level [Baek et al., 2008; Selbach et al., 2008]. Two more recent papers, which use translation profiling by monitoring polysome bound mRNA estimate that mRNA degradation explains 75-84% of miRNA-mediated changes in protein level [Guo et al., 2010; Hendrickson et al., 2009]. In summary, evidence for rapid mRNA degradation as the main mechanism of miRNA mediated regulation accumulates, which means that the effect of a miRNA on its targets should be measured on the level of mRNA abundance. This assumption will be important in the second part of my PhD project, when investigating the contribution of miRNAs to steady state mRNA level.

2.2.5 Readout of post-transcriptional events: mRNA half-life

After the mRNA gets transcribed and exported to the cytoplasm, the process of RNA degradation begins immediately. How fast a transcript is degraded is different for every mRNA as described in the previous sections. Depending on RNA sequence but also on the expression of interacting proteins a transcript will have a certain half-life, the time after which only half of the initial transcript will be existent in the cell. Transcript decay or degradation λ is indirectly proportional to half-life $t_{1/2}$,

$$t_{1/2} = \frac{ln(2)}{\lambda},$$

assuming an exponential decay process. Measuring abundance of a transcript at a time point t reflects the equilibrium between transcript synthesis and decay. To monitor only RNA decay, we therefore have to mask the synthesis process from our measurement. This

can be done in different ways:
The direct, however strongly invasive, method is to stop transcription in the cell. This can be done by arresting RNAP II by various chemicals, such as α-amanitin or actinomycin-D. From the moment of transcription arrest, no new transcript is being synthesized and one can measure the decrease of mRNA per gene over time. Typically, this time-course is not longer than a couple of hours because the RNAP II arresting chemicals also interfere with other cellular processes and may alter the speed of degradation [Dölken $\mathrm{et\ al.}$, 2008]. For fast dividing cells, such as ESC in our experiments, we even observe cell death after 8 hrs of actinomycin-D treatment. Nevertheless, this method has been widely used in genome-wide studies as it allows for a global quantification of the decaying mRNA pool by either microarray or RNA sequencing. Isolating RNA at each time point after transcription arrest from the exact same number of cells, results in a decreasing amount of total RNA obtained over time. This is precisely what we would like to monitor, however, both microarrays and sequencing technology require to use a specific amount of starting material (in this case RNA) for every experiment, which at time point t_0 can be obtained from half the amount of cells compared to $t_{1/2}$, where on average half of the mRNA is degraded. This can be solved by either 'filling up' the required RNA amounts by an artificial spike-in RNA or we can make use of the fact, that most of the RNA in a cell ($> 80\%$) actually comprises ribosomal RNA, which is known to have a long half-life ($\sim$ 5 days, Loeb $\mathrm{et\ al.}$ [1965]). In addition rRNA is transcribed by RNAP I, which is not inhibited by actinomycin-D. Consequently, we will not see rRNA decreasing during the time-course experiment of a few hours, however, the relative amount of mRNA in the RNA pool will decrease. Importantly, resulting microarray intensities from these experiments must not be normalized between arrays, as this would erase the signal of global mRNA decrease. For each transcript monitored on the array, one can infer a linear fit from the log transformed signal intensity depending on the time after transcription arrest. The slope of the regression line corresponds to the decay λ in the equation above and by plugging in the time interval of the experiment, one can obtain the half-life $t_{1/2}$ in hours.

Due to the side-effects of the transcription arrest, a less invasive, method has become state of the art measuring mRNA decay rates during the last years: metabolic labeling [Dölken $\mathrm{et\ al.}$, 2008; Rabani $\mathrm{et\ al.}$, 2011]. Here a 'label', for example a modified nucleotide, is added to the cell in excess for a certain time period, in which all mRNA synthesized will incorporate this label. One can then specifically separate labeled (newly synthesized) and

unlabeled (pre-existing) mRNA. With time the fraction of the labeled over unlabeled RNA increases until all pre-existing RNA is degraded and all mRNA is labeled. Quantitative measurement (microarray or RNA sequencing) is done between start of labeling and complete labeling for all three fractions: labeled, unlabeled and total RNA separately. Importantly, it is sufficient to do this measurement at one time point, because we know that at timepoint t_0 (before labeling) the ratio of $\frac{unlabeled}{total} = 1$. To calculate decay rates from the ratios obtained at this time point t_{0+x} use:

$$T_{1/2} = -t * \frac{ln(2)}{ln(1 - \frac{1}{1+\frac{(labeled/total)}{(unlabeled/total)}})},$$

again assuming exponential decay. Although, the advantage of this method is that the incorporation of a labeled nucleotide does not interfere with expression levels, a major downside is the IP based separation of labeled and unlabeled RNA. Depending on the labeling time this will enrich for a very small fraction of transcripts and is potentially subject to sequence biases. Further necessary purification steps add more potential steps for introduction of systematic errors. One has to be cautious when processing metabolic labeling data: The IP enriches for biotinylated labeled uridines (thio-U), the U frequency within a transcript will influence the enrichment, a newly transcribed mRNA with many Us will be more enriched than one with low U frequency even if both have been similarly transcribed. Therefore a U-normalization step is required before plugging in the $\frac{labeled}{total}$ fraction into the above formula.

Both methods measuring mRNA half-life have their individual shortcomings but unless one would have a reference of the 'real' mRNA half-life of each transcript it can not be decided which method is superior. A report which measures mRNA half-lives in parallel using both methods in the same system, shows a very low correlation meaning either one or both methods do not reflect actual decay rates. Assuming these experimental limitations, one has to take interpretations of mRNA half-life with a grain of salt, however, in theory the decay rate of a transcript should reflect its entire history from the moment it was transcribed, processed, exported and subject to decay or miRNA mediated silencing.

2.3 Introduction to the Theoretic Approach

The above two introductory sections were concerned with biological aspects that build the basis for my PhD Thesis. While this biological background is sufficient to raise the question of the thesis, a basic introduction to statistical methods, that I will use, is necessary to formulate the problem. In this section I will briefly introduce regression analysis with regard to the biological background of my PhD topic. This will help me to formulate the scope of my thesis in the following section 2.4.

2.3.1 Regression Analysis

Regression type problems were first considered in the 18th century to aid navigation. The method was almost exclusively used in physical sciences until later in the 19th century, where Francis Galton established the term 'regression to mediocracy' in 1875 and introduced r as the correlation between two variables x and y [Galton, 1890].
Galton used this definitions to explain a phenomenon called 'regression effect': the observation that sons of tall fathers tend to be tall but not as tall as their fathers and sons of short fathers tend to be short but not as short as their fathers. His work was later extended by Karl Pearson to a more general statistical context [Magnello, 1998]. In the 1950s and 1960s, economists used electromechanical desk calculators to calculate regressions and before 1970, it took up to 24 hours to receive the result from one regression. With the advent of high-speed computing regression methodology developed rapidly and as computing hardware improved the scope for this analysis has widened.

Simple Regression

In sections 2.1 and 2.2 I introduced transcriptional and post-transcriptional processes in the cell that play a role in gene expression regulation. Suppose we wish to forecast the abundance of a certain transcript in a cell, we now have plenty of information which factors are associated with variations in mRNA levels, chromatin, transcription, export, processing, degradation etc. For the time being let us restrict to one factor: call it RNA polymerase II occupancy at the promoter of the gene. Regression analysis with a single explanatory variable is termed *simple regression*.

We assume, possibly quite unrealistically, that mRNA level can be measured by a single attribute — RNA Polymerase II occupancy (RNAP). Initially in any regression study, one formulates a hypothesis about the relationship between the variables of interest, here, RNAP and mRNA level, based on, for example, mechanistic knowledge in the process of transcription. Thus, the tentative hypothesis is that higher levels of RNAP cause higher levels of mRNA, other things being equal. To investigate this we collect data from a number of genes in the genome, by RNA-sequencing and RNAP II-ChIP-seq. Because we have prior knowledge about the generation of the sequencing data, we know that it has to be logarithmically transformed before testing our hypothesis. We can now plot this information for all genes using a two-dimensional *scatter plot*, where each point represents one gene. The plot suggests that more RNAP indeed yields higher mRNA levels but at the same time the the relationship is not perfect. Regression analysis embraces the idea that other factors than RNAP influence mRNA levels. Thus the new hypothesis is that the mRNA level for each gene is determined by RNAP and an aggregation of omitted factors that we term 'noise'. The relationship can be written:

$$mRNA_i = \alpha + \beta RNAP_i + \epsilon,$$

where α is a constant, β the effect or 'coefficient' of RNAP, hypothesized to be positive and ϵ the 'noise' term reflecting other factors that influence mRNA level. The variable $mRNA_i$ called the *dependent variable* or *response* and $RNAP_i$ is the *independent or explanatory variable* or *predictor*. Note that the relationship between mRNA and RNAP is the equation for a line with an intercept α and a slope β. Regression estimates the line, which minimized the sum of squared errors (SSE), with error being the vertical distance of each gene from the regression line.

Multiple Regression

Plainly, mRNA levels, as described in previous section are affected by a variety of factors in addition to RNA polymerase occupancy, factors that were aggregated into the 'noise' term in the simple regression model above. *Multiple regression* allows additional factors (predictors) to enter the analysis separately so that the effect of each can be estimated. It is valuable for quantifying the impact of various simultaneous influences upon a single

response variable. For example histone marks are connected with transcription and may be incorporated in the regression. The modified model may be written:

$$mRNA_i = \alpha + \beta RNAP_i + \gamma H3K36 + \epsilon_i,$$

The task of estimating the parameters α, β, and γ is conceptually identical to the earlier task, in contrast we can no longer think of regression as choosing a line in a two-dimensional diagram. With two explanatory variables we need three dimensions, and instead of estimating a line we are estimating a plane. Multiple regression analysis is capable of dealing with an arbitrarily large number of explanatory variables, e.g. more histone modification measures may be included.

Another common statistic associated with regression analysis is the R^2, which will be used as an estimator of goodness of the model throughout the thesis. R^2 has a simple definition: it is equal to one minus the ratio of the sum of squared estimated errors (the deviation of the actual value of the dependent variable from the regression line, SSE_{fit}) to the sum of squared deviations about the mean of the dependent variable (SSE_{mean}).

$$R^2 = 1 - \frac{SSE_{fit}}{SSE_{mean}}$$

The R^2 statistic necessarily takes on a value between zero and one. A high value of R^2, suggesting that the regression model explains the variation in the dependent variable well, is obviously important if one wishes to use the model for predictive or forecasting purposes. The SSE about the regression line is a measure of the extent to which the regression fails to explain the response variable. Hence, the R^2 statistic is a measure of the extent to which the total variation of the response variable is explained by the regression.

Non-Linear Regression

In statistics, nonlinear regression is a form of regression analysis in which observational data are modeled by a function which is a non-linear combination of the model parameters and depends on one or more independent variables. The data are fitted by a method of successive approximations. In contrast to linear regression we can not estimate the optimal 'true' coefficients for each predictor. In non-linear regression the parameter fitting is an iterative 'try and error' process which terminates upon a stop criterion, for example if the SSE is below a certain threshold. In some cases where the relationship between the predictor

and the response variable is not linear but can be defined by another relation (exponential, logarithmic, trigonometric, power function...) one or both variables can be transformed to yield linear relation and use linear regression with the transformed variables. However, if the relation between the variables is more complex, one can employ non-linear regression. This way one might be able to catch plateau effects in biological measurements or other biases which are known to be technical. One should use non-linear modeling with caution because non-linear relationships between variables are much harder to interpret and it may be more useful to understand a biological process having a linear model with a higher SSE that is interpretable than a non-linear model, where the relationship between predictors is not clear.

2.4 Motivation, Idea and Scope of Thesis

"The formulation of the problem is often more essential than its solution which may be merely a matter of mathematical or experimental skill." [Albert Einstein]

All different cell types in a multicellular organism arise from one fertilized cell. During replication and differentiation the genetic information is static while differentiated cells show an enormous diversity in phenotype and function. This results from varying expression patterns of the genes in an organism's genome and the resulting protein pool in each cell, which is determined by the cell's gene regulation system.
Gene regulation is a multilayered process, which starts in the nucleus of the cell and continues in the cytoplasm (Figure 2.6). To regulate which gene is expressed at which time involves a complex interplay between proteins (transcription factors) that bind DNA in a sequence-specific manner as a genetic component, and the epigenetic state of the target sequence, defined at large by modifications of DNA and bound histones. In addition to this RNA synthesis determining steps, RNA degradation plays a role in setting up which transcripts will be available to the ribosome for translation. RNA binding proteins and particular small non-coding RNAs are well studies molecules mediating such post-transcriptional regulation.

When I started my PhD in 2008, genome-wide mapping of histone modifications switched from ChIP-chip technology to ChIP followed by deep-sequencing (ChIP-seq). The first high-resolution maps in mammals derived by ChIP-seq [Barski et al., 2007; Mikkelsen et al., 2007] together with deep sequencing studies of corresponding transcriptomes [Mortazavi et al., 2008; Wang et al., 2009] allowed to correlate mRNA expression with the epigenetic state of a certain cell type. This revealed a genome-wide contribution of active and repressive histone marks at promoter regions with transcription. At the same time the flood of new RNA sequencing data [Affymetrix ENCODE Transcriptome Project and Cold Spring Harbor Laboratory ENCODE Transcriptome Project, 2009] allowed detection of a large pool of RNA molecules previously masked by targeted microarray approaches and supported theories of pervasive transcription [Carninci et al., 2005; Pheasant, 2007; Taft et al., 2006] outside of protein coding-genes [Carninci, 2009; Core et al., 2008; Seila et al., 2008; Wang et al., 2009]. While both, epigenetic modifications of chromatin as well as regulatory RNAs, were

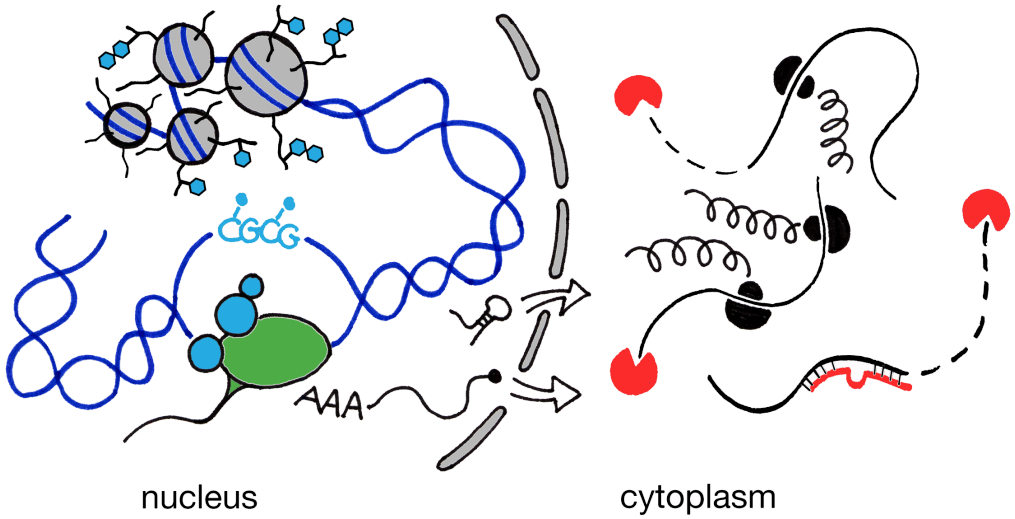


Figure 2.6: This sketch illustrates a simplified view of locally separated regulatory processes of RNA expression in an eukaryotic cell. Histone modifications, chromatin and transcription in the nucleus and different mechanisms of RNA degradation in the cytoplasm.

reported to be linked to mRNA expression and different phenotypes, there was no study comparing the contribution of these regulatory layers on a quantitative base.

To meet this challenge we made use of an *in vitro* differentiation system of ESC to terminal neurons [Bibel et al., 2007], where chromatin modification maps and transcription data was readily available from previous studies in the lab [Lienert et al., 2011; Mohn et al., 2008; Tiwari et al., 2012]. With this and public data in the same cell type [Mikkelsen et al., 2007] we initially identified the chromatin readouts most relevant for mRNA levels. We found that together with RNAP II occupancy and two other histone modification at the promoter region (H3K4me2 and H3K27me3), H3K36me3, a co-transcriptional histone mark, is most predictive for mRNA abundance. Hence, we generated high-resolution H3K36me3 maps by ChIP-sequencing in our *in vitro* differentiation system and used a regression model to integrate these maps with the other available ChIP-seq data in order to predict transcription. The idea of our study, is that given these chromatin-based transcription measures, it would be impossible to capture information from the post-transcriptional layer. Therefore any deviation between the chromatin-derived 'predicted transcription' and actual measured mRNA levels should be due to regulation that happens after the RNA is synthesized, meaning post-transcriptionally. In order to quantify the relative contributions of these layers of regulation

we primarily infer the explained variance of mRNA levels by the linear combination of histone marks and RNAP II occupancy and secondly, determine how much of the remaining variance might be explained by post-transcriptional processes. We aimed to quantify this effect of post-transcriptional regulation by measuring mRNA decay rates initially by transcription arrest in our system. Later we compared these results with metabolic labeling of RNA, which emerged as the method of choice to defined decay rates [Dölken et al., 2008]. Specifically, we wanted to describe the quantitative contribution of miRNAs to post-transcriptional decay of their respective target transcripts. To this end we inferred abundance of small RNAs throughout in vitro differentiation in our system by small RNA sequencing. We integrate the miRNA abundance with experimentally [Sinkkonen et al., 2008] and computationally predicted miRNA-target interactions [Gaidatzis et al., 2007] with the data derived on the chromatin level and compare their relative contributions to steady-state mRNA level as well as changes in mRNA abundance throughout differentiation.

Chapter 3

Results

3.1 Submitted Manuscript

Relative contributions of different regulatory layers to steady-state mRNA levels

Sylvia C. Tippmann[1,2,3], Robert Ivánek[1], Dimos Gaidatzis[1,3], Anne Schöler[1,2,3], Leslie Hoerner[1], Erik van Nimwegen[4], Peter F. Stadler[5,6,7,8,9,10,11], Michael B. Stadler[1,3,*], Dirk Schübeler[1,2,*]

* Corresponding Authors:
Michael B. Stadler (michael.stadler@fmi.ch), Dirk Schübeler (dirk@fmi.ch)

1. Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, 4058 Basel, Switzerland
2. University of Basel, Petersplatz 1 CH-4003 Basel, Switzerland
3. Swiss Institute of Bioinformatics, Maulbeerstrasse 66, 4058 Basel, Switzerland
4. Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, 4056 Basel, Switzerland
5. Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria
6. Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
7. Bioinformatics Group, Department of Computer Science, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
8. Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany
9. Fraunhofer Institut für Zelltherapie und Immunologie – IZI, Perlickstraße 1, D-04103 Leipzig, Germany
10. Center For Non-Coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark
11. Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

Subject Categories:	Chromatin and Transcription, RNA
Keywords:	histone modifications, chromatin, microRNA, RNA decay, transcriptional regulation
Running Title:	Prediction of transcription rate from chromatin
Character Count:	56.335

Abstract

Messenger RNA levels in eukaryotes are controlled by multiple consecutive regulatory processes, which can be classified into two layers: Primary transcriptional regulation at the levels of chromatin and secondary, co- and post-transcriptional regulation of the mRNA. To identify the individual contribution of these layers to steady-state RNA levels requires separate quantification. Using mouse as a model organism, we show that chromatin features are sufficient to model RNA levels but with different sensitivities in dividing versus post-mitotic cells. In both cases chromatin derived transcription rates explain over 80% of the observed variance in measured RNA levels. Further inclusion of measurements of mRNA half-life and microRNA expression data enabled the identification of a low quantitative contribution of RNA decay by either microRNA or general differential turnover to final mRNA levels. Together this establishes a chromatin based quantitative model for the contribution of transcriptional and posttranscriptional processes to steady-state levels of messenger RNA.

3.1.1 Introduction

Regulation of mRNA levels is a key mechanism that defines cell identity. Cellular homeostasis requires stable gene expression patterns, while differentiation events in metazoan development or responses to external stimuli involve resetting of the transcriptional program. During the lifespan of an mRNA from transcription over maturation, export, translation and decay, its activity and abundance is controlled by various mechanisms: histone modifications and DNA methylation determine the epigenetic state of the chromatin environment of a gene depending on the DNA accessibility the transcription machinery can bind and initiate transcription and thereby produce primary transcript at different rates [Bell et al., 2010; Segal and Widom, 2009]. This is modulated co-transcriptionally by splicing and poly-adenylation [Di Giammartino et al., 2011; Millevoi and Vagner, 2010; Nilsen and Graveley, 2010] and further regulated at the level of nuclear export. Once the mRNA is in the cytoplasm it is subject to further post-transcriptional processing, that can reduce the transcript level in a targeted manner. Two major post-transcriptional regulatory processes influencing the amount of mRNA molecules available for translation are general RNA decay and microRNA-mediated RNA interference. Single-gene experiments have provided examples of the involved regulatory mechanisms that include transcription factor binding but also what is currently referred to as epigenetic regulation. These summarize chromatin regulation of DNA

accessibility through active or repressive histone modifications [Kouzarides, 2007] or nucleosomal positioning [Kornberg and Lorch, 1999; Wyrick et al., 1999], transcriptional repression by DNA methylation of gene promoters [Bird, 2002; Eckhardt et al., 2006; Weber et al., 2007] and post- transcriptional regulation of RNA decay rates by non-coding small RNAs [Ambros, 2004]. Additionally, genome-wide studies successfully approximated mRNA levels with information of transcription factor binding and histone modification patterns at promoter proximal sequences [Cheng and Gerstein, 2011; Karlic et al., 2010; Ouyang et al., 2009]. mRNA abundance however, may be determined to different degrees by transcriptional and post-transcriptional events and the contribution of these layers may vary depending on how stable or how fast the expression change needs to be. At a quantitative level, there is only a limited understanding of the individual contributions of these regulatory layers. To understand these relations we abstract the many layers into two processes: primary regulation of synthesis or transcription on the level of chromatin and secondary, post- transcriptional degradation of mRNA. We assume that the change of mRNA level (dR/dt) depends linearly on mRNA synthesis and degradation,

$$\frac{dR}{dt} = tx_j[DNA] - d_j[RNA_j] \qquad (3.1)$$

where [RNAj] is the RNA concentration for gene j, $[DNA]$ is constant ($[DNA] = 1$), tx_j is the transcription rate and d_j is the degradation rate of gene j. For simplification, we initially assume the degradation rate to be constant, meaning independent of gene j. Therefore in steady state where $dR/dt = 0$, the RNA concentration of gene j is proportional to transcription and degradation rates of gene j. Subsequently when we investigate the contribution of post-transcriptional regulation, we allow d_j to depend on gene j (see supplemental information section 1 for details). Consequently, we can estimate the individual contribution of transcription and mRNA degradation, or mRNA decay, by correlating them to mRNA levels respectively. Here we explore quantitatively how a prediction of transcription based on chromatin characteristics relates to mRNA levels and how such an approach can quantify changes in mRNA abundance that occur during the course of cellular differentiation. We ask if pluripotent and differentiated cells differ in their regulatory behaviors, potentially relating to differences in cell cycle and the ability to set and propagate epigenetic marks or a different usage of posttranscriptional processes. As a biological model we use mouse stem cells that we differentiate into a highly pure neuronal population through a defined progenitor state [Bibel et al., 2007]. We focus our analysis on pluripotent embryonic stem cells (ES) and post-mitotic glutamatergic neurons (TN). In order to quantitate the contribution of different regulatory processes to observed mRNA levels, we cre-

ated a linear model for each cell type based on various measures from transcriptional and post- transcriptional layers. In these models, a measure that is a strong correlate of transcription is expected to be highly predictive of mRNA levels. We found that genome-wide measures of histone modifications and polymerase occupancy alone – measures which stand for the transcriptional layer of regulation - allowed accurate prediction of mRNA levels and explained most of the observed experimental variation in steady-state mRNA levels. In addition we measured transcript half-life and microRNA abundance in these cells, representing the post-transcriptional layer of regulation, and identified only a minor contribution to the determination of mRNA levels.

3.1.2 Results

Histone marks are predictive of transcription rate

In order to separately quantify transcriptional and post-transcriptional processes on a genome-wide level, we estimated transcription rates for individual genes. Transcription rate is a function of multiple factors: transcription factors bind influenced by the chromatin environment and concordantly determine the rate of transcription. We use chromatin correlates of transcription as readout, which can be measured genome-wide in a robust way by chromatin immunoprecipitation followed by deep sequencing (ChIP- seq). We created genome-wide maps for RNA polymerase II (Pol-II) and tri- methylation of lysines 4, 27 [Lienert et al., 2011; Tiwari et al., 2012] and 36 in histone H3 (H3K4me2, H3K27me3 and H3K36me3) in both dividing and post-mitotic cells (see materials and methods for details) and investigated the distribution of sequence reads along the gene body in reference to gene activity defined by mRNA abundance of representative transcripts (see supplemental information section 2 for details). Figure 3.1 summarizes average distributions of these marks for non-overlapping genes: Pol-II, H3K4me2 and H3K27me3 are located around the promoter of the gene [Boyer et al., 2006; Guenther et al., 2007; Mohn et al., 2008; Rahl et al., 2010; Young et al., 2011] while H3K36me3 is distributed over the gene body [Barski et al., 2007; Bell et al., 2007; Mikkelsen et al., 2007; Pokholok et al., 2005] steadily increasing within the first 2 kilo bases downstream of the transcription start site (TSS). Based on these observations, which are in accordance with previously published models [Bell et al., 2007; Edmunds et al., 2008; Hon et al., 2009; Vakoc et al., 2006], we selected the regions to quantify these marks for individual genes. While most of the histone marks have a functional impact close to the TSS, the abundance of H3K36me3 throughout the gene body is notably by far the most informative measure for transcription (Figure 3.2, supplemental information section 3 for details), as could be expected from its mechanistic link to tran-

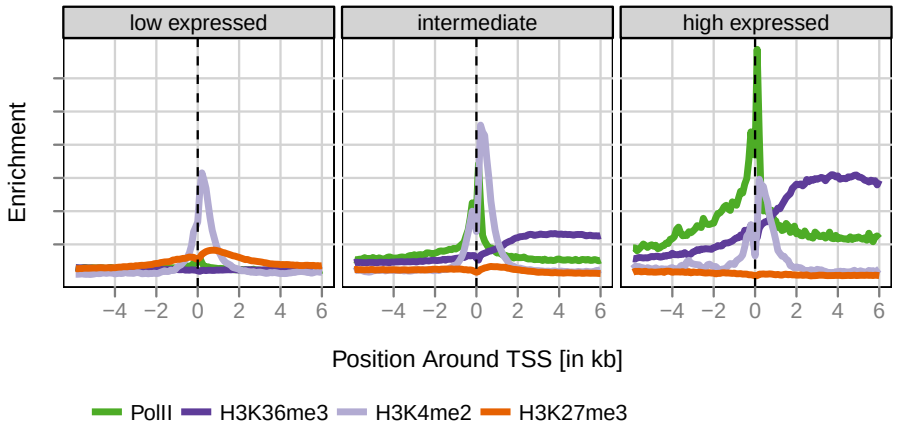


Figure 3.1: Using histone marks and RNA polymerase II to model mRNA levels. Metagene plot showing the distribution of histone marks along the gene body of genes aligned at their TSS with low, intermediate and high expression levels.

scription: H3K36me3 chromatin mark is set by a complex that associates with the active elongating RNA-polymerase-II [Joshi and Struhl, 2005; Keogh et al., 2005; Kizer et al., 2005; Krogan et al., 2003; Li et al., 2003, 2002; Pokholok et al., 2005; Strahl et al., 2002; Sun, 2005; Xiao et al., 2003; Yuan et al., 2009].

Using these marks as regressors (Figure 3.2) we infer a linear model, where mRNA measured by deep sequencing is the response variable (combining poly-A RNA and ribosomal-depleted RNA sequencing, for details see materials and methods) (Figure 3.2). The coefficients assigned to each of the regressors by the linear model reflect their function as active or repressive histone mark (sign of the coefficients) and their contribution to explaining transcription (absolute value of the coefficients). The correlation (controlled by a 2-fold cross-validation) between observed and predicted mRNA abundance is 0.92. This means that 84.6% of the observed differences in mRNA levels (variance) can be explained by this model (Figure 3.2, black bar) – exclusively based on measures from the transcriptional layer.

The remaining 15.4as measurement noise. While post-transcriptional effects could be explained by a more sophisticated model that includes additional experimental data from the post-transcriptional layer (see below), the technical and biological measurement noise cannot be predicted and thus defines an upper limit of prediction accuracy. We went on to partition this sum, by (A) estimating the noise, and thereby the maximum variance which can be explained by our regressors and (B) assigning relative contributions of two major post-transcriptional processes – microRNA-

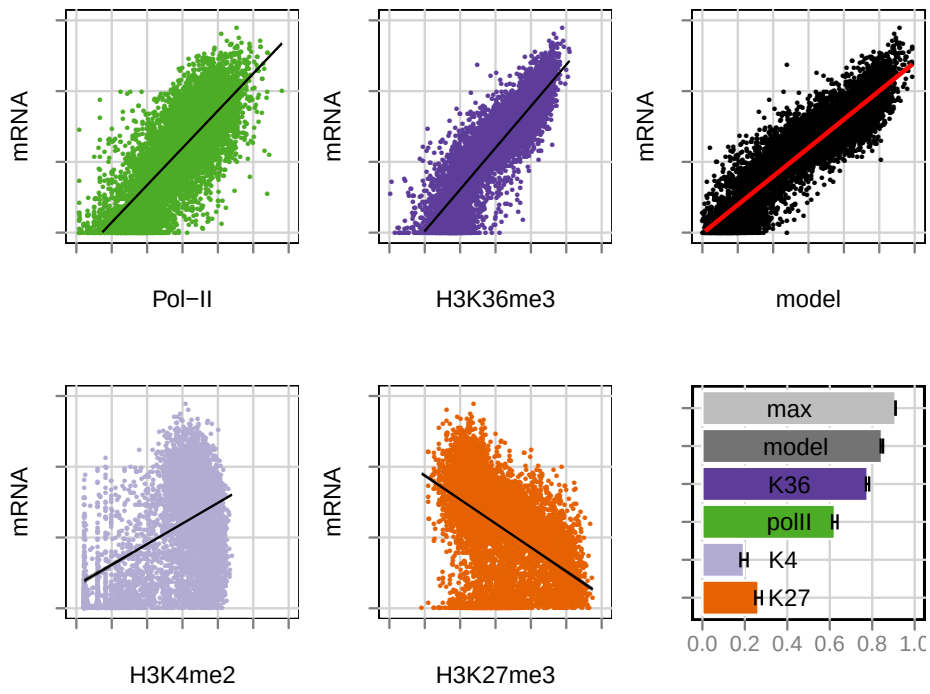


Figure 3.2: Using histone marks and RNA polymerase II to model mRNA levels. Scatter plot of RNA polymerase II (Pol-II, green) and three histone marks H3K36me3 (dark blue), H3K4me2 (light blue), H3K27me3 (orange) versus mRNA levels on the vertical axis. The number of reads aligned to either gene body (H3K36me3, mRNA) or at the TSS (H3K4me2, H3K27me3, Pol-II) is shown in logarithmic scale. Predicted transcription rate combining the four measures in a linear model versus mRNA level. Axes as in B. Bar plot showing the fraction of total variance in mRNA levels that is explained by each single histone mark, Pol-II occupancy or a linear combination of them (black). The maximally explainable variance (grey) is limited by the amount of measurement noise (see supplemental information section 4 for details).

mediated degradation and RNA decay - to final mRNA levels.

Estimating the upper bound of explained variance in RNA levels

Fluctuations in biological systems limit the explainable variance of mRNA through the variability between biological replicates. In order to determine how much of the remaining unexplained variance is due to such biological variability and measurement noise versus actual post- or co-transcriptional processes, we estimated the maximum variance to be explained given the variability in the data. In the linear model noise originates from both measurements of mRNA levels and measurements of chromatin marks. Since we use multiple regres-

sor measurements that each have independent noise, their individual noise adds up, which in turn sets the limits of explainable variance. To estimate its upper bound we follow the theory of noise propagation to calculate model noise based on replicates of RNA-seq and ChIP-seq experiments (see supplemental information section 4 for details). This approach sets the maximal explainable variance in mRNA levels to 91% (Figure 3.2, light-grey bar). The variance in RNA levels, which remains to be explained, is therefore the difference between this maximal to be explained variance and the variance that is already explained by the linear model using transcriptional information. In the case of ESC this difference is 6.4%.

The effect of degradation on steady-state mRNA level

Having estimated transcription rate and an upper bound for explainable variance we next explored the remaining 6.4% unexplained variance. We assumed that genes with lower measured RNA level than predicted by the transcription measures are degraded more rapidly than average due to post-transcriptional down-regulation of their transcript. To test this hypothesis we inferred the RNA decay rates of genes by measuring their abundance in a time-course after inhibition of transcription with actinomycin D (see methods and supplemental information section 5 for details). Transcript abundance was determined in replicates at 0, 1, 2, 4 and 8 hours after inhibition of transcription, but not later in order to reduce secondary effects due to long chemical treatment. From the degradation slope we calculate the RNA half-life according to Sharova et al. [Sharova et al., 2009], summarized in Figure 3.3 and Figure 3.4. The high correlation between biological replicates allowed us to extrapolate half-life times up to 20 hours and thus to include genes with slower decay rates. In accordance with a previous study in mouse ES cells [Sharova et al., 2009] we observe a mean half-life of around 8 hours with a distribution tailed towards longer half-lives (Figure 3.4). The extremely short-lived RNAs mostly belong to the class of non-polyadenylated genes, which are not protected from degradation (supplemental information section 5, supplemental figure 7). These genes are expected to show lower mRNA levels compared to other genes with the same predicted transcription rate. Indeed, short-lived RNAs are deviating negatively from the linear fit. This is particularly visible in the shift in the boxplots in Figure 3.4 in the 40-100% transcription bins, while there are hardly short lived genes in the low- transcribed bins (supplemental information section 5, supplemental figure 8). The degree to which the half-life explains additional variance in mRNA levels can be quantified by the correlation of the half-life with the residual of the linear fit. This correlation is 0.3; meaning of the 6.4% unexplained variance of mRNA levels in the transcriptional model, mRNA half-life

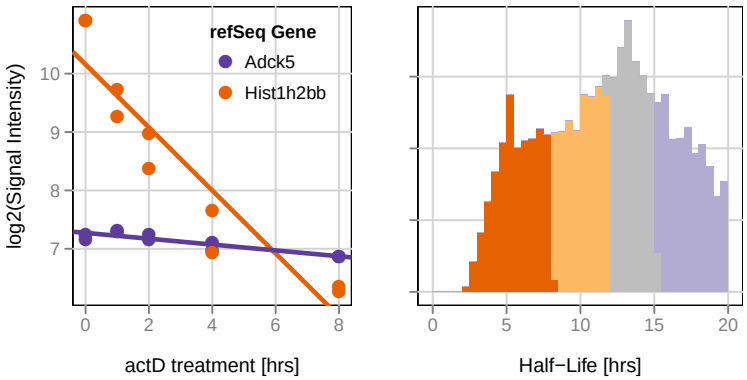


Figure 3.3: Effect of RNA half-life on mRNA levels. (A) Example genes (short-lived histone gene Hist1h2bb (orange) and the stable gene Adck5 (purple)) illustrating the inference of mRNA half-lives from expression data. Data points correspond to measured mRNA abundance at various time points after inhibition of transcription (time zero). (B) Half-life distribution of RefSeq genes with estimated mRNA decay rates. Half-lives of very stable genes were set to 21 hrs (the maximal inferable half-life given the experimental setup). (see supplemental information section 5 for details)

explains $0.32 = 9\%$ (supplemental information section 5, supplemental figure 9). As an alternative we can simply include the half-life as an additional feature in the linear model and infer the correlation to the measured mRNA levels again. Indeed the explained variance increases from 84.6% to 86%. To test if this result is independent from the experimental approach to measure half-life we next employed metabolic labeling of mRNA [Dölken et al., 2008; Rabani et al., 2011; Schwanhäusser et al., 2011]. After a short pulse of a modified ribonucleotide newly synthesized and pre-existing mRNA fractions are separated to determine their differential abundance in order to estimate a decay rate. This method has the advantage of not interfering with the transcriptional program, as does actinomycin D, and thus is less likely to cause indirect effects [Dölken et al., 2008]. However it is limited to a single time point. With this different approach we obtained a highly similar additional contribution of mRNA half-life to overall mRNA levels (total explained variance 85.9%; see supplemental information section 6 for details). Notably, the variance in mRNA levels explained by transcript half-life measures alone is between 11 and 12%, for thioU and actinomycinD derived half-lives respectively. This sets a theoretical upper bound for the relative contribution of transcript half-life to mRNA levels and further supports the observation of a minor contribution of mRNA half-life to steady-state levels inferred by different methods.

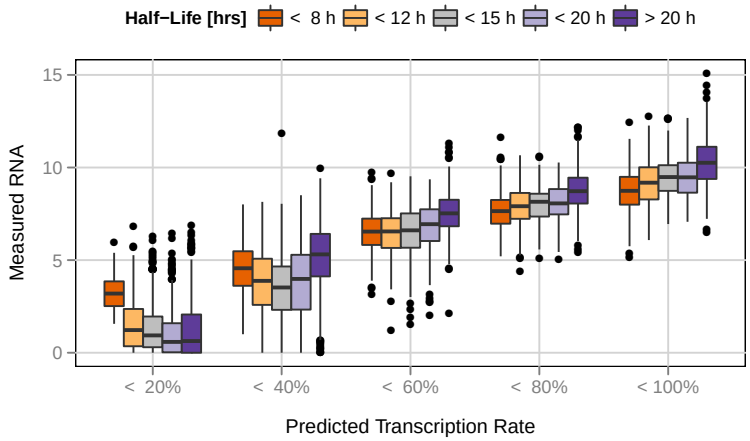


Figure 3.4: Effect of RNA half-life on mRNA levels. Genes are classified into five equal groups according to predicted transcription rate (0-100%), and within each group measured mRNA levels are shown as box-plots separately for genes with different mRNA half-life (color-coded). Within a transcription group with a sufficient number of genes, short-lived genes show less measurable mRNA than long-lived genes. In the two low transcription bins (0- 40%) mRNA levels are less well modeled and they are depleted of short-lived mRNAs. (see Supplemental Figure 7 for illustration)

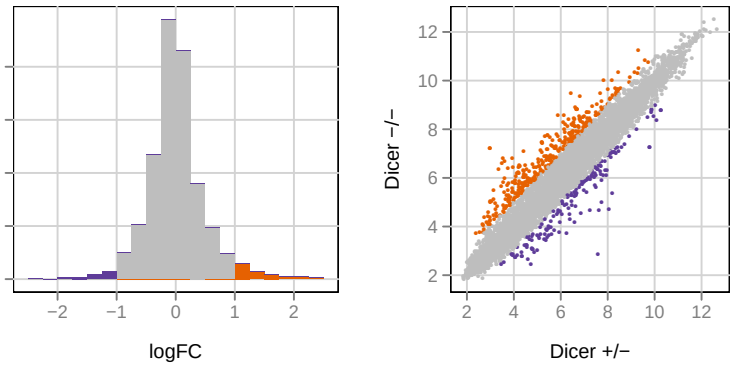


Figure 3.5: Effect of targeting by microRNAs on mRNA levels. (A) Scatter plot of $Dicer^{+/-}$ versus $Dicer^{-/-}$ ES cells inferred by microarray measurement [Sinkkonen et al., 2008]. Genes with increased mRNA levels in $Dicer^{-/-}$ are enriched for putative microRNA targets (orange), while genes with decreased mRNA levels are possibly affected by secondary effects (purple). (B) Distribution of the log fold-change (logFC) between $Dicer^{-/-}$ and $Dicer^{+/-}$.

The effect of microRNAs on steady-state mRNA level

Next we investigated whether we can attribute part of the observed mRNA half-life to the activity of microRNAs that target selected messages for degradation. To define the percentage of variance in mRNA

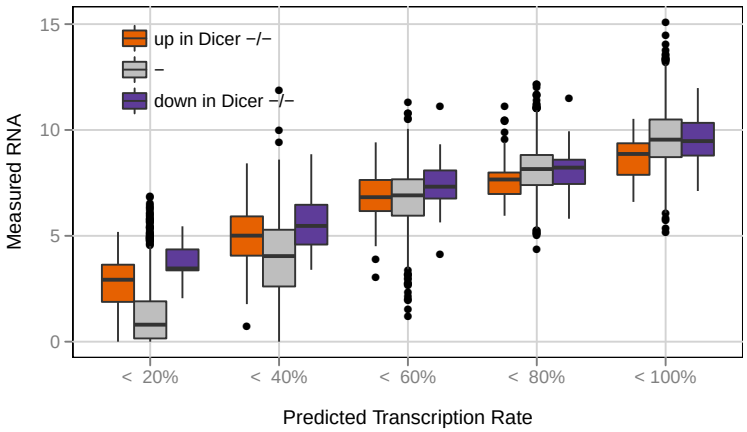


Figure 3.6: Effect of miRNAs on mRNA levels. Genes are classified into five equal groups according to predicted transcription rate (0-100shown as box-plots separately for genes with different log fold-change between $Dicer^{-/-}$ and $Dicer^{+/-}$ (color coded). Within the same transcription group, putative microRNA target genes (orange) show insignificantly different mRNA levels as non-target genes (pvalue=0.303).

level that can be explained by microRNA mediated degradation requires the identification of mRNAs that are regulatory targets of microRNAs. This can be attempted by identifying mRNAs bound to proteins involved in the RNAi pathway (such as Ago-IP [Beitzinger et al., 2007; Chi et al., 2009; Hafner et al., 2010; Landthaler et al., 2008] or by calculating the enrichment for motifs complementary to the microRNA within 3'-untranslated regions (UTR) of mRNAs [van Dongen et al., 2008] or by predicting targets using a combination of sequence, structure and conservation of the microRNA and its target mRNA site [Enright et al., 2003; Gaidatzis et al., 2007; Krek et al., 2005; Lewis et al., 2003; Rehmsmeier et al., 2004]. These methods share a high false-positive rate since actual targets are not only defined by sequence complementarity alone, but by additional sequence and structural constraints and other modulating factors that are currently only poorly understood. In order to circumvent these potential limitations we initially based our definition of microRNA-targets on mRNAs that increase in expression in ES cells that lack microRNAs due to a genetic deletion of the gene encoding Dicer [Hutvágner et al., 2001; Murchison et al., 2005]. An increased mRNA abundance in $Dicer^{-/-}$ cells suggests that these transcripts had been under negative control by microRNAs in wild-type ES cells (Figure 3.5). Consequently we correlate fold-changes in mRNA abundance between $Dicer^{+/-}$ and $Dicer^{-/-}$ cells with the deviation from the model in the linear fit (also referred to as 'residual of the linear fit'). This did not reveal a relationship between negative residuals indicative for post-

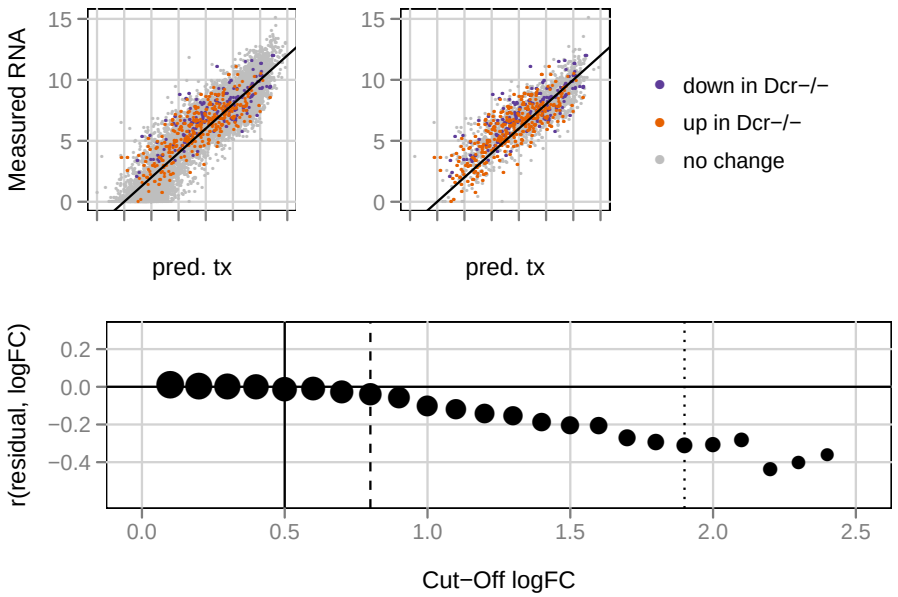


Figure 3.7: Focus on high-confidence microRNA target genes. (A) All RefSeq genes are classified into three color-coded groups according to up-regulated in Dicer-/- (orange), down-regulated in $Dicer^{-/-}$ (purple) and unchanged (grey). (B) Subset of all the genes in (A), where the absolute log fold-change (logFC) between $Dicer^{-/-}$ and $Dicer^{+/-}$ is higher than 0.5. This subset contains likely targets and non-targets. (C) Pearson correlation (r) between the residual of the linear model and the logFC between $Dicer^{-/-}$ and $Dicer^{+/-}$ as a function of cut-off in absolute logFC between $Dicer^{-/-}$ and $Dicer^{+/-}$. A logFC cut-off of zero corresponds to (A), and a cut-off of 0.5 (solid vertical line) corresponds to (B). Correlations are shown for subsets of genes for logFC cut-offs incremented in 0.1 intervals. The point size illustrates the number of genes at each cut-off. At 0.8, the subset contains 1000 genes (dashed line), the subset at 1.9 contains 100 genes (dotted line). Increasing logFC cut-offs select higher-confidence microRNA-target genes that can explain the residual of the linear fit increasingly better.

transcriptional regulation and the likelihood of an mRNA being a microRNA target (correlation between fold-change upon Dicer KO and residual is r = 0.01; Figure 3.6, supplemental information section 7). Importantly however, it has been shown that expression changes of mRNAs upon removal of all microRNAs in $Dicer^{-/-}$ cells are relatively small in general (2-fold) [Babiarz et al., 2008]. It is thus conceivable that such small effects are not detectable in the population of all mRNAs that consist of targets and non-targets, changing their expression both, through direct effects caused by the lack of microRNAs and indirect effects unrelated to microRNAs. To test this hypothesis we directly compared high-confidence targets (based on fold-change in abundance) with non- targets (Figure 3.7, B). We stepwise increase the cut-off applied

to the change in mRNA levels upon Dicer KO to define microRNA targets, thereby selecting a smaller and smaller subgroup and inferred for each of these subgroups the correlation of residual and fold-change (Figure **??**). In these groups of higher confidence microRNA-targets, we can detect a negative correlation with the residual (Figure **??**, dotted line, supplemental information section 7, supplemental figure 14). We thus conclude that genes that are likely microRNA targets have indeed less detectable transcript than expected based

Transcriptional and posttranscriptional regulation in dividing versus post-mitotic cells

Having established that chromatin and bound polymerase are highly predictive of mRNA levels in rapidly dividing stem cells we next asked if the same trend is observed in post-mitotic neurons that have exited the cell cycle. Consequently we differentiated stem cells first into neuronal progenitors (NPs), which show reduced proliferation and further into terminal neurons, which do not divide. Similarly to the analysis in ES we determined globally the abundance of mRNA, microRNA, Pol-II and of several histone marks and rebuild the linear model. This revealed that at all three stages chromatin data are comparably predictive for mRNA levels (Figure 3.8). To compare post-transcriptional contribution between cell-types we also derived mRNA half-life datasets at the TN stage. Including mRNA half-life in TN as regressor in the linear model increased explained variance (r2) of mRNA in TN about 1%, from 79% to 80%, revealing an equally low contribution of mRNA degradation in neurons as the one observed in dividing stem cells. Together this suggests that there is no general change in regulatory contributions once stem cells have exited the cell cycle and, in this particular case, gain neuronal functions. Having defined the relation between chromatin measures, RNA decay and mRNA abundance at individual cell states we next asked whether changes in transcription or changes in degradation between cell states are equally predictive for changes in mRNA levels. We fitted the linear model using the differences in measurements between two cell types, which reveals that changes in chromatin can indeed predict 67% of the change in mRNA levels. Similarly changes in transcript half-life can explain 1% of the remaining variance (see supplemental information section 9 for details). This illustrates that the experimental measurements in combination with the applied analytical approach enable quantification of the relative contributions of transcription and degradation to changes in mRNA levels.

Influence of cell division on the information content of transcription- coupled chromatin marks

H3K36me3 is set by a histone methyltransferase which interacts specifically with the

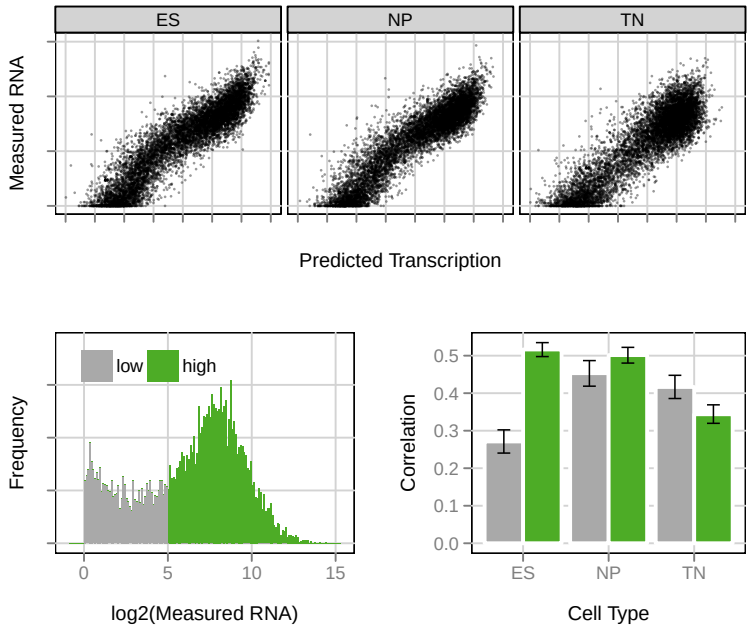


Figure 3.8: H3K36me3 explains most of the variance in mRNA level. Scatter plot of predicted transcription rate versus measured mRNA level for the ES, NP and TN. (B) Distribution of mRNA levels in ES, categorized into low and high expression groups. (C) Correlation (r) between H3K36me3 and mRNA for genes in expression groups from (B) in ES, NP and TN. The correlation of H3K36me3 with mRNA level differs between dividing cells and post-mitotic TN cells: In diving cells (ES and NP), it is best for high expressed genes, while in the post-mitotic TN, it is best for low expressed genes.

elongating RNA-polymerase-II [Joshi and Struhl, 2005; Keogh et al., 2005; Kizer et al., 2005; Krogan et al., 2003; Li et al., 2003, 2002; Pokholok et al., 2005; Strahl et al., 2002; Sun, 2005; Xiao et al., 2003; Yuan et al., 2009]. As a consequence H3K36me3 accumulates with repeated rounds of transcription explaining why this mark can not only predict sites but also rate of transcription [Barski et al., 2007; Bell et al., 2007; Buratowski and Kim, 2011; Edmunds et al., 2008; Mikkelsen et al., 2007; Pokholok et al., 2005; Wagner and Carpenter, 2012]. In dividing cells new nucleosomes that are not H3K36 trimethylated are deposited during genome replication. This is expected to dilute the prevalence of H3K36 methylation while this modification should further accumulate in non-dividing cells. In turn rate of cell division might influence the ability to predict mRNA levels from this modification. A potential accumulation of H3K36me3 in non-dividing cells could lead to higher sensitivity to predict transcription at weakly expressed genes and, in case all available residues are modified, to saturation and reduced predictive power at highly expressed genes. To

test the hypothesis of different H3K36me3 signal in dividing versus non-dividing cells we group genes according to their mRNA abundance into low and high expressed and correlate their mRNA levels with the abundance of the transcription coupled mark H3K36me3 along the gene body (Figure 3.8). In the dividing cell types ES and NP this mark shows highest predictive power for highly expressed and reduced sensitivity for lowly expressed genes. However in post-mitotic neurons there is a clear shift: in these cells predictability is now highest for low expressed genes in comparison to highly expressed genes. This is fully compatible with a model whereby chromatin modifications such as H3K36me3 integrate transcriptional activity over time and that the resulting signal is diluted with every cell division. In turn the sensitivity range changes in non-dividing cells, where signal for H3K36me3 accumulates above detection threshold for lowly expressed genes but also saturates for highly expressed genes.

Regulatory differences between tissue-specific and housekeeping genes

Genes can be classified according to their expression characteristics between cell types and tissues. Figure 3.9 shows a histogram of the number of tissues with detectable mRNA abundance (log2 intensity > 7) for the same set of genes studied in 72 tissues and cell types profiled in the SymAtlas project (Su et al., 2004). This reveals a clear bimodal distribution where genes show either widespread activity (expressed in most samples, also referred to as "housekeeping" genes) or selective activity in only up to five samples (also referred to as "tissue-specific"). This global behavior is also evident in the stem cell to neuron differentiation that we study here, where genes with widespread activity according to SymAtlas are enriched for genes that are expressed in both cell types, while tissue-specific genes tend to be expressed in either one or none of the two studied cell types (p-value $<$ 2.2e-16, see supplemental information section 10 for details). Importantly previous studies already noted that these two classes of genes differ in their regulation: housekeeping genes are mostly under the control of CpG rich promoters, while tissue-specific genes show a high frequency of CpG poor promoters [Mohn $\mathrm{et\ al.}$, 2008; **?**]. These two classes of genes are differentially occupied by histone modifications [She $\mathrm{et\ al.}$, 2009], show different exon density [Eisenberg and Levanon, 2003; Vinogradov, 2004] and differ in 3'UTR length and sequence composition making them unequal targets for microRNAs [Stark $\mathrm{et\ al.}$, 2005]. To ask if these classes of genes also differ in the relative regulatory contribution of transcriptional and posttranscriptional layers we compared the predictability of mRNA levels for tissue-specific and housekeeping genes using an identical linear model approach as described above. In this model, tissue-specific genes show more

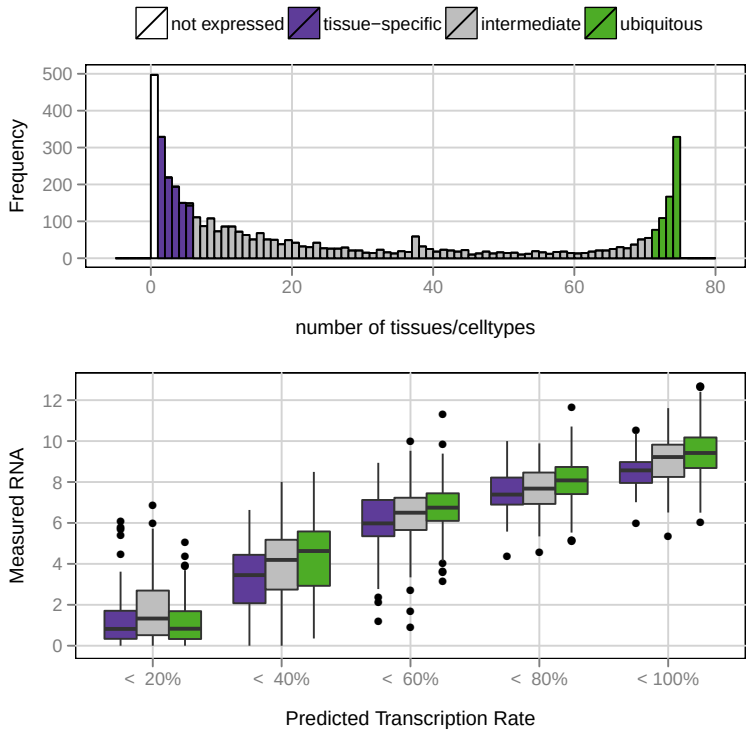


Figure 3.9: Post-transcriptional regulation in tissue-specific and ubiquitously expressed genes. (A) Histogram of the number of cell or tissue types with detectable expression of the analyzed genes. Genes are grouped in tissue-specific (expressed in 1-5 tissues, purple), intermediate (grey), and ubiquitously expressed (expressed in 70 or more tissues, green). (B) Genes are classified into five equal groups according to predicted transcription rate (0-100for genes with different tissue expression (as in (A), color- coded). At a given level of transcription tissue-specific genes have on average less measured mRNA than ubiquitously expressed genes, suggesting that the degree of post-transcriptional regulation is higher in tissue-specific genes.

negative deviation from the fit, corresponding to observed mRNA levels being lower than predicted based on the transcriptional features (Figure 3.9). We conclude that tissue-specific genes are more prominently controlled by post-transcriptional regulation than housekeeping genes.

3.1.3 Discussion

In our study we tried to quantify the relative contribution of transcriptional and post-transcriptional regulation to mRNA levels. We show that tri-methylation of lysine 36 of histone H3, a chromatin modification that is set co- transcriptionally, provides a quantitative measure of the process of RNA synthesis. We built a linear model that com-

bines H3K36 tri-methylation with other histone marks and Pol-II occupancy to predict transcription and to relate it to mRNA levels. This reveals a high correlation between predicted transcription based on chromatin and actual mRNA abundance in both dividing pluripotent cells and terminally differentiated neurons suggesting that transcription and mRNA levels are tightly linked at different cellular stages. These findings are consistent with two recent studies comparing direct measures of transcription with mRNA abundance [Rabani et al., 2011; Schwanhäusser et al., 2011]. Furthermore we investigated the predictive power of histone marks towards changes in mRNA levels between the two cell types and find similarly that transcription is also the main determinant when looking at genes that change their mRNA levels. Following the determination of transcriptional contribution we investigated the contribution of different post-transcriptional processes by extending the model to include information on microRNA targeting and transcript half-life. The effect of transcript half-life is indeed detectable on a genome-wide scale explaining minor additional variance of mRNA levels. Notably however we can also detect this minor contribution when we look at the predictive power of half-life towards changes in mRNA levels from ES to TN. Reliable reproduction of the effect of degradation for changes in mRNA levels suggests that the method to measure half-life is sensitive. Moreover, this supports that degradation indeed plays a small but measurable role in determining mRNA levels and changes. Targeted degradation of mRNA by the action of microRNAs affects actual half-lives of mRNAs [Guo et al., 2010]. Importantly however we could not detect the actual effect of microRNA at a genome-wide scale, but only in a subset of high-confidence microRNA targets. This precludes correct quantification of the contribution of microRNA regulation to total mRNA decay. However, when focusing only on those genes that are highly upregulated in cells that lack Dicer we observe that microRNAs can explain about 2.25% of the residual variance. Extrapolating this contribution to all genes as a fraction of the total measured mRNA decay effect, we can estimate that microRNAs contribute between 2.5 and 25% to the total mRNA decay. This effect is compatible with the notion that microRNAs generally cause small changes in mRNA abundance [Babiarz et al., 2008; Sinkkonen et al., 2008]. At the same time we foresee that the inherent complexity in correctly predicting microRNA targets leads to an underestimation of the actual effect. The relatively low contribution of post-transcriptional regulation on the mRNA levels and changes shows that that the lion's share of regulatory contribution is at the level of mRNA synthesis and predictable from chromatin alone. It is important to note that the identified quantitative contribution (the fraction of explained variance), while important

for understanding the regulatory principles, does not translate to functional relevance and thus should not be taken as a measure for biological importance. For example, the $Dicer^{-/-}$ cells used here to identify microRNA-targets lack the ability to differentiate into neurons. The low quantitative contribution of post-transcriptional processes is however compatible with the model that these mostly function in fine-tuning mRNA levels rather than functioning as on-off switches [Mukherji $\mathrm{et\ al.}$, 2011]. Our study shows that chromatin is highly predictive of transcriptional output, in particular methylation of lysine 36 of H3, a mark that is set throughout the gene body and depending on the elongating polymerase. Most other histone marks that are involved in transcription occur primarily at promoters and, such as K4 methylation of CpG islands, can even occur at a subclass of promoters without activity of the linked gene, which in turn limits their predictive power [Weber $\mathrm{et\ al.}$, 2007]. Interestingly H3K36me3 is a far better predictor than RNA polymerase itself. We believe that this reflects the fact that the histone mark is stable once it is set, while the polymerase rapidly elongates and thus is only present at the gene at low frequency. While it is inherently difficult to directly compare the performance of H3K36me3 with direct labeling approaches for ongoing transcription we note that the correlation between H3K36me3 and steady-state mRNA levels is higher at all three cell states than at recent reports using alternative approaches like GRO-seq ($r^2 = 0.62$ [Min $\mathrm{et\ al.}$, 2011]). One likely explanation for the high predictive power of H3K36me3 is that it increases with every round of transcription, which in turn means that it can eventually saturate, when all possible lysines are methylated. In dividing stem cells such saturation is not observed, likely due to the "dilution" of modified histones that occurs at every S-phase during genome duplication in addition to the general turnover of nucleosomes [Deal and Henikoff, 2011; Wirbelauer $\mathrm{et\ al.}$, 2005]. In post-mitotic cells however we indeed observe such saturation at highly expressed genes. At the same time the accumulation of signal increases the sensitivity for the detection of weakly expressed genes, which in the linear model compensates for the reduced predictability at highly expressed genes.

Cell Culture

Wild-type embryonic stem cells (129Sv-C57Bl/6) were cultured and differentiated as previously described (Bibel et al., 2007; Mohn et al., 2008).

Chromatin Immunoprecipitation (ChIP)

Cells were cross-linked in medium containing 1as described before (Mohn et al., 2008), starting with 70µg of chromatin and 5µg of the following antibodies: anti-dimethyl-H3K4 (Upstate, no. 07-030 (Lienert et al., 2011; Tiwari et al., 2012), anti-trimethyl-H3K36 (Abcam ab9050), anti-trimethyl- H3K27 (Upstate, no. 07-449) (Lienert et al., 2011; Tiwari et al., 2012) anti- RNA-polymerase-II (Santa Cruz Biotechnology, no. SC899) (Lienert et al., 2011; Tiwari et al., 2012). Chromatin was sonicated for 10 cycles of 30 sec using a Diagenode Bioruptor. Precipitated DNA was subjected to next generation sequencing.

Next generation sequencing

5 to 10 ng of precipitated DNA was prepared for Solexa Sequencing as described (Mikkelsen et al., 2007). Briefly, ChIP DNA was ligated to adapters and ligation products of about 250 bp were gel purified on 1.5 18 PCR cycles. DNA sequencing was carried out using the Illumina/Solexa Genome Analyzer II (GA2) sequencing system. In addition 2 lanes of non-enriched chromatin from ES cells were sequenced and pooled to serve as an input/background to calculate the enrichment of reads obtained from ChIP-seq experiments. The raw .srf and .wig files are accessible at GEO GSE33252 (reviewer link)

Genomic coordinates

The July 2007 M. musculus genome assembly (NCBI37/mm9) provided by NCBI `http://www.ncbi.nlm.nih.gov/genome/guide/mouse/` and the Mouse Genome Sequencing Consortium `http://www.sanger.ac.uk/Projects/M_musculus/` was used as a basis for all analyses. Annotation of known RefSeq transcripts was obtained from UCSC.

Read filtering, alignment and weighting

Low-complexity reads were filtered out based on their dinucleotide entropy as follows: For each read, the dinucleotide entropy was calculated according to the formula $H = \sum_i f_i log(f_i)$, where fi is the frequency of dinucleotide i in the read and the sum is over all dinucleotides (i from 1 to 16). The read was filter out if its H was less than half the dinucleotide entropy of the genome, typically removing less than 0.5% of the reads in a given sample. Alignments to the mouse genome were performed by the software bowtie (version 0.9.9.1) [Langmead et al., 2009] with parameters -v 2 -a -m 100, tracking up to 100 best alignment positions per query and allowing at most two mismatches. To

track genomically untemplated hits (e.g., exon-exon junctions or missing parts in the current assembly), the reads were also mapped to an annotation database containing known mouse sequences (microRNA from ftp://ftp.sanger.ac.uk/pub/mirbase/sequences/13.0, rRNA,snRNA, snoRNA and RefSeq mRNA from GenBank http://www.ncbi.nlm.nih.gov/sites/entrez, downloaded on July 16, 2009, tRNA from http://lowelab.ucsc.edu/GtRNAdb/ and piRNA from NCBI (accessions DQ539889 to DQ569912). In that case, all best hits with at most two mismatches were tracked. Each alignment was weighted by the inverse of the number of hits. In the cases where a read had more hits to an individual sequence from the annotation database than to the whole genome, the former number of hits was selected to ensure that the total weight of a read does not exceed one. All quantifications were based on weighted alignments. For generation of wiggle files samples were normalized for library size first and files were generated with a window size of 100 bps.

RNA-Sequencing

Poly-A-RNA-seq: RNA from ES cells, NP cells and TN was isolated using the Trizol (Invitrogen). The sequencing libraries were prepared according to mRNA-Seq Sample Preparation Guide (Illumina) starting from 1µg of total RNA and using oligo dT primers for selection of polyadenylated mRNAs. The libraries were sequenced on an Illumina GA II analyzer. Ribosome-depleted-RNA-seq: RNA was isolated from ES, cells NP cells and TN using Trizol (Invitrogen) followed by depletion of ribosomal RNA, starting with 2µg of total RNA and following the instructions of Ribo-Zero Kit (Epicentre). Strand specific RNA libraries were prepared according to pre-release version of the Directional mRNA-Seq Library Preparation guide (Illumina) and sequenced on an Illumina GA II analyzer. Reads were mapped to the Mus musculus transcriptome and normalized to transcript length and sequencing library size. The raw .srf and .wig files are accessible at GEO GSE33252

Small RNA sequencing

RNA of ES, NP and TN was isolated in triplicates from cell culture with mirVanaTM microRNA Isolation Kit (AM1560) according to the kit instructions. Small RNA was prepared for sequening with Illumina Small RNA Sequencing Kit (FC-102-1009) following the Small RNA Sample Prep v1.5.0 protocol.

Linear model to predict transcription rate

We used R (Team, 2011) and the function lm() to fit a linear model to describe transcription rate. For every gene we selected a representative transcript of median length. Only transcripts, which did not overlap with alternative transcripts with different TSS or transcripts in antisense di-

rection, were kept for further analysis (supplemental information section 2). ChIP-seq reads of Pol-II, H3K4me3 and H3K27me3 were mapped to the TSS (+/- 500 bp) of the representative transcript. H3K36me3 was mapped to 4 different regions along the gene-body: (i) exons within first 2kb of the transcripts, (ii) introns within the first 2kb of the transcripts, (iii) exons located 2kb downstream from the TSS, (iv) introns located 2kb downstream from the TSS, (supplemental information section 3). Input chromatin sequencing reads were mapped to the whole gene body and used as an additional regressor to account for amplification and sequencing biases caused by the DNA sequence itself. These 7 regressors were fitted to mRNA levels as response value (mean read count of poly-A-enriched RNA-sequencing and strand-specific-sequencing) with 2-fold-cross validation. The squared pearson correlation coefficient corresponds to the explained variance in the response variable (Achen, 1982).

Transcript half-life measurement

ES cells and TN of two independent biological replicates were treated with actinomycin D as previously described (Sharova et al., 2009). RNA was isolated from an equal number of cells with Trizol at 1,2,4, and 8 hrs after treatment. 100ng of extracted total RNA was amplified using the Ambion WT Expression kit (Ambion) and the resulting sense-strand cDNA was fragmented and labeled using the Affymetrix GeneChip WT Terminal Labeling kit (Affymetrix). Affymetrix GeneChip arrays were hybridized following the GeneChip Whole Transcript (WT) Sense Target Labeling Assay Manual (Affymetrix) with a hybridization time of 16h. The Affymetrix Fluidics protocol FS450-0007 was used for washing. Scanning was performed with Affymetrix GCC Scan Control v. 3.0.0.1214 on a GeneChip® Scanner 3000 with autoloader. Subsequently arrays were normalized with RMA, without in between normalization to preserve absolute mRNA abundance. Decay slope of every transcript was inferred with a linear model and only transcripts with reliably inferable slopes (R>=0.4) were kept for further analysis. Transcripts half-lives were calculated from the mRNA abundance over time according to (Sharova et al., 2009) (see supplemental information section 5 for detailed description). The raw .CEL files and a table with normalized expressions are accessible at GEO GSE33252. To confirm our results obtained by actinomycinD treatment we infer mRNA half-life by metabolic labeling of nascent RNA adapted from the protocol described in [Dölken et al., 2008]. RNA was isolated with trizol, using 30µg RNA (a final concentration of 120ng/µl) for the biotinylation, followed by 2 chloroform/IAA extractions on the bio tagged RNA. Non-denaturated RNA is used in the IP with Dynabeads M-280 Streptavidin (112.06D, Invitrogen) for pull down (50µl /30µg RNA), followed by one elution step with DTT. See supplemen-

tal information section 6, for experimental details, analysis and results. The raw .CEL files and a table with normalized expressions are accessible at GEO GSE33252

Data Accession

All the data used in this study is accessible at GEO in the superseries GSE33252
Reviewer Link to GEO superset GSE33252: `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=jzchtusugkqcwbo&acc=GSE33252`

Acknowledgements

We thank members of the Schübeler lab for feedback on the manuscript. SCT is supported by a predoctoral fellowship of the Boehringer Ingelheim Foundation. Research in the laboratory of D.S. is supported by the Novartis Research Foundation, by the European Union (NoE EpiGeneSys FP7- HEALTH-2010-257082, LSHG-CT-2006-037415), the European Research Council (ERC-204264) and the Swiss initiative in Systems Biology (SystemsX.ch).

Author Contributions

S.C.T., M.B.S., D.G. and D.S. conceived the study. R.I. performed and analyzed strand specific RNA sequencing. A.S. performed H3K36me3 ChIP experiments. L.H. performed the metabolic labeling experiments. E.v.N. performed and supervised the estimation of noise. P.F.S., M.B.S. and D.S. supervised the study. S.C.T., M.B.S. and D.S. wrote the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

3.2 Supplemental Information

3.2.1 Definition of the model

As a general model of production and degradation contribution to mRNA abundance we can formulate:

$$\frac{dR}{dt} = tx_j[DNA] - d_j[RNA_j] \tag{3.2}$$

Importantly all measures of mRNA abundance and chromatin readouts of transcription are log transformed to be able to use them in a linear regression.
Therefore tx is log10(transcription rate of gene j) and d is log10(degradation rate). $[DNA]$ and $[RNA]$ are both concentrations, where $[RNA]$, log10(RNA abundance of gene j), depends on the gene j and $[DNA] = 1$.

At equilibrium $\frac{dR}{dt} = 0$ and we can write:

$$tx_j[DNA] = d_j[RNA_j] \tag{3.3}$$

Because $[DNA] = 1$ we can write:

$$tx_j = d_j[RNA_j] \tag{3.4}$$

Therefore, if we are speaking about one cell type, where mRNA concentrations are not changing, we are in an equilibrium. In this case the transcription rate tx_j is proportional to RNA concentration $[RNA_j]$.
Note that for the first part, talking about transcriptional regulation, we assume the degradation rate to be independent of the gene j and therefore, for the first part d_j is a $constant$. In the second part, when we introduce post-transcriptional regulation into the model, we allow d_j to be dependent on the gene j and actually infer the gene dependent degradation rates experimentally.

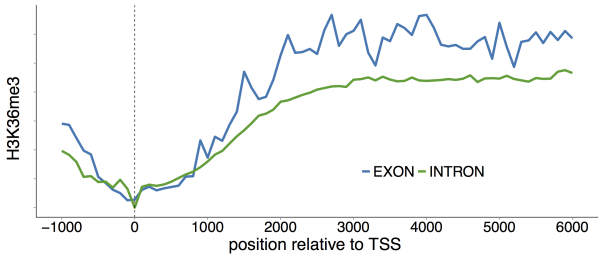
3.2.2 Selection of representative transcripts

For each RefSeq annotated gene in the mouse genome, a representative transcript of median length was selected. For the whole analysis described in the paper we only used RefSeq transcripts, which do not overlap with an anti-sense transcript and do not overlap with any transcript containing an alternative transcription start site. About half of the mouse genes fulfill these criteria (10.000 genes).

Supplemental Figure 3.1: Scheme illustrating transcript selection for the analysis performed in the paper.

Regions to infer regressors for the linear model

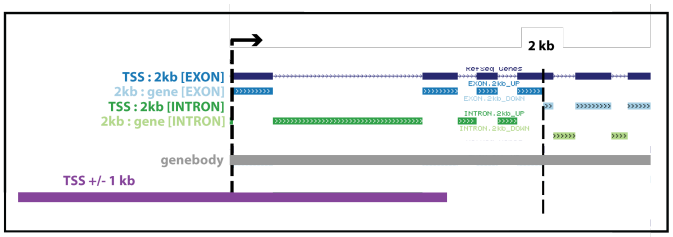
Sequencing reads from ChIP-seq experiments where mapped to different regions, depending on where the histone modification of interest is most predictive for mRNA levels. H3K27me3, H3K4me2 and pol-II were mapped to the TSS while H3K36me3 was mapped to 4 independent regions due to the distribution pattern of this modification (Figure 2).



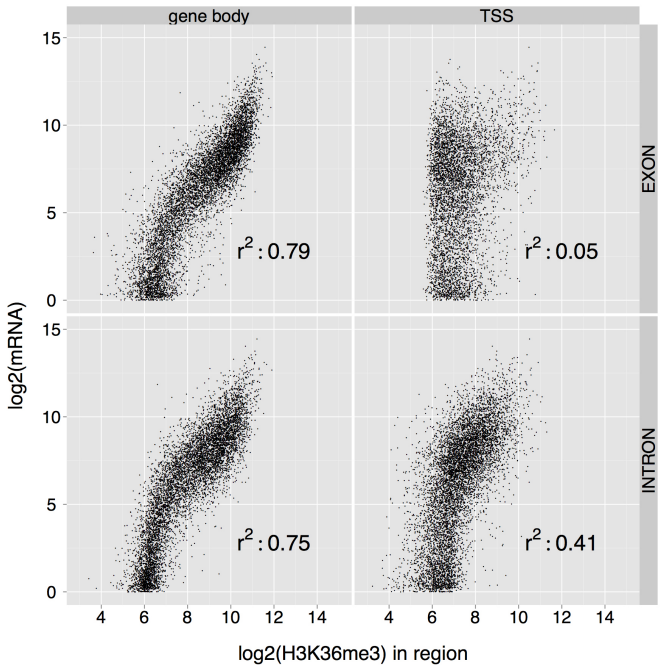
Supplemental Figure 3.2: Enrichment of H3K36me3 along the first 6 kb into the transcript. X-axis shows position relative to TSS, y-axis shows enrichment over input. The reads were separately mapped to exonic (blue) and intronic (green) regions. Enrichment in exons is generally higher than in introns, however because exons are shorter, they bear less total reads leading to a more noisy signal in the metagene plot.

H3K36me3 is set by the elongating RNA polymerase. The signal increases over the first 2 kb starting from TSS and remains throughout the gene body. Exons and introns have different H3K36me3 levels, possibly due to their different sequence compositions. We

account for this locally different H3K36me3 patterns by separating the signal in 4 different bins (Figure 3).



Supplemental Figure 3.3: Illustralion of regions relative to TSS used to map ChIP reads: Pol-II, H3K27me3 and H3K4me2 were mapped to TSS (purple). H3K36me3 was mapped separately to TSS proximal and genebody, to exons (blues) and introns (greens) respectively.



Supplemental Figure 3.4: H3K36me3 mapped to the 4 different regions within the genes (Figure 3) and correlated to the mRNA levels of the respective gene. H3K36me3 levels over the exons located 2kb upstream of the TSS to the end of the gene body are most predictive for mRNA levels.

3.2.3 Estimation of error in the linear model

Estimation of Error in the Linear Model in one Cell Type

The Problem

We want to model mRNA levels as linear function of the levels of 3 histone marks and RNA polymerase II occupancy at the corresponding gene's locus. To both fit and test this model, we measured mRNA expression, Pol2 occupancy, and the 3 histone marks in duplicate during differentiation of mouse embryonic stem cells into post-mitotic glutamatergic neurons. Importantly, a separate biological replicate differentiation was used for each measurement. In this note we use the variation across replicate measurements to estimate the noise on our estimates of mRNA and chromatin mark levels, and derive the maximal fraction of the observed variation in mRNA levels that could potentially be explained by the model, i.e. taking into account variation that is due to noise.

The linear model and its relative error

For the quantification of mRNA levels we count the reads over annotated refSeq genes. For histone marks H3K27me3, H3K4me2 and Pol-II-IPs we count the reads in a 1kb region around the TSS. H3K36me3 is measured in 4 separate regions: exonic and intronic TSS proximal region (0-2kb downstream of TSS) and exotic and intronic gene body region. All values are log2 transformed (pseudo count=1).

Let m_i denote the log2 transformed level of mRNA i, and let $h_{c,i}$ denote the log2 transformed level of histone mark c at gene i. We fit a linear model of the form

$$m_i = c + \sum_h \alpha_h h_{c,i}, \tag{3.5}$$

where c is a constant and α_h are the linear coefficients that we estimate when fitting the model. $c = \langle m \rangle - \sum_h \alpha_h \langle h_c \rangle$, where $\langle m \rangle$ is the mean mRNA level. At a given time point, the expression levels m_i show a variation that is given by

$$\mathrm{var}(m) = \frac{1}{N} \sum_{i=1}^{N} (m_i - \bar{m})^2 \,, \tag{3.6}$$

where N is the total number of genes and $\bar{m}$ is the average mRNA level

$$\bar{m} = \frac{1}{N} \sum_{i=1}^{N} m_i. \tag{3.7}$$

We now want to compare the variance $\mathrm{var}(m)$ with the average squared-deviation (the 'error') between the model and the true mRNA levels. This error is defined as

$$D^2 = \frac{1}{N}\sum_{i=1}^{N}\left(m_i - c - \sum_c \alpha_c h_{i,c}\right)^2. \tag{3.8}$$

The fraction of the variance f that is explained by the model can now be defined as

$$f = \frac{\mathrm{var}(m) - D^2}{\mathrm{var}(m)}. \tag{3.9}$$

Note that for a perfect model $D^2 = 0$ so that $f = 1$, and for a model that just predicts $m_i = \bar{m}$, i.e. just the average for every gene, we have $f = 0$.

Measurement and biological replicate noise

In equation (3.5), the quantities m_i and $h_{c,i}$ denote the 'true' mRNA and chromatin mark levels at a particular stage of differentiation, which can be thought of as the mean levels in the population of cells, averaged over a large number of biological replicates. However, we do not have direct access to these levels, we only have duplicate measurements from different experimental replicates. As a consequence, part of the deviations between the *measured* mRNA levels and the predicted levels in terms of the *measured* chromatin mark levels will be due to deviations between the true and measured levels.

Let m_i^1 and m_i^2 denote the duplicate measurements of the mRNA level of gene i. These values will differ from the 'true' mRNA level m_i by some unknown amount ϵ_i, i.e.

$$m_i^1 = m_i + \epsilon_i^1, \tag{3.10}$$

and similar for m_i^2. Note that the deviation ϵ_i^1 includes both biological 'noise' from variations in levels across the biological replicates, as well as measurement noise. Note also that, per definition, the expectation value of the deviation is zero

$$\langle \epsilon_i^1 \rangle = 0. \tag{3.11}$$

The size of the noise is characterized by the variance of the deviations, i.e. we define

$$\sigma_i^2 = \langle (\epsilon_i^1)^2 \rangle = \langle (\epsilon_i^2)^2 \rangle. \tag{3.12}$$

In our model we will allow different genes i to have different sized variations across the replicates.

For the chromatin marks we similarly write

$$h_{i,c}^1 = h_{i,c} + \epsilon_{i,c}^1, \tag{3.13}$$

and

$$\sigma_{c,i}^2 = \langle (\epsilon_{i,c}^1)^2 \rangle = \langle (\epsilon_{i,c}^2)^2 \rangle. \tag{3.14}$$

A key assumption that we will make is that, the values of the deviations ϵ_i^j and $\epsilon_{i,c}^j$ are all mutually independent. This is a highly reasonable assumption since these measurements derive from separate biological replicates. That is, all covariances are zero, e.g.

$$\langle \epsilon_i^j \epsilon_{i,c}^k \rangle = 0, \tag{3.15}$$

for all j, k and c.

Estimating the noise levels

We can use the replicate measurements to both estimate the true values m_i and $h_{i,c}$, as well as estimate the size of the noise σ_i^2 and $\sigma_{i,c}^2$. In particular, given the measured values m_i^1 and m_i^2, the expected value of m_i is simply given by the mean

$$\langle m_i \rangle = \frac{m_i^1 + m_i^2}{2}, \tag{3.16}$$

which we will also refer to as $\bar{m}_i$. Similarly, the expected variance $\langle \sigma_i^2 \rangle$ is given in terms of the difference of the measurements, i.e.

$$\langle (m_i^1 - m_i^2)^2 \rangle = \langle (\epsilon_i^1 - \epsilon_i^2)^2 \rangle = \langle (\epsilon_i^1)^2 \rangle + \langle (\epsilon_i^2)^2 \rangle + 2\langle (\epsilon_i^1 \epsilon_i^2) \rangle = 2\sigma_i^2, \tag{3.17}$$

where we have used that the covariance is zero, i.e. $\langle \epsilon_i^1 \epsilon_i^2 \rangle = 0$. From this we have the estimate

$$\sigma_i^2 = \frac{1}{2} \langle (m_i^1 - m_i^2)^2 \rangle. \tag{3.18}$$

In complete analogy, we have for the estimated chromatin mark levels

$$\langle h_{i,c} \rangle = \frac{h_{i,c}^1 + h_{i,c}^2}{2} = \bar{h}_{i,c}, \tag{3.19}$$

and for the noise levels of the chromatin marks

$$\sigma_{i,c}^2 = \frac{1}{2} \langle (h_{i,c}^1 - h_{i,c}^2)^2 \rangle. \tag{3.20}$$

Finally, we can define the average noise levels across all genes as

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2, \tag{3.21}$$

and

$$\sigma_c^2 = \frac{1}{N} \sum_{i=1}^{N} \sigma_{i,c}^2. \tag{3.22}$$

Estimating the variance in mRNA levels

We cannot directly measure the variance $\mathrm{var}(m)$ in mRNA levels, but we can calculate the observed variation V^2 in measured mRNA levels. Defining

$$\bar{m} = \frac{1}{N}\sum_{i=1}^{N} \bar{m}_i, \tag{3.23}$$

we have

$$V^2 = \frac{1}{N}\sum_{i=1}^{N} (\bar{m}_i - \bar{m})^2 \,. \tag{3.24}$$

Writing $\bar{m}_i$ in terms of the true level m_i and the deviations due to replicate fluctuations and measurement error, we have

$$\langle V^2 \rangle = \frac{1}{N} \langle \left(m_i - \bar{m} + \frac{\epsilon_i^1 + \epsilon_i^2}{2} \right)^2 \rangle. \tag{3.25}$$

Using the fact that the cross-correlations are zero we have

$$\langle V^2 \rangle = \mathrm{var}(m) + \frac{1}{2}\sigma^2. \tag{3.26}$$

As a technical note, we have here neglected the deviation of the measured average mRNA level $\bar{m}$ from the true average level $(1/N)\sum_i m_i$. Taking this into account would lead to corrections of order $1/N$, which are negligible in practice.

We can thus estimate the true variance $\mathrm{var}(m)$ in terms of the measured variance V^2 as

$$\mathrm{var}(m) = V^2 - \frac{1}{2}\sigma^2. \tag{3.27}$$

Estimating the error of the model

To estimate the error in the model, we compare the estimated mRNA levels $\bar{m}_i$ with the predicted ones based on the chromatin marks $\bar{h}_{i,c}$. We define the average squared-deviation as

$$T^2 = \frac{1}{N}\sum_{i=1}^{N} \left(\bar{m}_i - c - \sum_c \alpha_c \bar{h}_{i,c} \right)^2 . \tag{3.28}$$

We can write the expectation of this quantity $\langle T^2 \rangle$ in terms of the true deviation between model and mRNA levels for each gene i, i.e.

$$D_i = m_i - c - \sum_c \alpha_c h_{i,c}, \tag{3.29}$$

and the noise due to biological replicate variations and measurement errors. That is, we have

$$\langle T^2 \rangle = \frac{1}{N} \sum_{i=1}^{N} \langle \left(D_i + \frac{\epsilon_i^1 + \epsilon_i^2}{2} - \sum_c \alpha_c \frac{\epsilon_{i,c}^1 + \epsilon_{i,c}^2}{2} \right)^2 \rangle. \tag{3.30}$$

Using that the covariances between all different noise terms are zero, and making the final assumption that there are no correlations between the true deviations D_i and the noise levels, e.g.

$$\langle D_i \epsilon_{i,c}^j \rangle = 0, \tag{3.31}$$

we find that all cross-terms are zero and we have

$$\langle T^2 \rangle = D^2 + \frac{1}{2}\sigma^2 + \frac{1}{2} \sum_c \alpha_c^2 \sigma_c^2. \tag{3.32}$$

Since we can measure T^2, and we have above derived expressions for the noise levels σ^2 and σ_c^2 in terms of the duplicate measurements, we can thus estimate the true deviation D^2, i.e.

$$D^2 = T^2 - \frac{1}{2}\sigma^2 - \frac{1}{2} \sum_c \alpha_c^2 \sigma_c^2. \tag{3.33}$$

Putting it all together, we finally estimate the fraction of explained variance as

$$f = \frac{V^2 - T^2 + \frac{1}{2} \sum_c \alpha_c^2 \sigma_c^2}{V^2 - \frac{1}{2}\sigma^2}. \tag{3.34}$$

Error in the model of the expression changes across two cell types $\Delta TN, ES$

Instead of explaining absolute mRNA levels we also want to use a linear model to predict the changes in expression levels between the embryonic stem cell and neuron stage. Specifically, we will model the log fold-change $\Delta_{i,m}$ in mRNA expression level of each gene i. All values are log transformed, 'TN-ES' therefore stands for a log ratio, $log(TN) - log(ES) = log(TN/ES)$.

$$\frac{1}{N} \sum_i \left(m_i(TN) - m_i(ES) - \bar{m}(TN) + \bar{m}(ES) \right)^2 \tag{3.35}$$

Linear model of expression changes

$$\Delta_{i,m} = m_i(\mathrm{TN}) - m_i(\mathrm{ES}), \tag{3.36}$$

in terms of the changes $\Delta_{i,c}$ in chromatin marks

$$\Delta_{i,c} = h_{i,c}(\mathrm{TN}) - h_{i,c}(\mathrm{ES}) \tag{3.37}$$

using a linear model. That is, in complete analogy with our previous linear model we write

$$\Delta_{i,m} = \tilde{c} + \sum_c \tilde{\alpha}_c \Delta_{i,c}. \tag{3.38}$$

Measurement and replicate noise

We use the same replicate measurements to estimate the changes $\Delta_{i,m}$ and $\Delta_{i,c}$. Importantly, individual measurements are coming from separate biological replicates so that our assumption that the cross-correlation of deviations are expected to be zero still holds.

We thus estimate the change $\Delta_{i,m}$ by averaging over the duplicate measurements, i.e.

$$\bar{\Delta}_{i,m} = \frac{m_i^1(\mathrm{TN}) + m_i^2(\mathrm{TN}) - m_i^1(\mathrm{ES}) - m_i^2(\mathrm{ES})}{2}, \tag{3.39}$$

and similary for the chromatin marks

$$\bar{\Delta}_{i,c} = \frac{m_i^1(\mathrm{TN}) + m_i^2(\mathrm{TN}) - m_i^1(\mathrm{ES}) - m_i^2(\mathrm{ES})}{2}. \tag{3.40}$$

We estimate the noise levels in our estimates $\bar{\Delta}_{i,m}$ and $\bar{\Delta}_{i,c}$ at both time points using the replicates exactly as described above. That is, we have

$$\langle \sigma_{i,m}^2(\mathrm{ES}) \rangle = \frac{1}{2} \left(m_i^1(\mathrm{ES}) - m_i^2(\mathrm{ES}) \right)^2, \tag{3.41}$$

$$\langle \sigma_{i,m}^2(\mathrm{TN}) \rangle = \frac{1}{2} \left(m_i^1(\mathrm{TN}) - m_i^2(\mathrm{TN}) \right)^2, \tag{3.42}$$

$$\langle \sigma_{i,c}^2(\mathrm{ES}) \rangle = \frac{1}{2} \left(h_{i,c}^1(\mathrm{ES}) - h_{i,c}^2(\mathrm{ES}) \right)^2, \tag{3.43}$$

and

$$\langle \sigma_{i,c}^2(\mathrm{TN}) \rangle = \frac{1}{2} \left(h_{i,c}^1(\mathrm{TN}) - h_{i,c}^2(\mathrm{TN}) \right)^2. \tag{3.44}$$

The variance in expression changes

We again estimate the true variance of expression changes

$$\mathrm{var}(\Delta_m) = \frac{1}{N} \sum_{i=1}^{N} \left(\Delta_{i,m} - \bar{\Delta}_m \right)^2, \tag{3.45}$$

with

$$\bar{\Delta}_m = \frac{1}{N} \sum_{i=1}^{N} \Delta_{i,m}, \tag{3.46}$$

from the observed variance of the measured expression changes

$$V^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\bar{\Delta}_{i,m} - \bar{\Delta}_m\right)^2. \tag{3.47}$$

Writing the measured gene expression changes $\bar{\Delta}_{i,m}$ in terms of the true values $\Delta_{i,m}$ and the deviations, and using that cross-correlations are zero, we obtain

$$\langle V^2\rangle = \mathrm{var}(\Delta_m) + \frac{1}{2}\sigma^2_{i,m}(\mathrm{ES}) + \frac{1}{2}\sigma^2_{i,m}(\mathrm{TN}). \tag{3.48}$$

Using this we thus estimate the true variance in expression changes as

$$\mathrm{var}(\Delta_m) = V^2 - \frac{1}{2}\sigma^2_{i,m}(\mathrm{ES}) - \frac{1}{2}\sigma^2_{i,m}(\mathrm{TN}). \tag{3.49}$$

Error in the model

We again define the true deviation between predicted and true expression change for gene i as

$$D_i = \Delta_{i,m} - \tilde{c} - \sum_c \tilde{\alpha}_c \Delta_{i,c}, \tag{3.50}$$

and want to estimate the true average squared-deviation

$$D^2 = \frac{1}{N}\sum_{i=1}^{N}\left(D_i\right)^2. \tag{3.51}$$

The observed total deviation between measured and predicted levels is given by

$$T^2 = \frac{1}{N}\sum_{i=1}^{n}\left(\bar{\Delta}_{i,m} - \tilde{c} - \sum_c \tilde{\alpha}_c \bar{\Delta}_{i,c}\right)^2. \tag{3.52}$$

Writing the measured expression changes $\bar{\Delta}_{i,m}$ and chromatin mark changes $\bar{\Delta}_{i,c}$ in terms of the true changes and deviations, and using that the cross-correlations in the deviations are zero, we obtain

$$\langle T^2\rangle = D^2 + \frac{1}{2}\sigma^2_{i,m}(\mathrm{ES}) + \frac{1}{2}\sigma^2_{i,m}(\mathrm{TN}) + \frac{1}{2}\sum_c \tilde{\alpha}_c^2\left(\sigma^2_{i,c}(\mathrm{ES}) + \sigma^2_{i,c}(\mathrm{TN})\right). \tag{3.53}$$

From this we estimate the true average squared-deviation as

$$D^2 = T^2 - \frac{1}{2}\sigma^2_{i,m}(\mathrm{ES}) - \frac{1}{2}\sigma^2_{i,m}(\mathrm{TN}) - \frac{1}{2}\sum_c \tilde{\alpha}_c^2\left(\sigma^2_{i,c}(\mathrm{ES}) + \sigma^2_{i,c}(\mathrm{TN})\right). \tag{3.54}$$

Combining these results we finally estimate the fraction f of expression-change that can possibly be explained by the model as

$$f = \frac{V^2 - T^2 + \frac{1}{2}\sum_c \tilde{\alpha}_c^2\left(\sigma^2_{i,c}(\mathrm{ES}) + \sigma^2_{i,c}(\mathrm{TN})\right)}{V^2 - \frac{1}{2}\sigma^2_{i,m}(\mathrm{ES}) - \frac{1}{2}\sigma^2_{i,m}(\mathrm{TN})}. \tag{3.55}$$

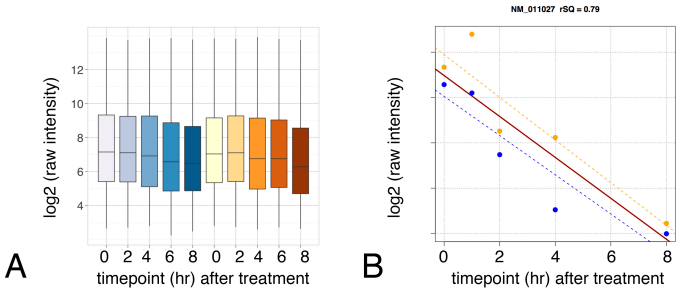
3.2.4 Calculation of transcript half-life by actinomycinD treatment

RNA was isolated from a fixed number of cells in culture (ES cells as well as Neurons) and subjected to Affymetrix ST 1.0 Mouse Gene Arrays. The raw data was processes with R's 'oligo' package, RMA was used without normalization. Expression values were aggregated on transcript level and degradation rates of each transcript were estimated using linear regression of the log (log2) transformed signal intensity values y versus time t.

$$y = n - mt \tag{3.56}$$

where t is time, m is the slope, n is the intercept and $d = m * ln(2)$ is the decay rate.

Using R's 'limma package' we calculate log2 fold-changes from t=0 to t=8 in duplicates



Supplemental Figure 3.5: (A) Duplicates of mRNA abundance measurements (orange, blue) in a time-course after actinomycinD treatment. For each experiment the log2 transformed values of all transcripts are summarized in a boxplot. Overall mRNA abundance decreases over time. (B) Raw mRNA abundance in log2 as a function of time after actinomycinD treatment for a single transcript. Measurement and linear regression in duplicates (orange and blue).

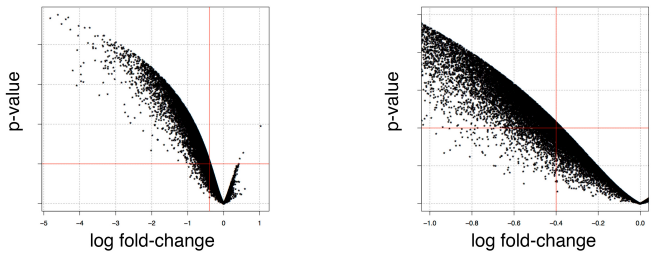
and infer the p-value for each. The resulting 'vulcano plot' is shown in Figure 2. Assuming a p-value of 0.01 (1 false in 100) the vast majority of genes with a log2 fold-change lower than -0.4 show significantly decreased levels in the time interval from 0 to 8 hrs to be considered. The slope m of these genes is calculated by:

$$m = \frac{\Delta y}{\Delta x} = \frac{logFC}{\Delta time} = \frac{-0.4}{8h} = -0.05 \tag{3.57}$$

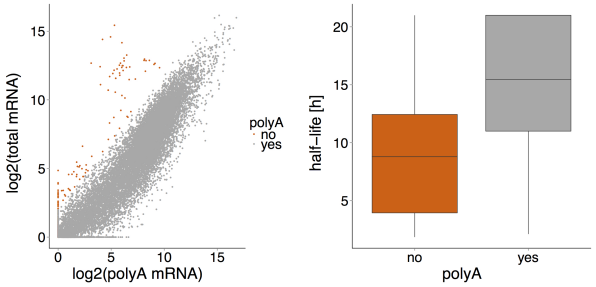
The corresponding half-life τ to the slope m of -0.05 is calculated by:

$$\tau = \frac{-1}{m} = \frac{-1}{-0.05} = 20h \tag{3.58}$$

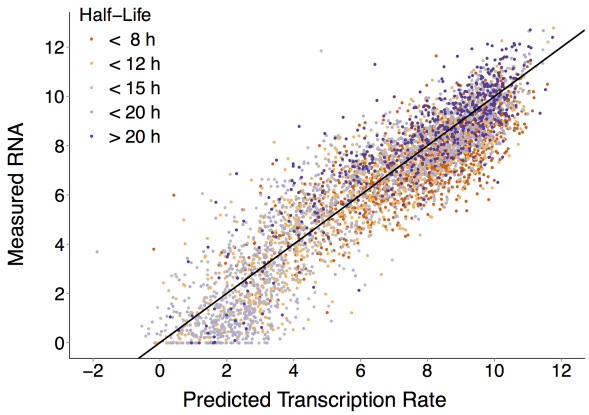
This value of 20 hrs corresponds to an upper limit that we select for the extrapolation of half-lives based on a time-course experiment covering an interval of 8 hours.



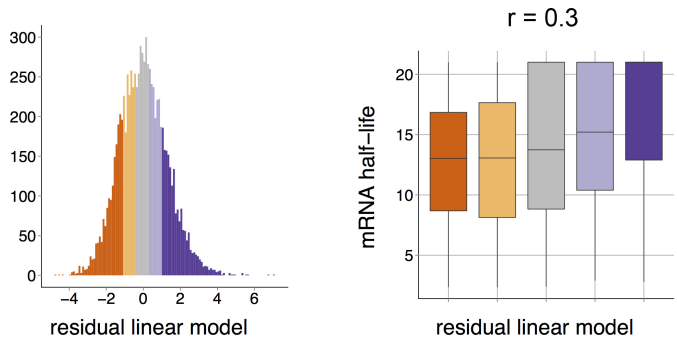
Supplemental Figure 3.6: Vulcano plot showing log fold-change from time point t=0 to time point t=8 on the x-axis versus corresponding p-values from replicates. Intersection of red lines at -0.4 log fold-change with p-value of 0.01.



Supplemental Figure 3.7: Short-lived transcripts are polyA-depleted. (A) PolyA depleted transcripts were identified contrasting mRNA sequencing with following polyA selection (x-axis) and mRNA sequencing following ribosomal RNA depletion without polyA enrichment (y-axis). Reads present in the non-polyA-selected experiment, which are not present in the polyA-selection are defined as non-polyA transcripts (orange). (B) Box-plot showing the shift in transcript half-life comparing polyandenylated (grey) and non-polyadenylated (orange) transcripts.



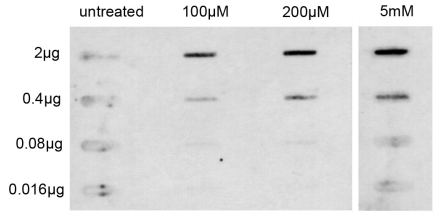
Supplemental Figure 3.8: Scatter plot of linear-model derived transcription rate (x-axis) and measured mRNA levels (y-axis). Both are log2 transformed read-counts. This plot illustrates i) that mRNA levels are less well modelled in low transcribed regions (corresponding to bin <20transcribed genes are depleted of short lived-genes (therefore the box-plot in Figure 2C in the leftmost bin can be misleading)



Supplemental Figure 3.9: Correlation of the model's residuals with the respective measure of post-transcriptional regulation. (A) Histogram of the residuals of the linear model. Colors indicate grouping of residulas in bins of equal size, this binning also applies for B,C and D. (B) Boxplot showing the correlation of mRNA half-life versus the residual of the linear model. Pearson correlation shown on top is 0.29, which means that almost 30% of the residual variation in mRNA levels can be explained by mRNA half-life.

3.2.5 Calculation of transcript half-life by metabolic labeling

We treated ESCs with medium containing thioU in a final concentration of 200μM for 1hr. RNA was isolated with Trizol. 4sU-labeled RNA was biotinylated using EZ-Link Biotin-HPDP (Pierce) and streptavidin IP was performed to separate the labeled RNA fraction. To recover the unlabeled RNA the flow-through was collected. RNA was recovered from the washing fractions and eluates using the RNeasy MinElute Spin columns (Qiagen).

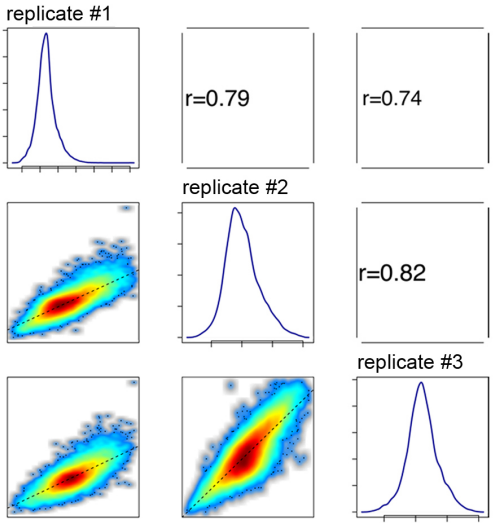


Supplemental Figure 3.10: Dot blot assay according to Doelken et al. 2008. Exposure 20 mins shown. Cells were treated for 60 mins with different concentrations of ThioU. We did not observe a difference in flourescence of labeled RNA between the two highest concentrations, therefore we choose 200μM as reported before in Doelken et al. 2008 for our experiment.

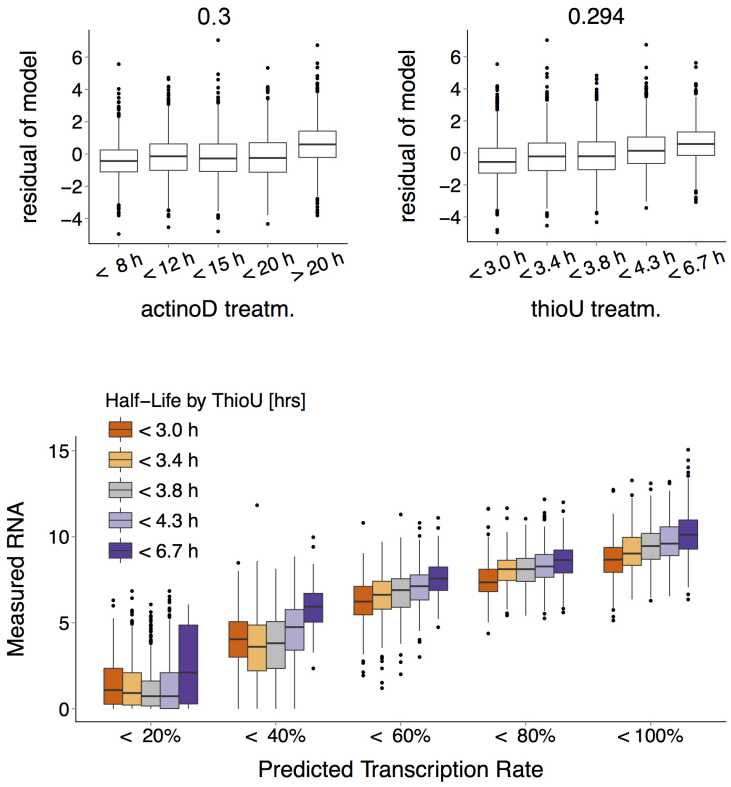
Experiment was done in biological triplicates. All three fractions of each replicate, RNA, total RNA, labeled and un-labeled RNA, were subjected Affymetrix Gene Arrays. All arrays (triplicates of total, labeled and unlabeled RNA) were normalized together by RMA and summarized on transcript level. Transcripts with RMA intensities less than 5 on linear scale were discared. To account for the relative measurement of the microarrays we calculate correction factors for the ratios (labeled/total RNA) and (unlabeled/total RNA) accoring to Doelken et al. 2008, for each of the triplicates separately. In addition we account for the U-bias in the IP (described in Schwannhaeuser et al. 2011) by normalizing to 'U' density of the transcript. RNA half-lives are subsequently calculated for each replicate assuming exponantial decay:

$$T_{1/2} = -t * \frac{ln(2)}{ln(1 - \frac{1}{1+\frac{(labeled/total)}{(unlabeled/total)}})} \tag{3.59}$$

were $-t$ is the labeleing time. As decribed in Doelken et al. we use t =55mins assuming the thioU labeling starts 5 mins after addition of thioU to the medium.

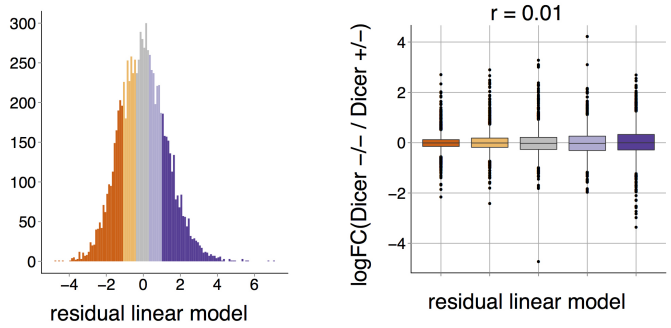


Supplemental Figure 3.11: Transcript half-life calculated from biological triplicates of total-, newly synthetized and preexisting RNA. Shown are pairwise comparisons and their pearson correlations.

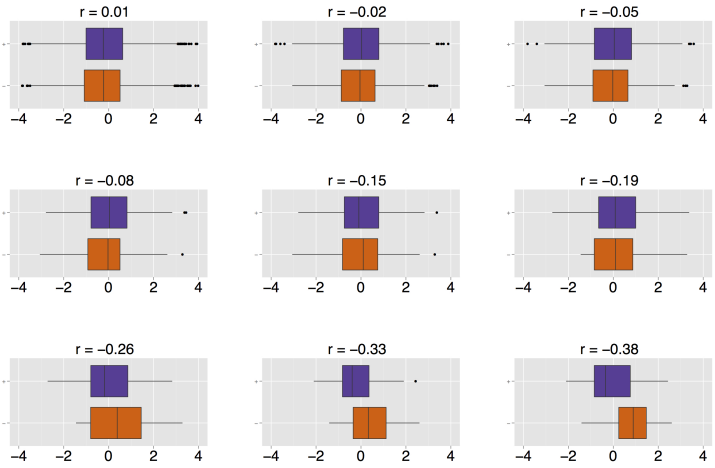


Supplemental Figure 3.12: RNA half-life derived by metabolic labeling was integrated in the linear model predicting transcription. The upper boxplots show half-lifes derived by actinomycinD treatment (left) and thioU treatment (right) agains the residual of the linear model. Both half-life measures can explain the remeining variance (residual) to the same extend, pearson correlation shown above the plot respectively. Correlation beweeen the two half-life measures shown in the scatterplot below. The histogram shows half-life distribution derived by thioU treatment. The lower plot shows mRNA half-life in the context of predicted transcription and measured mRNA level. Lower mRNA level in the same transcription bin can be explained by mRNA half-life derived by thioU treatment, in accordance with out actinomycinD derived data.

3.2.6 MicroRNA target determination by Dicer knockdown



Supplemental Figure 3.13: Correlation of the model's residuals with the respective measure of post-transcriptional regulation. Histogram of the residuals of the linear model. Colors indicate grouping of residulas in bins of equal size, this binning also applies for B,C and D. Boxplot showing the correlation of the log fold-change between Dicer knockout Dicer +/- cells versus the residual of the linear model. Pearson correlation shown on top is 0.01, which means that 1% of the residual variation in mRNA levels can be explained by a miRNA-target definition based on Dicer KO data.



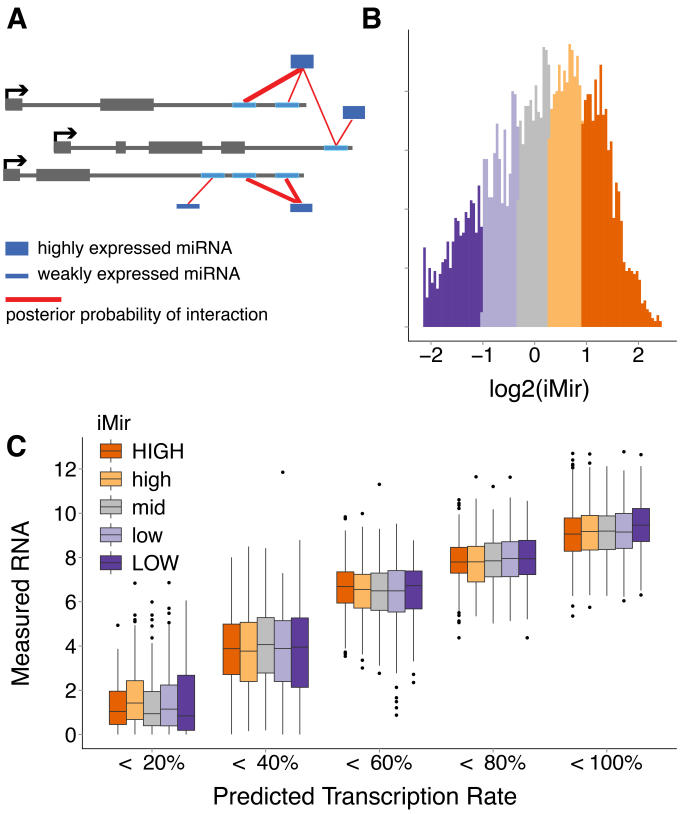
Supplemental Figure 3.14: Box Plots for different cut-offs of log fold-changes upon Dicer KO (y-axis) versus the residual of the linear model (x-axis). The color separates genes with positive (purple) or negative (orange) logFC upon Dicer KO. If the log fold-change is negative, the level of the respective mRNA is higher in Dicer -/- cells. These mRNA should be the ones affected by miRNAs and therefore we expect their residual to be negative. Hence we expect an increasing anti-correlation the higher the cut-off in absolute log fold-change. This increasing shift of the two groups can be observed for cut-offs of log fold- changes of: (A) 0, (B) 0.5, (C) 0.7, (D) 0.9, (E) 1.1, (F) 1.3, (G) 1.5, (H) 1.7, (I) 1.9

3.2.7 MicroRNA target determination by calculation of iMir score

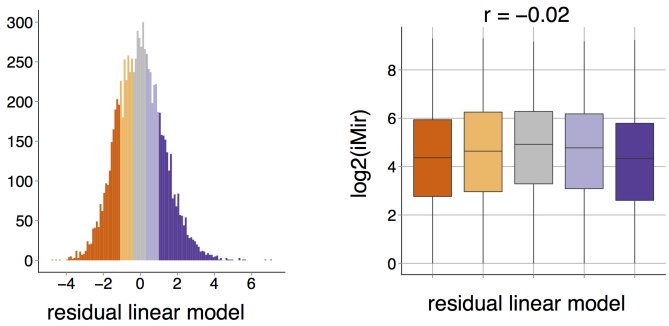
For the representative mRNA of each refSeq gene we calculated a score, which reflects the probability of this mRNA to be down regulated by miRNAs, here called iMir score. This score is influenced by two factors: (A) the posterior probability of a miRNA to bind a specific target sequence in the 3' UTR of the mRNA and (B) by the abundance of the miRNA in the respective sample. The posterior probability of a miRNA to bind a specific target sequence was adapted from the EIMMo algorithm (Gaidatzis et al. 2007). MiRNA abundance in the three cell types ES, NP and TN was measured in triplicates by small RNA sequencing (see methods). The iMir score is a sum of all the posterior probabilities of a miRNA target site in a mRNA weighted by the abundance of the miRNA and summed up for each mRNA. Formally we can write the $iMir$ score for one mRNA as:

$$iMir_{mRNA} = \sum_{n=1}^{N} pp_n * exp_n^{miRNA} \tag{3.60}$$

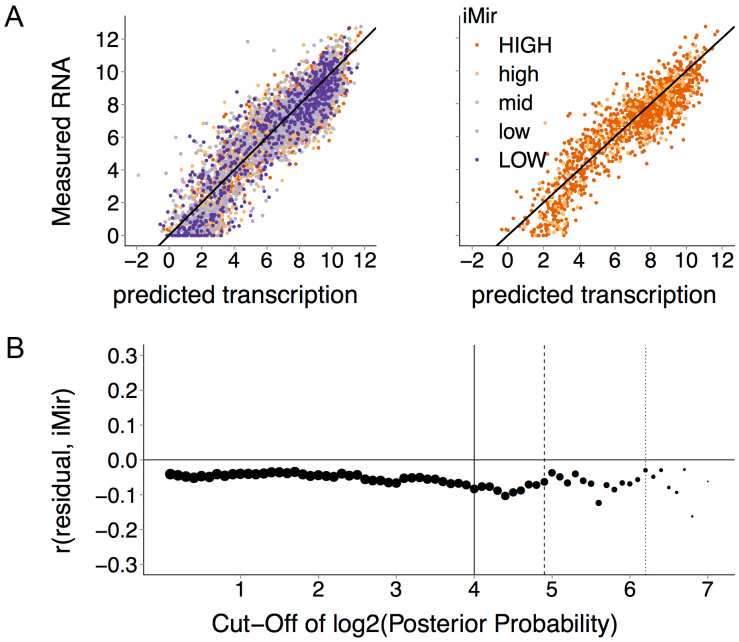
where N is the number of possible mRNA-miRNA interactions in a given 3' UTR and exp is the expression (abundance) of the respective miRNA involved in this interaction.
Consequently an mRNA will have a high $iMir$, or likely being down regulated by miRNAs, if there are many possible binding sites, the posterior probability of each binding site is high and the abundance of the possibly binding miRNAs is high.



Supplemental Figure 3.15: Effect of targeting by miRNAs on mRNA levels. (A) Scheme illustrating the components used in the calculation of the iMir value, a measure for the likelihood of an mRNA to be regulated by miRNAs. (B) Distribution of the logarithmic iMir value. Genes are grouped by iMir value into five equal groups indicated by color. (C) Genes are classified into five equal groups according to predicted transcription rate (0-100%), and within each group measured mRNA levels are shown as boxplots separately for genes with different iMir values (as in (B), color-coded). Within a given transcription group, predicted miRNA-target genes (iMir 'HIGH') have similar mRNA levels as other genes.

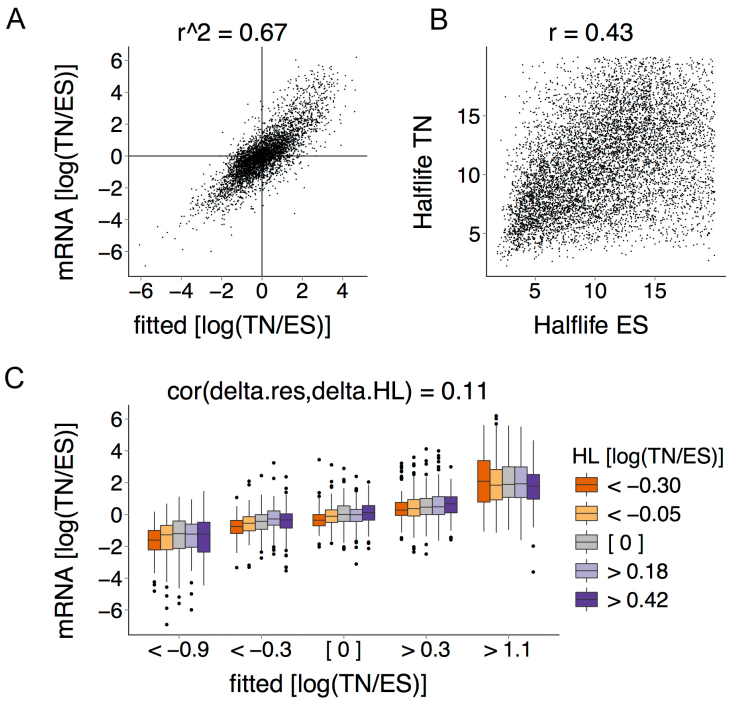


Supplemental Figure 3.16: Correlation of the model's residuals with the respective measure of post-transcriptional regulation. Histogram of the residuals of the linear model. Colors indicate grouping of residulas in bins of equal size, this binning also applies for B,C and D. Boxplot showing the correlation of the logarithmic iMir value versus the residual of the linear model. Pearson correlation shown on top is -0.04 (note that here we expect the correlation to be negative, as the more negative the residual, the higher the iMir value), which means that 4% of the residual variation in mRNA levels can be explained by a miRNA-target definition based on in-silico target prediction.



Supplemental Figure 3.17: Focus on high-confidence miRNA target genes. (A) Same plot as in Figure 1C. All refSeq genes are classified according to their iMir value as in Supplementary Figure 1B. (B) Subset of all the genes in (A) with log2(iMir) values higher than 4. This subset contains genes that are more likely to be miRNA-targets. (C) Pearson correlation (r) between the residual of the linear model and the log2(iMir) value as a function of cut-off applied to the posterior probability to be a miRNA target. A cut-off of zero corresponds to (A), and a cut-off of 4 (solid vertical line) corresponds to (B). Correlations are shown for subsets for log2 posterior probability cut-offs in 1.0 inter- vals. The point-size illustrates the number of genes at each cut-off. At 4.9, the subset contains 1000 genes (dashed line), the subset at 6.2 contains 100 genes (dotted line). Increasing cut-offs select higher-confidence miRNA-target genes that can explain the residual of the linear fit slightly better.

3.2.8 Prediction of mRNA abundance change between cell types



Supplemental Figure 3.18: Predictive power of chromatin and hal-life in changes of mRNA. (A) Scatter plot showing correlation between change predicted transcription rate (x-axis) and change in measured mRNA level (y-axis), changes from ES to TN respectively. (B) Scatter plot showing correlation between experimen- tally inferred half-life in ES (x-axis) and TN (y-axis). Transcript half-life changes between the cell types indi- cating a functional importance of RNA decay. (C) Potential of mRNA half-life changes to explain remaining changes in measured mRNA levels. Similarly to (A), correlation between changes in transcription (x-axis) and changes in measured mRNA (y-axis), colours indicate level of change in mRNA half-life. Negative values indicate genes which decrease in half-life during differentiation. These genes contribute to a decrease in mRNA levels during differentiation.

3.2.9 Tissue-specific expr.: test of independence or homogeneity

We test here whether ES or TN specific expressed genes are independent of the definition of tissue-specificity obtained from SymAtlas. Tissue specific expression was defined based on symAtlas expression over 75 tissues. For each refSeq gene we counted the number of tissues/cell types in which it is expressed (expression defined by a cut-off in log-transformed expression, $exp > 7$). RefSeqs are then classified accruing to the number of tissues in not expressed (0 tissues), tissues-specific expressed (1-5 tissues), intermediate (6-70 tissues) and ubiquitous expressed (71-75 tissues). ES or TN specific expression was defined given the mRNA sequencing data in our differentiation system, comparing ES and TN. We classified into expressed in [ES AND TN] or [ES OR TN] by a cut-off at $+/-2$ from the $x = y$ diagonal. Based on these groups we tested the if tissue-specific expression accruing to symAtlas is independent to the expression pattern observed in our system. We use the following matrix to perform a chi-squared test:

	tissues-specific	ubiquitous
ES or TN	279	661
ES and TN	386	2019

.

The p-value obtained is close to zero ($< 2.2e-16$). Hence we reject the hypothesis that the expression type according to symAtlas is the same for genes expressed in either or both of our cells types. Looking at the data, ubiquitous expressed genes are more enriched in genes expressed in both cell types.

3.2.10 A partially non-linear model

In a linear setting we have:

$$T_i = a_0 + \sum_j a_j C_{ij} \tag{3.61}$$

where C_{ij} are measured chromatin modifications of type j for gene i and a_j are the corresponding coefficients of the multilinear regression. The fitted response (mRNA level) is sigmoidal distorted (Figure 13, left).

$$\sigma(x) = sigmoid(a_0 + \sum_j a_j C_{ij}) \tag{3.62}$$

where $\sigma(x)$ is a sigmoidal function that captures the distortion of the linear relationship due the detection limit at the lower end and due to the saturation of 3K36me3 at the upper end of the mRNA levels. One of the simplest sigmoidal functions is:

$$\sigma(x) = c + \frac{d-c}{1+\exp(b*(x-e)} \tag{3.63}$$

It has the simple inverse:

$$\bar{\sigma}(y) = e + \frac{1}{b} * log(\frac{-d+y}{c-y}) \tag{3.64}$$

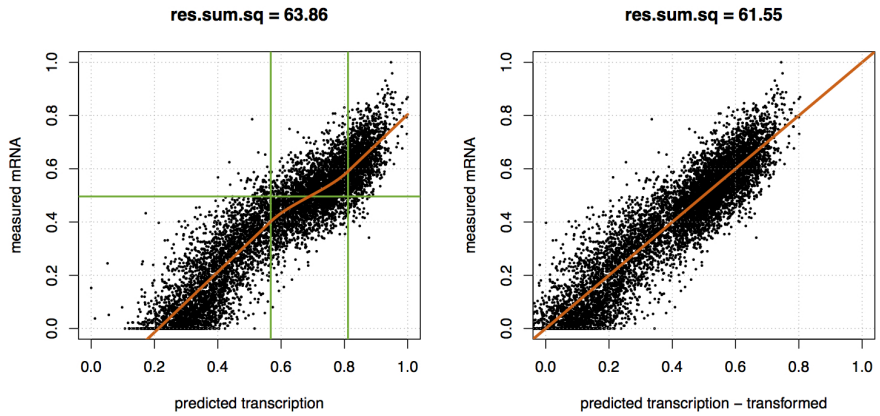
where e is the position of the largest inflection, b is the slope there.

However, $inSig$ is only defined between the saturation levels c and d. If we want to fit all our datapoints, some of which will be outside these boundaries, we have to use a complex function composed of an inverse sigmoid with a linear function attached to each side of the inverse sigmoid. For that we introduce a fifth parameter, δ that determines the 'attachment point' of these linear functions relative to the left and right boundaries c and d of the inverse sigmoid. The formula for the composite function can be written as:

$$\bar{\sigma}(x)^*(b,c,d,e,\delta) = \begin{cases} mx+n, & \text{if } x <= c+\delta \\ mx+n, & \text{if } x >= d-\delta \\ e + \frac{1}{b} * log(\frac{-d+x}{c-x}), & \text{else} \end{cases} \tag{3.65}$$

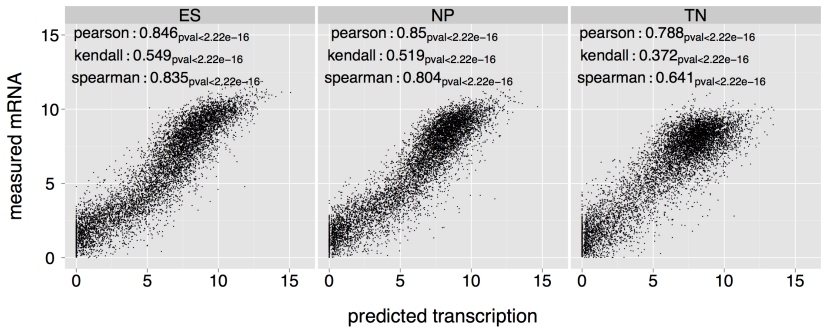
where the slope m of the 'attached' linear functions is the first derivative of $\bar{\sigma}(x)$ at $(c+\delta)$ and $(d-\delta)$. The y-intercept n for each of the linear functions can be calculated from x, m and $\bar{\sigma}(x)$.

Using a range of starting parameters for b, c, d, e and δ we do a non-linear model fitting. In case the nls converges, it always converges with the parameters $b = -16.39$, $c = 0.49$, $d = 0.87$, $e = 0.5$ and $\delta = 0.07$ (see Figure 13, left).
Indeed, the lower mRNA levels seem to relate linear to the predicted transcription, whereas the sigmoid distortion is detected only in a small intervall. TThe normalized covariance between the measured mRNA levels and the predicted transcription values after transformation by the inverse sigmoidal function is only marginally larger than for the linear fit (see Figure 13, 0.846 vs. 0.851).

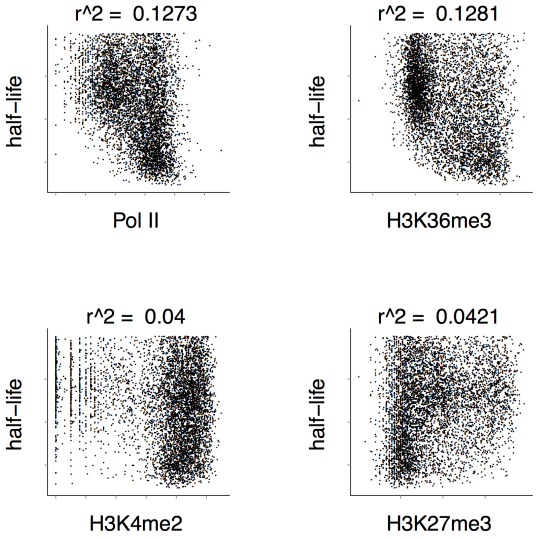


Supplemental Figure 3.19: Non-linear fitting of a composite function to the data, predicted transcription vs. measured mRNA. Left: A non-linear model composed of an inverse sigmoid with linear functions attached is fitted. The orange line shows the composite function. The green lines indicate the range of the invese sigmoid in the x-axis and the inflection point on the y-axis. Right: Predicted transciption was transformed accoding to the composite function and plotted against mRNA again. Explained variance of mRNA by original and transformed predicted transcription is shown above the plots respectively.

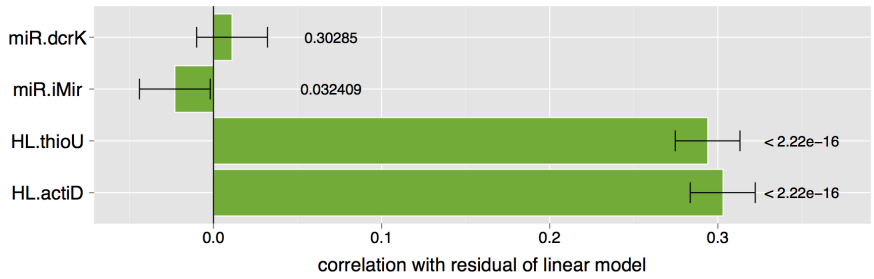
3.2.11 Additional supplemental figures



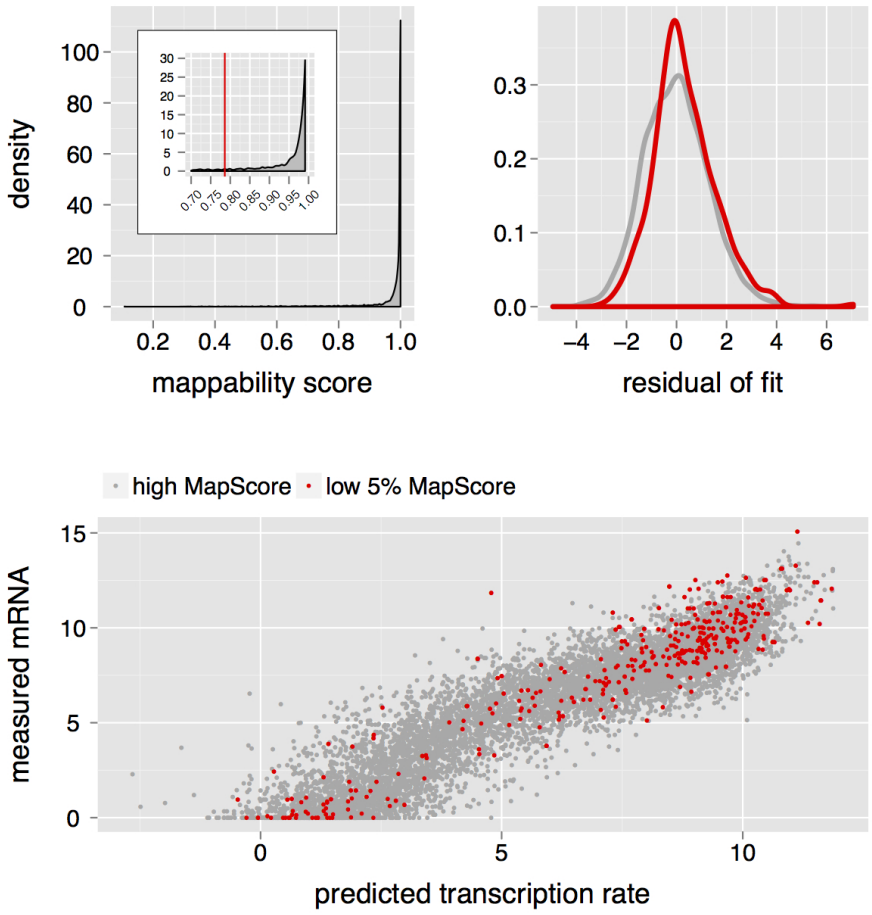
Supplemental Figure 3.20: Scatter plot of linear-model derived transcription rate (x-axis) and measured mRNA levels (y-axis). Both are log2 transformed read-counts. In additon to pearson correlation, we also show kendall and spearman correlations and corresponding p-values to show the significance of the high correlation between the two values despite their non-normal distribution.



Supplemental Figure 3.21: Scatter plots show correlation of the regressors in the linear model (histone modifications) with half-life. Shown are the respective regressors on the x-axes and half-life of the corresponding transcript on the y-axes.



Supplemental Figure 3.22: Bar plot represents the correlation (r) of different post-transcritional measures with the residual of the linear model. For mRNA half-life measurements by metabolic labeling (HL.thioU), as well as for the measures of being a miRNA target, pearson correlation was tested by bootstrapping using the repective contiuous variables. For the mRNA half-life measure derived by actinomycinD treatment (HL.actiD) we use a categorical variable (due to group of long lived genes) and infer r by linear regression. In all cases, error bars represent 95% confidence interval, p-value respectively for each test.



Supplemental Figure 3.23: Mapability of mRNA-sequencing reads. Foreach refSeq transcript a mapability score m was calculated as the ratio of uniquely mapable over all 36mer sequences of that transcript. To calculate mapablity of 36mers, each transcript was tiled in n 36mers, where $n = length_{transcript} - 35$. All 36mes were aligned to the mm9 genomic sequences as well as to refSeq transcripts and defined as unique if it did not match more than once in either of the references. The distribution of m is shown in the density plot on the left, inset plot contains only $m! = 1$. The scatter plot on the right shows the distribution of non-uniquely mapping transcripts (lower 5%, m <0.79) within the population of all representative transcripts.

Chapter 4

Conclusions

In the past 5 years much progress has been made understanding the mechanisms that influence steady-state or dynamic mRNA level at the level of chromatin, transcription or on the post-transcriptional level. However an integrated view of all these layers is needed to understand the interplay and relative contributions to gene expression in a cell.

At the beginning of my PhD thesis there were no genome-wide studies on global contributions of transcription, mRNA stability and other factors to mRNA levels [Cheadle et al., 2005] and my aim was to integrate genome-wide chromatin data available in the lab with measures of post-transcriptional regulation to determine the relative contributions of these layers to steady state mRNA levels. We found that the lion share of steady state mRNA level is set on the transcriptional level, and post-transcriptional contributions are quantitatively minor. Additionally, we describe histone marks, especially the co-transcriptional histone mark H3K36me3, as very good predictors of mRNA abundance in the cell.
In the following sections I would like to discuss our findings in the context of recent studies, published during the time-period of my PhD. These recent studies on the one hand make use of global measurements of numerous genomic features and attempt to predict mRNA abundance and on the other hand integrate different regulatory layers into a network explaining the connectivity in gene expression regulation.
Gene regulation in a broad sense however, goes from DNA not only until mRNA levels but ultimately to protein levels. Recent developments in quantitative mass spectrometry opened possibilities to measure protein abundance on a genome-wide scale allowing for the first time global comparisons of mRNA and protein level and consequently regulatory impact of translational and post-translational mechanism. I will also discuss some of those findings and how the results of my PhD project can be interpreted in light of these new developments.

4.1 A Longstanding Task: Decoupling Regulatory Layers

In my PhD project I asked the question: If regulation of gene expression in eukaryotes is a stepwise, multi-layered process, what are the relative contributions of these layers? What are the actual numbers as percentage from 'total' 100% determination of the steady state mRNA level?

To address this question we had to separate at least two layers: the transcriptional regulation, 'what get's transcribed and how efficient?', and the post-transcriptional regulation, as 'how long is the life-time of the initial transcript?'. To estimate what is being transcribed in the fist place, there are various methods. The most straightforward approach would be to monitor ongoing transcription directly, for which nuclear run-on (NRO) would be the closest method to do so [García-Martínez et al., 2004]. By pulse-labeling, one can integrate the labeled mRNA over time and have a readout what was transcribed [García-Martínez et al., 2004]. By integration of a modified, radiolabeled nucleotides, intensities can be detected on a blot, however, it does not allow for quantitative readout by microarray or deep-sequencing. Other, less direct, methods try to infer transcription from different genomic features: DNA sequence itself was shown to be predictive for gene expression. Predicted TF binding sites and their evolutionary conservation have been used to infer promoter activity [Hemberg and Kreiman, 2011; Irie et al., 2011]. But DNA sequence not only provides binding specificity for TFs but also largely determines nucleosome positioning [Segal et al., 2006]. DNA binding complexes, nucleosomes, TFs and the recruited RNAP II transcription machinery compete for binding around promoter regions and one can calculate probabilities of transcription initiation at promoters based on thermodynamic equilibrium (reviewed in Segal and Widom [2009]). However, experimentally derived TF binding makes a much better predictor [Cheng and Gerstein, 2011] since only a subset of predicted TF binding sites will be actually occupied *in vivo*. In the same line experimentally inferred histone modification data provides a powerful prediction of mRNA abundance of the linked gene. Several studies build on the histone modification maps in human T cells [Barski et al., 2010; Wang et al., 2008] and inferred patterns of histone modifications characteristic for actively transcribed promoters [Hon et al., 2009]. Based on this data the levels of histone modifications were found to be predictive for both protein-coding mRNA expression, depending on CpG density of the promoter [Karlic et al., 2010] and even for miRNA expression [Zhang and Zhang, 2011]. Notably, all of these studies restrict their readout of histone modification on a region around the TSS and thereby miss the high predictive power H3K36me3 in transcription elongation.

Moreover they find that a few histone modifications are sufficient to predict mRNA levels and that adding more histone marks does not improve modeling significantly.
We choose to monitor RNAP II together with a few histone marks, that are bi-uniquely connected with either active of repressed genes. In particular however, we do not restrict our analysis to the promoter region of the gene, where initiation is regulated but instead use a more downstream readout which is set during transcription elongation: H3K36me3. As reviewed in the introduction, this mark is set by the elongating polymerase and stays and accumulates until it gets diluted by cellular division. Therefore to measure H3K36me3 is actually more informative than RNAP II occupancy itself because RNAP II is only catch- and detectable at the promoter or in the moment when it 'runs' through the gene body. It is also the closest measure in terms of direct readout of transcription without using an invasive labeling method, IP on chromatin of normal grown cells is sufficient and enrichments yield high specificity. The variance in mRNA levels, which can be explained by H3K36me3 alone is therefore already higher than some of the complex predictors used in other studies.
Based on being able to predict more than 80% variance in mRNA levels by 3 histone marks and RNAP II occupancy, we inferred RNA decay rates transcriptome wide with two different methods (detailed in the introduction) and found, that although mRNA decay is measurable in terms of transcript abundance, it only shapes the steady-state level of an mRNA very little. The percentage contribution we assigned to RNA decay modulating mRNA levels is between 2 and 12%.
At the beginning of my PhD there was no study addressing the actual quantitative contributions of the different regulatory layers to mRNA levels with high throughput methods. The first systematic account to this question used transcription run-on (TRO) involving isolation of nuclei to measure ongoing transcription [García-Martínez $\mathrm{et\ al.}$, 2004]. RNA stability can then be calculated from mRNA levels and measured transcription rate. In yeast, Garcia-Martinez et al. found the median pearson correlation (r) of mRNA levels with transcription is 0.6 while r with mRNA stability is -0.24 and concluded that transcription is the main determinant of RNA levels. The first study in higher eukaryotes measured changes in both, transcription and mRNA half-life, during T cell activation in [Cheadle $\mathrm{et\ al.}$, 2005]. Using the same method as [García-Martínez $\mathrm{et\ al.}$, 2004], they observe a lack of detectable transcriptional regulation (change in newly transcribed RNA) of large numbers of changing mRNA levels and speculate that mRNA stability may account for as much as 50% of all changes in mRNA measured. However, this conclusion is largely driven by possible technical shortcomings of NRO (transcription in isolated nuclei) that bias the measurement and lead to a reduced correlation with mRNA level changes. To circumvent the downfalls of NRO, it is possible to measure mRNA decay directly as opposed to inferring it from transcription,

for example by RNAP II inhibition. A study in yeast showed the response of mRNA abundance and decay under different stress conditions and found that, depending on the type of stress (transient or enduring) RNA half-life can explain different amounts of changes in mRNA levels [Shalem et al., 2008]. Many following publications used the same method to infer mRNA half-life resulting in partially different conclusions (see further down). Using inhibition of RNAP II, or transcription arrest has the major problem, that the expression pattern of the cell will potentially change due to the effect of the stress imposed by arresting transcription.
Two studies in mouse cells recently revisited the question of relative contributions of different layers to gene expression [Rabani et al., 2011; Schwanhäusser et al., 2011]. Both studies used the less invasive method of metabolic labeling [Dölken et al., 2008] to measure transcription and calculate mRNA degradation. In contrast to our study, where transcription is modelled and half-life is experimentally measured, Rabani et al. measured transcription directly. Based on measured transcription and mRNA levels, the researchers test two different models: (I) assuming constant degradation for all gene their model can explain 78% of the variance in mRNA levels whereas a model (II) allowing gene dependent decay rates, has the capacity to explain 86% variance. This result is consistent with our findings and although using a different approach, almost yields the same percentages of relative contribution. This is equally consistent with the study by Schwannhäuser et al. [Schwanhäusser et al., 2011] which also predicts a minor contribution of mRNA degradation to steady-state levels and goes even a step further to predict relative regulatory contributions to protein levels (detailed discussion in section 4.3)

4.2 The Difficulty: Coupling of Regulatory Layers

The main idea of my thesis is based on the assumption that a decoupling of regulatory layers is fair and possible, because only if they do not depend on each other, we can assess the relative contribution of each layer.
In yeast, however, there is evidence for a direct coupling mechanism between mRNA transcription and degradation, mediated by two RNAP II subunits: Rpb4 and Rpb7. These subunits bind to the mRNA during transcription and escort the transcript from the nucleus to the cytoplasm. Thereby they can affect mRNA stability and modulate translation [Goler-Baron et al., 2008; Lotan et al., 2005, 2007]. Another complex linking production and degradation is CCR4-NOT, the major mRNA deadenylase in yeast, which controls the initial step of degradation (see introduction) [Chen et al., 2002; Tucker et al., 2001]. In

addition it is part of a multicomponent assembly containing diverse transcription initiation factors, such as members of the SAGA complex [Benson et al., 1998], subunits of RNAP II [Liu et al., 2001] and subunits of the transcription initiation factor TFIID [Lemaire and Collart, 2000; Sanders et al., 2002].
Most of the studies on coupling so far have been undertaken in yeast, because a lot is known about the proteins involved in a possible connection between transcription and degradation. Two principally different models of interconnection between transcription and degradation emerged, which I will refer to as 'co-operative' and 'non-co-operative' models. In the co-operative model transcription and degradation act in concert to achieve higher or lower levels of a certain mRNA, meaning if transcription rate of an mRNA is increased, it would also be stabilized post-transcriptionally. In the 'non-co-operative' model mRNA would be in contrast destabilized when transcriptionally induced. This would result in a 'balancing' mechanism where transcription and degradation 'buffer' each other to stabilize a certain level of mRNA abundance.

All studies on this topic where published within the last 3 years of my PhD, and show partially contradicting results. Evidence for the cooperative model comes from two studies in fission and bakers yeast, which monitor transcription by labeling with newly transcribed RNA with either 4sU [Amorim et al., 2010] or radioactive UTPs [Castells-Roca et al., 2011]. In response to heat shock Castells-Roca et al. found that changes in transcription rates and mRNA stabilities are mostly homo-directional, meaning induced transcription leads to higher stability of the mRNA. A similar observation was reported by Amorim et al. in induced meiotic differentiation of *S.pombe*. Here, the positive link between transcription and stability was shown to be a TF inducing the production of a stabilizing RBP Meu5p along with other genes, which are stabilized by Meu5p in the cytoplasm.
A cooperative mechanism would be advantageous in terms of an economic strategy for gene expression regulation. If in a specific steady-state low levels of an mRNA are sufficient a homo-directional regulation would similarly produce few mRNA, to avoid unnecessary degradation of wastefully produced transcript. However in cells of higher eukaryotes the energy spend to synthesize mRNA in terms of high energy phosphates is roughly one tenth of the energy consumption by translation [Schwanhäusser et al., 2011]. Therefore the cellular energy usage in transcription and post-transcriptional processes might not be a driving force to select a way of regulatory interaction. However the cooperative model may be sensible in terms of responsiveness to environment. If transcription and degradation of mRNA are able to act in concert to achieve higher or lower mRNA levels adaption to external stimuli will be fast.

4. CONCLUSIONS

Evidence for the non-cooperative model comes from both, yeast but also mammalian systems. When transcription and degradation are able to counter act, and thereby balance each other, noise is minimized when a precise level of expression is required.
Monitoring mRNA levels and decay rates in yeast under hyperosmotic stress conditions, Molin et al. observe genes with decreased transcription are stabilized in the cytoplasm while stress induced genes undergo destabilization [Molin et al., 2009]. This way the cell achieves a balancing effect, where final mRNA levels are changed only minimally. A similar effect was observed in yeast under oxidative stress [Shalem et al., 2011]: A wild type yeast could balance mRNA levels by a counteracting response in RNA stability, however a mutant strain, carrying a RNAP II, that poorly recruits Rpb4 and Rpb7, can not buffer the change in transcription. This is another evidence for the two RNAP II subunits Rpb4/Rpb7 being involved in coupling between transcription and degradation. Also in favor of the non-cooperative model is a study by Elkon et al. [Elkon et al., 2010] based on mRNA transcription data in mouse fibroblasts [Dölken et al., 2008]. As a response to interferons the mRNA stability is modulated according to the rapidity of gene induction: a higher induction leads to shorter half-life.
Within the last year two studies posed the question, if transcription and degradation appear to be coupled between conditions, is there an evolutionary connection between the two processes? Therefore they used two related yeast species respectively to investigate fold-changes in RNA synthesis and decay. Both studies, although using different techniques to infer mRNA decay, come to the same conclusion: there is coupled evolution between synthesis and decay [Dori-Bachash et al., 2011; Sun et al., 2012]. Moreover, Sun et al. used mutants of either RNAP II or the deadenylase Ccr4-Not and found that besides the expected decrease in transcription or degradation, the counteracting mechanism was buffering the mRNA levels respectively.
A previous paper from the same lab [Miller et al., 2011] investigating stress response in yeast, had shown that the interplay between mRNA synthesis and decay is largely dependent on the phase of the stress response. While there was non-cooperative behavior in the initial shock and induction phase, no correlation between production and degradation was observed.
With the data derived from our murine *in vitro* differentiation system, we see that predicted transcription at a single time point agrees with the non-cooperative model in that highly transcribed genes (high H3K36me3) are degraded fast and lowly transcribed genes have a long half-life (correlation of predicted transcription vs. degradation, $r = 0.36$). Based on the contradicting observations in yeast stress responses, it would be interesting to investigate how mRNA transcription and degradation rates change upon an external stimulus

to ESC or terminal neurons. Proteins like Rpb4/Rpb7 exist in mammals and it should be subject to investigation in the future if these proteins similarly build a physical link between transcriptional and post-transcriptional regulation.

4.3 mRNA to Protein

A key assumption in studying mRNA expression is that it is informative for the prediction of protein abundance. However, only recently studies have explored the mRNA-protein expression correlation in yeast or human tissues and the results have been relatively inconsistent [Guo et al., 2008].
Two early studies in yeast [Griffin et al., 2002; Gygi et al., 1999], which assessed this correlation, where restricted to a very low number of genes (< 250) due to the laborious work of mass spectrometry. Looking at steady state [Gygi et al., 1999] and fold-changes between yeast growth conditions [Griffin et al., 2002] both groups found a very low correlation between mRNA and protein. Apart from the low number of monitored genes, mass spectromety itself is thought to be difficult for quantitative studies because the efficiency with which peptides ionize and enter the mass spectrometer depends upon both their composition and the local chemical environment, producing variation in the MS signal intensity [Lu et al., 2007]. Lu et al., also in yeast, used a method to normalize for this effect (APEX) and found that 73% of the protein abundance is explained by mRNA abundance. This high correlation was confirmed by another lab which imposed osmotic stress on yeast and measured abundance of about 2500 proteins together with their coding mRNAs (pearson correlation $= 0.87$, Lee et al. [2011]).

Moreover, studies in mammalian systems are also in disagreement whether mRNA levels reflect protein abundance in a cell. Three studies in human monocytes [Guo et al., 2008], murine ES [Lu et al., 2009] and liver [Ghazalpour et al., 2011] cells report that a large proportion of changes in protein levels is not accompanied by analog changes in the expression of corresponding mRNAs, suggesting an important role for translational regulation. On the other hand, results in human cancer cell lines [Nagaraj et al., 2011; Vogel et al., 2010], which monitored up to 9207 genes by RNA-sequencing and microarrays and corresponding proteins levels, report a higher correlation ranging from 0.53 to 0.6. In human ESC, induced pluripotent stem cells and fibroblasts, the reported explained variance of protein by mRNA level was even higher ($r=0.7$, Munoz et al. [2011]).
Recent studies using the SILAC method (stable isotope labeling by amino acids in cell cul-

ture) to quantify protein abundance and sequencing to quantify mRNA abundance are in good agreement with a correlation between the two measures of more than 0.6 [Lundberg et al., 2010; Schwanhäusser et al., 2011].

In my PhD thesis I do not investigate the abundance of proteins at all and just ask whether transcription, or features related to transcription are predictive for mRNA levels. Following these more recent studies, which report a high correlation between mRNA and protein, our results would implicate that one can fairly estimate the actual protein output of a cell by measuring histone marks and RNAP II abundance alone. Many studies investigating mRNA levels, implicitly assume a high correlation to protein and extrapolate their findings the be relevant for the phenotype of the cell. This is challenged by reports showing a low explained variance of proteins on the mRNA level [Ghazalpour et al., 2011; Guo et al., 2008; Lu et al., 2009]. From the current knowledge and ongoing discrepancy about contributions to protein levels, and in the scope of my PhD project it would be too speculative to extrapolate to protein levels and phenotype but it would be an inevitable next step to investigate the quantitative contribution of chromatin marks not only to mRNA but to protein abundance.

4.4 Modeling in Biology

In my PhD thesis I attempt to explain the relative contributions of different regulatory layers in gene expression regulation to mRNA levels in a given cell. The last three conclusion chapters extend this aim in summarizing studies which investigate not only quantitative contributions but also the coupling between the regulatory processes. Some studies (section 4.3) even go further in trying to explain protein levels which implies the consideration of even more regulatory processes. The final goal of collecting this information on relationships between process is to build an abstract model which simplifies the complex biological processes. In general, an abstract scientific model can help to explain a system, to study the effects of different components, and to make predictions about unobserved data points. A statistical model, like the linear model used in my thesis, is a formalization of the relationships between variables in the form of mathematical equations.
Mathematical models can be classified differently, one of which is the distinction between deterministic and probabilistic (stochastic) models. A deterministic model is one in which every set of variable states is uniquely determined by parameters in the model and by sets of previous variable states. A stochastic model does not describe variables by unique values but rather by probability distributions. Due to the brownian motion in a cell for example, randomness is present and a stochastic model would likely reflect the situation better.

However, deterministic models always perform the same way for a given set of initial conditions, which is preferable when we want to make predictions. Amongst other possible classifications of models, we choose to use linear over non-linear models to describe our system. There may be cases where one can through biological reasoning assume linearity, as an example one could assume that one RNAP II produces 5 mRNAs/h, two RNAP II produce 10 mRNAs/h and so on. This would imply a linear relationship between RNAP II and the amount of mRNA produced. However in most cases this relationships between elements in a biological system are not known because quantitative measures are hard to obtain or vary largely either between molecules in a cell, between cells within a population or between replicates within an experiment. In our case we use the most simple model: a linear model.

Linear modeling, or linear regression, as introduced in the first chapter, was coined in the 18th century, where it was applied to observations about properties of peas and people [Galton, 1890]. At this time however, to make a single linear regression on a larger set of data could take days to solve manually. With the rapid development of computing in terms of memory and performance in the last decades regression problems for many data intensive purposes, such as in economics, are solved computationally. Since biological readout turned from 'blobs on films' to quantitative measurements, regression models started to find their application in this field, too. The advent of high-throughput methods for data quantification such as microarrays or deep-sequencing in the 1990s allowed parallel investigation of thousands of genes. Bringing together computational and technological advance enables us now to employ modeling on a new level. Due to the large amount of data, we can actually visualize if the relationship between two biological variables, is linear, or non-linear. We obtained measurements for more than 10.000 genes and visualized predicted transcription based on enrichment of histone modifications in relation to transcript abundance of the respective gene. From the scatter plot we can conclude that the relationship between the log-transformed values of these readouts is almost linear, which led us to employ linear regression. The non-linear behavior on the upper and lower end of the predicted transcription can be explained by the technical limitations of ChIP for a histone mark. We tried to account for this systematic deviation from the linear regression line by implementing a more complex model, composed of a linear and an inverse sigmoid part. Fitting the complex model to the data revealed that it is not much more powerful in predicting mRNA levels from histone modification data. In general, model complexity always involves a trade-off between simplicity and accuracy of the model. We therefore applied a principle particularly relevant to modeling, the essential idea being that among models with roughly equal predictive power, the simplest one is the most desirable (known as 'Occam's razor'). While

added complexity usually improves the predictive power of a model, it can make the model more difficult to interpret. In the special case of a linear model we can even conclude from the correlation coefficient to the explained variance, and coefficients of the predictors will return contributions to the explained variance. This is not possible with a non-linear model, therefore we chose the advantage of simplicity over the little improvement in explanatory power.

With models built on experimentally inferred high trough-put data from biological systems researchers can now start to investigate if concepts developed from single gene analyses hold true on genome wide level. Modeling allows us to reevaluate with a large amount of datapoints and potentially change these concepts on the way to understand gene regulation as a whole process. This follows the idea that natural systems and their properties, should be viewed as wholes, not as collections of parts. This holistic way of scientific research is reflected in the relatively young discipline of systems biology and will mostly be driven by the application of different modeling approaches.

Bibliography

Adkins, M. W., Howar, S. R. and Tyler, J. K. (2004). Chromatin disassembly mediated by the histone chaperone Asf1 is essential for transcriptional activation of the yeast PHO5 and PHO8 genes. Molecular cell *14*, 657–666. 14

Affymetrix ENCODE Transcriptome Project and Cold Spring Harbor Laboratory ENCODE Transcriptome Project (2009). Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. Nature *457*, 1028–1032. 35

Almer, A. and Hörz, W. (1986). Nuclease hypersensitive regions with adjacent positioned nucleosomes mark the gene boundaries of the PHO5/PHO3 locus in yeast. EMBO J. *5*, 2681–2687. 14

Ambros, V. (2004). The functions of animal microRNAs. Nature *431*, 350–355. 42

Amorim, M. J., Cotobal, C., Duncan, C. and Mata, J. (2010). Global coordination of transcriptional control and mRNA decay during cellular differentiation. Molecular systems biology *6*, 380. 95

Andersen, E. C. and Horvitz, H. R. (2007). Two C. elegans histone methyltransferases repress lin-3 EGF transcription to inhibit vulval development. Development (Cambridge, England) *134*, 2991–2999. 17

Antequera, F., Tamame, M., Villanueva, J. R. and Santos, T. (1984). DNA methylation in the fungi. J Biol Chem *259*, 8033–8036. 8

Appanah, R., Dickerson, D. R., Goyal, P., Groudine, M. and Lorincz, M. C. (2007). An unmethylated 3' promoter-proximal region is required for efficient transcription initiation. PLoS Genet *3*, e27. 9

Babiarz, J. E., Ruby, J. G., Wang, Y., Bartel, D. P. and Blelloch, R. (2008). Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. Genes & development *22*, 2773–2785. 50, 55

Baek, D., Villén, J., Shin, C., Camargo, F. D., Gygi, S. P. and Bartel, D. P. (2008). The impact of microRNAs on protein output. Nature *455*, 64–71. 28

Ball, M. P., Li, J. B., Gao, Y., Lee, J.-H., LeProust, E. M., Park, I.-H., Xie, B., Daley, G. Q. and Church, G. M. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol *27*, 361–368. 9

Bannister, A. J., Schneider, R., Myers, F. A., Thorne, A. W., Crane-Robinson, C. and Kouzarides, T. (2005). Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. J Biol Chem *280*, 17732–17736. 17

Barski, A., Chepelev, I., Liko, D., Cuddapah, S., Fleming, A. B., Birch, J., Cui, K., White, R. J. and Zhao, K. (2010). Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. Nature structural & molecular biology *17*, 629–634. 92

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837. 1, 11, 17, 35, 43, 52

Bauer, W. R., Hayes, J. J., White, J. H. and Wolffe, A. P. (1994). Nucleosome structural changes due to acetylation. J Mol Biol *236*, 685–690. 10

Behm-Ansmant, I., Rehwinkel, J., Doerks, T., Stark, A., Bork, P. and Izaurralde, E. (2006). mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. Genes & development *20*, 1885–1898. 28

Beitzinger, M., Peters, L., Zhu, J. Y., Kremmer, E. and Meister, G. (2007). Identification of human microRNA targets from isolated argonaute protein complexes. RNA biology *4*, 76–84. 49

Bell, O., Schwaiger, M., Oakeley, E. J., Lienert, F., Beisel, C., Stadler, M. B. and Schubeler, D. (2010). Accessibility of the Drosophila genome discriminates PcG repression, H4K16 acetylation and replication timing. Nature structural & molecular biology *17*, 894–900. 41

Bell, O., Wirbelauer, C., Hild, M., Scharf, A. N. D., Schwaiger, M., MacAlpine, D. M., Zilbermann, F., van Leeuwen, F., Bell, S. P., Imhof, A., Garza, D., Peters, A. H. F. M. and Schubeler, D. (2007). Localized H3K36 methylation states define histone H4K16 acetylation during transcriptional elongation in Drosophila. EMBO J. *26*, 4974–4984. 16, 17, 43, 52

Bender, L. B., Suh, J., Carroll, C. R., Fong, Y., Fingerman, I. M., Briggs, S. D., Cao, R., Zhang, Y., Reinke, V. and Strome, S. (2006). MES-4: an autosome-associated histone methyltransferase that participates in silencing the X chromosomes in the C. elegans germ line. Development (Cambridge, England) *133*, 3907–3917. 16

Benson, J. D., Benson, M., Howley, P. M. and Struhl, K. (1998). Association of distinct yeast Not2 functional domains with components of Gcn5 histone acetylase and Ccr4 transcriptional regulatory complexes. EMBO J. *17*, 6714–6722. 95

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L. and Lander, E. S. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326. 12

Bernstein, J. A., Lin, P.-H., Cohen, S. N. and Lin-Chao, S. (2004). Global analysis of Escherichia coli RNA degradosome function using DNA microarrays. Proceedings of the National Academy of Sciences of the United States of America *101*, 2758–2763. 24

Bibel, M., Richter, J., Lacroix, E. and Barde, Y.-A. (2007). Generation of a defined and uniform population of CNS progenitors and neurons from mouse embryonic stem cells. Nature protocols *2*, 1034–1043. 36, 42

Bibikova, M., Lin, Z., Zhou, L., Chudin, E., Garcia, E. W., Wu, B., Doucet, D., Thomas, N. J., Wang, Y., Vollmer, E., Goldmann, T., Seifart, C., Jiang, W., Barker, D. L., Chee, M. S., Floros, J. and Fan, J.-B. (2006). High-throughput DNA methylation profiling using universal bead arrays. Genome research *16*, 383–393. 8

Bilu, Y. and Barkai, N. (2005). The design of transcription-factor binding sites is affected by combinatorial regulation. Genome biology *6*, R103. 6

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes & development *16*, 6–21. 42

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. Nature *321*, 209–213. 8

Bird, A. P. and Wolffe, A. P. (1999). Methylation-induced repression–belts, braces, and chromatin. Cell *99*, 451–454. 8

Boeger, H., Griesenbeck, J., Strattan, J. S. and Kornberg, R. D. (2004). Removal of promoter nucleosomes by disassembly rather than sliding in vivo. Molecular cell *14*, 667–673. 14

Boyer, L. A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L. A., Lee, T. I., Levine, S. S., Wernig, M., Tajonar, A., Ray, M. K., Bell, G. W., Otte, A. P., Vidal, M., Gifford, D. K., Young, R. A. and Jaenisch, R. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349–353. 13, 43

Boyes, J. and Bird, A. (1992). Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. EMBO J. *11*, 327–333. 9

Brownell, J. E., Zhou, J., Ranalli, T., Kobayashi, R., Edmondson, D. G., Roth, S. Y. and Allis, C. D. (1996). Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. Cell *84*, 843–851. 10

Buratowski, S. (2003). The CTD code. Nature structural biology *10*, 679–680. 14

Buratowski, S. and Kim, T. (2011). The Role of Cotranscriptional Histone Methylations. Cold Spring Harbor symposia on quantitative biology *75*, 95–102. 52

Caput, D., Beutler, B., Hartog, K., Thayer, R., Brown-Shimer, S. and Cerami, A. (1986). Identification of a common nucleotide sequence in the 3'-untranslated region of mRNA molecules specifying inflammatory mediators. Proceedings of the National Academy of Sciences of the United States of America *83*, 1670–1674. 25

Carninci, P. (2009). Molecular biology: The long and short of RNAs. Nature *457*, 974–975. 1, 35

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. and others (2005). The transcriptional landscape of the mammalian genome. Science (New York, NY) *309*, 1559–1563. 35

Carrozza, M. J., Li, B., Florens, L., Suganuma, T., Swanson, S. K., Lee, K. K., Shia, W.-J., Anderson, S., Yates, J., Washburn, M. P. and Workman, J. L. (2005). Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. Cell *123*, 581–592. 12

Cary, P. D., Crane-Robinson, C., Bradbury, E. M. and Dixon, G. H. (1982). Effect of acetylation on the binding of N-terminal peptides of histone H4 to DNA. Eur J Biochem *127*, 137–143. 10

Castells-Roca, L., García-Martínez, J., Moreno, J., Herrero, E., Bellí, G. and Pérez-Ortín, J. E. (2011). Heat shock response in yeast involves changes in both transcription rates and mRNA stabilities. PLoS ONE *6*, e17272. 95

Cheadle, C., Fan, J., Cho-Chung, Y. S., Werner, T., Ray, J., Do, L., Gorospe, M. and Becker, K. G. (2005). Control of gene expression during T cell activation: alternate regulation of mRNA transcription and mRNA stability. BMC genomics *6*, 75. 91, 93

Chen, J., Chiang, Y.-C. and Denis, C. L. (2002). CCR4, a 3'-5' poly(A) RNA and ssDNA exonuclease, is the catalytic component of the cytoplasmic deadenylase. EMBO J. *21*, 1414–1426. 94

Cheng, C. and Gerstein, M. (2011). Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. Nucleic acids research *40*, 553–568. 2, 42, 92

Cheng, H., Dufu, K., Lee, C.-S., Hsu, J. L., Dias, A. and Reed, R. (2006). Human mRNA export machinery recruited to the 5' end of mRNA. Cell *127*, 1389–1400. 23

Chi, S. W., Zang, J. B., Mele, A. and Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature *460*, 479–486. 49

Clouaire, T. and Stancheva, I. (2008). Methyl-CpG binding proteins: specialized transcriptional repressors or structural components of chromatin? Cellular and molecular life sciences : CMLS *65*, 1509–1522. 9

Collins, C. A. and Guthrie, C. (2000). The question remains: is the spliceosome a ribozyme? Nature structural biology *7*, 850–854. 21

Core, L. J., Waterfall, J. J. and Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science (New York, NY) *322*, 1845–1848. 1, 35

Dahan, O., Gingold, H. and Pilpel, Y. (2011). Regulatory mechanisms and networks couple the different phases of gene expression. Trends Genet *27*, 316–322. 5

Darzacq, X., Shav-Tal, Y., de Turris, V., Brody, Y., Shenoy, S. M., Phair, R. D. and Singer, R. H. (2007). In vivo dynamics of RNA polymerase II transcription. Nature structural & molecular biology *14*, 796–806. 14

Deal, R. B. and Henikoff, S. (2011). Histone variants and modifications in plant gene regulation. Current opinion in plant biology *14*, 116–122. 56

Di Giammartino, D. C., Nishida, K. and Manley, J. L. (2011). Mechanisms and consequences of alternative polyadenylation. Molecular cell *43*, 853–866. 41

Ding, X. C. and Grosshans, H. (2009). Repression of C. elegans microRNA targets at the initiation level of translation requires GW182 proteins. EMBO J. *28*, 213–222. 27

Dölken, L., Ruzsics, Z., Rädle, B., Friedel, C. C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., Ghazal, P. and Koszinowski, U. H. (2008). High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. RNA (New York, NY) *14*, 1959–1972. 29, 37, 47, 59, 94, 96

Dori-Bachash, M., Shema, E. and Tirosh, I. (2011). Coupled evolution of transcription and mRNA degradation. PLoS biology *9*, e1001106. 96

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K. and Beck, S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. Nature genetics *38*, 1378–1385. 8, 42

Edmunds, J. W., Mahadevan, L. C. and Clayton, A. L. (2008). Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation. EMBO J. *27*, 406–420. 17, 43, 52

Eisenberg, E. and Levanon, E. Y. (2003). Human housekeeping genes are compact. Trends Genet *19*, 362–365. 53

Elkon, R., Zlotorynski, E., Zeller, K. I. and Agami, R. (2010). Major role for mRNA stability in shaping the kinetics of gene induction. BMC genomics *11*, 259. 96

Enright, A. J., John, B., Gaul, U., Tuschl, T. and Marks, D. S. (2003). MicroRNA targets in Drosophila. Genome biology *5*, R1. 49

Eulalio, A., Huntzinger, E., Nishihara, T., Rehwinkel, J., Fauser, M. and Izaurralde, E. (2009). Deadenylation is a widespread effect of miRNA regulation. RNA (New York, NY) *15*, 21–32. 28

Eulalio, A., Rehwinkel, J., Stricker, M., Huntzinger, E., Yang, S.-F., Doerks, T., Dorner, S., Bork, P., Boutros, M. and Izaurralde, E. (2007). Target-specific requirements for enhancers of decapping in miRNA-mediated gene silencing. Genes & development *21*, 2558–2570. 28

Ezhkova, E., Pasolli, H. A., Parker, J. S., Stokes, N., Su, I.-h., Hannon, G., Tarakhovsky, A. and Fuchs, E. (2009). Ezh2 orchestrates gene expression for the stepwise differentiation of tissue-specific stem cells. Cell *136*, 1122–1135. 13

Fan, X. C. and Steitz, J. A. (1998). Overexpression of HuR, a nuclear-cytoplasmic shuttling protein, increases the in vivo stability of ARE-containing mRNAs. EMBO J. *17*, 3448–3460. 25

Farh, K. K.-H., Grimson, A., Jan, C., Lewis, B. P., Johnston, W. K., Lim, L. P., Burge, C. B. and Bartel, D. P. (2005). The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science (New York, NY) *310*, 1817–1821. 28

Gaidatzis, D., van Nimwegen, E., Hausser, J. and Zavolan, M. (2007). Inference of miRNA targets using evolutionary conservation and pathway analysis. BMC Bioinformatics *8*, 69. 3, 37, 49

Galton, F. (1890). Kinship and correlation. 31, 99

García-Martínez, J., Aranda, A. and Pérez-Ortín, J. E. (2004). Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. Molecular cell *15*, 303–313. 92, 93

Gatfield, D. and Izaurralde, E. (2002). REF1/Aly and the additional exon junction complex proteins are dispensable for nuclear mRNA export. J Cell Biol *159*, 579–588. 23

Ghazalpour, A., Bennett, B., Petyuk, V. A., Orozco, L., Hagopian, R., Mungrue, I. N., Farber, C. R., Sinsheimer, J., Kang, H. M., Furlotte, N., Park, C. C., Wen, P.-Z., Brewer, H., Weitz, K., Camp, D. G., Pan, C., Yordanova, R., Neuhaus, I., Tilford, C., Siemers, N., Gargalovic, P., Eskin, E., Kirchgessner, T., Smith, D. J., Smith, R. D. and Lusis, A. J. (2011). Comparative analysis of proteome and transcriptome variation in mouse. PLoS Genet *7*, e1001393. 97, 98

Giardina, C., Pérez-Riba, M. and Lis, J. T. (1992). Promoter melting and TFIID complexes on Drosophila genes in vivo. Genes & development *6*, 2190–2200. 15

Giegerich, R. (1997). From Ukkonen to McCreight and Weiner: A unifying view of linear-time suffix tree construction. Algorithmica *1*, 1–10. 18

Giraldez, A. J., Mishima, Y., Rihel, J., Grocock, R. J., van Dongen, S., Inoue, K., Enright, A. J. and Schier, A. F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science (New York, NY) *312*, 75–79. 28

Godde, J. S., Nakatani, Y. and Wolffe, A. P. (1995). The amino-terminal tails of the core histones and the translational position of the TATA box determine TBP/TFIIA association with nucleosomal DNA. Nucleic acids research *23*, 4557–4564. 9, 10

Goler-Baron, V., Selitrennik, M., Barkai, O., Haimovich, G., Lotan, R. and Choder, M. (2008). Transcription in the nucleus and mRNA decay in the cytoplasm are coupled processes. Genes & development *22*, 2022–2027. 94

Gregory, T. R. (2001). Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biological reviews of the Cambridge Philosophical Society *76*, 65–101. 6

Griffin, T. J., Gygi, S. P., Ideker, T., Rist, B., Eng, J., Hood, L. and Aebersold, R. (2002). Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. Molecular & cellular proteomics : MCP *1*, 323–333. 97

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. and Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77–88. 12, 43

Guo, H., Ingolia, N. T., Weissman, J. S. and Bartel, D. P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835–840. 28, 55

Guo, Y., Xiao, P., Lei, S., Deng, F., Xiao, G. G., Liu, Y., Chen, X., Li, L., Wu, S., Chen, Y., Jiang, H., Tan, L., Xie, J., Zhu, X., Liang, S. and Deng, H. (2008). How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes. Acta biochimica et biophysica Sinica *40*, 426–436. 97, 98

Gygi, S. P., Rochon, Y., Franza, B. R. and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. Molecular and cellular biology *19*, 1720–1730. 20, 97

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M. and Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129–141. 49

Hemberg, M. and Kreiman, G. (2011). Conservation of transcription factor binding events predicts gene expression across species. Nucleic acids research *39*, 7092–7102. 92

Hendrickson, D. G., Hogan, D. J., McCullough, H. L., Myers, J. W., Herschlag, D., Ferrell, J. E. and Brown, P. O. (2009). Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. PLoS biology *7*, e1000238. 28

Hocine, S., Singer, R. H. and Grünwald, D. (2010). RNA processing and export. Cold Spring Harbor perspectives in biology *2*, a000752. 23

Hon, G., Wang, W. and Ren, B. (2009). Discovery and annotation of functional chromatin signatures in the human genome. PLoS computational biology *5*, e1000566. 43, 92

Huang, Y. and Steitz, J. A. (2005). SRprises along a messenger's journey. Molecular cell *17*, 613–615. 24

Huntzinger, E. and Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nature reviews Genetics *12*, 99–110. 26

Hutvágner, G., McLachlan, J., Pasquinelli, A. E., Bálint, E., Tuschl, T. and Zamore, P. D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. Science (New York, NY) *293*, 834–838. 49

Iglesias, N. and Stutz, F. (2008). Regulation of mRNP dynamics along the export pathway. FEBS Lett *582*, 1987–1996. 23

Imbalzano, A. N., Kwon, H., Green, M. R. and Kingston, R. E. (1994). Facilitated binding of TATA-binding protein to nucleosomal DNA. Nature *370*, 481–485. 10

Irie, T., Park, S.-J., Yamashita, R., Seki, M., Yada, T., Sugano, S., Nakai, K. and Suzuki, Y. (2011). Predicting promoter activities of primary human DNA sequences. Nucleic acids research *39*, e75. 92

Iwasaki, S., Kawamata, T. and Tomari, Y. (2009). Drosophila argonaute1 and argonaute2 employ distinct mechanisms for translational repression. Molecular cell *34*, 58–67. 27

Joshi, A. A. and Struhl, K. (2005). Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to Pol II elongation. Molecular cell *20*, 971–978. 44, 52

Kaneko, S. and Manley, J. L. (2005). The mammalian RNA polymerase II C-terminal domain interacts with RNA to suppress transcription-coupled 3' end formation. Molecular cell *20*, 91–103. 15

Karlic, R., Chung, H.-R., Lasserre, J. and Vlahovicek, K. (2010). Histone modification levels are predictive for gene expression. Proceedings of the National Academy of Sciences of the United States of America *107*, 2926–2931. 2, 42, 92

Kayne, P. S., Kim, U. J., Han, M., Mullen, J. R., Yoshizaki, F. and Grunstein, M. (1988). Extremely conserved histone H4 N terminus is dispensable for growth but essential for repressing the silent mating loci in yeast. Cell *55*, 27–39. 9

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature *423*, 241–254. 6

Keogh, M.-C., Kurdistani, S. K., Morris, S. A., Ahn, S. H., Podolny, V., Collins, S. R., Schuldiner, M., Chin, K., Punna, T., Thompson, N. J., Boone, C., Emili, A., Weissman, J. S., Hughes, T. R., Strahl, B. D., Grunstein, M., Greenblatt, J. F., Buratowski, S. and Krogan, N. J. (2005). Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. Cell *123*, 593–605. 12, 44, 52

Kim, S., Kim, H., Fong, N., Erickson, B. and Bentley, D. L. (2011). Pre-mRNA splicing is a determinant of histone H3K36 methylation. Proc Natl Acad Sci U S A *108*, 13564–13569. 18

Kizer, K. O., Phatnani, H. P., Shibata, Y., Hall, H., Greenleaf, A. L. and Strahl, B. D. (2005). A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation. Molecular and cellular biology *25*, 3305–3316. 16, 44, 52

Köhler, A. and Hurt, E. (2007). Exporting RNA from the nucleus to the cytoplasm. Nature reviews Molecular cell biology *8*, 761–773. 23

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X. S. and Ahringer, J. (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. Nature genetics *41*, 376–381. 17

Kornberg, R. D. and Lorch, Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. Cell *98*, 285–294. 42

Kornberg, R. D. and Thomas, J. O. (1974). Chromatin structure; oligomers of the histones. Science (New York, NY) *184*, 865–868. 7

Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693–705. 10, 42

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M. and Rajewsky, N. (2005). Combinatorial microRNA target predictions. Nature genetics *37*, 495–500. 49

Krogan, N. J., Kim, M., Tong, A., Golshani, A., Cagney, G., Canadien, V., Richards, D. P., Beattie, B. K., Emili, A., Boone, C., Shilatifard, A., Buratowski, S. and Greenblatt, J. (2003). Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. Molecular and cellular biology *23*, 4207–4218. 44, 52

Krol, J., Loedige, I. and Filipowicz, W. (2010). The widespread regulation of microRNA biogenesis, function and decay. Nature reviews Genetics *11*, 597–610. 26

Krützfeldt, J., Rajewsky, N., Braich, R., Rajeev, K. G., Tuschl, T., Manoharan, M. and Stoffel, M. (2005). Silencing of microRNAs in vivo with 'antagomirs'. Nature *438*, 685–689. 28

Lagos-Quintana, M., Rauhut, R., Lendeckel, W. and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. Science (New York, NY) *294*, 853–858. 26

Lam, F. H., Steger, D. J. and O'Shea, E. K. (2008). Chromatin decouples promoter threshold from dynamic range. Nature *453*, 246–250. 7

Landthaler, M., Gaidatzis, D., Rothballer, A., Chen, P. Y., Soll, S. J., Dinic, L., Ojo, T., Hafner, M., Zavolan, M. and Tuschl, T. (2008). Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. RNA (New York, NY) *14*, 2580–2596. 49

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25. 18, 57

Lau, N. C., Lim, L. P., Weinstein, E. G. and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science (New York, NY) *294*, 858–862. 26

Le Hir, H., Izaurralde, E., Maquat, L. E. and Moore, M. J. (2000). The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. EMBO J. *19*, 6860–6869. 22

Lee, D. Y., Hayes, J. J., Pruss, D. and Wolffe, A. P. (1993). A positive role for histone acetylation in transcription factor access to nucleosomal DNA. Cell *72*, 73–84. 9, 10

Lee, M. V., Topper, S. E., Hubler, S. L., Hose, J., Wenger, C. D., Coon, J. J. and Gasch, A. P. (2011). A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. Molecular systems biology *7*, 514. 97

Lee, R. C. and Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. Science (New York, NY) *294*, 862–864. 26

Lemaire, M. and Collart, M. A. (2000). The TATA-binding protein-associated factor yTafII19p functionally interacts with components of the global transcriptional regulator Ccr4-Not complex and physically interacts with the Not5 subunit. J Biol Chem *275*, 26925–26934. 95

Lewis, B. P., Shih, I.-h., Jones-Rhoades, M. W., Bartel, D. P. and Burge, C. B. (2003). Prediction of mammalian microRNA targets. Cell *115*, 787–798. 49

Li, B., Carey, M. and Workman, J. L. (2007). The role of chromatin during transcription. Cell *128*, 707–719. 10, 12

Li, B., Howe, L., Anderson, S., Yates, J. R. and Workman, J. L. (2003). The Set2 histone methyltransferase functions through the phosphorylated carboxyl-terminal domain of RNA polymerase II. J Biol Chem *278*, 8897–8903. 16, 44, 52

Li, J., Moazed, D. and Gygi, S. P. (2002). Association of the histone methyltransferase Set2 with RNA polymerase II plays a role in transcription elongation. J Biol Chem *277*, 49383–49388. 16, 44, 52

Li, Y., Trojer, P., Xu, C.-F., Cheung, P., Kuo, A., Drury, W. J., Qiao, Q., Neubert, T. A., Xu, R.-M., Gozani, O. and Reinberg, D. (2009). The target of the NSD family of histone lysine methyltransferases depends on the nature of the substrate. Journal of Biological Chemistry *284*, 34283–34295. 16

Lienert, F., Mohn, F., Tiwari, V. K., Baubec, T., Roloff, T. C., Gaidatzis, D., Stadler, M. B. and Schubeler, D. (2011). Genomic prevalence of heterochromatic H3K9me2 and transcription do not discriminate pluripotent from terminally differentiated cells. PLoS Genet *7*, e1002090. 36, 43

Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., Bartel, D. P., Linsley, P. S. and Johnson, J. M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature *433*, 769–773. 28

Lin, L.-J., Minard, L. V., Johnston, G. C., Singer, R. A. and Schultz, M. C. (2010). Asf1 can promote trimethylation of H3 K36 by Set2. Molecular and cellular biology *30*, 1116–1129. 16

Liu, H. Y., Chiang, Y. C., Pan, J., Chen, J., Salvadore, C., Audino, D. C., Badarinarayana, V., Palaniswamy, V., Anderson, B. and Denis, C. L. (2001). Characterization of CAF4 and CAF16 reveals a functional connection between the CCR4-NOT complex and a subset of SRB proteins of the RNA polymerase II holoenzyme. J Biol Chem *276*, 7541–7548. 95

Llave, C., Xie, Z., Kasschau, K. D. and Carrington, J. C. (2002). Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. Science (New York, NY) *297*, 2053–2056. 26

Loeb, J. N., Howell, R. R. and Tomkins, G. M. (1965). Turnover of ribosomal RNA in rat liver. Science (New York, NY) *149*, 1093–1095. 29

Lotan, R., Bar-On, V. G., Harel-Sharvit, L., Duek, L., Melamed, D. and Choder, M. (2005). The RNA polymerase II subunit Rpb4p mediates decay of a specific class of mRNAs. Genes & development *19*, 3004–3016. 94

Lotan, R., Goler-Baron, V., Duek, L., Haimovich, G. and Choder, M. (2007). The Rpb7p subunit of yeast RNA polymerase II plays roles in the two major cytoplasmic mRNA decay mechanisms. J Cell Biol *178*, 1133–1143. 94

Lu, P., Vogel, C., Wang, R., Yao, X. and Marcotte, E. M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol *25*, 117–124. 97

Lu, R., Markowetz, F., Unwin, R. D., Leek, J. T., Airoldi, E. M., MacArthur, B. D., Lachmann, A., Rozov, R., Ma'ayan, A., Boyer, L. A., Troyanskaya, O. G., Whetton, A. D. and Lemischka, I. R. (2009). Systems-level dynamic analyses of fate change in murine embryonic stem cells. Nature *462*, 358–362. 97, 98

Lucio-Eterovic, A. K., Singh, M. M., Gardner, J. E., Veerappan, C. S., Rice, J. C. and Carpenter, P. B. (2010). Role for the nuclear receptor-binding SET domain protein 1 (NSD1) methyltransferase in coordinating lysine 36 methylation at histone 3 with RNA polymerase II function. Proc Natl Acad Sci U S A *107*, 16952–16957. 16

Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Algenäs, C., Lundeberg, J., Mann, M. and Uhlen, M. (2010). Defining the transcriptome and proteome in three functionally different human cell lines. Molecular systems biology *6*, 450. 98

Luo, W. and Bentley, D. (2004). A ribonucleolytic rat torpedoes RNA polymerase II. Cell *119*, 911–914. 16

Lykke-Andersen, J. and Wagner, E. (2005). Recruitment and activation of mRNA decay enzymes by two ARE-mediated decay activation domains in the proteins TTP and BRF-1. Genes & development *19*, 351–361. 25

Magnello, M. E. (1998). Karl Pearson's mathematization of inheritance: from ancestral heredity to Mendelian genetics (1895-1909)., vol. 55,. The Wellcome Institute for the History of Medicine, London, UK. 31

Margueron, R., Li, G., Sarma, K., Blais, A., Zavadil, J., Woodcock, C. L., Dynlacht, B. D. and Reinberg, D. (2008). Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms. Molecular cell *32*, 503–518. 12

Maroney, P. A., Yu, Y., Fisher, J. and Nilsen, T. W. (2006). Evidence that microRNAs are associated with translating messenger RNAs in human cells. Nature structural & molecular biology *13*, 1102–1107. 27

Masuda, S., Das, R., Cheng, H., Hurt, E., Dorman, N. and Reed, R. (2005). Recruitment of the human TREX complex to mRNA during splicing. Genes & development *19*, 1512–1517. 23

Mathonnet, G., Fabian, M. R., Svitkin, Y. V., Parsyan, A., Huck, L., Murata, T., Biffo, S., Merrick, W. C., Darzynkiewicz, E., Pillai, R. S., Filipowicz, W., Duchaine, T. F. and Sonenberg, N. (2007). MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F. Science (New York, NY) *317*, 1764–1767. 27

Meissner, A., Mikkelsen, T., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B., Nusbaum, C., Jaffe, D., Gnirke, A., Jaenisch, R. and Lander, E. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *7205*, 766–770. 9

Merritt, C., Rasoloson, D., Ko, D. and Seydoux, G. (2008). 3' UTRs are the primary regulators of gene expression in the C. elegans germline. Curr Biol *18*, 1476–1482. 25

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S. and Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560. 1, 12, 17, 35, 36, 43, 52

Miller, C., Schwalb, B., Maier, K., Schulz, D., Dümcke, S., Zacher, B., Mayer, A., Sydow, J., Marcinowski, L., Dölken, L., Martin, D. E., Tresch, A. and Cramer, P. (2011). Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. Molecular systems biology *7*, 458. 96

Millevoi, S. and Vagner, S. (2010). Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. Nucleic acids research *38*, 2757–2774. 41

Min, I. M., Waterfall, J. J., Core, L. J., Munroe, R. J., Schimenti, J. and Lis, J. T. (2011). Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. Genes & development *25*, 742–754. 15, 56

Mohn, F., Weber, M., Rebhan, M., Roloff, T. C., Richter, J., Stadler, M. B., Bibel, M. and Schübeler, D. (2008). Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Molecular cell *30*, 755–766. 8, 13, 36, 43, 53

Molin, C., Jauhiainen, A., Warringer, J., Nerman, O. and Sunnerhagen, P. (2009). mRNA stability changes precede changes in steady-state mRNA amounts during hyperosmotic stress. RNA (New York, NY) *15*, 600–614. 96

Montero Llopis, P., Jackson, A. F., Sliusarenko, O., Surovtsev, I., Heinritz, J., Emonet, T. and Jacobs-Wagner, C. (2010). Spatial organization of the flow of genetic information in bacteria. Nature *466*, 77–81. 5

Moore, M. J. and Proudfoot, N. J. (2009). Pre-mRNA processing reaches back to transcription and ahead to translation. Cell *136*, 688–700. 22

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods *5*, 621–628. 1, 35

Mukherji, S., Ebert, M. S., Zheng, G. X. Y., Tsang, J. S., Sharp, P. A. and van Oudenaarden, A. (2011). MicroRNAs can generate thresholds in target gene expression. Nature genetics *43*, 854–859. 56

Munoz, J., Low, T. Y., Kok, Y. J., Chin, A., Frese, C. K., Ding, V., Choo, A. and Heck, A. J. R. (2011). The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. Molecular systems biology *7*, 550. 97

Murchison, E. P., Partridge, J. F., Tam, O. H., Cheloufi, S. and Hannon, G. J. (2005). Characterization of Dicer-deficient murine embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America *102*, 12135–12140. 49

Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S. and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Molecular systems biology *7*, 548. 97

Nilsen, T. W. and Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature *463*, 457–463. 41

Nottrott, S., Simard, M. J. and Richter, J. D. (2006). Human let-7a miRNA blocks protein production on actively translating polyribosomes. Nature structural & molecular biology *13*, 1108–1114. 27

Olsen, P. H. and Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. Developmental biology *216*, 671–680. 27

Ouyang, Z., Zhou, Q. and Wong, W. H. (2009). ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America *106*, 21521–21526. 42

Pal, M. and Luse, D. S. (2003). The initiation-elongation transition: lateral mobility of RNA in RNA polymerase II complexes is greatly reduced at +8/+9 and absent by +23. Proceedings of the National Academy of Sciences of the United States of America *100*, 5700–5705. 15

Parker, R. and Song, H. (2004). The enzymes and control of eukaryotic mRNA turnover. Nature structural & molecular biology *11*, 121–127. 25

Pei, Y., Schwer, B. and Shuman, S. (2003). Interactions between fission yeast Cdk9, its

cyclin partner Pch1, and mRNA capping enzyme Pct1 suggest an elongation checkpoint for mRNA quality control. J Biol Chem *278*, 7180–7188. 15

Petersen, C. P., Bordeleau, M.-E., Pelletier, J. and Sharp, P. A. (2006). Short RNAs repress translation after initiation in mammalian cells. Molecular cell *21*, 533–542. 27

Phatnani, H. P. and Greenleaf, A. L. (2006). Phosphorylation and functions of the RNA polymerase II CTD. Genes & development *20*, 2922–2936. 13

Pheasant, M. (2007). Raising the estimate of functional human sequences. Genome research *17*, 1245–1253. 35

Pillai, R. S., Bhattacharyya, S. N., Artus, C. G., Zoller, T., Cougot, N., Basyuk, E., Bertrand, E. and Filipowicz, W. (2005). Inhibition of translational initiation by Let-7 MicroRNA in human cells. Science (New York, NY) *309*, 1573–1576. 27

Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M., Hannett, N. M., Lee, T. I., Bell, G. W., Walker, K., Rolfe, P. A., Herbolsheimer, E., Zeitlinger, J., Lewitter, F., Gifford, D. K. and Young, R. A. (2005). Genome-wide map of nucleosome acetylation and methylation in yeast. Cell *122*, 517–527. 10, 11, 12, 14, 17, 43, 44, 52

Prohaska, S. J., Stadler, P. F. and Krakauer, D. C. (2010). Innovation in gene regulation: the case of chromatin computation. Journal of theoretical biology *265*, 27–44. 6

Proudfoot, N. (2004). New perspectives on connecting messenger RNA 3' end formation to transcription. Curr Opin Cell Biol *16*, 272–278. 22

Proudfoot, N. J., Furger, A. and Dye, M. J. (2002). Integrating mRNA processing with transcription. Cell *108*, 501–512. 20

Qiao, Q., Li, Y., Chen, Z., Wang, M., Reinberg, D. and Xu, R.-M. (2011). The structure of NSD1 reveals an autoregulatory mechanism underlying histone H3K36 methylation. Journal of Biological Chemistry *286*, 8361–8368. 16

Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., Amit, I. and Regev, A. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. Nat Biotechnol *5*, 436–442. 2, 3, 29, 47, 55, 94

Rahl, P. B., Lin, C. Y., Seila, A. C., Flynn, R. A., McCuine, S., Burge, C. B., Sharp, P. A. and Young, R. A. (2010). c-Myc regulates transcriptional pause release. Cell *141*, 432–445. 43

Rakyan, V. K., Hildmann, T., Novik, K. L., Lewin, J., Tost, J., Cox, A. V., Andrews, T. D., Howe, K. L., Otto, T., Olek, A., Fischer, J., Gut, I. G., Berlin, K. and Beck, S. (2004). DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. PLoS biology *2*, e405. 8

Rasmussen, E. B. and Lis, J. T. (1993). In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. Proceedings of the National Academy of Sciences of the United States of America *90*, 7923–7927. 15, 21

Rea, S., Eisenhaber, F., O'Carroll, D., Strahl, B. D., Sun, Z. W., Schmid, M., Opravil, S., Mechtler, K., Ponting, C. P., Allis, C. D. and Jenuwein, T. (2000). Regulation of chromatin structure by site-specific histone H3 methyltransferases. Nature *406*, 593–599. 10

Rehmsmeier, M., Steffen, P., Hochsmann, M. and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. RNA (New York, NY) *10*, 1507–1517. 49

Rehwinkel, J., Behm-Ansmant, I., Gatfield, D. and Izaurralde, E. (2005). A crucial role for GW182 and the DCP1:DCP2 decapping complex in miRNA-mediated gene silencing. RNA (New York, NY) *11*, 1640–1647. 28

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B. and Bartel, D. P. (2002). Prediction of plant microRNA targets. Cell *110*, 513–520. 26

Roh, T.-Y., Cuddapah, S., Cui, K. and Zhao, K. (2006). The genomic landscape of histone modifications in human T cells. Proceedings of the National Academy of Sciences of the United States of America *103*, 15782–15787. 12

Rollins, R. A., Haghighi, F., Edwards, J. R., Das, R., Zhang, M. Q., Ju, J. and Bestor, T. H. (2006). Large-scale structure of genomic methylation patterns. Genome research *16*, 157–163. 8

Rosner, A. and Rinkevich, B. (2007). The DDX3 subfamily of the DEAD box helicases: divergent roles as unveiled by studying different organisms and in vitro assays. Current Medicinal Chemistry *14*, 2517–2525. 22

Sanders, S. L., Jennings, J., Canutescu, A., Link, A. J. and Weil, P. A. (2002). Proteomics of the eukaryotic transcription machinery: identification of proteins associated with components of yeast TFIID by multidimensional mass spectrometry. Molecular and cellular biology *22*, 4723–4738. 95

Santos-Rosa, H., Schneider, R., Bernstein, B. E., Karabetsou, N., Morillon, A., Weise, C., Schreiber, S. L., Mellor, J. and Kouzarides, T. (2003). Methylation of histone H3 K4 mediates association of the Isw1p ATPase with chromatin. Molecular cell *12*, 1325–1332. 12

Saunders, A., Core, L. J. and Lis, J. T. (2006). Breaking barriers to transcription elongation. Nature reviews Molecular cell biology *7*, 557–567. 13, 15

Schübeler, D., Lorincz, M. C., Cimbora, D. M., Telling, A., Feng, Y. Q., Bouhassira, E. E. and Groudine, M. (2000). Genomic targeting of methylated DNA: influence of methylation on transcription, replication, chromatin structure, and histone acetylation. Molecular and cellular biology *20*, 9103–9112. 9

Schubeler, D., MacAlpine, D. M., Scalzo, D., Wirbelauer, C., Kooperberg, C., van Leeuwen, F., Gottschling, D. E., O'Neill, L. P., Turner, B. M., Delrow, J., Bell, S. P. and Groudine, M. (2004). The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. Genes & development *18*, 1263–1271. 10, 11, 12

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342. 2, 3, 47, 55, 94, 95, 98

Schwartz, S., Meshorer, E. and Ast, G. (2009). Chromatin organization marks exon-intron structure. Nature structural & molecular biology *16*, 990–995. 17, 18

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I. K., Wang, J.-P. Z. and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772–778. 92

Segal, E. and Widom, J. (2009). From DNA sequence to transcriptional behaviour: a quantitative approach. Nature reviews Genetics *10*, 443–456. 41, 92

Seggerson, K., Tang, L. and Moss, E. G. (2002). Two genetic circuits repress the Caenorhabditis elegans heterochronic gene lin-28 after translation initiation. Developmental biology *243*, 215–225. 27

Seila, A. C., Calabrese, J. M., Levine, S. S., Yeo, G. W., Rahl, P. B., Flynn, R. A., Young, R. A. and Sharp, P. A. (2008). Divergent transcription from active promoters. Science (New York, NY) *322*, 1849–1851. 1, 35

Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R. and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58–63. 28

Shalem, O., Dahan, O., Levo, M., Martinez, M. R., Furman, I., Segal, E. and Pilpel, Y. (2008). Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. Molecular systems biology *4*, 223. 94

Shalem, O., Groisman, B., Choder, M., Dahan, O. and Pilpel, Y. (2011). Transcriptome kinetics is governed by a genome-wide coupling of mRNA production and degradation: a role for RNA Pol II. PLoS Genet *7*, e1002273. 96

Sharova, L. V., Sharov, A. A., Nedorezov, T., Piao, Y., Shaik, N. and Ko, M. S. H. (2009). Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. DNA research : an international journal for rapid publication of reports on genes and genomes *16*, 45–58. 24, 46

Shaw, G. and Kamen, R. (1986). A conserved AU sequence from the 3' untranslated region of GM-CSF mRNA mediates selective mRNA degradation. Cell *46*, 659–667. 25

She, X., Rohl, C. A., Castle, J. C., Kulkarni, A. V., Johnson, J. M. and Chen, R. (2009). Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. BMC genomics *10*, 269. 53

Shogren-Knaak, M., Ishii, H., Sun, J.-M., Pazin, M. J., Davie, J. R. and Peterson, C. L. (2006). Histone H4-K16 acetylation controls chromatin structure and protein interactions. Science (New York, NY) *311*, 844–847. 10

Shuman, S. (2001). Structure, mechanism, and evolution of the mRNA capping apparatus. Progress in nucleic acid research and molecular biology *66*, 1–40. 21

Simon, J. A. and Kingston, R. E. (2009). Mechanisms of polycomb gene silencing: knowns and unknowns. Nature reviews Molecular cell biology *10*, 697–708. 12

Simpson, V. J., Johnson, T. E. and Hammen, R. F. (1986). Caenorhabditis elegans DNA does not contain 5-methylcytosine at any time during development or aging. Nucleic acids research *14*, 6711–6719. 8

Sinkkonen, L., Hugenschmidt, T., Berninger, P., Gaidatzis, D., Mohn, F., Artus-Revel, C. G., Zavolan, M., Svoboda, P. and Filipowicz, W. (2008). MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. Nature structural & molecular biology *15*, 259–267. 37, 48, 55

Spies, N., Nielsen, C. B., Padgett, R. A. and Burge, C. B. (2009). Biased chromatin signatures around polyadenylation sites and exons. Molecular cell *36*, 245–254. 17

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., Wirbelauer, C., Oakeley, E. J., Gaidatzis, D., Tiwari, V. K. and Schubeler, D. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495. 8, 9

Stark, A., Brennecke, J., Bushati, N., Russell, R. B. and Cohen, S. M. (2005). Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. Cell *123*, 1133–1146. 25, 28, 53

Stark, H. and Lührmann, R. (2006). Cryo-electron microscopy of spliceosomal components. Annual review of biophysics and biomolecular structure *35*, 435–457. 21

Strahl, B. D., Grant, P. A., Briggs, S. D., Sun, Z.-W., Bone, J. R., Caldwell, J. A., Mollah, S., Cook, R. G., Shabanowitz, J., Hunt, D. F. and Allis, C. D. (2002). Set2 is a nucleosomal histone H3-selective methyltransferase that mediates transcriptional repression. Molecular and cellular biology *22*, 1298–1306. 44, 52

Struhl, K. (1999). Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. Cell *98*, 1–4. 7

Sun, M., Schwalb, B., Schulz, D., Pirkl, N., Etzold, S., Larivière, L., Maier, K. C., Seizl, M., Tresch, A. and Cramer, P. (2012). Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. Genome research *0*, 0. 96

Sun, X. J. (2005). Identification and Characterization of a Novel Human Histone H3 Lysine 36-specific Methyltransferase. Journal of Biological Chemistry *280*, 35261–35271. 17, 44, 52

Suzuki, M. M. and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nature reviews Genetics *9*, 465–476. 8

Taft, R., Pheasant, M. and Mattick, J. (2006). The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays *3*, 288–299. 35

Taunton, J., Hassig, C. A. and Schreiber, S. L. (1996). A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. Science (New York, NY) *272*, 408–411. 10

Taverna, S. D., Ilin, S., Rogers, R. S., Tanny, J. C., Lavender, H., Li, H., Baker, L., Boyle, J., Blair, L. P., Chait, B. T., Patel, D. J., Aitchison, J. D., Tackett, A. J. and Allis, C. D. (2006). Yng1 PHD finger binding to H3 trimethylated at K4 promotes NuA3 HAT activity at K14 of H3 and transcription at a subset of targeted ORFs. Molecular cell *24*, 785–796. 12

Thomas, M. C. and Chiang, C.-M. (2006). The general transcription machinery and general cofactors. Critical reviews in biochemistry and molecular biology *41*, 105–178. 14

Tiwari, V. K., Stadler, M. B., Wirbelauer, C., Paro, R., Schubeler, D. and Beisel, C. (2012). A chromatin-modifying function of JNK during stem cell differentiation. Nature genetics *44*, 94–100. 36, 43

Tucker, M., Valencia-Sanchez, M. A., Staples, R. R., Chen, J., Denis, C. L. and Parker, R. (2001). The transcription factor associated Ccr4 and Caf1 proteins are components of the major cytoplasmic mRNA deadenylase in Saccharomyces cerevisiae. Cell *104*, 377–386. 94

Ujvári, A. and Luse, D. S. (2006). RNA emerging from the active site of RNA polymerase II interacts with the Rpb7 subunit. Nature structural & molecular biology *13*, 49–54. 15

Vakoc, C. R., Sachdeva, M. M., Wang, H. and Blobel, G. A. (2006). Profile of histone lysine methylation across transcribed mammalian chromatin. Molecular and cellular biology *26*, 9185–9195. 17, 43

Valencia, P., Dias, A. P. and Reed, R. (2008). Splicing promotes rapid and efficient mRNA export in mammalian cells. Proc Natl Acad Sci U S A *105*, 3386–3391. 22, 23, 24

van Dongen, S., Abreu-Goodger, C. and Enright, A. (2008). Detecting microRNA binding and siRNA off-target effects from expression data. Nature methods *5*, 1023–1025. 49

van Leeuwen, F., Gafken, P. R. and Gottschling, D. E. (2002). Dot1p modulates silencing in yeast by methylation of the nucleosome core. Cell *109*, 745–756. 12

van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. Trends Genet *19*, 479–484. 6

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. Nature reviews Genetics *10*, 252–263. 6

Vermeulen, M., Mulder, K. W., Denissov, S., Pijnappel, W. W. M. P., van Schaik, F. M. A., Varier, R. A., Baltissen, M. P. A., Stunnenberg, H. G., Mann, M. and Timmers, H. T. M. (2007). Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. Cell *131*, 58–69. 12

Vettese-Dadey, M., Grant, P. A., Hebbes, T. R., Crane-Robinson, C., Allis, C. D. and Workman, J. L. (1996). Acetylation of histone H4 plays a primary role in enhancing transcription factor binding to nucleosomal DNA in vitro. EMBO J. *15*, 2508–2518. 10

Vinogradov, A. E. (2004). Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet *20*, 248–253. 53

Vogel, C., Abreu, R. d. S., Ko, D., Le, S.-Y., Shapiro, B. A., Burns, S. C., Sandhu, D., Boutz, D. R., Marcotte, E. M. and Penalva, L. O. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Molecular systems biology *6*, 400. 97

Wagner, E. J. and Carpenter, P. B. (2012). Understanding the language of Lys36 methylation at histone H3. Nature reviews Molecular cell biology *13*, 115–126. 52

Wang, Z., Gerstein, M. and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics *10*, 57–63. 1, 35

Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M. Q. and Zhao, K. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nature genetics *40*, 897–903. 10, 92

Waterston, R., Lindblad-Toh, K., Birney, E., Rogers, J. and others (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–562. 6, 21

Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Pääbo, S., Rebhan, M. and Schübeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nature genetics *39*, 457–466. 8, 9, 12, 42, 56

Weiner, P. (1973). Linear pattern matching algorithms. Switching and Automata Theory *1*, 1–2. 18

Wen, Y. and Shatkin, A. J. (1999). Transcription elongation factor hSPT5 stimulates mRNA capping. Genes & development *13*, 1774–1779. 15

Wirbelauer, C., Bell, O. and Schubeler, D. (2005). Variant histone H3.3 is deposited at sites of nucleosomal displacement throughout transcribed genes while active histone modifications show a promoter-proximal bias. Genes & development *19*, 1761–1766. 56

Wolffe, A. P. and Pruss, D. (1996). Targeting chromatin disruption: Transcription regulators that acetylate histones. Cell *84*, 817–819. 10

Workman, J. L. and Kingston, R. E. (1998). Alteration of nucleosome structure as a mechanism of transcriptional regulation. Annual review of biochemistry *67*, 545–579. 7

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V. and Romano, L. A. (2003). The evolution of transcriptional regulation in eukaryotes. Molecular biology and evolution *20*, 1377–1419. 6

Wu, C.-H., Lee, C., Fan, R., Smith, M. J., Yamaguchi, Y., Handa, H. and Gilmour, D. S. (2005). Molecular characterization of Drosophila NELF. Nucleic acids research *33*, 1269–1279. 16

Wyrick, J. J., Holstege, F. C., Jennings, E. G., Causton, H. C., Shore, D., Grunstein, M., Lander, E. S. and Young, R. A. (1999). Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. Nature *402*, 418–421. 42

Xiao, T., Hall, H., Kizer, K. O., Shibata, Y., Hall, M. C., Borchers, C. H. and Strahl, B. D. (2003). Phosphorylation of RNA polymerase II CTD regulates H3 methylation in yeast. Genes & development *17*, 654–663. 16, 44, 52

Yamada, T., Yamaguchi, Y., Inukai, N., Okamoto, S., Mura, T. and Handa, H. (2006). P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation. Molecular cell *21*, 227–237. 16

Yang, E., van Nimwegen, E., Zavolan, M., Rajewsky, N., Schroeder, M., Magnasco, M. and Darnell, J. E. (2003). Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. Genome research *13*, 1863–1872. 24

Youdell, M. L., Kizer, K. O., Kisseleva-Romanova, E., Fuchs, S. M., Duro, E., Strahl, B. D. and Mellor, J. (2008). Roles for Ctk1 and Spt6 in regulating the different methylation states of histone H3 lysine 36. Molecular and cellular biology *28*, 4915–4926. 16

Young, M. D., Willson, T. A., Wakefield, M. J., Trounson, E., Hilton, D. J., Blewitt, M. E., Oshlack, A. and Majewski, I. J. (2011). ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. Nucleic acids research *39*, 7415–7427. 43

Yuan, W., Xie, J., Long, C., Erdjument-Bromage, H., Ding, X., Zheng, Y., Tempst, P., Chen, S., Zhu, B. and Reinberg, D. (2009). Heterogeneous nuclear ribonucleoprotein L Is a subunit of human KMT3a/Set2 complex required for H3 Lys-36 trimethylation activity in vivo. J Biol Chem *284*, 15701–15707. 17, 44, 52

Zhang, Z. and Gilmour, D. S. (2006). Pcf11 is a termination factor in Drosophila that dismantles the elongation complex by bridging the CTD of RNA polymerase II to the nascent transcript. Molecular cell *21*, 65–74. 16

Zhang, Z. and Zhang, M. Q. (2011). Histone modification profiles are predictive for tissue/cell-type specific expression of both protein-coding and microRNA genes. BMC Bioinformatics *12*, 155. 92

Acronyms

ARE AU rich element.

ChIP chromatin immunoprecipitation.

CTD carboxyl-terminal domain.

EEC early elongation complex.

EJC exon junction complex.

GTF general transcription factor.

H3K27me3 tri-methylation of lysine 27 at histone tail H3.

H3K36me3 tri-methylation of lysine 36 at histone tail H3.

H3K4me2 di-methylation of lysine 4 at histone tail H3.

HIV human immunodeficiency virus.

HMT histone methyl-transferase.

hnRNPL heterogenous ribonucleoprotein L.

IRES internal ribosome entry site.

ITC initially transcribing complex.

LMR low methylated region.

MBD methyl-CpG-binding domain protein.

MES-4 maternal effect sterile 4.

MET-1 histone-methyltransferase-like 1.

miRISC miRNA induced silencing complex.

miRNA micro RNA.

mRNA messenger RNA.

mRNP messenger ribonucleoproteins.

NPC nuclear pore complex.

NSD1 nuclear receptor binding SET domain protein 1.

P-TEFb positive transcription-elongation factor-b.

PIC pre-initiation complex.

PTM post-translational modifications.

RBP RNA binding protein.

RNAP II RNA polymerase II.

Rpb1 large subunit of RNAP II.

Set2 SET domain-containing.

SETD2 SET domain-containing 2.

snRNP small nuclear RNP.

SR serine/arginine-rich.

TBP TATA binding protein.

TF transcription factor.

TSS transcription start site.

UTR un-translated region.

Chapter 5

Curriculum vitae

Name:	Sylvia Tippmann
Institute:	Friedrich Miescher Institute for Biomedical Research
Address:	Maulbeerstrasse 66, CH-4058 Basel
Email:	sylvia.tippmann@fmi.ch
Date of birth:	15th May 1984
Place of birth:	Karl-Marx-Stadt (today Chemnitz), Germany

Education

01/2008 - present	**Ph.D Student at Friedrich Miescher Institute for Biomedical Research, Basel - Switzerland** Group Dirk Schuebeler - Propagation and dynamics of epigenetic states
01/2006 - 10/2006	**Bachelor of Science (Honours Class I) in Biochemistry at Univeristy of Queensland, Institute for Molecular Bioscience, Brisbane - Australia** Group John Mattick - Rnomics: noncoding RNA in mammalian evolution and development Thesis: 'Prediction of H/ACA-box snoRNAs in Drosophila melanogaster'
10/2002 - 10/2005	**Bachelor of Science in Bioinformatics and Genome Research at Bielefeld University - Germany** Group Robert Giegerich - Practical Computer Science Thesis: 'In Silico Dicer - Intrinsic and Extrinsic Prediction of mature miRNA'

Work Experiences

10/2007 - 12/2007	Staff Scientist at University of Leipzig, Insitute for Informatics, Department of Bioinformatics - Germany: Group Prof Peter Stadler
08/2007 - 09/2007	Staff Scientist at Institute for Theoretical Chemistry, University of Vienna - Austria: Group Ivo Hofacker
06/2007 - 07/2007	Staff Scientist at IMBA, Center for Integrative Bioinformatics, Vienna - Austria: Group Arndt von Haeseler
04/2007 - 05/2007	Laboratory Experience (ISH, RACE, Northern) at IMBA, Vienna - Austria: Group Javier Martinez
01/2007 - 03/2007	Laboratory Experience (RT-PCR, cell culture) at Max F. Perutz Laboratories, Vienna - Austria: Group Prof Renee Schroeder
10/2005 - 12/2005	Research Assistant at Bielefeld University - Germany: Group for Practical Informatics, Prof Robert Giegerich

Scholarships

09/2010 - present	PhD Scholarship Novartis Forschungsstiftung
09/2008 - 09/2010	Boehringer Ingelheim PhD Scholarship
01/2006 - 12/2006	Excellence Scholarship for Honours Program at Univeristy of Queensland Australia from DAAD (German Academic Exchange Service)

Conferences, Posters and Presentations

10/2011	"Frontiers in Epigenetics" Workshop [Baeza, ES] *Poster*
09/2011	Annual TBI Autumn Seminar [Decin, CZ] *Presentation*
08/2011	CSHL Meeting on Eukaryotic mRNA processing [NJ, USA] *Poster*
05/2011	Joint PhD Conference (FMI/CRG) [Pyrenees, ES] *Presentation*
02/2011	Annual TBI Winter Seminar [Bled, SI] *Presentation*
07/2010	ISMB/ECCB [Boston, USA] *Poster*
05/2010	Microsymposium on small RNAs [Vienna, AT]
09/2009	Otto Warburg International Summer School on Epigenomics [Berlin, D]
07/2009	Bioconductor Meeting [Seattle, USA]
06/2009	BC^2 Conference [Basel, CH]
06/2009	Joint PhD Conference (FMI/MRC) [Luzern, CH] *Poster*
05/2009	Boehringer Course: Communicating Science to the public [Lautrach, D]
10/2008	Annual TBI Autumn Seminar [Decin, CZ] *Presentation*
09/2008	ECCB [Sardinia, IT] *Poster*
03/2008	Bioconductor Training [Brixen, IT]
02/2008	Annual TBI Winter Seminar [Bled, SI] *Presentation*

Organized Conferences

05/2011	Member of Organizing Committee of a joint PhD Conference (FMI Basel and CRG Barcelona) [Pyrenees, ES]
06/2009	Member of Organizing Committee of a joint PhD Conference (FMI Basel and MRC London) [Luzern, CH]