COMPUTATIONAL ANALYSIS OF TRANSCRIPTIONAL AND POST-TRANSCRIPTIONAL REGULATION OF GENE EXPRESSION

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
ANDREAS JOHANNES GRUBER
aus
Österreich

Basel, 2017

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Dr. Mihaela Zavolan, Dr. Helge Grosshans

Basel, den 10.11.2015

Prof. Dr. Jörg Schibler
Dekan

ABSTRACT

The regulation of gene expression is fundamental to all life on Earth. Dynamic but precise control is vital to cell survival and function, and takes place at various tightly interwoven levels. In this thesis, we review and study the crosstalk between different types of regulators, including epigenetic regulators, transcription factors (TFs), RNA-binding proteins (RBPs) and microRNAs (miRNAs). First, we focus on the interplay between miRNAs and other types of regulators, in particular TFs and epigenetic regulators, both of which are strongly enriched among the predicted targets of miRNAs. Indeed, the direct interplay of miRNAs with other regulators that have genome-wide impact is one possible explanation for the reported importance of miRNAs to fundamental biological processes, including cell fate. We introduce a computational strategy that we apply in order to infer the transcription regulatory circuitries that act downstream of embryonic miRNAs. More precisely, we analyze genome-wide expression changes with an extended motif activity response analysis (MARA) model in order to identify transcriptional regulators that are direct targets of embryonic miRNAs and change in activity upon expression of the miRNAs. We experimentally validate our most promising predictions and integrate the extended MARA model into an automated system in order to make it available to other researchers. We demonstrate its application by modeling diverse high-throughput datasets, including paired liver biopsies of patients with chronic hepatitis C virus infections. Finally, we study alternative cleavage and polyadenylation, a process that impacts gene expression in various ways, including modulating the presence of *cis*-regulatory elements, such as miRNA and RBP binding sites, which tend to be located at the 3' ends of transcripts. We demonstrate that global shortening of untranslated transcript regions, which is associated with proliferative states, has a very limited effect on mRNA stability and protein output. By analyzing a large array of high-throughput 3' end sequencing data, we create comprehensive catalogs of 3' end processing sites for both human and mouse. Moreover, we identify novel *cis*-regulatory motifs that are involved in cleavage and polyadenylation, and point out a regulator, HNRNPC, that binds to one of the motifs, thereby globally impacting the usage of cleavage and polyadenylation sites.

PUBLICATIONS

Work discussed in this PhD thesis has appeared previously in the following publications:

1. Andreas J. Gruber and Mihaela Zavolan. Modulation of epigenetic regulators and cell fate decisions by miRNAs. *Epigenomics*, 5(6):671–683, December 2013. ISSN: 1750–1911 (Print), 1750–192X (Electronic). PMID: 24283881. doi: 10.2217/epi.13.65.

2. Michael T. Dill, Zuzanna Makowska, Gaia Trincucci, Andreas J. Gruber, Julia E. Vogt, Magdalena Filipowicz, Diego Calabrese, Ilona Krol, Daryl T. Lau, Luigi Terracciano, Erik van Nimwegen, Volker Roth, and Markus H. Heim. Pegylated ifn-α regulates hepatic gene expression through transient Jak/STAT activation. *The Journal of Clinical Investigation*, 124(4):1568–1581, April 2014. ISSN: 0021–9738 (Print), 1558–8238 (Electronic). PMID: 24569457. doi: 10.1172/JCI70408.

3. Piotr J. Balwierz, Mikhail Pachkov, Phil Arnold, Andreas J. Gruber, Mihaela Zavolan, and Erik van Nimwegen. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Research*, 24(5):869–884, May 2014. ISSN: 1088–9051 (Print), 1549–5469 (Electronic). PMID: 24515121. doi: 10.1101/gr.169508.113.

4. Andreas J. Gruber, William A. Grandy, Piotr J. Balwierz, Yoana A. Dimitrova, Mikhail Pachkov, Constance Ciaudo, Erik van Nimwegen, and Mihaela Zavolan. Embryonic stem cell-specific microRNAs contribute to pluripotency by inhibiting regulators of multiple differentiation pathways. *Nucleic Acids Research*, 42(14): 9313–9326, August 2014. ISSN: 0305–1048 (Print), 1362–4962 (Electronic). PMID: 25030899. doi: 10.1093/nar/gku544.

5. Andreas R. Gruber, Georges Martin, Philipp Mueller, Alexander Schmidt, Andreas J. Gruber, Rafal Gumienny, Nitish Mittal, Rajesh Jayachandran, Jean Pieters, Walter Keller, Erik van Nimwegen, and Mihaela Zavolan. Global 3' UTR shortening has a limited effect on protein abundance in proliferating T cells. *Nature Communications*, 5:5465, November 2014. ISSN: 2041–1723 (Electronic). PMID: 25413384. doi: 10.1038/ncomms6465.

6. Andreas J. Gruber, Ralf Schmidt, Andreas R. Gruber, Georges Martin, Manuel Belmadani, Walter Keller, Mihaela Zavolan. A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Research*, 26(8):1145–1159, August 2016. ISSN: 1088–9051 (Print), 1549–5469 (Electronic). PMID: 27382025. doi: 10.1101/gr.202432.115.

ACKNOWLEDGMENTS

This PhD thesis would not have been possible without the help and advice of many people. It is a pleasure to thank those who have positively contributed to my journey towards the successful conclusion of this work!

Great thanks to my family and friends, for their warm and helping support in everything I've done throughout the past years! Their contribution was crucial to this work!

I am grateful to Helge Grosshans, for his open-minded and helpful attitude, taking time out of his busy schedule to review my work and serve as member of my PhD committee. Thanks a lot!

Special thanks go to Erik van Nimwegen, who provided me the opportunity to contribute to his work and helped me to successfully complete several interesting projects. He made his support available in many ways, being always open to questions and responding with innovative ideas and clever solutions to all sorts of scientific and technical challenges. Without him my PhD would have taken another, certainly worse, direction.

Most of all, I would like to warmly thank Mihaela Zavolan, for receiving me in her lab and providing me with many broad-minded opportunities. Certainly, she has supported me in every sense, being always on the spot and creating an incredibly exciting atmosphere for conducting science. She let me explore my own interests, while bailing me out of difficulties whenever necessary. I always knew that I could rely on her, no matter what. Certainly, I'm positive about her caring and involved leadership, her progressive personality and her passion for science. The last few years were the so far most interesting of my life. Sincere thanks to Mihaela for all her support, making this possible!

Great thanks to Yvonne Steger, Sarah Güthe and Rita Manohar for their amazing administrative and personal support throughout the years of my PhD.

I'm grateful to Felix Naef, David Gatfield and Walter Keller for their help and advise on my projects!

Moreover, I would like to thank the people that have supported me with their expertise and a well working and robust high-performance computing environment, particularly Konstantin Arnold, Thierry Sengstag, Pablo Escobar López, Martin Jacquot and Jan Welker. Thanks!

Also, I would like to show my gratitude to the Werner-Siemens Foundation and the University of Basel for awarding me a Werner-Siemens Fellowship for Excellence. It provided me with independent funding and allowed me to rotate through a few research groups of my choice, before I had to commit myself to a group. Thanks for providing such an opportunity!

Thanks to my colleague and friend Jean Hausser, for accompanying me through the starting time of my PhD and staying in touch throughout the years!

Thanks to my colleague Chris Field and my friend Trey Nunley for being so kind and checking some parts of my thesis for linguistic errors!

I'm grateful to all my colleagues from the Zavolan, van Nimwegen and Schwede labs, for their help in various matters and for sharing their knowledge and views on scientific questions!

And last, but not least, many thanks to all my scientific collaborators!

To all I have mentioned and those I forgot: **Thanks a lot!**

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Although our planet exists for 4.5 billion years [1, 2] the oldest ecosystem discovered so far dates back to almost 3.5 billion years [3]. Hence, life has emerged relatively early in Earth's history. Various theories about the emergence of life from inanimate matter have been proposed. While many aspects of life's origin remain a matter of speculation, most models build on the pioneering work of Stanley Miller and Harold Urey published in the 1950s [4, 5]. The so called "Miller-Urey" experiment has demonstrated that inorganic chemicals, as have been present on the "primitive" Earth, favor chemical reactions that synthesize organic compounds, including amino acids. How life could have emerged from these compounds is not entirely clear. However, obviously today our planet is densely populated by living organisms of incredible diversity. And amazingly, despite all this variety, there exist fundamentals that are universal to all life on earth.
The minimum self-reproducing, basic building block of all living organisms is the cell. Although bacteria, animal and plant cells differ in many aspects, they all share the same basic components and mechanisms, which allow them to read, interpret and inherit their genetic information, that is stored in form of double-stranded deoxyribonucleic acid (DNA). Indeed, the code that enables the production of specific proteins from DNA is so universal, that one can take a piece of human DNA, place it into a bacterium and the code will still be successfully translated into protein and without difficulty inherited to its progeny. The fact that many basic, cellular components are so highly conserved, that their genetic information can be found within every cell on Earth, strongly suggests that all life on our planet, from bacteria, to flowers and human, stems from a common ancestor [6]. Studying and understanding cell biology enables the targeted development of effective drugs and therapies for all kinds of diseases. Moreover it provides us with answers to fundamental questions we have about our own existence. Thus, conducting my PhD was not solely about discovering and characterizing principles and mechanisms of cell biology, but also about gaining personal knowledge and insights on the nature of life, which has in addition helped me to shape thoughts and answers to deep, personal questions.

1.1 REGULATION OF GENE EXPRESSION

Parts of each cell's DNA – genes – serve as templates for the synthesis of ribonucleic acid (RNA) polymers, in a process called "transcription". Transcribed RNA can either be an operating cellular component itself, such as ribosomal RNA (rRNA), or serve as template for the production of amino acid polymers, i.e. peptides and proteins, in a process termed "translation".

In a broader sense, any process by which the information encoded in a gene is used to synthesize a gene product, can be referred to as gene expression [7]. The genome of a cell contains all the genetic information necessary to instruct the cell how to function. However, even though the cells of an individual have identical genome sequences, they can greatly differ in their phenotype and functional characteristics (e.g. hepatocytes versus neurons). That is, because distinct cell types express different sets of genes, enabling them to carry out cell type-specific tasks. Gene expression patterns can also be changed by extracellular stimuli that allow cells to adapt to new environments or react to signaling molecules. Thus, the regulation of gene expression is a highly dynamic and complex process that involves numerous levels, including chromatin accessibility, transcription and post-transcriptional/translational processes [6]. The work presented within this thesis touches multiple levels of gene expression regulation in order to shed light on interactions taking place between different types of regulators, including epigenetic regulators, transcription factors (TFs), RNA-binding proteins (RBPs) and microRNAs (miRNAs). In the following subsections, regulatory layers that are of relevance to this thesis, will be discussed in sufficient detail, so that the reader can easily place the conducted work discussed in the subsequent chapters in its wider context.

1.1.1 *Chromatin accessibility impacts gene expression*

Eukaryotic DNA is organized in a set of chromosomes. Each chromosome consists of an extremely long, unbranched DNA polymer and associated proteins, some of which help to pack the thread of DNA into a more compact structure, referred to as chromatin. At the first level of packaging, DNA is wrapped around cores of histone proteins, thereby forming bead-like units, termed nucleosomes [6] (Figure 1.1). Nucleosomes can be compacted into 30-nm fibers, that can be further folded and packed, which enables DNA to condense into very dense structures, such as metaphase chromosomes [8]. Importantly, DNA packaging is a highly dynamic process, that is able to prevent or permit access to defined regions of the DNA [9]. The degree of accessibility of specific DNA loci, such as transcription initiation regions (promoters), is known to affect the ability of regulators, such as TFs, to associate with their binding sites. Thus, the local chromatin state is able to modulate the access of TFs on their targets. And TF binding to DNA has been demonstrated to feed back on the chromatin state [9, 10]. Accordingly, promoter activity has been shown to correlate with the presence/absence of chemical modifications (marks) associated with specific chromatin states [11–14]. Such modifications are mainly found on DNA or histones and are referred to as epigenetic marks, while the term “epigenetics” broadly refers to heritable changes in gene expression that are not due to alterations of the DNA sequence [15]. However, there are ongoing discussions regarding a complete and valid definition of this term [16].

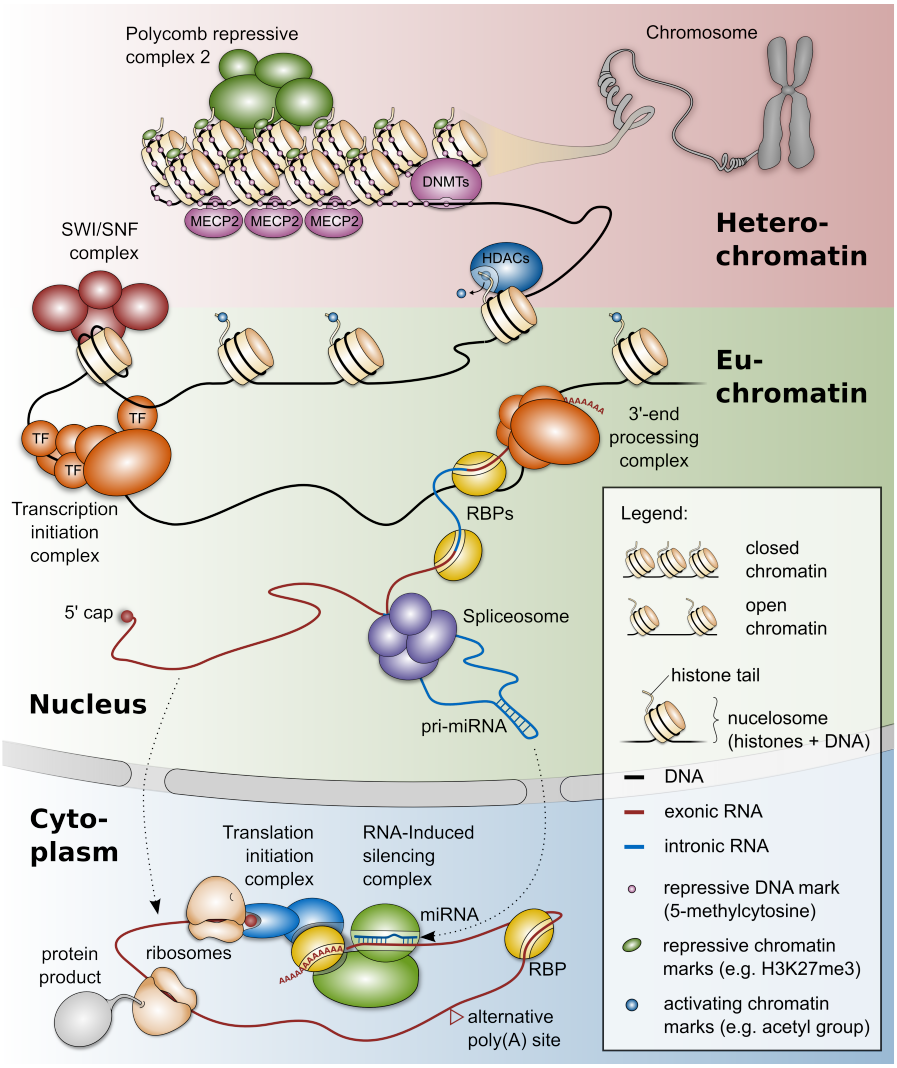


Figure 1.1: Gene expression regulation. Gene expression is modulated by epigenetic regulators (e.g. HDACs, SWI/SNF complexes, DNMTs, Polycomb repressive complexes) which alter chromatin accessibility. In contrast to "closed" chromatin (Heterochromatin), where loci are transcriptionally silent, Euchromatin is accessible to the transcriptional machinery. Transcribed pre-mRNAs, consisting of exons and introns, are bound by various RNA-binding proteins (RBPs) that act in pre-mRNA processing (e.g. alternative splicing) as well as in mRNA transport. Spliced-out introns can contain regulatory RNAs, such as pri-miRNAs, which are further processed and then exported from the nucleus to the cytoplasm where they finally act as post-transcriptional regulators (miRNAs). mRNAs harbor additional *cis*-regulatory elements at their 3' ends, which enable additional regulation of mRNA stability, localization and translation. These can be dynamically modulated through the usage of alternative poly(A) sites. Finally, translated protein products may be regulators, such as RBPs, transcription factors (TFs) or epigenetic regulators, that act at the indicated regulatory layers.

Chapter 2 extensively reviews epigenetic regulators, associated marks and their effect on DNA accessibility. At this point are discussed only a few selected examples in order to introduce the basic mechanisms and the corresponding terminology.

A relatively well described example of an epigenetic mark is DNA methylation, which consists of methylation of cytosines by DNA methyltransferases (DNMTs) [17]. In human, DNA methylation of promoter regions with a high content of CG dinucleotides (CpGs) comes along with transcriptional inactivity of the corresponding genes [18]. It is thought that methylation marks generally weaken TF–DNA interactions [19–21]. Moreover, methylated DNA has been shown to attract proteins, such as the Methyl CpG Binding Protein 2 (MECP2) [22], that in turn recruit epigenetic repressors, such as histone deacetylases (HDACs) [23], thereby making promoter regions inaccessible to the transcriptional machinery [19–21] (Figure 1.1). Trimethylation of H3 histones at lysine 27 (H3K27me3), a mark mediated by Polycomb repressive complex 2 (PRC2), was found to correlate with CpG-rich regions [24]. Interestingly, PRC2 has been shown to recruit DNMTs. Thus, there exists a regulatory crosstalk between DNA methylation and Polycomb-mediated repression [25]. However, even though exogenous CG-rich sequence elements seem to be sufficient to attract PRC2 [26], the mechanisms of PRC2 recruitment appear to be more complex [27–32] (see Chapter 2 for further details).
While repressive marks, such as DNA methylation or H3K27me3 chromatin modifications contribute to "closed", transcriptionally silent chromatin (named Heterochromatin), marks associated with transcriptionally active chromatin (termed Euchromatin) also exist. For instance, acetylation of histone tails is thought to reduce the affinity of histones for DNA thereby promoting an "open", transcriptionally active state of the chromatin [33]. Consistently, by removing the acetyl group from acetyl-lysine, HDACs act as transcriptional corepressors [34] (Figure 1.1).
Chromatin structure can be altered by nucleosome remodeling complexes, such as the SWItch/Sucrose NonFermentable (SWI/SNF) complexes. These complexes impact chromatin by sliding as well as by adding or removing histones (Figure 1.1). SWI/SNF complexes impact gene expression [35] and have been found to interact synergistically or antagonistically with other regulators, such as Polycomb repressive complexes [36] and HDACs [37].

1.1.2 *Gene expression regulation by transcription factors*

Albeit the interactions between epigenetic regulators and TFs are by far not understood in all details, it is well established that Euchromatin is accessible to components of the transcription initiation complex, including RNA polymerase II (RNAP II) and TFs present in the nucleus (Figure 1.1). TFs can either repress or activate transcription [38]. Many TFs require to be post-transcriptionally "switched" into an active state before they can shuttle to the nucleus and modulate the expression of their target genes [39–41]. This "switch" in TF activity allows rapid signal transduction and thus fast response to intra- or extracellular stimuli. TF activation/inactivation takes place in response to ligand binding [42, 43], phosphorylation [44] or interaction with other proteins [45, 46]. A relatively well studied example, in which the latter two mechanisms co-occur, illustrating the regulatory complexity of transcrip-

tion factor activation, is the interferon (IFN)–mediated immune response in human. At first, in IFN-induced JAK/STAT signaling [47, 48], type I IFNs associate with the IFN-α Receptor (IFNAR) located at the plasma membrane, thereby activating the receptor–associated tyrosine kinases Janus Kinase 1 (JAK1) and Tyrosine Kinase 2 (TYK2). The active kinases in turn activate the two TFs Signal Transducer and Activator of Transcription 1 (STAT1) and 2 (STAT2), respectively [49]. Activated STAT1 molecules can form homodimers that translocate to the nucleus and bind to gamma-activated sequences (GAS) in order to induce transcription of IFN-stimulated genes (ISGs). Alternatively, activated STAT1 can associate with STAT2 and IFN Regulatory Factor 9 (IRF9), thereby forming the IFN-Stimulated Gene Factor 3 (ISGF3), which can shuttle to the nucleus and activate hundreds of target genes that have IFN-stimulated response elements (ISRE) in their promoter region. Thus, IFN-induced JAK/STAT signaling triggers the formation of at least two different transcription factor complexes, both of which contain activated STAT1 protein, but feature distinct binding specificities (GAS or ISRE, respectively).

The example of IFN-induced JAK/STAT signaling illustrates the point that for many transcriptional regulators, it is not suitable to use the abundance of the regulator estimated with RNA sequencing or proteomics as a measure of the regulatory impact of the regulator on its target genes, because in many cases, the cellular concentration of the regulator will be very different from the concentration of its active form, present in the nucleus.

During my PhD I got the opportunity to work with Erik van Nimwegen, a leading scientist in the field of modeling gene expression regulation at genome-wide scale. And in collaboration with the research group lead by Markus Heim we studied the impact and dynamics of IFN-α treatment making use of Motif Activity Response Analysis (MARA) [50], a method described in Chapters 3 and 4 in great detail. Briefly, in contrast to standard transcriptome analyses, which for instance attempt to identify differentially expressed genes (including transcriptional regulators), MARA aims to infer the impact (also referred to as "activity") of regulator's binding motifs (such as ISRE or GAS) by modeling gene expression measurements as a linear function of the unknown activity of each motif and the number of predicted motif binding sites occurring in gene promoter regions.

Chapter 5 presents the outcome of the study conducted in collaboration with the Heim group. Analyzing IFN-α–induced signaling in paired liver biopsies obtained from 18 patients with chronic hepatitis C virus infections, we were able to shed light on the relative contribution of transcription factor binding motifs to the global changes in gene expression observed during the first week after IFN-α injection.

1.1.3 *Co- and post-transcriptional regulation of gene expression*

Post-transcriptional gene regulation (PTGR) comprises many processes that shape the transcriptome in a highly dynamic manner. After transcription initi-

ation, RNAP II synthesizes precursor messenger RNAs (pre-mRNAs). These nascent RNAs undergo various processing steps before they reach their mature form (mRNA), that is finally exported to the cytoplasm [51] (Figure 1.1). Processing of pre-mRNAs into mature mRNAs has three main steps: (i) capping, (ii) splicing and (iii) polyadenylation [52]. The latter two processes are highly dynamic and involve regulatory layers that significantly contribute to transcriptome diversity (see 1.1.3.2 and 1.1.3.3, below) [53].
Beyond pre-mRNA processing, PTGR further includes processes that take place during mRNA transport, translation and metabolism [54]. Importantly, regulatory factors involved in PTGR largely consist of RBPs that associate with RNA thereby forming ribonucleoprotein (RNP) complexes [51]. Below, post-transcriptional processes and corresponding regulators that are relevant to this thesis will be discussed in more detail. However, the interested reader is referred to more extensive literature on these topics (see e.g. [6]).

1.1.3.1 *Capping*

"Capping" refers to the attachment of a methylated guanosine (7-methylguanosine) to the 5' end of a newly transcribed RNA. The resulting 5' cap protects mature mRNAs against degradation by exonucleases and is required for their efficient translation into proteins. Moreover it has been demonstrated that the cap structure promotes the export of mRNAs from the nucleus to the cytoplasm [55].

1.1.3.2 *Splicing*

The vast majority of eukaryotic pre-mRNAs have their regions which encode for parts of mature mRNAs (called exons) interrupted by intervening regions (called introns) [56]. Thus, in order to obtain a continuous and meaningful protein-coding mRNA sequence that can be translated into protein, intronic regions need to be removed from nascent transcripts by a process called "splicing". Importantly, splicing is a highly dynamic process that adds an additional regulatory layer, allowing the production of different isoforms from individual pre-mRNA species [57]. This is done through the joining of exon sequences to each other in a dynamic process called "alternative splicing". Various types of alternative splicing events have been described, including (i) skipping or inclusion of exons (so called cassette exons), (ii) selection of one of multiple possible exons (mutually exclusive splicing), (iii) elongation/shortening of exons by selection of different 5' or 3' splice sites and (iv) intron retention [58, 59]. Ultimately, alternative splicing contributes to proteome diversity [60] and evolution [61–63].
Splicing is catalyzed by the so called spliceosome (Figure 1.1), a multisubunit complex that is composed of numerous proteins and small nuclear ribonucleoprotein particles (snRNPs) [64], which belong to the core components of the spliceosome [65]. snRNPs are composed of specific proteins and small nuclear RNAs (snRNAs), the latter guiding the interaction with short pre-mRNA located motifs, such as the branch point or the 5' splice site con-

sensus sequences [66]. The sequence elements recognized by snRNPs are found within nearly all potential splice sites, including sites that form rarely, if ever, spliced pseudo-exons [67, 68]. Consequently, additional *cis*-acting sequence features are required for high fidelity splicing events. Based on their location and regulatory function, these additional elements are classified into intronic and exonic splicing enhancers (ISEs and ESEs, respectively) and silencers (ISSs and ESSs, respectively)[69]. Enhancer elements act predominantly as binding sites for members of the serine-arginine-rich (SR) protein family, that help to recruit the snRNP subunits to the splice site and to assemble the spliceosome [65]. In contrast, silencer elements are often bound by heterogeneous nuclear RNPs (hnRNPs), which thereby inhibit spliceosome build-up and consequently splice site usage [64]. A relatively well studied example of antagonistic splicing regulators are hnRNP A1 and the SR protein SF2/ASF [70–72].

1.1.3.3 *Cleavage and polyadenylation*

Endonucleolytic cleavage and polyadenylation of nascent transcripts defines the 3' end boundary of mRNAs [73]. It is mediated by the 3'-end processing complex (Figure 1.1), which consists of more than 80 proteins [74], including the subcomplexes: Cleavage stimulation Factor (CstF), Cleavage Factors I (CFIm) and II (CFIIm), the Cleavage and Polyadenylation Stimulation Factor (CPSF) and further factors such as the Nuclear Poly(A) Binding Protein 1 (PABPN1), RNAP II and nuclear Poly(A) Polymerase (PAP) [75, 76].
CstF is a trimeric complex that is composed of CstF-50, CstF-77 and CstF-64 or its paralog CstF-64τ. The latter, CstF-64 and CstF-64τ, enable CstF to recognize U/GU-rich downstream sequence elements (DSEs) [77–79]. The CFIm subcomplex is a tetramer that consists of two CFIm25 and two further subunits made up of CFIm68 and/or CFIm59. CFIm binds to "UGUA" motifs via CFIm25 [80, 81] and CFIm25-knockdown results in a transcriptome-wide increase in proximal poly(A) site usage, demonstrating its decisive role in poly(A) site usage [81, 82]. The CFIIm subcomplex includes PCF11 and CLP1 [83]. PCF11 has been demonstrated to be involved in transcription termination, whereas CLP1 interacts with the CFIm and CPSF subcomplexes [84]. The CPSF subcomplex, consists of WDR33, FIP1, CPSF30, CPSF100, CPSF160 and CPSF73 [85–87], the latter being the endonuclease that cleaves the nascent RNA within a relatively small sequence window [88, 89]. CPSF recognizes the so called polyadenylation signal, a *cis*-regulatory element located approximately 10-30 nucleotides upstream of the cleavage site [90]. Recent studies suggest that the canonical polyadenylation signal ("AAUAAA") is recognized by the CPSF30 and WDR33 subunits [91, 92]. Moreover, FIP1 contributes to the interaction of CPSF with RNA, by binding to U-rich sequence elements upstream of the polyadenylation signal [81, 87, 91, 93] and together with CPSF160 it recruits PAP to the cleavage site [87, 94]. Finally, PAP adds a poly(A) tail to the nascent transcript, the precise length of which is determined by PABPN1 [95, 96]. In recent years, high-throughput measurements have pointed out that the majority (approximately 70-80%) of pre-

mRNAs have multiple possible cleavage and polyadenylation (poly(A)) sites [97, 98]. Thus, alternative cleavage and polyadenylation (APA) turns out to be a widespread phenomenon that is influenced by the abundance of a number of RBPs [75], thereby adding an additional layer of complexity to transcriptome diversity [99]. In detail, APA events can be classified into four different types: (i) APA at tandem 3' untranslated region (UTR) sites (ii) APA at alternative terminal exons (iii) APA at intronic sites and (iv) APA at exonic sites located within protein coding regions. The latter three types are able to change the coding DNA sequence (CDS), thereby giving rise to distinct protein isoforms. In contrast, the first mentioned APA type, although the most common, can only lead to transcript isoforms with different 3' UTR length [99]. Importantly, 3' UTRs usually harbor a multitude of *cis*-regulatory elements that serve as binding sites for *trans*-acting factors, such as RBPs and miRNAs (see 1.1.3.4, below).

Interestingly, it has been shown that fast proliferating cells undergo a systematic shift towards shorter 3' UTRs [100, 101]. Consistently, a similar shift has been demonstrated to take place upon activation of T lymphocytes [102]. However, the cause and functional consequences of such global changes in 3' UTR length are largely unknown. Early studies suggested that global shortening of 3' ends enable the cell to bypass the regulation mediated by miRNA binding sites located within the lost 3' UTR regions, thereby causing increased mRNA stability and protein levels [101, 102]. However, a more recent genome-wide analysis reported surprisingly small changes in the mRNA stability and translation rates of isoforms having different 3' UTR length [103]. Consistent with these findings, in Chapter 6 we present a study in which we find very limited changes taking place upon systematic 3' end shortening, at both, the mRNA and the protein level. In detail, we have used large-scale 3' end sequencing in combination with high-throughput proteomics measurements to characterize the significance of shortened 3' UTRs on mRNA stability and protein output in activated compared to naive T cells of human and mouse. These findings raise the question of the functional significance of APA. Besides mRNA stability and translation, APA has previously been demonstrated to influence mRNA [104] as well as protein localization [105]. Thus, regulatory effects might take place at other levels. For instance, in a recent study it has been shown that via 3' UTR-dependent protein localization (UDPL), APA is able to determine whether transmembrane proteins are transported from the endoplasmic reticulum to the plasma membrane. This is due to the addition/removal of alternative 3' UTR regions that contain binding sites for the ELAVL1 (also called HuR) RBP. Thus, rather unexpectedly, alternative cleavage and polyadenylation has been demonstrated to regulate the localization of proteins without changing their amino acid sequence [105].

In Chapter 7 we have put efforts in analyzing a huge array of previously published 3' end sequencing libraries in order to (i) generate a comprehensive and reliable catalog of poly(A) sites and (ii) to identify novel *cis*-regulatory elements that impact cleavage and polyadenylation. These efforts resulted in an extended list of known poly(A) signals by 6 variants that are conserved in hu-

man and mouse. Moreover, we found that the binding motif of the HNRNPC RBP, poly(U), has a specific positional profile around cleavage sites. Following this up experimentally, we showed that knockdown of HNRNPC entails global changes in poly(A) site usage, including 3' UTR extension/shortening. The alternatively regulated regions are rich in binding sites of the ELAVL1 RBP and include the region that has recently been demonstrated to be decisive for UDPL of the Cluster of Differentiation 47 (CD47) protein [105]. We conclude that HNRNPC is a potent regulator of 3' end processing, presumably modulating the A/U-rich element interactome.

1.1.3.4 *Gene expression regulation by microRNAs*

miRNAs are small (∼22 nucleotides long), regulatory RNAs that inhibit the expression of their target genes by reducing mRNA stability [106] and/or mRNA translation [107–109]. The majority of miRNAs are produced by RNAP II, whereupon the primary miRNA precursors (pri-miRNAs) are transcribed as independent transcripts or as part of other genes (intronic miRNAs, Figure 1.1) [110]. After transcription, the pri-miRNAs are processed into mature miRNAs, which are finally loaded into the RNA-induced silencing complex (RISC) [111]. Loaded RISC is able to bind target mRNAs, which are partially complementary to the sequence of the miRNA [112]. In Chapter 2 we review the mechanism of action and functions of miRNAs in detail. We highlight that other types of regulators, in particular TFs and epigenetic regulators, are strongly enriched among the predicted targets of miRNAs. Indeed, the targeting of other regulators that have genome-wide impact is one possible explanation for the documented importance of miRNAs to biological processes, such as development, as has been impressively demonstrated by experiments with embryonic stem cell lines that are entirely deficient in mature miRNAs due to the lack of a vital miRNA biogenesis factor, such as Dicer [113] or Dgcr8 [114]. Both, Dicer$^{-/-}$ and Dgcr8$^{-/-}$ embryonic stem cells (ESCs) exhibit highly similar proliferation defects, suggesting that the defect is caused by the deficiency in mature miRNAs rather than the lack of one of the two factors. The vital role of ESC-specific miRNAs to cell identity suggests that they target key regulators, which in turn, themselves are crucial to cell identity and fate.
In Chapter 3 we present a study, in which we have made use of MARA (see above) to gain insights into the transcription regulatory networks that lie immediately downstream of embryonic miRNAs. Towards this goal, we have extended the MARA model to also account for miRNA effects on mRNA stability. Applying this approach to transcriptome profiling data of cells that do or do not express embryonic miRNAs, we have identified transcriptional regulators that are direct targets of the miRNAs and whose activities were significantly altered, as inferred from genome-wide expression changes. In particular, we have shown that embryonic miRNAs target chromatin and cell cycle regulators at multiple levels. Moreover, they impact Irf2-dependent transcription and canonical NF-κB signaling.

2

MODULATION OF EPIGENETIC REGULATORS AND CELL FATE DECISIONS BY MIRNAS

2.1 ABSTRACT

The work discussed in this chapter was published in Epigenomics in 2013 (see reference [115]).

Mammalian gene expression is controlled at multiple levels by a variety of regulators, including chromatin modifiers, transcription factors and miRNAs. The latter are small, ncRNAs that inhibit the expression of target mRNAs by reducing both their stability and translation rate. In this review, we summarize the recent work towards characterizing miRNA targets that are themselves involved in the regulation of gene expression at the epigenetic level. Epigenetic regulators are strongly enriched among the predicted targets of miRNAs, which may contribute to the documented importance of miRNAs for pluripotency, organism development and somatic cell reprogramming.

2.2 INTRODUCTION

miRNAs are small RNAs, approximately 22 nucleotides long that regulate the expression of target mRNAs. Through high-throughput sequencing and computational analyses, thousands of miRNA-encoding loci have been identified in the human genome [116]. The identification of their targets, although proceeding at a fast pace, to some extent lags behind. The TarBase database catalogs miRNA targets for which supporting experimental evidence has been obtained [117]. Much about the principles of miRNA-target interactions has been inferred through computational analysis, and especially through comparative genomics [118, 119]. Although the first characterized function of a miRNA was the inhibition of target mRNA translation [120, 121], more recent evidence from high-throughput studies points to the impact of miRNAs on mRNA decay [106]. Translation inhibition appears to be a rather early outcome of miRNA–target interactions [107, 108] and/or an outcome that depends on the location of miRNA-binding sites within transcripts [119]. It has also been clear since their discovery, that miRNAs are necessary for organism development [122]. It is generally believed that miRNAs confer robustness to biological processes through a variety of mechanisms, including reinforcement of transcriptional programs that are essential for the establishment of cell lineages [123]. Not surprisingly, deregulated expression of miRNAs has been observed in a variety of pathological conditions, including cancers [124]. Epigenetic silencing is one of the mechanisms behind deregulated expression of miRNAs, which has been reviewed in a number of publications [125–128]. Surprising in light of their demonstrated involvement in cell fate determination is the relative paucity of reports on the epigenetic regulators whose

expression is in turn regulated by miRNAs. This is the topic of the present article.

2.3 MECHANISM OF ACTION & FUNCTIONS OF MIRNAS

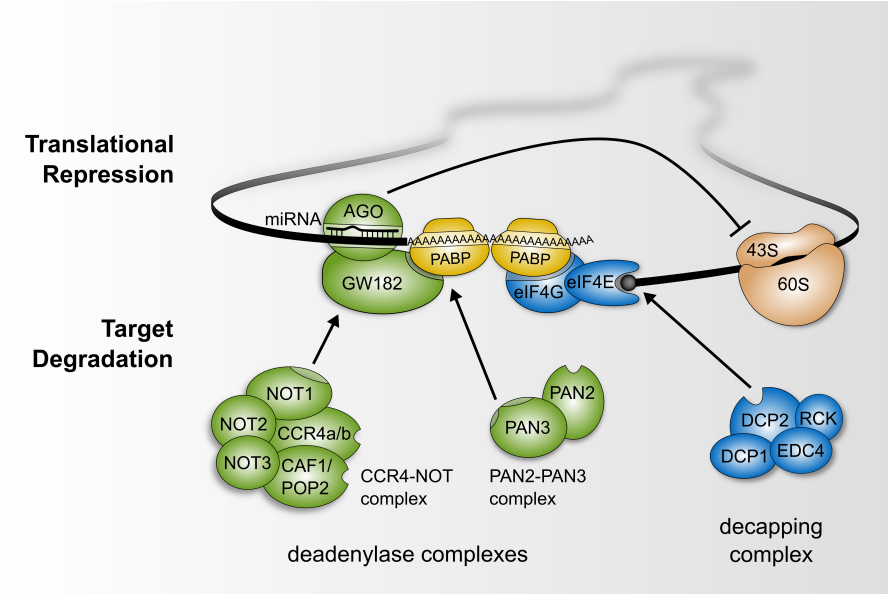


Figure 2.1: Mechanism of action of miRNAs. Molecular mechanisms underlying miRNA-dependent inhibition of gene expression. miRNAs are incorporated in Argonaute proteins, which they guide to mRNAs. Through the interactions of Argonaute proteins with various protein complexes, miRNAs induce translational repression, decapping and deadenylation of the target mRNAs.

The vast majority of human miRNAs are transcribed by RNA polymerase II, either as part of the introns of other genes or as independent transcripts [110]. Two endonucleolytic complexes, the first composed of Drosha and DGCR8 proteins acting in the nucleus [129] and the second composed of Dicer and the HIV TARBP2, or TRBP acting in the cytoplasm [130], come into play to produce the miRNA mature forms from the primary transcripts. Numerous examples of post-transcriptional regulation of miRNA processing leading to tissue-specific expression of mature miRNAs have been reported (see e.g., [131] for a review). At the molecular level, evidence has been provided for a role of miRNAs in translation, as well as in mRNA degradation (Figure 2.1). Very recently, an in vitro study employing rabbit reticulocyte lysate demonstrated that miRNAs specifically repress translation by inhibiting 43S ribosomal scanning [132]. This study further concluded that inhibition of miRNA-dependent translation depends on both the PABP and eIF4G. However, another study in Drosophila melanogaster did not find PABP to be necessary [133], and in an in vitro translation system involving extracts of ascites from Krebs2 mice, it was the cap recognition process that was inhibited by miRNAs [134]. The literature on miRNA-dependent deadenylation and degradation is much more extensive (see [135] for a recent re-

view). A current model is that miRNA-induced silencing complexes induce deadenylation of their targets through the PAN2-PAN3 and CCR4-NOT complexes, and the decapping of the mRNAs through the DCP1-DCP2 complex [109]. A variety of approaches have been employed towards the discovery of miRNA targets and functions, from individual gene reporter assays to high-throughput profiling of mRNA expression following changes in the expression of the miRNA [106, 136]. The latter approach takes advantage of the effect of miRNAs on mRNA decay rates. Methods that evaluate the change in protein levels [137, 138] or in mRNA translation [139] upon changes in miRNA concentration have also been proposed. More recently, it has become possible to directly identify transcriptome-wide miRNA-interacting sites, through crosslinking and immunoprecipitation of the Argonaute proteins [140–142]. The consequences of miRNA–target interactions reflect the diversity of miRNA targets. Comparative genomic analysis suggested that miRNAs recognize their targets predominantly through their 5'end, which is typically referred to as the "seed" sequence [118]. Over 60% of human genes are predicted to be miRNA targets [143]. Surprisingly, however, the impact of miRNAs on the mRNA and protein level of their conserved targets is typically small [137, 138]. This observation led to the view that miRNAs have qualitatively distinct effects on their targets, sometimes acting as on-off switches, and sometimes simply "fine-tuning" gene expression [112]. Although the typically small effects would suggest that miRNAs rarely act as switches, the different sensitivity of phenotypes to the dosage of different genes makes it difficult to estimate the prevalence of switch versus fine-tuning targets. Another hypothesis regarding the mechanisms through which miRNAs affect gene expression stems from the interplay they have with transcription factors. That transcription factors are preferred targets of miRNAs has been noted from the initial stages of miRNA target prediction [144]. Furthermore, it has been observed that gene expression regulatory networks are enriched in small network motifs known as feed-forward loops, in which a transcription factor and a miRNA act on a common target [145–147]. A recent computational study showed that when the transcription factor and the miRNA are part of an "incoherent" feed-forward loop (i.e., when the transcription factor upregulates the expression of the miRNA and the common target while the miRNA represses target gene expression), small changes in target gene expression coupled with a reduced response to fluctuations in the levels of upstream regulators can be achieved [148]. Although seemingly consistent with the observation that miRNAs typically induce small changes in target gene expression, this model would in fact predict a large change in the level of the miRNA target in response to a large perturbation in miRNA concentration, as achieved in transfection experiments. Thus, the mechanism underlying the limited response of miRNA targets to miRNA perturbations remains unclear. Moreover, it remains difficult to envision how these small, 'fine-tuning' effects on individual miRNA targets lead to distinguishable phenotypes on which evolutionary selection can act so that the individual miRNA target sites remain conserved

over large evolutionary distances. Interestingly, transcription factors, as well as epigenetic regulators, are highly probable miRNA targets (Figure 2.2).

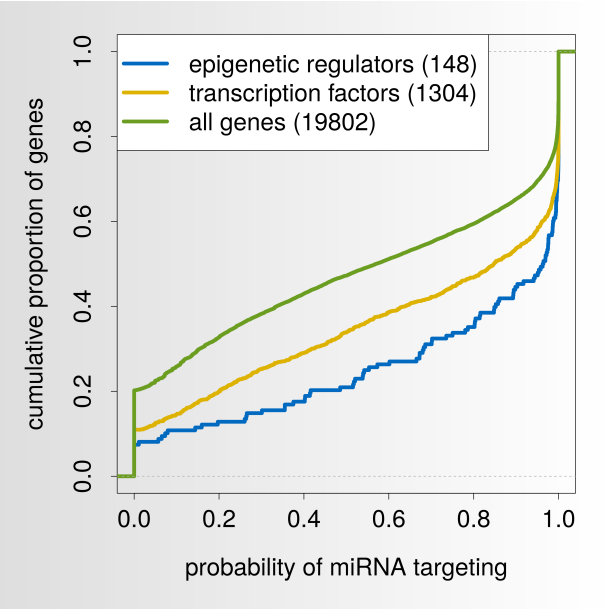


Figure 2.2: Preferred targets of miRNAs. Transcription factors and epigenetic regulators are preferred targets of miRNAs. We calculated the probability of each gene in a specific functional class being targeted by miRNAs and then plotted the cumulative distribution of the probability values for genes in individual functional categories. Briefly, targeting scores for each human gene and every miRNA seed family were obtained by first averaging the TargetScan aggregate probability of conserved targeting (P_{CT}) scores [143] of all transcripts that are associated with the gene. The probability of a gene to be targeted by at least one miRNA (P_{gene}) was calculated as: $P_{gene} = 1 - \prod (1 - P_{miR,gene})$, where $P_{miR,gene}$ is the miR probability that the gene is targeted by a specific miRNA miR, calculated as described above. Transcription factors were obtained from the DBD database [149] and epigenetic regulators by extracting genes that were annotated with epigenetic-related gene ontology terms from the AmiGO database [150]. The number of genes in each category is indicated in parentheses. Compared with all genes, epigenetic regulators and transcription factors tend to have higher probabilities of being miRNA targets.

2.4 EPIGENETIC REGULATION OF GENE EXPRESSION

Epigenetic modifications are thought to be central to cell fate and organism development (see [151, 152] for recent reviews). Thus, substantial efforts have been dedicated to the mapping of epigenetic marks in a variety of cell types. The method of choice is chromatin immunoprecipitation (i.e., employing antibodies directed towards specific types of histone and DNA modifications) [153]. Additionally, methods such as FAIRE-Seq [154], Sono-Seq [155] and DNaseI-Seq [156, 157] have been developed for mapping "open" chromatin regions that are free of nucleosomes and are accessible to micrococcal nuclease digestion [158, 159]. Such regions are typically associated with promoters. Furthermore, metabolic labeling of histones is used to in-

vestigate the kinetics of nucleosome turnover [160]. The plethora of chromatin marks and their significance for gene expression have been reviewed extensively elsewhere [161]. Here, we will only introduce the aspects that are relevant for understanding how miRNAs may exert their roles by acting on specific epigenetic regulators.

2.4.1 *DNA methylation*

The majority of cytosines that occur within CG dinucleotides (CpGs) in the human genome are methylated. The ease with which 5-methylcytosines are deaminated and mutated through error-prone repair is thought to be the reason for the general depletion of CpGs in the genome (see [162] for a recent review). However, CpGs are not uniformly distributed across the genome, but are specifically enriched in so-called CpG islands (CGIs), regions that are, on average, 1 kbp in length and are typically associated with promoters. Within CGIs, CpG dinucleotides tend to be maintained in a demethylated state [163]. Appropriate DNA methylation is important for many processes, including embryonic development [164]. It plays a role in transcriptional regulation [18], genomic imprinting [165] and X-chromosome inactivation [166]. Interestingly, DNA methylation has also been associated with mRNA splicing [167], as have other chromatin marks [168, 169]. While the methylation state of low CpG-content promoters is not indicative of the activity of the corresponding genes, methylation of high CpG-content promoters is generally associated with transcriptional inactivity [18]. The repressive effect of methylation at CpG-dense regions is thought to be mediated by methyl-CpG-binding domain (MBD) proteins (e.g., MECP2 [22]) that further recruit repressive chromatin modifiers, such as histone deacetylases (HDACs) [23], as well as by a general reduction of transcription factor–DNA interactions at methylated DNA sites [19–21]. In mammals, the transfer of a methyl group to the C5 position of cytosines is catalyzed by DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b (Figure 2.3A) [17]. Once established, DNA methylation patterns are maintained through cell divisions. The mechanism involves the interaction of the NP95 protein with hemimethylated DNA via the SET and RING–associated domain of NP95 [170], the formation of complexes that also include the proliferating cell nuclear antigen (PCNA) [171, 172], and recruitment of Dnmt1 by the NP95 protein at replication forks. Consistent with its function, expression of Dnmt1 is weak in resting cells, but high in dividing cells [173], particularly in the S-phase of the cell cycle [174]. By contrast, Dnmt3a and Dnmt3b are de novo methyltransferases that are essential to embryonic development [175]. Accordingly, Dnmt3b and Dnmt3a knockout mice exhibit developmental defects or die shortly after birth, respectively [176]. Recent studies suggest that ten–eleven translocation dioxygenases (TET1, TET2 and TET3) initiate the removal of methylation marks by oxidizing 5-methylcytosine to 5-hydroxymethylcytosine. 5-hydroxymethylcytosine is subsequently converted into further derivatives and finally replaced with

unmodified cytosine via base excision repair (see [177] for a recent review).

miRNA	Target	Targeting	Reference
hsa-miR-302	MECP2	3'UTR	[178]
hsa-miR-148a	DNMT3B	CDS	[179]
mmu-miR-290–295 cluster	Dnmt3a Dnmt3b	Indirect via Rbl2 Indirect via Rbl2	[180, 181]
hsa-miR-29a/b/c	DNMT3A DNMT3B	3'UTR 3'UTR	[182, 183]
hsa-miR-29a	TET1 TET2 TET3	3'UTR 3'UTR 3'UTR	[184]
hsa-miR-29b	DNMT1	Indirect via SP1	[183]
hsa-miR-152 hsa-miR-148a	DNMT1	3'UTR 3'UTR	[185]
hsa-miR-302	DNMT1	Indirect via AOF2	[178]
hsa-miR-26a	EZH2	3'UTR	[186]
hsa-miR-101	EZH2	3'UTR	[187, 188]
mmu-miR-137	Ezh2	3'UTR	[189]
mmu-miR-214	Ezh2	3'UTR	[190]
hsa-miR-128	BMI-1	3'UTR	[191]
mmu-miR-128 mmu-miR-183 mmu-miR-200c mmu-miR-203	Bmi-1	3'UTR 3'UTR 3'UTR 3'UTR	[192]
mmu-miR-294	Baf170	3'UTR	[193]
hsa-miR-199a	BRM	3'UTR	[194]
mmu-miR-9* mmu-miR-124	Baf53a	3'UTR 3'UTR	[195]
hsa-miR-449a	HDAC1	3'UTR	[196]
mmu-miR-1	Hdac4	3'UTR	[197]
mmu-miR-29b	Hdac4	3'UTR	[198, 199]
mmu-miR-206	Hdac4	3'UTR	[198, 200]
hsa-miR-433	HDAC6	3'UTR	[201]

Table 2.1: miRNAs that target epigenetic regulators. CDS: Coding DNA sequence; hsa: Homo sapiens; mmu: Mus musculus; UTR: Untranslated region.

DNA methylation is regulated directly and indirectly by multiple miRNAs (Table 2.1 & Figure 2.3A). DNMT1 was found to be targeted by miR-148a and miR-152, miRNAs that also reduce cell proliferation by indirectly up-regulating the cell-cycle inhibitors CDKN2A and RASSF1A. Thus, miR-

148a and miR-152 may couple cell-cycle progression and maintenance of DNA methylation [185]. Interestingly, miR–148a was found to also regulate the *de novo* DNA methyltransferase DNMT3B [179], which suggests that miR-148a, more generally, antagonizes DNA methylation. Similarly, the miR-29 family miRNAs were found to directly target DNMT3A and DNMT3B and, indirectly, through the SP1 transcription factor, to downregulate DNMT1. Consequently, miR-29 overexpression leads to global DNA demethylation [182, 183]. A surprising recent finding is that the miR-29 family miRNAs also target the TET enzymes, which would initiate the removal of methylation marks [184]. The functional consequences of these antagonistic miRNA effects on DNA methylation remains to be determined. During somatic cell reprogramming, miR-302 contributes to global DNA demethylation by downregulating several epigenetic regulators, including MECP2, as well as DNMT1 via AOF2 [178].

Finally, the mouse-specific miR-290–295 cluster miRNAs target the Rbl2 repressor to indirectly regulate the expression of the *de novo* DNA methyltrasferases [180, 181].

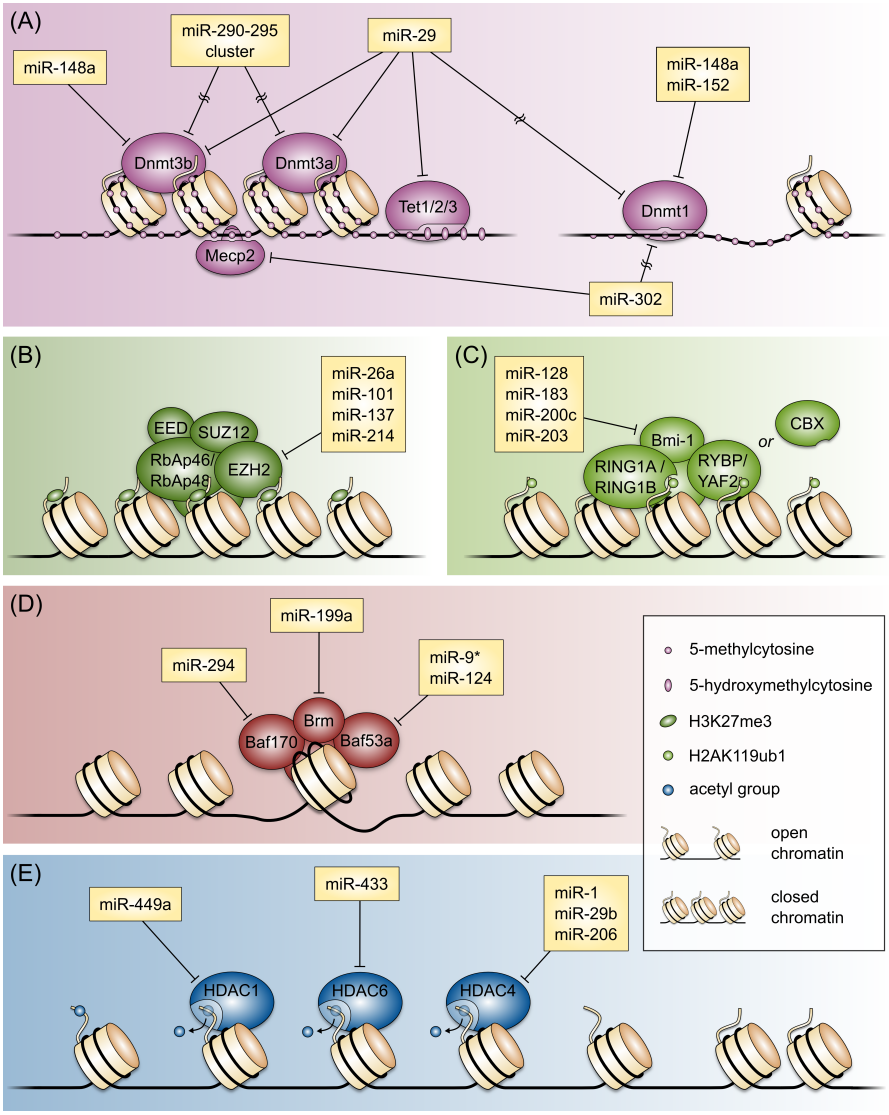


Figure 2.3: Epigenetic regulators that are targeted by miRNAs. **(A)** miRNAs that directly or indirectly regulate targets in the DNA methylation/demethylation pathways. **(B-E)** miRNAs that target a catalytic subunit of **(B)** PCR2, **(C)** PCR1, **(D)** SWI/SNF complex subunits and **(E)** histone deacetylases.

2.4.2 *Polycomb group proteins*

Polycomb repressive complexes (PRCs) form another class of epigenetic regulators that contribute to stable gene repression. Two major PRC complexes, PRC1 and PRC2, exist in mammals (see [202] for a recent review). PRC2 consists of four core subunits: RbAp48, SUZ12, EZH2 and EED (Figure 2.3B). EZH2 can be replaced by EZH1, RbAp48 by RbAp46 and EED has multiple isoforms. EZH2 and EZH1 are the catalytic subunits with which PRC2 methylates lysine 27 (K27) of histone H3. The resulting H3K27me3 histone mark is associated with Polycomb group (PcG)-mediated gene silencing and was found to correlate with CpG-rich regions, such as CGIs [24]. Interestingly, exogenous GC-rich sequence elements were found to be

sufficient to recruit PRC2 [26]. Accordingly, high-throughput experiments revealed that many CGIs are bound by PRC2 and marked with the associated H3K27me3 mark [24, 203–206]. Moreover, EZH2 was shown to recruit DNA methyltransferases, providing a direct link between PcG-mediated repression and DNA methylation [25]. The mechanisms by which PRCs are guided to their target regions appear to be very complex. Apart from CGIs, other DNA elements (PRE-kr and D11.12 [27, 28]) have been reported to recruit PRC2. Interestingly, a recent study reported on the REST transcription factor-dependent recruitment of PRC2 to gene loci during cell differentiation [207]. Furthermore, lncRNAs, such as HOTAIR and Xist were found to play a role in PcG-complex recruitment [29–32]. In contrast to PRC2, PRC1 has no methyl-transferase activity. Rather, PRC1 silencing is associated with the ubiquitylation of histone H2A at lysine 119 (H2AK119ub1) [208], which is brought about by a two-subunit PRC1 core complex formed by RING1A or RING1B and a Polycomb group ring finger (PCGF) [209] protein such as Bmi-1. The two-subunit PRC1 core complex further associates with either CBX or RYBP (or the RYBP homolog YAF2, respectively) (Figure 2.3C). In contrast to RYBP and YAF2, CBX binds H3K27me3 [202]. Accordingly, CBX-containing PRC1 complexes colocalize with H3K27me3, whereas the recruitment of RYBP/YAF2-containing PRC1 complexes was found to be H3K27me3-independent [209, 210]. A lncRNA, ANRIL, was found to contribute to transcriptional repression by PRC1 at the INK4b/ARF/INK4a locus [211].

Some PRC components appear to be heavily regulated by miRNAs (Table 2.1 & Figure 2.3B & C). For example, Bmi-1 is regulated by miR-128, miR-183, miR-200c and miR-203 [192], and miR-128 expression reduces H3K27me3 marks and self-renewal of glioma cells [191]. Moreover, multiple miRNAs including miR-26a, miR-101, miR-137 and miR-214 regulate the EZH2 catalytic component of PRC2. Interestingly, the upregulation of EZH2 expression by the pluripotency factor c-Myc appears to occur through reduction of miR-26a expres- sion [186]. EZH2 is upregulated in many cancers, where it silences several tumor suppressor genes. Thus, EZH2 targeting miRNAs, such as miR-26a and miR-101, may act as tumor suppressors by preventing the inactivation of other tumor suppressors [187, 188]. Modulation of Ezh2 expression by additional miRNAs occurs during stem cell differentiation. In adult neural stem cells, reduced miR-137 expression leads to increased Ezh2 expression and promotes differentiation [189]. On the other hand, miR-214 is expressed during differentiation of skeletal muscle cells, negatively regulating Ezh2 and thereby supporting its own expression and cell differentiation [190]. Consistently, miR-214 was found to reduce the efficiency of somatic cell reprogramming [212].

2.4.3 *SWI/SNF complexes*

The SWI/SNF proteins (also called BAF) form multisubunit complexes, with ATP-dependent nucleosome remodeling activity, in which the catalytic activ-

ity resides in either Brg1 (also called Smarca4) or Brm (also called Smarca2; see [35] for a recent review). SWI/SNF proteins have been shown to interact synergistically or antagonistically with other chromatin regulators, including HDACs [37] and Polycomb group proteins [36]. Several subunits of SWI/SNF complexes are expressed in a lineage-specific manner, probably contributing to lineage-specific gene expression. Consistently, SWI/SNF complex proteins were found to play important roles in development. For example, Baf60c is specifically expressed in the heart and somites of the mouse embryo and is required for the recruitment of SWI/SNF complexes to heart-specific enhancers, and its knockdown causes defective heart development [213]. On the other hand, neural stem and progenitor cells express a specific SWI/SNF complex (termed neural progenitor-specific BAF complex, brief npBAF), which contains Baf45a and Baf53a, both units being required for self-renewal. During the differentiation of progenitor cells into postmitotic neurons, Baf45a and Baf53a are replaced by Baf45b, Baf45c and Baf53b, forming a neuron-specific BAF complex (termed nBAF). This subunit switch seems to be important for neural development, since its inhibition results in reduced neuronal differentiation [214]. Finally, an embryonic stem cell (ESC)-specific SWI/SNF complex (termed esBAF) was described, containing Brg, Baf155 and Baf60a, but not Brm, Baf170 and Baf60c. esBAF is essential for pluripotency and self-renewal [215] and ChIP-Seq analysis revealed that esBAF colocalizes with the pluripotency factors Nanog, Oct4 and Sox2, as well as with Stat3 and Smad1 [216]. Interestingly, esBAF exerts both opposing, as well as synergistic, effects with respect to the Polycomb repressive complex in the maintenance of pluripotency [36]. Overexpression of Baf155 along with Oct4, Sox2 and Klf4 has been found to enhance the efficiency of somatic cell reprogramming and inclusion of Brg1 provides an even stronger improvement [217]. Although Baf155 has approximately a 60% sequence identity with Baf170, they seem to have different functions. Baf155 knockdown in ESCs results in reduced expression of the pluripotency marker Oct4, decreased proliferation and increased apoptosis, demonstrating its functional importance. This proliferative defect could only be rescued by Baf155, but not by Baf170 expression [215]. The BRM catalytic subunit of the SWI/SNF complex appears to be a target of miR-199a (Table 2.1 & Figure 2.3D), a miRNA that is expressed in ESCs, as well as in various organs such as the ovary, uterus, testis, prostate, kidney and heart [218]. The miRNA modulates the activity of SWI/SNF, which in turn modulates the expression of the miRNA through a double-negative feedback loop. Namely, miR-199a reduces the expression of BRM, a negative regulator of EGR1, leading to increased EGR1 expression and EGR1-dependent transcription activation from the miR-199a-2 locus, from which miR-199a and miR-214 are expressed [194]. This type of network architecture is known to generate bistable gene expression patterns. Consistently, epithelial tumor cell lines appear to be either EGR1-miR-199a-high and BRM-low, or BRM-high and EGR1-miR-199a-low [194].

2.4.4 *Histone acetylation/deacetylation*

Histone acetylation refers to the transfer of an acetyl group to the ϵ-amino group of a lysine residue catalyzed by histone acetyltransferases [34]. This reduces the affinity of histones for DNA, thereby promoting an "open", transcription-permissive chromatin structure. Histone acetyltransferases therefore act as transcriptional coactivators and histone acetylation positively correlates with transcriptional activity [33]. By contrast, HDACs, which catalyze the reverse reaction removing the acetyl group from acetyllysine, act as transcriptional corepressors (Figure 2.3E) [34]. Surprisingly, it has been found that in ESCs, activating H3K9 acetylation marks co-occur with repressive H3K27 trimethylation marks at many promoters [219]. It is now thought that this pattern is characteristic to lineage-specific genes that are maintained in a "poised" state, ready to be expressed.
Several HDACs are targeted by miRNAs (Table 2.1 & Figure 2.3E). HDAC1 was shown to be a target of miR-449a. The expression of this miRNA is low in prostate cancer compared with normal tissue, and overexpression of miR-449a in prostate cancer cells results in cell cycle arrest and apoptosis [196]. miR-1, miR-29 and miR-206 promote myogenesis by targeting Hdac4, a repressor of skeletal muscle genes [197, 198, 200]. Through Hdac4, as well as other inhibitors of osteoblast differentiation, miR-29b also promotes osteogenesis [199]. Finally, HDAC6 appears to be regulated by miR-433 as a mutation in the putative miR-433 target site of HDAC6 causes a specific form of chondrodysplasia [201].

2.5 MIRNA-DEPENDENT MODULATION OF PLURIPOTENCY, DIFFERENTIATION & SOMATIC CELL REPROGRAMMING

Consistent with the first reported function of a miRNA, lin-4, in the regulation of *Caenorhabditis elegans* development [120], animal models and ESC lines deficient in expression of miRNA biogenesis factors revealed that miRNAs are essential for both the maintenance of pluripotency and for embryonic development. Dicer-deficient mice die early in development, and lack pluripotent stem cells [220]. Dicer$^{-/-}$ ESCs, although viable, lack mature miRNAs and show several differentiation defects *in vitro* as well as *in vivo*, [221]. The similar proliferation defects observed in Dicer$^{-/-}$ [113] and Dgcr8-deficient [114] ESCs suggest that the underlying cause is the lack of mature miRNAs rather than the loss of either of these proteins *per se*. The miRNA population of mouse ESCs consists largely of members of the miR-290–295 cluster [222, 223]. Three out of the seven miRNAs that are expressed from this cluster, namely miR-291a-3p, miR-294-3p and miR-295-3p, share the seed sequence AAGUGCU, as do many other miRNAs with an embryonic pattern of expression (e.g., miR-302–367 cluster miRNAs). In contrast to the miR-290–295 cluster miRNAs that are only found in mouse, the miR-302–367 cluster is also present in human. The AAGUGCU-seed miRNAs from the miR-290–295 cluster appear to play an important role in the maintenance

of pluripotency. They target the cyclin E–Cdk2 pathway at multiple levels, forcing the G1–S transition and rescuing the proliferation defect observed in miRNA-deficient ESCs [224]. These miRNAs have also been found to promote induced pluripotency [225]. Although it has been reported that the miR-302–367 cluster miRNAs are able to reprogram somatic cells into induced pluripotent stem cells, without any additional exogenous transcription factors in both mouse and human [226], these findings were recently challenged [227]. The AAGUGCU family miRNAs regulate several processes that are known to be important for somatic cell reprogramming, including cell proliferation, epithelial–mesenchymal transition and epigenetic remodeling [212, 228–230]. Among their targets, Lats2 and Cdkn1a [224] are involved in establishing the cell cycling pattern specific to pluripotent stem cells, which have a truncated G1 phase and a lengthened S phase [231]. By targeting the TGF-b receptor 2 these miRNAs also downregulate TGF-b signaling, ultimately promoting the mesenchymal–epithelial transition [228, 229] that is crucial to the reprogramming of embryonic fibroblasts [230]. Consistent with their role in the maintenance of pluripotency, the AAGUGCU family miRNAs regulate multiple epigenetic regulators. The positive correlation of the expression of miR-290–295 cluster miRNAs and the expression of *de novo* DNA methyltransferases in ESCs [180, 181] suggested that these enzymes are indirect miRNA targets. Indeed, the AAGUGCU family miRNAs repress the expression of Rbl2 [180], which is part of the DREAM repressor complex [232]. Rbl2 was shown to directly repress the expression of Dnmt1 in both mouse and human [233], and the previously mentioned studies indicate that the *de novo* DNA methyltransferases are under a similar control. The picture that emerged from these studies is that the embryonic stem cell-specific miR-290–295 cluster miRNAs maintain a low Rbl2 expression level, which in turn allows the expression of *de novo* methyltransferases (Dnmt3a and Dnmt3b) and an appropriate deposition of methylation marks. A more recent study further determined that during somatic cell reprogramming, the AAGUGCU family miRNAs downregulate AOF1, AOF2, MECP1-p66 and MECP2, as well as Dnmt1 via AOF2, thereby contributing to global DNA demethylation [178]. The AAGUGCU family miRNAs also appear to target multiple components of the SWI/SNF chromatin-remodeling complexes. For example, Subramanyam et al. found that these miRNAs appear to target Baf170 (also called Smarcc2) [228], which is specific to the differentiated cell BAF complex (dBAF) [234]. Consistently, Baf170 is downregulated during somatic cell reprogramming, while the expression of Baf155, the component specific to esBAF, is increased [217]. In recent work [193], we confirmed that the AAGUGCU family miRNA, miR-294, directly targets Baf170. Interestingly, Baf155 is a high-confidence predicted target of the let-7 miRNA [143], which has been shown to antagonize the effects of the miR-290–295 cluster miRNAs in early differentiation [235]. The interplay between let-7 and the miR-290–295 cluster miRNAs appears to be extremely complex and important for the maintenance/loss of pluripotency. Let-7 family miRNAs are broadly expressed in somatic tissues, but not in ESCs [236], where their

processing is inhibited by the pluripotency–associated RNA-binding protein Lin28 [237–239]. In turn, let-7 targets several genes crucial to pluripotent stem cells, including n-Myc, c-Myc and Sall4 [235, 240]. Let-7 miRNAs also target Lin28, thereby maintaining their own expression in differentiating cells [241]. Finally, miRNAs miR-9* and miR-124 synergistically target the SWI/SNF complex component, Baf53a. Neurons in which the target sites of these miRNAs in the 3' untranslated region of Baf53a were mutated, exhibited defective dendritic outgrowth. By contrast, overexpressing the miRNAs in neural progenitor cells decreased proliferation [195]. The importance of miR-9* and miR-124 for neuronal fate determination was further underscored by a study in which human fibroblasts were converted into neurons through miR-9/9* and miR-124 expression [242]. Consistent with their antiproliferative and prodifferentiation effects, both miRNAs were found to act as tumor suppressor miRNAs in various types of cancer, including glioblastoma multiforme [243–245]. A summary of the epigenetic regulators that are targeted by miRNAs is shown in Table 2.1.

2.6 CONCLUSION

miRNAs form an extensive layer of post-transcriptional regulators of gene expression. Among their preferred targets are transcription factors and epigenetic regulators, which in turn regulate the expression of individual miRNAs. Much of the work on miRNA-containing gene expression regulatory networks has focused on transcription factors, which have been found to be enriched among miRNA targets from the initial stages of miRNA target prediction. The equally strong enrichment of epigenetic regulators appears to have been underappreciated thus far, though it may be related to the observation that many of the known phenotypes brought about by the loss of miRNAs are of a developmental nature. Consistent with the computational predictions, many epigenetic regulators have been identified as miRNA targets in developmental contexts.

2.7 FUTURE PERSPECTIVE

Given the intricacy of the regulation of processes touched upon in this review, predicting the dynamics of gene expression in these conditions is challenging, and quantitative models will be necessary to evaluate the relative contribution of individual players. Nonetheless, it is tempting to speculate that the impact of miRNAs on processes such as development stems, in part, from their targeting of other types of regulators that amplify the miRNA effects at the level of the targets. With the availability of methods for experimental identification of miRNA-binding sites *in vivo*, it has become possible to reconstruct the miRNA–target interaction networks in individual cell types and during specific processes such as development. These data should contribute to our improved understanding of the cell fate specification mechanisms. This, in

turn, will enable more precise modulation of cell fates, which has numerous applications in regenerative medicine.

2.8 EXECUTIVE SUMMARY

MECHANISM OF ACTION & FUNCTIONS OF MIRNAS

- miRNAs guide Argonaute protein-containing silencing complexes to target mRNAs.
- Target mRNAs are predominantly recognized through the miRNAs' "seed" sequence (nucleotides ∼1–8 from the miRNA 5' end).
- The majority of human genes are predicted targets of miRNAs.
- The outcome of miRNA–target interactions is an increased rate of target mRNA degradation and/or translation inhibition.
- Quantitatively, the effects of miRNAs range from "fine-tuning" to off-switching of gene expression.
- Epigenetic regulators and transcription factors are among the preferred targets of miRNAs.

EPIGENETIC REGULATION OF GENE EXPRESSION

- Many methods have been developed to map genome-wide epigenetic modifications. Histone and DNA modifications can be detected with ChIP-Seq, and open chromatin regions with FAIRE-Seq, Sono-Seq or DNaseI-Seq. Nucleosome turnover kinetics can be studied by metabolic histone-labeling experiments.
- Specific changes in the epigenetic marks were associated with organism development.

DNA METHYLATION

- DNA methylation is involved in many processes such as transcriptional regulation, genomic imprinting and embryonic development.
- DNA methylation is established by the *de novo* DNA methyltransferases (Dnmt3a, Dnmt3b) and maintained by Dnmt1 throughout cell proliferation. Active DNA demethylation is probably carried out by the ten–eleven translocation enzymes (TET1, TET2 and TET3).
- miR-148a, miR-29 and miR-152 were found to regulate the expression of DNA methyltransferases, as well as ten–eleven translocation enzymes.

POLYCOMB GROUP PROTEINS

- The Polycomb repressive complexes, PRC1 and PRC2, contribute to stable gene repression.
- PRC2-mediated repression is associated with the methylation of lysine 27 of histone H3 (H3K27me3). The EZH2 catalytic component of this complex is regulated by numerous miRNAs, including miR-26a, miR-101, miR-137 and miR-214.
- PRC1-mediated repression is associated with ubiquitylation of histone H2A at lysine 119 (H2AK119ub1). The Bmi-1 component of this complex is targeted by many miRNAs, including miR-128, miR-183, miR-200c and miR-203.

SWI/SNF COMPLEXES

- SWI/SNF (BAF) complexes are ATP-dependent nucleosome-remodeling complexes, with lineage-specific subunit compositions.
- Several BAF complex subunits are regulated by miRNAs, including Baf170 by miR-294, BRM by miR-199a, and Baf53a by miR-9* and miR-124.

HISTONE ACETYLATION/DEACETYLATION

- Acetylation reduces the affinity of histones for DNA, thereby promoting an "open", transcription-permissive chromatin structure, while histone deacetylases (HDACs) remove acetyl groups from histones, thereby acting as transcriptional corepressors.
- Many HDACs were found to be targeted by miRNAs, including HDAC1 by miR-449a, HDAC6 by miR-433 and Hdac4 by miR-1, miR-29b and miR-206.

MIRNA-DEPENDENT MODULATION OF PLURIPOTENCY, DIFFERENTIATION & SOMATIC CELL REPROGRAMMING

- miRNAs are essential for pluripotency, as well as for embryonic development.
- AAGUGCU seed family miRNAs contribute to pluripotency in various ways. For example, the miRNAs force G1–S cell-cycle transition and allow the expression of *de novo* DNA methyltransferases, thereby promoting appropriate DNA methylation in embryonic stem cells.

2.9 AUTHORS INFORMATION

2.9.1 *List of authors*

The following authors have contributed to the work discussed in Chapter 2:

1. Andreas Johannes Gruber[1] (Abbr.: AJG) &

2. Mihaela Zavolan[1] (Abbr.: MZ)

whereat author affiliations are as follows:

1 Biozentrum, University of Basel, Klingelberstrasse 50-70, CH-4056 Basel, Switzerland

2.9.2 *Author contributions*

The first author (AJG) contributed most (author abbreviations are defined in the previous subsection (i.e. 2.9.1)). In detail, AJG created all figures and Table 2.1. Moreover he performed the analysis discussed in Figure 2.2. MZ wrote the majority of sections 2.3, 2.6 and 2.7. AJG wrote the majority of sections 2.4, 2.5 and 2.8. Both authors wrote the abstract and the introduction and approved the final manuscript.

2.10 FUNDING

AJG was supported by a Werner Siemens Fellowship.

3

EMBRYONIC STEM CELL-SPECIFIC MICRORNAS CONTRIBUTE TO PLURIPOTENCY BY INHIBITING REGULATORS OF MULTIPLE DIFFERENTIATION PATHWAYS

3.1 ABSTRACT

The work discussed in this chapter was conducted in collaboration with the van Nimwegen and the Ciaudo labs and published in Nucleic Acids Research in 2014 (see reference [193]).

The findings that microRNAs (miRNAs) are essential for early development in many species and that embryonic miRNAs can reprogram somatic cells into induced pluripotent stem cells (iPSCs) suggest that these miRNAs act directly on transcriptional and chromatin regulators of pluripotency. To elucidate the transcription regulatory networks immediately downstream of embryonic miRNAs, we extended the motif activity response analysis (MARA) approach that infers the regulatory impact of both transcription factors and miRNAs from genome-wide expression states. Applying this approach to multiple experimental data sets generated from mouse embryonic stem cells (ESCs) that did or did not express miRNAs of the ESC-specific miR-290-295 cluster, we identified multiple transcription factors (TFs) that are direct miRNA targets, some of which are known to be active during cell differentiation. Our results provide new insights into the transcription regulatory network downstream of ESC-specific miRNAs, indicating that these miRNAs act on cell cycle and chromatin regulators at several levels and downregulate TFs that are involved in the innate immune response.

3.2 INTRODUCTION

Embryonic stem cells (ESCs) originate from the inner cell mass of mammalian blastocysts. Due to their ability to self-renew as well as differentiate into various specialized cell types, they hold the promise of medical applications such as stem cell therapy and tissue engineering. Therefore, the regulatory mechanisms behind pluripotency, stem cell fate and renewal are of great interest.

MiRNAs are short ($\sim$22 nt long), single-stranded RNAs that post-transcriptionally regulate the expression of target genes [246]. Computational and high-throughput studies suggest that a single miRNA can regulate hundreds of target genes [106, 137] and that the majority of human mRNAs are regulated by miRNAs [143]. Several studies found that the expression of ESC-specific miRNAs is required for initiation of stem cell differentiation and normal embryonic development [113, 114, 221]. The ESC-specific miR-290-295 cluster accounts for approximately 50 percent of the miRNA population of *mouse* embryonic stem cells [247–250] and its expression is downregulated relatively rapidly during differentiation [223, 248]. Interest-

ingly, three of the seven miRNAs that are co-expressed from the miR-290-295 cluster, namely miR-291a-3p, miR-294 and miR-295, are sufficient to force a G1→S transition [224] and promote induced pluripotency [225]. All of these miRNAs, as well as those of another ESC-specific miRNA cluster, miR-302-367 [223, 251], have the same sequence "AAGUGCU" at positions 2-8 (also called the "seed") which define a family of miRNAs with related targets [143]. In contrast to the miR-290-295 cluster, miR-302-367 is also present in human and has been used to reprogram fibroblasts into induced pluripotent stem cells (iPSCs) [226]. The reprogramming of differentiated cells into pluripotent stem cells entails large gene expression and phenotypic changes that are likely to be due to regulatory cascades that involve several regulators. To identify *transcriptional regulators* that are immediate targets of the AAGUGCU seed family miRNAs, we analyzed data obtained in several previous studies that aimed to uncover the function of the miR-290-295 cluster. These data consist of microarray-based measurements of mRNA expression in ESCs that were either deficient in miRNAs or expressed subsets of ESC-specific miRNAs (Supplementary Table A.1). Sinkkonen et al. [180] analyzed mRNA expression of ESCs that express miRNAs ($Dicer^{+/-}$), ESCs that do not express miRNAs ($Dicer^{-/-}$) as well as $Dicer^{-/-}$ ESCs transfected with the miR-290-295 cluster miRNAs (miR-290, miR-291a-3p, miR-292-3p, miR-293, miR-294 and miR-295 mimics). The study showed that the expression profile of ESCs can be restored to a large extent in $Dicer^{-/-}$ ESCs through transfection of miR-290-295 cluster miRNAs, and that these miRNAs are important for appropriate *de novo* DNA methylation in differentiating ESCs. Hanina et al. [252] profiled mRNA expression in $Dicer^{-/-}$ ESCs as well as in $Dicer^{-/-}$ ESCs transfected with miR-294. Combining these expression data with a biochemical approach to isolate Argonaute 2 (Ago2)-bound mRNAs, the study identified miR-294 targets in ESCs. It further concluded that miR-294 regulates a subset of genes that are also targeted by the Myc transcriptional regulator and that some of the effects of miR-294 expression may be due to the indirect upregulation of pluripotency factors such as Lin28. Employing mRNA expression profiling of $Dgcr8^{-/-}$ ESCs, as well as miR-294-transfected $Dgcr8^{-/-}$ ESCs, Melton et al. [235] showed that self-renewal and differentiation of ESCs is regulated in an antagonistic manner by miR-294 and let-7. Finally, Zheng et al. [250] profiled mRNA expression of miRNA expressing ESCs and $Dicer^{-/-}$ ESCs and uncovered a pro-survival, anti-apoptotic function of the miR-290-295 cluster of miRNAs. Altogether, these studies provide five separate experimental data sets that can be used to investigate the function of AAGUGCU seed family miRNAs in ESCs. They all determined mRNA expression profiles of ESCs with impaired miRNA expression (due to knockout of either Dgcr8 or Dicer components of the miRNA biogenesis pathway), as well as of ESCs that expressed miRNAs of the AAGUGCU seed family. The latter were either ES cells which expressed the full complement of miRNAs, or miRNA-deficient ESCs that were transfected with either miRNAs of the miR-290-295 cluster, or only miR-294. Although it has been observed that these studies resulted in sets of miRNA tar-

gets that are only partially overlapping [249], a meta-analysis that combines these data sets to identify the pathways that are most reproducibly targeted by the AAGUGCU miRNAs has not been performed. In our study, we aimed to infer transcriptional regulators that are directly and consistently targeted by the AAGUGCU family of miRNAs, the pathways that these regulators control and the interactions that they have with each other. Towards this end, we modeled genome-wide mRNA expression in terms of computationally predicted target sites of both transcription factors and miRNAs. This approach allowed us to identify a number of transcriptional regulators whose activity is consistently altered by miRNAs of the AAGUGCU seed family and that could contribute to the maintenance of pluripotency. Through reporter assays we validated these regulators as targets of AAGUGCU seed family miRNAs. Employing Dicer$^{-/-}$ mouse ES cells we showed that the expression of the IRF2 transcription factor is strongly upregulated in the absence of miRNAs and that the nuclear concentration of the RelA component of the NF-κB pathway upon stimulation with TNF-α is also increased. Our results give new insights into the functions of miRNAs in the regulatory circuitry of ESCs.

3.3 RESULTS

3.3.1 *General relationship between data sets*

A common, though perhaps naive expectation is that combining data from experiments that have been independently performed in different labs, with different experimental procedures, allows one to identify essential properties of the system that are invariant with respect to details of the experimental approach. In our case, in any given experiment, confounding effects may have led to some genes being spuriously identified as targets of AAGUGCU miRNAs (false positives), and true targets of AAGUGCU miRNAs being missed (false negatives). For example, because it is unclear whether the miRNA processing enzymes solely function in this pathway, it is important to analyze data from ESCs in which the miRNA biogenesis has been impaired at different levels (Dicer in the studies of Sinkkonen et al. [180] and Hanina et al. [252] and Dgcr8 in the study of Melton et al. [235]). Furthermore, although ESCs expressing the full complement of miRNAs provide the most physiological reference point for the function of the miR-290-295 cluster miRNAs in normal, unstressed cells, the effect of these miRNAs in these cells is confounded by the effects of other co-expressed miRNAs. Similarly, if the profiled cell population was heterogeneous with respect to the pluripotency/differentiation status, the let-7 miRNAs may have masked the effect of miR-294, because these miRNAs have antagonistic effects [235]. Requiring targets to show consistent downregulation across multiple data sets can reduce the number of false positive miR-294 targets. On the other hand, requiring perfect consistency across a large number of experiments is likely to lead to too many false negatives, simply because different experiments have different levels of accuracy or confounding effects. Thus, we first investigated

the relationship of gene-level expression changes between ESCs that did or did not express embryonic miRNAs in all pairs of experiments. Although pairwise Pearson correlation coefficients were as low as 0.11 (Supplementary Fig. A.1), three of the five experimental data sets (Fig. 3.1A), covering all described conditions (expression of miR-294, miR-290-295 cluster miRNAs, or the entire complement of embryonically expressed miRNAs in a miRNA-deficient background) gave reasonably high pairwise correlation coefficients (Fig. 3.1B). We therefore focused our discussion on these datasets, and for completeness, we present the results of a similar analysis of all five data sets in the Supplementary material. Of the approximately $4000 - 5000$ genes that were downregulated in a single experiment, a little less than 2000 genes were downregulated in all three experiments. Importantly, the proportion of predicted AAGUGCU seed family targets among downregulated genes increased when intersecting an increasing number of data sets (Fig. 3.1C), indicating that the approach of a combined analysis of these data sets does have the potential to reveal important regulators that are immediately downstream of the AAGUGCU family of miRNAs. 252 of the genes downregulated in all three experiments were predicted AAGUGCU seed family targets [143] (Supplementary Table A.2).

(A)

Publication	Abbr.	GEO Id	First Condition	Second Condition
Melton et al. (2010)	A	GSE18840	Dgcr8$^{-/-}$	miR-294 transfected Dgcr8$^{-/-}$
Sinkkonen et al. (2008)	B	GSE8503	Dicer$^{-/-}$	miR-290 cluster transfected Dicer$^{-/-}$
Sinkkonen et al. (2008)	C	GSE7141	Dicer$^{-/-}$	Dicer$^{+/-}$

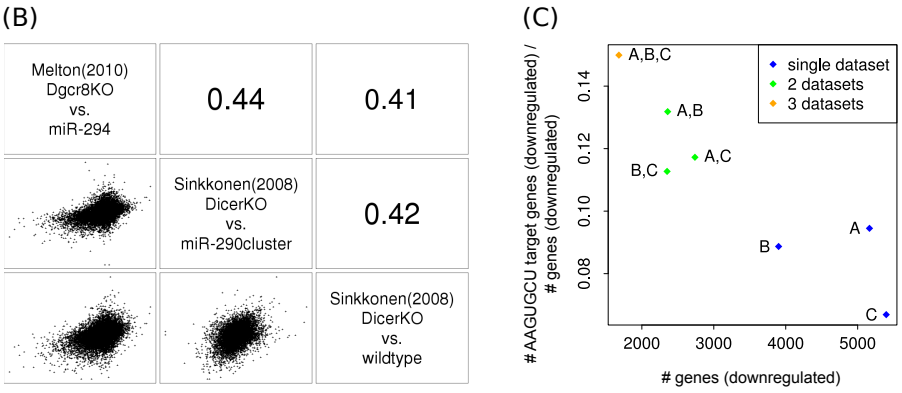


Figure 3.1: Overview of the mRNA expression data sets. **(A)** Data sources. **(B)** Matrix of scatter plots (below diagonal) and Pearson correlation coefficients (above diagonal) of per-gene $\log_2$ fold changes in pairs of experiments. The names of the individual data sets are shown on the diagonal. **(C)** Proportion of predicted targets of the AAGUGCU seed family of miRNAs (TargetScan aggregate P_{CT} score based predictions [143]) among genes that are consistently downregulated in all three (orange), pairs (green) or individual data sets (blue) (indicated by the labels, key given in the "Abbr." column of the table in panel (A)), plotted against the number of genes that are consistently downregulated in all of the considered data sets.

3.3.2 *The transcriptional network regulated by the miRNAs of the AAGUGCU seed family in ESCs*

As mentioned in the Introduction, the main aim of our study was to identify *transcriptional regulators* that are targeted by the AAGUGCU seed family and at the same time can account for the largest fraction of gene expression changes that are observed in cells that do or do not express the miRNAs. We therefore built on the motif activity response analysis (MARA) approach [50] that we recently made available in the form of an easy to use web application ([253], see Chapter 4). In contrast to standard transcriptome analyses that strive to find genes (including transcription regulators) whose expression changes significantly between conditions, MARA aims to infer changes of the *regulatory impact* (also referred to as "activity") of binding motifs. This is achieved by modeling gene expression as a linear function of the number of regulatory motif binding sites occurring in the promoter (for TFs) and 3' UTR (for miRNAs) of the gene and the unknown activity of each motif. The change in activity of a specific binding motif (e.g. of the Irf2 transcription factor) in a specific condition (e.g. transfection of miR-294) is inferred from the expression changes of all (predicted) *targets* of this motif (determined by transcriptome profiling), taking into account the occurrences of sites for other regulators in these targets. For example, a decrease in Irf2 activity is inferred when the predicted Irf2 targets consistently show a decrease in expression that cannot be explained by the occurrence of binding sites for other regulatory motifs in the promoters or 3' UTRs of these targets. This means that MARA can uncover gene expression changes that are due not only to changes in the mRNA expression level of a regulator, but also to changes in the *active form* (e.g. for TFs through post-translational modifications such as phosphorylation) of the regulator. MARA was initially developed for the characterization of transcription regulatory networks [50], and we have recently extended it to also model miRNA-dependent changes in mRNA stability ([253], see Chapter 4). For this study we further extended the MARA approach to identify regulators whose activity not only changes most significantly between samples but also reproducibly across multiple data sets. Our approach is described in detail in the Materials and Methods section. To verify that MARA can indeed uncover the key regulator in these experiments, namely the miRNAs of the AAGUGCU seed family, we first applied MARA taking into account all TFs and miRNA seed families (see Supplementary Table A.4). In subsequent analyses however, we performed the MARA analysis with only the AAGUGCU seed family motif added to the full complement of transcription factor motifs. This was because when all miRNAs are included in the analysis, MARA will also infer nonzero activities for other miRNAs, e.g. those with significantly overlapping sets of targets [254]. MARA quantifies the extent to which the activity of each motif varies across conditions by a z-statistic, that roughly corresponds to the ratio between the average deviation of the motif activity from zero and the standard deviation of the motif activity (see Materials and Methods). Supplementary Table A.3 shows all motifs ranked by their abso-

lute z-values. MARA also predicts which promoters or 3' UTRs are targeted by each motif, quantifying the confidence in each predicted motif-target interaction by a posterior probability (see Materials and Methods). We used these probabilities to construct a regulatory network of motif-motif interactions (Fig. 3.2) that provides a synthetic view of the regulatory impact of the AAGUGCU seed family of miRNAs on the transcriptional network of pluripotent stem cells. An arrow was drawn from motif A to motif B whenever motif A was predicted by MARA to regulate a transcription factor b whose binding specificity is represented by motif B. Only motif-TF interactions that were predicted in all data sets and that involved motifs with high significance ($z > 5$) are shown.

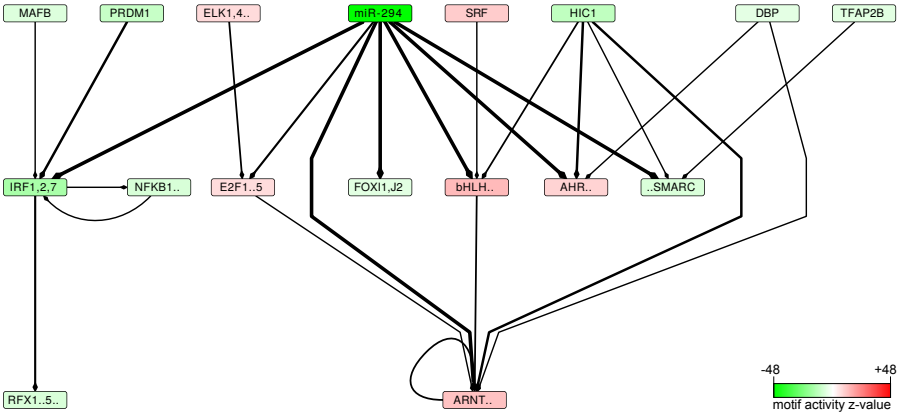


Figure 3.2: The transcriptional network inferred to be affected by the miRNAs of the AAGUGCU seed family (represented by miR-294). A directed edge was drawn from a motif A to a motif B if A was consistently (across data sets) predicted to regulate a transcription factor b whose sequence specificity is represented by motif B. The thickness of the edge is proportional to the product of the probabilities that A targets b. For the clarity of the figure, only motifs with absolute z-values > 5 and only edges with a target probability product > 0.3 are shown. The intensity of the color of a box representing a motif is proportional to the significance of the motif (the corresponding z-values can be found in Suppl. Table A.3). Red indicates an increase and green a decrease in activity, corresponding to increased and decreased expression, respectively, of the tagets of the motif when the miRNAs are expressed. The full motif names as well as the corresponding transcription factors are listed in Supplementary Table A.7.

The motif corresponding to the AAGUGCU seed family (represented by the dark green "miR-294" motif in Fig. 3.2) is by far the most significantly changing motif (see also Supplementary Table A.3). Its negative change in activity upon miRNA expression is consistent with the destabilizing effect of the miRNA on its targets. The motif with the second most significant change in activity, "IRF1,2,7", is bound by the interferon regulatory factors. MARA predicts that this motif is directly targeted by miR-294, in line with previous suggestions that the interferon regulatory factors are targets of the miR-290 cluster miRNAs [252]. We present a more detailed analysis of this motif in the next section.

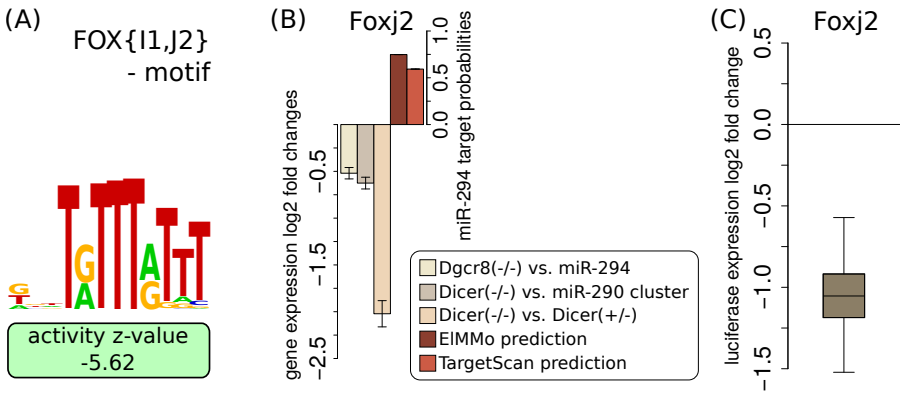


Figure 3.3: Foxj2 is a direct target of miR-294. (A) The *"FOX{I1,J2}"* motif shows a negative change in activity in the presence of miR-294. **(B)** Foxj2 mRNA $\log_2$ fold changes ($\pm$1.96*SEM; n=3) in the Melton et al. Dgcr8$^{-/-}$ versus miR-294 transfection (yellow), Sinkkonen et al. Dicer$^{-/-}$ versus miR-290-295 cluster transfection (dark brown) and Dicer$^{-/-}$ versus Dicer$^{+/-}$ (light brown) data sets, as well as the prediction scores for these genes as targets of miR-294 as given by ElMMo [255] (dark red) and TargetScan (aggregate P_{CT}) [143] (light red). **(C)** A luciferase reporter construct carrying the 3' UTR of Foxj2 is downregulated upon co-transfection with miR-294 relative to a construct carrying the Foxj2 3' UTR but with a mutated miR-294 target site (n=9).

A second motif whose activity decreases significantly upon miRNA expression is "FOX{I1,J2}" (Fig. 3.3A). Of the TFs associated with this motif, Foxj2 is predicted within all data sets to be directly regulated by miR-294 (Fig. 3.2). Consistently, Foxj2 is downregulated upon miRNA expression on the mRNA level (Fig. 3.3B). In order to validate that Foxj2 is a direct target of the miRNAs, as predicted by both Elmmo and TargetScan (Fig. 3.3B), we cloned the 3' UTR of Foxj2 downstream of a luciferase reporter and co-transfected this construct together with miR-294 in the murine mammary gland cell line NMuMG. For comparison, we generated a construct in which the presumed miRNA-294 target site was mutated and we performed similar co-transfection experiments. The results of this experiment clearly show that Foxj2 is indeed a functional target of miR-294 (Fig. 3.3C). We carried out similar transfection experiments with control siRNAs, that do not target the reporter, and a standard analysis of these data is presented in Supplementary Fig. A.3. Little is known about the function of Foxj2 in cell fate. It appears to be expressed very early in development [256], but its overexpression has a negative effect on embryogenesis [257]. Our results suggest that the AAGUGCU seed family of miRNAs contributes to the maintenance of an adequate expression of Foxj2 in pluripotent stem cells. The third most significant changing motif, basic-helix-loop-helix (referred to as "bHLH.." in Fig. 3.2), can be bound by many transcription factors (reviewed in [258]), some of which are predicted direct targets of miR-294.

To further elucidate the transcription regulatory network downstream of the AAGUGCU seed family of miRNAs, we analyzed in-depth the transcription factors whose associated motif had the most significant activity change (*z*-

value > 5) and that were consistently predicted by MARA to be direct targets of the miR-294 seed family miRNAs across the multiple data sets (Table 3.1).

Name	Motif	Motif Abbreviation	Activity z-value
Irf2	IRF1,2,7.p3	IRF1,2,7	-16.29
Mxd3	bHLH_family.p2	bHLH..	13.00
Clock	bHLH_family.p2	bHLH..	13.00
Arnt2	ARNT_ARNT2_BHLHB2_MAX_MYC_USF1.p2	ARNT..	11.60
Arnt2	AHR_ARNT_ARNT2.p2	AHR..	8.39
BAF170	DMAP1_NCOR{1,2}_SMARC.p2	..SMARC	-6.98
E2f5	E2F1..5.p2	E2F1..5	6.62
Foxj2	FOX{I1,J2}.p2	FOXI1,J2	-5.62

Table 3.1: AAGUGCU seed family transcription factor targets as predicted by combined MARA. Transcription factors consistently predicted by MARA to be a direct target of miR-294 and whose absolute motif activity z-value is > 5 (due to the presence of AAGUGCU seed family miRNAs).

We found that the majority of these direct target TFs fall into three categories that have previously been associated with pluripotency: NF-κB-related interferon response factors that control NF-κB signalling, cell cycle regulators, and epigenetic regulators.

3.3.3 *AAGUGCU seed family miRNAs modulate Irf2-dependent transcription*

The "IRF1,2,7" motif shows the second strongest activity change upon changes in miR-294 expression (Fig. 3.4A and Suppl. Table A.3). Of the individual factors associated with this motif, Irf2 is the one that was consistently predicted by our analysis to be a direct target of the AAGUGCU seed family miRNAs across data sets (Table 3.1), consistent with the predictions of both ElMMo and TargetScan (Fig. 3.4B). Irf2 was down-regulated at the mRNA level across all analyzed data sets (Fig. 3.4B). Consistently, we found that Irf2 is strongly down-regulated in $DCR^{flox/flox}$ compared to $DCR^{-/-}$ ESCs, both at the mRNA level (Fig. 3.4C) as well as at the protein level (Fig. 3.4D). To validate Irf2 as a direct target of miR-294, we conducted luciferase assays as described above for Foxj2. Our results demonstrate that Irf2 is indeed targeted by miR-294 (Fig. 3.4E). Although relatively little is known about the function of this factor in ESCs, a recent study showed that Irf2 overexpression causes differentiation of ESCs [259]. The strong impact of AAGUGCU miRNAs on Irf2 levels and the relatively large impact of the "IRF1,2,7" motif on gene expression suggest that this regulatory connection plays an important role in maintaining ESC pluripotency.

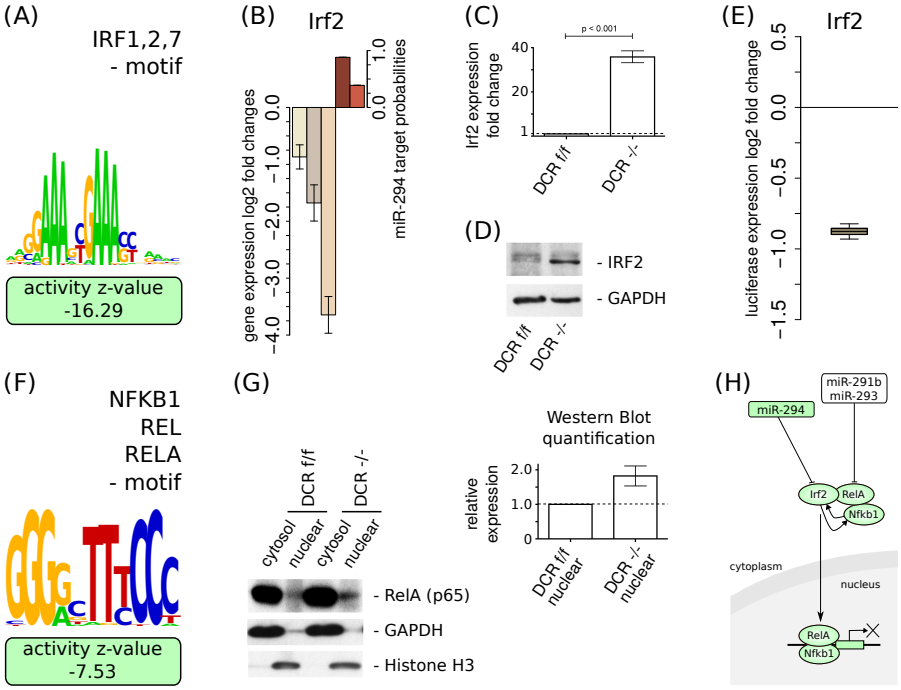


Figure 3.4: miR-294 targets the Irf2 transcription factor and modulates "IRF1,2,7" and "NFKB1_REL_RELA" activities. **(A)** The activity of the "IRF1,2,7" motif is strongly decreased in the presence of miR-294. **(B)** The expression of Irf2 is downregulated within all analysed data sets (±1.96*SEM; n=3) and Irf2 is predicted by ElMMo and TargetScan to be a direct target of miR-294 (color scheme as in Figure 3.3). Low levels of Irf2 mRNA **(C)** and **(D)** protein in $DCR^{flox/flox}$ ES cells compared to miRNA deficient $DCR^{-/-}$ ESCs are observed with qRT-PCR and Western blot, respectively. qRT-PCR experiments were run in triplicate (±SEM; n=3). **(E)** The luciferase reporter construct carrying the Irf2 3' UTR shows a strong response to miR-294 co-transfection compared to a similar construct but with a mutated Irf2 target site (n=9). **(F)** Sequence logo [260] of the "NFKB1_REL_RELA" motif that is associated with the canonical NF-κB pathway and that exhibits a significant decrease in activity in the presence of miR-294. **(G)** Western blots of RelA, GAPDH and Histone H3 in nuclear and cytoplasmic fractions in ESCs that do and do not express miRNAs. The densitometric quantification indicates an increased level of nuclear RelA in the $DCR^{-/-}$ ESCs compared to $DCR^{flox/flox}$ ESCs (±SEM; n=3). **(H)** Proposed model of the inhibitory effect of miR-290-295 cluster miRNAs on the canonical NF-κB pathway in pluripotent stem cells. Regulatory motifs are denoted by colored rectangles and individual genes by ovals. See text for the evidence of individual interactions.

Like the "IRF1,2,7" motif, the "NFKB1_REL_RELA" motif also exhibits a significantly lower activity when the embryonic miRNAs are expressed (Fig. 3.4F). Western blot confirms that after stimulation with tumor necrosis factor α (TNFα), $DCR^{flox/flox}$ ESCs have lower levels of nuclear NF-κB pathway-associated marker RelA compared with miRNA-deficient $DCR^{-/-}$ ES cells (Fig. 3.4G). This observation is consistent with a decreased activity of the canonical NF-κB signalling pathway in the presence of the miRNAs, which has been shown to be important for maintaining ESCs in a pluripotent state yet poised to undergo differentiation [261, 262]. Indeed, the Nanog

pluripotency factor directly interacts with components of the NF-κB complex, inhibiting its transcriptional activity [261]. Combining our results with recent reports that link the expression of the miR-290-295 cluster to signalling through the canonical NF-κB pathway and the latter to Irf2, the following model of the involvement of the miR-290-295 cluster in the regulation of NF-κB signalling emerges. Expression of the RelA component of the NF-κB complex is repressed post-transcriptionally by the miR-290-295 cluster members miR-291b-5p and miR-293 both of which do not belong to the AAGUGCU seed family of miRNAs [262]. In humans, RelA recruitment to the nucleus, which is a pre-requisite for NF-κB complex-dependent transcription, appears to depend on IRF2 [263], whose knockdown interferes with transcriptional activation via NF-κB [263]. Here we found that in mouse, IRF2 expression is also repressed by other members of the miR-290-295 cluster, namely the AAGUGCU family of miRNAs. Thus, the miRNAs of the miR-290-295 cluster may act in concert to inhibit the canonical NF-κB signalling in ESCs (Fig. 3.4H).

3.3.4 *miRNAs of the AAGUGCU seed family impact the cell cycle at multiple levels*

AAGUGCU seed family members of the miR-290-295 cluster were previously shown to accelerate the G1→S transition and promote proliferation of ESCs by targeting the cyclin E-Cdk2 regulatory pathway [224]. Consistently, we found that these miRNAs increase the activity of transcription regulatory motifs associated with activation of the cell cycle (Fig. 3.5), in particular the "ARNT-ARNT2-BHLHB2-MAX-MYC-USF1" motif that is bound by Myc. This TF was previously found to increase upon miR-294 transfection [235]. How the miRNAs, with intrinsically repressive function, increase the Myc activity on its targets is unknown. Our analysis suggests a few hypotheses.

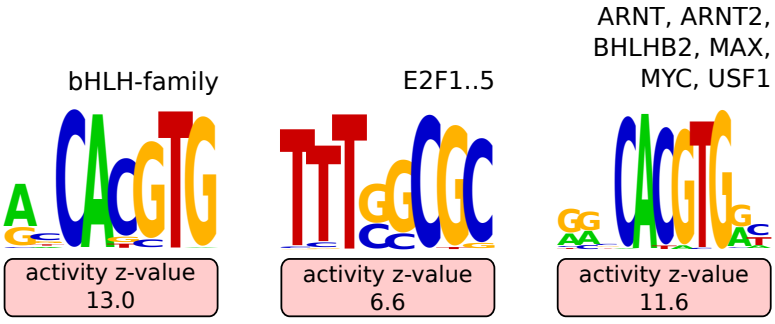


Figure 3.5: miR-294 impacts cell cycle associated motifs. MARA analysis reveals that miR-294 induces positive activity changes of multiple motifs involved in cell cycle regulation. Shown are the sequence logos [260] and the corresponding activity z-values (red shapes represent positive motif activity changes) of these motifs: the Myc- and Arnt2-associated motif "ARNT-ARNT2-BHLHB2-MAX-MYC-USF1", the putative Myc-regulating "E2F1..5" motif and the Mxd3-associated "bHLH-family" motif. Shapes scheme is as in Figure 3.4.

Specifically, luciferase assays show that three cell cycle-associated TFs, namely Mxd3 (also known as Mad3), E2f5 and Arnt2 are not only predicted but also experimentally confirmed direct targets of the AAGUGCU seed family miRNAs (Fig. 3.6A,B and Table 3.1). Mxd3 is one of the so-called 'Mad' partners of the Max protein (reviewed in [264]). In contrast to Myc, which forms a heterodimeric complex with Max in proliferating cells [265], the Mad factors Mad1, Mad3 (i.e. Mxd3) and Mad4 are primarily expressed and form complexes with Max in differentiating, growth-arrested cells [266]. Mxd3 was further shown to specifically regulate the S-phase [267]. Second, we found that E2f5, one of the TFs associated with the "E2F1..5" motif, was consistently downregulated at the mRNA level in all analyzed data sets (Fig. 3.6A) and luciferase assays further confirm that E2f5 is a target of miR-294 (Fig. 3.6B), albeit with a small response to the miRNA. Consistently, E2f5 expression is increased in $DCR^{-/-}$ ES cells compared to $DCR^{flox/flox}$ ESCs (Fig. 3.6C). The positive activity change of the E2F1..5 motif in the presence of the miRNAs (Fig. 3.5) suggests that this TF acts predominantly as repressor (as proposed before, reviewed in [268]). Notably, Myc is among the predicted targets of E2F1..5, providing an indirect path to the upregulation of Myc upon the presence of the miRNAs (Fig. 3.6A,C).

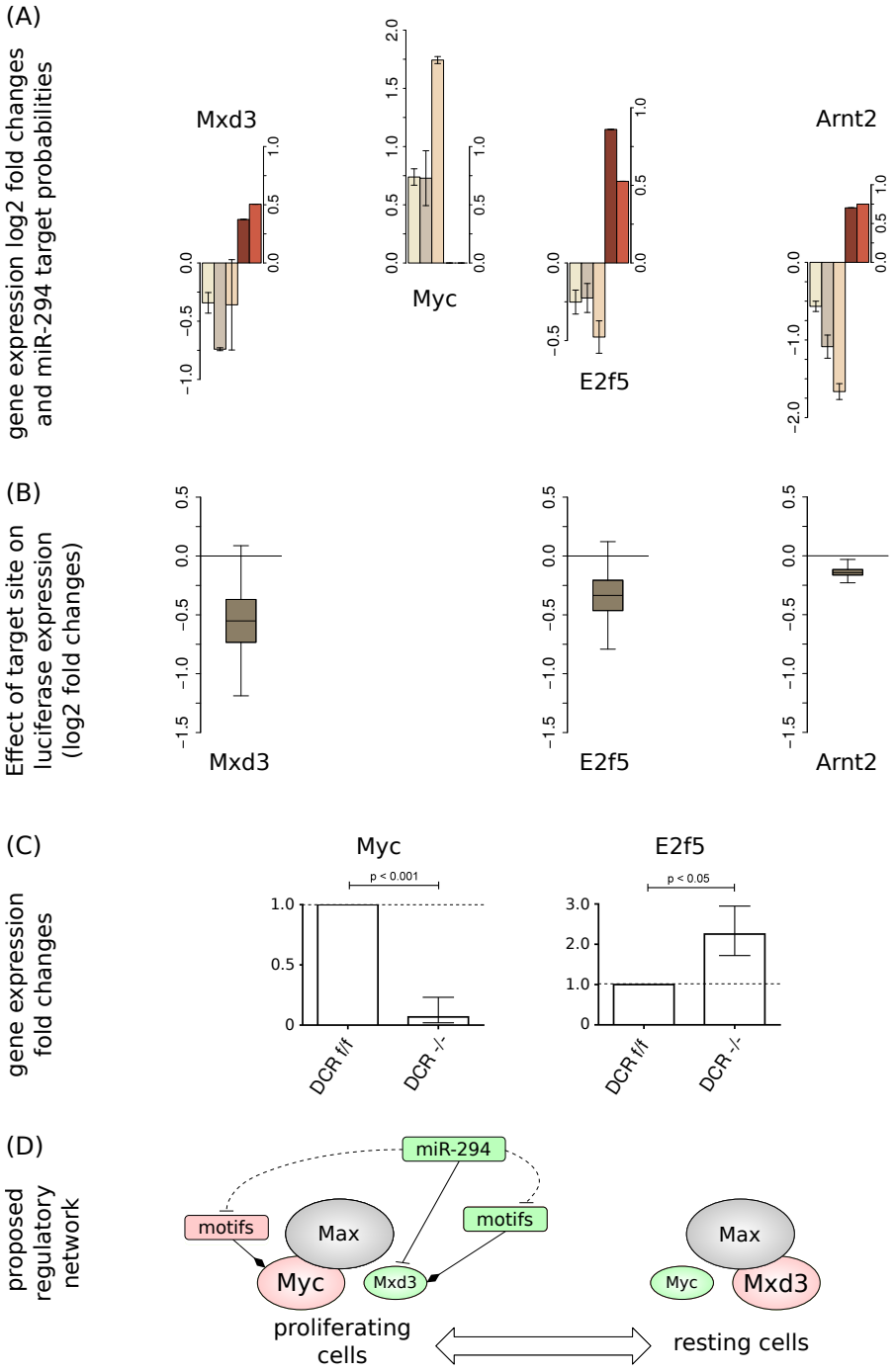


Figure 3.6: miR-294 impacts cell cycle regulation at multiple levels. **(A)** $\log_2$ mRNA fold changes ($\pm$1.96*SEM; n=3) of Myc, Arnt2, E2F5 and Mxd3 (color scheme as in Figure 3.3) in the analyzed data sets. **(B)** Luciferase constructs carrying the 3' UTR of Arnt2, E2f5 or Mxd3 respectively, are downregulated upon co-transfection with miR-294 relative to constructs carrying the same 3' UTRs but with mutated miR-294 binding sites (n=9). **(C)** qRT-PCR show decreased expression of Myc and increased expression of E2f5 in $DCR^{-/-}$ ES cells relative to $DCR^{flox/flox}$ ESCs. qRT-PCR experiments were run in triplicate ($\pm$SEM; n=3). **(D)** Proposed model of miR-294-dependent regulation of the Myc-Max/Mxd-Max network. Shapes scheme is as in Figure 3.4. Green or red shapes represent negative or positive changes (in motif activities or gene expression fold changes) respectively. Dashed lines indicate indirect and solid lines direct regulatory links between motifs/-genes.

Finally, Arnt2, a TF associated with the "ARNT-ARNT2-BHLHB2-MAX-MYC-USF1" motif, but also with the "AHR-ARNT-ARNT2" motif that corresponds to the complex of Arnt2 and Ahr, is also a predicted direct target of the AAGUGCU seed family which we validated in a luciferase assay (Fig. 3.6B). This TF forms heterodimers with the aryl-hydrocarbon receptor (AHR) [269] and appears to be involved in the differentiation of ESCs into endothelial cells under hypoxic conditions [270], but otherwise little is known about its function. Given that Arnt2 and Myc [271] share the same binding motif, an interesting hypothesis is that Arnt2 competes with Myc for binding to targets and that its downregulation by AAGUGCU miRNAs allows Myc to act at promoters which would otherwise be bound by Arnt2. This hypothesis is again consistent with a positive Myc activity in ESCs, in which these miRNAs are expressed. The model that we propose based on these results is that miRNAs of the AAGUGCU family regulate the cell cycle and the G→S transition through multiple pathways that come together in the increased expression of the crucial Myc regulator (Fig. 3.6D). The miRNAs are able to downregulate the Mxd3 antagonist of Myc, the E2F5 repressor which would in turn result in the increased expression of E2F5 targets including Myc, and can downregulate Arnt2 which may compete with Myc for binding to regulatory sites.

3.3.5 *miRNAs of the AAGUGCU seed family control multiple epigenetic regulators*

As TFs, epigenetic regulators are also enriched among the targets of miRNAs [115]. A role for the miR-290-295 cluster in epigenetic regulation was already proposed by Sinkkonen et al. [180], who found that expression of retinoblastoma-like 2 (Rbl-2) protein, a known repressor of the *de novo* methyltransferases, is controlled by these miRNAs. Through our analysis we found that the AAGUGCU miRNAs directly target the epigenetic regulator BAF170 (Smarcc2), a component of ATP-dependent, BAF (BRG1-associated factor) complexes (also known as SWI/SNF complexes) that remodel the nucleosome structure and thereby regulate gene expression (reviewed in [35]). The activity of the BAF170 motif changed significantly upon AAGUGCU miRNA expression in miRNA-deficient ESCs (Fig. 3.7A, Table 3.1), accompanied by consistent downregulation of BAF170 mRNA (Fig. 3.7B). Comparing constructs with and without the putative miR-294 binding site in the BAF170 3' UTR in a luciferase assay we found that BAF170 is significantly downregulated by miR-294 (Fig. 3.7C), indicating that BAF170 is indeed a direct target of miR-294. Recently, it was shown that BAF170 is downregulated during miR-302-367-based reprogramming and that BAF170 knockdown increases the number of iPSC colonies in somatic cell reprogramming [228]. As miRNAs of the miR-302-367 cluster share the seed sequence with miR-294, it is likely that miR-294 has similar effects on BAF170 expression and pluripotency.

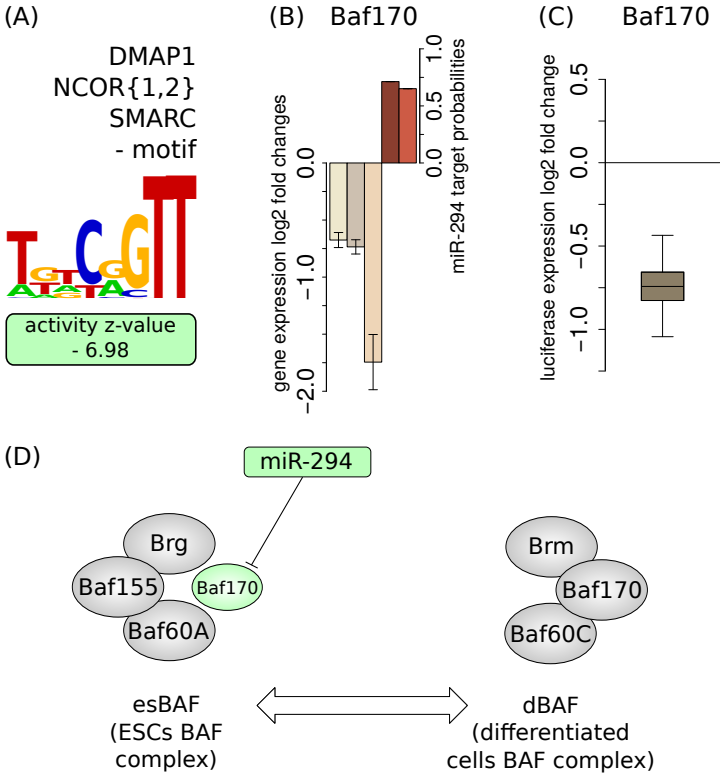


Figure 3.7: The BAF170 (Smarcc2) component of the dBAF chromatin remodeling complex is a direct target of miR-294. **(A)** MARA analysis reveals a negative activity change of the "DMAP1-NCOR{1,2}-SMARC" motif in the presence of miR-294. **(B)** Expression of BAF170 (Smarcc2) is consistently downregulated in the presence of miR-294 in all considered experimental data sets (±1.96*SEM; n=3; color scheme as in Figure 3.3). **(C)** A luciferase construct carrying the BAF170 3' UTR is downregulated upon co-transfection with miR-294 relative to a construct carrying a mutated 3' UTR (n=9). **(D)** Model of the possible involvement of miR-294 in the maintenance of the ESC-specific chromatin remodeling complex esBAF. The miRNA-induced reduction in BAF170 levels may contribute to the maintenance of appropriate levels of esBAF complexes in ESCs thereby maintaining self-renewal and proliferation [215]. Color, shapes and lines scheme is as in Figure 3.6.

The model that emerges from these studies is that the AAGUGCU family of miRNAs may play a role in the remodeling of BAF complexes. In ESCs, the BAF complex (esBAF), which contains a BAF155 subunit, shares a large proportion of target genes with the pluripotency-associated transcription factors Oct4, Sox2 and Nanog [216] and is required for the self-renewal and maintenance of pluripotency in mouse ESCs [215]. Consistently, overexpression of esBAF components was found to promote reprogramming [217]. In differentiated cells however, the so-called differentiated cell BAF complex (dBAF) [234], contains the BAF170 and not the BAF155 subunit [215]. The fact that induced BAF170 expression in ESCs decreases the level of BAF155 protein suggested that BAF170 can displace BAF155 from esBAF, thereby increasing its degradation rate [215]. By preventing expression of BAF components that are specific to differentiated cells and that antagonize embyonic state-

specific BAF (Fig. 3.7D) the AAGUGCU family of miRNAs may promote an ESC-specific epigenetic state.

3.4 DISCUSSION

It has been established that ESC-specific miRNAs that share an AAGUGCU seed region are among the regulatory factors that are necessary to maintain a pluripotent embryonic stem cell state. Strikingly, over-expression of a cluster of ESC-specific miRNAs was found sufficient for inducing reprogramming of differentiated cells into induced pluripotent stem cells. This suggests that the miRNAs can set into motion an entire regulatory cascade that leads to cell reprogramming. Several studies determined the gene expression profiles of ESCs that did and did not express AAGUGCU family miRNAs. An insight emerging from these studies were that miR-290-295 miRNAs regulate the cell cycle and apoptosis, either directly or indirectly. To better understand how the direct regulatory factor targets of these miRNAs contribute to pluripotency, we made use of a recently developed method, called Motif Activity Response Analysis, that models gene expression in terms of computationally predicted regulatory sites. The approach originates in regression models that were first proposed by Bussemaker et al. [272] for inferring regulatory elements from gene expression data. However, MARA's goal is different. It uses predicted regulatory sites in combination with a linear model to infer from gene expression data the activities of transcriptional regulators. The first application of MARA [50] to the reconstruction of the core transcriptional regulatory network of a differentiating human cell line, demonstrated that the method can successfully infer key regulatory interactions *ab initio*. Notably, it was found that MARA accurately infers the activities of the key regulatory motifs, in spite of computational predictions of regulatory sites being error-prone, and of gene expression likely being a much more complex function of the regulatory sites. The power of the method stems from the fact that motif activities are inferred from the *statistics* of expression of hundreds to thousands of putative target genes of each regulatory motif. Here we have used an extended version of the MARA model, which also includes predicted miRNA binding sites, to infer both transcriptional and post-transcriptional regulators of mRNA expression levels. A similar approach was recently applied by Setty et al. [273] to reconstruct the regulatory networks in glioblastoma. The TF targets of the AAGUGCU miRNAs that we identified with the extended MARA model had the following properties:

1. the activity of their corresponding motif changed significantly upon expression of the AAGUGCU miRNA(s), meaning that the predicted targets of these regulators showed, on average, consistent expression changes.

2. their expression was consistently downregulated at the mRNA level upon expression of the AAGUGCU miRNA(s).

3. they were predicted as direct targets of the AAGUGCU family of miRNAs by miRNA target prediction programs.

4. they were consistently (i.e. within every analysed data set) predicted by MARA to be directly regulated by the AAGUGCU seed family of miRNAs on the basis of the dependence on their expression changes on the presence of the miRNA binding sites in their 3' UTRs.

5. they could be confirmed as AAGUGCU miRNA targets with luciferase assays.

Altogether, these lines of evidence firmly establish these transcriptional regulators as direct targets of the AAGUGCU seed family miRNAs, forming the first layer downstream of this miRNAs in the regulatory network of pluripotency. First, our analysis suggests that AAGUGCU miRNAs target the cell cycle, and in particular the G1→S transition, through multiple pathways. By targeting the repressive cell cycle regulator E2f5, the miRNAs might directly promote the G1→S transition. In addition, the miRNAs seem to increase the activity of the proliferation-associated TF Myc through multiple indirect routes, including shifting the balance between Myc and its antagonist Mxd3 within transcription regulatory complexes that act on Myc target genes. Second, we found that the AAGUGCU miRNAs may affect the balance between chromatin remodeling complexes that are active in ESCs and in differentiated cells, a function probably important for keeping specific genomic regions from being silenced through heterochromatin formation. Third, we found that the AAGUGCU miRNAs directly target the interferon regulatory factor Irf2, whose expression is strongly increased in $DCR^{-/-}$ cells, consistent with a significant change in the regulatory impact that we inferred for this factor. Finally, our analysis uncovers a few transriptional regulators that have previously not been connected to the transcriptional network of pluripotent stem cells, including Foxj2, whose expression is strongly affected by the miRNAs and the Clock (circadian locomotor output cycles kaput) TF. Interestingly, circadian oscillations are not present in mouse ES cells, but are switched on during differentiation, and then disappear again upon reprogramming of differentiated cells into iPSCs [274]. It is thus tempting to speculate that circadian oscillations in ESCs may be actively suppressed by the AAGUGCU miRNAs and that downregulation of these miRNAs during development may be necessary for the establishment of circadian rhythms. However, the response of the 3' UTR of Clock in luciferase assays was very variable in our hands, and we were not able to unambiguously validate it as a direct target of miR-294. As mentioned before, the AAGUGCU seed motif is not unique to miRNAs of the mouse-specific miR-290-295 cluster. It also occurs in the miR-302 family of miRNAs that is present in human and in a shifted version (at positions 3-9 instead of 2-8) it occurs in the miR-17/20a miRNAs of the oncogenic miR-17-92 cluster. Although miR-19 has been reported to be the key oncogenic component of this cluster [275], the strong effects that AAGUGCU miRNAs exert on the cell cycle raise the question of whether miR-17 and miR-20a may not play a role similar to miR-294 in malignant cells. In summary, our

analysis demonstrates that combining accurate predictions of regulatory elements with analysis of transcriptome-wide mRNA expression changes in response to specific manipulations is a general and powerful approach to uncovering key regulators within gene expression networks. In the future, incorporation of measurements of miRNA expression as well as of predictions of transcription factor binding sites in miRNA genes will enable identification of feedback loops between miRNAs and transcription factors that are known to operate in many systems.

3.5 MATERIALS AND METHODS

3.5.1 *Experimental data sets*

Supplementary Table A.1 summarizes the data sets that we obtained from the GEO database of NCBI and that we have used in our study. Each data set covers at least two distinct experimental conditions, with three replicates per condition. The first condition of every data set corresponds to an embryonic stem cell line deficient in mature microRNAs due to Dicer- or Dgcr8-knockout. The second condition corresponds to either an embryonic stem cell line expressing the entire complement of embryonically expressed microRNAs or the knockout cell line transfected with miR-294 or with mimics of the miR-290 cluster miRNAs (mir-290, mir-291a-3p, mir-292-3p, mir-293, mir-294 and mir-295).

3.5.2 *Microarray analysis*

3.5.2.1 *Computational analysis of Illumina MouseWG-6 v2.0 Expression BeadChips from Hanina et al. (2010)*

We downloaded the processed data from the GEO database of NCBI (accession no. GSE20048). Probe-to-gene associations were made by mapping the probe sequences (provided by the authors) to the set of mouse transcript sequences (downloaded 2011-02-19 from the UCSC Genome Bioinformatics web site). We computed average gene expression levels as weighted averages of the signals of all probes that perfectly matched to at least one transcript of the gene. Whenever a probe mapped to multiple genes, a weight of $1/n$ was assigned to each of the n genes to which the probe matched. For a given replicate experiment, the $\log_2$ expression fold change of each gene was then determined by subtracting the $\log_2$-average expression of the gene in the first condition (control) from the $\log_2$-average expression in the second condition (treatment).

3.5.2.2 *Computational analysis of Affymetrix Mouse Genome 430 2.0 chips from Sinkkonen et al. (2008) and Zheng et al. (2011)*

We downloaded the data from the GEO database (accessions GSE8503, GSE7141 and GSE30012) and analyzed the CEL files with the R software

(http://www.R-project.org) using the BioConductor affy package [276]. We used the GCRMA algorithm [277] for background correction and the MClust R package [278] to fit a two-component Gaussian mixture model to the $\log_2$-probe intensities and classify probes as expressed or not expressed. A probe was considered for further analysis if it was consistently classified as expressed in all three replicates of at least one of the two experimental conditions. The remaining probes were quantile normalized across all conditions and replicates of a particular experiment. Probe-to-gene associations were made by mapping probe sequences (provided on the Affymetrix website, http://www.affymetrix.com) to mouse transcript sequences (as used by motif activity response analysis (MARA), downloaded from UCSC Genome Bioinformatics web site as described above). We then computed $\log_2$-gene expression fold changes as described for Illumina Expression BeadChips (see above).

3.5.2.3 *Computational analysis of Affymetrix Mouse Gene 1.0 ST chips from Melton et al. (2010)*

We downloaded the data from the GEO database (accession no. GSE18840) and analyzed the CEL files with the R Bioconductor oligo package [279]. We used the RMA algorithm [280] for background adjustment. The rest of the analysis, including the classification of probes into expressed/not expressed, the quantile normalization, and the calculation of $\log_2$ gene expression fold changes, was carried out as described above.

3.5.2.4 *Proportions of AAGUGCU miRNA seed family targets among genes that are consistently downregulated in multiple experiments*

For each gene and each experiment, we calculated the standard error in its $\log_2$ fold change across the replicates. A gene was considered significantly downregulated when it was down-regulated more than 1.96 standard-errors. We then determined the intersection set of significantly downregulated genes for every possible subset of the experiments *S={MeltonDGCR8KOVs294, SinkkonenDicerKOVs290, SinkkonenDiverKOVsWT}*. Subsequently, for every obtained intersection set, the proportion of AAGUGCU miRNA seed family targets (TargetScan aggregate P_{CT} score predictions [143]) was determined and plotted against the size of the corresponding intersection set.

3.5.2.5 *Combined motif activity response analysis of TFs and miRNAs*

We carried out the MARA [253] separately for each experimental data set. MARA relates the expression level E driven by individual promoters (measured by microarrays) to the number of binding sites N that various regulators have in the promoters using a simple linear model:

$$E_{ps} = \tilde{c}_s + c_p + \sum_m N_{pm} A_{ms}, \tag{3.1}$$

where c_p is a term reflecting the basal expression of promoter p, $\tilde{c}_s$ reflects the mean expression in sample s, and A_{ms} is the (unknown) *activity* of binding motif m in sample s (where with "sample" we refer to any individual replicate of any condition of a data set, see section 3.5.1). That is, using the predicted site-counts N_{pm} and the measured expression levels E_{ps} we used an approximation (4.1) to infer the activities A_{ms} of all motifs across all samples by ridge regression. In our analyses, we considered a curated set of 189 transcription factor binding motifs (for detailed information about the motifs and the corresponding transcription factors see Supplementary Table A.7). Furthermore, we included the binding sites in the 3' UTRs of mRNAs of 85 miRNA families by incorporating aggregate P_{CT} scores as provided by TargetScan[143] (predictions downloaded on the 27th of March 2012 from the TargetScan website, http://www.targetscan.org). miRNAs are grouped into families by their seed sequences and in particular the *AAGUGCU* seed family corresponds to the following miRNAs: *mmu-miR-291a-3p*, *mmu-miR-294*, *mmu-miR-295*, *mmu-miR-302a*, *mmu-miR-302b* and *mmu-miR-302d*. A aggregate P_{CT} score was assigned to a promoter by averaging the aggregate P_{CT} scores of transcripts associated with this promoter. For a given motif m, MARA provides for each sample s motif activities A^*_{ms} and associated errors σ_{ms}. More specifically, marginalizing over all other motifs, the likelihood $P(D|A_{ms})$ of the expression data D given the activity of a given motif is proportional to a Gaussian

$$P\left(D \mid A_{ms}\right) \propto \exp\left[-\frac{1}{2}\frac{\left(A_{ms}-A^*_{ms}\right)^2}{\sigma^2_{ms}}\right]. \tag{3.2}$$

Given that all analysed experiments were performed in multiple replicates we were interested in averaging motif activities across replicates and we used the following Bayesian approach. For each motif m separately, we assumed that the activities across a group g of replicates belonging to a specific condition of an experiment (see section 3.5.1) are normally distributed around some (unknown) mean $\bar{A}_{mg}$ with (unknown) variance σ^2_{mg}

$$P\left(A_{ms} \mid \bar{A}_{mg}, \sigma_{mg}\right) = \frac{1}{\sqrt{2\pi}\sigma_{mg}} \exp\left[-\frac{1}{2}\frac{\left(A_{ms}-\bar{A}_{mg}\right)^2}{\sigma^2_{mg}}\right]. \tag{3.3}$$

By combining the prior (3.3) with the likelihood (3.2) for each replicate sample $s \in g$ and integrating out the (unobserved) true activities A_{ms} in each of the replicates, we obtained the probability of the form

$$P\left(D \mid \bar{A}_{mg}, \sigma_{mg}\right) = \prod_{s\in g} \frac{1}{\sqrt{2\pi(\sigma^2_{mg}+\sigma^2_{ms})}} \exp\left[-\frac{(A^*_{ms}-\bar{A}_{mg})^2}{2(\sigma^2_{mg}+\sigma^2_{ms})}\right]. \tag{3.4}$$

Formally, we would next integrate out the unknown standard deviation of activities in the group σ_{mg} of this likelihood. Unfortunately, this integral cannot

be performed analytically. We thus approximated the integral by the value of the integrand at its maximum, i.e. we numerically found the value of σ_{mg} that maximizes (3.4). Assuming an uniform prior over mean activity $\bar{A}_{mg}$, we obtained the expression for $P\left(\bar{A}_{mg} \mid D\right)$ to be again a Gaussian with mean

$$\bar{A}^*_{mg} = \frac{\sum_{s\in g} \frac{A^*_{ms}}{(\sigma^*_{mg})^2+(\sigma_{ms})^2}}{\sum_{s\in g} \frac{1}{(\sigma^*_{mg})^2+(\sigma_{ms})^2}}, \tag{3.5}$$

and error

$$\bar{\sigma}^*_{mg} = \sqrt{\frac{1}{\sum_{s\in g} \frac{1}{(\sigma^*_{mg})^2+(\sigma_{ms})^2}}}. \tag{3.6}$$

where σ^*_{mg} is the maximum likelihood estimate of (3.4). We call (3.5) and (3.6) averaged activities and averaged errors, respectively. To identify motifs that consistently change in their activities across experiments, we wanted to further average motif activities across these experiments. However, because of the inherent variations in the *scale* of expression variations in the different experiments, the motif activities also varied in scale across the experiments. Thus, before averaging we first standardized the motif activities across the two conditions a and b. That is, for a given experiment we defined a scale L

$$L = \sqrt{\frac{\left(\bar{A}^{*b}_{mg}\right)^2 + \left(\bar{A}^{*a}_{mg}\right)^2}{2}}, \tag{3.7}$$

and rescaled the activities

$$\tilde{A}^*_{mg} = \frac{\bar{A}^*_{mg}}{L} \tag{3.8}$$

and their errors

$$\tilde{\sigma}^*_{mg} = \frac{\bar{\sigma}^*_{mg}}{L}. \tag{3.9}$$

These condition-specific, averaged and rescaled activities ($\tilde{A}^*_{mg}$) and errors ($\tilde{\sigma}^*_{mg}$) from the different experiments were then combined into two groups, i.e. the group of a conditions and the group of b conditions, and for each group we again averaged the activities exactly as described above for the replicates. To rank the activity changes between two different experimental conditions (presence/absence of miRNAs) we determined a z-value for every motif m by dividing the change in averaged activities between the two different conditions a and b by the averaged errors as follows:

$$z = \frac{\tilde{A}^{*b}_{mg} - \tilde{A}^{*a}_{mg}}{\sqrt{\left(\sigma^{*b}_{mg}\right)^2 + \left(\sigma^{*a}_{mg}\right)^2}}. \tag{3.10}$$

Consequently, from the results of equation (3.10) we obtained a global z-value-based ranking of the motifs.

3.5.2.6 *Motif-motif interaction network*

To uncover which transcription factors were targeted by a particular motif m, we focused only on those transcription factor genes, whose promoters were consistently (in all experiments) predicted by MARA to be targets of motif m. MARA computes a target score S for each potential target promoter of motif m. S corresponds to the log-likelihood ratio of the data D assuming the promoter is indeed a target, and assuming the promoter is independent of the regulator, i.e

$$S = \log\left[\frac{P(D|\text{target})}{P(D|\text{nottarget})}\right]. \tag{3.11}$$

Assuming a uniform prior of $1/2$ that the promoter is indeed a target, the posterior probability p that the promoter is a target given the data is:

$$p = \frac{1}{1+\frac{1}{e^S}}. \tag{3.12}$$

To obtain a combined probability p_c that a gene is a target of a particular motif across N different experiments the probability product was calculated by multiplying the probabilities p_n obtained in individual experiments n, i.e.

$$p_c = \prod_{n=1}^{N} p_n. \tag{3.13}$$

3.5.3 *Evaluating miR-294 targets with luciferase assays*

3.5.3.1 *Cloning, cell culture and luciferase Assay*

We PCR-amplified 3' UTRs fragments of the putative target genes from Normal Murine Mammary Gland (NMuMG) genomic DNA and cloned them into pGEM-T Easy vector (Promega; Cat. No.A1360). We used site-directed mutagenesis and the QuickChange II kit (Stratagene; Cat. No.200524-5) to generate deletion mutant constructs that differed in a few nucleotides in the miR-294 seed-matching region from the wild type construct. All constructs, wild-type and mutated, were verified by sequencing and then subcloned into the empty psiCHECK-2 vector (Promega; Cat. No.C8021) at XhoI - NotI restriction sites. The sequences of the primers used for cloning and mutagenesis can be found in Supplementary Tables A.9 and A.10, respectively. NMuMG cells were reverse-transfected with Lipofectamine2000 reagent (Invitrogen; Cat. No.11668019), and the corresponding psiCHECK-2 constructs in the presence of 50nM Syn-mmu-miR-294-3p mimic (QIAGEN; Cat. No. MSY0000372), or 50nM of non-targeting negative control siRNA (Microsynth). Between 36 and 48 hours post-transfection cells were collected and both renilla and firefly luciferase activities were measured using Dual Glo Luciferase Assay System (Promega; Cat. No.E2940). For each gene, expression was measured for both constructs in 3 separate experiments, and each experiment contained 3 technical replicates.

3.5.3.2 *Analysis of the luciferase data*

We denote by w_{ir} the logarithm (base 2) of the expression level of the luciferase construct containing the wild type 3' UTR in experiment i replicate r and by m_{ir} the analogous expression for the mutant construct. For each gene the data thus consist of 9 values w and 9 values m. We took into account two sources of variability, namely true expression variability across experiments and 'measurement noise' between replicates. We first describe the measurement noise. Assuming the true expression of the wild type was w_i, we assumed the probability to measure expression level w_{ir} (in a given replicate r) is given by a Gaussian distribution with a certain variance τ_i:

$$P(w_{ir}|w_i,\tau_i) = \frac{1}{\tau_i\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{w_{ir}-w_i}{\tau_i}\right)^2\right], \tag{3.14}$$

thus allowing for the possibility that each experiment i has a *different* level of noise τ_i between replicates. The probability of the wild type data of experiment i, assuming that τ_i is given, is simply the product of expressions $P(w_{ir}|w_i,\tau_i)$ over the three replicates $r = 1$ through 3. Using $\langle w_i\rangle$ and $\mathrm{var}(w_i)$ to denote the mean and variance of the measurement across the replicates, we can rewrite this as

$$P(\{w_{ir}\}|w_i,\tau_i) \propto \frac{1}{\tau_i^3} \exp\left[-\frac{3}{2}\left(\frac{w_i-\langle w_i\rangle}{\tau_i}\right)^2 - \frac{3}{2}\frac{\mathrm{var}(w_i)}{\tau_i^2}\right]. \tag{3.15}$$

Integrating over the unknown variable τ_i from 0 to infinity with a scale prior $P(\tau_i) \propto 1/\tau_i$ we obtain

$$P(\{w_{ir}\}|w_i) \propto \left(1+\frac{(w_i-\langle w_i\rangle)^2}{\mathrm{var}(w_i)}\right)^{3/2}. \tag{3.16}$$

Approximating this Student-t distribution by a Gaussian, that is, approximating the probability of the data in experiment i by a Gaussian with mean $\langle w_i\rangle$ and variance $\mathrm{var}(w_i)$, we have

$$P(\{w_{ir}\}|w_i) \approx \sqrt{\frac{3}{\mathrm{var}(w_i)}} \exp\left[-\frac{3(w_i-\langle w_i\rangle)^2}{2\mathrm{var}(w_i)}\right]. \tag{3.17}$$

Since the variability between replicates is much smaller than the variability across experiments, this approximation will have a negligible effect on the final outcome. For the true variability between experiments, we denote by w the 'true' average expression of the wild type construct. We assume that the deviation of the level w_i in experiment i from the mean w follows a Gaussian distribution with variance σ. We thus have

$$P(w_i|w,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{w_i-w}{\sigma}\right)^2\right]. \tag{3.18}$$

To obtain the probability of the data given w we multiply $P(\{w_{ir}\}|w_i)$ by $P(w_i|w,\sigma)$ and integrate over the unknown expression level w_i. We then obtain

$$P(\{w_{ir}\}|w,\sigma) \propto \frac{1}{\sqrt{\sigma^2+\mathrm{var}(w_i)/3}} \exp\left[-\frac{(\langle w_i\rangle-w)^2}{2(\sigma^2+\mathrm{var}(w_i)/3)}\right]. \tag{3.19}$$

The interpretation of this formula is straight-forward. The deviation between the mean $\langle w_i \rangle$ of the observations in experiment i, and the average level w is Gaussian-distributed with a variance that is the sum of the variability σ^2 across experiments, and the variability $\mathrm{var}(w_i)/3$ associated with estimating w_i from the 3 replicate measurements due to measurement noise. For the measurements of the mutant construct in experiment i we obtain an analogous equation

$$P(\{m_{ir}\}|m,\tilde{\sigma}) \propto \frac{1}{\sqrt{\tilde{\sigma}^2 + \mathrm{var}(m_i)/3}} \exp\left[-\frac{(\langle m_i \rangle - m)^2}{2(\tilde{\sigma}^2 + \mathrm{var}(m_i)/3)}\right], \tag{3.20}$$

where we have introduced the variability $\tilde{\sigma}$ of the true expression of the mutant construct across replicates. What we are interested in is the *difference* $w - m$ in log-expression of the wild type and mutant construct. To this end we define $\mu = w - m$ and $y = (m + w)/2$ and integrate over y. We then obtain

$$P(\{w_{ir}\},\{m_{ir}\}|\mu,\sigma,\tilde{\sigma}) \propto \frac{1}{\sqrt{\sigma^2 + \tilde{\sigma}^2 + \mathrm{var}(w_i)/3 + \mathrm{var}(m_i)/3}} \exp\left[-\frac{(\langle w_i \rangle - \langle m_i \rangle - \mu)^2}{2(\sigma^2 + \tilde{\sigma}^2 + \mathrm{var}(w_i)/3 + \mathrm{var}(m_i)/3)}\right]. \tag{3.21}$$

This is again a Gaussian with mean $\langle w_i \rangle - \langle m_i \rangle$ and a variance that is the sum of all variances σ^2, $\tilde{\sigma}^2$, $\mathrm{var}(w_i)/3$, and $\mathrm{var}(m_i)/3$. Clearly, although both σ^2 and $\tilde{\sigma}^2$ are unknown, the only variable that enters in our equations is their sum. We thus simplify the notation by defining this sum as

$$\gamma^2 = \sigma^2 + \tilde{\sigma}^2. \tag{3.22}$$

Similarly, we redefine the variance associated with measurement noise as

$$t_i^2 = \mathrm{var}(w_i)/3 + \mathrm{var}(m_i)/3, \tag{3.23}$$

which leads to

$$P(\{w_{ir}\},\{m_{ir}\}|\mu,\gamma) \propto \frac{1}{\sqrt{\gamma^2 + t_i^2}} \exp\left[-\frac{(\langle w_i \rangle - \langle m_i \rangle - \mu)^2}{2(\gamma^2 + t_i^2)}\right]. \tag{3.24}$$

We now combine the data from the different experiments and remove the final unknown variable γ. The probability of all data given the variable of interest μ and unknown variability parameter γ is simply the product

$$P(D|\mu,\gamma) = \prod_{i=1}^{3} P(\{w_{ir}\},\{m_{ir}\}|\mu,\gamma). \tag{3.25}$$

To obtain the probability of the data D given μ we multiply this expression with a scale prior for γ, i.e. $P(\gamma) = 1/\gamma$, and integrate over γ:

$$P(D|\mu) = \int_0^\infty P(D|\mu,\gamma)\frac{d\gamma}{\gamma}. \tag{3.26}$$

We performed the integration numerically with Mathematica to obtain $P(D|\mu)$, and used Bayes' theorem to compute the posterior distribution of the parameter μ, $P(\mu|D)$ as $P(D|\mu)/\int_{-\infty}^{\infty} P(D|\mu)d\mu$. Finally, we determined the 5 percentile, the 25 percentile, the median, the 75 percentile, and the 95 percentile of this distribution again with the Mathematica software.

3.5.4 *mouse ESC (mESC) culture*

The generation of Dicer(DCR)$^{flox/flox}$ and DCR$^{-/-}$ mouse ES cell lines has been described elsewhere [281]. The cells were routinely screened for both pluripotency and differentiation markers (see Supplementary Figure A.4). Both mESC cell lines were maintained in Dulbecco's Modified Eagles Medium (DMEM) (Gibco; 41966-029) supplemented with 15% of a special batch of fetal bovine serum tested for optimal growth of mESCs. In addition, the DMEM contained 1000 U/ml of a homegrown recombinant LIF (a kind gift of Thomas Grentzinger), 0.1mM 2β-mercaptoethanol (Millipore; ES-007-E), 1x L-Glutamine (Gibco; 25030-024), 1x Sodium Pyruvate (Gibco; 11360), and 1x Minimum Essential Medium, Non-Essential Amino Acids (MEM, NEAA) (Gibco; 11140-35). The cells were grown on gelatin-coated (Sigma; G1393) dishes. The medium was changed daily, and the cells were sub-cultured every 2-3 days. To induce NF-κB signaling, mouse ESCs were treated with 20ng/ml TNFα (Cell Signaling Technology; 5178) for 24 hrs.

3.5.5 *Quantitative RT-PCR*

Total RNA was extracted from mESCs using Tri Reagent (Sigma; T9424) following the supplier's protocol. Contaminating DNA was removed using the RQ1 RNase-Free DNase kit (Promega; M6101). The resulting DNA-free RNA was then purified using the RNeasy MinElute Cleanup kit (Qiagen; 74204) and quantified using Nanodrop. Superscript III (Invitrogen; 18080) was then used to create cDNA following the manufacturer's recommendations. The cDNA was finally purified using QIAquick PCR Purification kit (Qiagen; 74204), quantified using Nanodrop, and diluted to 8 ng/μl. Each qRT-PCR reaction was run using 2μl of the purified cDNA in triplicate (n=3) using Power SYBR Green PCR Master Mix (Applied Biosystems; 4367659) on a StepOne Plus RT-PCR System (Applied Biosystems). The following primer pairs were used in this study:

- Mouse IRF2 Fwd: 5'-CTG GGC GAT CCA TAC AGG AAA-3'
- Mouse IRF2 Rev: 5'-CTC AAT GTC GGG CAG GGA AT-3'
- Mouse E2F5 Fwd: 5'-GTT GTG GCT ACA GCA AAG CA-3'
- Mouse E2F5 Rev: 5'-GGC CAA CAG TGT ATC ACC ATG A-3'

- Mouse c-Myc Fwd: 5'-GTT GGA AAC CCC GCA GAC AG-3'
- Mouse c-Myc Rev: 5'-ATA GGG CTG TAC GGA GTC GT-3'
- Mouse GAPDH Fwd: 5'-CAT CAC TGC CAC CCA GAA GAC TG-3'
- Mouse GAPDH Rev: 5'-ATG CCA GTG AGC TTC CCG TTC AG-3'

qRT-PCR data were normalized using GAPDH expression and evaluated using the $2^{-\Delta\Delta Ct}$ method [282]. Significant changes in gene expression were identified based on Student's t-test.

3.5.6 *Western Blots*

To extract total proteins from mESCs, Radioimmunoprecipitation assay (RIPA) buffer supplemented with 1x Complete, EDTA-free protease inhibitor cocktail (Roche; 11873580001) was used to lyze cell pellets. Cytosolic and nuclear protein fractions were enriched using a series of lysis buffers as follows:

- Lysis Buffer 1 (LB1): 50 mM Hepes-KOH, pH 7.5; 140 mM NaCl; 1 mM EDTA, pH 8.0; 10% v/v Glycerol; 0.5% v/v NP-40; 0.25% v/v Triton X-100.
- Lysis Buffer 2 (LB2): 10 mM Tris-HCl, pH 8.0; 200 mM NaCl; 1mM EDTA, pH 8.0; 0.5 mM EGTA, pH 8.0.
- Lysis Buffer 3 (LB3): 10mM Tris-HCl, pH 8.0; 100 mM NaCl; 1 mM EDTA, pH 8.0; 0.5 mM EGTA, pH 8.0; 0.1% v/v Na-Deoxycholate; 30% v/v N-Lauroylsarcosine.

All lysis buffers were supplemented with the protease inhibitor cocktail immediately before use. The cytosolic fraction was extracted by lyzing the cell pellets in LB1 that leaves the nuclear membrane intact. The nuclei were then pelleted (1,350 x g; 4°C; 5 mins), washed with LB2, pelleted once more and finally lyzed with LB3 to release the nuclear contents. All protein lysates were quantified using the BCA Protein Assay kit (23227; Pierce). The following antibodies (dilution 1:1000) were used in this study:

- Anti-IRF2 (Center) rabbit IgG (Abgent; AP11225c)
- Anti-NF-κB p65 (D14E12) XP rabbit IgG (Cell Signaling Technology; 8242)
- Anti-GAPDH (6C5) mouse IgG (Santa Cruz Biotechnology; sc-32233)
- Anti-Histone H3 (C-16) goat IgG (Santa Cruz Biotechnology; sc-8654)
- HRP-conjugated Polyclonal swine Anti-Rabbit (Dako; P0217)
- HRP-conjugated Polyclonal rabbit Anti-Mouse (Dako; P0260)

- HRP-conjugated Polyclonal rabbit Anti-Goat (Dako; P0449)

Western blot signals were visualized with the enhanced chemiluminescence (ECL) blotting detection reagents (RPN2106; GE Healthcare). Cytosolic enrichment was confirmed via a postive GAPDH signal, while nuclear enrichment was confirmed by Histone H3. Western blot quantifications were performed using the ImageJ software by quantifying the pixels of each band and normalizing against a housekeeper such as Histone H3.

3.6 AUTHORS INFORMATION

3.6.1 *List of authors*

The following authors have contributed to the work discussed in Chapter 3:

1. Andreas Johannes Gruber[1] (Abbr.: AJG),
2. William A. Grandy[1] (Abbr.: WAG),
3. Piotr J. Balwierz[1] (Abbr.: PJB),
4. Yoana A. Dimitrova[1] (Abbr.: JAD),
5. Mikhail Pachkov[1] (Abbr.: MP),
6. Constance Ciaudo[2] (Abbr.: CC),
7. Erik van Nimwegen[1] (Abbr.: EvN) &
8. Mihaela Zavolan[1] (Abbr.: MZ)

whereat author affiliations are as follows:

1 Biozentrum, University of Basel, Klingelberstrasse 50-70, CH-4056 Basel, Switzerland

2 ETH Zürich, Otto-Stern-Weg 7, CH-8093 Zürich, Switzerland

3.6.2 *Author contributions*

The listing of authors in the previous subsection (3.6.1) was performed according to the authors' contributions, whereat the first author (AJG) contributed most and subsequent authors decreasingly. However, the last three authors are principal investigators and thus their listing follows the opposite ranking (the last contributed the most and the preceding two authors decreasingly).
In detail, using the abbreviations specified in the previous subsection (i.e. 3.6.1): MZ and AJG designed the project. AJG, PJB and MP implemented the extended MARA approach and AJG performed the analysis with help

from EvN and MZ. CC provided the ESCs and helped to culture it. YAD and WAG carried out the experiments and EvN analyzed the data from the luciferase assays. AJG, EvN and MZ wrote the manuscript. All authors read and approved the final manuscript.

3.7 SUPPLEMENTARY MATERIALS

Supplementary materials can be found in Appendix A.

3.8 ACKNOWLEDGMENTS

We thank the members of the Zavolan group for feedback on the manuscript.

3.9 FUNDING

Using the abbreviations specified in subsection 3.6.1: The work discussed in this chapter was supported by the Swiss National Science foundation grant #31003A_127307, by a European Research Council Starting Grant to MZ and a Werner Siemens Fellowship to AJG.

4

ISMARA: AUTOMATED MODELING OF GENOMIC SIGNALS AS A DEMOCRACY OF REGULATORY MOTIFS

The work discussed in this chapter was conducted in collaboration with the van Nimwegen lab and published in Genome Research in 2014 (see reference [253]).

4.1 ABSTRACT

Accurate reconstruction of the regulatory networks that control gene expression is one of the key current challenges in molecular biology. Although gene expression and chromatin state dynamics are ultimately encoded by constellations of binding sites recognized by regulators such as transcriptions factors (TFs) and microRNAs (miRNAs), our understanding of this regulatory code and its context-dependent read-out remains very limited. Given that there are thousands of potential regulators in mammals, it is not practical to use direct experimentation to identify which of these play a key role for a particular system of interest.

We developed a methodology that models gene expression or chromatin modifications in terms of genome-wide predictions of regulatory sites, and completely automated it into a web-based tool called ISMARA (Integrated System for Motif Activity Response Analysis), located at http://ismara.unibas.ch. Given as input only gene expression or chromatin state data across a set of samples, ISMARA identifies the key TFs and miRNAs driving expression/chromatin changes and makes detailed predictions regarding their regulatory roles. These include predicted activities of the regulators across the samples, their genome-wide targets, enriched gene categories among the targets, and direct interactions between the regulators.

Applying ISMARA to data sets from well-studied systems, we show that it consistently identifies known key regulators *ab initio*. We also present a number of novel predictions including regulatory interactions in innate immunity, a master regulator of mucociliary differentiation, TFs consistently disregulated in cancer, and TFs that mediate specific chromatin modifications.

4.2 INTRODUCTION

Since the seminal work of Jacob and Monod [283], much has been learned about the molecular mechanisms by which gene expression is regulated, and the molecular components involved. Historically, most work has focused on transcription factors (TFs), arguably the most important regulators of gene expression, which bind to cognate sites in the DNA, and regulate the rate of transcription initiation. However, more recently it has become clear that the state of the chromatin, which can be modulated through modifications of the DNA nucleobases and of the histone tails of nucleosomes, also plays a crucial role. For example, the local chromatin state affects the ability of

TFs to access their binding sites, and the chromatin state can in turn be modified through TF-guided recruitment of chromatin modifying enzymes. Furthermore, an entirely new layer of post-transcriptional regulation has been uncovered in recent years in the form of microRNAs (miRNAs) [112]. These guide RNA-induced silencing complexes to target mRNAs, inhibiting their translation and accelerating their decay [284]. In spite of these many insights, our current understanding of the function of genome-wide gene regulatory networks in mammals is still rudimentary. For example, we only know the sequence specificity of less than half [285–287] of the approximately 1500 [288] TFs in mammalian genomes. Our knowledge of how TF binding is affected by chromatin state, of the combinatorial interactions between TFs and their co-factors, and the impact of post-translational modifications on TF activity, is even more fragmentary. Our understanding of the transcriptome-wide effects of miRNAs on gene expression remains similarly limited. Given that we are clearly still far from being able to develop realistic quantitative models of genome-wide gene regulatory dynamics, the most constructive contribution that computational approaches can currently provide is to develop models that help guide experimental efforts. Due to the dramatic decrease in high-throughput measurement costs, it has become relatively straight forward to measure gene expression (i.e. with micro-array or RNA-seq) or chromatin state (with ChIP-seq) genome-wide across a set of samples for a particular system of interest. Consequently, researchers interested in a particular developmental or cellular differentiation process, or in the response of a tissue to a particular perturbation, have increasingly turned to genome-wide profiling of expression and various chromatin marks, with the aim of using such data to elucidate the key regulatory circuitry acting in their system. However, deriving insights into regulatory circuitry from high-throughput data requires sophisticated computational analysis methods. Over the last years comparative genomic methods have been developed that allow relatively accurate computational prediction of regulatory sites for hundreds of TFs and miRNAs on a genome-wide scale [143, 289, 290]. In addition, through extensive experimental efforts, genome-wide annotations of transcript structures [291, 292] and promoters [293] have become available. Capitalizing on these developments, we recently presented a general method, called Motif Activity Response Analysis (MARA) for inferring key gene regulatory circuitry from genome-wide gene expression data by modeling the observed gene expression dynamics in terms of computationally predicted regulatory sites. We showed that this method can reconstruct core transcription regulatory networks in a human differentiation system *ab initio* [50]. Furthermore, several recent studies confirm that computational modeling of observed expression and chromatin dynamics is a powerful approach to reconstructing regulatory circuitry [294, 295] (to give just two examples), and show that MARA-like approaches can be extended to include miRNA regulation [273] and the dynamics of genome-wide histone modifications [207]. Unfortunately, applying MARA-like methods to high-throughput data is technically challenging and requires the expertise of dedicated computational biology groups. Thus, whereas many labs

are now routinely producing high-throughput data-sets, and methodologies for analyzing such data have been described in the literature, the vast majority of groups that produce data have to develop collaborations with expert computational groups to apply these methods. Indeed, over the last years our group applied MARA to a large range of mammalian systems studied by various experimental collaborators, and experimentally validated predicted regulatory circuitry in these systems [296–306]. Although these studies further validated the power of the method, they required a considerable investment of time and effort for the analysis of each new data-set. Through these experiences we became convinced that lack of easy access to such computational analysis procedures is currently a major bottle-neck in the field, and decided to invest our efforts into developing a completely automated system for performing MARA. Here we present ISMARA (Integrated System for Motif Activity Response Analysis), a completely automated computational tool that aims to make the computational reconstruction of regulatory circuitry from high-throughput data easily accessible to any researcher. Given as input a set of genome-wide gene expression or chromatin state measurements across a number of samples, ISMARA uses motif activity response analysis to identify the key regulators (i.e. TFs and miRNAs) driving gene expression/chromatin state changes across the samples, the activity profiles of these regulators, their target genes, and the sites on the genome through which these regulators act. The analysis combines pre-calculated annotations of regulatory sites for hundreds of regulators across genes in mammalian genomes with automated processing of input data, modeling and parameter inference, and post-processing to provide a large collection of analysis results. To use ISMARA, users only need to upload their data to the web-server http://ismara.unibas.ch/ and submit it to the system, without the need of setting or tuning any parameters. All results are presented through a user-friendly graphical web-interface. In ISMARA the motif activity response analysis has been extended to model not only gene expression data from various platforms (micro-array, RNA-seq), but essentially any sequencing data reflecting a genomic mark (ChIP-seq) including chromatin modifications or TF binding. In addition, ISMARA models not only the effects of TFs on mammalian gene expression, but also the effects of miRNAs. Below we first outline the methodologies that we developed for automating the computational modeling, and provide an overview of all results that ISMARA provides by applying it to RNA-seq data of a human tissue-atlas. After this, we further demonstrate ISMARA using a number of example data sets that highlight different aspects of the method.

4.3 RESULTS

As schematically depicted in Figure 4.1, ISMARA takes raw gene expression (micro-array or RNA-seq) or chromatin state (ChIP-seq) data from any number of samples and automatically models this data in terms of computationally predicted regulatory sites, thereby predicting the genome-wide regulatory interactions that drive the observed expression or chromatin state

changes across the samples. ISMARA is available through a web interface http://www.ismara.unibas.ch as part of our SwissRegulon resources [287]. Users can directly upload unprocessed micro-array (CEL files), RNA-seq, or ChIP-seq data (BED or BAM files) which are then analyzed automatically without the need for any additional input from the user (Figure 4.1B). The results are made available through a web interface and can also be downloaded in flat-file format.

In order to be able to provide such completely automated analysis ISMARA makes use of pre-calculated genome-wide annotations of promoters, sets of transcripts associated with each promoter, multiple alignments of promoter regions across 7 mammals, a curated collection of mammalian regulatory motifs, TFBS predictions for all motifs across all promoters, and predicted target transcripts of miRNAs (Figure 4.1A). Additionally, we developed a substantial number of analysis procedures in order to automatically process and normalize the raw input data (Figure 4.1B) and transform them into a standardized format to which the motif activity response analysis can be applied (Figure 4.1C). The analysis procedures involved in all these steps are outlined in the Methods and detailed in the Supplementary Methods (see Appendix B).

4.3.1 *Overview of the analyses performed by ISMARA*

To give an overview of the analysis results that ISMARA automatically provides for any data-set, and to outline how these analyses are performed, we applied ISMARA to an example RNA-seq data-set of expression profiles across 16 human cell types, i.e. data from the Illumina Body Map 2 (Geo Accession GSE30611) (IBM2). The results as obtained after submitting the raw RNA-seq data to ISMARA are available at http://ismara.unibas.ch/supp/dataset1_IBM/ismara_report/.

As described in the Methods, ISMARA infers the motif activities according to a linear model (Figure 4.1D) using a Bayesian procedure. Importantly, a Gaussian prior on motif activities is used to avoid over-fitting and the parameter of this prior is fit automatically by ISMARA for each input data-set using a cross-validation scheme. Motif activities are fitted from 80% of the promoters and the performance of the model, i.e. the fraction of the variance in E_{ps} explained by the model, is assessed on the remaining 20% of promoters.

Although our model fits E_{ps}, it is important to note that it is not the model's aim to provide an accurate fit of the signals E_{ps}. As discussed in the introduction, we do not expect the highly simplified linear model to provide an accurate fit to the signal E_{ps} at individual promoters. Indeed, the model explains 7.7% of the variance in E_{ps} for the IBM2 data, and across the data-sets studied here, we find that the model typically captures $5-15\%$ of the variance of E_{ps} across samples (Supplementary Figure B.2). Although these fractions are modest, given that tens of thousands of promoters are involved, they are extremely significant, i.e. using randomization of the association between site-count and expression we estimate the p-value for explaining 7.7%

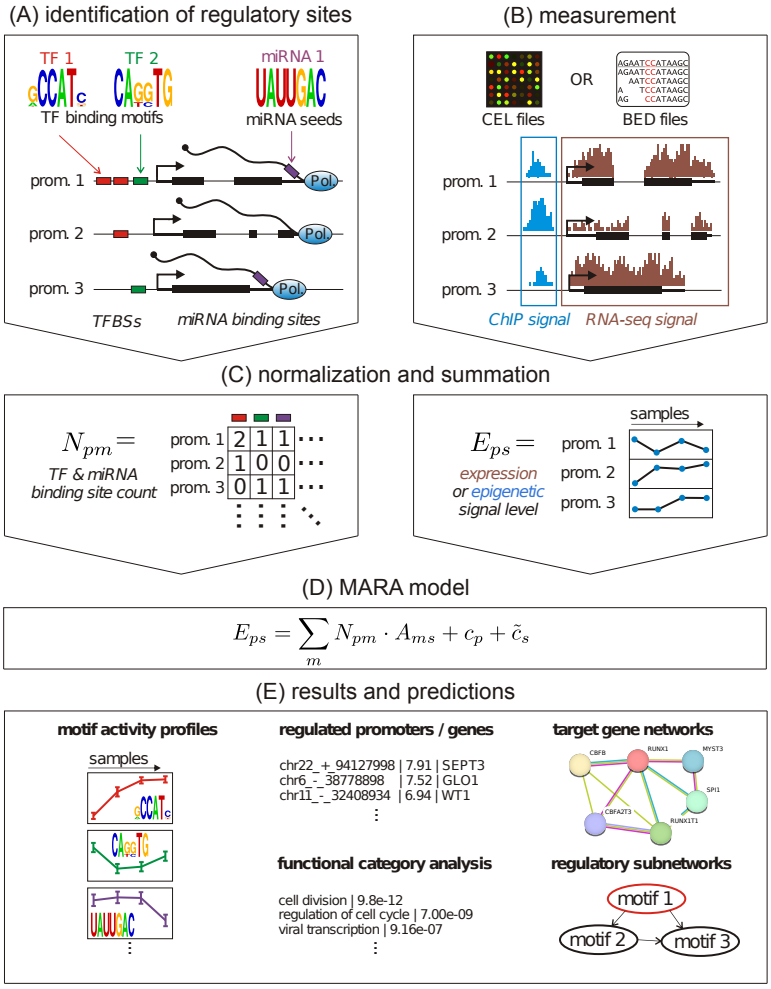


Figure 4.1: Outline of the Integrated System for Motif Activity Response Analysis.
(A) ISMARA starts from a curated genome-wide collection of promoters and their associated transcripts. Using a comparative genomic Bayesian methodology [290], transcription factor binding sites (TFBSs) for ≈ 200 regulatory motifs are predicted in proximal promoters. Similarly, miRNA target sites for ≈ 100 seed families are annotated in the 3' UTRs of transcripts associated with each promoter [143]. **(B)** Users provide measurements of gene expression (micro-array, RNA-seq) or chromatin state (ChIP-seq). The raw data are processed automatically and a signal is calculated for each promoter in each sample. For ChIP-seq data, the signal is calculated from the read density in a region around the transcription start. For gene expression data, the signal is calculated from read densities across the associated transcripts (RNA-seq) or intensities of associated probes (micro-array). **(C)** The site predictions and measured signals are summarized in two large matrices. The components N_{pm} of matrix N contain the total number of sites for motif m (TF or miRNA) associated with promoter p. The components E_{ps} of matrix E contain the signal associated with promoter p in sample s. **(D)** The linear MARA model is used to explain the signal levels E_{ps} in terms of bindings sites N_{pm} and unknown motif activities A_{ms}, which are inferred by the model. The constants c_p and $\tilde{c}_s$ correspond to basal levels for each promoter and sample, respectively. **(E)** As output, ISMARA provides the inferred motif activity profiles A_{ms} of all motifs across the samples s, sorted by the significance of the motifs. A sorted list of all predicted target promoters is provided for each motif, together with the network of known interactions between these targets (provided by the String database, http://string-db.org), and a list of Gene Ontology categories that are enriched among the predicted targets. Finally, for each motif, a local network of predicted direct regulatory interactions with other regulators is provided.

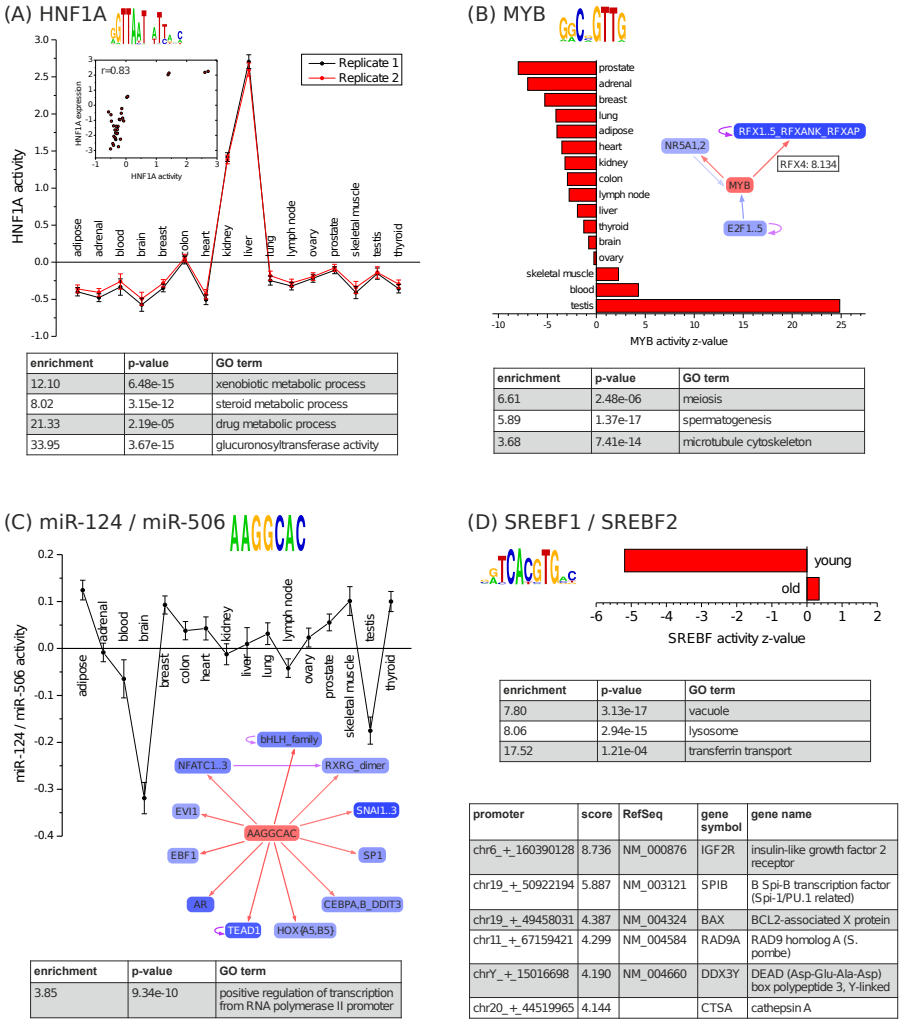
of the variance by chance is approximately 10^{-235} (Supplementary Methods and Supplementary Figure B.3).

ISMARA's main aim is to identify which regulatory motifs m play an important role, and how these motifs contribute to E_{ps} across the samples. ISMARA's output first of all lists all regulatory motifs sorted by a z-score which summarizes the importance of the motif for explaining the expression variation across the samples. This score roughly corresponds to the average number of standard-deviations the motif activity is away from zero (see Methods and Supplementary Methods). Besides the z-score of each motif, the list also displays the set of TFs or miRNAs that bind to sites of the motif, a thumbnail of its activity across the input samples, and a sequence logo for each motif (Supplementary Figure B.4). Following the link from the motif name leads to a page with a large number of predictions regarding the motif's precise regulatory role. To illustrate these, Figure 4.2 shows some of ISMARA's results for the HNF1A, MYB, hsa-miR-124/hsa-miR-506, and the SREBF motifs.

HNF1A was the most significant motif for the IBM2 data-set and its predicted activity is highly tissue-specific, being almost entirely restricted to liver and kidney (Figure 4.2A, and Supplementary Figures B.5 and B.6). The associated transcription factor hepatocyte nuclear factor 1 homeobox A (HNF1A) is relatively well-studied and indeed known to be mainly expressed in liver, kidney, stomach and intestine [307, 308], where it is essential for organ function [309]. Figure 4.2A also illustrates that the inferred motif activities are highly reproducible. In fact, motif activities are more reproducible than the expression profiles from which the motif activities were inferred (Supplementary Figure B.16). The reason for this high reproducibility of motif activities is that each motif m typically targets hundreds to thousands of promoters and that the inferred motif activities A_{ms} are statistical averages of the behaviors of a large number of promoters. This averaging causes the complexities at individual promoters to effectively cancel out and ensures that the overall influence of a motif can still be reliably inferred.

For many of the regulatory motifs there are multiple TFs that can bind to the sites of the motif and it is not *a priori* clear which of the TFs is most responsible for the motif activity in a given system. ISMARA infers motif activities from the behavior of the predicted *targets* of the motif. That is, roughly speaking, an increased activity is inferred when its targets show on average an increase in expression, that cannot be explained by the presence of other motifs in their promoters. The *mRNA expression profiles* of the TFs associated with a motif thus provide independent information about the link between the TFs and the motif activities, and ISMARA provides an analysis of the correlation between motif activities and the expression profiles of the associated TFs. For HNF1A, there is a good correlation between mRNA expression of the TF and the inferred motif activity (Figure 4.2A inset). However, for the fourth most significant motif (POU2F), only one of the 3 POU2F factors, POU2F2 (also known as OCT2), shows significant correlation of its mRNA level with motif activity, and it is the most highly expressed. This suggests that POU2F2 is mainly responsible for the motif activity in these tissues

(Supplementary Figures B.7 and B.8). The fact that the correlation is positive also strongly suggests that POU2F2 acts as an activator. In contrast, whenever a negative correlation between motif activity and TF expression is observed, the TF most likely acts as a repressor, e.g. as observed for the known repressor ZHX2 [310] (Supplementary Figure B.9). However, it should be noted that motif activity does not need to be a direct function of TF expression, i.e. the effect of a TF on its targets will not only depend on its expression, but possibly on post-translational modifications, on cellular localization, and on the presence of specific co-factors. Therefore, although a strong correlation between TF expression and motif activity is a good indication that the TF is responsible for the motif activity, the absence of such a correlation does not imply that the TF is not involved in the motif's activity.



enrichment	p-value	GO term
12.10	6.48e-15	xenobiotic metabolic process
8.02	3.15e-12	steroid metabolic process
21.33	2.19e-05	drug metabolic process
33.95	3.67e-15	glucuronosyltransferase activity

enrichment	p-value	GO term
6.61	2.48e-06	meiosis
5.89	1.37e-17	spermatogenesis
3.68	7.41e-14	microtubule cytoskeleton

enrichment	p-value	GO term
3.85	9.34e-10	positive regulation of transcription from RNA polymerase II promoter

enrichment	p-value	GO term
7.80	3.13e-17	vacuole
8.06	2.94e-15	lysosome
17.52	1.21e-04	transferrin transport

promoter	score	RefSeq	gene symbol	gene name
chr6_+_160390128	8.736	NM_000876	IGF2R	insulin-like growth factor 2 receptor
chr19_+_50922194	5.887	NM_003121	SPIB	B Spi-B transcription factor (Spi-1/PU.1 related)
chr19_+_49458031	4.387	NM_004324	BAX	BCL2-associated X protein
chr11_+_67159421	4.299	NM_004584	RAD9A	RAD9 homolog A (S. pombe)
chrY_+_15016698	4.190	NM_004660	DDX3Y	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked
chr20_+_44519965	4.144		CTSA	cathepsin A

Figure 4.2: Results for the Illumina Body Map 2. Each panel corresponds to a motif (indicated with name and sequence logo) and shows the inferred motif activities across the 16 tissues (activities with error-bars in panels A and C, and activity z-values in panels B and D). Tables show Gene Ontology categories enriched among predicted targets of each motif, and individual target promoters (panel D). The networks (panels B and C) show direct regulatory interactions between the motif and other regulators. **(A)** Red and black curves correspond to motif activities from two replicate measurements. The inset shows the correlation between motif activity and HNF1A mRNA levels. **(B)** The inset shows that MYB is predicted to directly target the RFX4 promoter with target score 8.134. **(C)** The regulatory network inset and GO table show that hsa-miR-124/hsa-miR-506 is predicted to directly target many TFs. **(D)** The red bars show z-values of the average motif activity of the SREBF motif for samples coming from older (age 58-86) and younger (age 19-47) donors.

ISMARA predicts individual target promoters p for each motif m by calculating the difference S_{pm} of the log-likelihood of the model with the original site-count matrix N and the log-likelihood of the model in which only the binding sites for motif m in promoter p have been removed (Methods and Supplementary Methods). For each motif, a searchable and resizable list is provided of all target promoters, their associated transcripts, and associated genes (Supplementary Figure B.10). For HNF1A, the accuracy of ISMARA's

target predictions is suggested by the fact that most of the top predicted targets are supported by the literature, including some of the oldest known direct targets of HNF1A [311]. For each target promoter, ISMARA provides a link to the genome browser view of the promoter (Supplementary Figure B.11), showing the precise genomic location of the predicted regulatory site. To provide the user with a more intuitive picture of the predicted list of targets of the motif, a link is provided to a network view of the target genes as provided by the STRING database [312], where network links indicate known functional associations between the genes. For HNF1A, the STRING network reveals a large, highly connected cluster of predicted targets that are known to be involved in the metabolism of drugs and toxins in the liver (Supplementary Figure B.12). As another means to provide insights into the pathways targeted by a given motif, ISMARA also provides lists of enriched Gene Ontology categories [313] (Figure 4.2 and Supplementary Figure B.13), which in this case confirms that HNF1A targets genes involved in the metabolism of drugs and xenobiotics.

To gain insight in the transcription regulatory networks that control expression profiles, it is of particular interest to identify direct regulatory connections between the TFs themselves. In ISMARA, a direct regulatory interaction from motif m to m' is predicted when motif m is predicted to target a promoter of one of the TFs associated with m'. To visualize the predicted direct regulatory interactions between regulators, ISMARA provides, for each motif m, a local network picture that shows all predicted regulatory connections between m and promoters of TFs that are associated with other motifs (Supplementary Figure B.14). The user can interactively change the cut-off on the target score S_{pm} to draw this picture. For HNF1A we find that the strongest predicted targets are *HNF4A*, *FOXA2*, *NR5A2*, and *HNF1A* itself (Supplementary Figure B.14). In addition, HNF4A and FOXA2 are predicted to target the *HNF1A* promoter as well. Remarkably, all these predictions are supported by independent experimental evidence [314–319].

ISMARA predicts that the MYB motif is by far most active in testis, and that it targets genes are involved in meiosis and spermatogenesis (Figure 4.2B). In addition, the MYB motif is predicted to target the *RFX4*, *RFX2* and *NR5A1* promoters. A literature search reveals that MYBL1, a close homolog of MYB that binds to the same regulatory sites, is a master regulator of male meiosis and spermatogenesis [320, 321]. Moreover, *RFX2* has been implicated as a direct target of MYBL1 in spermatogenesis [322]. ISMARA's prediction that *RFX4* is also regulated by the MYB motif (presumably through MYBL1) is novel to our knowledge. Finally, ISMARA's prediction that the *MYB* promoter is targeted by the E2F motif is also supported by the literature [323].

To illustrate ISMARA's predictions of the regulatory role of miRNAs, Figure 4.2C shows results for the second most significant miRNA seed family, hsa-miR-124/hsa-miR-506. This seed family has strongest negative activity in brain and its targets are highly enriched for TFs (Figure 4.2C). Indeed, hsa-miR-124 is a well-known brain-specific miRNA [236]. Moreover, of the top 9 predicted TF target genes of hsa-miR-124, 6 (*TEAD1*, *CEBPA*, *AR*,

SP1, *SNAI2*, *NFATC1*) are supported by independent experimental evidence [106, 138, 324–326] again confirming the high accuracy of ISMARA's target predictions.

Of course, most of the results highlighted in Figure 4.2, such as the function of HNF1A in liver and the brain-specific role of hsa-miR-124 are well-known from the literature. However, all these results, including very specific predictions of the precise targets of each regulator, were obtained by a completely automated analysis of RNA-seq data from 16 human tissues, without any free parameters or specific processing of the data. Moreover, they constitute only a small selection of the predictions made by ISMARA.

By default ISMARA focuses on regulatory motifs that explain *changes* in expression levels across the input samples. However, some users may be interested in regulators that are predictive for a consistently high or consistently low expression level across all samples. To address this, ISMARA also fits the absolute expression levels of the promoters, i.e. averaged over all input samples, in terms of "mean activities" (Methods and Supplementary Methods). For the IBM2 data-set we find that the TFs YY1 and NRF1 are most predictive of high average expression, whereas the known repressors REST and RREB1 are most predictive for low average expression (Supplementary Figure B.15).

Experiments are often performed in multiple replicates and ISMARA implements procedures for specifically identifying motifs that behave reproducibly across the replicates. The ISMARA results page links to a section where users can provide batch and replicate annotation for their samples, which is then used by ISMARA to calculate motif activity profiles that are averaged over replicates using a rigorous Bayesian procedure (Supplementary Methods). In addition, updated motif z-scores quantify to what extent a motif's activity varies across samples in a way that is reproducible across the replicates (Supplementary Methods). As an example, the replicate-averaged results for the IBM2 data-set are available at http://ismara.unibas.ch/supp/dataset1_IBM/averaged_replicates/averaged_report/.

Apart from replicate-averaging, this procedure can further be used to calculate contrasts between subsets of samples. To illustrate this, we noted that the samples of the 16 tissues of the IBM2 data-set derived from donors of different ages, and we investigated whether any motifs have consistently different activities between samples from older and younger individuals. We divided the samples into those deriving from donors aged 19-47 and those deriving from donors aged 58-86. We then directed ISMARA to calculate averaged activities for "young" and "old" samples for each motif (results at http://ismara.unibas.ch/supp/dataset1_IBM/averaged_age/averaged_report/). We found that only the SREBF motif is significantly differently regulated between old and young samples (Figure 4.2D). The targets of SREBF are up-regulated in older tissues relative to the younger ones and are highly enriched for lysosomal genes. Lysosomes are responsible for degradation of many macromolecules including proteins and increase in lysosomal mass is a well-known characteristic of aging and senescence in cells [327, 328]. In

addition, evidence is increasing that a progressive decrease in the efficiency of autophagy and lysosomes with age plays a key role in aging-associated degenerative changes in mammals [329]. Several recent findings support that SREBP TFs play a key role in these processes. SREBF1 expression increases with age in rat brains [330], SREBF1 mediated lipogenesis is involved in senescence [331], SREBF2 regulates autophagy [332], and SREBF activity is regulated by mTOR complex 1 [333]. It is remarkable that, simply by contrasting motif activities in tissues from younger and older donors, ISMARA was able to automatically identify SREBF as a key regulator of aging-related changes in expression of lysosomal genes.

As another example of the power of motif activity contrasts across sets of samples, we searched for motifs consistently disregulated in cancer by joint analysis of the human GNF atlas of 79 tissues and cell lines [334] and the NCI-60 reference cancer cell lines [335] (full results at http://ismara.unibas.ch/supp/dataset2/ismara_report/). Supplementary Tables B.2 and B.3 show the motifs that are most consistently up-regulated or down-regulated in tumors, including miRNAs. As discussed in the supplementary material, many of the top dis-regulated motifs, such as HIF1A and hsa-miR-205 miRNA (Supplementary Figure B.17), are well-known in cancer biology, again supporting the accuracy of ISMARA's predictions. Besides well-known oncogenes and tumor suppressors, ISMARA also makes several novel predictions of regulators consistently disregulated in cancers, including the TFs HAND1, KLF12, BPTF, FOXD3, and ZNF143.

4.3.2 *Inferring motif activity dynamics: inflammatory response*

To illustrate ISMARA's analysis of time series data, we applied it to a time series of expression data obtained after activation of human umbilical vein endothelial cells (HUVECs) with tumor necrosis factor (TNF, also known as TNFα). Messenger RNA expression was measured every 15 minutes for the first 4 hours after treatment, and every 30 minutes for the next 4 hours [336]. Whereas the original study focused solely on nascent transcription, we here show that standard application of ISMARA to this data set (http://ismara.unibas.ch/supp/dataset3/ismara_report/) uncovers the transcription regulatory network involved in this inflammatory response in remarkable detail.

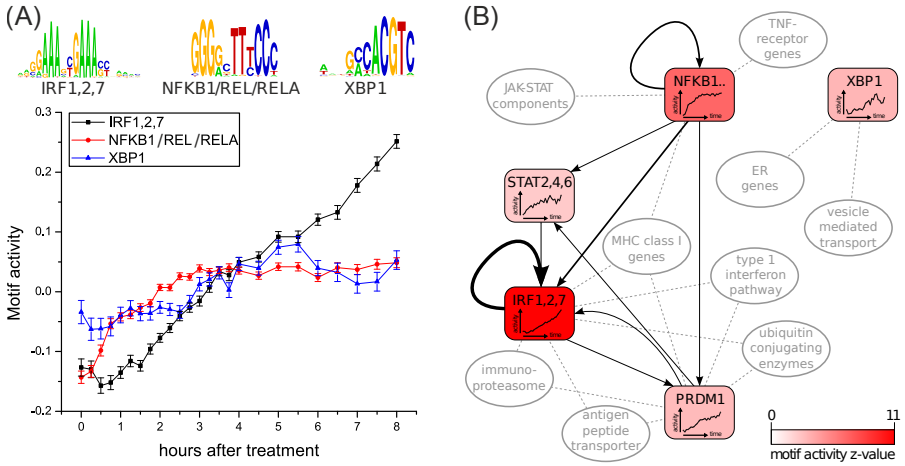


Figure 4.3: Analysis of an inflammatory response time series of human umbilical vein endothelial cells responding to TNF. (A) Time-dependent activities of the 3 most significant motifs, i.e. NFKB1/REL/RELA (red), IRF1/2/7 (black), and XBP1 (blue). Error-bars denote standard-deviations of the inferred activities. **(B)** Summary of the inferred core regulatory network. Selected top motifs are shown together with interactions between them and pathways/functional categories that are enriched among the targets of these motifs. The intensity of the color corresponds to the z-score of the motif, its time-dependent activity is indicated inside the node, and the thickness of each edge corresponds to its target score S_{pm}.

The response of endothelial cells to TNF is known to be mediated by TFs of the NFκB family, GATA2, IRF1, and JUN [337] TFs. TFs of the NFκB family in particular are crucial for the resulting inflammatory response [338]. Indeed, ISMARA infers that the two most significant motifs are IRF1,2,7 and NFKB1/REL/RELA. The activity of NFKB1/REL/RELA increases sharply in the first 45 minutes and slower afterwards, until it reaches a steady activity after 3 hours. The activity of the IRF1,2,7 motif increases steadily starting at 30 to 45 minutes after treatment until the end of the time course (Figure 4.3A). As shown by NFKB1/REL/RELA's local network figure (Figure 4.3B and on the ISMARA results website), ISMARA predicts that *IRF1* is activated directly at the level of transcription by these regulators, which is confirmed by the experimental literature [339]. Other predicted targets of NFKB1/REL/RELA that are also significantly upregulated in this process are TNF receptor genes, components of the JAK-STAT pathway (note that STAT2,4,6 is the 11th most significant motif, indicating that STAT activity changes, affecting the level of *its* targets) and MHC class I genes. The latter are also predicted to be regulated by IRF1,2,7, which is confirmed by experimental data [340]. ISMARA makes the novel predictions that both NFKB1/REL/RELA and IRF1,2,7 activate the 5th most significant motif, PRDM1, which is an important developmental regulator in the B-cell and T-cell lineages and is required for the secretory pathway in B-cells [341]. PRDM1 activity increases, like that of IRF, across the entire time course, and these two regulators appear to share many of their predicted targets, including type 1 interferon pathway genes, the immuno-proteasome [342], ubiquitin conjugating enzymes, antigen peptide transporters, and MHC class I genes.

These targets suggest that the IRF and PRDM1 TFs may be responsible for activation of the antigen presenting pathway.

We note that, although our TFBS predictions incorporate cross-species conservation analysis, this does not mean that the predicted targets must be conserved across mammals. For example, the third most significant TF target of the IRF motif is the ATF5 promoter, which is targeted through a TFBS that is primate-specific (Supplementary Figure B.18).

To provide an example assessment of the accuracy of ISMARA's genome-wide target predictions, we compared the predicted targets of NFKB1/REL/RELA with targets identified through ChIP-seq in lymphoblastoid cell lines derived from 10 individuals of African, European, and Asian ancestry [343]. We find that almost two-thirds of the top 50 targets, more than 50% of the top 150 targets, and about 40% of the top 300 targets are supported by ChIP-seq binding at the promoter (Supplementary Figure B.19). To put these numbers in perspective, we compared the validation of ISMARA's targets with the variability in NFKB1/REL/RELA binding across individuals and replicate samples. We used the ChIP-seq data from each sample to predict target promoters, and then 'validated' these 'predictions' using the other ChIP-seq data-sets in complete analogy to the way we validated ISMARA's targets. The typical validation rate for the ChIP-seq data was higher than for the ISMARA target predictions, i.e. $60-70\%$ versus $40-66\%$. This is not surprising given that all ChIP-seq data were obtained in the same lymphoblastoid cell type, which differs from the HUVEC cells. Still, we found significant variability across the ChIP-seq data-sets, and the targets from some ChIP-seq data-sets had lower intersection with the other ChIP-seq data-sets than ISMARA's targets (Supplementary Figure B.19). This analysis shows that ISMARA's genome-wide predictions can reach accuracies comparable to those obtained from a ChIP-seq study.

Finally, the 3rd most significant motif is XBP1, which is activated only after 2.5 hours. Its predicted targets are highly over-represented for endoplasmic reticulum (ER) genes and genes involved in vesicle-mediated and Golgi transport, consistent with the fact that XBP1 is a major regulator of ER stress and the unfolded protein response (UPR) [344]. Moreover, several studies support that the UPR is a general characteristic resulting from inflammation or TNF activation in endothelial cells [345, 346]. Interestingly, the induction of XBP1's activity occurs at the same time as the NFKB1/REL/RELA activity stops increasing which is in line with studies showing that the UPR can attenuate the induction of inflammation as mediated by TFs of the NFκB family [347–349]. The induction of XBP1's activity is not reflected in the expression of XPB1 itself, which is almost constant across the time course (Supplementary Figure B.20). This underscores that ISMARA infers a motif's activity from the expression of its predicted targets and does not use the regulator's own expression. Indeed, it has been established that XBP1 activity is regulated post-transcriptionally through alternative splicing [350, 351]. Together these results demonstrate that ISMARA reconstructs the core regulatory cir-

cuitry of the innate immune response in HUVEC cells (Figure 4.3B) *ab initio* using only time course expression data.

4.3.3 *Identifying novel master regulators: Mucociliary differentiation of bronchial epithelial cells*

Next, we turned to an example system for which much less is known, namely the mucociliary differentiation of bronchial epithelial cells on an air-liquid interface. Aiming to elucidate the regulation of bronchial development, Ross et al. [352] performed differentiation experiments in triplicate over a period of 28 days with cells from three separate donors. This data was then analyzed with commonly used bioinformatic procedures, i.e. genes were clustered into co-expression clusters and the clusters were analyzed for over-represented gene ontology categories and pathways. This analysis uncovered clusters associated with TGF-beta pathway genes, extra-cellular adhesion genes, and genes associated with the microtubule cytoskeleton, but no key regulators or regulatory interactions that drive these expression changes were identified.
In contrast, applying ISMARA to this gene expression data set, we obtain the prediction that by far the most important regulatory motif in this system is RFX, whose activity is strongly increasing over the period from roughly day 4 to day 10 in all 3 donors (Figure 4.4A, http://ismara.unibas.ch/supp/dataset4/ismara_report/). The predicted targets of RFX are highly enriched in genes known to be associated with cilium assembly, axoneme, and the microtubule cytoskeleton genes (Figure 4.4B) suggesting that RFX directs ciliogenesis in bronchial epithelial cells.

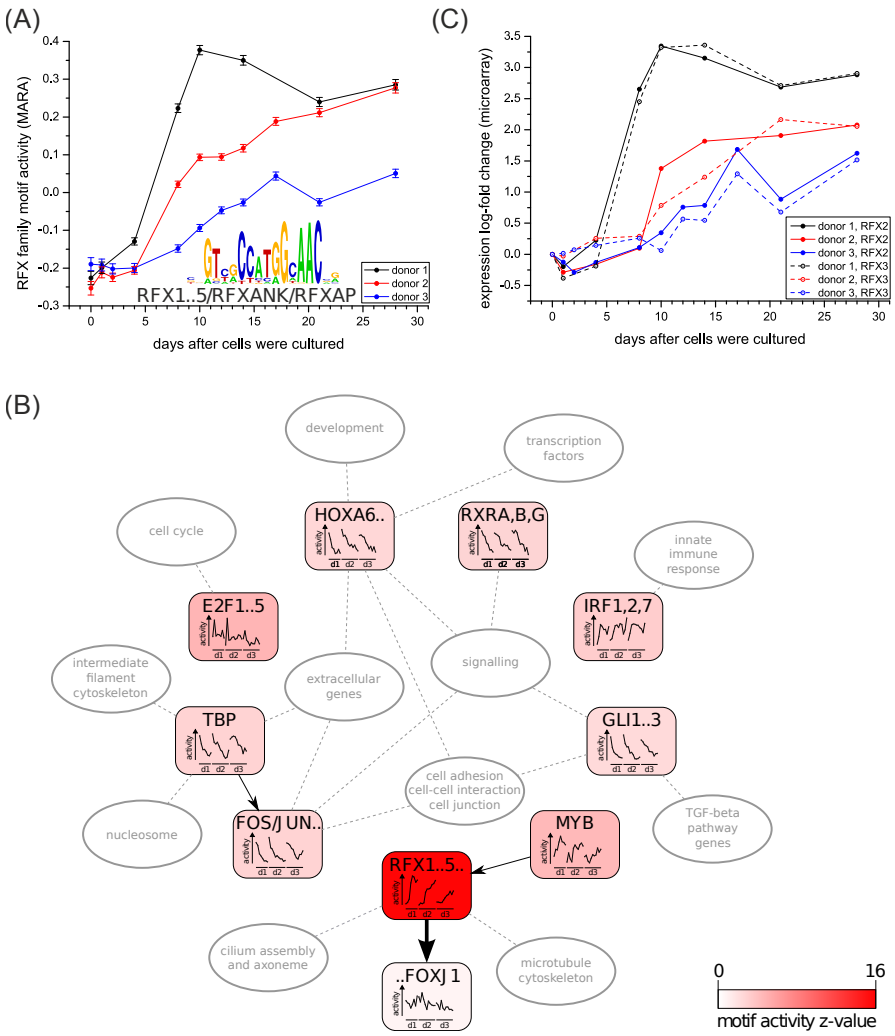


Figure 4.4: Mucociliary differentiation. (A) Inferred RFX motif activity profile in mucociliary differentiation of bronchial epithelial cells from three independent donors (black, red, and blue lines). **(B)** Key predicted regulators and their targets in this system. Selected top motifs are shown together with predicted interactions between them and pathways/functional categories that are enriched among predicted targets of these motifs. The intensity of the color corresponds to the z-score of the motif, its time-dependent activity for each donor is indicated inside the node, and thickness of the edges corresponds to the target score S_{pm}. **(C)** mRNA expression profiles of the *RFX2* (solid) and *RFX3* (dashed) genes across the differentiation (colors of the donors as in panel (A)).

The RFX family of TFs contains 7 members and it is not *a priori* clear which of these are driving the bronchial differentiation. Comparison of the mRNA expression profiles with activity profiles shows that two of the family members, RFX2 and RFX3 exhibit a striking correlation in their expression with the motif activity (Figure 4.4A and C). Together these results strongly suggest that the TFs RFX2/3 are master regulators of ciliogenesis in this system. This prediction is consistent with previous studies that have shown that Rfx3 is necessary for the ciliogenesis of nodal cilia in mouse embryonic develop-

ment [353] and during ciliogenesis of motile cilia in a mouse cell-culture system [354].
Strikingly, ISMARA's results on the IBM2 data-set also identified the RFX motif as the key regulator of ciliogenesis in spermatogenesis. As discussed above, in that system ISMARA predicted that the *RFX2* and *RFX4* promoters were directly targeted by the MYB motif (most likely through the MYBL1 TF). We here find that ISMARA predicts MYB to target the *RFX2* promoter in the mucociliary differentiation system as well (Figure 4.4B). In addition, ISMARA's prediction that RFX directly upregulates *FOXJ1* in this system was also made in the results on the IBM2 data-set. Indeed, RFX3 was found to activate *FOXJ1* during ciliogenesis in the mouse cell-culture system mentioned above [354]. These observations suggest that the core regulatory network involved in ciliogenesis, with MYBL1 targeting RFX promoters and RFX TFs targeting *FOXJ1*, is conserved across multiple mammalian systems.
As indicated in Figure 4.4B, ISMARA additionally predicts that, in this system, IRF1,2,7 upregulates innate immune response genes, and that a short spike of E2F activity up-regulates cell-cycle genes at day one. Finally, there is a group of motifs (TBP, FOS_FOS{B,L1}_JUN{B,D}, RXR{A,B,G}, HOX{A6,A7,B6,B7}, and GLI1..3) whose targets are progressively down-regulated across the differentiation time course. The targets of these motifs are generally enriched for extracellular proteins involved in cell adhesion, cell-cell junctions, and signaling. More specifically, targets of GLI1..3 involve genes from the TGF-beta pathway, targets of TBP involve nucleosomal and intermediate filament cytoskeletal genes, and targets of the homeodomain motif (HOX{A6,A7,B6,B7}) are enriched for developmental genes and transcription factors. The genes in these pathways are most likely involved in the transition of the tissue from squamous to columnar epithelial that occurs during differentiation. Thus, in contrast to the methods used in the original study [352], ISMARA predicts which regulators are directing various aspects of the differentiation process, including ciliogenesis, the innate immune response, and the transition from squamous to stratified epithelial. As far as we are aware, these predictions of the core regulatory network controlling mucociliary-differentiation are all novel.

4.3.4 *Interactions between TFs and miRNAs: epithelial-mesenchyme transition*

To illustrate ISMARA's ability to integrate the role of both TFs and miRNAs in the gene regulatory network, we took advantage of data from a system in which miRNAs are known to play important regulatory roles: the epithelial-to-mesenchymal transition (EMT). We applied ISMARA to expression measurements from epithelial and mesenchymal subpopulations [355] (results at http://ismara.unibas.ch/supp/dataset5/ismara_report/) and used replicate-averaging to identify regulators that explain the differences between epithelial and mesenchymal cells (results at http://ismara.unibas.ch/supp/dataset5/averaged_report/). As discussed in the Supplementary Materials and Supple-

mentary Figure B.21, ISMARA automatically inferred much of the key regulatory interactions between TFs and miRNAs involved in EMT (reviewed by Polyak and Weinberg [356]) using only the gene expression data.

4.3.5 *TF activities affecting chromatin state: analysis of ChIP-seq data*

Beyond analyzing gene expression data, motif activity response analysis can be applied to modeling any signal along the genome in terms of the local occurrence of TFBSs. Indeed, in recent work [207] we applied the MARA approach to ChIP-seq data mapping the dynamics of tri-methylation at lysine 27 of histone 3 (H3K27me3) and identified TFs involved in recruiting this epigenetic mark that is set by the Polycomb system. In ISMARA the analysis of ChIP-seq data has now been completely automated. In particular, given a ChIP-seq data set, ISMARA quantifies the signal at all promoters across all samples and models this in terms of the TFBSs at each promoter. For the details of ISMARA's processing and normalization of the ChIP-seq data we refer to the Methods and Supplementary Methods. Note that, like for the transcriptomic data, ISMARA thus by default focuses on the variation in ChIP-seq signals at *promoters* only. However, the approach can easily be applied genome-wide and to allow expert users to apply MARA to any collection of genomic regions the ISMARA website includes an 'expert mode' that allows users to upload their own signal and site-count matrices and apply MARA with these matrices.

To illustrate ISMARA's results on ChIP-seq data, we make use of data from the ENCODE Project in which, besides gene expression, 9 different chromatin marks were measured across 8 different cell types [357] (all modifications and cell types are listed in Supplementary Tables B.4 and B.5). We first ran ISMARA separately on each of the 10 data sets, i.e. expression and 9 chromatin modifications (see Supplementary Table B.6 for the URLs of the results on all data sets). We observed that motifs that are highly significant for explaining differences in levels of a particular chromatin mark across tissues, were often also highly significant for explaining *mRNA expression* differences. This was particularly the case for methylation of lysine 4 on histone H3 (H3K4me2, H3K4me3), for acetylation of histone H3 (H3K9ac, H3K27ac), and for tri-methylation of lysine 36 on histone H3 (H3K36me3). For example, Figure 4.5A shows the activity profiles for these marks for the SNAI1..3 motif, which is recognized by the Snail TFs (see Supplementary Figure B.22 for additional examples). As is clear from these figures, for these motifs the activity profile for expression is highly similar to those of all of these histone marks. Indeed, this reflects that these chromatin marks are associated with promoter activity [11], and several recent studies have shown that the levels of these marks can be used to predict gene expression levels [12–14].

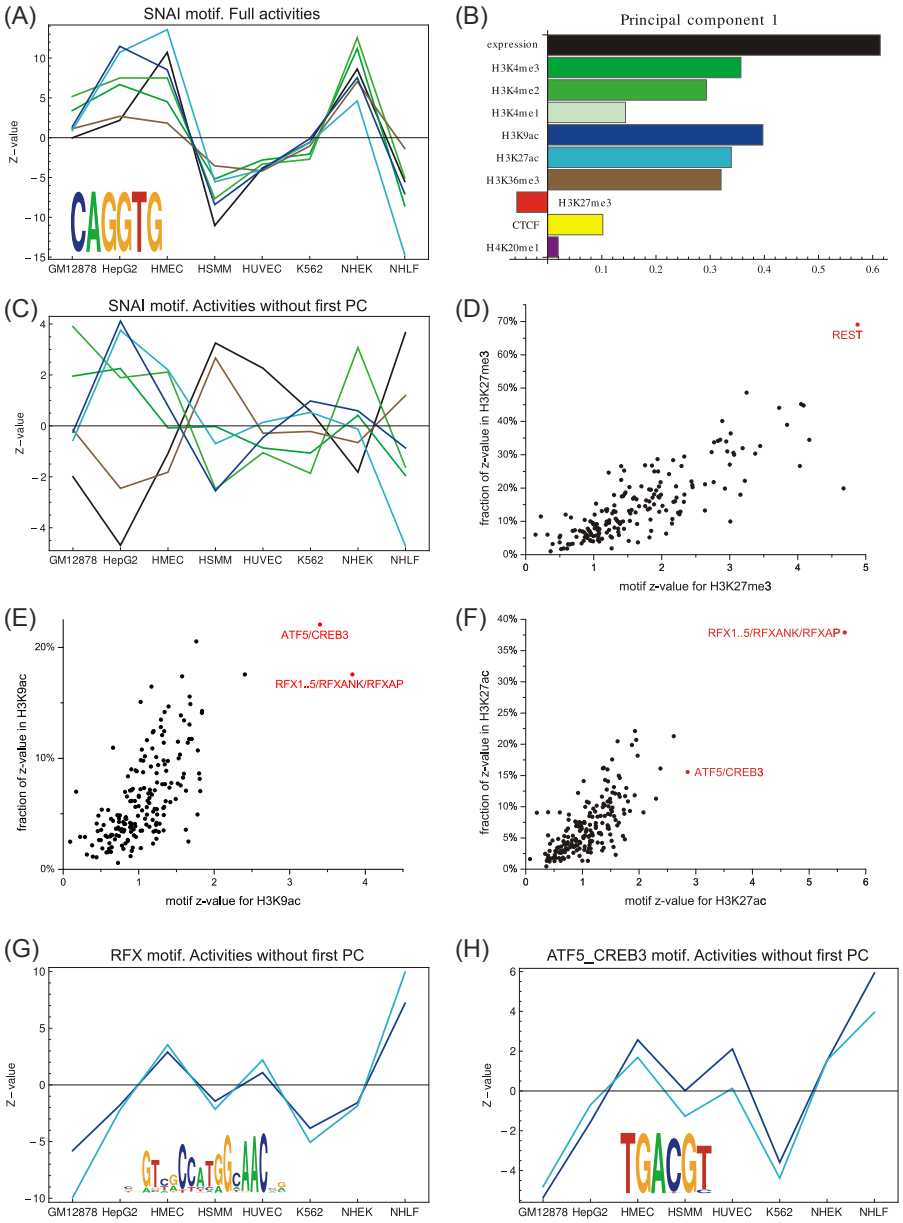


Figure 4.5: ISMARA predicts TFs involved in recruiting specific chromatin marks. **(A)** Activity across cell types of the SNAI1..3 motif in explaining expression (black), and levels of the chromatin marks H3K4me3 (dark green), H3K4me2 (light green), H3K9ac (dark blue), H3K27ac (light blue), and H3K36me3 (brown). **(B)** First principal component explaining the majority of variation in chromatin mark levels across all cell types. The bars indicate the relative contributions to the principal component of each mark. **(C)** Motif activities of the SNAI1..3 motif, as in panel A, but after removal of the first principal component. **(D)** z-values and specificities (see text) of motifs for explaining H3K27me3 levels. The REST motif, with both highest z-value and highest specificity, is indicated in red. **(E)** As in panel D, for H3K9ac levels. The two most significant motifs are shown in red. **(F)** As in panels D and E, for H3K27ac levels. **(G)** Activity, after removal of the first principal component, of the RFX motif for explaining H3K9ac (dark blue) and H3K27ac (light blue) levels. **(H)** As in panel G, for the ATF5_CREB motif.

To investigate the correlations between the levels of the different chromatin marks more quantitatively, we performed principal component analysis (PCA) of the levels of the 10 different marks across all promoters, separately for each sample (Supplementary Methods). Strikingly, we find that in each sample the first PCA component explains the majority of the variance across promoters, typically explaining around 60% of the total variance (Supplementary Figure B.23). Moreover, we find that the first PCA component looks virtual identical for each sample (Supplementary Figure B.23) and Figure 4.5B shows the first principal component obtained using PCA on the pooled data from all cell types. The first principal vector has its highest positive component along the expression axis, and the activation-associated marks H3K4me3, H3K4me2, H3K9ac, H3K27ac, and H3K36me3, also all have a strong positive component in this vector, whereas the known repressive mark H3K27me3 has a negative component. These findings strongly suggest that variation along the first principal vector corresponds roughly to variation in "promoter activity". In addition, the fact that this first principal vector is identical in all tissues suggests that the relative levels of the different marks in this first principal vector result not from tissue-specific but from general factors, e.g. conceivably they may result from the general transcription machinery recruiting chromatin modifying enzymes.

Because the variation in promoter activity captures almost two-thirds of the variation in all 10 measured levels at the promoter, any motif explaining variation in expression will also appear to explain variation in all chromatin marks associated with promoter activity, and confounds identification of TFs that are involved in affecting specific marks. To address this, for each motif we discarded the part of its activity profile along the first PCA component, retaining only variation in motif activities orthogonal to promoter activity. As illustrated in Figure 4.5C and Supplementary Figure B.22, after removal of the first principal component, there are no longer any obvious correlations in the remaining motif activity profiles for different activating marks.

We next analyzed the remaining motif activities and calculated, for each motif and each mark, a z-value quantifying the motif's contribution to explaining the mark's levels and also a "specificity" that measures the fraction of a motif's overall significance that is associated with a given mark (Supplementary Methods). Strikingly, we find that for many of the marks, the motifs that most significantly affect the mark are also among the most specific for that mark. For example, REST is the motif with the highest z-value for H3K27me3 levels, and is also by far most specific for H3K27me3 (Figure 4.5D). Indeed, in recent work [207] we showed that REST is involved in recruiting this mark during the differentiation of murine embryonic stem cells into pyramidal neurons, specifically at the neural progenitor state. With respect to the two acetylation marks, i.e. H3K9ac and H3K27ac, we find that the same two motifs, i.e. RFX and ATF/CREB, are most significant for both these marks (Figure 4.5E and F). It is well known that ATF/CREB TFs can recruit histone acetylases (HATs) such as CREB binding protein (CREBBP) and EP300 [358], and for RFX TFs it has also been established that they can recruit HATs at partic-

ular promoters [359]. Our results thus suggest that recruitment of HATs by TFs bound to ATF/CREB and RFX motifs make an important contribution to genome-wide histone acetylation at promoters. Moreover, the activity profiles of these motifs for H3K9ac and H3K27ac are highly similar, suggesting that these two marks may be recruited through a common or highly overlapping pathways. Supplementary Figure B.24 shows the most significant motifs for each of the other marks. Among the additional predictions made by ISMARA is that the PITX motif is associated with both mono- and di-methylation of lysine 4 of histone 3. This prediction is supported by recent biochemical evidence that PITX2 can recruit methyltransferases that methylate H3K4 [360]. As expected, CTCF is the most significant motif explaining CTCF binding. ISMARA also makes several predictions that are completely novel, as far as we have been able to determine: It predicts that the hepatocyte nuclear factors HNF1A and HNF4A have the most significant effect on the levels of the H3K36me3 mark, which is known to be set by elongating RNA polymerase [361, 362], and that YY1 and the NF-Y complex (consisting of NFYA, NFYB, and NFYC) most significantly explain variations in H4K20me1 levels.

4.4 DISCUSSION

The advent of high-throughput technologies now allows the routine measurement of genome-wide mRNA expression across conditions, and such data in principle provide the opportunity to systematically investigate gene regulation on a genome-wide scale across different models systems. However, a major bottle-neck in the field is that such investigations require sophisticated computational approaches that are not available to most experimental researchers. Here we have presented ISMARA, a completely automated system that enables any researcher to apply sophisticated computational modeling, on data from their system of interest, and obtain concrete predictions on the key regulators acting in their system, their activities, their genome-wide targets, and so on.

That the computational model at the core of ISMARA, i.e. motif activity response analysis, is a powerful method for reconstructing regulatory interactions from high-throughput data has already been demonstrated, not only in its original application [50], but in a substantial number of recent studies across a wide range of mammalian systems [207, 296–306]. In each of these studies, MARA successfully inferred key regulators and their regulatory interactions *ab initio*. The applications in this work not only further confirm that, in systems where key regulatory interactions are already known, ISMARA successfully infers them, but also provides a large collection of novel regulatory predictions across different systems in human and mouse, e.g. novel regulators that are disregulated in cancers, novel regulatory interactions in the inflammatory response, and the core regulatory circuitry involved in mucociliary differentiation and ciliogenesis. We believe that, by empowering experimental researchers to automatically apply this approach to their own data,

ISMARA can make a substantial contribution to the study of gene regulatory networks.

The applications we presented highlighted several of ISMARA's advantages. First, by inferring a regulator's activity from the behavior of its targets, ISMARA does not rely on changes in a TF's expression to infer activity changes, and readily detects activity changes due to alternative splicing, post-translation modifications, changes in cellular localization, etcetera. Second, when motif activity is transcriptionally regulated, comparing motif activity with TF expression allows ISMARA to identify the relevant TF(s), i.e. as illustrated by the identification of RFX2 and RFX3 as the key regulators of mucociliary differentiation. Such comparisons can also indicate whether a regulator acts as a repressor or an activator. An important goal of ISMARA is to provide predictions that are amenable to direct experimental follow-up. In this respect, the GO enrichment and STRING network analysis are typically very helpful in identifying the biological processes and pathways targeted by each motif, often suggesting potential markers for experimentally validating their predicted regulatory roles. Similarly, ISMARA's predictions of direct regulatory interactions between the key regulatory motifs provide concrete hypotheses regarding the regulatory circuitry that is acting in a given system, e.g. the predicted regulatory feedbacks between NFKB1/REL/RELA, IRF TFs, and PRDM1, or the prediction that MYBL1 is an upstream activator of RFX TFs in ciliogenesis. Moreover, the links to the individual binding sites on the genome [287] allow for targeted validation of such individual regulatory interactions. There are many indications that the actions of miRNAs and TFs are tightly integrated [145, 363, 364] and ISMARA's incorporation of miRNA regulation allows for the automated identification of regulatory interactions between TFs and miRNAs, as demonstrated by the analysis of the EMT data. Finally, gene expression regulation involves a tight interplay between the actions of TFs and changes in chromatin state. ISMARA's ability to not only model expression data, but any ChIP-seq signal at promoters genome-wide, allows for the identification of key TFs that are involved in dynamic regulation of chromatin state, as exemplified here by the analysis of ChIP-seq data from the ENCODE project which predicted, among other things, regulatory factors involved in recruiting histone acetylations.

There are of course several limitations to ISMARA's current approach which we aim to address in future work. First, using a simple linear model [365] has the advantage of being exactly solvable, but it ignores saturation effects that undoubtedly occur in reality. Second, the approach currently assumes that a given TF acts either mainly as an activator or mainly as a repressor, whereas it is clear that some TFs can act as an activator on some targets and as a repressor on others. Indeed, it has been recently shown [366] that allowing such dual function of TFs can significantly increase correlation between model predictions and measurement. Explicitly considering higher order constellations of TFBSs, e.g the occurrence of pairs or triplets of TFBSs for particular combinations of TFs, is another extension that we are currently evaluating. The regulatory motifs currently included in ISMARA represent approximately 350

of the roughly 1500 mammalian TFs. However, through developments in protein array technology [367] and the decreasing cost in ChIP-seq experiments, regulatory motifs for a rapidly increasing number of additional mammalian TFs have recently become available. We are currently working on curating a new, highly extended set of regulatory motifs, which we expect to incorporate into ISMARA in the near future.

Finally, ISMARA currently focuses solely on predicted TFBSs in proximal promoters, ignoring the effects of distal enhancers. In contrast to promoters, accurate genome-wide maps of enhancers have not been available until recently. However, the discovery that active enhancers exhibit characteristic chromatin modification patterns [368], DNA methylation patterns [369], and more generally DNA accessibility patterns [157], has now led to the first genome-wide mappings of enhancers in specific cell types [370]. If a set of relevant enhancers for a particular system of interest is available, it is in principle straight-forward to predict TFBSs in these enhancers and we are currently developing methodology for automatically incorporating the effects of TFBSs at distal enhancers into MARA. However, enhancers are highly cell-type specific and, in many cases, the data that users upload to ISMARA may come from systems for which no accurate mappings of distal enhancers are available. Therefore, automated incorporation of the effects of distal enhancers into ISMARA will only be possible when general methods for mapping active enhancers in any system have become available. Of course, the dynamics of chromatin accessibility and enhancer activity are themselves also controlled by constellations of regulatory sites on the genome, and our ultimately goal is to develop computational models that are able to predict genome-wide DNA accessibility and enhancer activity in terms of local constellations of regulatory sites.

4.5 METHODS

In this section we outline the methods that were used for automated processing and modeling of the data. More detailed descriptions of all procedures are provided in the supplementary methods.

4.5.1 *Promoteromes and regulatory site predictions*

For each model organism of interest (in this work we will focus exclusively on data from human and mouse) ISMARA relies on two pre-calculated resources: a genome-wide annotation of promoters, and a comprehensive collection of transcription factor binding site (TFBS) predictions in all promoters (Figure 4.1A, C). The genome-wide annotation of promoters in human and mouse, i.e. so-called “promoteromes”, were constructed primarily from deep sequencing data of transcription start sites (deepCAGE data [371]) using Bayesian methods that we described previously [293]. To infer expression levels of promoters from micro-array of RNA-seq data it is necessary to associate all promoters with the transcripts that they drive. We thus collected the

5' ends of all known mRNA mappings from the UCSC Genome Database, filtered these for mapping quality, and clustered all promoters and 5' ends that are within 150 base pairs (bps). In this way we obtained comprehensive sets of promoters and their associated transcripts for both human ($36'383$ promoters) and mouse ($34'050$ promoters). We also classified the promoters into CpG-island and non-CpG island promoters based on their CG and CpG content.

We next comprehensively predicted TFBSs in the proximal promoter regions of all promoters. Briefly, we curated a collection of 190 WMs representing $\approx$ 350 mammalian TFs using data from the JASPAR [285] and TRANSFAC [286] databases, additional motifs from the literature, and our own analysis of ChIP-chip and ChIP-seq data. For each promoter, we extracted 500 bps upstream and downstream of the TSS, and orthologous segments in 6 other mammals. The 7 orthologous sequences were then multiply aligned using T-Coffee [372]. Using the 190 regulatory motifs and a phylogenetic tree of the species (Supplementary Figure B.1) as input, we then applied our MotEvo algorithm [290] to predict functional TFBSs for all TF regulatory motifs across all promoters in human and mouse (Figure 4.1A,C). MotEvo is a Bayesian algorithm which considers all possible ways in which configurations of binding sites for all motifs, as well as additional conserved elements of unknown function, can be assigned to the input alignments, calculating likelihoods for all configurations using a rigorous model of the evolution of TFBSs and neutral sequence across the phylogeny. Since different motifs show different positioning preferences and abundances relative to TSS, which differ between CpG and non-CpG promoters, we also incorporated position-dependent prior probabilities for all motifs, separately for CpG and non-CpG promoters. We summarize the TFBS predictions in a matrix $\mathbf{N}$, where N_{pm} is the sum of the posterior probabilities of all predicted TFBSs for motif m in promoter p.

When modeling expression levels in terms of regulatory sites using a linear model, it is relatively straight-forward to extend the modeling to not only include effects of TFBSs but also the effects of miRNA regulation, e.g. as recently introduced in a supervised learning scheme for modeling regulation in glioblastomas [273]. In ISMARA the effects of miRNA regulation have been incorporated into a completely automated procedure that can be applied to any expression data-set. Specifically, we used miRNA target site predictions from TargetScan using preferential conservation scoring (P_{CT}) [143], which assigns target scores for 86 miRNA seed families to all RefSeq transcripts. To associate a target score N_{pm} for miRNA seed family m targeting promoter p, we average TargetScan's scores over all transcripts associated with promoter p.

4.5.2 *Processing of raw micro-array, ChIP-seq, and RNA-seq data*

To perform ISMARA analysis, the user only needs to upload raw micro-array (i.e. CEL files), RNA-seq, or ChIP-seq (BED or BAM files) data. The latter should contain the genomic mappings of the raw sequencing reads. The first

part of ISMARA's analysis consists of processing these raw data into a matrix **E**, where E_{ps} denotes the "signal" associated with promoter p for sample s.
When gene expression data is provided in the form of micro-arrays, ISMARA first automatically detects the particular type of micro-array used, and then applies corrections for background and unspecific binding tailored to that micro-array type. Micro-Array platforms currently supported by ISMARA are listed in Supplementary Table B.1. Using Gaussian mixture modeling, probes are classified into "expressed" and "non-expressed" for each sample. Probes that are consistently non-expressed are removed and the intensities of the remaining probes are quantile normalized. Instead of relying on annotation of the manufacturer we map all probe sequences to all transcripts associated with our promoters. The final log-expression of a given promoter is given by a weighted average of the log-intensities of all probes mapping to the transcripts associated with the promoter.
In many applications of next-generation sequencing data a main aim of the analysis is to detect genomic regions that are significantly enriched, or transcripts that are significantly differentially expressed, so that the analysis crucially depends on the noise statistics of sequencing data [293, 373]. In contrast, ISMARA aims to model the variation in "signal" E_{ps}, i.e. the amount of chromatin immuno-precipitation or the amount of expression, across promoters p and samples s in terms of predicted TFBSs. Our aim is thus not to assess the statistical significance of changes in the signal, but to estimate the relative strength of the signal across promoters and conditions. When processing ChIP-seq data, the signal E_{ps} is calculated as the estimated logarithm of the fraction of reads in sample s that map to a 2 kilobase region centered on promoter p. To avoid large fluctuations in E_{ps} at promoters with low signal due to sequencing noise, this estimate involves using a uniform prior distribution across the genome.
When processing RNA-seq data, the mapped reads are first mapped to our transcript set in a weighted manner. That is, when a read maps to n separate transcripts, each transcript's read-count is incremented by $1/n$. The expression of each transcript is then estimated by dividing its read count by transcript length, and the expression of a promoter is calculated by summing the expression of the transcripts associated with it. The final level E_{ps} is the logarithm of the estimated number of transcripts per million transcripts in the cells of sample s that derived from promoter p.

4.5.3 *Inference of motif activities*

At the core of ISMARA is the MARA model [50] which, similar to previous linear modeling approaches [365, 374], assumes that the "signal" at each promoter p is a linear function of its binding sites N_{pm}:

$$E_{ps} = \tilde{c}_s + c_p + \sum_m N_{pm} A_{ms} + \text{noise}, \tag{4.1}$$

where c_p is a term reflecting the average activity of promoter p across the samples, $\tilde{c}_s$ reflects the total expression in sample s, and the A_{ms} are the

(unknown) *activities* of each motif m in each sample s, which the model will infer. We set the constants $\tilde{c}_s$ and c_p to their maximum likelihood values.
As a result, equation (4.1) is renormalized into

$$E'_{ps} = \sum_m N'_{pm} A'_{ms} + \text{noise}, \tag{4.2}$$

where the matrix $\mathbf{E}'$ is obtained by subtracting the row and column averages from the entries of $\mathbf{E}$. Similarly, $\mathbf{N}'$ is obtained by subtracting the column averages, i.e. the average number of sites $\langle N_m \rangle$ for each motif m, from the entries of $\mathbf{N}$. Finally, the activities $\mathbf{A}'$ are obtained by subtracting the average motif activities A_m across the samples from the activities A_{ms}. That is, in equation (4.2) the expression *changes* across the samples and promoters are modeled in terms of changes in site counts across promoters and changes in motif activities across the samples.
As explained in the supplementary methods, the noise term in the above equation is dominated not by measurement or biological replicate noise, but by the *error* in the model and we assume these errors are Gaussian distributed with an unknown variance σ^2, that is integrated out of the likelihood. To infer the activities, ISMARA uses a Bayesian procedure which combines the Gaussian likelihood model for the difference between the measured signal E'_{ps} and the predicted signal, with a Gaussian prior distribution for the activities. This prior distribution, which *a priori* favors small activities is used to avoid overfitting. Its parameter is estimated automatically using 80/20 cross-validation: The activities are inferred on a randomly chosen 'training set' of 80% of the promoters, and the prior's parameter is set so as to maximize the fit of the predicted expression profiles on the 'test set' consisting of the remaining 20% of the promoters. In this way, ISMARA automatically adapts its prior to each data-set that is submitted.
The final posterior distribution of motif activities is a multi-variate Gaussian which is determined using singular value decomposition (see Supplementary Methods). By projecting the multi-variate Gaussian onto individual motifs ISMARA also calculates standard-deviations $\delta A'_{ms}$ on all motif activities. Finally, the overall significance of each motif m in explaining variations in E'_{ps} is summarized by a z-like statistic,

$$z_m = \sqrt{\frac{1}{S} \sum_{s=1}^{S} \left(\frac{A'_{ms}}{\delta A'_{ms}} \right)^2}, \tag{4.3}$$

where S is the number of samples. The z-scores calculate how many standard-deviations away from zero on average the inferred motif activities are.
Popular alternatives to a Gaussian prior include Laplacian priors, also referred to as Lasso regularization [375], or a product of Gaussian and Laplacian priors, also referred to as elastic net regularization [376]. These priors are often considered attractive because they induce sparsity, i.e. a subset of the fitted parameters will be set strictly zero. However, since ISMARA by default sorts motifs by their significance z_m, motifs with weak activities move to the bottom of the list, where they will be ignored by most users. Moreover, in some

cases a user might be interested in the inferred activity of a particular motif, even if its significance is weak, and the Gaussian prior ensures that a nonzero motif activity profile is inferred for every motif.

Although users will typically be primarily interested in motif activity changes that explain expression changes across the conditions, in some situations it would also be interesting to fit the *average* expression $\langle E_p \rangle$ of each promoter, i.e. averaged across all samples, in terms of average motif activities A_m. ISMARA fits such average activities using the same procedure, using a separate prior for the average motif activities A_m, and fitting this prior separately using cross-validation.

4.5.4 *Target predictions*

ISMARA also predicts which individual promoters are regulated by each motif m. As detailed in the Supplementary Methods, for each promoter with predicted TFBSs for the motif (i.e. $N_{pm} > 0$) ISMARA estimates the log-likelihood ratio S_{pm} of the entire model with the TFBSs for m in p present, and the model in which the entry N_{pm} has been set to zero. That is, we *in silico* mutate the promoter p such that its TFBSs for motif m are removed, and then recalculate the probability of the data $\mathbf{E}$ with this mutated site-count matrix, integrating over all unknown activities. Thus, S_{pm} rigorously quantifies how much removal of the sites for m in p decreases the fit of the model to the data.

Finally, enrichment of targets within particular Gene Ontology categories is done by selecting all targets where inclusion of motif m substantially helps predicting the expression levels ($S_{pm} > 1$) and performing a standard hypergeometric test. Target networks between motifs are constructed by drawing a link from motif m to m' whenever m is predicted to target one of the promoters associated with a TF that is associated with motif m'.

4.5.5 *Materials*

The publically available data sets of gene expression profiling were obtained from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo): time course of HUVEC after TNF treatment (GSE9055), mucociliary differentiation of airway epithelial cells (GSE5264), Novartis (GNF) SymAtlas (GSE1133), epithelial and mesenchymal subpopulations within immortalized human mammary epithelial cells (GSE28681), ENCODE ChIP-seq (GSE26386) and expression profiling (GSE26312) in human cell lines, and the Illumina Body Map 2 (GSE30611). Micro-Array files from the NCI-60 samples were downloaded from the project web page (http://genome-www.stanford.edu/nci60/).

4.6 AUTHORS INFORMATION

4.6.1 *List of authors*

The following authors have contributed to the work discussed in Chapter 4:

1. Piotr J. Balwierz[1] (Abbr.: PJB),
2. Mikhail Pachkov[1] (Abbr.: MP),
3. Phil Arnold[1] (Abbr.: PA),
4. Andreas Johannes Gruber[1] (Abbr.: AJG),
5. Mihaela Zavolan[1] (Abbr.: MZ) &
6. Erik van Nimwegen[1] (Abbr.: EvN)

whereat author affiliations are as follows:

1 Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Klingelberstrasse 50-70, CH-4056 Basel, Switzerland

4.6.2 *Author contributions*

The listing of authors in the previous subsection (4.6.1) was performed according to the authors' contributions, whereat the first author (PJB) contributed most and subsequent authors decreasingly. However, the last two authors are principal investigators and thus their listing follows the opposite ranking (the last contributed the most and the preceding author less).
In detail, using the abbreviations specified in the previous subsection (i.e. 4.6.1) AJG contributed with the following work: AJG benchmarked the performance of different miRNA prediction algorithms on MARA results and helped to integrate miRNA target predictions (from TargetScan [143]) into ISMARA. AJG, PJB and EvN performed the analysis on TF/miRNA interactions in EMT and wrote the corresponding subsections (4.3.4 and B.9). Finally, AJG contributed figures 4.3B and 4.4B.

4.7 SUPPLEMENTARY MATERIALS

Supplementary materials can be found in Appendix B.

4.8 ACKNOWLEDGMENTS

We would like to thank Jean Hausser for the help in probe-wise processing of micro-array data, and members of the van Nimwegen research group for helpful discussions.

4.9 FUNDING

Using the abbreviations specified in subsection 4.6.1: EvN, PA, and PJB were supported by grant SNF 3100A0-118318 of the Swiss National Science Foundation and by SystemsX.ch, the Swiss Initiative in Systems Biology, through the CellPlasticity project. AJG was supported by a Werner Siemens Fellowship. MP was supported by a grant from the Swiss Institute of Bioinformatics.

5

PEGYLATED IFN-α REGULATES HEPATIC GENE EXPRESSION THROUGH TRANSIENT JAK/STAT ACTIVATION

The work discussed in this chapter was conducted in collaboration with the van Nimwegen and the Heim labs and published in The Journal of Clinical Investigation in 2014 (see reference [377]).

5.1 ABSTRACT

The use of pegylated interferon-α (pegIFN-α) has replaced unmodified recombinant IFN-α for the treatment of chronic viral hepatitis. While the superior antiviral efficacy of pegIFN-α is generally attributed to improved pharmacokinetic properties, the pharmacodynamic effects of pegIFN-α in the liver have not been studied. Here, we analyzed pegIFN-α–induced signaling and gene regulation in paired liver biopsies obtained prior to treatment and during the first week following pegIFN-α injection in 18 patients with chronic hepatitis C. Despite sustained high concentrations of pegIFN-α in serum, the Jak/STAT pathway was activated in hepatocytes only on the first day after pegIFN-α administration. Evaluation of liver biopsies revealed that pegIFN-α induces hundreds of genes that can be classified into four clusters based on different temporal expression profiles. In all clusters, gene transcription was mainly driven by IFN-stimulated gene factor 3 (ISGF3). Compared with conventional IFN-α therapy, pegIFN-α induced a broader spectrum of gene expression, including many genes involved in cellular immunity. IFN-induced secondary transcription factors did not result in additional waves of gene expression. Our data indicate that the superior antiviral efficacy of pegIFN-α is not the result of prolonged Jak/STAT pathway activation in hepatocytes, but rather is due to induction of additional genes that are involved in cellular immune responses.

5.2 INTRODUCTION

Interferons (IFNs) are central mediators of immune responses to viral infections [378]. They exert their antiviral activity by inducing the expression of hundreds of genes that together establish an "antiviral state", which restricts the spread of virus among neighboring cells [379]. Type I IFNs (all IFN-αs and IFN-β) bind to the IFN-α receptor (IFNAR) and activate the receptor–associated tyrosine kinases Jak1 and Tyk2, which in turn activate signal transducer and activator of transcription 1 (STAT1) and STAT2 by phosphorylation of a tyrosine in the C-terminal domain [49]. Activated STAT1 combines with STAT2 and IFN regulatory factor 9 (IRF9) to form IFN-stimulated gene factor 3 (ISGF3). ISGF3 translocates into the nucleus, binds to IFN-stimulated response elements (ISREs) in gene promoters and induces the transcription of hundreds of genes. Activated STAT1 can also form homodimers that bind to γ-activated sequences (GASs) and induce an overlapping but distinct set of

IFN-stimulated genes (ISGs). IFN-induced Jak/STAT signaling is tightly controlled by negative regulators. Suppressor of cytokine signaling 1 (SOCS1) and SOCS3 are rapidly induced and strongly inhibit STAT1 phosphorylation at the receptor-kinase complex within hours [380]. SOCS proteins are also rapidly degraded and in most cells become undetectable within hours after their induction. However, IFN signaling remains refractory for days in many cell types [381]. In the liver of mice repeatedly injected with IFN-α, a long-lasting upregulation of ubiquitin-specific peptidase 18 (USP18) was found to be responsible for prolonged unresponsiveness of liver cells to IFN-α [382]. For more than 25 years, recombinant IFN-α has been used for the treatment of hepatitis C virus (HCV) infections [383]. HCV is a parenterally transmitted positive-strand RNA virus that replicates in human hepatocytes and can cause chronic hepatitis with progressive fibrosis, leading to cirrhosis and hepatocellular carcinoma [384]. Initially, unmodified recombinant IFN-α2a or -α2b was used alone or in combination with the antiviral compound ribavirin. In 2001, pegylated IFN-α (pegIFN-α) became the standard of care because of its superior efficacy [385, 386]. The covalent attachment of polyethylene glycol (PEG) molecules to IFN-α produces a biologically active molecule with a longer half-life. The delayed clearance allows once-weekly injections, compared with three times a week for conventional IFN-α. It is generally assumed that the sustained high serum concentrations of pegIFN-α provide for uninterrupted antiviral activity through a permanent stimulation of the IFN signaling pathways, whereas the serum concentrations of standard IFN-α (with an elimination half-life of 4 to 10 hours) decline below pharmacologically active levels in the second half of each 48-hour dosing interval [387, 388]. However, there is no experimental evidence supporting prolonged pharmacodynamic effects of pegIFN-α. On the contrary, the refractoriness of Jak/STAT signaling in mouse liver challenges the concept that pegIFN-α is more effective because of prolonged stimulation of IFN signaling pathways [382]. We previously investigated pegIFN-α–induced signaling and gene regulation in the liver of 16 patients who started treatment of their chronic hepatitis C (CHC) [389]. All patients had a pretreatment liver biopsy during the routine workup for CHC and a second liver biopsy 4 hours after the first subcutaneous injection of pegIFN-α. Six patients had an induction of ISGs already before treatment and showed no further activation of IFN signal transduction or ISG expression in response to pegIFN-α. None of these patients responded to therapy [389]. It is now firmly established that patients with an activated endogenous IFN system are poor responders to IFN-α–based therapies [389–392], and quantification of the expression of a limited number of ISGs from liver biopsies allows the most accurate prediction of response to pegIFN-α and ribavirin [393]. In the 10 patients without a preactivation of the hepatic IFN system, pegIFN-α induced phosphorylation and nuclear translocation of STAT1 and the expression of hundreds of ISGs within 4 hours [389]. Nine patients had a sustained virological response (SVR) later and were cured of CHC, and 1 patient had a virological response during treatment, but later relapsed. In the present work, again using a paired biopsy approach, we extended the pharma-

codynamic analysis of pegIFN-α to the entire 1-week dosing interval in an additional 12 patients. Three patients each had a second liver biopsy 16, 48, 96, and 144 hours after the first injection of pegIFN-α2b. This unique analysis of the molecular effects of pegIFN-α in human liver revealed that Jak/STAT signaling occurs in the first 24 hours and then becomes refractory in hepatocytes for the entire dosing interval despite persistently high pegIFN-α serum concentrations. Compared with conventional IFN-α, we found that pegIFN-α induced a broader spectrum of ISGs, including many genes involved in cellular immune responses. The initial activation of ISGF3 was the main driver of ISG transcription during the entire week after the first injection of pegIFN-α. The induction of secondary transcription factors and of unphosphorylated STAT1 (U-STAT1) had negligible effects. We conclude that the superior therapeutic efficacy of pegIFN-α is not caused by a sustained activation of the Jak/STAT pathway in hepatocytes, but rather by the sustained induction of ISGs in liver-infiltrating immune cells.

5.3 RESULTS

5.3.1 *pegIFN-α2b induced STAT1 phosphorylation and ISG expression in the liver*

We studied pegIFN-α2b–induced STAT1 phosphorylation and gene regulation in 18 patients who underwent treatment for CHC with pegIFN-α and ribavirin. All patients had a first liver biopsy before treatment during the routine clinical CHC work-up. A second biopsy was taken 4 hours (n = 6), 16 hours (n = 3), 2 days (n = 3), 4 days (n = 3), and 6 days (n = 3) after the first injection of pegIFN-α2b. The 6 patients whose second liver biopsy was performed 4 hours after injection were selected from among the 16 patients who had already been included in the previous study described above [389], because they had no preactivation of the endogenous IFN system in the liver and a normal response to pegIFN-α2b. The patients were selected in a two-step procedure for the later time points. First, liver biopsies from patients with CHC who agreed to donate part of their liver biopsy for research were analyzed with a previously developed and validated four-gene classifier to predict their likelihood of responding to pegIFN-α [393]. Patients with a high probability of an unimpaired, normal response to pegIFN-α were then asked to participate in our study and to consent to a second liver biopsy. This two-step selection process was necessary, because in patients with preinduced hepatic ISGs, the Jak/STAT signaling pathway is refractory in liver cells [389], and a second liver biopsy would have been of little use for the study of pegIFN-α pharmacodynamic effects in the liver. Indeed, the selection process with the four-gene classifier was highly accurate in predicting a good response to pegIFN-α: all patients were treatment responders, and apart from 1 patient who relapsed after treatment, all patients were cured of their HCV infection (Table 5.1).

Patient no.	Age (yr)	Sex	HCV GT	Viral load, log IU/mL			Response			METAVIR	IL28B GT	Time point	Medication	IFN conc. (pg/mL)
				Baseline	4-wk	12-wk	4-wk	12-wk	Follow-up					
1	52	m	3	7.14	neg	neg	RVR	cEVR	SVR	A2/F2	CC	4h	PegIFN-α-2b	138
2	37	m	3	4.9	neg	neg	RVR	cEVR	SVR	A1/F2	CT	4h	PegIFN-α-2b	530
3	54	f	2	4.95	neg	neg	RVR	cEVR	SVR	A3/F3	CT	4h	PegIFN-α-2b	214
4	57	m	3	5.25	2.15	neg	Non-RVR	cEVR	Relapse	A3/F4	CC	4h	PegIFN-α-2b	702
5	38	m	4	4.08	1.66	neg	Non-RVR	cEVR	SVR	A2/F2	CT	4h	PegIFN-α-2b	241
6	51	f	1	6.82	3.52	neg	Non-RVR	cEVR	SVR	A1/F2	CT	4h	PegIFN-α-2b	419
7	26	m	3	4.58	neg	neg	RVR	cEVR	SVR	A1/F1	TT	16h	PegIFN-α-2b	1194
8	42	f	3	5.49	neg	neg	RVR	cEVR	SVR	A1/F2	CT	16h	PegIFN-α-2b	774
9	41	m	3	5.66	neg	neg	RVR	cEVR	SVR	A1/F2	CT	16h	PegIFN-α-2b	973
10	30	m	3	7.07	neg	neg	RVR	cEVR	SVR	A2/F2	CT	48h	PegIFN-α-2b	356
11	57	f	1	5.95	neg	neg	RVR	cEVR	SVR	A2/F2	CC	48h	PegIFN-α-2b	414
12	37	m	3	6.72	1.28	neg	Non-RVR	cEVR	SVR	A3/F2	CT	48h	PegIFN-α-2b	887
13	62	m	4	7.16	neg	neg	RVR	cEVR	SVR	A3/F4	CT	96h	PegIFN-α-2b	1567
14	43	m	1	5.6	1.63	neg	Non-RVR	cEVR	SVR	A3/F2	CC	96h	PegIFN-α-2b	155
15	40	m	1	5.16	1.41	neg	Non-RVR	cEVR	SVR	A3/F4	CT	96h	PegIFN-α-2b	186
16	25	f	1	2.64	neg	neg	RVR	cEVR	SVR	A2/F2	CC	144h	PegIFN-α-2b	NA
17	70	m	2	6.86	1.84	neg	Non-RVR	cEVR	SVR	A2/F3	CT	144h	PegIFN-α-2b	NA
18	34	m	3	5.56	neg	neg	RVR	cEVR	SVR	A2/F2	CT	144h	PegIFN-α-2b	233
19	57	f	2	5.18	neg	neg	RVR	cEVR	SVR	A2/F2	CC	144h	PegIFN-α-2a	6564
20	57	m	1	6.54	4.59	3.33	Non-RVR	EVR	interrupted	A3/F4	CT	144h	PegIFN-α-2a	6146
21	38	f	4	6.32	5.07	neg	Non-RVR	cEVR	SVR	A3/F4	CC	144h	PegIFN-α-2a	15986

Table 5.1: Patient characteristics. conc, concentration; GT, genotype; Neg, negative; RVR, rapid virological response (undetectable viral load at 4 weeks); cEVR, complete early virological response (undetectable viral load at 12 weeks); EVR, early virological response (>log_2 reduction of viral load at 12 weeks); SVR, sustained virological response (undetectable viral load 24 weeks after end of treatment); METAVIR, liver histology score for grading inflammation (A1–A3) and staging fibrosis (F0–F4).

We analyzed pegIFN-α2b–induced Jak/STAT signaling by immunohistochemistry (IHC) with phosphorylated STAT1–specific (p-STAT–specific) antibodies. Because we selectively included only patients who had no pretreatment induction of the endogenous IFN system, STAT1 was not activated in the biopsies obtained before treatment (Figure 5.1A).

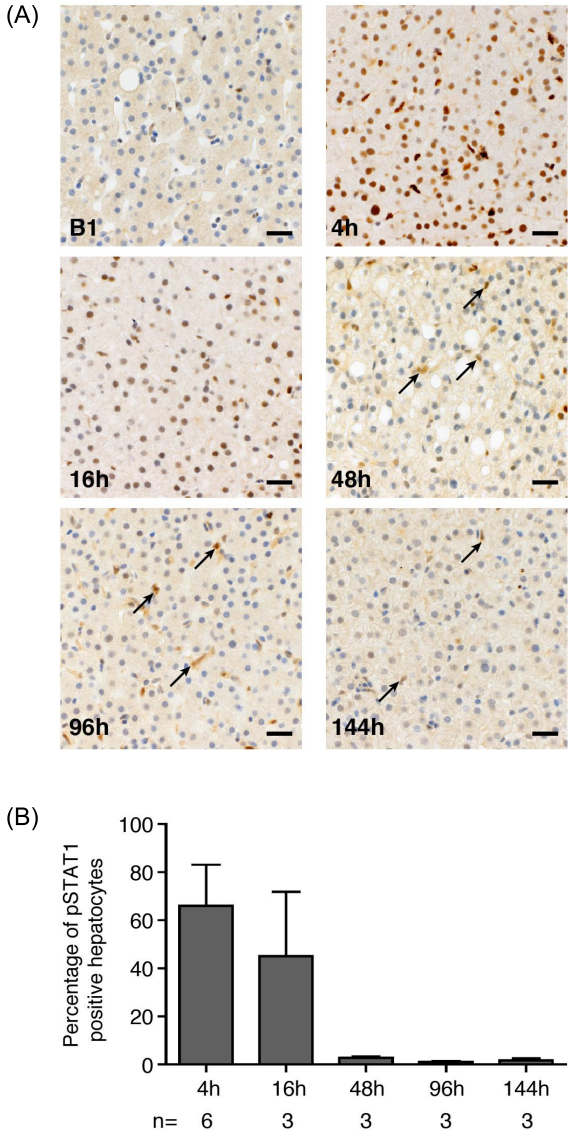


Figure 5.1: pegIFN-α2b transiently induces the Jak/STAT pathway in the liver. **(A)** Representative images of IHC analysis of p-STAT1 in liver biopsies obtained before treatment (B1) and at several time points after the first injection of pegIFN-α2b. Strong nuclear p-STAT1 signals were present at the 4- and 16-hour time points, but not at later time points, where the signals were localized in nonparenchymal cells (arrows). Scale bars: 20 μm. **(B)** Quantitative analysis of the mean percentage of p-STAT1–positive hepatocyte nuclei (5 $\times$ 100 cells counted per sample; the number of samples is indicated) per time point. Bars show the mean with SEM.

Following the first pegIFN-α2b injection, we observed a rapid and strong activation of STAT1 already 4 hours later, with nuclear p-STAT1 signals detected in more than 60% of hepatocytes (Figure 5.1). p-STAT1 signals were still strong after 16 hours, but then rapidly declined. In liver biopsies obtained after 2, 4, or 6 days, p-STAT1 signals in hepatocytes were weak and were detected in less than 5% of hepatocytes. In nonparenchymal cells, we detected p-STAT1 signals at all time points. To further address the kinetics of

ISG induction by pegIFN-α2b, we adapted a highly sensitive and specific *in situ* hybridization (ISH) method (QuantiGene ViewRNA) that allowed the detection of ISG mRNAs in fresh-frozen liver biopsy samples. We detected MX1 mRNA already 4 hours after the injection of pegIFN-α2b and found that it peaked at the 16-hour time point and then rapidly declined (Figure 5.2A). IFI27 mRNA expression peaked at 16 hours and declined at a much slower rate. Of note, the intensity of the signals declined in all hepatocytes, and at later time points we did not detect hepatocytes with the signal intensities found at the 16-hour point. Together with the absence of strong nuclear p-STAT1 signals in hepatocytes at later time points (Figure 5.1A), these data do not support the hypothesis that hepatocytes recover asynchronously from the refractory state and that they are, in part, restimulated by pegIFN-α2b circulating at high concentrations during the entire dosing interval (Table 5.1).

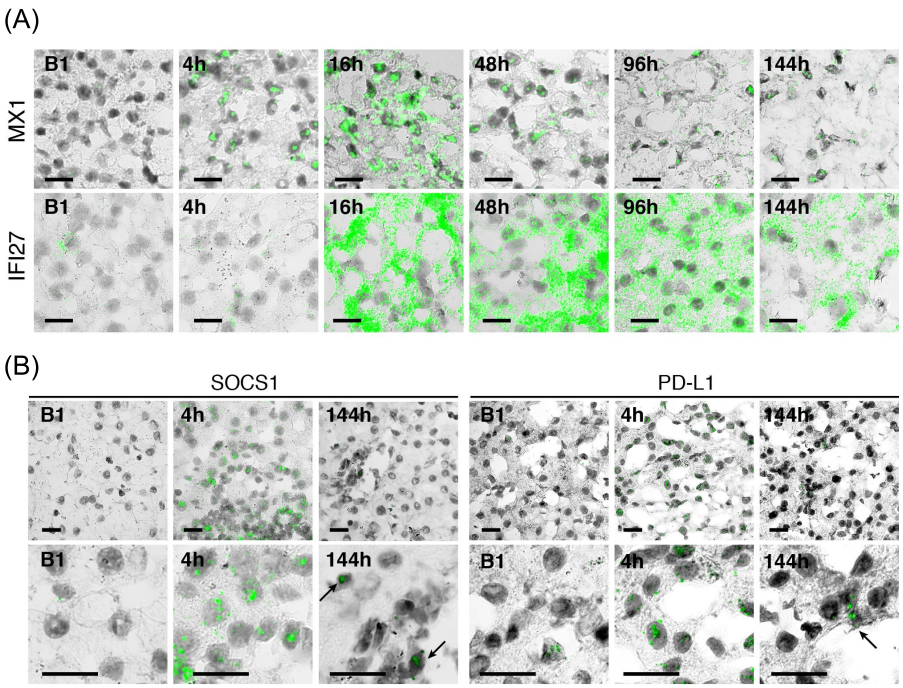


Figure 5.2: ISH reveals distinct expression patterns of ISG mRNAs at different time points. (A) Representative examples of ISH staining (green) in liver biopsies for MX1 and IFI27 mRNA showing that ubiquitous expression gradually declined over time with distinct kinetics. **(B)** ISH staining (green) in liver biopsies for SOCS1 and PDL1 mRNA revealed expression in hepatocytes and in nonparenchymal cells at 4 hours. At 144 hours, SOCS1 and PDL1 were detected only in nonparenchymal cells (black arrows). Scale bars: 20 μm.

In contrast to hepatocytes, we found that nonparenchymal cells showed strong nuclear p-STAT1 signals also at later time points (Figure 5.1A, arrows). Accordingly, SOCS1 and PDL1 mRNAs, two ISGs that are only transiently induced in hepatocytes, were also expressed at the 144-hour time point in nonparenchymal cells (Figure 5.2B). We conclude that in hepatocytes, pegIFN-α2b induces a transient activation of the Jak/STAT signaling pathway during the first day, but not during the entire 1-week dosing interval, and this despite

sustained high serum concentrations of pegIFN-α2b at all time points (Table 5.1). We found that nonparenchymal cells remained IFN-α sensitive at all time points investigated.

5.3.2 *Induction of negative regulators of Jak/STAT signaling*

We then assessed the induction of negative regulators of IFN signaling in the liver biopsies. On the mRNA level, SOCS1 was strongly induced at 4 hours and 16 hours, but then returned to pretreatment expression levels (Figure 5.3A).

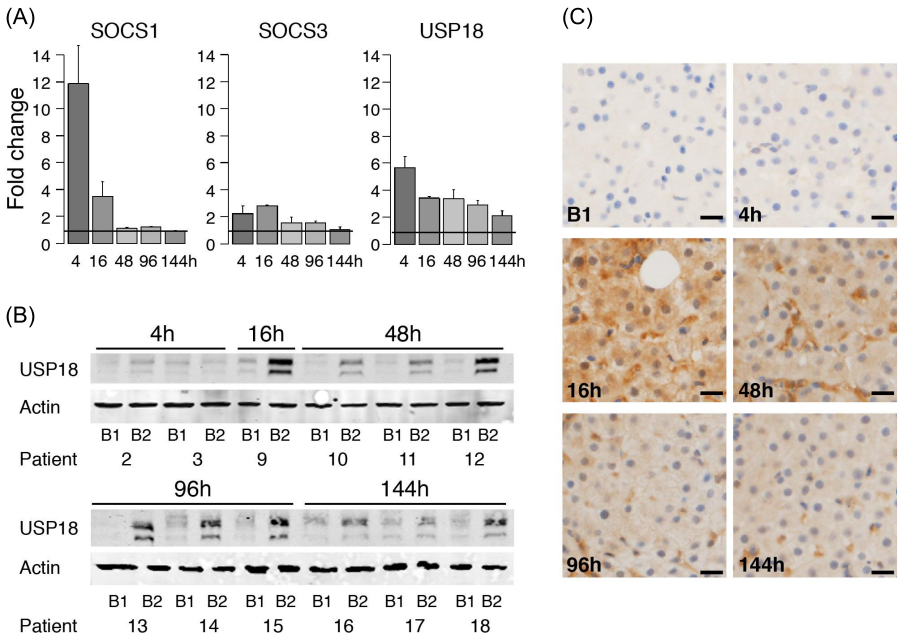


Figure 5.3: The negative regulator USP18 is continuously upregulated during the entire week after pegIFN-α2b injection. (A) Bar plot indicating the mRNA expression fold change between the pretreatment biopsy (B1) and the on-treatment biopsy (B2) of SOCS1, SOCS3, and USP18. Data represent the mean with SEM (n = 6 for the 4-hour time point; n = 3 for all other time points). The black line indicates the baseline measured in pretreatment biopsies from the same patients (n = 18). **(B)** USP18 protein expression by Western blot analysis using whole-cell extracts of liver samples from B1 and B2. Patients are numbered according to Table 1. **(C)** Representative images of IHC for USP18 of liver biopsies obtained before treatment (B1) and at several time points after the first injection of pegIFN-α2b as indicated. Scale bars: 20 μm.

SOCS3 was also upregulated in the first 16 hours, albeit to a lesser extent (up to 2.5-fold) and remained slightly elevated for up to 4 days. USP18 was also rapidly induced, but unlike SOCS1 and SOCS3, the expression level of USP18 mRNA remained persistently high during the entire week (Figure 5.3A). Accordingly, USP18 protein was detectable from 16 hours on at all time points by Western blot and IHC analyses (Figure 5.3B and C). Presumably for technical reasons, we could not detect SOCS1 or SOCS3 proteins at

any time point, despite testing several different antibodies. We conclude that pegIFN-α2b induces transient activation of the Jak/STAT signaling pathway in hepatocytes because of the rapid induction of SOCS1, SOCS3, and USP18 and that the signaling pathway remains refractory to ongoing stimulation by circulating pegIFN-α2b because of the persistent induction of USP18.

5.3.3 *pegIFN-α2b–induced genes fall into four robust classes with distinct temporal expression patterns*

We assessed pegIFN-α2b–regulated gene expression with transcriptome analysis using Affymetrix U133 Plus 2.0 arrays. Pairwise comparison of pretreatment and on-treatment biopsies revealed a greater than 2-fold induction in two-thirds of samples of hundreds of genes, with a peak at 16 hours (Figure 5.4A and Supplemental Table C.1).

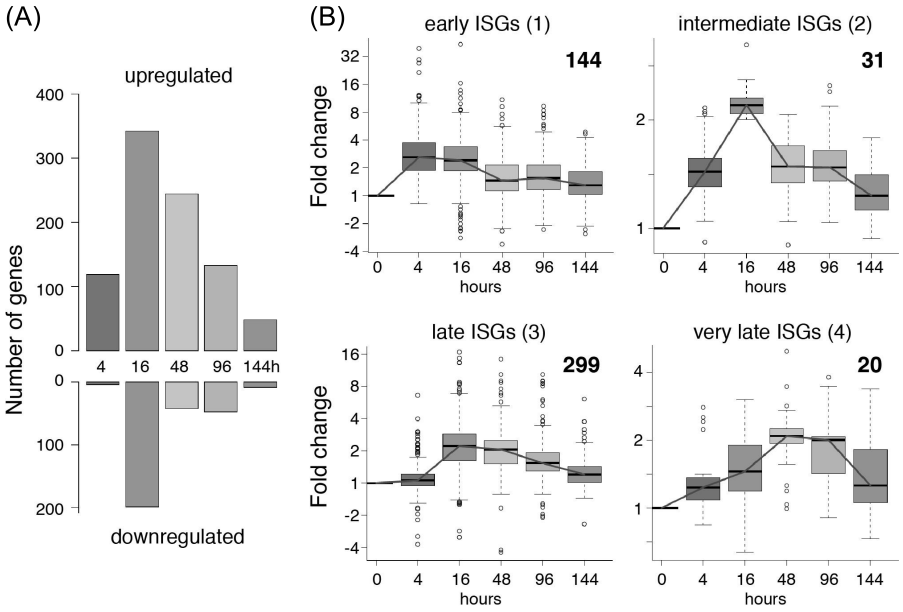


Figure 5.4: pegIFN-α2b–induced genes fall into four robust classes with distinct temporal expression patterns. (A) Number of genes greater than 2-fold up- or downregulated in two-thirds of the patients at each time point. **(B)** Clustering analysis of the upregulated genes produced four robust clusters (numbers 1–4) composed of early, intermediate, late, and very late ISGs. Boxes represent the quartiles, and whiskers represent 1.5 times the interquartile range. Bold line indicates the median expression value, and the number of genes in each cluster is indicated.

Likewise, up to 200 genes were downregulated (Supplemental Table C.2). To gain insight into the temporal expression patterns of ISGs induced by pegIFN-α2b in the human liver, we analyzed the transcriptome data using a Bayesian clustering algorithm. The algorithm produced four robust clusters of upregulated genes, which were termed early (144 genes), intermediate (31 genes), late (299 genes), and very late ISGs (20 genes) (Figure 5.4B and Supplemental Table C.1). For over 95% of all upregulated genes, the peak mRNA levels occurred 4 or 16 hours after injection, followed by a steady decline over

the remaining 128 hours of treatment (Figure 5.4B). Because of the limited amount of tissue obtained by percutaneous liver biopsies, we could not comprehensively analyze ISG protein expression. We therefore measured the protein expression of three exemplary ISGs. USP18 protein expression peaked at the 16-hour time point and then gradually declined, but remained induced up to the 144-hour time point. We found that USP18 mRNA expression peaked already at 4 hours, but was also persistently induced up to the 144-hour time point (Figure 5.3). STAT1 mRNA was induced up to the 96-hour time point, whereas STAT1 protein expression was still increased at 144 hours (Supplemental Figure C.1). We found a very good correlation between IP10 mRNA expression in the liver and IP-10 protein concentration in the serum (Supplemental Figure C.1). Taken together, we found a reasonably good correlation between mRNA and protein expression of ISGs in this limited set of exemplary ISGs.

5.3.4 *Compared with conventional IFN-α, pegIFN-α2b induces a broader range of genes including many ISGs involved in cellular immune responses*

Given the known superior antiviral efficacy of pegIFN-α, we could not treat our study patients with conventional IFN-α. To compare IFN-α and pegIFN-α–induced gene regulation, we therefore made use of previously published transcriptome data obtained 24 hours after the injection of conventional IFN-α [394]. Fortunately, the samples were analyzed on the same Affymetrix U133 Plus 2.0 arrays, allowing a direct comparison of the data. The discrepant time points after injection between the pegIFN-α and the IFN-α studies were a potential pitfall, but unsupervised hierarchical clustering of the combined data positioned the IFN-α2a samples properly between the 16-hour time point and the 48-hour time point of the pegIFN-α samples. Importantly, the magnitude of mRNA upregulation in the IFN-α samples was comparable to that in the pegIFN-α samples from the 16- and 48-hour time points. The most striking difference between IFN-α and pegIFN-α was the number of genes that were induced more than 2-fold in two-thirds of the samples (Figure 5.5A).

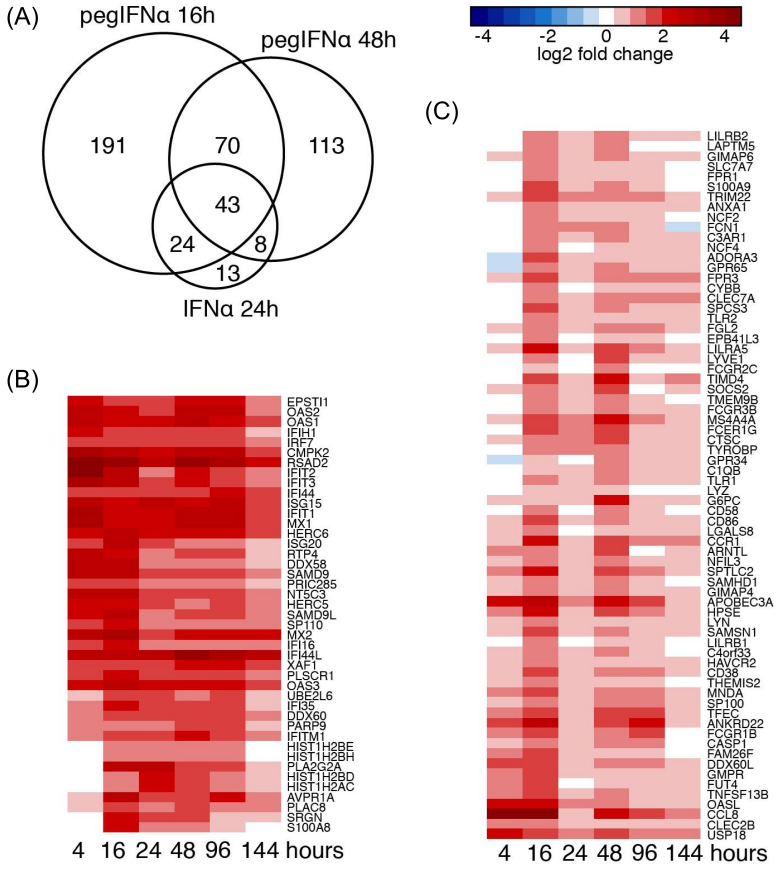


Figure 5.5: IFN-α2a induces mainly "classical" ISGs, while pegIFN-α2b leads to transcription of additional immune cell–associated genes. **(A)** Venn diagram of genes identified as being upregulated by more than 2-fold in two-thirds of the patients at 16 or 48 hours after pegIFN-α2b injection (n = 3 each) or 24 hours after conventional IFN-α injection (n = 6). **(B and C)** Heatmaps show expression patterns (mean log_2 fold change compared with paired pretreatment biopsies) of genes upregulated after IFN-α injection. **(B)** 43 ISGs were upregulated by conventional IFN-α2a at 24 hours as well as by pegIFN-α2b at both 16 and 48 hours. **(C)** 70 ISGs were upregulated by pegIFN-α2b at both 16 and 48 hours, but not by conventional IFN-α2a at 24 hours.

We found that most of the genes upregulated by IFN-α were also induced by pegIFN-α, but a substantially larger number of genes were induced more than 2-fold exclusively by pegIFN-α. Gene ontological (GO) analysis revealed these to be genes associated with immune cells and adaptive immunity, whereas the genes upregulated by both IFN-α and pegIFN-α fell into the "classical" ISG group (Figure 5.5B and C, and Supplemental Table C.4). We conclude that while a common subset of ISGs is upregulated within the first 1–2 days independently of the IFN-α formulation, an additional set of genes associated with cellular immune responses is more markedly induced by pegIFN-α.

5.3.5 *pegIFN-α2a and pegIFN-α2b induce overlapping sets of genes in the liver 144 hours after injection despite their different pharmacokinetic properties*

Two different formulations of pegIFN-α2 with distinct pharmacokinetic properties are approved for the treatment of viral hepatitis: pegIFN-α2b (PegIntron) and pegIFN-α2a (PEGASYS). While the single-chain PEG moiety of pegIFN-α2b is subject to hydrolysis, which leads to release of IFN-α2b into the human body and faster elimination of the drug, pegIFN-α2a is not hydrolyzed, has a lower absorption rate, and is eliminated at a much slower rate [395]. pegIFN-α2a achieves maximal serum levels of 7,000 pg/ml about 80 hours after administration, and the peak extends up to 168 hours after injection [396], as opposed to a much earlier peak (15–44 hours) and a more rapid decline of pegIFN-α2b. In order to investigate whether these distinct pharmacokinetic properties result in distinct pharmacodynamic effects, we included 3 additional patients in our study who were treated with pegIFN-α2a and obtained a second liver biopsy at the end of the 1-week dosing interval. As expected, pegIFN-α2a serum concentrations were still high at the end of the first week, whereas pegIFN-α2b concentrations declined in the second half of the dosing interval (Table 5.1). However, despite the difference in serum concentration between pegIFN-α2a and pegIFN-α2b, we found that the number of genes upregulated by greater than 2-fold in two-thirds of the patients in each group was not significantly different (59 versus 49 genes, respectively). Furthermore, we observed a considerable overlap of the gene sets, with 26 genes being upregulated by both pegIFN-α2a and pegIFN-α2b, and these common genes comprised most of the typical ISGs (Supplemental Table C.3). We conclude that the different pharmacokinetic properties of the two pegIFN-α2 formulations do not cause significant differences in ISG expression at the end of a 1-week dosing interval.

5.3.6 *pegIFN-α2b–induced gene transcription is mainly driven by IFN-stimulated response element motifs during the entire dosing interval*

Among the hundreds of genes induced by IFN-α, one also finds several transcription factors such as IFN regulatory factors (IRFs), cytokines and chemokines that could directly or indirectly activate additional signal transduction pathways and transcriptional programs (Supplemental Table C.1). Such "secondary" transcription factors could be the drivers of gene transcription at later time points when pegIFN-α2b–induced Jak/STAT signaling is refractory. We therefore analyzed the relative contribution of transcription factor–binding motifs to global gene expression at 4 hours, 16 hours, 2 days, 4 days, and 6 days using a recently developed method called motif activity response analysis (MARA) [50]. MARA infers the activities of transcription regulators by modeling genome-wide expression profiles in terms of computationally predicted binding sites for a large array of mammalian regulatory motifs such as IFN-stimulated response element (ISRE). Roughly speaking,

MARA infers that a regulatory motif increases in activity when its predicted target promoters show an overall increase in expression that cannot be explained by the occurrence of other regulatory motifs in these promoters. In our current application, we used MARA to calculate changes in the activity of motifs across paired samples (pretreatment versus on-treatment). This analysis revealed ISRE as the most substantially changing motif across all time points up to 6 days (Figure 5.6A and B). We observed a strong positive ISRE motif activity change in all patients (Figure 5.6A). MARA identified additional motifs that contribute to gene expression changes such as GAS, DMAP1_NCOR{1,2}_SMARC (DMAP1), PRRX1,2, and ATF6. However, the changes in their activities were relatively minor in comparison with ISRE (Figure 5.6B). MARA results of the transcription factor–binding site (TFBS) analysis were confirmed by motif discovery analysis using HOMER software [397]. In each of the four ISG clusters (Figure 5.4B), ISRE was by far the most significantly enriched motif (Supplemental Figure C.2).

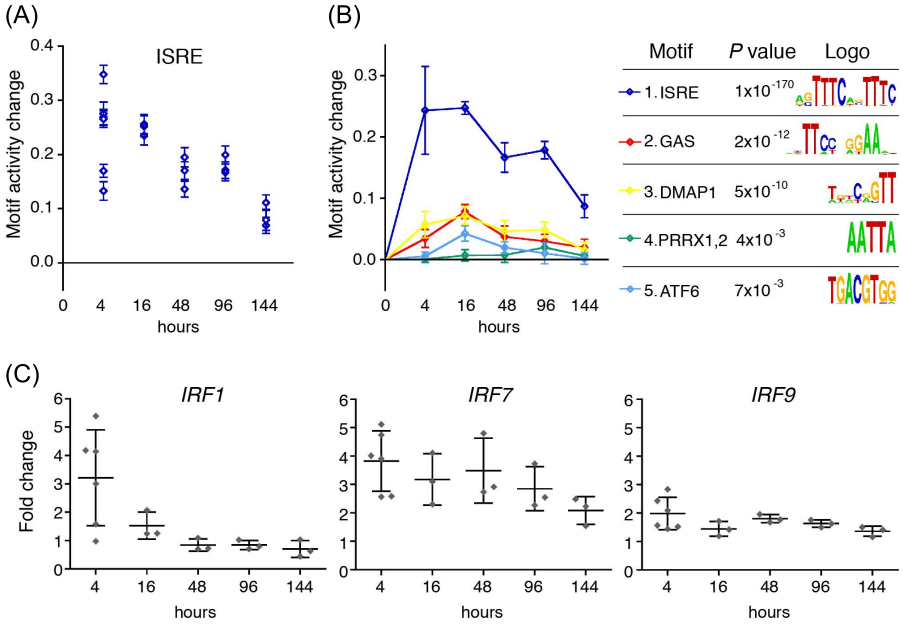


Figure 5.6: MARA reveals ISRE as the most significantly upregulated motif. **(A)** Activity changes of the ISRE motif in each patient, as inferred by MARA, showed a significant increase in ISRE activity for every patient at every time point. Shown are inferred activity changes (points) ± 1 SD (bars). **(B)** Motif activity profiles of the top five motifs with the most significant positive activity changes. Shown are the mean activity changes per time point (lines) ± 1 SEM as well as the P values and sequence logos of the motifs. **(C)** Fold change of IRF1, IRF7, and IRF9 mRNA expression for every patient. Shown are the mean values with SEM at each time point.

ISRE motifs are the binding sites for ISGF3 and also IRFs. ISGF3 is activated by IFN-α–induced phosphorylation of STAT1 and STAT2. IRFs are transcriptionally induced and are also regulated by phosphorylation [398]. We therefore measured the expression of IRF mRNAs. Of the nine IRFs, only IRF1, IRF7, and IRF9 were upregulated by pegIFN-α2b in the liver (Figure 5.6C).

IRF1 is transiently induced at the 4- and 16-hour time points. IRF9 is part of the ISGF3 complex, and its transcriptional activity depends on p-STAT1 and p-STAT2. IRF7 is upregulated during the entire dosing interval of pegIFN-α2b and could also be involved in ISRE-mediated gene transcription. However, the transcriptional activity of IRF7 depends on serine phosphorylation by IKK-α [399], a downstream component of cellular sensory pathways that are activated by viral pathogen–associated molecular patterns (PAMPs).

5.3.7 *Unphosphorylated STAT1 does not prolong ISG induction*

The central IFN-α–induced signal transducer and transcription factor STAT1 is itself one of the most strongly induced ISGs. Indeed, we found STAT1 mRNA strongly induced in the first 4 days after pegIFN-α2b injection (Supplemental Table C.1). STAT1 protein was even upregulated during the entire 1-week dosing interval (Supplemental Figure C.1). The functional significance of the expression of large amounts of U-STAT1 protein is unclear, but, intriguingly, a recent paper described a role of U-STAT1 as an active transcription factor that prolongs gene transcription after dephosphorylation of p-STAT1 [400]. In that work, thirty ISGs were found to be upregulated by U-STAT1–driven transcription [400]. We therefore hypothesized that U-STAT1 could be involved in the prolonged ISG induction by pegIFN-α2b. However, when we took the list of U-STAT1–induced genes and investigated their expression during the first week of pegIFN-α2b therapy, we did not find them to be overrepresented in clusters 3 and 4 with late and very late induced ISGs, respectively (data not shown). We therefore decided to address the potential of U-STAT1 to induce gene transcription in a more rigorous way. To that end, STAT1-deficient U3A cells [401] were stably transfected with STAT1 wild-type (STAT1-WT) or a mutant STAT1 with a phospho-tyrosine acceptor site at position 701 mutated to phenylalanine (STAT1-Y701F). For both STAT1-WT and STAT1-Y701F, three clones with different STAT1 expression levels were selected. One clone each expressed the transfected STAT1 at levels usually present in unstimulated parental 2fTGH cells, one clone each expressed the constructs at levels found after maximal STAT1 expression obtained in 2fTGH cells stimulated with IFN-α for 24 hours, and one clone each expressed the transfected constructs at intermediate levels (Figure 5.7A).

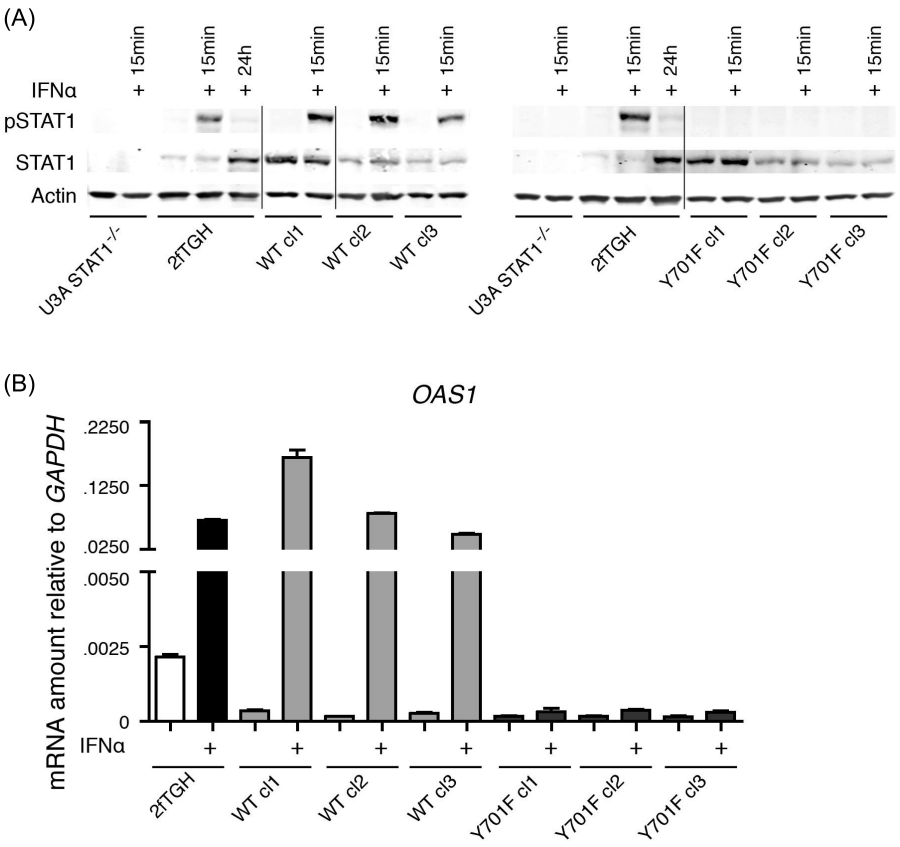


Figure 5.7: U-STAT1 does not induce ISGs. (A) Three clones with different expression levels of WT (WT cl1, WT cl2, and WT cl3) or mutated STAT1 (Y701F cl1, Y701F cl2, and Y701F cl3) were stimulated with IFN-α. STAT1-deficient U3A cells and STAT1 WT parental 2fTGH cells were used as controls. IFN-α induced STAT1 phosphorylation in 2fTGH and in all three WT clones. Actin is shown as a loading control. The cells were either untreated or treated for 15 minutes with 1,000 U/ml of IFN-α. WT cl1 and Y701F cl1 express STAT1 in an amount similar to that induced by 2fTGH cells treated for 24 hours with 1,000 U/ml of IFN-α. Shown are representative blots each from three independently performed experiments (black lanes separate blots that were derived from the same gel, but were noncontiguous). **(B)** IFN-α–induced OAS1 mRNA expression relative to GAPDH was assessed by qRT-PCR. Cells were treated with 1,000 U/ml IFN-α for 8 hours. Upregulation of OAS1 was found only in cells with WT STAT1 after IFN-α treatment. Expression of maximal amounts of Y701F-mutated STAT1 in U3A cells did not induce ISG expression. Shown are the mean values with SEM of three replicate experiments.

As we expected, IFN-α treatment of U3A cells transfected with STAT1-WT induced ISGs. In contrast, we observed no ISG induction in U3A cells transfected with STAT1-Y701F (Figure 5.7B and Supplemental Figure C.3). These results do not support a role for U-STAT1 in prolonged ISG expression.

5.3.8 *Ongoing gene transcription and lower mRNA decay rates both contribute to prolonged expression of "late" ISGs*

Since ISRE seems to be the main TFBS in all transcription clusters and U-STAT1 was not able to induce ISGs, we next hypothesized that the genes belonging to the late ISG clusters might show prolonged expression due to lower mRNA degradation rates, since such a mechanism was recently proposed to play an important role in temporal expression patterns of genes induced by TNF-α [402]. Decay of mRNAs can be regulated by specific microRNA recognition sequences present in the 3' untranslated regions (UTRs) of mRNAs [403]. We therefore analyzed our transcriptome datasets for specific binding sites of microRNAs to test whether the four ISG clusters defined by our unbiased clustering approach (Figure 5.4) have distinct microRNA binding sites in their 3'UTRs. However, we could not identify biologically meaningful microRNA binding patterns that would predict or explain the differences in decay rates of the four clusters (data not shown). We also analyzed the decay rates of mRNAs experimentally in IFN-α–treated Huh7 cells by inhibition of gene transcription with actinomycin D. Relative to GAPDH mRNA, early ISGs (RSAD2, USP18) showed faster mRNA decay, while the late ISGs (IFI27, LGALS3BP) decayed more slowly than GAPDH (Figure 5.8A).

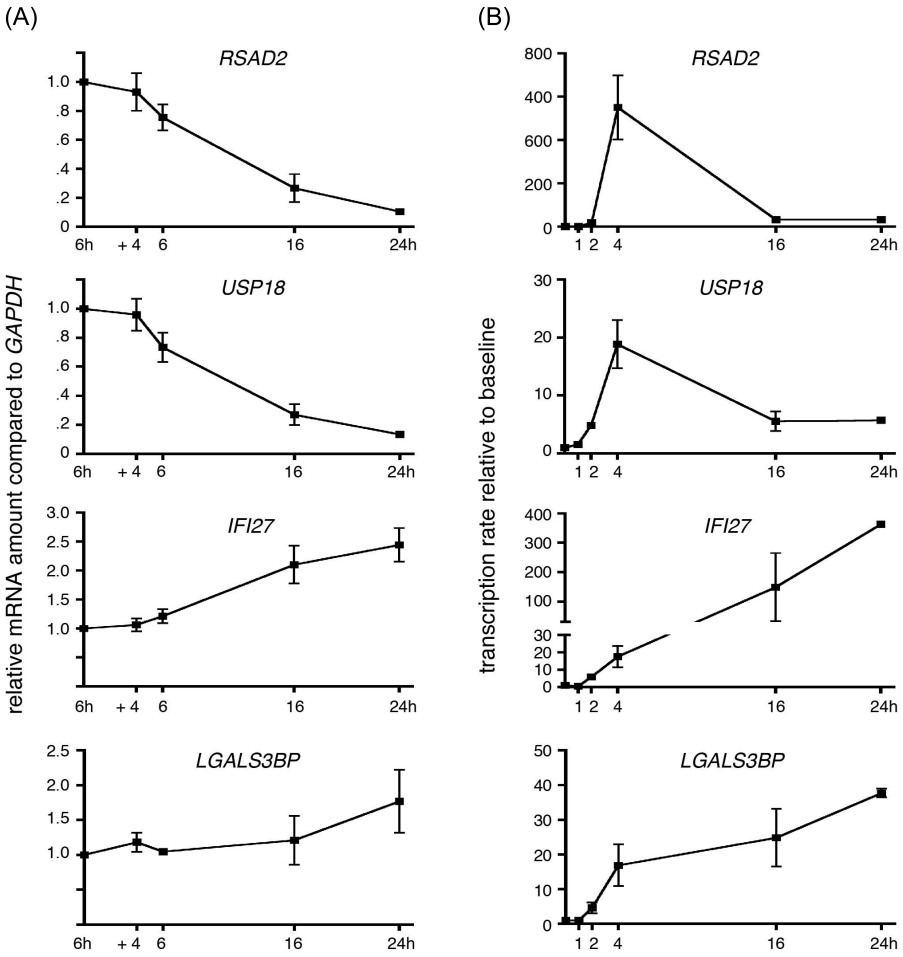


Figure 5.8: Late ISGs show a more prolonged transcriptional induction and a slower mRNA degradation rate than early ISGs in vitro. **(A)** mRNA degradation of early (RSAD2, USP18) and late (IFI27, LGALS3BP) ISGs was assessed in Huh7 after induction with 1,000 U/ml of IFN-α for 6 hours at the indicated time points. Transcription was blocked with actinomycin D, and the mRNA degradation over time was compared with GAPDH by qRT-PCR. Results from two independent experiments run in duplicate are shown. **(B)** Transcription rates of early (RSAD2, USP18) and late (IFI27, LGALS3BP) ISGs over time in Huh7 cells treated with 1,000 U/ml of IFN-α for the indicated time points. In vitro transcription in isolated nuclei was performed for 45 minutes. Newly transcribed mRNA labeled with biotinylated UTP was isolated and assessed by qRT-PCR. Results depicted as relative transcription compared with untreated baseline are shown from two independent experiments run in duplicate.

However, a delayed mRNA decay rate cannot readily explain the expression peaks at later time points such as those observed in cluster 4 genes (Figure 5.3). We therefore also analyzed the transcription of representative early (RSAD2, USP18) and late (IFI27, LGALS3BP) ISGs using a nuclear run-on assay. Nuclei were isolated from Huh7 cells after 1, 2, 4, 16, and 24 hours of stimulation with 1,000 IU/ml IFN-α and were then incubated with biotin-labeled UTP for 45 minutes. The newly transcribed mRNA was purified on streptavidin beads and quantified by quantitative PCR (qPCR). We found a

markedly prolonged transcription of late versus early ISGs (Figure 5.8B). We conclude that the temporal expression patterns of ISGs are determined by the duration of gene transcription as well as by different mRNA decay rates.

5.4 DISCUSSION

Arguably, no other cytokine has been used in clinical medicine more extensively than recombinant (peg)IFN-α. Hundreds of thousands of patients with chronic hepatitis B and CHC have been treated worldwide in the past 20 years. Despite this clinical success story, amazingly little is known about the mechanism of action and the pharmacodynamic effects of IFN-α and pegIFN-α. In principle, IFN-α exerts its antiviral effect both through the induction of antiviral effector systems in infected cells and through the regulation of immune cells such as natural killer cells, dendritic cells, and T cells [378]. Immunomodulatory effects of IFN-α have been investigated in cells isolated from blood in patients with CHC undergoing therapies with (peg)IFN-α. Indeed, several studies reported that HCV-specific T cell reactivity is increased by IFN-α treatment and correlates with treatment response [404–406]. However, other investigators have shown no association [407, 408]. The direct effects of (peg)IFN-α on hepatocytes are more difficult to study because of the requirement of liver biopsies from patients undergoing pegIFN-α treatments. In a previous study including 16 patients with CHC, we analyzed pegIFN-α2b–induced Jak/STAT signaling and global gene expression in paired liver biopsies obtained before treatment and 4 hours after the first injection of pegIFN-α2b [389]. Unexpectedly, in 6 patients we found an upregulation of hundreds of ISGs already in pretreatment biopsies. In these "preactivated" patients, we found no significant further increase in the number or expression level of ISGs induced by pegIFN-α2b and no increase in the p-STAT1 nuclear signal intensity. Apparently, the constitutive activation of the endogenous IFN system not only fails to eliminate the virus, but also inhibits Jak/-STAT signaling and thereby inhibits a response to pegIFN-α2b treatments [389]. In a follow-up study, we developed and validated a classifier based on the expression of four genes in the liver that allows one to predict a response to (peg)IFN-α in individual patients [393]. In the present study, we made use of this classifier to screen patients with CHC and included 12 patients who did not have an activated endogenous IFN system in the liver. This selection process allowed us to exclude patients with refractory Jak/STAT signaling pathways. The results from our present study most likely reflect IFN responses in general, because the patients with CHC who were included had normal responsiveness of IFN-α signaling pathways in the liver. Ever since the introduction of pegIFN-α into therapeutic regimens for CHC, the prevailing explanation for the superior efficacy of pegIFN-α compared with that of conventional IFN-α was centered on the prolonged high serum concentration of pegIFN-α molecules. In this paradigm, the permanently high serum levels of pegIFN-α were equated with a permanent stimulation of the target cells, i.e., the infected hepatocytes. The inferior efficacy of IFN-α was explained

by the short serum half-life that caused serum levels to return to baseline in the second half of the 2-day dosing interval, leaving the infected hepatocytes unstimulated and thereby enabling a periodic resurgence of HCV replication. Based on the results of our present study, we refute this model. Our data show an activation of the Jak/STAT pathway in the liver only during the first day, despite prolonged high serum concentrations of pegIFN-α2b. This finding is in agreement with experimental data from studies in chimpanzees that showed only transient induction of ISGs after pegIFN-α2a injection [409]. The molecular mechanisms that temporally limit IFN-α signaling most likely involve two negative regulators of IFN-α–induced Jak/STAT signaling. We observed in our study that within hours after pegIFN-α2b injection, SOCS1 and USP18 were induced in the liver, and USP18 remained strongly upregulated during the entire 1-week dosing interval. Solid evidence from experiments with genetically modified mice shows that SOCS1 and USP18 have a central role in inhibiting IFN-α–induced Jak/STAT signaling [382, 410, 411]. We therefore conclude that SOCS1 and USP18 upregulation in the liver of patients treated with pegIFN-α2b restricts Jak/STAT signaling during the first day of the 1-week dosing interval. This conclusion is further supported by the fact that we did not observe a significant increase in the number or the expression level of ISGs induced at the 144-hour time point in patients treated with pegIFN-α2a compared with those treated with pegIFN-α2b, despite the high serum concentrations of pegIFN-α2a in all 3 patients. The refractoriness of Jak/STAT signaling pathways in the liver apparently overrides the potential benefits of the prolonged serum half-life of pegIFN-α2a. Indeed, in a large clinical study, pegIFN-α2a had no superior antiviral efficacy compared with that of pegIFN-α2b [412]. Taken together, the assumption that increasing the serum half-life of IFN-α formulations necessarily improves their antiviral efficacy because of an uninterrupted stimulation of IFN-α responses in hepatocytes cannot be sustained. The comparison of the gene sets induced by conventional IFN-α versus pegIFN-α supports a different mechanism: pegIFN-α induces a more sustained upregulation of a set of genes involved in cellular immune responses. The superior antiviral efficacy is most likely caused by an increased stimulation of the cellular immune response to HCV. It remains to be clarified which immune cells are critically involved in pegIFN-α–induced antiviral activities. It also remains to be clarified why pegIFN-α can induce this broader set of genes. For the 75 genes found in the intersection of conventional and pegIFN-α (Figure 5.5A), the magnitude of mRNA expression and the fold induction over baseline were equal for both IFN-α and pegIFN-α. Therefore, for the induction of those classical ISGs in hepatocytes, IFN-α is not less potent compared with pegIFN-α. However, IFN-α's short half-life of 6 to 8 hours might become important for nonparenchymal cells, i.e., liver-resident immune cells that were found to be responsive to pegIFN-α during the entire week after pegIFN-α injection. Based on our data, we propose a model in which the superior antiviral efficacy of pegIFN-α is the result of continuous stimulation of immune cells and is not due to continuous stimulation of the Jak/STAT pathway in HCV-infected hepatocytes. A large body of fundamental knowledge

about the key signaling pathways and the biological role of IFN-α has been acquired through cell culture experiments and mouse models with genetic deletions of IFNs, IFN receptors, or components of the Jak/STAT pathway [413]. On the other hand, the IFN-α–induced effects in target organs of human pathogens have been little investigated. Not surprisingly, the molecular mechanisms responsible for the antiviral activity of (peg)IFN-α against HCV are still not known. In the Huh7 cell–based HCV replicon system, overexpression and siRNA interference screens identified several ISGs involved in the inhibition of replication, among them: IRF1, IRF2, IRF7, IFN-induced helicase C domain–containing protein 1 (IFIH1, also known as MDA5), retinoic acid–inducible gene 1 (RIGI, also known as DDX58), mitogen-activated protein kinase kinase kinase 14 (MAP3K14), IFN-induced protein with tetratricopeptide repeats 3 (IFIT3), IFN-induced transmembrane protein 1 (IFITM1), IFITM3, phospholipid scramblase 1 (PLSCR1), TRIM14, RNASEL, and inducible nitric oxide synthase (INOS, also known as NOS2) [414, 415]. With the exception of MAP3K14 and NOS2, all of these ISGs were indeed upregulated by pegIFN-α2b in the liver and can be considered *bona fide* candidate antiviral effector genes. However, IRF1, IRF2, and IRF7 are transcription factors, and IFIH1 (MDA5), RIGI, IFIT3, and TRIM14 are involved in sensory pathways that activate IFN-β in infected cells [416–418]. These seven ISGs are most likely not direct-acting antiviral effector proteins. IFITM1 has been recently shown to be a tight-junction protein expressed in hepatocytes and has been found to inhibit HCV entry [419]. IFITM3 is an important restriction factor for the influenza virus and also acts through inhibition of cell entry [420]. PLSCR1 restricts RNA viruses, probably by enhancing the induction of a subset of ISGs including IFIT1 and IFIT2, two antiviral effectors that inhibit translation at the ribosome by binding to eIF3 [421]. Finally, RNaseL is a nonspecific antiviral effector that degrades viral and host RNAs upon activation by 2'-5'oligoadenylates [379]. Based on their proven direct antiviral effector functions, their identification in the above-mentioned siRNA interference and overexpression screens [414, 415], and their pegIFN-α2b–induced upregulation in the liver, IFITM1, IFITM3, PLSCR1, and RNaseL are prime candidates for anti-HCV effectors in humans. Most likely, however, many more of the upregulated ISGs are involved in an orchestrated antiviral effector program that can eliminate HCV from chronically infected patients. On a more fundamental level, our study also provides for the first time important insights into how IFN-α regulates gene induction over a prolonged observation period of 1 week. Our analysis of global gene expression data obtained from biopsies performed at five time points up to 6 days after pegIFN-α2b injection with an unbiased mathematical model, using an infinite Gaussian mixture model with a Dirichlet process prior, produced four robust clusters of upregulated ISGs with distinct kinetic patterns. Surprisingly, the ISRE promoter element was by far the most important TFBS motif in all the clusters. This clearly demonstrates that ISGs with late or delayed maximal expression are not induced by a different set of transcription factors that could be upregulated by the primary IFN-α–induced transcription factors ISGF3 and STAT1 homodimers and that

could then stimulate a second (and third) wave of gene transcription. Based on our nuclear run-on assays and mRNA decay rate measurements in cell culture experiments, we propose that a different duration of gene transcription as well as a different mRNA stability are responsible for the distinct kinetic expression profiles of ISG clusters. We examined the relative contribution of transcription factor–binding motifs to the global gene expression by MARA, which revealed ISRE to be the most significantly changing motif across all time points up to 6 days. ISRE motifs are the binding sites for ISGF3 and also IRFs. ISGF3 is activated by IFN-α–induced phosphorylation of STAT1 and STAT2. In hepatocytes, signaling through the Jak/STAT pathway becomes refractory within the first day after injection. We observed that the ISRE motif activity indeed peaked at the 4-hour and 16-hour time points, but remained increased even at later time points (Figure 5.6B). The persistent activation of ISRE sites might be caused by ongoing activation of ISGF3 in hepatocytes at lower levels that are not readily detectable by p-STAT1 immunoblotting. Alternatively, it might reflect the persistent activation of the Jak/STAT pathway in nonparenchymal cells that do not become refractory. Persistent ISRE motif activity could also be driven by IRF7. IRF7 mRNA was induced during the entire 1-week dosing interval of pegIFN-α (Figure 5.6C), and it is likely that IRF7 protein was upregulated as well. However, the transcriptional activity of IRF7 is tightly regulated by serine phosphorylation by IKK-α, a downstream component of cellular sensory pathways that are activated by viral PAMPs [398, 399, 422, 423]. IRF7-mediated gene induction would occur only in HCV-infected cells. The strong upregulation of IFI27 mRNA in more than 90% of hepatocytes at the 96-hour time point (Figure 5.2A) is not likely to be caused by activated IRF7, because HCV rarely infects more than 50% of hepatocytes [424], although we cannot rule out an alternative activation of IRF7 in uninfected cells in the context of IFN treatment. Finally, our work also sheds light on the role of U-STAT1 as a transcriptional activator that has been proposed to be important in prolonging IFN-α–induced gene transcription. We hypothesized that U-STAT1 target genes would also be strongly expressed at later time points during the 1-week pegIFN-α2b dosing interval, because U-STAT1 was indeed strongly upregulated during the entire week after pegIFN-α2b injection, whereas STAT1 phosphorylation occurred only during the first day. However, the U-STAT1 target genes identified by Cheon and Stark [400] had expression kinetics not different from other p-STAT1–driven ISGs. We therefore addressed the transcriptional activity of U-STAT1 on a STAT1-null background by expressing a mutant tyrosine 701 full-length STAT1 in U3A cells that lack STAT1. Despite very high expression levels, U-STAT1 did not induce ISGs in these cells. We conclude that in cells that lack a WT STAT1, U-STAT1 cannot induce ISG transcription. These findings do not support a role for U-STAT1 in prolonging pegIFN-α–induced gene transcription in the liver. Nevertheless, we would like to point out that these findings might be specific for U3A cells, and therefore we cannot formally exclude that U-STAT1 is driving gene transcription in human hepatocytes. In conclusion, pegIFN-α induces a transient activation of Jak/STAT signaling in

hepatocytes that is terminated by the prolonged upregulation of USP18. The predominant transcription factor is ISGF3. Hundreds of genes were induced and can be classified into four robust clusters with distinct kinetic expression patterns. ISGs with peak expression levels at later time points were not induced by secondary transcription factors, and we could not substantiate a role for U-STAT1 in prolonged ISG induction. Our data do not support the prevailing explanation for the superior antiviral efficacy of pegylated versus conventional IFN-α, i.e., that the constantly high serum levels of pegIFN-α cause permanent stimulation of the IFN signal transduction pathways and prolonged IFN-stimulated gene expression in infected hepatocytes. Rather, we found that pegIFN-α induced a broader range of genes, including many genes involved in cellular immune responses. The prolonged serum half-life of pegIFN-α permits a continuous stimulation of nonparenchymal cells in the liver which, contrary to hepatocytes, do not become refractory, but remain sensitive to pegIFN-α during the entire 1-week dosing interval. We therefore propose that the superior efficacy of pegIFN-α is caused by an indirect mechanism involving infiltrating or liver-resident immune cells.

5.5 METHODS

5.5.1 *Patients*

The patients were recruited between March 2006 and April 2010 at the Hepatology Outpatient Clinic of the University Hospital Basel. Patients with CHC who underwent a biopsy for diagnostic purposes (B1) and provided written informed consent were screened for hepatic ISG expression. The four-gene classifier was used to assess the probability of an SVR [393]. Patients with a high probability of achieving an SVR were asked to participate in the study, which included a second biopsy (B2) taken at a particular time point after the first therapeutic injection of pegIFN-α. We included 3 patients for each of the following time points: 16, 48, 96, and 144 hours, and additionally, the analysis included data on 6 patients from a previous study who had a biopsy at 4 hours (patients 1, 2, 6, 7, 8, and 9) [389]. The patients received 1.5 μg/kg body weight pegIFN-α2b (Essex Chemie). Weight-adjusted ribavirin treatment was initiated only after the second biopsy to avoid confounding effects. An additional 3 patients treated with 180 μg of pegIFN-α2a (Roche) were included for the 144-hour time point study. Blood for serum analysis was taken at the time of the first and second biopsies. Serum HCV RNA was quantified using the COBAS AmpliPrep/COBAS TaqMan HCV Test and the COBAS AMPLICOR Monitor (Roche Molecular Systems). Details regarding the 6 patients treated with IFN-α2a have been described previously [394]. The patients included in the present analysis correspond to patients 2–7 in the original publication [394].

5.5.2 *IL28B genotyping*

DNA extraction and genotyping for the single nucleotide polymorphism rs12979860 near the IL28B gene were performed as described previously [393].

5.5.3 *Measurement of serum proteins*

Serum was collected before the first injection of pegIFN-α and at the time of the second biopsy. Serum levels of IFN-α2b and pegIFN-α2a were measured with an ELISA kit (Verikine 41100; PBL InterferonSource). Standard curves were prepared separately for pegIFN-α2a and -2b by a serial dilution starting at 12.5 pg/ml. The patient serum samples were diluted 10 times in sample diluent. IP-10 serum levels were measured with an ELISA (BD OptEIA Set Human IP-10, 2732KI; BD Biosciences) according to the manufacturer's instructions.

5.5.4 *IHC*

Four-micrometer-thick serial sections were cut from formalin-fixed, paraffin-embedded liver biopsy specimens, rehydrated, pretreated for 20 minutes in ER2 solution, incubated with a monoclonal rabbit antibody against p-STAT1 (dilution 1:200, no. 9167; Cell Signaling Technology) or USP18 (1:100, catalog 4813; Cell Signaling Technology), and counterstained with hematoxylin. Standard indirect immunoperoxidase procedures were used for IHC (ABC-Elite; Vectra Laboratories). The staining procedure was performed with an automated stainer (Bond; Vision BioSystems).

5.5.5 *RNA extraction and microarray hybridization*

Total RNA was extracted from human liver tissue using QIAzol reagent and the RNeasy Mini Kit (QIAGEN) according to the manufacturer's instructions. Gene expression was assessed by microarray analysis using Affymetrix Human Genome U133 Plus 2.0 arrays. Total RNA (1 μg) from each sample was reverse transcribed using a Genechip 3'IVT Express Kit (Affymetrix) according to the manufacturer's instructions. The Hybridization and Wash Kit (Affymetrix) was used to hybridize the samples. All original array data are deposited in the NCBI's Gene Expression Omnibus (GEO GSE48445).

5.5.6 *RNA ISH*

For the present study, we adapted a highly sensitive and specific ISH system (QuantiGene ViewRNA; Affymetrix). OCT-embedded and shock-frozen biopsies were cryosectioned (10-μm-thick sections) in a cryostat and mounted on Superfrost Plus Gold glass slides (Thermo Fisher Scientific).

Upon fixation (4% formaldehyde, 16–18 hours at 4°C), washing, and dehydration in ethanol, the sections were pretreated by boiling for 1 minute in Pretreatment Solution, followed by a 10-minute digestion in Protease QF (both from Affymetrix). Sections were hybridized for 2 hours at 40°C with QuantiGene ViewRNA probes against MX1, IFI27, SOCS1, and PDL1 (Affymetrix). Bound probes were preamplified and subsequently amplified according to the manufacturer's instructions. Labeled oligonucleotide probes conjugated with alkaline phosphatase (LP-AP) type 1 or type 6 were added, followed by the addition of fast red or fast blue substrate used to detect ISG mRNAs. Finally, the slides were counterstained with Meyer's hematoxylin and embedded with DAPI-containing aqueous mounting medium (Roti-Mount FluorCare DAPI; Roth). Random images were acquired using a laser scanning confocal microscope (LSM710; Zeiss) and Zen2 software (Zeiss). All images were acquired with identical settings and saved in the Zeiss confocal file format (.lsm).

5.5.7 *MARA*

Here, we provide a brief description of MARA and its particular use in this work. For a detailed description of the general approach, the reader is referred to the FANTOM Consortium study [50]. To model the activity $A_{m,s}$ of a motif m in sample s, MARA uses a simple linear model that relates the number of binding sites $N_{p,m}$ in promoter p, for each of a large number of regulatory motifs m, to the expression $E_{p,s}$ of promoter p in samples:

$$E_{p,s} = \tilde{c}_s + c_p + \sum_m N_{p,m} A_{m,s}, \tag{5.1}$$

where c_s represents the mean expression in sample s, and c_p is the basal expression of promoter p. To determine promoter expression levels, we first computed the transcript expression levels by averaging weighted probeset signals (preprocessed as described above) over all probesets that matched a particular transcript as annotated by Affymetrix (Affymetrix annotation, Release 31 [NM accession RefSeqs only]). In this averaging, a probeset's signal was weighted by the number of transcripts it matches. Subsequently, transcript expression levels were mapped to the human promoterome by averaging the weighted expression levels over all transcripts associated with a particular promoter. In this averaging, each transcript's signal was weighted by the inverse of the number of promoters that express this particular transcript. To predict the TFBSs in each promoter, we used a curated set of transcription factor–binding motifs from SwissRegulon [425] with minor changes: since the SwissRegulon IRF1,2,7 motif only covered the IRF core consensus sequence, we replaced it with the ISRE (ThioMac-LPS-exp) motif to better cover the ISGF3 and IRF9 binding sites in our analysis. In addition, we replaced the three highly redundant STAT motifs from Swiss-Regulon with a single, high-quality GAS motif (HelaS3-STAT1-ChIP-Seq, as the GAS motif representative). Both motifs were obtained with HOMER software [397]. By applying MARA, we obtained for every motif m in each sample s an expected activity $A_{m,s}$ and a corresponding error $\sigma_{m,s}$.

5.5.8 *Determining donor-specific motif activity changes due to IFN-α treatment*

For every donor and every motif m, we calculated the difference between the motif activity before IFN-α treatment (A_m^b) and the motif activity after IFN-α treatment (A_m^a):

$$A_m^\Delta = A_m^a - A_m^b \tag{5.2}$$

as well as the corresponding error:

$$\sigma_m^\Delta = \sqrt{(\sigma_m^a)^2 + (\sigma_m^b)^2}. \tag{5.3}$$

Thus, for every motif m and each donor d, the expression data D imply an expected activity change $A_{m,d}^\Delta$ with corresponding error $\sigma_{m,d}^\Delta$. Consequently, the probability of the data D, assuming a true (unobserved) activity change $\tilde{A}_{m,d}^\Delta$, is a Gaussian with the expected mean $A_{m,d}^\Delta$ and error $\sigma_{m,d}^\Delta$:

$$P(D|\tilde{A}_{m,d}^\Delta) = \frac{1}{\sqrt{2\pi}\sigma_{m,d}^\Delta} \exp\left[-\frac{1}{2}\frac{\left(\tilde{A}_{m,d}^\Delta - A_{m,d}^\Delta\right)^2}{(\sigma_{m,d}^\Delta)^2}\right]. \tag{5.4}$$

5.5.9 *Determining mean motif activity changes due to pegIFN-α treatment at certain time points*

To obtain mean activity changes for every group of donors g∈G whose second biopsy was taken at an equal time point after pegIFN-α treatment, we assumed that the activity changes A_m^Δ of motif m were Gaussian distributed (with mean $A_{m,g}^\Delta$ and variance $(\sigma_{m,d}^\Delta)^2$). Accordingly, the probability of an activity change $\tilde{A}_{m,d}^\Delta$ in donor d is:

$$P(\tilde{A}_{m,d}^\Delta|A_{m,g}^\Delta, \sigma_{m,d}^\Delta) = \frac{1}{\sqrt{2\pi}\sigma_{m,d}^\Delta} \exp\left[-\frac{1}{2}\frac{\left(\tilde{A}_{m,d}^\Delta - A_{m,g}^\Delta\right)^2}{(\sigma_{m,d}^\Delta)^2}\right]. \tag{5.5}$$

For each donor d∈g, we combine Equations 5.4 and 5.5 and integrate out all unknown (true activity changes) $\tilde{A}_{m,d}^\Delta$ so that we can calculate the probability of the data D given the mean activity of the group $A_{m,g}^\Delta$ and the corresponding error $\sigma_{m,g}^\Delta$:

$$P(D|A_{m,g}^\Delta, \sigma_{m,g}^\Delta) = \prod_{d\in g}\left[\int_{-\infty}^{\infty} P(D|\tilde{A}_{m,d}^\Delta) P(\tilde{A}_{m,d}^\Delta|A_{m,g}^\Delta, \sigma_{m,g}^\Delta)\mathbf{d}\tilde{A}_{m,d}^\Delta\right]. \tag{5.6}$$

Solving these integrals analytically gives:

$$P(D|A_{m,g}^\Delta, \sigma_{m,g}^\Delta) = \prod_{d\in g} \frac{1}{\sqrt{2\pi((\sigma_{m,g}^\Delta)^2 + (\sigma_{m,d}^\Delta)^2)}} \exp\left[-\frac{(A_{m,d}^\Delta - A_{m,g}^\Delta)^2}{2((\sigma_{m,g}^\Delta)^2 + (\sigma_{m,d}^\Delta)^2)}\right]. \tag{5.7}$$

We then numerically determine the value $\sigma^{\Delta*}_{m,g}$ that maximizes Equation 5.7. Assuming a uniform prior for $A^{\Delta}_{m,g}$, we obtain an expression for the posterior probability $P(A^{\Delta}_{m,g}|D)$, which is a Gaussian with mean

$$\bar{A}^{\Delta}_{m,g} = \frac{\sum_{d\in g} \frac{A^{\Delta}_{m,d}}{(\sigma^{\Delta*}_{m,g})^2+(\sigma^{\Delta}_{m,d})^2}}{\sum_{d\in g} \frac{1}{(\sigma^{\Delta*}_{m,g})^2+(\sigma^{\Delta}_{m,d})^2}}, \tag{5.8}$$

and error

$$\bar{\sigma}^{\Delta}_{m,g} = \sqrt{\frac{1}{\sum_{d\in g} \frac{1}{(\sigma^{\Delta*}_{m,g})^2+(\sigma^{\Delta}_{m,d})^2}}}, \tag{5.9}$$

where $\sigma^{\Delta*}_{m,g}$ is the maximum likelihood estimate of Equation 5.9. We call $\bar{A}^{\Delta}_{m,g}$ the mean activity change for group g and $\bar{\sigma}^{\Delta}_{m,g}$ the corresponding error. To obtain a measure for the significance of the mean activity change for group g, we calculate a corresponding z value:

$$\bar{z}^{\Delta}_{m,g} = \sqrt{\frac{\bar{A}^{\Delta}_{m,g}}{\bar{\sigma}^{\Delta}_{m,g}}}, \tag{5.10}$$

Finally, a global z value considering all time points is given by:

$$\bar{z}^{\Delta}_{m} = \sqrt{\frac{\sum_{g\in G} \left(\bar{z}^{\Delta}_{m,g}\right)^2}{|G|}}, \tag{5.11}$$

To calculate a P value for a calculated z value $\bar{z}^{\Delta}_{m}$, we used the null hypothesis that the z statistic $\bar{z}^{\Delta}_{m,g}$ in each group g, i.e., the ratio between the motif activity change and its error, was drawn from a Gaussian with mean zero and variance 1. Under this null hypothesis, the distribution of the statistic,

$$(\bar{z}^{\Delta}_{m})^2 * G = \sum_{g=1}^{G} (\bar{z}^{\Delta}_{m,g})^2 \tag{5.12}$$

with G representing the number of groups (wherein each group represents a time point), is Gamma-distributed, and we used this distribution to calculate the P value corresponding to the z value of each of our motifs.

5.5.10 *TFBS analysis*

TFBS analysis was carried out using HOMER software [397] (http://biowhat.ucsd.edu/homer/motif/index.html). Briefly, promoter regions (2 kbp upstream and 500 bp downstream of the transcription start site) of all genes within each cluster were screened for known TFBS. Enrichment of TFBS in our gene lists relative to all human promoter regions was assessed by hypergeometric tests.

5.5.11 *Quantitative real-time RT-PCR*

RNA was reverse transcribed by Moloney murine leukemia virus reverse transcriptase (Promega) in the presence of random primers (Promega) and deoxynucleoside triphosphate. The samples were incubated for 5 minutes at 70°C and then for 1 hour at 37°C. The reaction was stopped by heating at 95°C for 5 minutes. SYBR real-time PCR was performed using the SYBR Green PCR Master Mix (Applied Biosystems). Intron-spanning primers for GAPDH, HERC6, IFI27, IFI44L, ISG15, LGALS3BP, MX1, OAS1, OAS2, RSAD2, and USP18 were used (Supplemental Table C.5). All reactions were performed in duplicate on an ABI 7500 Real-Time PCR System (Applied Biosystems). mRNA expression levels of the transcripts were normalized to GAPDH using the ΔCt method.

5.5.12 *Western blot analysis*

Whole-cell extracts and blotting of human liver samples and cells were performed as described [389]. The membranes were incubated with primary antibodies against p-STAT1 (1:1,000, catalog 9171; Cell Signaling Technology), STAT1 (1:1,000, catalog 610116; BD Transduction Laboratories), USP18 (1:1,000, no. 4813; Cell Signaling Technology), and β-actin (1:2,000, A5441; Sigma-Aldrich) diluted in Tris-buffered saline containing Tween-20 (TBST) overnight at 4°C. After three washes with TBST, membranes were incubated for 1 hour at room temperature with fluorescent secondary goat anti-mouse (IRDye 680) or anti-rabbit (IRDye 800) antibodies (both from LI-COR Biosciences). Blots were scanned using the Odyssey Infrared Imaging System (LI-COR Biosciences).

5.5.13 *Cell culture*

Huh7 cells were maintained in DMEM (Gibco) supplemented with 10% FBS. 2fTGH and U3A STAT1$^{-/-}$ cells were maintained in DMEM with 10% FBS and 250 μg/ml of hygromycin B (Sigma-Aldrich). The stably transfected U3A STAT1$^{-/-}$ cells were selected with 800 μg/ml of G418 (catalog 345810; Calbiochem). Cells were treated with 1,000 U/ml human IFN-α (Roferon; Roche) and/or with 5 μg/ml actinomycin D (Sigma-Aldrich).

5.5.14 *Site-directed mutagenesis and transfection*

The STAT1-flag-pcDNA3 (STAT1-WT) was provided by J.E. Darnell (Rockefeller University, New York City, New York, USA). STAT1 (Y701F)-flag-pcDNA3 was generated from STAT1-flag-pcDNA3 using the method described by Mikaelian and Sergeant [426]. Briefly, two consecutive PCR reactions with 30 cycles were performed using 20 ng of template DNA, 200 μM dNTP, 1 U of *Pfu* DNA polymerase (Promega), and 5 μM of each of the following primers: 5'-CTGGCACCAGAACGAATGA-3'; 5'-

ATTTAGGTGACACTATAG-3'; 5'-GGAACTGGATTCATCAAGACTGAG-3'; and 5'-CTCAGTCTTGATGAATCCAGTTC-3', in a final volume of 25 μl. The amplified products were loaded on a 1.5% agarose gel, excised, digested by BlpI and ApaI, and ligated into STAT1-flag-pcDNA3, previously cut with the same restriction enzymes. Mutation of Tyr701 to Phe was confirmed by sequencing. U3A STAT1$^{-/-}$ cells were transfected with 1 ÎŒg of the respective plasmid using Fugene HD (Roche) according to the manufacturer's instructions. Cells were selected with 800 μg/ml of hygromycin B (Roche) for 15 days, and single clones were chosen.

5.5.15 *Nuclear run-on assay*

After IFN-α treatment, cells were washed with 1$\times$ PBS, treated with 0.25% trypsin for 4 minutes (Gibco), suspended in 10 ml of ice-cold diethylpyrocarbonate-treated (DEPC-treated) 1$\times$ PBS, and concentrated by centrifugation (160 g, 10 minutes). Cells were then washed once with ice-cold buffer 1 containing 10 mM Tris-HCl (pH 7.4), 150 mM KCl, and 8 mM Mg acetate, centrifuged at 530 g for 10 minutes, and subsequently lysed with buffer 1 with the addition of 0.5% Igepal (Sigma-Aldrich) for 10 minutes at 4°C. Nuclei were then isolated by a sucrose gradient (600 mM), then washed and suspended in buffer containing 40% glycerol. Nuclei were immediately used for the run-on assay. Nuclei (5×10^6) were incubated in reaction buffer containing 5 mM Tris-HCl (pH 8.0), 2.5 mM MgCl 2, 150 mM KCl, and 2.5 mM each of ATP, GTP, CTP, UTP, and biotin-16-UTP (Roche) for 45 minutes at 30°C. RNA was then isolated with TRIzol according to the manufacturer's instructions. Subsequently, biotinylated RNA was purified with streptavidin-coupled beads (Dynabeads M-280; Invitrogen) according to the manufacturer's instructions, and RNA was again isolated with TRIzol.

5.5.16 *Statistics*

Microarray analysis was performed with Bioconductor packages within the R statistical environment [427]. Data were preprocessed using the standard RMA algorithm. Batch effects observed between the human liver samples that were processed and hybridized at different times were corrected using the ComBat algorithm [428]. Probesets with very low expression intensities (below 80 in the highest-expressing sample) as well as the control probesets were excluded from the subsequent analyses. The list of significantly regulated probesets was compiled as follows: (a) probesets showing more than a 2-fold difference in levels between the B1 and B2 samples taken from the same patient were selected; (b) for every time point, the probesets that changed in two-thirds of the patients were retained. Probesets fulfilling those criteria were included in the clustering analysis. The expression data were normalized so that total expression levels did not affect the grouping of the probesets. An infinite Gaussian mixture model with a Dirichlet process prior was used to produce the gene clusters. This nonparametric model suggests

a growing number of Gaussians to describe the gene expressions. With the special choice of a Dirichlet process prior, the number of clusters need not be fixed in advance, but is adaptively chosen based on the observed data. The results were tested for robustness by moderately changing the hyperparameters that control the Dirichlet process. Enrichment of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and GO biological process terms were assessed using DAVID software, version 6.7. Additional statistical analyses using a 2-tailed Student's t test were carried out using GraphPad Prism software, version 6.0 (GraphPad Software). A P value of less than 0.05 was considered significant.

5.5.17 *Study approval*

All patients provided written informed consent to participate in the study, which was approved by the ethics committee of Basel.

5.6 AUTHORS INFORMATION

5.6.1 *List of authors*

The following authors have contributed to the work discussed in Chapter 5:

1. Michael T. Dill[1,2] (Abbr.: MTD),
2. Zuzanna Makowska[1] (Abbr.: ZM),
3. Gaia Trincucci[1] (Abbr.: GT),
4. Andreas Johannes Gruber[3] (Abbr.: AJG),
5. Julia E. Vogt[4] (Abbr.: JEV),
6. Magdalena Filipowicz[1,2] (Abbr.: MF),
7. Diego Calabrese[1] (Abbr.: DC),
8. Ilona Krol[1] (Abbr.: IK),
9. Daryl T. Lau[5] (Abbr.: DTL),
10. Luigi Terracciano[6] (Abbr.: LT),
11. Erik van Nimwegen[3] (Abbr.: EvN),
12. Volker Roth[4] (Abbr.: VR) &
13. Markus H. Heim[1,2] (Abbr.: MHH),

whereat author affiliations are as follows:

1 Department of Biomedicine, Hepatology Laboratory, University of Basel, Basel, Switzerland

2 Division of Gastroenterology and Hepatology, University Hospital Basel, Basel, Switzerland

3 Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Basel, Switzerland

4 Computer Science Department, University of Basel, Basel, Switzerland

5 Liver Center, Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts, USA

6 Institute of Pathology, University Hospital Basel, Basel, Switzerland

5.6.2 *Author contributions*

The listing of authors in the previous subsection (5.6.1) was performed according to the authors' contributions, whereat the first author (MTD) contributed most and subsequent authors decreasingly. However, the last three authors are principal investigators and thus their listing follows the opposite ranking (the last contributed the most and the preceding two authors decreasingly).
In detail, using the abbreviations specified in the previous subsection (i.e. 5.6.1) AJG contributed with the following work: AJG performed the MARA analysis described in section 5.3.6 with exception of the P value calculation as described in methods section 5.5.9. Moreover AJG performed the microRNA binding site analysis of the four ISG clusters discussed in section 5.3.8.

5.7 SUPPLEMENTARY MATERIALS

Supplementary materials can be found in Appendix C.

5.8 ACKNOWLEDGMENTS

We thank the patients who donated liver biopsy specimens for this study. We are grateful to Francois H.T. Duong for helpful discussions and Philippe Demougin (Life Sciences Training Facility, Pharmazentrum, Basel, Switzerland) for providing technical help with the processing of microarrays. We also thank Sylvia Ketterer for excellent technical assistance.

5.9 FUNDING

Using the abbreviations specified in subsection 5.6.1: This work was supported by Swiss National Science Foundation (SNF) grant 320030-130243 (to MHH), SNF grant 323500-123714 (to MTD), and a Werner Siemens Fellowship (to AJG).

6

GLOBAL 3' UTR SHORTENING HAS A LIMITED EFFECT ON PROTEIN ABUNDANCE IN PROLIFERATING T CELLS

The work discussed in this chapter was conducted in collaboration with the van Nimwegen and the Pieters labs and published in Nature Communications in 2014 (see reference [429]).

6.1 ABSTRACT

Alternative polyadenylation is a cellular mechanism that generates messenger RNA (mRNA) isoforms differing in their 3' untranslated regions (3' UTRs). Changes in polyadenylation site usage have been described upon induction of proliferation in resting cells, but the underlying mechanism and functional significance of this phenomenon remain largely unknown. To understand the functional consequences of shortened 3' UTR isoforms in a physiological setting, we used 3' end sequencing and quantitative mass spectrometry to determine polyadenylation site usage, mRNA and protein levels in murine and human naive and activated T cells. Although 3' UTR shortening in proliferating cells is conserved between human and mouse, orthologous genes do not exhibit similar expression of alternative 3' UTR isoforms. We generally find that 3' UTR shortening is not accompanied by a corresponding change in mRNA and protein levels. This suggests that although 3' UTR shortening may lead to changes in the RNA-binding protein interactome, it has limited effects on protein output.

6.2 INTRODUCTION

Expression of messenger RNA (mRNA) precursors transcribed by RNA polymerase II requires recognition and processing of signals in the pre-mRNA by the cleavage and polyadenylation factors to guide proper formation of 3' ends. Most mammalian genes have multiple polyadenylation (poly(A)) sites [97, 98], whose regulated selection leads to the production of alternative mRNA forms that differ in localization, stability and/or protein-coding potential. A systematic shift towards coding region-proximal 3' end processing sites, leading to an overall shortening of 3' untranslated regions (3' UTRs) was recently observed in activated compared with naive lymphocytes [102] as well as in cells that proliferate rapidly [100, 101]. The differentiation of embryonic stem cells and, conversely, the induction of pluripotency in somatic cells, are associated with changes in opposite directions in 3' UTR lengths [97, 430, 431]. The functional significance of this regulation is not well understood. Initial studies suggested that the lack of microRNA (miRNA)-binding sites in the shortened 3' UTRs leads to an increased stability of the mRNAs and an increased protein output [101, 102]. This conjecture was later refuted by a transcriptome-wide analysis that was carried out in mouse embryonic fibroblasts, where small differences in the relative stability of 3' UTR iso-

forms were found [103]. MiRNAs are only one class of regulators that act on 3' UTRs, guiding the RNA-induced silencing complexes to target mRNAs to increase their decay rate and reduce translation [112, 432]. The 3' UTRs contain binding sites for many RNA-binding proteins (RBPs) and integrate a variety of signals for mRNA localization, decay and translation. Of the hundreds of RBPs that bind and potentially regulate various aspects of mRNA metabolism [433, 434], at least some, such as the human antigen R (HuR) [435] and the A/U-rich-element-binding factor-1 (also known as the heterogeneous nuclear ribonucleoprotein D or hnRNPD) [436], have been reported to increase mRNA stability. A very recent study that was carried out in yeast [437] found that the stability of transcripts is not correlated with the length of their 3' UTRs. However, the protein that appeared to cause the largest difference in decay rates between 3' UTR isoforms in this study, Puf3, had an overall destabilizing effect (presumably on the longer 3' UTR isoforms, containing additional Puf3-binding sites compared with the shorter 3' UTR isoforms). Thus, the functional relevance of the observed systematic reduction in 3' UTR lengths in relation to cell proliferation remains unclear. To determine the consequences and functional relevance of 3' UTR shortening during lymphocyte activation, we undertook a systematic investigation of the changes in the poly(A) site usage and in the protein output of the corresponding genes in mouse and human T cells.

6.3 RESULTS

6.3.1 *Activated T cells express mRNAs with shortened 3' UTRs*

We focused our investigation on the T cell activation system, in which the 3' UTR shortening was initially described [102]. We dissected spleen and lymph nodes from C57BL/6 mice and isolated T cells by MACS purification. T cells were activated for 72 h with CD3/CD28 Dynabeads and IL-2 (see 6.5) and libraries of 3' ends of mature mRNAs were prepared and processed with the A-seq protocol as described previously [81]. Excluding reads that may result from internal priming, we identified 269,751 high-confidence poly(A) sites in the mouse genome, to which a total of 32,388,835 reads mapped (Supplementary Table D.1). For each library, more than 70% of the reads mapped to terminal exons of transcripts and only a small fraction (less than 0.5%) to other exons (Fig. 6.1A). Compared with resting T cells, the density of 3' end sequencing reads in terminal exons showed a clear and highly reproducible shift towards the 5' end of terminal exons in the activated T cells (Fig. 6.1B).

To further validate the quality and reproducibility of our results, we compared libraries on gene-by-gene basis using the number of reads mapped to terminal exons of transcripts assigned to a particular gene as a proxy for the expression level of a gene. Requiring a minimal expression level of five reads per million in at least one of the four libraries, we identified 9,928 genes as being expressed in naive and activated murine T cells. Biological replicates

of both naive and activated T cells showed very high correlation ($r \geq 0.94$; Supplementary Fig. D.1). The genes that were upregulated in activated T cells showed a clear enrichment of cell cycle-associated Gene Ontology (GO) terms, whereas immune system-related GO terms were most enriched among downregulated genes (Supplementary Table D.2). These results are consistent with the physiological state of the cells and further demonstrate that our 3' end sequencing data accurately reflect transcript-level changes. To investigate the dynamics of 3' end processing, we first clustered the poly(A) sites that were very closely spaced and probably the result of imprecise 3' end cleavage and identified distinct poly(A) sites [81]. The nucleotide distribution in regions flanking the inferred sites and the presence of upstream polyadenylation signals indicate that our strategy allowed us to identify genuine poly(A) sites (Supplementary Fig. D.2A,B). We then restricted our analysis to tandem poly(A) sites that were located in the same terminal exon, as was done in a previous study [102]. Overall, we inferred that 3,116 genes undergo alternative polyadenylation (APA) at tandem poly(A) sites (Supplementary Table D.3). The number of reads assigned to poly(A) sites in the 5' half and the 3' half of terminal exons showed good reproducibility between biological replicates indicating that our data can be used to analyse the relative use of alternative poly(A) sites in different conditions (Supplementary Fig. D.2C).

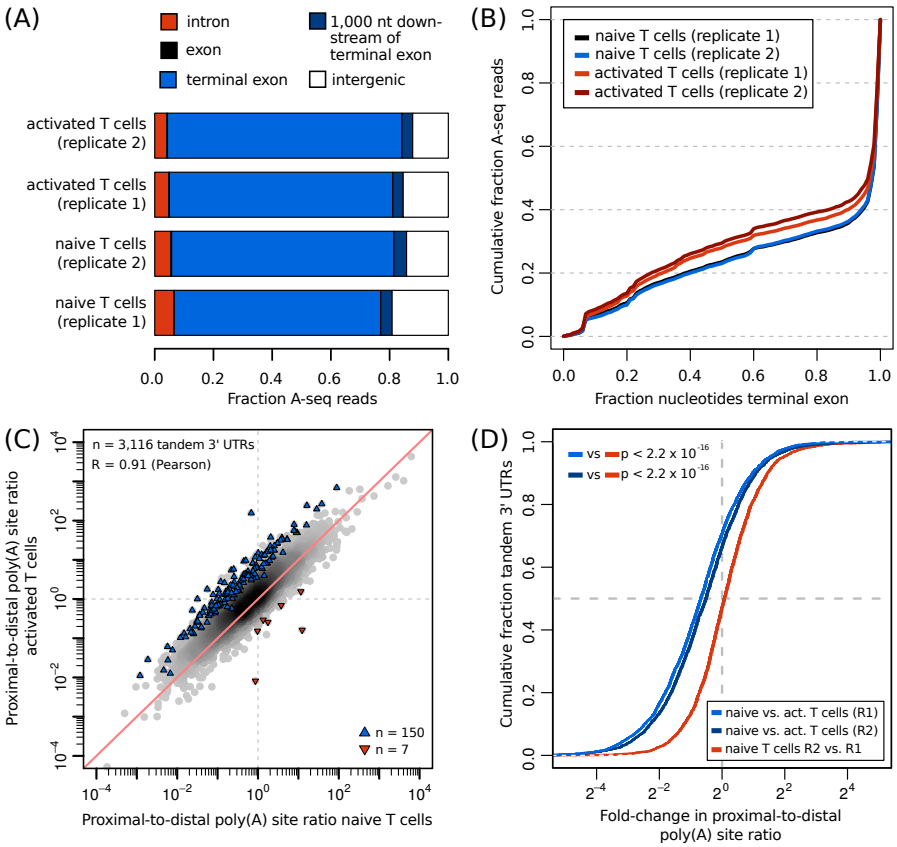


Figure 6.1: 3' end sequencing reveals increased proximal poly(A) site usage upon activation of murine T cells. (A) Annotation of reads obtained from mRNA 3' end sequencing of naive and activated T cells. Two biological replicates were used for each condition. **(B)** Coverage of terminal exons by 3' end sequencing reads as a function of the distance from the exon start. Activated T cells show increased coverage of the 5' compared with the 3' region of terminal exons, resulting from preferential use of proximal poly(A) sites. **(C)** Contour plot of the relative use of proximal and distal poly(A) sites in naive and activated T cells. Only genes with tandem poly(A) sites are shown. Genes that were identified to undergo significant changes in poly(A) site use are marked by coloured triangles with blue indicating increased use of the proximal poly(A) site and red indicating increased use of the distal poly(A) site in activated T cells. **(D)** Cumulative distribution function of the change in proximal vs distal poly(A) sites between pairs of samples (data shown in C). The shift towards increased usage of proximal poly(A) sites is statistically highly significant (P-values obtained by Mann–Whitney test on 3,116 genes with tandem poly(A) sites).

As expected from a previous study [102], there was a marked shift towards increased usage of proximal poly(A) sites in activated compared with naive T cells (Fig. 6.1C). The shift is not restricted to a small subset of genes, but affected the entire transcriptome (Fig. 6.1D). That is, more than 70% of genes showed an increased use of proximal poly(A) sites upon T cell activation. We used the DEXSeq software to analyse the differential use of poly(A) sites [438]. A total of 157 genes showed a significant difference (P-value$\leq$0.05) in the proximal-to-distal poly(A) site use between naive and activated T cells (Supplementary Table D.4), and 150 of these genes (96%) had an increased

abundance of proximal transcript isoforms in activated T cells. This set includes Bcl2, Creb1 and Tnfrsf9 (CD137), genes that are known to influence proliferation [439–441]. Whereas the increased proximal poly(A) site use of Bcl2 and Tnfrsf9 is associated with increased expression, consistent with the initial reports of the effect of 3' UTR shortening on gene expression, the expression of Creb1, is rather reduced, at least at the mRNA level. An example of a gene with a marked shift towards increased proximal poly(A) site use upon T cell activation is shown in Supplementary Fig. D.3A, which depicts a CLIPZ [442] genome browser screenshot of the 3' UTR of Reep5. It is well known that upon activation, T cells undergo a dramatic remodelling of the cytoskeleton [443]. Some of the genes that are involved in this process also show a significantly higher use of proximal poly(A) sites. These are Pak1 and Prkca (PKCa), which are involved in signalling transduction cascades, as well as Wasf2 (WAVE2) [444], Marcks [445] or Jmy [446], which interact directly with actin. Again in contrast to the expectation that the stability of short 3' UTR isoforms is higher compared with the long 3' UTR isoforms [101], all of these immune response-related genes are downregulated at the mRNA level, in spite of their increased use of proximal poly(A) sites. Moreover, when we analysed separately genes that are significantly downregulated or upregulated at the mRNA level upon T cell activation, we found that downregulated genes showed a more pronounced 3' UTR shortening than upregulated genes (Supplementary Fig. D.3B). This motivated us to investigate the relation between the change in poly(A) site use and the change in mRNA/protein abundance in more detail.

6.3.2 *Regulatory element content of 3' UTR isoforms*

Among the regulatory elements that are lost when proximal poly(A) sites are used more frequently, are binding sites for miRNAs that in naive T cells could contribute to the repression of gene expression [102]. To assess the consequence of 3' UTR shortening on the miRNA–mRNA interactome of T cells, we retrieved miRNA target predictions from the ElMMo database [255] and counted the number of target sites for each miRNA seed family in the common and alternative parts of the 3' UTRs of genes with tandem poly(A) sites (Supplementary Table D.5). MiRNA target sites that are located between the most proximal and most distal poly(A) site in the alternative 3' UTR region constitute a significant fraction of all predicted target sites. This is a reflection of the large change in 3' UTR length that is associated with T cell activation. For example, 65% of all the target sites predicted for the miR-29a seed family are located in the alternatively processed region of the 3' UTRs. To examine which miRNA regulators would be most affected by 3' UTR shortening, we carried out the following test. Each site predicted by ElMMo has an associated probability of being under evolutionary selection. By summing the probabilities of individual binding sites in a 3' UTR, we obtained an expected number of sites that are under selection. Performing this computation for individual transcript isoforms with their corresponding ex-

pression levels estimated based on 3' end sequencing, we obtained expected numbers of sites in the isoforms produced in a specific condition (activated and naive T cells). As expected from a global shortening of 3' UTRs, we find a net loss of target sites for all miRNAs. However, the loss of target sites by 3' UTR shortening does not affect all miRNAs to a similar degree. To identify which miRNAs would be most affected by the 3' UTR shortening, we randomized the predicted interactions involving the alternative parts of 3' UTRs. This amounts to randomizing the "labels" that indicate which miRNA binds an individual site. Computing the z-score of the observed change in the number of sites relative to what would be expected from the randomized data set, we found large differences between miRNAs. In particular, miRNAs that have been implicated in the regulation of cell proliferation appear at the extreme of the z-score range, some (miR-17, miR-365 and miR-135a [447–449]) losing more sites than expected and others (miR-26a, miR-103 [450, 451]) losing less sites than expected (Fig. 6.2A). MiRNAs with a cell type-specific expression show a less extreme pattern of site loss. These results indicate that APA at proximal poly(A) sites in proliferating cells does impact the susceptibility of the corresponding genes to regulation by miRNAs that themselves are involved in cell proliferation.

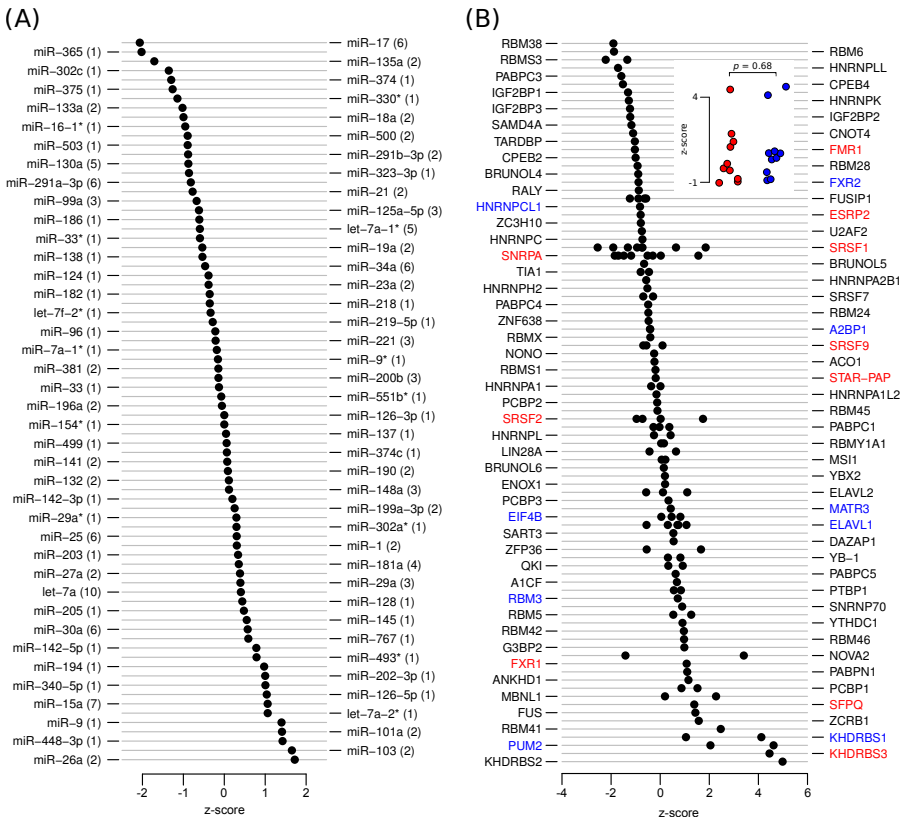


Figure 6.2: Quantification of the loss of regulatory elements upon 3' UTR shortening. Influence of 3' UTR shortening on miRNA **(A)** and RBP **(B)** target sites. The x axis represents the z-score of the loss of binding sites obtained by comparing the observed loss of target sites with what would be expected from random permutations of target sites across the set of alternative 3' UTRs. A negative z-score indicates that the loss in binding sites is greater than if sites were placed randomly in alternative regions of 3' UTRs. RBPs with a stabilizing effect on their transcript targets as assigned in Ray et al. [452] are marked in blue, whereas RBPs with a destabilizing effect are marked in red. The inset shows z-scores obtained for representative PWMs of RBPs with stabilizing and destabilizing effects (P-value obtained from a Wilcoxon rank sum test).

We similarly evaluated the change in the susceptibility of transcripts to regulation by RBPs. Binding motifs for a relatively large set of RBPs were recently published in the form of positional weight matrices (PWMs) [452]. Based on these PWMs and on the alignments of a number of genomes (see Methods), we predicted evolutionarily conserved RBP-binding sites in the 3' UTRs of expressed genes with a method that was introduced previously [290]. Applying the procedure that we described above for miRNAs starting from the inferred probabilities of RBP-binding sites to be under evolutionary selection, we determined which RBPs lose the most or least sites upon increased use of proximal poly(A) sites, compared with what we would expect by chance (Fig. 6.2B). As for miRNAs our analysis revealed large differences between individual RBPs. The expected impact of this regulation is, however, more complex because in contrast to miRNAs, for which the evidence for target destabilization is overwhelming [432], RBPs have a variety of functions. An

individual RBP frequently acts at multiple levels of gene regulation, including APA [453]. Nonetheless, Ray et al. [452] have already associated a few of the RBPs from their study with changes in mRNA stability. These proteins are indicated by the red (destabilizers) and blue (stabilizers) colours in Fig. 6.2B. We do not observe a clear trend of destabilizing RBPs losing more sites and stabilizing RBPs losing less sites than expected by chance, which would be consistent with the small bias of shorter 3' UTR isoforms being more stable than the corresponding long 3' UTR isoforms [103]. Two proteins with the best established function in mRNA stabilization, ELAVL1 (HuR) and PUM2 (Pumillio 2), appear to be losing less sites than expected. To facilitate further investigations into the impact of 3' UTR shortening on the fate of individual mRNAs, we have summarized the transcripts that are predicted to be regulated by those regulators whose impact on the transcriptome is most affected by the systematic change in poly(A) site use (Supplementary Data D.9).

6.3.3 *The impact of 3' UTR shortening on mRNA abundance*

The above analysis suggests that, consistent with the conclusions drawn from the initial studies of 3' UTR shortening [101, 102], preferential processing at proximal poly(A) sites in proliferating cells leads to an overall loss of destabilizing sequence elements. This would be expected to lead to increased expression of genes with tandem poly(A) sites, yet it is not what we observed in our initial analysis of genome-wide gene expression changes. We next focused on genes with a simple pattern of polyadenylation, considering only genes with tandem poly(A) sites and genes with a single poly(A) site, and excluding genes with more complex patterns of APA (such as alternative terminal exons). We further restricted our set to genes for which spurious A-seq reads in the rest of the gene body including cryptic intronic sites accounted for at most 10% of the reads that were assigned to main poly(A) sites. We estimated the overall mRNA expression level of each gene as the sum of A-seq reads assigned to poly(A) sites located in the terminal exon. Comparing the change in total mRNA levels between naive and activated T cells for genes with two, three or four tandem poly(A) sites relative to genes with a single poly(A) site, we found a slight trend of upregulation of genes with multiple poly(A) sites in activated cells (Fig. 6.3A). Considering genes with precisely two poly(A) sites, we asked whether the change in total mRNA level can be attributed to the change in the relative use of proximal and distal poly(A) sites. If gene expression was mainly regulated through APA with the short 3' UTR isoform being significantly more stable than the long isoform, we would expect a positive correlation between the change in the total mRNA level and the change in proximal vs distal polyadenylation site use. As shown in Fig. 6.3B, we did not detect such a relationship.

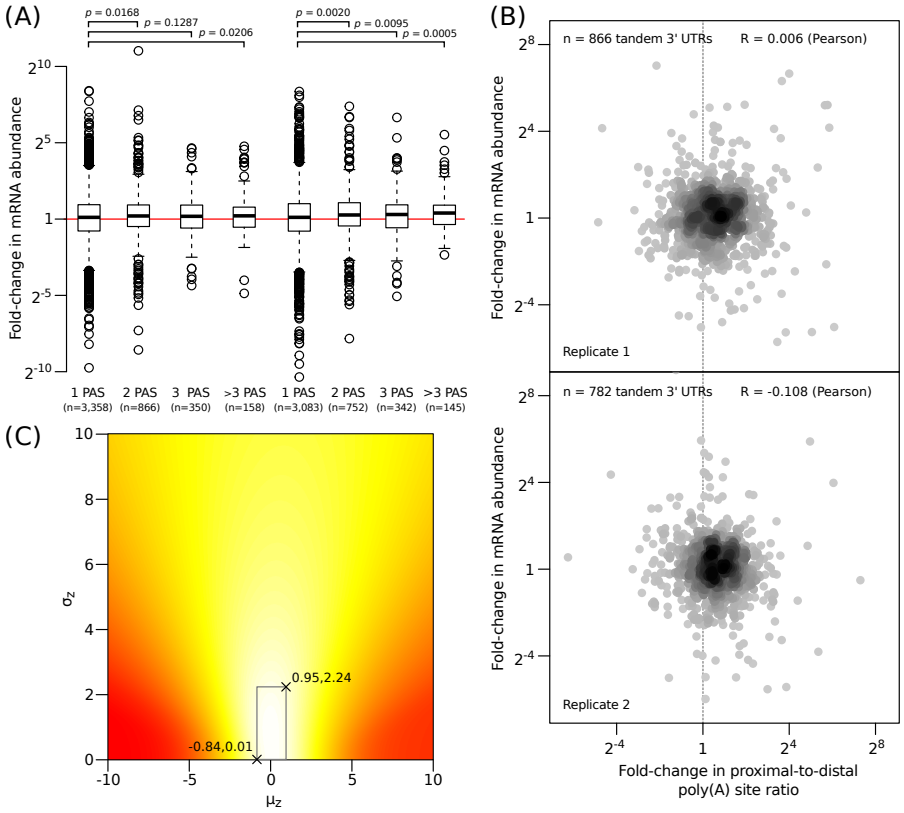


Figure 6.3: Evaluation of changes in mRNA levels in naive and activated murine T cells with respect to changes in poly(A) site usage. (A) Comparison of fold-changes in mRNA abundance for genes with a single or multiple poly(A) sites (PAS; P-values obtained from a one-sided t-test). **(B)** Correlation between changes in mRNA abundance and changes in poly(A) site usage. The centre of mass of the cloud of points is at a positive x-value, reflecting the noted increase in proximal poly(A) site use in dividing cells. **(C)** Contour plot of the log-likelihood of the data as a function of the mean and standard deviation of the log ratio of decay rates (z) of the short and long 3' UTR isoforms. The colour gradient ranging from red to white describes the log-likelihood obtained under the model, with white values showing the better fit. The box marks the 95% posterior probability interval of the parameters (μ_z, σ_z).

The observed abundance of short and long 3' UTR isoforms in different conditions depends not only on the relative rates of polyadenylation at the two sites, but also on the overall rates of transcription and the relative rates of decay of the two transcript forms in the two conditions. To estimate the relative decay rates of short and long isoforms from the 3' end sequencing data, eliminating the effect of confounding factors, we developed a mathematical model (described in Methods) based on the assumption that the distribution of changes in the transcription rate of genes between activated and naive T cells is similar for genes with a single poly(A) site and genes with two tandem poly(A) sites.

The contour plot of the log-likelihood of the data for the 712 genes with two tandem poly(A) sites under the model, as a function of the mean and standard deviation of the log-ratio of the decay rates of long and short isoforms is shown in Fig. 6.3C. We infer that μ_z, the average of the log-ratio of de-

cay rates is located between -0.84 and 0.95. Thus, consistent with a recent study in which the decay rates of long and short isoforms were estimated in mouse fibroblasts [103], we found little evidence for short 3' UTR isoforms being generally more stable compared with long 3' UTR isoforms. Because 3' UTR-mediated interactions with RBPs may also affect the translation rates of mRNAs, we next evaluated the protein output of genes with tandem poly(A) sites in naive and activated T cells.

6.3.4 *The impact of 3' UTR shortening on protein abundance*

To quantify dynamic protein changes on a system-wide level, we combined high mass accuracy mass spectrometry with isobaric tandem mass tagging (TMT) [454] and extensive off-gel electrophoresis sample fractionation (Fig. 6.4A) [455].

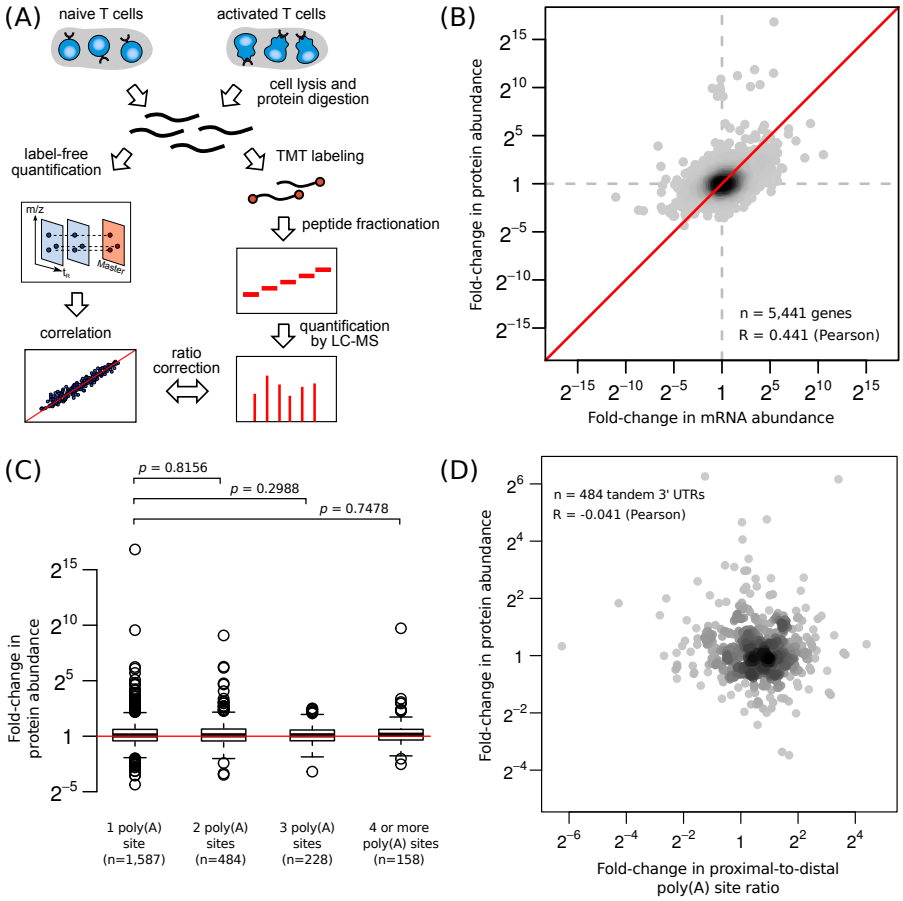


Figure 6.4: Influence of 3' UTR shortening on protein levels in murine T cells. **(A)** Quantitative mass spectrometry (liquid chromatography-mass spectrometry (LC-MS))-based proteomics workflow. Proteins extracted from naive and activated mouse and human T cells were digested and subjected to label-free and tandem mass tag (TMT) quantification, respectively. The TMT-labelled peptides were further fractionated using isoelectric focusing before LC-MS analysis to increase proteome coverage. Finally, the ratios obtained by the TMT approach were correlated with the label-free quantities to correct for possible ratio distortion effects in the final TMT-based quantitative data sets, which comprised more proteins than the data sets based on LFQ. **(B)** Correlation of mRNA and protein abundance changes between activated and naive T cells. **(C)** Change in protein abundance between activated and naive T cells for genes with one to four or more tandem poly(A) sites. **(D)** Correlation between the change in proximal poly(A) site use and the change in protein level.

Performing our experiment in biological duplicates we obtained a total of 138,816 peptide-spectral matches, 48,113 unique peptides and overall quantified 6,187 protein clusters/genes at 1% false discovery rate. It has been reported that ratio compression arising from co-isolated peptides is prevalent with TMT-labelled peptides and needs to be controlled to achieve accurate protein quantification [456]. Therefore, we carried out additional unbiased, label-free quantifications (LFQs) of all samples. As reported previously [457, 458], we observed a good linear correlation of TMT and LFQ protein ratios (Supplementary Fig. D.5A,B). This suggested that the ratio compression of TMT can be largely corrected by an average compression factor [457]

and we therefore recalculated all TMT ratios accordingly (see Supplementary Methods).
Analysis of protein expression data revealed the expected high correlation between replicate measurements (Supplementary Fig. D.4A), and quantitative western blots confirmed the protein-level changes between naive and activated mouse T cells for ten randomly selected proteins (see Supplementary Table D.6). Furthermore, direct comparison of mRNA levels to protein levels (iBAQ derived from LFQ) shows correlations comparable to those recently reported for a variety of tissues (Supplementary Fig. D.5C) [459]. The changes in total mRNA levels determined as described in the previous section also correlated well with the changes in the corresponding protein levels between activated and naive T cells (Fig. 6.4B). We then asked whether genes with tandem poly(A) sites show a systematic change in protein levels compared with genes with a single poly(A) site (Fig. 6.4C) and found that for none of the groups with multiple tandem poly(A) sites could a significant difference be detected. Moreover, similar to the results obtained from the corresponding analysis on the mRNA level, we found no correlation between the change in proximal vs distal poly(A) site use and the change in protein levels (Fig. 6.4D). These data indicate that 3' UTR shortening does not have the same consequence on the mRNA/protein abundance of all affected genes.

6.3.5 *Weak evolutionary conservation of APA*

To determine whether the regulation of polyadenylation at tandem poly(A) sites is evolutionarily conserved, we performed 3' end sequencing and quantitative proteomics on naive and activated T cells obtained from a human blood donor. We found that similar to the mouse T cells, human T cells also had a pronounced increase in the use of proximal polyadenylation sites upon activation (Supplementary Fig. D.6A,B). Also similar to the mouse T cells, the change in polyadenylation site use does not correlate with the change in gene expression, neither at the mRNA (Fig. 6.5A) nor at the protein level (Fig. 6.5B).

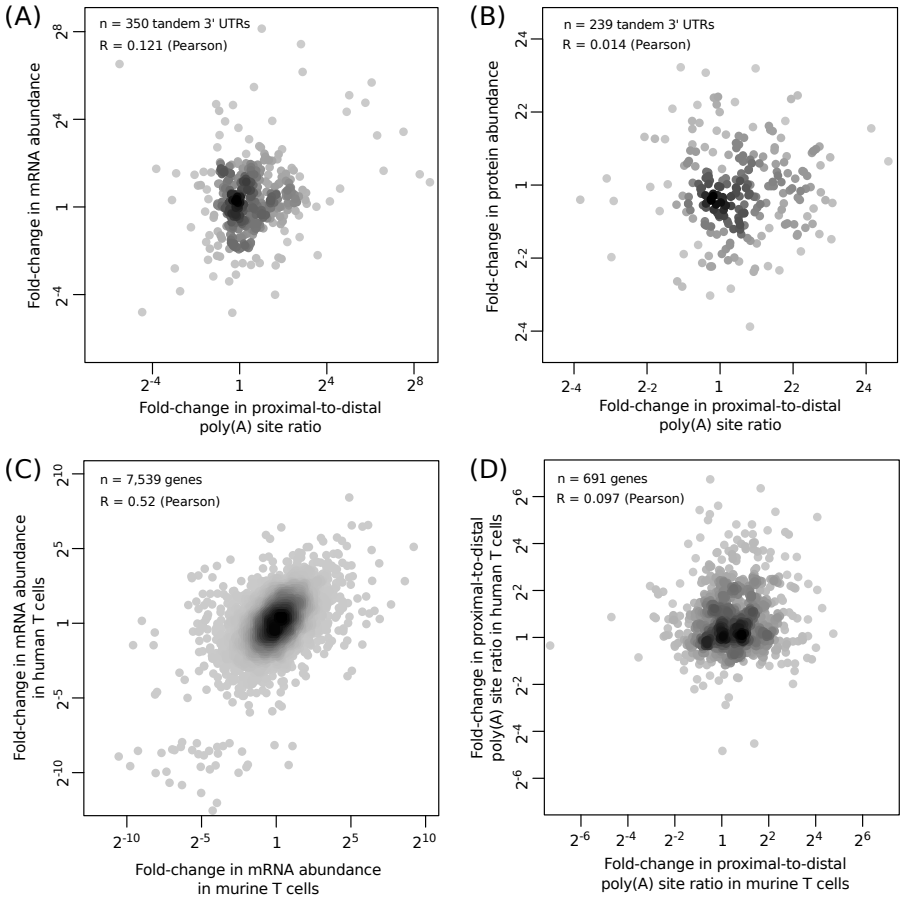


Figure 6.5: Evolutionary conservation of alternative polyadenylation at tandem poly(A) sites. Assessment of the impact of 3' UTR shortening on mRNA **(A)** and protein **(B)** levels in naive and activated human T cells. **(C)** Comparison of changes in mRNA levels upon T cell activation in murine and human T cells. **(D)** Comparison of changes in poly(A) site use upon T cell activation in murine and human T cells. Only orthologous genes that showed the same number of alternative poly(A) sites in mouse and human were considered.

We then investigated to what extent the changes in gene expression in the T cell activation system are conserved between mouse and human. We used the NCBI HomoloGene database [460] to infer mouse–human orthologous genes. Comparing the changes in expression of orthologous genes, we found a good correlation at both the mRNA (Fig. 6.5C) and the protein (Supplementary Fig. D.6C) level. However, the change in the relative use of tandem poly(A) sites is not conserved. Of the 1,734 genes that show regulation by a tandem poly(A) site mechanism in both murine and human T cells, 691 genes share the same number of tandem poly(A) sites in both species. Analysis of this set of genes (Fig. 6.5D) and the bigger set of genes that generally undergo APA in both species (Supplementary Fig. D.6D) did not reveal a correlation between the changes in the relative use of alternative poly(A) sites of orthologous genes. The set of genes that showed a strong (at least 4-fold, n=64) shift to proximal poly(A) site use in both human and murine T cells is not significantly enriched in any particular GO category. However, genes that are

most enriched in this set are related to "stem cell division" (adj. P=0.26) and "positive regulation of mitotic cell cycle" (adj. P=0.28). Last, we investigated whether the same RBP and miRNA regulators would be affected by 3' UTR shortening in human and mouse. We therefore predicted target sites of RBPs and miRNAs in the alternatively processed human 3' UTRs and compared the z-scores of individual regulators between human and mouse (Supplementary Fig. D.6E,F). Indeed, we observed some degree of conservation at this level. For example, we inferred that miRNAs of the Mir17 cluster lose more sites than expected in both systems (mouse: z-score$= -2.1$, human: z-score$= -1.0$), whereas the RBP PUM2, which has a stabilizing effect on its targets, retains more sites than expected (mouse: z-score$= 2.1$, human: z-score$=1.3$).

Our analysis thus indicates that, although the process of 3' UTR shortening in dividing cells is conserved between mouse and human, it is not highly conserved at a quantitative level and on a gene by gene basis. Furthermore, our results indicate that APA at tandem poly(A) sites contributes little to the mRNA and protein output of individual genes. Rather, what appears to be conserved is the restructuring of the RBP and miRNA interactome. That is, even though the 3' UTR shortening of orthologous genes is poorly conserved, the RBPs and miRNAs whose targetome changes most significantly as a result of APA are the same between human and mouse.

6.4 DISCUSSION

Sequencing of animal genomes revealed surprisingly small differences in gene numbers. In the years that followed, much emphasis has been placed on other factors underlying the transcriptome and proteome complexity. MiRNAs and RBPs form a vast regulatory layer whose dynamics has recently come into focus. The discovery of the systematic 3' UTR shortening in proliferating compared with resting cells raised the question of whether this mode of RNA processing serves to bypass repression by miRNAs and generally upregulates the expression of genes with tandem poly(A) sites. Here we have combined measurements of relative polyadenylation site use with measurements of protein levels to investigate this hypothesis in the context of mouse and human T cell activation. Although we were able to demonstrate the systematic 3' UTR shortening in both systems, we did not find a correlation between the extent of proximal polyadenylation site use and the mRNA or the protein levels. Further inferring the relative rates of decay of short and long 3' UTR isoforms, we found that short 3' UTR isoforms have a slightly lower decay rate compared with their long 3' UTR isoforms. Nonetheless, the difference between isoforms appears to be small, consistent with what has been observed in mouse embryonic fibroblasts [103] and it is not systematic.

Analysing the process of 3' UTR shortening upon proliferation of mouse and human T cells, we found that although the phenomenon is conserved, there is little conservation in the set of genes that exhibit 3' UTR shortening and

in the relative change in proximal/distal processing ratios. In this respect, APA resembles alternative splicing where individual events are also poorly conserved between species such as mouse and human [461]. Two questions remain at this point unanswered. The first concerns the molecular mechanism that underlies the systematic change in polyadenylation sites upon cell proliferation. Compelling evidence has been presented that the U1 snRNP acts as a protective factor that prevents premature polyadenylation and that transient limitations in U1 snRNP abundance in specific cellular states lead to polyadenylation at proximal sites [462, 463]. Other proteins such as the mammalian cleavage and polyadenylation factor I (CFIm) components CFIm25 and CFIm68 [81, 82] and the poly(A)-binding protein nuclear 1 [453] appear to have similar effects, a reduction in their concentration leading to polyadenylation at proximal sites. In our data, the CFIm factors as well as PABPN1 appear to be downregulated at the protein level in both mouse and human T cells (Supplementary Data D.10 and D.11). However, whether these or other factors are at work in proliferating cells remains to be determined.

The second question concerns the ultimate consequence and functional relevance of the change in polyadenylation sites. Although systematic differences in the decay rates of short and long isoforms were not identified, transcript stability may still be regulated via APA, with some short isoforms having higher and others lower stability relative to their corresponding long isoforms (see for example, Gupta et al. [437]). However, the fact that many of the proteins that lose the most or least binding sites upon APA are splicing and RNA transport factors suggests that regulatory effects may be expected at other levels. For example, one of the factors that appear to preferentially lose sites is YB-1, a marker of stress granules and processing bodies [464]. On the other hand, poly(A)-binding proteins that are involved in the nuclear export of mRNAs and many other cytoplasmic processes [465], lose substantially fewer sites than expected.

Finally, one hypothesis to consider is that 3' UTR shortening does not have specific consequences on gene regulation. Rather, it could be a complex regulatory system, acting on long 3' UTRs that may no longer be needed when cells are engaged in a very defined state of active proliferation. In those circumstances, the 3' UTR shortening may act both to conserve energy, as well as to prevent the interference of complex, cell type-specific post-transcriptional regulatory networks, with the cell cycle programme. With the availability of systems that allow genetic modification of mammalian cells [466], it may soon become possible to modify the poly(A) signals and to test the effect of expressing solely the long 3' UTR isoforms in cells that are induced to proliferate.

6.5 METHODS

6.5.1 *Isolation and activation of T cells*

For mouse T cells, spleen and lymph nodes were dissected from 8- to 10-week-old female C57BL/6 mice and total, untouched T cells were isolated by MACS purification (Pan T cell isolation kit from Miltenyi or the mouse T cell isolation kit from Stem Cell Technologies) according to the manufacturer's protocol. T cells were activated for 72 h with mouse T-Activator CD3/CD28 Dynabeads (Gibco/Life Technologies) and 30 U of recombinant IL-2 (Peprotech). Corresponding unstimulated cells were from the same T cell preparations. Human T cells from single donor human blood samples were isolated with the Pan T cell isolation kit from Miltenyi and either left untreated or were stimulated with human T-Activator CD3/CD28 Dynabeads (Life Technologies) and 30 U of recombinant IL-2 from Peprotech.

6.5.2 *3' End sequencing and inference of poly(A) sites*

Murine T cell 3' end sequencing libraries were prepared according to the original A-seq protocol [81]. To circumvent the frequently cumbersome size selection step in this protocol, we developed an improved 3' end sequencing protocol (A-seq2) that we used for the preparation of the human 3' end sequencing libraries (see Supplementary Methods for details). The Gene Expression Omnibus (GEO) accession number for the A-seq libraries is GSE54950. Sequencing reads were preprocessed to remove 3' adapter sequences and mapped to the mouse genome (mm9) and human genome (hg19), respectively, with CLIPZ [442]. To ensure that only genuine 3' ends are considered, we only used A-seq reads that contained at least four nucleotides of the adapter sequence. Based on the precise mapping of the 3' end of reads that mapped to a unique position in the genome, we computed putative cleavage sites and their abundance at nucleotide resolution. Putative cleavage sites that had at least seven genomically encoded A nucleotides in the eight nucleotide region immediately downstream were considered likely internal priming events and were not used in further analyses. Finally, closely spaced 3' end sites located in terminal exons of transcripts were grouped into poly(A) site clusters by applying single-linkage clustering with a distance threshold of seven nucleotides. Only those clusters that showed a minimal abundance of five A-seq reads per million were further analysed. For the mouse genome, we thereby inferred a total of 15,068 clusters with an average cluster span of 16 nucleotides. For each cluster, a representative cleavage site was chosen by ranking individual sites by their expression value in each A-seq library and then determining the overall top ranked site (majority vote over all A-seq libraries) [81]. For 3' end sequencing of human mRNAs, we used a slightly modified A-seq procedure (A-seq2, see Supplementary Methods for details). Sequencing reads were first processed based on their expected structure from this protocol. First, only A-seq2 reads that contained three T residues at posi-

tions 5–7 (indicating the beginning of the poly(A) tail) were selected for further use. From these, randomized nucleotides at positions 1–4 at the 5' end of the reads (needed for cluster coordination in Illumina sequencing) were trimmed together with the three Ts, thus removing seven nucleotides. The reverse complement of the remaining sequences, presumably representing mRNA 3' ends, was then mapped to the genome. The rest of the analysis was carried out identically to the murine A-seq sequences. A total 18,918 poly(A) site clusters were inferred for the human T cell samples. Summary statistics on the number of mapped reads and tandem poly(A) sites are provided in Supplementary Tables D.7 and D.8, respectively. The nucleotide profile flanking the inferred cleavage sites (Supplementary Fig. D.7A) closely resembles the profile obtained in murine T cells (Supplementary Fig. D.2A) and previous studies [81]. Also the distribution of polyadenylation motifs upstream of the cleavage site corresponds to the pattern observed in murine T cells (Supplementary Fig. D.7B). In order to identify genes that show a marked change in the use of proximal and distal poly(A) sites, we first divided the region between the most proximal and most distal poly(A) site into two parts of equal length and pooled read counts of poly(A) sites in the 5' half and in the 3' half. For murine T cell samples, we next employed the statistical framework DEXseq version 1.8 [438] to identify genes that showed a change in the usage pattern between the proximal and the distal poly(A) site. A total of 157 genes were identified to undergo a significant change (adjusted P-value$\leq$0.05) in the use of the proximal poly(A) site.

6.5.3 *Differential gene expression and GO analysis*

For each gene, A-seq reads mapping to terminal exons of its associated transcript isoforms were counted. Differential gene expression analysis was performed with DESeq version 1.10 [373]. Genes that showed a log twofold differential regulation and an adjusted P-value$\leq$0.01 were considered as changing significantly. GO analysis of up- and downregulated genes was performed with Ontologizer version 2.0 [467].

6.5.4 *Prediction of miRNA and RBP target sites in murine 3' UTRs*

MiRNA target predictions were obtained from the ElMMo server release 5 (http://www.mirz.unibas.ch). We restricted the set of target sites to conserved sites by choosing a minimal ElMMo score of 0.5. For ease of use, we mapped 3' UTR sequences to the mouse or human genome using GMAP [468], and converted predicted transcript coordinates of miRNA target sites to genomic coordinates. A weighted target site score was then calculated as the sum over all genes with tandem poly(A) sites, with the probability of each target site for the miRNA multiplied by the abundance of the gene's 3' UTR isoform in which the predicted target site was present. For each miRNA, we recorded the log_2 fold-change (x) of the sum of weighted target site scores in alternative 3' UTR regions for activated over naive T cells. To assess the significance

of the fold-change, we shuffled the labels of the miRNA target sites (corresponding to the cognate miRNAs) that were located in alternative 3' UTR regions. We performed 500 randomizations and obtained the mean (μ_x) and standard deviation (σ_x) of the log_2 fold changes across randomized data sets. We estimated the significance of the observed log_2 fold change by the z-score defined as $z = (x - \mu_x)/\sigma_x$.

PWMs of the binding motifs of RBPs were obtained from the CISBP-RNA database (http://cisbp-rna.ccbr.utoronto.ca) [452]. Only PWMs with category annotation 'direct evidence' in mouse or human were considered. MotEvo was used to scan murine 3' UTRs (using a background prior of 0.99 and an UFE prior of 200) to predict evolutionarily conserved motif matches [290]. As input we provided multiple sequence alignments of nine mammalian species generated with the ElMMo pipeline [255]. Computations were performed the same way as done for miRNAs only replacing the ElMMo score with the score obtained from MotEvo. In case more than one PWM was present for a given RBP in the database, we evaluated the predicted binding sites for each of them individually, and used as background for the site randomization the sites predicted for only one representative PWM for each RBP. The representative was the PWM with the highest information content. We only report the results for PWMs that had a minimum of ten predicted sites in the alternatively processed 3' UTR regions.

6.5.5 *Estimation of relative mRNA decay rates of short and long transcript isoforms*

We used the following model to estimate the relative stability of 3' UTR isoforms. Let us assume that mRNAs are transcribed from their corresponding gene at rate c, and are processed at either the proximal or the distal poly(A) site with frequencies f and $1 - f$, respectively. Let μ_S and μ_L be the decay rates of the short and long 3' UTR isoforms, respectively. With the dynamics of the short (M_S) and long (M_L) isoforms being described by the following equations $dM_S/dt = cf - \mu_S M_S$ and $dM_L/dt = cf - \mu_L M_L$, and denoting by superscripts A and N the variables corresponding to activated and naive T cells, we obtain at steady state $M_S^N = \frac{c^N f^N}{\mu_S}$, $M_L^N = \frac{c^N(1-f^N)}{\mu_L}$, $M_S^A = \frac{c^A f^A}{\mu_S}$, $M_L^A = \frac{c^A(1-f^A)}{\mu_L}$. From our 3' end sequencing experiments, we obtain ratios of proximal-to-distal site use in the two conditions, that is, $R^N = \frac{M^N}{M_L^N}$ and $R^A = \frac{M^A}{M_L^A}$ as well as the ratio in the overall mRNA expression between the two conditions, $Q = \frac{M_S^N + M_L^N}{M_S^A + M_L^A}$. With the further notation $\alpha = \mu_L/\mu_S$ and $\beta = c^N/c^A$, we can express these measured quantities in terms of the variables of the model defined above as follows $R^N = \frac{f^N}{1-f^N}\alpha$, $R^A = \frac{f^A}{1-f^A}\alpha$, $Q = \beta\frac{f^N\alpha + 1 - f^N}{f^A\alpha + 1 - f^A}$. Note, that we can express the *unknown* frequencies of 3' UTR processing at proximal and distal sites in the two conditions in terms of the measured ratios of proximal-to-distal site use, that is, $f^N = \frac{R^N}{R^N + \alpha}$ and

$f^A = \frac{R^A}{R^A+\alpha}$. Further defining $x = log(Q)$, $y = log(\beta)$, $z = log(\alpha)$ and the function $g(R^N, R^A, z) = log\left(\frac{(R^N+1)(R^A+e^z)}{(R^A+1)(R^N+e^z)}\right)$, we obtain:

$$x = y + g(R^N, R^A, z). \tag{6.1}$$

That is, the observed log fold-change in total mRNA levels (x) is a result of the log fold-change in transcription rate (y) and a log fold-change in the decay rate ($g(R^N, R^A, z)$). The latter is a function of the observed ratios of short vs long isoforms in naive and activated T cells, and of the log-ratio z of decay rates of the short and long isoforms. Note, that whereas x has been measured, the variables y and z are both unknown. Thus, we cannot uniquely determine the relative decay rates for a particular gene, without knowing the relative transcription rates for that gene in the two conditions. Reasoning that genes that are only regulated at the level of transcription and not through polyadenylation provide an upper bound on transcriptional changes, we estimate the distribution of transcription log fold-changes y from the set of genes that have only a single isoform. We found that to a good approximation, the distribution of y is a Gaussian with mean $\mu_y = 0.00154$ and standard deviation $\sigma_y = 0.92691$, that is, $P(y|\mu_y, \sigma_y) = \frac{1}{\sigma_y\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right)$. We further assume that the log-ratio of decay rates can also be approximated by a Gaussian distribution, that is, $P(z|\mu_z, \sigma_z) = \frac{1}{\sigma_z\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_z)^2}{2\sigma_z^2}\right)$. Finally, we estimate the parameters μ_z and σ_z by assuming that both y and z were drawn from their respective distributions and comparing the observed mRNA fold-changes with those expected using equation 6.1. In particular, the log fold-changes x were measured in duplicate for each gene and this allows us to estimate the measurement error of these measurements. Let (x_i^1, x_i^2) denote the pair of replicate measurements for gene i. Assuming that measurement errors are Gaussian distributed, we can estimate the variance of the measurement errors as $\tau^2 = \frac{1}{2(n-2)} \sum_i (x_i^1 - x_i^2)^2$. Given particular values of y and z, the probability to measure a given log fold-change x is given by $P(x|y, z, R_R, R_A, \tau) = \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{(x-y-g(z,R_R,R_A))^2}{2\tau^2}\right)$. By multiplying this conditional probability by the prior probabilities $P(y|\mu_y, \sigma_y)$ and $P(z|\mu_z, \sigma_z)$ and integrating over both y and z we obtain the probability $P(x|R_R, R_A, \mu_y, \mu_z, \sigma_y, \sigma_z, \tau) = \int d_y d_x P(x|y, z, R_R, R_A, \tau) P(y|\mu_y, \sigma_y) P(z|\mu_z, \sigma_z)$. Note that the integral over y can easily be performed analytically to obtain:

$$\begin{aligned} P(x|R_R, R_A, \mu_y, \mu_z, \sigma_y, \sigma_z, \tau) = & \\ & \int d_z \frac{1}{2\pi\sigma_z\sqrt{\tau^2+\sigma_y^2}} \\ & \exp\left(-\frac{(z-\mu_z)^2}{2\sigma_z^2}\right. \\ & \left.-\frac{\left(x-\mu_y-g(z,R_R,R_A)\right)^2}{2(\tau^2+\sigma_y^2)}\right). \end{aligned} \tag{6.2}$$

This integral, however, cannot be performed analytically and we therefore carried out the integration numerically over the range $\mu_z \pm \times\sigma_z$ with MATLAB

version R2012B. In order to evaluate combinations of μ_z and σ_z, we calculated the log likelihood of the data by a grid approach sampling values for μ_z and σ_z with a step size of 0.01 from -10 to 10 and 0.01 to 10, respectively.

6.6 AUTHORS INFORMATION

6.6.1 *List of authors*

The following authors have contributed to the work discussed in Chapter 6:

1. Andreas R. Gruber[1] (Abbr.: ARG),
2. Georges Martin[1] (Abbr.: GM),
3. Philipp Mueller[2,3] (Abbr.: PM),
4. Alexander Schmidt[4] (Abbr.: AS),
5. Andreas Johannes Gruber[1] (Abbr.: AJG),
6. Rafal Gumienny[1] (Abbr.: RG),
7. Nitish Mittal[1] (Abbr.: NM),
8. Rajesh Jayachandran[3] (Abbr.: RJ),
9. Jean Pieters[3] (Abbr.: JP),
10. Walter Keller[1] (Abbr.: WK),
11. Erik van Nimwegen[1] (Abbr.: EvN) &
12. Mihaela Zavolan[1] (Abbr.: MZ)

whereat author affiliations are as follows:

1 Computational and Systems Biology, Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Basel CH-4056, Switzerland

2 Department of Biomedicine, University of Basel and University Hospital Basel, Hebelstrasse 20, Basel CH-4031, Switzerland

3 Infection Biology, Biozentrum, University of Basel, Basel CH-4056, Switzerland

4 Proteomics Core Facility, Biozentrum, University of Basel, Basel CH-4056, Switzerland

6.6.2 *Author contributions*

The listing of authors in the previous subsection (6.6.1) was performed according to the authors' contributions, whereat the first two authors (ARG and GM) contributed equally and most to this work and subsequent authors decreasingly. However, the last four authors are principal investigators and thus their listing follows the opposite ranking (the last contributed the most and the preceding three authors decreasingly).
In detail, using the abbreviations specified in the previous subsection (i.e. 6.6.1): ARG designed the experiments, performed the analysis and wrote the paper. GM designed and executed the Aseq2 method, contributed to T cell isolation and activation and wrote the paper. PM performed the human and mouse T cell isolation and activation. AS designed and performed the mass spectrometric analysis and wrote the paper. AJG and RG contributed the predictions of RBP- and miRNA-binding sites, respectively. NM and RJ contributed to the experiments. WK designed the experiments and wrote the paper. JP designed the experiments. EvN developed the inference of mRNA decay rates. MZ conceived the project, designed the experiments, analysed the data and wrote the paper.

6.7 SUPPLEMENTARY MATERIALS

Supplementary materials can be found in Appendix D.

6.8 ACKNOWLEDGMENTS

We thank Beatrice Dimitriades for technical assistance. We also thank Christian Beisel, Ina Nissen and Manuel Kohler from the D-BSSE of the ETH Zurich for the support to accomplish the Illumina sequencing.

6.9 FUNDING

Using the abbreviations specified in subsection 6.6.1: This work was supported by the University of Basel, the Louis-Jeantet Foundation for Medicine and the Swiss National Science Foundation (grant no. 31003A-143977 to WK; 31003A-146463 and CRS133_124819 to JP). RJ received support from the Cloetta foundation. AJG was funded by a Werner Siemens Fellowship.

7

A COMPREHENSIVE ANALYSIS OF 3' END SEQUENCING DATA SETS REVEALS NOVEL POLYADENYLATION SIGNALS AND THE REPRESSIVE ROLE OF HETEROGENEOUS RIBONUCLEOPROTEIN C ON CLEAVAGE AND POLYADENYLATION

The work discussed in this chapter was published in Genome Research in 2016 (see reference [469]).

7.1 ABSTRACT

Alternative polyadenylation (APA) is a general mechanism of transcript diversification in mammals, which has been recently linked to proliferative states and cancer. Different 3' untranslated region (3' UTR) isoforms interact with different RNA-binding proteins (RBPs), which modify the stability, translation, and subcellular localization of the corresponding transcripts. Although the heterogeneity of pre-mRNA 3' end processing has been established with high-throughput approaches, the mechanisms that underlie systematic changes in 3' UTR lengths remain to be characterized. Through a uniform analysis of a large number of 3' end sequencing data sets, we have uncovered 18 signals, six of which are novel, whose positioning with respect to pre-mRNA cleavage sites indicates a role in pre-mRNA 3' end processing in both mouse and human. With 3' end sequencing we have demonstrated that the heterogeneous ribonucleoprotein C (HNRNPC), which binds the poly(U) motif whose frequency also peaks in the vicinity of polyadenylation (poly(A)) sites, has a genome-wide effect on poly(A) site usage. HNRNPC-regulated 3' UTRs are enriched in ELAV-like RBP 1 (ELAVL1) binding sites and include those of the CD47 gene, which participate in the recently discovered mechanism of 3' UTR-dependent protein localization (UDPL). Our study thus establishes an up-to-date, high-confidence catalog of 3' end processing sites and poly(A) signals, and it uncovers an important role of HNRNPC in regulating 3' end processing. It further suggests that U-rich elements mediate interactions with multiple RBPs that regulate different stages in a transcript's life cycle.

7.2 INTRODUCTION

The 3' ends of most RNA polymerase II–generated transcripts are generated through endonucleolytic cleavage and the addition of a polyadenosine tail of 70–100 nucleotides (nt) median length [470]. Recent studies have revealed systematic changes in 3' UTR lengths upon changes in cellular states, either those that are physiological [102, 463] or those during pathologies [471]. 3' UTR lengths are sensitive to the abundance of specific spliceosomal proteins [462], core pre-mRNA 3' end processing factors [81, 82], and polyadenylation factors [453]. Because 3' UTRs contain many recognition elements for

RNA-binding proteins (RBPs) that regulate the subcellular localization, intracellular traffic, decay, and translation rate of the transcripts in different cellular contexts (see, e.g., [472]), the choice of polyadenylation (poly(A)) sites has important regulatory consequences that reach up to the subcellular localization of the resulting protein [105]. Studies of presumed regulators of polyadenylation would greatly benefit from the general availability of comprehensive catalogs of poly(A) sites such as PolyA_DB [473, 474], which was introduced in 2005 and updated 2 years later. Full-length cDNA sequencing offered a first glimpse on the pervasiveness of transcription across the genome and on the complexity of gene structures [475]. Next-generation sequencing technologies, frequently coupled with the capture of transcript 5' or 3' ends with specific protocols, enabled the quantification of gene expression and transcript isoform abundance [476]. By increasing the depth of coverage of transcription start sites and mRNA 3' ends, these protocols aimed to improve the quantification accuracy ([371, 477–479]). Sequencing of mRNA 3' ends takes advantage of the poly(A) tail, which can be captured with an oligo-dT primer. More than 4.5 billion reads were obtained with several protocols from human or mouse mRNA 3' ends in a variety of cell lines [479, 480], tissues [98, 481], developmental stages [482, 483], and cell differentiation stages [484], as well as following perturbations of specific RNA processing factors [81, 82, 453, 485, 486]. Although some steps are shared by many of the proposed 3' end sequencing protocols, the studies that employed these methods have reported widely varying numbers of 3' end processing sites. For example, 54,686 [474], 439,390 [98], and 1,287,130 [480] sites have been reported in the human genome. The current knowledge about sequence motifs that are relevant to cleavage and polyadenylation (for review, see [487]) goes back to studies conducted before next-generation sequencing technologies became broadly used [90, 488, 489]. These studies revealed that the AAUAAA hexamer, which recently was found to bind the WDR33 and CPSF4 subunits of the cleavage and polyadenylation specificity factor (CPSF) [91, 92] and some close variants, is highly enriched upstream of the pre-mRNA cleavage site. The A[AU]UAAA *cis*-regulatory element (also called poly(A) signal) plays an important role in pre-mRNA cleavage and polyadenylation [490] and is found at a large proportion of pre-mRNA cleavage sites identified in different studies [489, 491, 492]. However, some transcripts that do not have this poly(A) signal are nevertheless processed, indicating that the poly(A) signal is not absolutely necessary for cleavage and polyadenylation. The constraints that functional poly(A) signals have to fulfill are not entirely clear, and at least 10 other hexamers have been proposed to have this function [90]. Viral RNAs as, for example, from the simian virus 40 have been instrumental in uncovering RBP regulators of polyadenylation and their corresponding sequence elements. Previous studies revealed modulation of poly(A) site usage by U-rich element binding proteins such as the heterogeneous nuclear ribonucleoprotein (hnRNP) C1/C2 [493, 494], the polypyrimidine tract binding protein 1 [494, 495], FIP1L1, and CSTF2 [494], and by proteins that bind G-rich elements–cleavage stimulation factor CSTF2 [496] and HNRNPs F

and H1 [497]–or C-rich elements–poly(rC)-binding protein 2 [486]. Some of these proteins are multifunctional splicing factors that appear to couple various steps in pre-mRNA processing, such as splicing, cleavage, and polyadenylation [498]. The sequence elements to which these regulators bind are also frequently multifunctional, enabling positive or negative regulation by different RBPs [496]. A first step toward understanding the regulation of poly(A) site choice is to construct genome-wide maps of poly(A) sites, which can be used to investigate differential polyadenylation across tissues and the response of poly(A) sites to specific perturbations.

7.3 RESULTS

7.3.1 *Preliminary processing of 3' end sequencing data sets*

Protocol-specific biases as well as vastly different computational data processing strategies may explain the discrepancy in the reported number of 3' end processing sites, which ranges from less than 100,000 to over 1 million [98, 474, 480] for the human genome. By comparing the 3' end processing sites from two recent genome-wide studies [98, 481], we found that a substantial proportion was unique to one or the other of the two studies (Supplemental Table E.1). This motivated us to develop a uniform and flexible processing pipeline that facilitates the incorporation of all published sequencing data sets, yielding a comprehensive set of high-confidence 3' end processing sites. From public databases we obtained 78 human and 110 mouse data sets of 3' end sequencing reads (Supplemental Tables E.2, E.3), generated with nine different protocols, for which sufficient information to permit the appropriate preprocessing steps (trimming of 5' and 3' adapter sequences, reverse-complementing the reads, etc., as appropriate) was available.

We preprocessed each sample as appropriate given the underlying protocol and then subjected all data sets to a uniform analysis as follows. We mapped the preprocessed reads to the corresponding genome and transcriptome and identified unique putative 3' end processing sites. Because many protocols employ oligo-dT priming to capture the pre-mRNA 3' ends, internal priming is a common source of false-positive sites, which we tried to identify and filter out as described in the Methods section. From the nearly 200 3' end sequencing libraries, we thus obtained an initial set of 6,983,499 putative 3' end processing sites for human and 8,376,450 for mouse. The majority of these sites (76% for human and 71% for mouse) had support in only one sample, consistent with our initial observations of limited overlap between the sets of sites identified in individual studies and mirroring also the results of transcription start site mapping with the CAGE technology [499]. Nevertheless, we developed an analysis protocol that aimed to identify bona fide, independently regulated poly(A) sites, including those that have been captured in a single sample. To do this, we used not only the sequencing data but also information about poly(A) signals, which we therefore set to comprehensively identify in the first step of our analysis.

7.3.2 *Highly specific positioning with respect to the pre-mRNA cleavage site reveals novel poly(A) signals*

To search for signals that may guide polyadenylation, we designed a very stringent procedure to identify high-confidence 3' end processing sites. Pre-mRNA cleavage is not completely deterministic but occurs with higher frequency at "strong" 3' end processing sites and with low frequency at neighboring positions [489]. Therefore, a common step in the analysis of 3' end sequencing data is to cluster putative sites that are closely spaced and to report the dominant site from each cluster [81, 489, 500]. To determine an appropriate distance threshold, we ranked all the putative sites first by the number of samples in which they were captured and then by the normalized number of reads in these samples. By traversing the list of sites from those with the strongest to those with the weakest support, we associated lower-ranking sites located up to a specific distance from the higher-ranked site with the corresponding higher-ranking site. We scanned the range of distances from 0 to 25 nt upstream of and downstream from the high-ranking site, and we found that the proportion of putative 3' end processing sites that are merged into clusters containing more than one site reached 40% at $\sim$8 nt and changed little by further increasing the distance (for details, see 7.10 Methods). For consistency with previous studies [489], we used a distance of 12 nt. To reduce the frequency of protocol-specific artifacts, we used only clusters that were supported by reads derived with at least two protocols, and to allow unambiguous association of signals to clusters, for the signal inference we only used clusters that did not have another cluster within 60 nt. This procedure resulted in 221,587 3' end processing clusters for human and 209,345 for mouse. By analyzing 55-nt-long regions located immediately upstream of the center of these 3' end processing clusters (as described in the 7.10 Methods section), we found that the canonical poly(A) signals AAUAAA and AUUAAA were highly enriched and had a strong positional preference, peaking at 21 nt upstream of cleavage sites (Fig. 7.1A), as reported previously [90, 489]. We therefore asked whether other hexamers have a similarly peaked frequency profile, which would be indicative of their functioning as poly(A) signals. The 12 signals that were identified in a previous study [90] served as controls for the procedure. In both mouse and human data, the motif with the highest peak was, as expected, the canonical poly(A) signal AAUAAA, which occurred in 46.82% and 39.54% of the human and mouse sequences, respectively. Beyond this canonical signal, we found 21 additional hexamers, the second most frequent being the close variant of the canonical signal AUUAAA, which was present in 14.52% and 12.28% of the human and mouse 3' sequences, respectively. All 12 known poly(A) signals [90] were recovered by our analysis in both species, demonstrating the reliability of our approach. Further supporting this conclusion is the fact that six of the 10 newly identified signals in each of the two species are shared. All of the conserved signals are very close variants (1 nt difference except for AACAAG) of one of the two main poly(A) signals,

AAUAAA and AUUAAA. Strikingly, all of these signals peak in frequency at 20–22 nt upstream of the cleavage site (Fig. 7.1A).

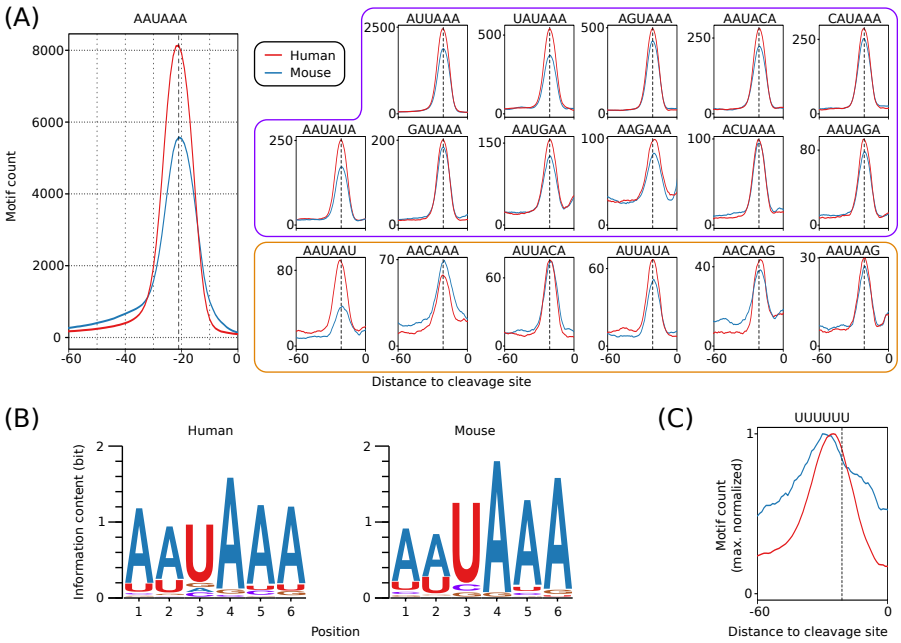


Figure 7.1: Hexamers with highly specific positioning upstream of human and mouse pre-mRNA 3' end cleavage sites. (A) The frequency profiles of the 18 hexamers that showed the positional preference expected for poly(A) signals in both human and mouse. The known poly(A) signal, AAUAAA, had the highest frequency of occurrence (left). Apart from the 12 signals previously identified (AAUAAA and motifs with the purple frame) [90], we have identified six additional motifs (orange frame) whose positional preference with respect to poly(A) sites suggests that they function as poly(A) signals and are conserved between human and mouse. **(B)** Sequence logos based on all occurrences of the entire set of poly(A) signals from the human (left) and mouse (right) atlas. **(C)** The $(U)_6$ motif, which is also enriched upstream of pre-mRNA cleavage sites, has a broader frequency profile and peaks upstream of the poly(A) signals, which are precisely positioned 20–22 nt upstream of the pre-mRNA cleavage sites (indicated by the dashed, vertical line).

Experimental evidence for single-nucleotide variants of the AAUAAA signal (including the AACAAA, AAUAAU, and AAUAAG motifs identified here) functioning in polyadenylation was already provided by Sheets et al. [501]. The four signals identified in only one of each species also had a clear peak at the expected position with respect to the poly(A) site, but they had a larger variance (Supplemental Fig. E.1). Altogether, these results indicate a genuine role of the newly identified signals in the process of cleavage and polyadenylation.

Of the 221,587 high-confidence 3' end processing clusters in human and 209,345 in mouse, 87% and 79%, respectively, had at least one of the 22 signals identified above in their upstream region. Even when considering only the 18 signals that are conserved between human and mouse, 86% of the human clusters and 75% of the mouse clusters had a poly(A) signal.

Thus, our analysis almost doubles the set of poly(A) signals and suggests that the vast majority of poly(A) sites does indeed have a poly(A) signal that is positioned very precisely with respect to the pre-mRNA cleavage site. The dominance of the canonical poly(A) signal is reflected in the sequence logos constructed based on all annotated hexamers in the human and mouse poly(A) site atlases, generated as described in the following section and in the 7.10 Methods section (Fig. 7.1B).

7.4 A COMPREHENSIVE CATALOG OF HIGH-CONFIDENCE 3' END PROCESSING SITES

Based on all of the 3' end sequencing data sets available (for more details about the protocols that were used to generate these data sets, see Supplemental Material) and the conserved poly(A) signals that we inferred as described above, we constructed a comprehensive catalog of strongly supported 3' end processing sites in both the mouse and human genomes. We started from the 6,983,499 putative cleavage sites for human and 8,376,450 for mouse. Although in many data sets a large proportion of putative sites was supported by single reads and did not have any of the expected poly(A) signals in the upstream region, the incidence of upstream poly(A) signals increased with the number of reads supporting a putative site (Supplemental Fig. E.2). Thus, we used the frequency of occurrence of poly(A) signals to define sample-specific cutoffs for the number of reads required to support a putative cleavage site. We then clustered all putative sites with sufficient read support, associating lower-ranked sites with higher-ranking sites that were located within at most 12 nt upstream or downstream, as described above. Because in this set of clusters we found cases where the pre-mRNA cleavage site appeared located in an A-rich region upstream of another putative cleavage site, we specifically reviewed clusters in which a putative cleavage site was very close to a poly(A) signal, as these likely reflect internal priming events [98, 429, 479]. These clusters were either associated with a downstream cluster, retained as independent clusters, or discarded, according to the procedure outlined in the Methods section. By reasoning that distinct 3' end processing sites should have independent signals to guide their processing, we merged clusters that shared all poly(A) signals within 60 nt upstream of their representative sites, clusters whose combined span was <25 nt, and clusters without annotated poly(A) signals that were closer than 12 nt to each other and had a combined span of at most 50 nt. Clusters >50 nt and without poly(A) signals were excluded from the atlas. This procedure (for details, see the 7.10 Methods section) resulted in 392,912 human and 183,225 mouse 3' end processing clusters. Of note, even though 3' end processing sites that were within 25 nt of each other were merged into single clusters, the median cluster span was very small, 7 and 3 nt for mouse and human, respectively (Supplemental Fig. E.3). Supplemental Figures E.4A and E.5A show the frequency of occurrence of the four nucleotides as a function of the distance to the cleavage sites for

sites that were supported by a decreasing number of protocols. These profiles exhibited the expected pattern [81, 489, 502], indicating that our approach identified bona fide 3' end processing sites, even when they had limited experimental support. The proportion of clusters located in the terminal exon increased with an increasing number of supporting protocols (Supplemental Fig. E.4B, E.5B), probably indicating that the canonical poly(A) sites of constitutively expressed transcripts are identified by the majority of protocols, whereas poly(A) sites that are only used in specific conditions were captured only in a subset of experiments. Although in constructing our catalog we used most of the reads generated in two recent studies (>95% of the reads that supported human 3' end processing sites in these two data sets mapped within the poly(A) site clusters of our human catalog) [98, 481], only 61.82% [481] and 41.38% [98] of the unique processing sites inferred in these studies were located within poly(A) clusters from our human catalog. This indicated that a large fraction of the sites that were cataloged in previous studies is supported by a very small number of reads and lacks canonically positioned poly(A) signals. We applied very stringent rules to construct an atlas of high-confidence poly(A) sites, and the entire set of putative cleavage sites that resulted from mapping all of the reads obtained in these 3' end sequencing studies is available as Supplemental Data E.9 (human) and E.10 (mouse), as well as online at http://www.polyasite.unibas.ch, where users can filter sites of interest based on the number of supporting protocols, the identified poly(A) signals, and/or the genomic context of the clusters.

7.4.1 *3' end processing regions are enriched in poly(U)*

Of the human and mouse 3' end processing sites from our poly(A) atlases, 76% and 75%, respectively, possessed a conserved poly(A) signal in their 60 nt upstream region. That ∼25% did not may support the hypothesis that pre-mRNA cleavage and polyadenylation do not absolutely require a poly(A) signal [503]. Nevertheless, we asked whether these sites possess other signals, with a different positional preference, which may contribute to their processing. To answer this question, we searched for hexamers that were significantly enriched in the 60 nt upstream of cleavage sites without an annotated poly(A) signal. The two most enriched hexamers were poly(A) (P-value of binomial test $<1.0 \times 10^{-100}$), which showed a broad peak in the region of -20 to -10 upstream of cleavage sites, and poly(U) (P-value $<1.0 \times 10^{-100}$), which also has a broad peak around -25 nt upstream of cleavage sites, particularly pronounced in the human data set (Fig. 7.1C). The poly(U) hexamer is very significantly enriched (P-value of binomial test $<1.0 \times 10^{-100}$) in the 60 nt upstream regions of all poly(A) sites, not only in those that do not have a common poly(A) signal (11th most enriched hexamer in the human atlas and 60th most enriched hexamer in the mouse atlas) (Supplemental Tables E.4, E.5). Although the A- and U-richness of pre-mRNA 3' end processing regions have been observed before [490], their relevance for polyadenylation and the regulators that bind these motifs have been characterized only partially. For

example, the core 3' end processing factor FIP1L1 can bind poly(U) [87, 93], and its knock-down causes a systematic increase in 3' UTR lengths [93, 504].

7.5 HNRNPC KNOCK-DOWN CAUSES GLOBAL CHANGES IN ALTERNATIVE CLEAVAGE AND POLYADENYLATION

Several proteins (ELAVL1, TIA1, TIAL1, U2AF2, CPEB2 and CPEB4, HNRNPC) that regulate pre-mRNA splicing and polyadenylation, as well as mRNA stability and metabolism, have also been reported to bind U-rich elements [452]. Of these, HNRNPC has been recently studied with crosslinking and immunoprecipitation (CLIP) and found to bind the majority of protein-coding genes [505], with high specificity for poly(U) tracts [452, 505–508]. HNRNPC appears to nucleate the formation of ribonucleoprotein particles on nascent transcripts and to regulate pre-mRNA splicing [505, 506] and polyadenylation at Alu repeats [509]. We therefore hypothesized that HNRNPC binds to the U-rich regions in the vicinity of poly(A) sites and globally regulates not only splicing but also pre-mRNA cleavage and polyadenylation. To test this hypothesis, we generated two sets of pre-mRNA 3' end sequencing libraries from HEK 293 cells that were transfected either with a control siRNA or with an siRNA directed against HNRNPC. The siRNA was very efficient, strongly reducing the HNRNPC protein expression, as shown in Supplemental Figure E.6. To evaluate the effect of HNRNPC knock-down on polyadenylation, we focused on exons with multiple poly(A) sites. We identified 12,136 such sites in 4405 exons with a total of 22,698,094 mapped reads (Supplemental Table E.6). We calculated the relative usage of a poly(A) site in a given sample as the proportion of reads that mapped to that site among the reads mapping to any 3' end processing site in the corresponding exon. We then computed the change in relative use of each poly(A) site in si-HNRNPC–treated cells compared with control siRNA–treated cells. We found that HNRNPC knock-down affects a large proportion of transcripts with multiple poly(A) sites, reminiscent of what we previously reported for the 25- and 68-kDa subunits of the cleavage factor I (CFIm) [81, 82]. Out of the 5152 poly(A) sites that showed consistent behavior across replicates, we found 1402 poly(A) sites (27.2%) to increase in usage, 1378 poly(A) sites (26.7%) to decrease in usage, and 2372 poly(A) sites (46.0%) to undergo only a minor change in usage upon knock-down of HNRNPC. To find out whether HNRNPC systematically increases or decreases 3' UTR lengths, we examined the relative position of poly(A) sites whose usage increases or decreases most strongly in response to HNRNPC knock-down, within 3' UTRs. The results indicated that poly(A) sites whose usage increased and decreased upon HNRNPC knock-down tended to be located distally and proximally, respectively, within exons (Fig. 7.2A).

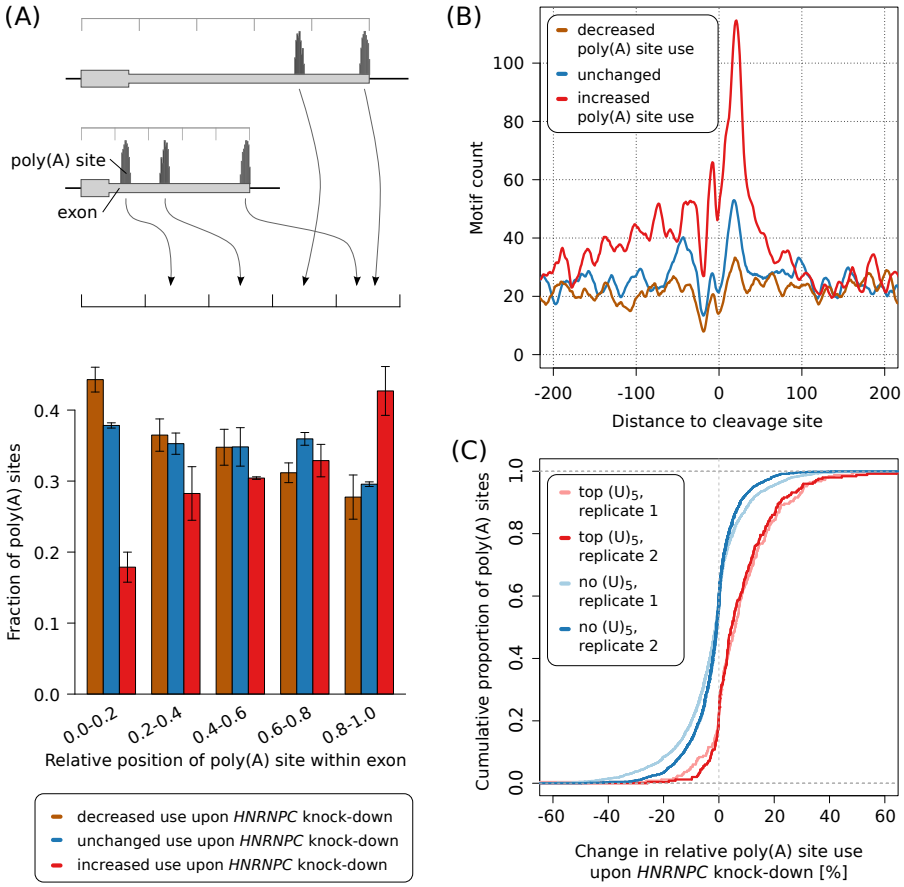


Figure 7.2: siRNA-mediated knock-down of HNRNPC leads to increased use of distal poly(A) sites. (A) Relative location of sites whose usage decreased (brown), did not change (blue) or increased (red) in response to HNRNPC knock-down within 3' UTRs. We identified the 1000 poly(A) sites whose usage increased most, the 1000 whose usage decreased most, and the 1000 whose usage changed least upon HNRNPC knock-down; divided the associated terminal exons into five bins, each covering 20% of the exon's length; and computed the fraction of poly(A) sites that corresponded to each of the three categories within each position bin independently. Values represent means and SDs from the two replicate HNRNPC knock-down experiments. **(B)** Smoothened ($\pm$5 nt) density of nonoverlapping $(U)_5$ tracts in the vicinity of sites with a consistent behavior (increased, unchanged, decreased use) in the two HNRNPC knock-down experiments. **(C)** Cumulative density function of the percentage change in usage of the 250 poly(A) sites with the highest number of $(U)_5$ motifs within $\pm$50 nt around their cleavage site (red) and of poly(A) sites that do not contain any $(U)_5$ tract within $\pm$200 nt (blue), upon HNRNPC knock-down.

We confirmed the overall increase in 3' UTR lengths upon HNRNPC knock-down by comparing the proximal-to-distal poly(A) site usage ratios of exons that had exactly two polyadenylation sites (replicate 1 P-value: 1.1×10^{-19}; replicate 2 P-value: 3.1×10^{-61}; one-sided Wilcoxon signed-rank test) (Supplemental Figs. E.7, E.8). It was noted before that distal poly(A) sites are predominantly used in HEK 293 cells [81]. Indeed, the proportion of dominant (>50% relative usage) distal sites was 61.75% and 62.58%, respectively, in the two control siRNA–treated samples. However, this proportion

increased further in the si-HNRNPC–treated samples to 64.16% and 65.67%, respectively, consistent with HNRNPC decreasing, on average, the lengths of 3' UTRs. Nevertheless, many 3' UTRs became shorter upon this treatment as will be discussed in more detail in the analysis of terminal exons with exactly two poly(A) sites (tandem poly(A) sites) below. As HNRNPC binds RNAs in a sequence-specific manner, one expects an enrichment of HNRNPC binding sites in the vicinity of poly(A) sites whose usage is affected by the HNRNPC knock-down. Indeed, this is what we observed. The density of $(U)_5$ tracts, previously reported to be the binding sites for HNRNPC [452, 505, 508], was markedly higher around poly(A) sites whose usage increased upon HNRNPC knock-down compared with sites whose relative usage did not change or decreased upon HNRNPC knock-down (Fig. 7.2B). No such enrichment emerged from a similar analysis of untransfected versus si-Control transfected cells (Supplemental Fig. E.9)). To exclude the possibility that this profile is due to a small number of regions that are very U-rich, we also determined the fraction of poly(A) sites that contained $(U)_5$ tracts among the poly(A) sites whose usage increased, decreased, or did not change upon HNRNPC knock-down (Supplemental Fig. E.10). We found, consistent with the results shown in Figure 7.2B, a higher proportion of $(U)_5$ tract-containing poly(A) sites among those whose usage increased upon HNRNPC knock-down compared with those whose usage decreased or was not changed. To further validate HNRNPC binding at the derepressed poly(A) sites, we carried out HNRNPC CLIP and found, indeed, that derepressed sites have a higher density of HNRNPC CLIP reads compared with other poly(A) sites (Supplemental Fig. E.11). Finally, we found that poly(A) sites with the highest density of $(U)_5$ tracts in the 100-nt region centered on the cleavage site were reproducibly used with increased frequency upon HNRNPC knock-down relative to poly(A) sites that did not contain any binding sites within 200 nt upstream or downstream (replicate 1 P-value: 2.4×10^{-36}; replicate 2 P-value: 1.9×10^{-42}; one-sided Mann-Whitney U test) (Fig. 7.2C). We therefore concluded that HNRNPC's binding in close proximity of 3' end processing sites likely masks them from cleavage and polyadenylation.

7.6 BOTH THE NUMBER AND THE LENGTH OF THE URIDINE TRACTS CONTRIBUTE TO THE HNRNPC-DEPENDENT POLY(A) SITE USAGE

If the above conclusions were correct, the effect of HNRNPC knock-down should decrease with the distance between the poly(A) site and the HNRNPC binding sites. Thus we determined the mean change in usage of sites with high densities of poly(U) tracts at different distances with respect to the cleavage site, upon HNRNPC knock-down. As shown in Figure 7.3A, we found that the largest change in poly(A) site use is observed for poly(A) sites that have a high density of poly(U) tracts in the 100-nt window centered on the cleavage site.

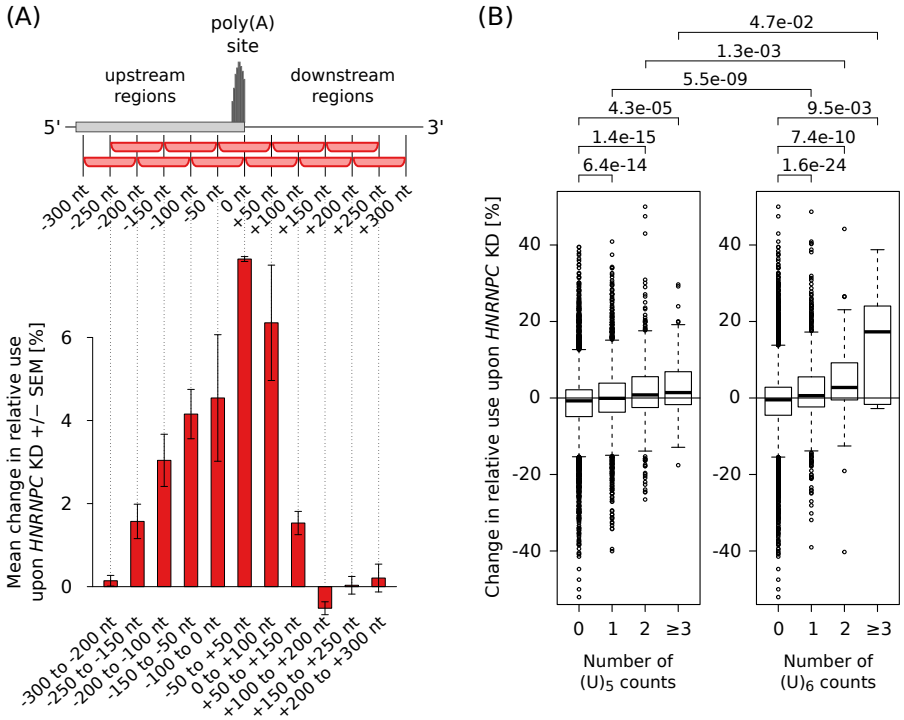


Figure 7.3: The length, number, and location of poly(U) tracts with respect to poly(A) sites influence the change in poly(A) site use upon HNRNPC knock-down. (A) Mean change in the use of sites containing the highest number of $(U)_5$ motifs within 100-nt-long regions located at specific distances from the cleavage site (indicated on the x-axis) upon HNRNPC knock-down (KD). Shown are mean $\pm$SEM in the two knock-down experiments. Two hundred fifty poly(A) sites with the highest density of $(U)_5$ motifs at each particular distance were considered. **(B)** Mean changes in the relative use of poly(A) sites that have 0, 1, 2, or more ($\geq$3) nonoverlapping poly(U) tracts within $\pm$50 nt from their cleavage site. Distributions of relative changes in the usage of specific types of sites were compared, and the P-values of the corresponding one-sided Mann-Whitney U tests are shown at the top of the panel.

The apparent efficacy of HNRNPC binding sites in modulating polyadenylation decreased with their distance to poly(A) sites and persisted over larger distances upstream (approximately -200 nt) of the poly(A) site compared with regions downstream (approximately +100 nt) from the poly(A) site (Fig. 7.3A). Although the minimal RNA recognition motif of HNRNPC consists of five consecutive uridines [452, 507, 508], longer uridine tracts are bound with higher affinity [505–507]. Consistently, we found that, for a given length of the presumed HNRNPC binding site, the effect of the HNRNPC knock-down increased with the number of independent sites and that, given the number of nonoverlapping poly(U) tracts, the effect of HNRNPC knock-down increased with the length of the sites (Fig. 7.3B).

7.7 ALTERED TRANSCRIPT REGIONS CONTAIN ELAVL1 BINDING SITES THAT MEDIATE UDPL

As demonstrated above, binding of HNRNPC to U-rich elements that are located preferentially distally in terminal exons seems to promote the use of proximal 3' end processing sites. Analysis of a conservative set of tandem poly(A) sites showed that among the poly(A) sites that were derepressed upon HNRNPC knock-down and that had at least one $(U)_5$ motif within -200 to +100 nt, two-thirds (390 sites, 67.2%) were located distally, leading to longer 3' UTRs, whereas the remaining one-third (190 sites, 32.8%) were located proximally leading to shorter 3' UTRs (for examples, see Supplemental Figs. E.12, E.13). The altered 3' UTRs contain U-rich elements with which a multitude of RBPs such as ELAVL1,(also known as Hu Antigen R, or HuR) could interact to regulate, among others, the stability of mRNAs in the cytoplasm [435]. To determine whether the HNRNPC-dependent alternative 3' UTRs indeed interact with ELAVL1, we determined the number of ELAVL1 binding sites (obtained from a previous ELAVL1 CLIP study) [510] that are located in the 3' UTR regions between tandem poly(A) sites. As expected, we found a significant enrichment of ELAVL1 binding sites in 3' UTR regions whose inclusion in transcripts changed in response to HNRNPC knock-down compared with regions whose inclusion did not change (Fig. 7.4A).

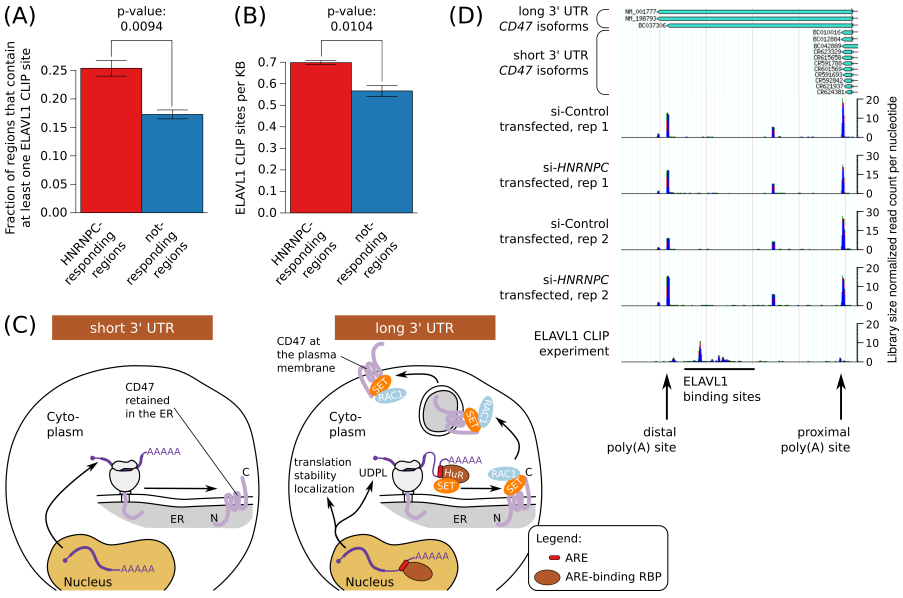


Figure 7.4: HNRNPC-responsive 3' UTRs are enriched in ELAVL1 binding sites. **(A)** Fraction of HNRNPC-responding and not-responding 3' UTR regions that contain one or more ELAVL1 CLIP sites. The P-value of the one-sided t-test is shown. **(B)** Density of ELAVL1 CLIP sites per kilobase (kb) in the 3' UTR regions described above. The P-value of the one-sided t-test is shown. **(C)** Model of the impact of A/U-rich elements (ARE) in 3' UTR regions on various aspects of mRNA fate [105]. **(D)** Density of A-seq2 reads along the CD47 3' UTR in cells, showing the increased use of the distal poly(A) site in si-HNRNPC compared with si-Control transfected cells. The density of ELAVL1 CLIP reads in this region is also shown.

Moreover, the density of ELAVL1 binding sites and not only their absolute number was enriched across these 3' UTR regions (Fig. 7.4B). Our results thus demonstrate that the HNRNPC-regulated 3' UTRs are bound and probably susceptible to regulation by ELAVL1. Recently, a new function has been attributed to the already multifunctional ELAVL1 protein. Work from the Mayr laboratory [105] showed that 3' UTR regions that contain ELAVL1 binding sites can mediate 3' UTR-dependent protein localization (UDPL). The ELAVL1 binding sites in the 3' UTR of the CD47 molecule (CD47) transcript were found to be necessary and sufficient for the translocation of the CD47 transmembrane protein from the endoplasmic reticulum (ER) to the plasma membrane, through the recruitment of the SET protein to the site of translation. SET binds to the cytoplasmic domains of the CD47 protein, translocating it from the ER to the plasma membrane via active RAC1 (Fig. 7.4C; [105, 511]). By inspecting our data, we found that the region of the CD47 3' UTR that mediates UDPL is among those that responded to HNRNPC knock-down (Fig. 7.4D). Sashimi plots generated based on mRNA-seq experiments of HEK 293 cells transfected with si-Control or si-HNRNPC, respectively, confirmed the increased abundance of the long 3' UTR isoform of CD47 upon knock-down of HNRNPC. This analysis also verified that the increased relative usage of distal poly(A) sites cannot be explained by alternative splicing events (Supplemental Fig. E.14) but are the consequence of increased usage of the distal poly(A) site upon knock-down of HNRNPC (Fig. 7.4D). To find out whether HNRNPC can act as an upstream regulator of UDPL, we quantified the level of CD47 at the plasma membrane of cells that underwent siRNA-mediated knock-down of HNRNPC and cells that were treated with a control siRNA. Strikingly, we found that the CD47 level at the plasma membrane increased upon HNRNPC knock-down (Fig. 7.5A; Supplemental Fig. E.15).

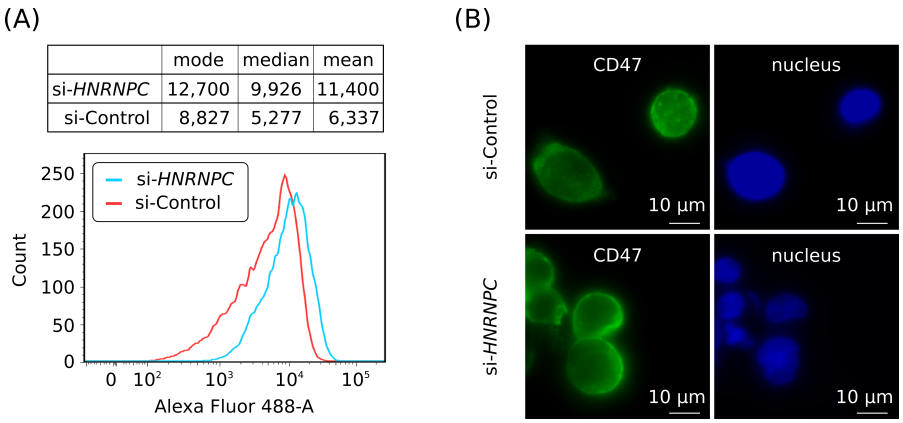


Figure 7.5: The knock-down of HNRNPC affects CD47 protein localization. **(A)** Indirect immunophenotyping of membrane-associated CD47 in HEK 293 cells that were treated either with an si-HNRNPC (blue) or with si-Control (red) siRNA. Mean, median, and mode of the Alexa Fluor 488 intensities computed for cells in each transfection set (top), with histograms shown in the bottom panel. **(B)** Immunofluorescence staining of permeabilized HEK 293 cells with CD47 antibody (left) or nuclear staining with Hoechst (right). Top and bottom panels correspond to cells that were treated with control siRNA and si-HNRNPC, respectively.

Western blots for CD47 that were performed in HNRNPC and control siRNA-treated cells ruled out the possibility that the increase in membrane-associated CD47 upon HNRNPC knock-down was due to an increase in total CD47 levels (Supplemental Fig. E.16). We also carried out an independent immunofluorescence analysis of CD47 in these two conditions and again observed that the HNRNPC knock-down led to an increase in the plasma membrane CD47 levels (Fig. 7.5B). Overall, our results suggest that HNRNPC can function as an upstream regulator of UDPL.

7.8 HNRNPC REPRESSES CLEAVAGE AND POLYADENYLATION AT INTRONIC, TRANSCRIPTION START SITE-PROXIMAL POLY(A) SITES

Up to this point, we focused on alternative polyadenylation (APA) sites that are located within single exons. However, given that HNRNPC binds to nascent transcripts, we also asked whether HNRNPC affects other types of APA, specifically at sites located in regions that in the GENCODE v19 set of transcripts [512] are annotated as intronic. Indeed, we found that the HNRNPC knock-down increased the use of intronic poly(A) sites that are most enriched in putative HNRNPC-binding $(U)_5$ motifs within ± 50 nt compared with sites that do not have $(U)_5$ tracts within ± 200 nt (P-values of the one-sided Mann-Whitney U test for the data from the two replicate knock-down experiments are 1.4×10^{-30} and 5.1×10^{-29}) (Fig. 7.6A).

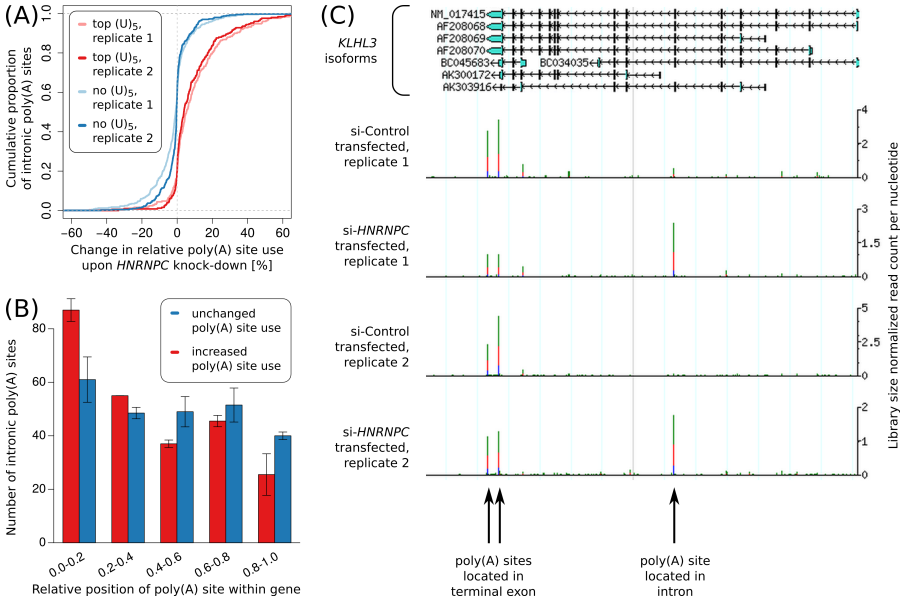


Figure 7.6: HNRNPC knock-down leads to increased usage of intronic poly(A) sites. (A) The change in the relative use of intronic poly(A) sites that did not contain any $(U)_5$ within $\pm$200 nt and of the top 250 intronic poly(A) sites according to the number of $(U)_5$ motifs within $\pm$50 nt around the cleavage site, upon HNRNPC knock-down. **(B)** Relative location within the gene of the top 250 most-derepressed intronic poly(A) sites that have HNRNPC binding motifs within -200 to +100 nt around their cleavage site and of the 250 intronic poly(A) sites that changed least upon HNRNPC knock-down. **(C)** Screenshot of the KLHL3 gene, in which intronic cleavage and polyadenylation was strongly increased upon HNRNPC knock-down.

These sites are predominantly associated with cryptic exons that are spliced in upon HNRNPC knock-down as opposed to exons whose splice site fails to be recognized by the spliceosome leading to exon extension in HNRNPC-depleted cells (Supplemental Fig. E.17). Importantly, only the intronic sites that responded to HNRNPC knock-down were strongly enriched in $(U)_5$ tracts immediately downstream from the poly(A) site (Supplemental Fig. E.18). This indicates that these poly(A) site–associated motifs contribute to the definition of these terminal exons. To further characterize the "masking" effect of HNRNPC on intronic poly(A) sites, we binned poly(A) sites into five groups based on their relative position within the host gene and asked how the position of sites within genes relates to their usage upon HNRNPC knock-down. As shown in Figure 7.6B, we found that intronic poly(A) sites that are most derepressed upon HNRNPC knock-down are preferentially located toward the 5' ends of genes. We conclude that HNRNPC tends to repress the usage of intronic cleavage and polyadenylation sites whose usage leads to a strong reduction of transcript length. Figure 7.6C shows the example of the Kelch Like family member 3 (KLHL3) gene, which harbors one of the most derepressed intronic poly(A) sites.

7.9 DISCUSSION

Studies in recent years have shown that pre-mRNA cleavage and polyadenylation is a dynamically regulated process that yields transcript isoforms with distinct interaction partners, subcellular localization, stability, and translation rate (for review, see, e.g., [513]). Specific polyadenylation programs seem to have evolved in relation with particular cell types or states. For example, APA and 3' UTR lengths are developmentally regulated [430, 431, 514], and short 3' UTRs are generated in proliferating and malignant cells [102, 474, 515]. The key regulators of these polyadenylation programs are unknown. Reduced expression of the U1 snRNP [463] or of the mammalian cleavage factor I (CFIm) components NUDT21 and CPSF6 [81, 82] can cause a systematic reduction in 3' UTR lengths, but only limited evidence about the relevance of these factors in physiological conditions has been provided [463, 471]. Other factors that are part of the 3' end processing machinery and have systematic effects on polyadenylation are the poly(A) binding protein nuclear 1 [453], which suppresses cleavage and polyadenylation; the 64-kDa cleavage stimulation factor subunit 2 (CSTF2) component of the 3' end cleavage and polyadenylation complex, whose expression correlates with the preferential use of short 3' UTRs in cancer cells [515]; and the retinoblastoma binding protein 6, whose reduced expression results in reduced transcript levels and increased use of distal poly(A) sites [516]. Many experimental protocols to capture transcript 3' ends and enable studies of the dynamics of polyadenylation have been developed (for review, see [517]), and consequently, a few databases of 3' end processing sites are available [98, 474, 481]. However, none of these databases has used the entire set of 3' end sequencing data available to date, and thus, their coverage is limited. In this study, we have developed a procedure to automatically process heterogeneous data sets generated with one of nine different protocols, aiming to identify bona fide poly(A) sites that are independently regulated. Although most of the reads that were used to construct the currently available databases [98, 481] map within the poly(A) site clusters that we constructed, the differences at the level of reported processing sites are quite large. This is largely due to the presence of many sites with very limited read support and no upstream poly(A) signals in previous data sets. For example, focusing on the terminal exons of protein-coding genes and lincRNAs from the UCSC GENCODE v19 Basic Set annotation, the human atlas that we constructed has a higher fraction of exons with assigned poly(A) sites compared with previous databases; 71.12% of all terminal exons of protein coding genes in our atlas have at least one annotated poly(A) site in contrast to 66.26% and 62.69% for the studies of Derti et al. [98] and You et al. [481], respectively. The coverage of the terminal exons of lincRNAs is smaller overall but is clearly higher in our atlas (37.59%) compared with those of Derti et al. [98] and You et al. [481] (29.57% and 24.51%, respectively) (Supplemental Fig. E.19). The lower coverage of lincRNAs is probably due to their lower expression in comparison with protein-coding genes [518] and to the fact that some of them are bi-

morphic, appearing in both the poly(A)$^+$ and poly(A)$^-$ fraction [519], and cannot be captured efficiently with protocols that require the presence of a poly(A) tail. Although for the mouse we did not have lincRNA annotations, the general trend of higher coverage in our atlas compared with existing ones holds also for mouse genes (Supplemental Fig. E.20; for detailed numbers, see Supplemental Tables E.7, E.8). The 3' end processing sites reported by other studies [98, 481] but missing from our atlas have, on average, a substantially lower read support. Some were only documented by multimapping reads, had features indicative of internal priming, or originated in regions from which broadly scattered reads were generated. By building upon a large set of 3' end sequencing samples, we have analyzed the sequence composition around high-confidence poly(A) sites to identify elements that may recruit RBPs to modulate polyadenylation. We have identified sequence motifs that exhibit a positional preference with respect to 3' end cleavage sites almost identical to the canonical poly(A) signal AAUAAA. Six of the 10 novel motifs that we found in each human and mouse data set are shared. Not all the poly(A) sites in the atlas that we constructed have one of the 18 conserved signals, which suggests that the set of poly(A) signals is still incomplete. However, with a more comprehensive set of poly(A) signals, we have been able to more efficiently use data from many heterogeneous experiments, thereby achieving a higher coverage of terminal exons and annotated genes by poly(A) sites. Even though the poly(A) and poly(U) motifs are also strongly enriched around poly(A) sites, they were not annotated as poly(A) signals due to positional profiles divergent from what is expected for poly(A) signals. The general A- and U- richness in the vicinity of cleavage and polyadenylation sites has been observed before [490], but the RBP interactors and their role in polyadenylation remain to be characterized. Here we hypothesized that HNRNPC, a protein that binds poly(U) tracts [452, 507, 508] and has a variety of functions including pre-mRNA splicing [505] and mRNA transport [520], also modulates the processing of pre-mRNA 3' ends. HNRNPC has originally been identified as a component of the HNRNP core particle [521, 522] and found to form stable tetramers that bind to nascent RNAs [523]. Systematic evolution of ligands by exponential enrichment (SELEX) experiments have shown that HNRNPC particles bind to uninterrupted tracts of five or more uridines [524], and studies employing CLIP indicated that longer tracts are bound with higher affinity [505]. By sequencing mRNA 3' ends following the siRNA-mediated knock-down of HNRNPC, we found that transcripts that contain poly(U) tracts around their poly(A) sites respond in a manner indicative of HNRNPC masking poly(A) sites. This is reminiscent of the U1 snRNP protecting nascent RNAs from premature cleavage and polyadenylation, in a mechanism that has been called "telescripting" [462, 463]. Indeed, HNRNPC seems to have at least in part a similar function, because the knock-down of HNRNPC increased the incidence of cleavage and polyadenylation at intronic sites, with a preference for intronic sites close to the transcription start. It should be noted that these intronic sites are not spurious but have experimental support as well as polyadenylation signals. Thus, the short transcripts

that terminate at these sites could be functionally relevant, either through the production of truncated proteins or through an effective down-regulation of the functional, full-length transcript forms. In terminal exons, U-rich poly(A) sites whose usage increased upon HNRNPC knock-down tended to be located distally. In these transcripts, HNRNPC may function to "mask" the distal, "stronger" signals, allowing the "weaker" proximal poly(A) sites to be used [525]. Interestingly, the competition between HNRNPC and U2AF2 appears to regulate exonization of Alu elements [506] and, furthermore, impacts polyadenylation at Alu exons [509]. These studies have emphasized the complex cross-talk between regulators that come into play during RNA splicing and polyadenylation [487]. They also illustrate the striking multifunctionality of U-rich and A/U-rich elements that are bound by various proteins at different stages to modulate processes ranging from transcription termination [485] up to protein localization [105]. Initial studies that reported 3' UTR shortening in dividing cells hypothesized that shortened 3' UTRs harbor a reduced number of miRNA binding sites, the corresponding mRNAs being more stable and having an increased translation rate [101, 102]. However, genome-wide measurements of mRNA and protein levels in dividing and resting cells revealed that systematic 3' UTR shortening has a relatively minor impact on mRNA stability, translation, and protein output [103, 429]. Instead, evidence has started to emerge that 3' UTR shortening results in the loss of interaction with various RBPs, whose effects are not limited to mRNA stability and translation [437] but reach as far as the transport of transmembrane proteins to the plasma membrane [105]. The CD47 protein provides a striking example of 3' UTR-dependent protein localization. However, the upstream signals and perhaps additional targets of this mechanism remain to be uncovered. Here we have demonstrated that HNRNPC can modulate polyadenylation of a large number of transcripts, leading to the inclusion or removal of U-rich elements. When these elements remain part of the 3' UTRs, they can be subsequently bound by a variety of U-rich element binding proteins, including ELAVL1, which has been recently demonstrated to play a decisive role in the UDPL of CD47 [105]. Indeed, we found that the knock-down of HNRNPC promoted the expression of the long CD47 3' UTR that is accompanied by an increased membrane localization of the CD47 protein. Although HNRNPC did not appear to target any particular class of transcripts, nearly one-quarter (>23%) of the HNRNPC-responsive transcripts encoded proteins that were annotated with the Gene Ontology category "integral component of membrane" (GO:0016021). Thus, our results provide an extended set of candidates for the recently discovered UDPL mechanism. In conclusion, PolyAsite, available at http://www.polyasite.unibas.ch, is a large and extendable resource that supports investigations into the polyadenylation programs that operate during changes in cell physiology, during development, and in malignancies.

7.10 METHODS

7.10.1 *Uniform processing of publicly available 3' end sequencing data sets*

Publicly available 3' end sequencing data sets were obtained from the NCBI GEO archive (www.ncbi.nlm.nih.gov/geo) and from NCBI SRA (www.ncbi.nlm.nih.gov/sra). To ensure uniform processing of 3' end sequencing data generated by diverse 3' end sequencing protocols, we developed the following computational pipeline (Supplemental Fig. E.21). First, raw sequencing files were converted to FASTA format. For samples generated with protocols that leave a 5' adapter sequence in the reads, we only retained the reads from which the specified adapter sequence could be trimmed. Next, we trimmed the 3' adapter sequence, and when the protocol captured the reverse complement of the RNAs, we reverse complemented the reads. Reads were then mapped to the corresponding genome assembly (hg19 and mm10, respectively) and to mRNA and lincRNA-annotated transcripts (GENCODE v14 release for human [512] and Ensembl annotation of mouse [526], both obtained from UCSC [527] in June 2013). The sequence alignment was done with segemehl with default parameters [528]. In cases where the sex of the organism from which the sample was prepared was female, mappings to the Y Chromosome were excluded from further analysis. For each read, we only kept the mappings with the highest score (smallest edit distance). Mappings overlapping splice junctions were only retained if they covered at least 5 nt on both sides of the junction and they had a higher score compared with any mapping of the same read to the genomic sequence. Based on the genome coordinates of individual exons and the mapping coordinates of reads within transcripts, next we converted read-to-transcript mapping coordinates into read-to-genome mapping coordinates. For generating a high-confidence set of pre-mRNA 3' ends, we started from reads that consisted of no more than 80% of adenines and that mapped uniquely to the genome such that the last 3 nt of the read were perfectly aligned. Furthermore, we required that the 3' end of the read was not an adenine and collapsed the 3' ends of the sequencing reads into putative 3' end processing sites. Finally, we filtered out those sites that showed one of the following patterns: one of the AAAA, AGAA, AAGA, or AAAG tetramers immediately downstream from the apparent cleavage site; or six consecutive or more than six adenines within the 10 nt downstream from the apparent cleavage site. We empirically found that these patterns were associated with many spurious poly(A) sites (for details on the entire pipeline, see Supplemental Fig. E.21).

7.10.2 *Clustering of closely spaced 3' end sites into 3' end processing regions*

Putative 3' end processing sites identified as described above were used to construct clusters to (1) identify poly(A) signals, (2) derive sample-specific

cutoffs for the number of reads necessary to support a site, and (3) determine high-confidence 3' end processing sites in the human and mouse genomes. In clustering putative 3' end processing sites from multiple samples, as done for analyses 1 and 3, we first sorted the list of 3' end sites by the number of supporting samples and then by the total normalized read count (read counts were normalized per sample as reads per million [RPM], and for each site a total RPM was obtained by summing these numbers over all samples). In contrast, to generate clusters of putative reads from individual samples (analysis 2), we only ranked genomic positions by RPM. Clusters were generated by traversing the sorted list from top to bottom and associating lower-ranking sites with a representative site of a higher rank, if the lower-ranked sites were located within a specific maximum distance upstream (d_u) of, or downstream (d_d) from, the representative site (Supplemental Fig. E.22). To determine a maximum distance between sites that seem to be under the same regulatory control, we applied the above-described clustering procedure for distances d_u and d_d varying between 0 and 25 nt and evaluated how increasing the cluster length affects the number of generated clusters that contain more than one site (Supplemental Fig. E.23). Consistent with previous observations, we found that at a distance of 8 nt from the representative site, ∼40% of the putative 3' end processing sites are part of multisite clusters; this proportion increases to 43% for a distance of 12 nt and reaches 47% at a distance of 25 nt. For consistency with previous studies, we used $d_u = d_d = 12$ nt [481, 489]. Only for the clustering of putative 3' end processing sites in individual samples, we used a larger distance, $d_u = d_d = 25$, resulting in a more conservative set of clusters, with a maximum span of 51 nt.

7.10.3 *Identification of poly(A) signals*

To obtain a set of high-confidence 3' end processing sites from which to identify poly(A) signals, we filtered the preliminary 3' end clusters, retaining only those that were supported by data from at least two protocols. For clusters with at least two putative sites, we took the center of the cluster as the representative cleavage site. Then, we constructed the positional frequency profile in the -60 to -5 nt region upstream of the representative cleavage sites for each of the 4096 possible hexamers (Supplemental Fig. E.24A). We did not consider the 5 nt upstream of the putative cleavage sites to reduce the impact of artifacts originating from internal priming at poly(A) nucleotides, which are very close in sequence to the main poly(A) signal, AAUAAA (see below for details on "PAS priming sites"). Before fitting a specific functional form to the frequency profiles, we smoothed them, taking at each position the average frequency in a window of 11 nt centered on that position, and we subtracted a motif-specific "background" frequency which we defined as the median of the 10 smallest frequencies of the motif in the entire 55-nt window. To identify motifs that have a specific positional preference upstream of the cleavage site, we fitted a Gaussian density curve to the background-corrected frequency profile with the "nls" function in R [529], assessing the quality of

the fit by the r^2 value and by the height:width ratio of the fitted peak, where the width was defined as the standard deviation of the fitted Gaussian density (Supplemental Fig. E.24A). Alternative poly(A) signals should have the same positional preference as the main signal, AAUAAA. However, when considering 60 nt upstream of the cleavage site, poly(A) signals can occur not only at -21 nt, which seems to be the preferred location of these signals, but also at other positions, particularly when the poly(A) signal is suboptimal and co-occurs with the main signal. Thus, we started from motifs that peaked in the region upstream of the cleavage site ($r^2 \geq 0.6$ for the fit to the Gaussian and a height:width ratio $\geq$5) but allow a permissive position of the peak, between -40 to -10 nt. Putative poly(A) signals were then determined according to the following iterative procedure (Supplemental Fig. E.24B). We sorted the set of putative signals by their strength. The strongest signal was considered to be the one with the lowest P-value of the test that the peak frequency of the motif could have been generated by Poisson sampling from the background rate inferred as the mean motif frequency in the regions of 100 to 200 nt upstream of and downstream from the cleavage site. As expected, in both human and mouse data sets, the most significant hexamer was the canonical poly(A) signal AAUAAA. Before every iteration, we removed all sequences that contained the most significant signal of the previous iteration in the -60-nt window upstream of the cleavage sites and repeated the procedure on the remaining set of sequences. Signals with an r^2 value of the fit to a Gaussian $\geq$0.9 and a height:width ratio $\geq$4 were retained and the most significant added to the set of potential signals. The fitted Gaussian densities of almost all of the putative poly(A) signals recovered with this procedure had highly similar peak positions and standard deviations. Therefore, only signals that peaked at most 1 nt away from the most significant hexamer, AAUAAA, were retained in the final set of poly(A) signals. The only hexamers that did not satisfy this condition were the AAAAAA hexamer in the mouse and AAAAAA as well as UUAAAA in the human.

7.10.4 *Treatment of putative 3' end sites originating from internal priming*

Priming within A-rich, transcript-internal regions rather than to the poly(A) tail is known to lead to many false-positive sites with most of the existing 3' end sequencing protocols. We tried to identify and eliminate these cases as described above. An underappreciated source of false positives seems to be the annealing of the poly(T) primer in the region of the poly(A) signal itself, which is A-rich and close to the poly(A) site [489, 525]. Indeed, a preliminary inspection of cleavage sites that seemed to lack poly(A) signals revealed that these sites were located on or in the immediate vicinity of a motif that could function as a poly(A) signal. To reduce the rate of false positives generated by this mechanism, we undertook an additional filtering procedure as follows (Supplemental Fig. E.25). First, every 3' end site that was located within a poly(A) signal or had a poly(A) signal starting within 5 nt downstream from the apparent cleavage site was marked initially as "PAS priming

site". Then, during the clustering procedure, each cluster that contained a "PAS priming site" was itself marked as putative internal priming candidate, and the most downstream position of the cluster was considered as the representative site for the cluster. Finally, internal priming candidate clusters were either (1) merged into a downstream cluster, if all annotated poly(A) signals of the downstream cluster were also annotated for the internal priming candidate, or (2) retained as valid poly(A) cluster when the distance between the representative site to the closest poly(A) signal upstream was at least 15 nt or (3) discarded, if neither condition (1) nor (2) was met.

7.10.5 *Generation of the comprehensive catalog of high-confidence poly(A) sites*

7.10.5.1 *Annotating poly(A) signals*

The procedure outlined in the sections above yielded 18 signals that showed a positional preference similar to AAUAAA in both mouse and human. These signals were used to construct the catalog of 3' end processing sites. We started again from all unique apparent cleavage sites from the 78 human and 110 mouse samples (Supplemental Tables E.2, E.3), amounting to 6,983,499 and 8,376,450 sites, respectively. For each of these sites, we annotated all occurrences of any of the 18 poly(A) signals within -60 to +5 nt relative to the apparent cleavage site.

7.10.5.2 *Identification of 3' end processing clusters expressed above background in individual samples*

For each sample independently, we constructed clusters of 3' end processing sites as described above. At this stage, we did not eliminate "PAS priming sites" but rather used a larger clustering distance, of $d_u = d_d = 25$, to ensure that "PAS priming sites" were captured as well. We kept track of whether any 3' end processing site in each cluster had an annotated poly(A) signal or not. Next, we sorted the clusters by the total number of reads that they contained, and by traversing the sorted list from top (clusters with most reads) to bottom, we determined the read count c at which the percentage of clusters having at least one annotated poly(A) signal dropped below 90%. We then discarded all clusters with $\leq c$ read counts as not having sufficient experimental support (for outlines how to determine sample-specific cutoffs, Supplemental Fig. E.26). This allowed for an efficient filtering of reads presumably representing background noise.

7.10.5.3 *Combining poly(A) site clusters from all samples into a comprehensive catalog of 3' end processing sites*

By starting from the sites identified in at least one of the samples, we first normalized the read counts to the total number of reads in each sample to compute expression values as RPM and then merged all sites into a unique

list that we sorted first by the number of protocols supporting each individual site and then by the total RPM across all samples that supported the site. These sites were clustered, and then internal priming candidates were eliminated as described above. Closely spaced clusters were merged (1) when they shared the same poly(A) signals or (2) when the length of the resulting cluster did not exceed 25 nt. The above procedure could result in poly(A) clusters that were still close to each other but with a combined length exceeding the maximum cluster size and that did not have any poly(A) signal annotated. To retain from these the most likely and distinct poly(A) sites, we merged clusters without poly(A) signals with an inter-cluster distance $\leq$12 nt and retained those whose total cluster span was $\leq$50 nt. A small fraction of the clusters had a span $\geq$50 nt, with some even wider than 100 nt. These clusters were not included in the atlas. Finally, the position with the highest number of supporting reads in each cluster was reported as the representative site of the cluster (Supplemental Fig. E.27). The final set of clusters was saved in a BED-formatted file, with the number of supporting protocols as the cluster score. A cluster obtained support by a protocol if any of the reads in the clusters originated from that protocol. We used the protein-coding and lincRNA annotations from the UCSC GENCODE v19 Basic Set for human and the Ensembl mm10 transcript annotation from UCSC for mouse to annotate the following categories of clusters, listed here in the order of their priority (which we used to resolve annotation ambiguity):

- TE: terminal exon,
- EX: any other exon except the terminal one,
- IN: any intron,
- DS: up to 1000 nt downstream from an annotated gene,
- AE: antisense to an annotated exon,
- AI: antisense to an annotated intron,
- AU: antisense and within 1000 nt upstream of an annotated gene, and
- IG: intergenic

7.10.5.4 *Supplemental atlas versions*

To provide more details on different aspects of the inferred poly(A) site clusters, additional versions of the human and mouse atlas with extended information were generated. For human, we established a version that annotated one of the above categories to every poly(A) site cluster based on the UCSC GENCODE v19 Comprehensive Set annotation (not limited to protein-coding and lincRNA-encoding genes). Moreover, for mouse and human, a version with additional information about the tissues/cell types in which each poly(A) site was identified was constructed. All versions are publicly available and online at http://www.polyasite.unibas.ch.

7.10.5.5 *Sequence logos of the identified poly(A) signals*

The procedure described above was used again, this time to construct a version of the human and mouse poly(A) site atlases that incorporated the entire set of 22 organism-specific poly(A) signals, not just the 18 signals that were shared between species. Frequencies of all annotated poly(A) signals (possibly more than one per poly(A) cluster) across all identified clusters were calculated for the human and mouse catalog independently. FASTA files with poly(A) signals, including their multiplicities in the data, were used with the Weblogo program [530] version 3.3, with default settings, to generate the sequence logos for human and mouse, respectively.

7.10.5.6 *Hexamer enrichment in upstream regions of 3' end clusters*

We calculated the significance (P-value) of enrichment of each hexamer in the set of 3' end clusters (and their 60 nt upstream regions) of our human and mouse atlas relative to what would be expected by chance, assuming the mononucleotide frequencies of the sequences and a binomial distribution of motif counts.

7.10.5.7 *Annotation of poly(A) sites with respect to categories of genomic regions*

We used the genomic coordinates of the protein-coding genes and lincRNAs from the UCSC GENCODE v19 Basic Set (human) and the Ensembl mm10 (mouse) annotations to annotate our and previously published sets of poly(A) sites with respect to genomic regions with which they overlap. A poly(A) site was assigned to an annotated feature if at least one of its genomic coordinates overlapped with the genomic coordinates of the feature. **PolyAsite:** For every poly(A) cluster annotated in our catalog, the entire region of the cluster was used to test for an overlap with annotated genomic features.
PolyA-seq: Processed, tissue-specific data were downloaded as a BED file (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30198). Poly(A) sites from nine and five different samples were downloaded for human and mouse, respectively [98]. Mouse genome coordinates were converted to the coordinates of the Ensembl mm10 annotation through LiftOver [531]. The genomic coordinates of all poly(A) sites (one position per poly(A) site) were intersected with the annotation features.
APASdb: Processed, tissue-specific data for human poly(A) sites were downloaded from http://mosas.sysu.edu.cn/utr/download_datasets.php. This included poly(A) sites from 22 human tissues [481]. The genomic coordinates of all poly(A) sites (one position per poly(A) site) were intersected with the annotation features.

7.10.6 *Analysis of 3' end libraries from HNRNPC knock-down experiments*

7.10.6.1 *Sequencing of A-seq2 libraries and quantification of relative poly(A) site usage*

We considered all high-confidence A-seq2 [429] reads that mapped to a unique position in the human genome (hg19) and that had 5' ends that were located in a cluster supported by two or more protocols. For our A-seq2 protocol, high-confidence reads are defined as sequencing reads that do not contain more than two ambiguous bases (N), have a maximum A-content of 80%, and the last nucleotide is not an adenine. By using our atlas of poly(A) sites that was constructed considering the 18 conserved poly(A) signals, we calculated the relative usage of poly(A) sites. We considered in our analysis all exons that had multiple poly(A) clusters expressed at $>$3.0 RPM in one or more samples. There were 12,136 such clusters. We considered as "consistently" changing poly(A) sites those that had a change of at least 5% in the same direction in both replicates. We considered as "consistently" unchanged poly(A) sites those whose mean change and standard deviation across replicates were $<$2%.

7.10.6.2 *Determination of ELAVL1 binding sites that are affected by APA events taking place upon HNRNPC knock-down*

Determination of 3' UTR regions that respond to HNRNPC knock-down: To identify putative HNRNPC regulated regions, we have selected exons that had exactly two poly(A) sites, one of which showing an increase in relative usage by at least 5% upon HNRNPC knock-down and harboring a putative HNRNPC binding site ($(U)_5$) within a region of -200 to 100 nt relative to the cleavage site. We considered as unchanged regions exons with exactly two poly(A) sites, both of which changing $<$5% upon HNRNPC knock-down. ELAVL1 binding site extraction from PAR-CLIP: We used data from a previously published ELAVL1 CLIP experiment [510], Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) database accession GSM714641. Enriched binding sites were determined by applying the mRNA site extraction tool available on CLIPZ [442, 532] using the mRNA-seq samples with GEO accessions GSM714684 and GSM714685 as background. CLIP sites with an enrichment score *geq*5.0 were translated into genome coordinates (hg19) using GMAP [468]. To identify ELAVL1 CLIP sites located within transcript regions that are included/excluded through APA, we intersected the set of enriched ELAVL1 CLIP sites with genomic regions enclosed by tandem poly(A) sites (located on the same exon) using BEDTools [533].

7.10.6.3 *Determination of intronic poly(A) sites*

To make sure that we can capture premature cleavage and polyadenylation events that might occur spontaneously upon knock-down of HNRNPC and are therefore observable in the HNRNPC knock-down samples only, for each

sample we created clusters as described above, using conserved poly(A) signals only. By analogy to tandem poly(A) sites within exons, we calculated the relative usage of clusters within genes by considering all genes having multiple poly(A) clusters that were expressed at $>$3.0 RPM in one or more samples. There were 22,498 such clusters, 2454 of which were annotated to be intronic. Finally, we determined the set of sites that showed a consistent change upon HNRNPC knock-down as described above.

7.10.7 *Experiments*

7.10.7.1 *Cell culture and RNAi*

HEK 293 cells (Flp-In-293, from Life Technologies) were grown in Dulbecco's modified Eagle's medium (DMEM; from Sigma) supplemented with 2 mM L-glutamine (Gibco) and 10% heat-inactivated fetal calf serum (Gibco). Transfections of siRNA were carried out using Lipofectamine RNAiMAX (Life Technologies) following the manufacturer's protocol. The following siRNAs were used: negative-control from Microsynth (sense strand AGGUAGUGUAUCGCCUUGTT) and si-HNRNPC1/2 (sc-35577 from Santa Cruz Biotechnologies), both applied at 20 nM in 2.5 mL DMEM on six-well plates.

7.10.7.2 *Western blotting*

Cells were lysed in 1 $\times$ RIPA buffer, and protein concentration was quantified using BCA reagent (Thermo Scientific). A stipulated amount of the sample (usually 10 μg) was then used for SDS gel separation and transferred to ECL membrane (Protran, GE Healthcare) for further analysis. Membranes were blocked in 5% skim milk (Migros) in TN-Tween (20 mM Tris-Cl at pH 7.5, 150 mM NaCl, 0.05% Tween-20). The following antibodies were used for Western blots: Actin, sc-1615 from Santa Cruz Biotechnology; hnRNP C1/C2 (N-16), sc-10037 from Santa Cruz Biotechnology (used at 1:1000 dilution); CD47, AF-4670 from R&D Systems (used at 1:200 dilution). HRP-conjugated secondary antibodies were applied at 1:2000 dilution. After signal activation with ECL Western blotting detection reagent (GE Healthcare), imaging of Western blots was performed on an Azure c600 system. Signal quantification was done with ImageJ software.

7.10.7.3 *Immunofluorescence*

For the immunofluorescence analysis, HEK 293 cells were transfected with either control siRNA or siRNAs targeting HNRNPC as described under Cell Culture and RNAi, 48 h post transfection cells were fixed with 4% paraformaldehyde for 30 min, permeabilized, and blocked with PBS containing 1% BSA and 0.1% Triton X-100 for 30 min. Primary anti-CD47 antibody (sc-59079 from Santa Cruz Biotechnology) was incubated for 2 h at room temperature at a dilution of 1:100 in the same buffer. To visualize CD47 in cells, secondary antibody conjugated with Alexa Fluor 488 was applied,

while the nucleus was labeled with Hoechst dye. Imaging was performed with a Nikon Ti-E inverted microscope adapted with a LWD condenser (WD 30mm; NA 0.52), Lumencor SpectraX light engine for fluorescence excitation LED transmitted light source. Cells were visualized with a CFI Plan Apochromat DM 60 × lambda oil (NA 1.4) objective, and images were captured with a Hamatsu Orca-Flash 4.0 CMOS camera. Image analysis and edge detection was performed with NIKON NIS Elements software version 4.0. All images were subsequently adjusted uniformly and cropped using Adobe Photoshop CS5.

7.10.7.4 *FACS analysis*

FACS analyses of siRNA transfected cells were performed similar to immunofluorescence studies (see above) except that cells were not permeabilized prior to the treatment with antibody against CD47 (sc-59079 from Santa Cruz Biotechnology). Analysis of Alexa Fluor 488 signal and counts was carried out on a BD FACS Canto II instrument, and data were analyzed with the FLOWJO software. An equal pool of siRNA samples from each transfection set was mixed for the IgG control staining to rule out nonspecific signals.

7.10.7.5 *PAR-CLIP and A-seq2 libraries*

A-seq2 libraries were generated as previously described [429] and sequenced on an Illumina HiSeq 2500 sequencer. The HNRNPC PAR-CLIP was performed as previously described [81] with a modification consisting of preblocking of the Dynabeads-Protein A (Life Technologies), resulting in reduced background and higher efficiency of library generation. To this end, Dynabeads were washed three times with PN8 buffer (PBS buffer with 0.01% NP-40), and incubated in 0.5 mL of PN8-preblock (1 mM EDTA, 0.1% BSA from Sigma [A9647], and 0.1 mg/mL heparin from Sigma [H3393], in PN8 buffer) for 1 h on a rotating wheel. The preblock solution was removed and replaced by the antibody in 0.2 mL preblock solution and rotated for 2-4 h. We used the goat polyclonal antibody sc-10037 against HNRNPC (Santa Cruz Biotechnology). The 5' adapter was GTTCAGAGTTCTACAGTCCGACGATC and the 3' adapter was TGGAATTCTCGGGTGCCAAGG.

7.10.8 *HNRNPC PAR-CLIP analysis*

The raw data were mapped using CLIPZ [510]. For each poly(A) site, the uniquely mapping reads that overlapped with a region of ±50 nt around the cleavage site were counted and normalized (divided) by the expression level (RPKM) of the poly(A) sites host gene using the mRNA-seq samples with GEO accession GSM714684. For Supplemental Figure E.11, normalized CLIP read counts of poly(A) sites belonging to different categories of consistently behaving poly(A) sites across replicates, as defined above, were used.

7.10.9 *Analysis of mRNA-seq libraries from HNRNPC knock-down experiments*

Publicly available libraries of HNRNPC knock-down and control experiments (two replicates) that have been published recently [508] were downloaded from the sequence read archive (SRA) database of the National Center for Biotechnology Information (accession numbers SRX699496/GSM1502498, SRX699497/GSM1502499, SRX699498/GSM1502500, and SRX699499/GSM1502501). After adapter removal, the FASTQ file containing the reads sequenced in sense direction was mapped using the STAR aligner with default settings [534].

7.10.9.1 *Evaluation of novel exon vs. extended internal exon contribution to intronic poly(A) sites*

First we identified all poly(A) sites that were located in introns according to gene structures reflected in the GENCODE v19 (human) transcript set and that were putative HNRNPC targets. That is, they were consistently derepressed upon knock-down of HNRNPC (see above) and contained putative HNRNPC-binding $(U)_5$ motifs within -200 to +100 nt around their cleavage site. For each of these intronic sites, we determined the closest upstream exon, here referred to as u-exon. To find out whether this type of poly(A) sites represented the 3' ends of novel terminal exons or of extended versions of the u-exon, we calculated the ratio $R = \frac{S+1}{C+1}$, where C is the number of reads that map over the 3' end of the u-exon (extending by at least 10 nt in the downstream region), and S is the number of reads that map across a splice boundary, the 5' splice site (ss) being within ± 3 nt of the 3' end of the u-exon and the 3' end of the read mapping upstream of the intronic poly(A) site. The C type of reads provide evidence for the extension of the u-exon, whereas the S type of reads provide evidence for a novel terminal exon. In order to prevent artifacts that may result from poorly expressed transcripts, we required the u-exon to intersect with at least 10 reads within a sample, and we only included regions for which we had at least three reads of either C or S type (or both). We used a pseudo-count of one for both read types.

7.11 AUTHORS INFORMATION

7.11.1 *List of authors*

The following authors have contributed to the work discussed in Chapter 7:

1. Andreas Johannes Gruber[1] (Abbr.: AJG),
2. Ralf Schmidt[1] (Abbr.: RS),
3. Andreas R. Gruber[1] (Abbr.: ARG),
4. Georges Martin[1] (Abbr.: GM),

5. Souvik Ghosh[1] (Abbr.: SG),
6. Manuel Belmadani[1] (Abbr.: MB),
7. Walter Keller[1] (Abbr.: WK) &
8. Mihaela Zavolan[1] (Abbr.: MZ)

whereat author affiliations are as follows:

1 Biozentrum, University of Basel, Klingelberstrasse 50-70, CH-4056 Basel, Switzerland

7.11.2 *Author contributions*

The listing of authors in the previous subsection (7.11.1) was performed according to the authors' contributions, whereat the first author (AJG) contributed most and subsequent authors decreasingly. However, the last two authors are principal investigators and thus their listing follows the opposite ranking (the last contributed the most and the preceding author less).
In detail, using the abbreviations specified in subsection 7.11.1: MZ, AJG, ARG, and WK designed the project. ARG, RS, and AJG collected data sets and created the catalog of poly(A) sites. MB developed the PolyAsite web interface. RS and AJG identified poly(A) signals. GM performed the HNRNPC PAR-CLIP and A-seq2 experiments. AJG analyzed the data with help from RS. GM and SG performed the experiments. MZ supervised the project. AJG, MZ, RS, and ARG wrote the manuscript.

7.12 ACKNOWLEDGMENTS

We thank Erik van Nimwegen for his input on data analysis, Beatrice Dimitriades for technical assistance, and Josef Pasulka for suggestions on the analysis.

7.13 DATA ACCESS

The sequencing data from this study have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) under accession number SRP065825.

7.14 SUPPLEMENTARY MATERIALS

Supplementary materials can be found in Appendix E.

7.15 FUNDING

The work discussed in this chapter was supported by the Swiss National Science Foundation grant 31003A-143977 to Walter Keller and by the Swiss National Science Foundation NCCR project "RNA & Disease" (51NF40_141735).

8

CONCLUSIONS

In Chapter 2 we have reviewed the mechanism of action and the functions of microRNAs (miRNAs). In particular, we pointed out that both epigenetic regulators and transcription factors (TFs) are among the preferred targets of miRNAs. We have highlighted epigenetic regulators whose regulation by miRNAs has been demonstrated experimentally, and introduced their mode of action and impact on gene expression. The targeting of other regulators with genome-wide impact is one possible explanation for the decisive role of miRNAs in fundamental biological processes, such as the establishment of cell fate. In this context, we further reviewed the importance of miRNAs for development, somatic cell reprogramming and pluripotency. A few studies support the idea that the vital role of miRNAs in the establishment of cell fate stems, at least in part, from their targeting of other important regulators that in turn amplify the miRNA effect. For instance, embryonic stem cell (ESC)-specific miRNAs have been demonstrated to regulate the expression of *de novo* DNA methyltransferases, thereby contributing to appropriate DNA methylation levels in ESCs [180, 181].

In Chapter 3 we have presented a study in which we have made use of a computational approach, motif activity response analysis (MARA) [50], which infers the impact of regulatory motifs by modeling genome-wide expression changes as a linear function of the unknown activity of each motif and the number of motif binding sites that are predicted within promoter regions. A related method that was published before, termed "REDUCE", aims to identify *cis*-regulatory elements by modeling gene expression changes in terms of k-mer counts and the unknown activity of each k-mer [272]. However, in contrast to the REDUCE approach, MARA makes use of a curated set of position weight matrices (motifs) that have been inferred from experimental data and reflect the presumed binding specificities of hundreds of transcriptional regulators. Thus, in contrast to REDUCE, MARA is able to infer activities for specific regulatory motifs and designate associated regulators. In its first application, MARA was used to infer the regulatory circuitry that controls differentiation and growth arrest in a human myeloid leukemia cell line [50]. Since then, MARA has been applied to various mammalian systems and several transcriptional networks predicted by MARA have been confirmed experimentally [296–306]. In order to use the MARA approach to study the impact of embryonic miRNAs on the transcriptional landscape of ESCs, in Chapter 3 we have extended the MARA model to include the regulatory effects of miRNAs. Applying this approach to data from ESCs that did or did not express miRNAs we have identified transcriptional regulators that are direct targets of the miRNAs and whose activities also changed significantly as a result. We have identified seven putative miRNA targets, six of which we have confirmed with reporter constructs. In particular, we have pointed

out that the ESC-specific miRNAs regulate multiple cell cycle and epigenetic regulators. Our analysis provides novel insights into the transciption regulatory circuitries that are triggered by and act downstream of miRNAs in ESCs. Moreover, we have experimentally validated one of the transcription regulatory networks that was predicted by MARA, and showed that the miRNAs inhibit NF-κB signaling in ESCs. In summary, our results demonstrate that the extended MARA model is a powerful tool to infer miRNA activities, functional miRNA targets and downstream transcriptional networks.

However, implementing MARA-like approaches is time-consuming and requires the expertise of computational biologists. Thus, in order to make our analysis approach available to other scientists, we have integrated the miRNA-extended MARA model into a fully automated system that has been developed in the van Nimwegen lab over the last years. In Chapter 4 we have introduced this system, called ISMARA (Integrated System for Motif Activity Response Analysis), describing its application to various high-throughput data sets. The system enables researchers that do not have the resources and/or expertise to implement MARA-like models to analyze their own high-throughput data sets, almost as they come out of the sequencing hardware. ISMARA takes no additional input other than the mapped sequencing reads and provides users with a huge array of analysis results, such as activities for transcriptional regulators and their predicted targets, fully annotated within the STRING [312] and gene onthology frameworks [313].

In Chapter 5 we have presented a study which demonstrates that MARA performs also well when applied to data as generated in the clinic. In detail, we have modeled paired liver biopsies from 18 patients, obtained prior to pegylated interferon-α (pegIFN-α) treatment and during the first week post injection, in order to characterize the relative contribution of transcription factor binding motifs to global gene expression changes triggered by pegIFN-α. We have identified the Interferon-stimulated response element (ISRE) to be the most substantially changing motif within all patients. Its activity peaks at 4 and 16 hours post pegIFN-α injection and stays persistently activated within the entire 1-week dosing interval. Besides ISRE, we have found further motifs, including GAS and ATF6, to significantly contribute to the observed gene expression changes triggered by pegIFN-α treatment. As ISRE, GAS has also been shown to play a role in the host response to HCV infection [535, 536]. Furthermore, ATF6 has previously been demonstrated to act in endoplasmic reticulum stress and activation of autophagy upon HCV infection [537–540]. However, in comparison with ISRE, activity changes of other motifs were relatively small. These results are extremely exciting because they demonstrate the utility of ISMARA in the analysis of patient data towards the discovery of molecular mechanisms of disease.

In the future, improved sets of regulatory motifs in combination with more accurate target prediction methods and the inclusion of other regulatory regions such as enhancers will increase the predictive power of MARA-like approaches. Moreover, considering regulatory binding sites in promoter regions of miRNA genes and the integration of small RNA sequencing data

will allow the identification of feedback loops between TFs and miRNAs. Indeed, diverse miRNA-TF network motifs have previously been reported to be critical to various systems and disease, including cancer [541–545]. Up to the present, one bottleneck in this approach was that miRNA promoters are relatively poor characterized, because their processing from pri-miRNAs precludes the capture of promoter-proximal transcript regions. In future the use of Drosha/DGCR8 knockout cells should allow for the enrichment of pri-miRNAs, thereby facilitating accurate identification of miRNA promoter regions.

Importantly, alternative cleavage and polyadenylation (APA), has the potential to determine the presence/absence of miRNA-binding sites within 3' untranslated regions (UTRs) through the choice of alternative 3' end processing sites. Initially, it has been suggested that systematic 3' end shortening, as has been observed in rapidly proliferating cells [100, 101], leads to enhanced mRNA stability and increased protein levels [101, 102]. However, this supposition was later on challenged by a large-scale analysis that reported a surprisingly small impact of 3' end choice on mRNA stability and translational efficiency [103]. Thus, the functional impact of APA on mRNA stability and especially protein output remains poorly understood. In Chapter 6 we have presented the results of a study in which we have characterized the consequences of global 3' end shortening on mRNA and protein levels under physiological conditions. With 3' end sequencing and quantitative proteomics measurements of naive and activated human and murine T cells, we observed the expected systematic global shift towards shorter 3' UTRs taking place within both species [102]. However, we found that shifts towards proximal poly(A) sites generally cause minor changes in mRNA and protein levels. From these findings, the question of the functional impact of APA remains. Apart from mRNA stability and translation, APA has previously been related to mRNA localization [104]. Strikingly, a recent study has demonstrated that, via the 3' UTR-dependent protein localization (UDPL) mechanism, APA is even able to regulate the localization of proteins without changing their amino acid sequence [105]. Consequently, alternative cleavage and polyadenylation likely affects further regulatory levels.

In order to study APA related processes, it is important to have reliable quantifications of poly(A) site usage available. Although many different protocols have been developed that enable mapping and quantification of transcript 3' ends at genome-wide scale, each of these protocols comes along with its specific technical biases and limitations. Consequently, it is crucial to have available a comprehensive and valid catalog of poly(A) site annotations that allow the identification and exclusion of false positive sequencing reads from downstream computational analyses. In Chapter 7 we have introduced such a resource. Incorporating the majority of large-scale 3' end sequencing libraries available to date, we have created a comprehensive catalog of high-confidence cleavage and polyadenylation sites for human and mouse. We have further developed a computational approach, that allowed the identification of 18 conserved variants of the polyadenylation signal, 6 of

which are novel, and all of which exhibit highly specific positioning with respect to cleavage and polyadenylation sites. Moreover, we have identified the well characterized binding motif of HNRNPC, poly(U) [505, 506], to have a specific positional profile around cleavage and polyadenylation sites. Experimentally we have shown that knockdown of HNRNPC causes genome-wide changes in poly(A) site usage. Alternatively regulated 3' UTRs are rich in binding sites for the ELAVL1 RBP and include the region of the CD47 transmembrane protein-encoding transcript that has recently been demonstrated to mediate UDPL [105]. Our results reveal that HNRNPC globally regulates cleavage and polyadenylation and possibly UDPL. Moreover, we provide a comprehensive catalog of cleavage and polyadenylation sites, supplemented with novel signals.

In the future, our "PolyAsite" resource will allow accurate quantification of poly(A) site usage, using annotated sites that were reproducibly identified with multiple protocols. Combining these quantifications with potential *cis*-regulatory elements located in close proximity to cleavage sites, will allow the development of REDUCE-like approaches, which model the changes in poly(A) site usage as a linear function of *cis*-regulatory elements and the unknown activity of these elements. Indeed, during my PhD I started to work on such an approach (not discussed within this thesis). I have implemented an initial version and tested its performance on data sets of poly(A) site usage upon the knock-down of various regulators. It turns out that this REDUCE-like model performs extremely well, being able to identify novel *cis*-regulatory elements (sequence motifs) on which a particular regulator is acting. This suggests a very general approach for the discovery of sequence motifs that are relevant at different levels of gene regulation from appropriate readouts that are obtained upon the knock-down of individual regulators.

In Chapter 4 we have shown that motif activities appear to be more reproducible across replicates than the gene expression levels from which the activities have been inferred. Moreover, our analysis of paired liver biopsies taken from 18 patients (Chapter 5) has impressively demonstrated that inferred activities of key regulatory motifs are strikingly similar across individuals. Thus, it appears that MARA-like models could become powerful tools for the identification of regulators that are involved in particular diseases. Certainly, a major advantage of sophisticated, automatable computational approaches is that they can easily be performed on standard RNA sequencing libraries that can be prepared from biopsies, as by now frequently obtained in the clinic. Hence, MARA-like approaches are also promising to personalized medicine, possibly feeding into diagnosis and/or the development of personalized drugs and therapies. Particularly, the diagnosis and treatment of highly heterogenous disease, such as various types of cancer [546–549], will profit from analysis approaches that are able to infer potential key regulators and associated circuitry from standard sequencing libraries. Furthermore, in combination with recently developed and since then constantly improved single-cell RNA sequencing technologies [550, 551], MARA-like approaches will contribute to

a better understanding of complex, disease associated phenomena, such as tumor heterogeneity [552, 553].

A

SUPPLEMENTARY MATERIAL TO CHAPTER 3

Publication	GEO Id	First Condition	Second Condition
Hanina et al. [252]	GSE20048	Dicer$^{-/-}$	miR-294 transfected Dicer$^{-/-}$
Melton et al. [235]	GSE18840	Dgcr8$^{-/-}$	miR-294 transfected Dgcr8$^{-/-}$
Sinkkonen et al. [180]	GSE8503	Dicer$^{-/-}$	miR-290 cluster transfected Dicer$^{-/-}$
Sinkkonen et al. [180]	GSE7141	Dicer$^{-/-}$	Dicer$^{+/-}$ (miRNA expressing ESCs)
Zheng et al. [250]	GSE30012	Dicer$^{-/-}$	Dicer$^{+/-}$ (miRNA expressing ESCs)

Table A.1: Experimental data sets available for the analysis of miR-290-295 cluster function in embryonic stem cells. Each data set consists in measurements of mRNA expression in a first condition (an ES cell line deficient in mature miRNAs) and a second condition (the respective cell line transfected with specific miRNAs or the cell line expressing the full complement of miRNAs). Accession numbers of the data sets in the Gene Expression Omnibus database [554] are specified.

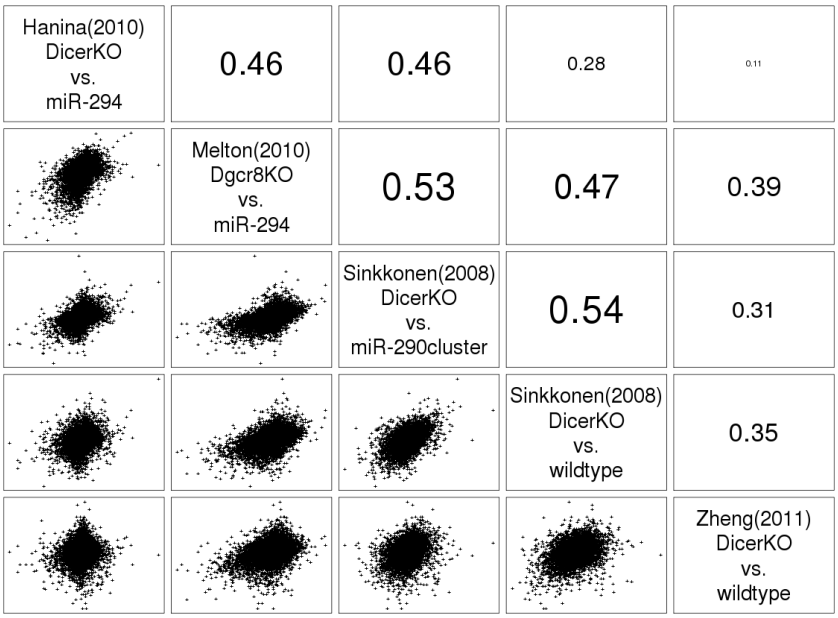


Figure A.1: Summary of the relationship between all the data sets that we considered for our study. Scatter plot / correlation coefficient matrix: The identities of the data sets are indicated on the diagonal. Below the diagonal are scatter plots of $\log_2$ - gene expression fold changes in any pair of experiments. Above the diagonal are the corresponding Pearson correlation coefficients.

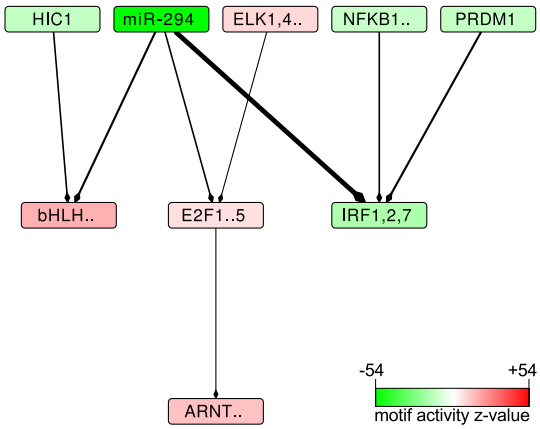


Figure A.2: The transcriptional network inferred through a combined MARA analysis of all five experimental data sets (Supplementary Table A.1) to be affected by the miRNAs of the AAGUGCU seed family (represented by miR-294). A directed edge was drawn from a motif A to a motif B if A was consistently (across data sets) predicted to regulate a transcription factor b whose sequence specificity is represented by motif B. The thickness of the edge is proportional to the product of the probabilities that A targets b across data sets. For the clarity of the figure, only edges with a target probability product > 0.135 and only motifs with absolute z-values > 5 are shown. The intensity of the color of a box representing a motif is proportional to the significance of the motif (the corresponding z-values can be found in Suppl. Table A.5). Red indicates an increase and green a decrease in activity, corresponding to an increased and decreased, respectively, expression of the targets of the motif in the presence of the miRNAs. The full motif names as well as the corresponding transcription factors are listed in Supplementary Table A.7.

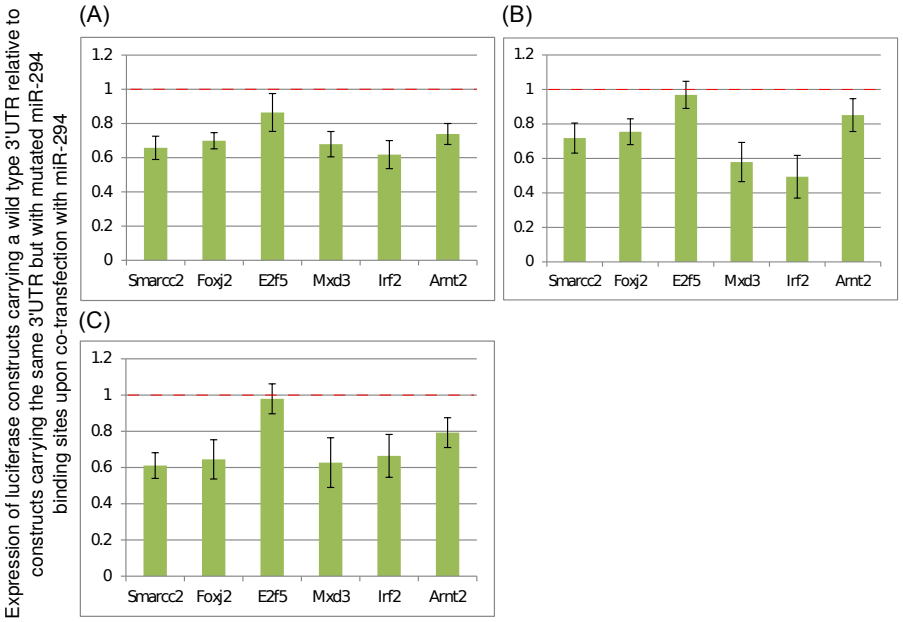


Figure A.3: Results of luciferase assays. (A), **(B)** and **(C)** show the results of three independent experiments in which the effect that the miR-294 mimic has on the expression of constructs containing the 3' UTRs of predicted miR-294 targets downstream of the coding region of the Renilla luciferase relative to a control siRNA (si-Ctrl) was measured. The Firefly luciferase was used as a transfection control. Each experiment consisted in three technical replicates. For each of the wildtype and mutant constructs we calculated the mean (μ) and the standard deviation (σ) of the ratios of Renilla-to-Firefly expression ratio in all pairs of miR-294 vs. si-Ctrl transfections that were done in an individual experiment. We thus obtained, for each target gene, estimates of the mean and standard deviation of the response of the wild type 3'UTR (μ_{wt}, σ_{wt}) and the mutated 3'UTR (μ_{mut}, σ_{mut}). Finally, we calculated the ratio of these means ($r_{wt/mut} = \frac{\mu_{wt}}{\mu_{mut}}$) and the corresponding standard deviation $\sigma_{wt/mut} = r_{wt/mut} * \sqrt{(\sigma_{wt}/\mu_{wt})^2 + (\sigma_{mut}/\mu_{mut})^2}$. The estimates obtained from the three independent experiments are shown in panels **(A)** to **(C)** of the figure.

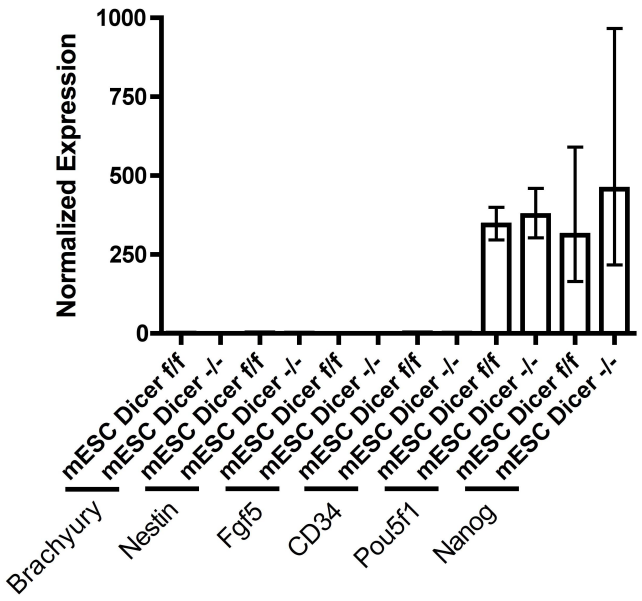


Figure A.4: Expression of markers that are indicative of the pluripotent and differentiation state of mESCs. Total RNA was extracted from the DCR$^{flox/flox}$ and DCR$^{-/-}$ mESCs as described in the methods section. qRT-PCR reactions were run in triplicate using the following primer pairs as described in [555]. Brachyury, Nestin, Fgf5, and CD34 all serve as markers of differentiation and were all very lowly expressed in both stem cell lines, whereas the pluripotency markers Pou5f1 (Oct4) and Nanog were very highly expressed in comparison (±SEM; n=3). Both DCR$^{flox/flox}$ and DCR$^{-/-}$ cells express similar amounts of each marker showing that DCR$^{-/-}$ are not differentiated.

Due to its overlength, the table is not printed here.
Please request it from the author or access it online at:
http://nar.oxfordjournals.org/content/suppl/2014/07/15/gku544.DC1/nar-01400-x-2014-File002.pdf

Table A.2: Genes predicted to be a direct target of the miR AAGUGCU seed family (by TargetScan) and that are significantly downregulated within all three experiments.

Due to its overlength, the table is not printed here.
Please request it from the author or access it online at:
http://nar.oxfordjournals.org/content/suppl/2014/07/15/gku544.DC1/nar-01400-x-2014-File002.pdf

Table A.3: Three experiments combined MARA results **including only the miR AAGUGCU seed family** ranked by absolute activity z-values.

Due to its overlength, the table is not printed here.
Please request it from the author or access it online at:
http://nar.oxfordjournals.org/content/suppl/2014/07/15/gku544.DC1/nar-01400-x-2014-File002.pdf

Table A.4: Three experiments combined MARA results **including all miRNA seed families** ranked by absolute activity z-values.

Due to its overlength, the table is not printed here.
Please request it from the author or access it online at:
http://nar.oxfordjournals.org/content/suppl/2014/07/15/gku544.DC1/nar-01400-x-2014-File002.pdf

Table A.5: Five experiments combined MARA results **including only the miR AAGUGCU seed family** ranked by absolute activity z-values.

Due to its overlength, the table is not printed here.
Please request it from the author or access it online at:
http://nar.oxfordjournals.org/content/suppl/2014/07/15/gku544.DC1/nar-01400-x-2014-File002.pdf

Table A.6: Five experiments combined MARA results **including all miRNA seed families** ranked by absolute activity z-values.

Motif abbreviation	Motif name	Motif binding genes
TEAD1	TEAD1.p2	Tead2, Tead1, Tead4, Tead3
EOMES	EOMES.p2	Eomes
KLF4	KLF4.p3	Klf4
POU5F1..	POU5F1_SOX2{dimer}.p2	Sox2, Pou5f1
ADNP..	ADNP_IRX_SIX_ZHX.p2	Zhx3, Adnp, Zhx1, Irx4, Zhx2, Six2, Irx5, Six5
HOXA4,D4	HOX{A4,D4}.p2	Hoxd4, Hoxa4
BACH2	BACH2.p2	Bach2
POU3F1..4	POU3F1..4.p2	Pou3f4, Pou3f2, Pou3f3, Pou3f1
ZNF143	ZNF143.p2	Zfp143
MYFfamily	MYFfamily.p2	Myog, Myf5, Myf6, Myod1
NFE2	NFE2.p2	Nfe2
T	T.p2	T
NKX3-1	NKX3-1.p2	Nkx3-1
FOXA2	FOXA2.p3	Foxa2
CEBPA,B..	CEBPA,B_DDIT3.p2	Cebpb, Ddit3, Cebpa
EHF	EHF.p2	Ehf
HNF4A..	HNF4A_NR2F1,2.p2	Hnf4a, Nr2f1, Nr2f2
CRX	CRX.p2	Crx
TFDP1	TFDP1.p2	Tfdp1
GFI1B	GFI1B.p2	Gfi1b
STAT1,3	STAT1,3.p3	Stat3, Stat1
NKX3-2	NKX3-2.p2	Nkx3-2
STAT2,4,6	STAT2,4,6.p2	Stat4, Stat2, Stat6
RBPJ	RBPJ.p2	Rbpj
RREB1	RREB1.p2	Rreb1
FOS..	FOS_FOS{B,L1}_JUN{B,D}.p2	Junb, Fosl1, Fos, Fosb, Jund
SREBF1,2	SREBF1,2.p2	Srebf2, Srebf1
FOXL1	FOXL1.p2	Foxl1
HSF1,2	HSF1,2.p2	Hsf2, Hsf1
ARID5B	ARID5B.p2	Arid5b
GTF2I	GTF2I.p2	Gtf2i
ESRRA	ESRRA.p2	Esrra
ONECUT1,2	ONECUT1,2.p2	Onecut1, Onecut2
NKX2-3..	NKX2-3_NKX2-5.p2	Nkx2-5, Nkx2-3
GATA6	GATA6.p2	Gata6
ZBTB16	ZBTB16.p2	Zbtb16
AIRE	AIRE.p2	Aire
SMAD1....	SMAD1..7,9.p2	Smad4, Smad5, Smad9, Smad7, Smad6, Smad2, Smad3, Smad1
BPTF	BPTF.p2	Bptf
CDC5L	CDC5L.p2	Cdc5l
NKX6-1,2	NKX6-1,2.p2	Nkx6-2, Nkx6-1
NANOG	NANOG.p2	Nanog
PAX4	PAX4.p2	Pax4
NFYA,B,C	NFY{A,B,C}.p2	Nfya, Nfyb, Nfyc
EP300	EP300.p2	Ep300
TAL1..	TAL1_TCF{3,4,12}.p2	Tcf4, Tcf12, Tcf3, Tal1
GZF1	GZF1.p2	Gzf1
ATF4	ATF4.p2	Atf4
PAX3,7	PAX3,7.p2	Pax7, Pax3
NKX2-2,8	NKX2-2,8.p2	Nkx2-2, Nkx2-9
AHR..	AHR_ARNT_ARNT2.p2	Ahr, Arnt, Arnt2
NFE2L1	NFE2L1.p2	Nfe2l1
MYB	MYB.p2	Myb
NFIX	NFIX.p2	Nfix
ELK1,4..	ELK1,4_GABP{A,B1}.p3	Elk1, Gabpa, Gabpb1, Elk4
KLF12	KLF12.p2	Klf12
SPZ1	SPZ1.p2	Spz1
GFI1	GFI1.p2	Gfi1
MAZ	MAZ.p2	Maz
PRDM1	PRDM1.p3	Prdm1
PAX8	PAX8.p2	Pax8

HIC1	HIC1.p2	Hic1
FOSL2	FOSL2.p2	Fosl2
EWSR1-F..	EWSR1-FLI1.p2	Fli1, Ewsr1
TP53	TP53.p2	Trp53
FOXN1	FOXN1.p2	Foxn1
JUN	JUN.p2	Junb, Jun, Jund
AR	AR.p2	Ar
ZNF423	ZNF423.p2	Zfp423
FOXC1,C2	FOX{C1,C2}.p2	Foxc1, Foxc2
VSX1,2	VSX1,2.p2	Vsx1, Vsx2
MZF1	MZF1.p2	Mzf1
SP1	SP1.p2	Sp1
ELF1,2,4	ELF1,2,4.p2	Elf4, Elf1, Elf2
SOX5	SOX5.p2	Sox5
XBP1	XBP1.p3	Xbp1
REST	REST.p3	Rest
TFAP4	TFAP4.p2	Tcfap4
ETS1,2	ETS1,2.p2	Ets2, Ets1
NR3C1	NR3C1.p2	Nr3c1
HMGA1,2	HMGA1,2.p2	Hmga1, Hmga2
EN1,2	EN1,2.p2	En2, En1
ALX1	ALX1.p2	Alx1
ZNF384	ZNF384.p2	Zfp384
HOXA5,B5	HOX{A5,B5}.p2	Hoxa5, Hoxb5
IKZF1	IKZF1.p2	Ikzf1
TCF4..	TCF4_dimer.p2	Tcf4
CUX2	CUX2.p2	Cux2
IRF1,2,7	IRF1,2,7.p3	Irf2, Irf1, Irf7
HBP1..	HBP1_HMGB_SSRP1_UBTF.p2	Ssrp1, Hmgb3, Hmgb2, Hbp1, Ubtf
TLX2	TLX2.p2	Tlx2
SRF	SRF.p3	Srf
HMX1	HMX1.p2	Hmx1
MEF2A,B..	MEF2{A,B,C,D}.p2	Mef2a, Mef2c, Mef2d, Mef2b
GLI1..3	GLI1..3.p2	Gli2, Gli1, Gli3
NKX2-1,4	NKX2-1,4.p2	Nkx2-4, Nkx2-1
FOXQ1	FOXQ1.p2	Foxq1
TBP	TBP.p2	Tbp
TLX1..3..	TLX1..3_NFIC{dimer}.p2	Tlx1, Tlx3
PBX1	PBX1.p2	Pbx1
EGR1..3	EGR1..3.p2	Egr1, Egr2, Egr3
PDX1	PDX1.p2	Pdx1
ESR1	ESR1.p2	Esr1
LHX3,4	LHX3,4.p2	Lhx4, Lhx3
ZNF148	ZNF148.p2	Zfp148
HLF	HLF.p2	Hlf
FOXI1,J2	FOX{I1,J2}.p2	Foxi1, Foxj2
SPIB	SPIB.p2	Spib
FEV	FEV.p2	Fev
NFIL3	NFIL3.p2	Nfil3
NFE2L2	NFE2L2.p2	Nfe2l2
MAFB	MAFB.p2	Mafb
ATF2	ATF2.p2	Atf2
HOXA6,A..	HOX{A6,A7,B6,B7}.p2	Hoxb6, Hoxb7, Hoxa6, Hoxa7
GTF2A1,2	GTF2A1,2.p2	Gtf2a1, Gtf2a2
TFEB	TFEB.p2	Tcfeb
SRY	SRY.p2	Sry
TGIF1	TGIF1.p2	Tgif1
RORA	RORA.p2	Rora
SNAI1..3	SNAI1..3.p2	Snai2, Snai3, Snai1
HNF1A	HNF1A.p2	Hnf1a
RXRA,B,G	RXR{A,B,G}.p2	Rxrb, Rxrg, Rxra
TFCP2	TFCP2.p2	Tcfcp2
IKZF2	IKZF2.p2	Ikzf2
NR1H4	NR1H4.p2	Nr1h4

ZIC1..3	ZIC1..3.p2	Zic3, Zic2, Zic1
SPI1	SPI1.p2	Sfpi1
POU1F1	POU1F1.p2	Pou1f1
NR5A1,2	NR5A1,2.p2	Nr5a2, Nr5a1
NFKB1..	NFKB1_REL_RELA.p2	Rel, Rela, Nfkb1
CDX1,2,4	CDX1,2,4.p2	Cdx1, Cdx2, Cdx4
FOXF1,F..	FOX{F1,F2,J1}.p2	Foxj1, Foxf1a, Foxf2
TFAP2A,C	TFAP2{A,C}.p2	Tcfap2a, Tcfap2c
NFATC1..3	NFATC1..3.p2	Nfatc2, Nfatc1, Nfatc3
ZFP161	ZFP161.p2	Zfp161
ATF6	ATF6.p2	Atf6
MYOD1	MYOD1.p2	Myod1
ARNT..	ARNT_ARNT2_BHLHB2 _MAX_MYC_USF1.p2	Myc, Bhlhe40, Arnt, Arnt2, Usf1, Max
STAT5A,B	STAT5{A,B}.p2	Stat5a, Stat5b
FOXD1,D2	FOX{D1,D2}.p2	Foxd1, Foxd2
POU2F1..3	POU2F1..3.p2	Pou2f2, Pou2f1, Pou2f3
HES1	HES1.p2	Hes1
EVI1	EVI1.p2	Mecom
PRRX1,2	PRRX1,2.p2	Prrx2, Prrx1
FOXD3	FOXD3.p2	Foxd3
RXRA..	RXRA_VDR{dimer}.p2	Vdr
ZBTB6	ZBTB6.p2	Zbtb6
RUNX1..3	RUNX1..3.p2	Runx1, Runx2, Runx3
ZNF238	ZNF238.p2	Zfp238
PATZ1	PATZ1.p2	Patz1
LMO2	LMO2.p2	Lmo2
DBP	DBP.p2	Dbp
CTCF	CTCF.p2	Ctcf
PAX6	PAX6.p2	Pax6
PAX5	PAX5.p2	Pax5
PPARG	PPARG.p2	Pparg
PAX2	PAX2.p2	Pax2
CREB1	CREB1.p2	Creb1
NRF1	NRF1.p2	Nrf1
ZEB1	ZEB1.p2	Zeb1
YY1	YY1.p2	Yy1
NHLH1,2	NHLH1,2.p2	Nhlh1, Nhlh2
SOX17	SOX17.p2	Sox17
TBX4,5	TBX4,5.p2	Tbx5, Tbx4
..SMARC	DMAP1_NCOR{1,2}_SMARC.p2	Ncor2, Ncor1, Smarca1, Smarca5, Smarcc2, Dmap1
MYBL2	MYBL2.p2	Mybl2
RXRG..	RXRG_dimer.p3	Pparg, Rxrb, Rxrg, Ppara, Rxra, Nr1h2
PITX1..3	PITX1..3.p2	Pitx3, Pitx2, Pitx1
HOXA9..	HOXA9_MEIS1.p2	Meis1, Hoxa9
ATF5..	ATF5_CREB3.p2	Atf5, Creb3
SOX8,9,10	SOX{8,9,10}.p2	Sox8, Sox9, Sox10
POU6F1	POU6F1.p2	Pou6f1
POU5F1	POU5F1.p2	Pou5f1
MSX1,2	MSX1,2.p2	Msx2, Msx1
bHLH..	bHLH_family.p2	Hey2, Hey1, Olig2, Heyl, Id1, Mxi1, Mitf, Clock, Tcfe3, Mnt, Arntl, Mlxipl, Olig1, Mxd4, Arntl2, Npas2, Mxd3, Hes6
GATA1..3	GATA1..3.p2	Gata3, Gata1, Gata2
TFAP2B	TFAP2B.p2	Tcfap2b
EBF1	EBF1.p2	Ebf1
MTF1	MTF1.p2	Mtf1
FOXO1,3,4	FOXO1,3,4.p2	Foxo4, Foxo1, Foxo3
E2F1..5	E2F1..5.p2	E2f2, E2f5, E2f3, E2f4, E2f1
LEF1..	LEF1_TCF7_TCF7L1,2.p2	Lef1, Tcf7l2, Tcf7l1, Tcf7
HIF1A	HIF1A.p2	Hif1a
RFX1..5..	RFX1..5_RFXANK_RFXAP.p2	Rfx3, Rfxank, Rfx1, Rfx2, Rfxap, Rfx4, Rfx5
SOX2	SOX2.p2	Sox2
NANOGmo..	NANOG{mouse}.p2	Nanog

NR6A1	NR6A1.p2	Nr6a1
FOXP3	FOXP3.p2	Foxp3
HAND1,2	HAND1,2.p2	Hand2, Hand1

Table A.7: Motif information table. Motif name, motif name abbreviation and transcription factor(s) binding the motif.

Name	Motif	Motif abbreviation	z-value
Irf2	IRF1,2,7.p3	IRF1,2,7	-17.49
Mxd3	bHLH_family.p2	bHLH..	16.54
Clock	bHLH_family.p2	bHLH..	16.54
E2f5	E2F1..5.p2	E2F1..5	6.98

Table A.8: AAGUGCU seed family transcription factor targets as predicted by combined MARA. Transcription factors consistently predicted (within all five experiments) by MARA to be a direct target of miR-294 and whose absolute motif activity z-value is > 5 (due to the presence of AAGUGCU seed family miRNAs).

Gene	Orientation	Cloning primers
Arnt2	Forward	GAATTCCTCGAGCCACTGGCAACCAAGCACA
Arnt2	Reverse	CAGGACATAAGCGGCCGCCTCAAGTTGATCAATTACCAA
E2f5	Forward	GAATTGCTCGAGATAATGAAGGAGTTTGTGATCTGTT
E2f5	Reverse	AGGACATATGCGGCCGCTCAGTTGTAATACAGAGATAGTGCAT
Foxj2	Forward	GATGTAGCTCGAGATTCGAGAGAGGGAAATCTCACTT
Foxj2	Reverse	GAGTGAAATGCGGCCGCGGAAACCAAAGGGAAGCTGT
Irf2	Forward	GTATCTCTCGAGCCCGTGTCAAGAGCTGTTAA
Irf2	Reverse	CAGGAGAAATGCGGCCGCCATGATTACGCTCTAAGTAGACACA
Mxd3	Forward	GAATTCCTCGAGCGGGAGCACAGCTACTCA
Mxd3	Reverse	CAGGACATAAGCGGCCGCAGTGCGCAGGTGATAGACTGTA
Smarcc2	Forward	GAATTCCTCGAGCTCAGCCTGAAGAGTTCATCACTA
Smarcc2	Reverse	GACCAGAAAGCGGCCGCCTGCCTAGCAGCCACAGCTAA
Zfp238	Forward	CAAACTCTCGAGGTACGGTCTAAAAGCAGTCTTGTT
Zfp238	Reverse	TAAACTGCGGCCGCCAAATCTGTTGTGCGACTAT

Table A.9: Primers used for cloning.

Gene	Orientation	Mutagenesis primers
Arnt2	Forward	GGCTTCCAAGAACAGCAAACTCGTCTCTCTCTTAGCC
Arnt2	Reverse	GGCTAAGAGAGAGACGAGTTTGCTGTTCTTGGAAGCC
E2f5	Forward	GTGAAGTGCCTTCTGTTTTAGAAGCCTATCAGTTTGTTGAC
E2f5	Reverse	GTCAACAAACTGATAGGCTTCTAAAACAGAAGGCACTTCAC
Foxj2	Forward	GCAGTTCACTAAAGAGTTATTTCTTTGTAAGG
Foxj2	Reverse	CCTTACAAAGAAATAACTCTTTAGTGAACTGC
Irf2	Forward	GCACCTTATCTTGAAGTACAATAGGCCTTCTTG
Irf2	Reverse	CAAGAAGGCCTATTGTACTTCAAGATAAGGTGC
Mxd3	Forward	GGAATTTCATGTAGCGGCCCTGCTTTGCTGC
Mxd3	Reverse	GCAGCAAAGCAGGGCCGCTACATGAAATTCC
Smarcc2	Forward	CCTCAAGTTTGAAAAGCAGCACCTACTTTTGACAG
Smarcc2	Reverse	CTGTCAAAAGTAGGTGCTGCTTTTCAAACTTGAGG
Zfp238	Forward	GTTGGGATCTTAACTTGTTTTTGTAGAATAATAGCATGAGAATCTCAC
Zfp238	Reverse	GTGAGATTCTCATGCTATTATTCTACAAAAACAAGTTAAGATCCCAAC

Table A.10: Primers used for mutagenesis.

B

SUPPLEMENTARY MATERIAL TO CHAPTER 4

B.1 SUPPLEMENTARY METHODS

B.1.1 *Human and mouse promoteromes*

The central entities whose regulation is modeled by ISMARA are *promoters*. When analyzing expression data, be they micro-array or RNA-seq, ISMARA estimates and models the expression profiles of individual promoters, and when analyzing ChIP-seq data ISMARA models the chromatin state of genomic regions centered on promoters. Thus, the first step in the analysis consists of the construction of reference sets of promoters in human and mouse. To make a comprehensive list of promoters we used two sources of data: deepCAGE data, i.e. next-generation sequencing data of 5' ends of mRNAs [556, 557], and the 5' ends of all known mRNAs listed in GenBank.
Using CAGE data from a considerable set of human and mouse tissues, we recently constructed genome-wide human and mouse 'promoteromes' [293] consisting of a hierarchy of individual transcription start sites (TSSs), transcription start clusters (TSCs) of nearby co-regulated TSSs, and transcription start regions (TSRs), which correspond to clusters of TSCs with overlapping proximal promoter regions. As the basis of our promoter sets we started with the sets of TSCs, i.e. local clusters of TSSs whose expression profiles are proportional to each other to within experimental noise, as identified by deepCAGE.
As the currently available CAGE data do not yet cover all cell types in human and mouse, a substantial number of cell type-specific promoters are not represented within this set of TSCs. We thus supplemented the TSCs with all 5' ends of mRNAs, using the BLAT [558] mappings from UCSC Genome Browser web site [559]. To avoid transcripts whose 5' ends are badly mapped, we filtered out those for which more than 25 bases at the 5' end of the transcript were unaligned. We then produced reference promoter sets by iteratively clustering the TSCs with the 5' ends of mRNAs as follows: Initially each TSC and each 5' end of an mRNA forms a separate cluster. At each iteration the pair of nearest clusters are clustered, with the constraint that there can be at most one TSC per cluster. That is, we never cluster two TSCs together as our previous analysis in [293] has already established that each TSC is independently regulated. Here the distance between two clusters is defined as the distance between the nearest pair of TSSs of the two clusters, i.e. the distance between the rightmost TSS of the left cluster and leftmost TSS of the right cluster. This iteration is repeated until the distance between the closest pair of clusters is larger than 150 base pairs. Note that we thus chose the length of sequence wrapped by a single nucleosome, i.e. roughly

150 base pairs, as an *ad hoc* cut-off length for two TSSs to belong to a common promoter. The reasoning behind this choice of cut-off, is that, on the one hand, we have empirically observed that co-expressed TSSs can spread over roughly this length-scale and, on the other hand, that it is not implausible that the ejection of a single nucleosome near the TSS may be responsible for setting this length scale. In any case, the resulting promoters are not sensitive to the precise setting of this cut-off (data not shown). Finally, inspection of the results showed, especially in ubiquitously expressed genes, many apparent TSSs from Genbank that appear downstream of both the TSSs identified by deep-CAGE and the annotated RefSeq transcripts. It is highly likely that many of these apparent TSSs are due to cDNA sequences that were not full length. Indeed, only a small fraction of the transcripts in the database of mRNAs underwent expert curation, and truncated transcripts are likely common. To avoid such spurious apparent TSSs we removed all clusters which did not contain at least one curated transcript (RefSeq) or a TSC. Finally, since a sequence of at least one associated transcript is necessary to estimate a promoter's expression level from either RNA-seq or micro-array data, we also discarded all promoters that consisted solely of a TSC.

For human, the resulting reference promoter set had $36'383$ promoters, of which $13'265$ contained both a TSC and at least one RefSeq transcript, $14'538$ contained only a TSC together with non-RefSeq transcripts, and $8'580$ had at least one RefSeq transcript and potentially non-RefSeq transcripts, but no TSC. For the mouse genome, the corresponding numbers are: $34'050$ promoters in total, $8'578$ RefSeq-only, $12'303$ TSC-only, and $13'169$ with both a TSC and at least one RefSeq transcript. These reference promoters sets cover almost all known protein-coding genes in human and mouse.

Finally, as we discussed in [293], mammalian promoters clearly fall into two classes associated with high and low content of CpG dinucleotides, and these promoter classes have clearly distinct architectures, i.e. different lengths, different numbers of TSSs per promoters, and different distributions of transcription factor binding sites (TFBSs). We classified all promoters into a high-CpG and low-CpG class based on both the CG content and the CpG content in the proximal promoter, as described in [293]. In the TFBS prediction below we perform separate predictions for high-CpG and low-CpG promoters.

B.1.2 *A curated set of regulatory motifs*

We use standard position dependent weight matrices (WMs) to represent regulatory motifs, i.e. the sequence specificities of TFs. Each WM is named for the TFs that are annotated to bind its site. For some motifs the names correspond to multiple TFs which are all assumed to bind to the same sites. We used a partly manual curation procedure whose details were first described in [50]. For completeness, we here also give a description of this curation procedure.

For a number of reasons regarding data quality and annotation ambiguities, the construction of a set of position-specific weight matrices (WMs) for mam-

malian transcription factors is rife with problems that, in our opinion, do not currently have a clean solution. Therefore, our procedures necessarily involve several subjective choices, judgments, and hand-curation, which are certainly far from satisfactory.
Our main objectives were

1. To remove redundancy, we aim to have no more than 1 WM representing any given TF. Whenever multiple TFs have WMs that are statistically indistinguishable or when their DNA binding domains are virtually identical, then we use only one WM for that set of TFs.

2. To associate WMs with TFs based on the sequences of their DNA binding domains. That is, we obtain lists of TFs that can plausibly bind to the sites of a given WM by comparison of DNA binding domain sequences of TFs known to bind to the sites with those of all other TFs.

3. Re-estimation of WMs using genome-wide predictions of regulatory sites in the proximal promoters of CAGE TSSs.

The input data for our WM construction consisted of

1. The collection of JASPAR vertebrate WMs plus, for each WM, the amino acid sequence of the TF that JASPAR associates with the WM [560].

2. The collection of TRANSFAC vertebrate WMs (version 9.4) and the amino acid sequences of all vertebrate TFs in TRANSFAC that are associated with those WMs [286].

3. A list of 1322 human TFs (Entrez gene IDs) and their amino acid sequences (from RefSeq).

4. A list of 483 Pfam IDs corresponding to DNA binding domains and their Pfam profiles [561].

We decided not to include 6 TRANSFAC motifs, which were constructed out of less than 8 sites: M00326 (PAX1, PAX9), M00619 (ALX4), M00632 (GATA4), M00634 (GCM1, GCM2), M00630 (FOXM1), M00672 (TEF). TRANSFAC often associates multiple WMs with a single human TF. Although there undoubtedly are cases where a single TF can have multiple distinct modes of binding DNA, and could therefore be realistically represented by multiple WMs, we believe that for the very large majority of TFs it is more realistic to describe the DNA binding specificity of the TF with a single WM. Indeed, a manual inspection of cases in which TRANSFAC associated multiple WMs with a single TF shows that these WMs are typically highly similar and appear redundant. Therefore, we decided to remove this redundancy and for each TF with multiple WMs in TRANSFAC we choose only a single ‘best’ WM based on TRANSFAC’s own matrix quality annotation, or WM information score when there were multiple WMs with the same quality score. The information score of a WM is given by 2 times the length of the WM minus its entropy in bits.

We next aimed to obtain, for each human TF, a list of WMs from JASPAR/-TRANSFAC, that can potentially be associated to this TF. To do this we aim to find, for each TF, which motifs from JASPAR/TRANSFAC are associated with a TF that has a highly similar DNA binding domain. To this end we we ran Hmmer [562] with the DNA binding domain (DBD) profiles from Pfam to extract the DBDs from all TFs (E-value cut-off 10^{-9}) associated with either JASPAR or TRANSFAC matrices. We then represented each such TF with the union of its DNA binding domain sequences. Next we used BLAT to map the DBDs of all TFs associated with JASPAR/TRANSFAC matrices against the entire protein sequences of all human TFs. For each human TF we then extracted a list of all JASPAR/TRANSFAC matrices for which the DBDs of at least one associated TF has a significant BLAT hit (default parameters) against the TF sequence. For each human TF the associated WMs were ordered by the percent identity of the hit, i.e. the fraction of all amino acids in the DBDs that map to matching amino acids in the TF.
From this data we created a list of 'necessary WMs' as follows. For each human TF we obtain the JASPAR WM with the highest percent identity in the DBDs of an associated TF. If there is a TRANSFAC WM with a higher percent identity than any JASPAR TF we record this WM as well. Thus, the necessary WMs are those that are the best match for at least one human TF. This list yielded 381 WMs representing 980 human TFs (often the same WM is the best match for multiple TFs). Manual inspection indicated that a lot of redundancy (essentially identical looking WMs) remained in this list. First we often have both a TRANSFAC and a JASPAR WM for the same TF and moreover often there are multiple TFs, each with its own WM, that look essentially identical. We thus want to fuse WMs in the following situations

1. Different WMs for TFs with identical or near identical DBDs.

2. WMs that are statistically indistinguishable, predict highly overlapping sets of sites, and are associated with TFs that have similar DBDs.

For each pair of WMs we obtained three similarity measurements

1. The percent identity of the DBDs of the TFs associated with the WMs. If there are multiple TFs associated with a WM we take the maximum over all TF pairs.

2. The overlap of the binding sites predicted by each WM. We use MotEvo to predict TFBSs in all proximal promoters and we calculate what fraction of predicted TFBS positions are shared between the two WMs.

3. A statistical measure of the similarity of the two WMs. Here we take the two sets of sites that define the two WMs and calculate the likelihood-ratio of these sites assuming they either derive from a single underlying WM or assuming that the set of sites for each WM derives from an independent WM.

For each of these three criteria we set a cut-off: 95% identity of the DBDs, 60% overlap of predicted TFBSs, and a likelihood-ratio of e^{40}. Using single-linkage clustering, we cluster all WMs whose similarity is over the cut-off for at least 1 of these three criteria. The resulting clusters were then all checked manually and whenever the linkage was dubious we split the cluster. That is, we took a conservative attitude towards removing redundancy and only kept clusters when we were convinced the WMs were essentially identical. For each cluster we then constructed a new WM by aligning the WMs in the cluster so as to optimize the information content of the resulting fused WM, which is obtained by simply summing the counts across each column in the alignment.

Finally, we used MotEvo [290] to predict TFBSs for all WMs in the multiple-species alignments of all human proximal promoters. We then constructed new WMs from the list of predicted TFBSs for each WM, weighing each predicted site with its posterior probability (which incorporates position-specific prior probabilities, as described below). The number of top-scoring sites was chosen manually for each motif and was between 100 and 4000 sites, in most cases being 200 or 500 sites.

At this point we excluded one TRANSFAC motif M00395 (HOXA3, HOXB3, HOXD3) which had very low information content and predicted predicted only very low-probability sites. We additionally excluded the motifs M00480 (TOPORS) and M00987 (FOXP1), which were unrealistically specific and (in case of M00987) predicted stretches of poly(T).

For a few TFs we obtained more recent WMs from the literature (SP1, OCT4, NANOG, SOX2, XBP1, PRDM1, and the RXRG dimer) and we used these to replace the corresponding WM in the list.

We improved several motifs by running MotEvo on TF ChIP-seq data: SRF, STAT1/3, REST and ELK1/4/GABPA/GABPB1. Some other motifs were obtained by predicting *de novo* using the Phylogibbs algorithm [563] on ChIP-seq data: SPI1, CTCF, OCT4, SOX2 and NANOG.

For a few motifs JASPAR has recently updated or introduce new motifs which were based on high-throughput data and we included these motifs.This is the case for FOXA2, KLF4, EWSR1-FLI1, FEV, NR4A2. We also removed MA0118, as it had been discarded from the JASPAR data base.

Our final list contains 189 WMs. For each final WM there is an ordered list of associated human TFs, ordered by percent identity of the DBDs of TFs known to bind sites of the WM and the DBDs of the TF. We then checked this list of associations by hand and for each WM cut-off the list of associated human TFs manually. In total 340 human TFs are associated with our 189 WMs. The corresponding mouse orthologous TFs were selected using the MGI data base [564]. The entire set of WMs and mapping to associated TFs is available from the SwissRegulon website (http://www.swissregulon.unibas.ch).

B.1.3 *Transcription factor binding site predictions*

After creating reference promoter sets and curating a set of mammalian regulatory motifs we next predicted TFBSs in the proximal promoter regions of each promoter. Analysis of sequence conservation in the neighborhood of TSSs (see [293]) and experimentation with TFBS prediction in regions of different lengths around TSSs indicated that a reasonable balance between sensitivity (i.e. including relevant binding sites) and specificity (avoiding too many false positive predictions) can be obtained by predicting TFBSs in a 1 kilobase region around the TSSs of each promoter.
For each promoter, we thus extended the promoter sequence spanned by its cluster of TSSs by 500 bp upstream and 500 bp downstream. We denote this as the *proximal promoter region* of a promoter. We then extracted the sequence of the reference species, i.e. human or mouse and orthologous regions from 6 other mammals (human or mouse, rhesus macaque, cow, dog, horse, and opossum) using pairwise BLASTZ [565] alignments. For each promoter, we multiply aligned the orthologous regions using T-Coffee [372].
To obtain a phylogenetic tree for these mammalian species, with branch lengths corresponding to the expected number of substitutions per neutrally evolving site, we used methods described previously [566]. Briefly, we first obtained the topology of the tree from the UCSC Genome Browser [567]. Then, for each pair of species we made pairwise alignments of the coding regions of orthologous genes and extracted all third positions in fourfold-degenerate codons of amino acids that are conserved between the two species. Using these fourfold-degenerate positions we estimated a pairwise distance for each pair of species. Finally, we estimated the lengths of the branches in the phylogenetic tree as those that minimize the square-deviations between the implied pairwise distances and the pairwise distances estimated from the fourfold-degenerate positions. The resulting tree structure is shown in Suppl. Fig. B.1.

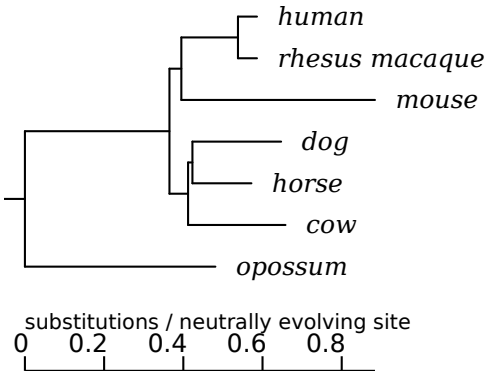


Figure B.1: The phylogenetic tree used by MotEvo for the transcription factor binding site predictions that are used by ISMARA.

The multiple sequence alignments were then used together with the phylogenetic tree and the collection of WMs as an input for TFBS predictions using

the MotEvo algorithm [290]. Given a multiple alignment, MotEvo considers all ways in which the sequence of the reference species can be segmented into 'background' positions, 'binding sites' for one of the supplied WMs, and 'unknown functional elements' (UFEs). The likelihood of alignment columns assigned to background are calculated under a model of neutral evolution along the specified phylogenetic tree. The likelihood of alignment segments assigned to be a site for a given WM are calculated by first estimating which of the species have retained a site for the WM (based on the WM scores of the individual sequences) and then applying an evolutionary model in which substitution rates are set so as to match the sequence preferences of the WM. Finally, segments assigned to be UFEs are assumed to evolve under *unknown* purifying selection constraints on the sequence, which is implemented by treating them as sites for an unknown WM. Each unknown WM column is a nuisance parameter that is integrated out of the likelihood. Finally, MotEvo assigns, at each position of the alignment and for each WM, a posterior probability that a site for the corresponding WM occurs at this position.

Since most motifs show clear positional preferences relative to TSS, we implemented, separately for each motif, a distribution of position-dependent prior probabilities of site occurrence as a function of position relative to the TSS and fitted these distributions by maximum likelihood using expectation-maximization. In addition, since high-CpG and low-CpG promoters have highly distinct configurations of TFBSs, we estimated the position-dependent prior probability distributions separately for high-CpG and low-CpG promoters.

The final result of this analysis is a matrix $\mathbf{N}$, with N_{pm} the total number of predicted sites for motif m in promoter p, i.e. the sum of the posterior probabilities of the individual sites. To reduce the probability of spurious predictions, we set $N_{pm} = 0$ whenever the sum of the posteriors of all sites combined was less than 0.1.

B.1.4 *Associating miRNA target sites with each promoter*

Apart from incorporating the effects of TFBSs in promoters, ISMARA also integrates the effects of miRNAs in its modeling of expression levels. To this end, we needed to obtain a set of predicted miRNA target sites for each promoter. We base our predictions on the miRNA target predictions of TargetScan using preferential conservation scoring (aggregate P_{CT}) [143] which has shown consistently high performance in various benchmark tests. As opposed to focusing on individual miRNAs, TargetScan groups miRNAs that have identical subsequences at positions 2 through 8 of the miRNA, i.e. the 2-7 seed region plus the 8th nucleotide, and provides predictions for each such seed motif. We will treat these seed motifs exactly like the regulatory motifs (WMs) for TFs, i.e. a miRNA seed motif can be associated with multiple miRNAs. TargetScan provides predictions for 86 mammalian miRNA seed motifs in total.

TargetScan P_{CT} provides a score for each seed motif and each RefSeq transcript. To obtain a 'site count' N_{pm} for the number of sites of miRNA seed motif m associated with promoter p we average the TargetScan P_{CT} scores of all RefSeq transcripts associated with the promoter p. Finally, the miRNA seed motif site counts N_{pm} are simply added as columns to the site count matrix $\mathbf{N}$ with site counts of TFBSs.

B.1.5 *Expression data processing*

When using expression data from oligonucleotide microarrays, the raw probe intensities are corrected for background and unspecific binding using the Bioconductor packages affy [568], oligo [279], and gcrma [277], depending on the type of the particular microarray used. The micro-arrays that are currently supported by ISMARA are listed in supplementary table B.1.

Microarray	Organism	Producer
HG-U133A	*Homo sapiens*	Affymetrix
HG-U133B	*Homo sapiens*	Affymetrix
HG-U133_Plus_2	*Homo sapiens*	Affymetrix
HG-U133A_2	*Homo sapiens*	Affymetrix
HuGene-1_0-st-v1	*Homo sapiens*	Affymetrix
HuGene-1_1-st-v1	*Homo sapiens*	Affymetrix
HuGene-2_0-st	*Homo sapiens*	Affymetrix
HuGene-2_1-st	*Homo sapiens*	Affymetrix
HT_HG-U133A	*Homo sapiens*	Affymetrix
HT_HG-U133B	*Homo sapiens*	Affymetrix
HT_HG-U133_Plus_PM	*Homo sapiens*	Affymetrix
Mouse430_2	*Mus musculus*	Affymetrix
Mouse430A_2	*Mus musculus*	Affymetrix
MOE430A	*Mus musculus*	Affymetrix
MOE430B	*Mus musculus*	Affymetrix
MoGene-1_0-st-v1	*Mus musculus*	Affymetrix
MoGene-1_1-st-v1	*Mus musculus*	Affymetrix
HT_MG-430A	*Mus musculus*	Affymetrix
HT_MG-430B	*Mus musculus*	Affymetrix
MG_U74Av2	*Mus musculus*	Affymetrix
MG_U74Bv2	*Mus musculus*	Affymetrix
MG_U74Cv2	*Mus musculus*	Affymetrix

Table B.1: Microarrays currently supported by ISMARA.

For its further analysis, ISMARA uses the logarithms of the probe intensities. For a given sample, the histogram of log-intensities is generally bimodal, with the modes corresponding to probes of non-expressed and expressed genes.

The probes are classified as expressed or non-expressed in each sample separately by fitting a two-component Gaussian mixture model to the log-intensity data using the Mclust R package [569, 570]. Probes that are consistently non-expressed are filtered out from further processing; a probe is considered not to be expressed if in *all* the samples the probability of it belonging to the expressed class is below 0.4. Subsequently, the intensity values of the remaining probes are quantile normalized across all input samples.

Microarray probes can hybridize to multiple transcripts, belonging to different genes, or different isoforms of one gene, and we decided not to rely on transcript annotations of a micro-array producer. Instead, we comprehensively mapped the probe sequences to the set of all transcripts that are associated with our reference set of promoters. Note that we thus also ignore the annotation of probes into probe sets. To calculate the expression of a promoter we average the log-expression levels of all probes that map to one (or more) of the transcripts associated with the promoter (i.e. the start of the transcript is a member of the cluster of starts that defines the promoter). The expression level of the promoter is then a weighted average of the expression levels of these probes, where a probe that maps to n different transcripts obtains a weight $1/n$. That is, in general, a probe can map to multiple transcripts.

When ISMARA uses RNA-seq for input expression data, it expects the RNA-seq data to be provided as genome alignments of the reads to the hg19 or mm9 genome assembles in BED or BAM format. The loci of the mapped reads are then intersected with the genome alignments of all transcripts that are associated with reference promoters. A read is associated with a particular transcript if its mapping falls entirely into its exons. Note that, more recent RNA-seq data in some cases involved reads that are so long that, frequently, the read overlaps two rather than a single exon of the transcript. To take this into account, recent mapping algorithms allow the start and end of the read to map to different genomic loci. The ISMARA pipe-line associates such a mapping with a given transcript when both the start and end piece map to one of its exons. In the future ISMARA may be extended to include the mapping of raw reads themselves.

To obtain an expression level for each promoter ISMARA calculates a weighted average over all reads mapping to the transcripts associated with the promoter. The weighting results from multiple mappings at two levels. Firstly, a read can map to multiple genomic loci and, secondly, a single locus may intersect multiple transcripts that are associated with multiple promoters. When a read maps to n genomic loci, we assign a weight of $1/n$ to each locus. If that locus intersects transcripts of m different promoters, then this read contributes a final weight of $1/(nm)$ to the expression of the transcript. Each transcript t is assigned a total weight w_t that consist of the sum of the weights of all reads mapping to it. Note that the expected value of w_t is both proportional to the average number of mRNAs per cell this transcript t has as well as proportional to the length l_t of the transcript. The normalized weight $\tilde{w}_t = w_t/l_t$ is proportional to the number of mRNAs per cell of transcript t. The expression of a promoter p is measured in terms of the total number of

mRNAs deriving from this promoter. Thus, for each promoter p, we calculate a total weight w_p by summing $\tilde{w}_t$ over all transcripts t that are associated with the promoter, i.e. $w_p = \sum_t \tilde{w}_t$. We obtain such a weight w_{ps} for each promoter p and each sample s. Promoters that have weights $w_{ps} = 0$ in all samples are discarded. There will be some promoters that have zero weights in some, but not all, of the samples. In order to define log-expression values for all promoters we add a small pseudo-count to the weights w_{ps}. For each sample s, we rank the promoters with nonzero weight by their weight w_{ps} and calculate the 5th percentile pc_s. We then add this weight pc_s as a pseudo-count to all weights w_{ps} of promoters, including promoters that had zero weights in sample s. Finally, we normalize the w_{ps} and log-transform them as follows:

$$E_{ps} = \log_2 \left[10^6 \frac{w_{ps}}{\sum_{p'} w_{p's}} \right]. \tag{B.1}$$

Note that the resulting expression level E_{ps} corresponds to the logarithm (base 2) of the number of mRNAs deriving from promoter p, per million mRNAs in the cell. Note that this weighting procedure for calculating promoter expression levels is robust to redundancy in the transcript sets. For example, when a promoter is associated with k highly overlapping transcripts, then a read mapping within the exons of these transcripts will get assigned to all these transcripts, with a weight $1/k$ for each. When the total weight w_{ps} of the promoter is calculated, these k are then summed back and will in the end contribute precisely 1 read.

B.1.6 *ChIP-seq data processing*

Apart from modeling expression dynamics, ISMARA can also process ChIP-seq data to automatically model chromatin state (or TF binding) changes at promoters genome-wide. Examples of such chromatin state data include histone occupancy, histone modifications, TF binding and DNAse1 hypersensitivity in promoter regions. After several experiments, we found that integrating the chromatin signal from a region of 2000 bps centered on the TSS of each promoter gives the most robust results. To obtain a chromatin state level E_{ps} of promoter p in sample s, we calculate the sum r_{ps} of the reads that map entirely within this region around promoter p and transform to the log-space after adding a pseudocount:

$$E_{ps} = \log_2 \left(r_{ps} + \frac{N_s l}{L} \right), \tag{B.2}$$

where the second term is a pseudo-count, N_s is the total number of reads mapped to the genome in sample s (the number of lines in the BED file), $l = 2000$ is the length of the regions, and L is the total length of the genome. Note that this pseudo-count is precisely the number of reads that would be expected if all N_s reads where distributed uniformly over the genome. We set to pseudo-count to this value to make the pseudo-count roughly of the same size as the

read-count from background reads in regions where the chromatin mark in question does not appear. The rational is that, in regions where there are only background reads, statistical fluctuations may cause the read-counts r_{ps} to change significantly from sample to sample. By adding a constant pseudo-count of roughly the same size, these fluctuations are effectively dampened. More formally, this pseudo-count results within a Bayesian context if we use a Dirichlet prior with an expected density l/L for each region.

B.1.7 *Motif activity fitting.*

We model the log-expression (or ChIP-seq signal) value E_{ps} of a promoter p in sample s as a linear function of the site-counts N_{pm} for all motifs m associated with the promoter, i.e. either TFBSs in the proximal promoter region or miRNA binding sites in the 3' UTRs of associated transcripts. In each sample s, the contribution of the sites N_{pm} to E_{ps} is given by the (unknown) *motif activity* A_{ms}. That is, we fit a model of the form:

$$E_{ps} = \tilde{c}_s + c_p + \sum_m N_{pm} A_{ms} + \text{noise}, \tag{B.3}$$

where $\tilde{c}_s$ and c_p are sample and promoter-dependent constants.
The last 'noise' term corresponds to the difference between the signal that the model predicts, and the signal E_{ps} that was actually measured. This difference generally results from multiple sources. First, there are measurement errors in E_{ps}. Second, there is 'biological noise, i.e. uncontrolled fluctuations in the true state of the biological system. Third, and most importantly, there is the error in the model. Regarding the distribution of measurement errors and biological noise there has been a considerable amount of work in the literature. For microarray measurements, after background correction and normalization, the sum of biological and measurement noise in log-expression levels can be reasonably approximated by a Gaussian. For next-generation sequencing data such as RNA-seq and ChIP-seq data, which is intrinsically digital in nature, we have previously studied the distribution of biological and measurement noise using data from replicate experiments [293]. This analysis showed that on a normal (i.e. non-logarithmic) scale, the noise distribution can be well approximated by Poisson sampling with a rate that is itself log-normally distribution (with some variance σ^2). As we showed in [293], in log-scale this distribution can be well-approximated by a normal distribution with a variance that is the sum of the variance σ^2 of the original log-normal, and a term $1/n$, where n is the total read-count, which results from the Poisson sampling noise. An alternative model of biological and replicate noise that has been used in the literature is the negative binomial distribution [373, 571]. A negative binomial is obtained when there is Poisson sampling noise with a rate that is itself Gamma distributed. Like the distribution derived in [293], this distribution also has the property that, in log-scale, the contribution to the variance due to Poisson sampling decreases with absolute expression level.
However, as mentioned above, besides uncontrolled fluctuations in the state of the biological system and measurement noise, the 'noise' term in equation

(B.3) also contains a contribution from the *error* of the model. That is, even if experimentalists could perfectly control the state of the biological system (i.e. no biological noise) and make measurements without any errors (i.e. no measurement noise) then, because of the simplicity of our model, there would still be a large difference between the predicted signal levels of each promoter, and the true signal levels. Indeed, our model typically only captures a modest fraction of the variance in expression and ChIP-seq levels, meaning that the error in the model is generally much larger than the biological and measurement noise. That is, the noise term in equation (B.3) is *dominated* by the error in the model. Consequently, the relevant noise distribution is not the distribution of biological and measurement noise, but the distribution of model errors. Since we have no specific information regarding the form of the distribution of modeling errors we will make the assumption that the noise is Gaussian distributed with an unknown variance σ^2 that is the same for all promoters and in all samples.

Under these assumptions we find the following expression for the likelihood of the expression data given the site-counts, motif activities and sample and promoter-dependent constants:

$$P\left(E \mid A, c, \tilde{c}, N, \sigma\right) \propto \prod_{p,s} \frac{1}{\sigma} \exp\left[-\frac{\left(E_{ps} - \tilde{c}_s - c_p - \sum_m N_{pm} A_{ms}\right)^2}{2\sigma^2}\right] \tag{B.4}$$

We first maximize this expression with respect to all the constants c_p and $\tilde{c}_s$, and substitute these with their *maximum likelihood* estimates. After doing this we obtain:

$$P\left(E \,\middle|\, A', N, \sigma\right) \propto \sigma^{-PS} \exp\left[-\frac{\sum_{ps}\left(E'_{ps} - \sum_m N'_{pm} A'_{ms}\right)^2}{2\sigma^2}\right], \tag{B.5}$$

where P is the number of promoters, S is the number of samples, the N'_{pm} are a motif-normalized site-counts $N'_{pm} = N_{pm} - \langle N_m \rangle$, with $\langle N_m \rangle$ the average site-count per promoter for motif m, the A'_{ms} are sample-normalized activities $A'_{ms} = A_{ms} - \langle A_m \rangle$, i.e. with $\langle A_m \rangle$ the average activity of motif m across the samples, and the E'_{ps} are sample- and promoter-normalized expression values $E'_{ps} = E_{ps} - \langle E_p \rangle - \langle E_s \rangle + \langle\langle E \rangle\rangle$. That is the log-expression matrix E'_{ps} is normalized such that all its rows and columns sum to zero, the activities A'_{ms} are normalized such that the average activity over all samples is zero, i.e. $\sum_s A'_{ms} = 0$, and the site-counts N'_{pm} are normalized such that the average count over all promoters is zero, i.e. $\sum_p N'_{pm} = 0$.

To avoid over-fitting we assign a symmetric Gaussian prior to each motif activity, i.e. the joint prior for all activities is given by:

$$P\left(A' \,\middle|\, \lambda, \sigma\right) \propto \prod_{ps} \exp\left[-\frac{\lambda^2}{2\sigma^2} \sum_m A'^2_{ms}\right], \tag{B.6}$$

where the constant λ^2 sets the width of prior distribution relative to the width of the likelihood function. Using this prior with the likelihood derived above, the posterior distribution of motif activities takes the form:

$$P\left(A^{'} \;\middle|\; E,N,\sigma,\tau\right) \propto \sigma^{-PS} \exp\left[-\frac{\sum_{p,s}\left(\left(E^{'}_{ps}-\sum_m N^{'}_{pm}A^{'}_{ms}\right)^2+\lambda^2\sum_m A^{'2}_{ms}\right)}{2\sigma^2}\right]. \tag{B.7}$$

Since equation (B.7) factorizes into independent expressions for the different samples, it is enough to consider one sample at a time. The posterior distribution for the motif activities in a particular sample takes the general form of a multi-variate Gaussian centered around $A^{'*}_{ms}$:

$$P\left(A^{'}_{s} \;\middle|\; E,N,\sigma\right) \propto \sigma^{-P} \exp\left[-\frac{\sum_{m\tilde{m}}\left(A^{'}_{ms}-A^{'*}_{ms}\right)W_{m\tilde{m}}\left(A^{'}_{\tilde{m}s}-A^{'*}_{\tilde{m}s}\right)+\chi^2_s}{2\sigma^2}\right], \tag{B.8}$$

where the χ^2_s is the unexplained part of variance in sample s

$$\chi^2_s = \sum_p \left(E^{'}_{ps} - \sum_m N^{'}_{pm}A^{'*}_{ms}\right)^2, \tag{B.9}$$

and the matrix W is given by

$$W_{m\tilde{m}} = \sum_p \left(N^{'}_{pm}N^{'}_{p\tilde{m}} + \lambda^2\delta_{m\tilde{m}}\right). \tag{B.10}$$

Finally, the *maximum a posteriori* (MAP) estimates $A^{'*}_{ms}$ can be found by minimizing the expression in the numerator of equation (B.7) using standard numerical procedures for ridge regression. ISMARA performs this calculation by singular value decomposition of the $N^{'}$ matrix.

B.1.7.1 *Setting λ through cross-validation*

Both the MAP estimates $A^{'*}_{ms}$, and the matrix $W_{m\tilde{m}}$ are functions of λ. The constant λ^2 represents the ratio between the *a priori* expected variance of activities, to the average squared-deviation of the model from the expression data (which results from both error in the model, noise in the expression measurements, and biological noise). In general λ will depend on the measurement platform used, i.e. microarray, RNA-seq, or ChIP-seq, and also on the samples used, because the true variance in motif activities will depend on the variance in the E_{ps} across the samples. Thus, the appropriate value of λ will generally not be known in advance and ISMARA therefore includes a method for automatically setting λ from the data. To determine the optimal λ ISMARA uses a 80/20 cross-validation scheme. The set of promoters is divided randomly into two sets, with one containing 80% of all promoters (the 'training set') and the other the remaining 20% (the 'test set'). The training set of promoters is used for fitting the motif activities while the quality

of the fit is evaluated on the test set. ISMARA then finds the value of λ that minimizes the average squared-deviation of the expression levels in the test set from those predicted by the model. We denote this optimal value of λ by λ^*.

B.1.7.2 *Error bars on motif activities*

Apart from the MAP estimates A'^*_{ms} ISMARA also determines the uncertainties associated with these estimates. Since σ in Eq. B.8 is not known, we integrate it out with a suitable scale-invariant prior $P(\sigma) \propto \frac{1}{\sigma}$.

$$\begin{aligned} P\left(A'_s \,\middle|\, E,N,\lambda\right) &= \int_{\sigma=0}^{\infty} P\left(A'_s \,\middle|\, E,N,\sigma,\lambda\right) P(\sigma)\mathbf{d}\sigma \\ &\propto \frac{\Gamma\left(\frac{P}{2}\right)}{\left[\sum_{m\tilde{m}} \left(A'_{ms}-A'^*_{ms}\right) W_{m\tilde{m}} \left(A'_{\tilde{m}s}-A'^*_{\tilde{m}s}\right)+\chi^2_s\right]^{\frac{P}{2}}} \\ &\propto \exp\left[-\frac{P\sum_{m\tilde{m}} \left(A'_{ms}-A'^*_{ms}\right) W_{m\tilde{m}} \left(A'_{\tilde{m}s}-A'^*_{\tilde{m}s}\right)}{2\chi^2_s}\right], \end{aligned} \tag{B.11}$$

where the last proportionality is a very good approximation when the number of promoters is large. Note that this is again a multi-variate Gaussian distribution. The covariance matrix of this Gaussian posterior distribution is given by:

$$C_{m\tilde{m};s} = \frac{\left(W^{-1}\right)_{m\tilde{m}} \chi^2_s}{P} \tag{B.12}$$

As is well known, given this multi-variation Gaussian form, the marginal distribution for a single motif activity A'_{ms} will be Gaussian distributed with standard-deviations $\delta A'_{ms}$ given by the square root of the corresponding diagonal term of the covariance matrix, i.e.

$$\delta A'_{ms} = \sqrt{C_{mm;s}} \tag{B.13}$$

We define the overall *significance* of a motif m as the average squared ratio between fitted activities and their standard deviations (z-values)

$$z_m = \sqrt{\frac{1}{S}\sum_s \left(\frac{A'^*_{ms}}{\delta A'_{ms}}\right)^2}. \tag{B.14}$$

B.1.7.3 *Fitting mean activities*

By introducing a promoter-dependent basal expression level c_p in equation (B.3) we effectively ensure that the average expression of each promoter is accounted for, i.e. only the *changes* in expression of each promoter across the samples are fitted by the motif activities A'_{ms}. Consequently, the fitted motif activities all average to zero, i.e. $\sum_s A'_{ms} = 0$. Although, typically, users would indeed be most interested in explaining expression *changes* across the samples, in some cases users might also be interested in knowing to what extent the absolute *average* levels of the promoters across the samples can

be fit in terms of 'mean activities' of the motifs, i.e. to learn which motifs are most predictive of consistently high or low absolute expression across the replicates.
To fit mean activities we start from equation (B.3) and set $c_p = 0$ for all promoters p. In addition, we explicitly write the activity in terms of a sample-dependent part that averages to zero, and a mean activity, i.e.

$$A_{ms} = A'_{ms} + \bar{A}_m. \tag{B.15}$$

Defining the sample-corrected average expression values as

$$\tilde{E}_p = \frac{1}{S} \sum_s \left(E_{ps} - \langle E_s \rangle \right), \tag{B.16}$$

and again the motif-normalized site counts $N'_{pm} = N_{pm} - \langle N_m \rangle$, it is straightforward to show that the mean activities $\bar{A}_m$ are optimized when the expression

$$\tilde{\chi}^2 = \sum_p \left(\tilde{E}_p - \sum_m N'_{pm} \bar{A}_m \right)^2, \tag{B.17}$$

is minimized. We fit the mean activities $\bar{A}_m$ in exact analogy with the fitting of the activities A'_{ms}. We introduce a separate Gaussian prior for the mean activities $\bar{A}_m$, with its own parameter $\tilde{\lambda}$, and again set $\tilde{\lambda}$ using 80/20 cross-validation. We also determine error-bars $\delta\bar{A}_m$ on the mean activities $\bar{A}_m$. Finally, we also define z-scores for the mean activities, i.e.

$$\tilde{z}_m = \frac{\bar{A}_m}{\delta\bar{A}_m}. \tag{B.18}$$

Motifs with the highest positive $\tilde{z}_m$ are the most significant predictors of consistently high expression across the samples, whereas motifs with highly negative $\tilde{z}_m$ are the most significant predictors of consistently low expression across the samples.

B.1.8 *Processing of replicates*

Careful studies typically involve experimental replicates to account for the part of variability in the readout which is not under direct experimental control. ISMARA allows users to indicate which samples correspond to replicates and will automatically calculate averaged motif activities and error bars across these replicates. To perform this analysis the user should first upload all samples and perform the standard ISMARA analysis. On the results page ISMARA provides a link to a page where users can interactively annotate which samples are replicates. In addition, if the replicates came in clearly defined batches, for example, when a time-course was performed multiple times, then the user can also indicate this. Once all samples are annotated ISMARA can then perform motif activity averaging across the replicates. Note that this

approach can easily be extended beyond replicates, i.e. the user can arbitrarily divide the samples into groups and ISMARA will automatically calculate average motif activities and associated standard-deviations for each group of samples.

Here we describe how activities within a group are averaged. For a given group G of samples and a particular motif, we assume that its activities A_s in samples $s \in G$ are given by a mean activity $\bar{A}^g$ plus some deviation δ_s, i.e

$$A_s = \bar{A}^g + \delta_s, \tag{B.19}$$

where we assume that the prior probability of δ_s is Gaussian distributed with (unknown) standard-deviation σ_g, i.e

$$P(\delta_s|\sigma_g) = \frac{1}{\sqrt{2\pi}\sigma_g} \exp\left[-\frac{1}{2}\frac{\delta_s^2}{\sigma_g^2}\right]. \tag{B.20}$$

Thus, given the mean activity $\bar{A}^g$ in the group, the prior probability to have activity A_s in a particular sample s from the group is

$$P(A_s|\bar{A}^g,\sigma_g) = \frac{1}{\sqrt{2\pi}\sigma_g} \exp\left[-\frac{1}{2}\frac{(A_s-\bar{A}^g)^2}{\sigma_g^2}\right]. \tag{B.21}$$

Using the input data, ISMARA has inferred the motif activity A_s to have expected value A_s^* with standard-error δA_s for each sample s. That is, once the dependence on all other activities is integrated out, the probability of the expression data D conditioned on the motif activity A_s is a Gaussian with standard-deviation δA_s, i.e.

$$P(D|A_s) = \frac{1}{\sqrt{2\pi}\delta A_s} \exp\left[-\frac{1}{2}\frac{(A_s-A_s^*)^2}{(\delta A_s)^2}\right]. \tag{B.22}$$

Using the expressions for $P(D|A_s)$ and $P(A_s|\bar{A}^g,\sigma_g)$ we can calculate the probability of the data D given the mean activity $\bar{A}^g$ and standard-deviation σ_t by integrating over all unknown A_s:

$$P(D|\bar{A}^g,\sigma_g) = \prod_{s\in G}\left[\int_{-\infty}^{\infty} P(D|A_s)P(A_s|\bar{A}^g,\sigma_g)\mathbf{d}A_s\right]. \tag{B.23}$$

These integrals can be performed analytically and we obtain

$$P(D|\bar{A}^g,\sigma_g) = \prod_{s\in G}\frac{1}{\sqrt{2\pi(\sigma_g^2+\sigma_s^2)}} \exp\left[-\frac{(A_s^*-\bar{A}^g)^2}{2(\sigma_g^2+\sigma_s^2)}\right]. \tag{B.24}$$

Although, formally, we should integrate this expression over the unknown standard-deviation σ_g as well, this integral unfortunately cannot be performed analytically. Therefore, we estimate the integral simply by finding the value σ_g^* that maximizes $P(D|\bar{A}^g,\sigma_g)$. Assuming a uniform prior for the mean activity $\bar{A}^g$ of the samples in the group, we then finally obtain an expression for the posterior probability $P(\bar{A}^g|D)$ which we characterize by its mean $\langle\bar{A}^g\rangle$

and standard-deviation $\delta\bar{A}^g$. That is, $\langle\bar{A}^g\rangle$ is the inferred average motif activity for the samples within the group, and $\delta\bar{A}^g$ is the error-bar on this average activity. This mean and error-bar of the activity for the 'group' of samples are given by

$$\langle\bar{A}^g\rangle = \frac{\sum_{s\in G}\frac{A^*_s}{(\sigma^*_g)^2+\sigma^2_s}}{\sum_{s\in G}\frac{1}{(\sigma^*_g)^2+\sigma^2_s}}, \tag{B.25}$$

and

$$\delta\bar{A}^g = \sqrt{\frac{1}{\sum_{s\in G}\frac{1}{(\sigma^*_g)^2+\sigma^2_s}}}. \tag{B.26}$$

Finally, we assign significances z_m to each motif completely analogously as before, but now averaging over all groups, i.e.

$$z_m = \sqrt{\frac{1}{|G|}\sum_g\left(\frac{\langle\bar{A}^g\rangle}{\delta\bar{A}^g}\right)^2}, \tag{B.27}$$

where $|G|$ is the number of groups. A motif will have a high significance z_m when its motif activities vary relatively little within each group, and vary by a large amount across groups.

B.1.9 *Target predictions*

In order to infer motif activities A_{ms}, ISMARA assumes that all promoters with predicted target sites for a motif m will respond to changes in motif activity, i.e. in proportion to the predicted number of sites N_{pm}. This is a reasonable assumption when inferring motif activities, as the activities A_{ms} depend on the statistics of all promoters with sites for motif m. However, in a given condition or system, it is likely that only a subset of the promoters with sites for a motif m are in fact regulated by this regulator. This might be due to a limited accessibility, dependence on particular co-factors, weaker affinity of a site, and other context-dependent factors. Thus, when we aim to predict individual target promoters of a given motif m, we not only use the binding site predictions N_{pm}, but also evaluate at which promoters the activities A_{ms} contribute to explaining the profiles E_{ps}.

To quantify if a given promoter p is targeted by a motif of interest m we first demand that there exists a TFBS prediction, i.e $N_{pm} > 0$. Second, we quantify the contribution of m to the fit of the expression/chromatin state profile E_{ps}. The most rigorous approach to quantifying the effect of motif m on promoter p is to calculate both the probability of the entire data set, i.e. the profiles E_{ps} across all promoters and samples, with the original site-count matrix $\mathbf{N}$, and a site-count matrix $\tilde{\mathbf{N}}$ where only the sites for motif m in promoter p are set to zero. To calculate this probability we treat all the unknown motif activities A_{ms} as well as the standard-deviation σ as nuisance

parameters that are integrated out of the likelihood. That is, we formally want to calculate the ratio of probabilities

$$R_{pm} = \frac{\int_{-\infty}^{\infty} \mathbf{d}A \int_{0}^{\infty} \mathbf{d}\sigma P(E|N, A, \sigma)}{\int_{-\infty}^{\infty} \mathbf{d}A \int_{0}^{\infty} \mathbf{d}\sigma P(E|\tilde{N}, A, \sigma)}, \tag{B.28}$$

where the integrals are over all motif activities A_{ms}, and over the standard-deviations σ. Note that, when we set $N_{pm} = 0$ for promoter p and motif m, we make a very small change to the site-count matrix. That is, as there are tens of thousands of promoters and close to 200 motifs, we are changing only one of the millions of entries in the matrix. As a consequence, the inferred motif activities A'^{*}_{ms} that result from the mutated matrix $\tilde{N}$ are generally very close to those that result from the original matrix N. Similarly, the inverse covariance matrix W of the mutated matrix is also very close to that of the original matrix and, finally, the optimal values of the constants c_p, $\tilde{c}_s$, and the prior constant λ^* will also change very little under mutation of the matrix. To make the calculation more tractable we will make the approximation that all these quantities are *unchanged* upon mutation of the matrix. Under that approximation we have

$$P(E|A, N, \sigma, \lambda^*) \propto \sigma^{-PS} \exp\left[-\frac{\sum_{s,m,\tilde{m}} \left(A'_{ms} - A'^{*}_{ms}\right) W_{m\tilde{m}} \left(A'_{\tilde{m}s} - A'^{*}_{\tilde{m}s}\right) + \sum_{p,s} \chi^2_{ps}}{2\sigma^2}\right], \tag{B.29}$$

where χ^2_{ps} is the squared-deviation between the observed value E'_{ps} and the predicted value, i.e.

$$\chi^2_{ps} = \left(E'_{ps} - \sum_m N'_{pm} A'^{*}_{ms}\right)^2. \tag{B.30}$$

For the probability of the data with the mutated site-count matrix we have

$$P(E|A, \tilde{N}, \sigma, \lambda^*) = P(E|A, N, \sigma, \lambda^*) \exp\left[-\frac{\sum_s (\chi^2_{psm} - \chi^2_{ps})}{2\sigma^2}\right], \tag{B.31}$$

where χ^2_{psm} is the squared-deviation for promoter p and sample s when motif m is removed, i.e.

$$\chi^2_{psm} = \left(E'_{ps} - \sum_{m'} \tilde{N}'_{pm'} A'^{*}_{m's}\right)^2. \tag{B.32}$$

In this form the integrals over the motif activities and σ can be easily performed and we find for the ratio of the probabilities

$$R_{pm} = \left(\frac{\sum_{p',s} \chi^2_{p's}}{\sum_{p',s} \chi^2_{p's} - \sum_s (\chi^2_{psm} - \chi^2_{ps})}\right)^{S(P-M)}, \tag{B.33}$$

where M is the total number of motifs. Since $P \gg M$ we approximate $P - M \approx P$ and we find approximately

$$R_{pm} = \exp\left[\frac{\sum_s(\chi^2_{psm} - \chi^2_{ps})}{\langle\chi^2\rangle}\right], \tag{B.34}$$

where we have defined the average squared-deviation per sample/promoter combination

$$\langle\chi^2\rangle = \frac{1}{PS}\sum_{p,s}\chi^2_{ps}, \tag{B.35}$$

and made use of the fact that $[1 - x/(SP)]^{-SP} \approx e^x$ for large SP.
In the results shown in the web-server we show, for each predicted target, the logarithm of the likelihood ratio, i.e. the score S_{pm} for motif m targeting promoter p is

$$S_{pm} = \frac{\sum_s \chi^2_{psm} - \chi^2_{ps}}{\langle\chi^2\rangle}. \tag{B.36}$$

Note that this result has a straightforward interpretation: The difference $\chi^2_{psm} - \chi^2_{ps}$ is the amount by which the square deviation between the predicted and observed signal increases when the sites for motif m are removed from promoter p, and the ratio $(\chi^2_{psm} - \chi^2_{ps})/\langle\chi^2\rangle$ is the relative increase in square-deviation, i.e. relative to the average squared-deviation between the predicted and observed signals. The score S_{pm} is obtained by summing this relative change in χ^2 over all samples. By default ISMARA reports all target promoters for which this score is positive, i.e. where removing the motif from the promoter reduces the quality of the fit.

B.1.9.1 *Enriched Gene Ontology categories*

To analyze whether there are any Gene Ontology categories whose genes are over-represented among the targets of a motif, we use the "GO::TermFinder" Perl module [572]. The ontology files and associations between genes and categories were taken from the Gene Ontology (GO) Consortium web-site [313]. As a set of target genes for motif m we include all genes associated with promoters that have a target score $S_{pm} > 0$. For microarray chips we create a background set from all the genes which have probes present on the microarray, i.e. according to our mappings of the probes (see Expression data processing). For RNA-seq data we take as a background set all genes associated with promoters which have mapped reads. In the web results we display all GO categories with a p-value of 0.05 or less. These p-values are corrected for multiple testing using a simple Bonferroni correction, i.e. multiplied by the number of tests performed.

B.1.10 *Principal component analysis of the activities explaining chromatin mark levels*

We first performed standard ISMARA analysis on the $n = 10$ data sets measuring expression and 9 different chromatin marks (ChIP-seq), across $S = 8$

cell types [357]. For each motif m, and each mark i, we thus obtained estimated activities A^i_{ms}.

We performed principal component analysis (PCA) of the expression and chromatin mark levels across all promoters, separately for each cell type. For a given sample s, let E_{pi} denote the level of mark i at promoter p (suppressing the label s for notational simplicity). We have here already column normalized these levels, i.e.

$$\sum_p E_{pi} = 0, \tag{B.37}$$

for all marks i.

Using singular value decomposition, the matrix $E = U \cdot D \cdot V^T$ can be uniquely decomposed into an orthonormal matrix U (of size $P \times n$), a diagonal positive-semidefinite matrix D (of size $n \times n$), and an orthonormal matrix V (of size $n \times n$) as:

$$E_{pi} = \sum_{k=1}^{n} U_{pk} D_{kk} V_{ik}, \tag{B.38}$$

where k denotes the index of each component, the column vectors $\vec{V}_k$ with components V_{ik} contain the principal components, and D^2_{kk} is the fraction of the variance in the E_{pi} values, i.e.

$$\mathrm{var}(E) = \frac{1}{nP} \sum_{p,i} \left(E_{pi}\right)^2, \tag{B.39}$$

that is explained by component k.

The first principal component $\vec{V}_1$, shown in the top panels of Suppl. Fig. B.23, is virtually identical in all cell types and captures approximately 60% of the collective behavior of the expression and 9 chromatin marks (8 histone modification and CTCF binding) across promoters in each sample. As discussed in the main text, this first principal component appears to capture the combination of chromatin mark levels associated with the general 'activity' of a promoter. As a consequence, the effect of a given TF on a specific chromatin mark is confounded by its effect on general promoter activity and we therefore decided to subtract it from the activity profiles of all TFs.

For the purpose of removing the first principal component from the motif activities, we will treat each motif m separately and ignore the covariances in the inferred motif activities, i.e. as we assumed previously when calculating the error bars on the motif activities in (B.13). We perform the removal one sample (cell line) at a time. A careful probabilistic analysis must be performed in order to calculate the error bars.

Let's focus on a given motif m in sample s and denote by A the vector of activities across the marks, i.e. A_i is the activity associated with mark i. In addition, let δA_i denote the standard-deviation (error-bar) of this activity. The posterior distribution $P(A|D)$ of this activity vector given the data is given by a Gaussian, i.e. as in (B.12), of the form

$$P(A|D) \propto \exp\left[-\frac{1}{2}\sum_i \frac{(A_i - A_i^*)^2}{\delta A_i^2}\right], \tag{B.40}$$

where A_i^* is the MAP estimate of the motif activity of mark i. If we introduce a diagonal matrix containing the inverse of the standard-deviation, we can write this expression in matrix-vector form:

$$P(A|D) \propto \exp\left[-\frac{1}{2}(A-A^*)^T \cdot \mathrm{diag}\left(\frac{1}{\delta A^2}\right) \cdot (A-A^*)\right], \quad \text{(B.41)}$$

where A^* is a $n \times 1$ vector of the MAP estimates and $\mathrm{diag}\left(\frac{1}{\delta A^2}\right)$ is a 10×10 diagonal precision matrix which elements are set to the inverses of motif activity variances.
Using principal components V of E (B.38) and their orthonormality $V \cdot V^T = \mathbb{1}$ this distribution can be rewritten as

$$P(A|D) \propto \exp\left[-\tfrac{1}{2}(A-A^*)^T \cdot V \cdot V^T \cdot \mathrm{diag}\left(\tfrac{1}{\delta A^2}\right) \cdot V \cdot V^T \cdot (A-A^*)\right]. \quad \text{(B.42)}$$

We can rewrite the activities in the basis of the principal vectors as $B \equiv V^T \cdot (A-A^*)$ and the precision matrix in the same basis as $M \equiv V^T \cdot \mathrm{diag}\left(\frac{1}{\delta A^2}\right) \cdot V$. In this basis the probability distribution takes the form:

$$P(B|D) \propto \exp\left[-\frac{1}{2}B^T \cdot M \cdot B\right]. \quad \text{(B.43)}$$

Note that in this basis, the inverse covariance matrix M contains off-diagonal terms.
We want to integrate out the activities along the first principal component, therefore we separate elements of B and M in the following way

$$B = \begin{pmatrix} b_1 \\ \begin{pmatrix} b_2 \\ \vdots \\ b_n \end{pmatrix} \end{pmatrix} \equiv \begin{pmatrix} b_1 \\ B_y \end{pmatrix} \quad \text{(B.44)}$$

$$M = \begin{pmatrix} m_{11} & \begin{pmatrix} m_{12} & \cdots & m_{1n} \end{pmatrix} \\ \begin{pmatrix} m_{21} \\ \vdots \\ m_{n1} \end{pmatrix} & \begin{pmatrix} m_{22} & \cdots & m_{2n} \\ \vdots & \ddots & \vdots \\ m_{n2} & \cdots & m_{nn} \end{pmatrix} \end{pmatrix} \equiv \begin{pmatrix} m_{11} & M_y^T \\ M_y & M_w \end{pmatrix}, \quad \text{(B.45)}$$

and the last equivalency holds because the matrix M is symmetric.
Using these definitions, eq. (B.43) can be expanded and rewritten to obtain:

$$\begin{aligned} P(B|D) &\propto \exp\left[-\frac{1}{2}\left(b_1^2 m_{11} + 2b_1 B_y^T \cdot M_y + B_y^T \cdot M_w \cdot B_y\right)\right] \quad \text{(B.46)} \\ &= \exp\left[-\frac{1}{2}\left(m_{11}\left(b_1 + \frac{B_y^T \cdot M_y}{m_{11}}\right)^2 + B_y^T \cdot M_w \cdot B_y - \frac{B_y^T \cdot M_y \cdot M_y^T \cdot B_y}{m_{11}}\right)\right] \end{aligned}$$

Where we reordered terms and completed the square to bring out that this posterior is proportional to a Gaussian with respect to b_1. It is now straightforward to integrate this probability distribution along the first principal direction:

$$\begin{aligned} P(B_y|D) = \int_{b_1=-\infty}^{\infty} P(B|D)\mathbf{d}b_1 \quad &\propto \quad \exp\left[-\frac{1}{2}\left(B_y^T \cdot M_w \cdot B_y - \frac{B_y^T \cdot M_y \cdot M_y^T \cdot B_y}{m_{11}}\right)\right] \cdot \\ &\cdot \int_{b_1=-\infty}^{\infty} \exp\left[-\frac{1}{2} m_{11}\left(b_1 + \frac{B_y^T \cdot M_y}{m_{11}}\right)^2\right] \mathbf{d}b_1 \quad \text{(B.47)} \\ &\propto \quad \exp\left[-\frac{1}{2} B_y^T \cdot \left(M_w - \frac{M_y \cdot M_y^T}{m_{11}}\right) \cdot B_y\right], \end{aligned}$$

The last proportionality holds because the Gaussian integral yields a constant (with respect to B_y). Since the covariance matrix is the inverse of the precision matrix, the covariance matrix W in the reduced $(n-1)$-dimensional space (i.e. without the first principal direction) has the form:

$$W = \left(M_w - \frac{M_y \cdot M_y^T}{m_{11}}\right)^{-1} \tag{B.48}$$

Finally, this covariance matrix W needs to be transformed back from the principal component basis to the original basis. To this end we use the principal components contained in columns 2 through n of the V matrix. We obtain for the final covariance matrix K in the original basis

$$K_{ij} = \sum_{k,l=2}^{n} V_{ik} W_{kl} V_{jl}. \tag{B.49}$$

The standard deviation of activities of the i^{th} mark is given by square root of the corresponding diagonal element of this matrix

$$\delta\tilde{A}_i = \sqrt{K_{ii}}. \tag{B.50}$$

The corrected MAP activities are obtained by first defining

$$B^* = V^T \cdot A^*, \tag{B.51}$$

and then transforming back to the original basis using only the components along principal vectors 2 through n:

$$\tilde{A} = \sum_{k=2}^{n} V_{ik} B_k^*. \tag{B.52}$$

The reported z-value of the i^{th} mark (we introduce back the indices for motif m and sample s omitted previously) is given by

$$z_{ms}^i = \frac{\tilde{A}_i}{\delta\tilde{A}_i} \tag{B.53}$$

After removing the contribution of the first principal component to the motif activities, we re-calculated significance z-values z_m^i for each motif m and each mark i (x-axis in the Suppl. Fig. B.24)

$$z_m^i = \sqrt{\frac{\sum_{s'} \left(z_{ms'}^i\right)^2}{S}}. \tag{B.54}$$

In addition, we calculated a specificity s_m^i which measures the fraction of the overall significance that is associated with mark i (y-axis in the Suppl. Fig. B.24)

$$s_m^i = \frac{z_{mk}^2}{\sum_{k'} z_{mk'}^2}. \tag{B.55}$$

That is, a motif m will be highly specific for mark i if it has a high z-value z_m^i, and low z-values for all other marks.

B.2 FRACTION OF VARIANCE EXPLAINED BY THE FIT

The total variance V in a data set is given by the sum of the squared normalized expression values

$$V = \frac{1}{PS} \sum_{p,s} (E'_{ps})^2. \tag{B.56}$$

After fitting the model, the average squared deviation left unexplained is given by the average of χ_{ps}^2 across all promoters and samples, i.e. as defined by equations (B.30) and (B.35). The fraction of the variance f explained by the fit is thus

$$f = 1 - \frac{\langle \chi^2 \rangle}{V}. \tag{B.57}$$

For the data-sets that we analyze in this study, the fraction of explained variance ranges from slightly less than 2% to almost 17%, with a median of 7.7%. Suppl. Fig. B.2 shows a histogram of the fraction of variance explained across all samples.

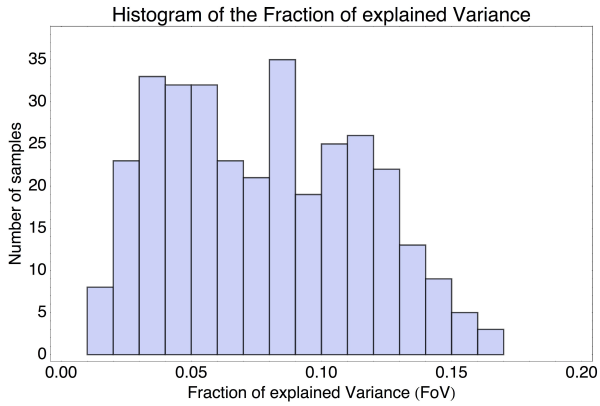


Figure B.2: Histogram of the fraction of variance explained for all the gene expression samples analyzed in this study (datasets 1 through 5 and dataset 6.1).

For the first data-set, the Illumina Body Map 2, we find that 7.71% of the variance is explained by the model. To assess the statistical significance of this fraction, we performed 10′000 randomization experiments in which we randomized the association between promoter expression profiles E_{ps} and site-counts N_{pm}, i.e. we randomly shuffled the rows of the matrix $\mathbf{N}$ while leaving

the matrix **E** unchanged. For each of the 10′000 randomizations, we then fitted the model, including fitting the parameter of the Gaussian prior through cross-validation so as to maximize the fraction of explained variance on the test-set. The left panel of Suppl. Fig. B.3 shows the distribution of fraction of explained variance f for the 10′000 randomizations. As the figure shows, there is a roughly exponential distribution of f and the highest observed f was $f = 0.0012$. If we extent the exponential fit to the distribution of f values in randomization experiments (right panel of Suppl. Fig. B.3) we see that the observed fraction of variance $f = 0.0771$ on the unshuffled promoters corresponds roughly to a p-value of $1.3 * 10^{-235}$.

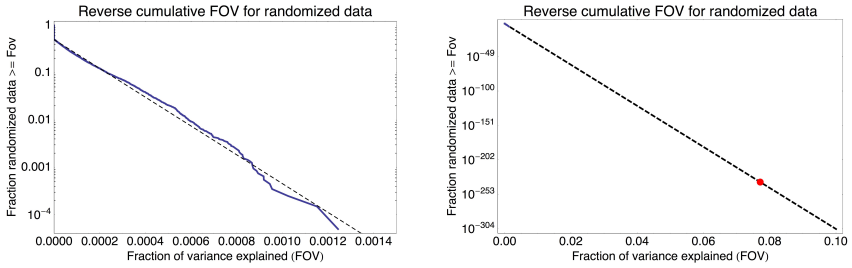


Figure B.3: Left panel: Reverse-cumulative distribution of the fraction of explained variance f on the Illumina Body Map 2 data-set on 10′000 randomizations between promoter expression and site-counts (solid line). The dashed line shows an exponential fit. **Right panel:** The exponential fit of the left panel extended to the observed fraction of explained variance ($f = 0.0771$, red dot) of the original, i.e. non-randomized, data-set. The estimated p-value of the observed fraction of variance is approximately $1.3 * 10^{-235}$.

B.3 OVERVIEW OF RESULTS PRESENTED IN THE WEB-INTERFACE

To illustrate the results that ISMARA provides, we here present a number of figure that show examples of results on the RNA-seq data of the Illumina Body Map 2 [573]. Note that almost all of these figures are screen shots from the actual web-interface. All the full results for the Illumina Body Map are available at http://ismara.unibas.ch/supp/dataset1_IBM/ismara_report/.

The main page of results that ISMARA provides for a given data set centers around a list of motifs, sorted by their significance, showing for each motif its significance, the associated TFs, a sequence logo of the motif, and a thumbnail image of its inferred activity across the samples. Supplementary Fig. B.4 shows an excerpt from this list of motifs.

Each motif name in this list is in fact a link to a separate page with much more extensive results for the motif. Among these more extensive results is, first of all, a figure showing the inferred motif activity (and error bars) across all samples, where the samples are ordered according from left to right, according to the user's input.

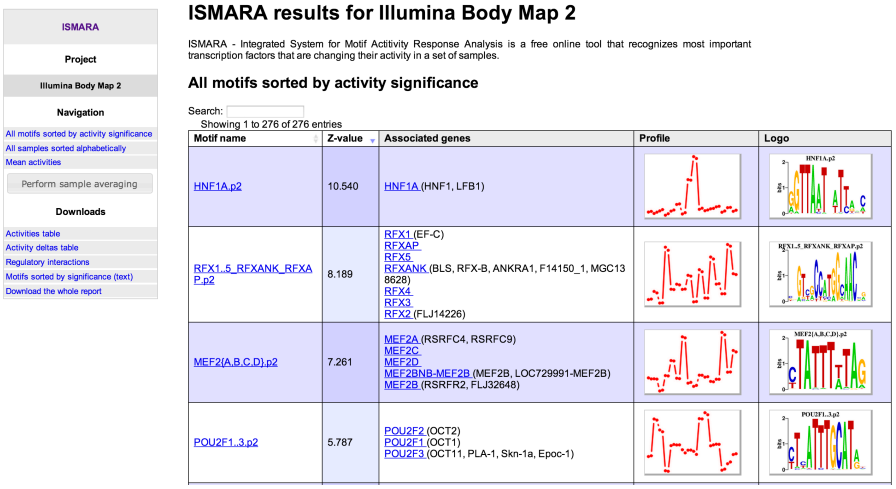


Figure B.4: Fragment of the list of regulatory motifs sorted by their significance (z-score). The motifs are sorted from top to bottom. Shown for each motif are, from left to right, the name of the motif (which is a link to a separate page with results for the motif), its z-score, a list of associated TFs (links to NCBI pages for these genes), a thumbnail of the inferred motif activity profile, and the sequence logo of the motif.

Supplementary Fig. B.5 shows the activity profile of the HNF1A motif across the Illumina Body Map 2 samples. Note that such a lexicographic ordering of motif activities across samples is especially helpful when the samples come from a time course, in which case the graph shows the motif activity across time.

However, in many cases, including the Illumina Body Map analyzed here, there is no preferred natural ordering of the samples. In those cases it is more natural to present the motif activities with samples sorted from those in which the motif is most significantly upregulated, to those where it is most significantly downregulated. ISMARA provides such a list of motif z-values, with samples sorted from largest to smallest z-value, as shown in Suppl. Fig. B.6 for the HNF1A motif. In this case, HNF1A activity is highly specific to liver and kidney.

For many of the motifs incorporated into the ISMARA analysis, there is more than one TF that can potentially bind to sites for the motif. As a consequence, it is not always clear which individual TFs are responsible for the observed motif activity in a particular system. To help determine which TFs are most likely involved in the activity of a given motif, ISMARA provides an analysis of the correlation of motif activity and mRNA expression of the associated TFs. In particular, a table is provided showing the Pearson correlation between the motif's activity profile and the mRNA expression profiles of each of the TFs that can bind to the sites of the motif. The TFs in the list are sorted by their p-value. Supplementary Fig. B.7 shows the list of correlations for the POU2F TFs.

For each of the correlations a link is also provided to a scatter plot showing the mRNA expression levels and motif activities across the samples. Supplementary Fig. B.8 shows example scatter plots for the TFs POU2F1, POU2F2,

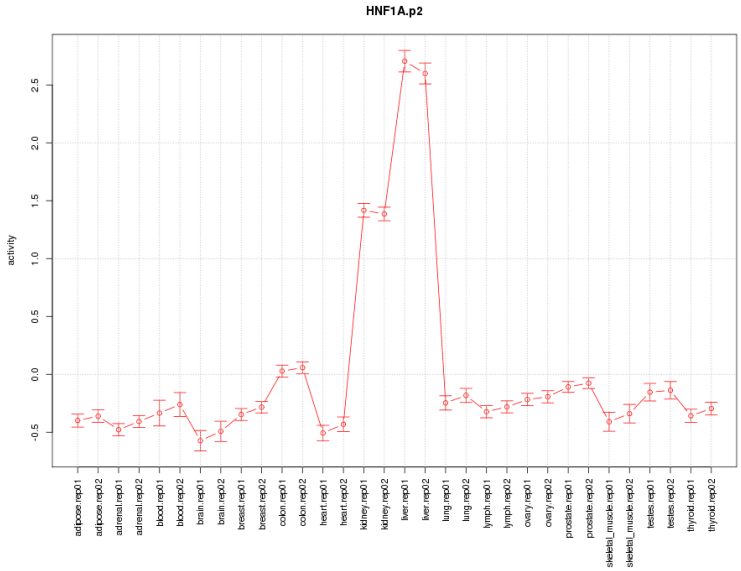


Figure B.5: Inferred activities of the HNF1A motif on the tissues of the Illumina Body Map 2. The samples are ordered, from left to right, in lexicographic order, according to sample names input by the user. Note that this causes the two replicate samples from the same tissue to appear next to each other. The red circles show the estimated activities A^*_{ms} and the error-bars δA_{ms} are shown as red vertical bars. Samples names are indicated on the bottom.

and POU2F3. Note that only POU2F2's expression is significantly correlated with the motif activity, suggesting that it is this TF that is mainly responsible for the motif activity in these samples. In addition, the fact that the TF's mRNA expression correlates *positively* with motif activity strongly suggests that this TF act as an activator, i.e. as its mRNA levels go up, the expression of target genes is affected positively.

To show an example of the opposite behavior, Suppl. Fig. B.9 shows the mRNA expression levels of the TF ZHX2 against its inferred motif activities across the Illumina Body Map 2 samples. The clear negative correlation strongly suggests that ZHX2 acts as a *repressor* of its targets, and this matches what has been reported in the literature [310].

The next important information provided for each motif, is a predicted list of target promoters. ISMARA provides the target promoters p for a motif m sorted by their target score S_{pm} (see section B.1.9). As an example, the list of top targets for the HNF1A motif is shown in Suppl. Fig. B.10. Each row in the table corresponds to one target promoter and information shown includes the promoter ID, its score S_{pm}, associated transcripts and Entrez gene symbol, and the gene's name. Note that all these pieces of information are links that take the user to additional information on the promoter, the associated transcripts, and the gene. To keep the page easily viewable, by default only the top 20 targets are shown. However, the user can interactively change the number of targets shown in the list. In addition, a search box

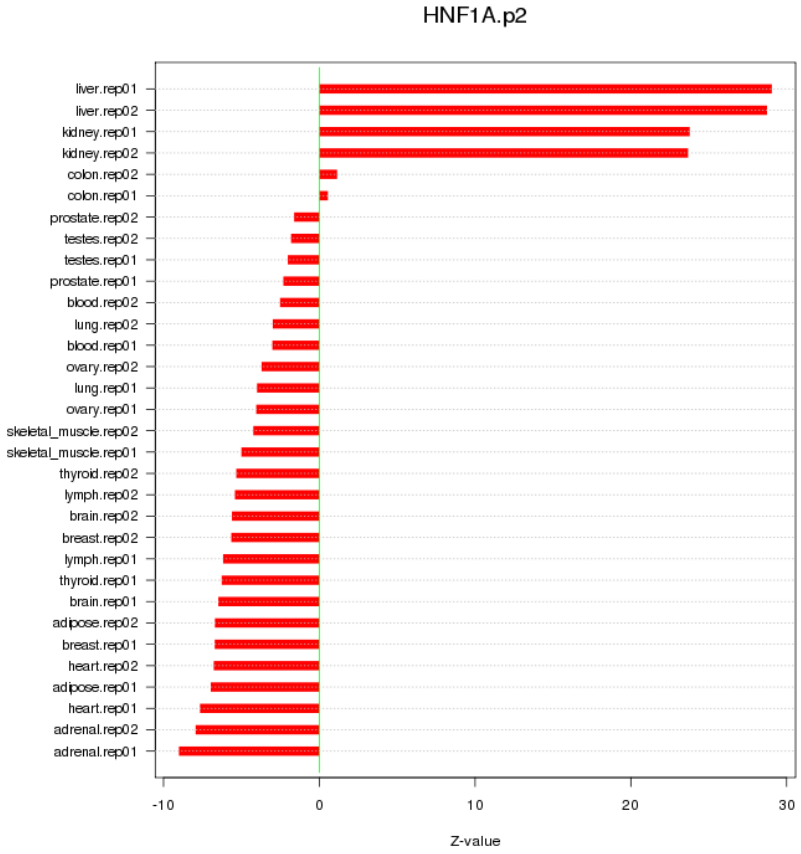


Figure B.6: Sorted list of z-values for the HNF1A motif across all samples of the Illumina Body Map 2. Note that the replicate samples from liver and kidney have much higher z-value then all other samples.

Activity-expression correlation:

Gene Symbol	Promoter	Pearson corr. coef.	P-value	Plot
POU2F2	hg19_v1_chr19_-_42636559_42636624	0.69	1.2e-05	Click!
POU2F1	hg19_v1_chr1_+_167298280_167298317	-0.24	1.9e-01	Click!
POU2F3	hg19_v1_chr11_+_120107348_120107348	0.03	8.5e-01	Click!

Figure B.7: Correlations between the HNF1A motif activity and mRNA expression profiles of TFs that can bind to sites of the motif. The table shows the names of the associated TF genes, the IDs of the associated promoters of these genes, the Pearson correlation coefficient, the p-value for the correlation, and a link to a figure showing a scatter of the motif activity and mRNA expression levels across the samples (Suppl. Fig. B.8) below.

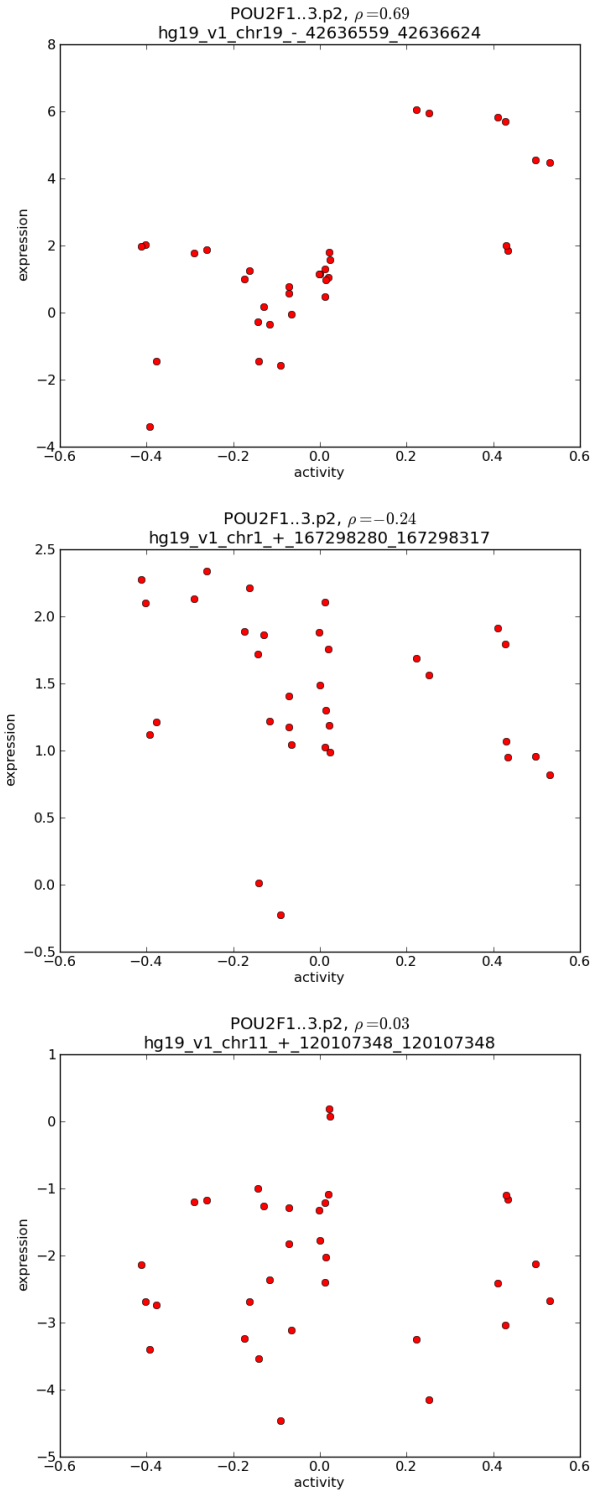


Figure B.8: Example scatter plots showing the correlations between HNF1A motif activity and the mRNA expression of POU2F2 left panel), POU2F1 (middle panel), and POU2F3 (right panel) TFs, across the samples of the Illumina Body Map 2. Each dot corresponds to one sample. The estimated expression levels correspond to the log_2 of the number of mRNAs per million mRNAs. At the top of the panel the Pearson correlation coefficient ρ and the ID of the promoter are shown.

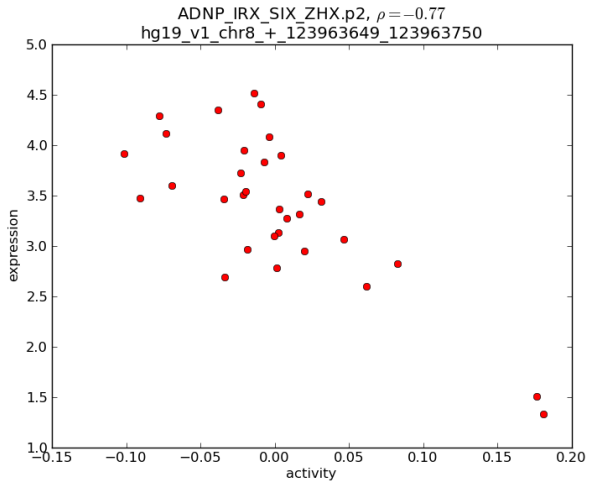


Figure B.9: Scatter plot showing the correlation between the ADNP_IRX_SIX_ZHX motif activity and the mRNA expression of the ZHX2 TF, across the samples of the Illumina Body Map 2. Each dot corresponds to one sample. The expression levels are shown on a logarithmic scale. At the top of the panel the Pearson correlation coefficient ρ and the ID of the promoter are shown.

allows the user to search whether a particular promoter, transcript, or gene of interest occurs within the full list of targets.

Of particular interest is the additional information provided about each promoter, through the links with the promoter IDs. Following this link takes the user to the genome browser of our SwissRegulon database [425], showing the section containing the proximal promoter region (500 base pairs up-stream and down-stream of the major TSS of the promoter). In this browser the user is shown all the predicted TFBSs that are used by ISMARA in its modeling of expression or ChIP-seq data. This thus allows the user to determine the precise locations of the TFBSs on the genome, through which a particular TF is predicted to target a given promoter. Supplementary Fig. B.11 shows, as an example, the promoter of the Albumin gene, which is the among the top 10 targets of HNF1A and is in fact a well-known target gene of HNF1A.

Beyond a list of individual targets, a user would typically like to gain some intuition of the pathways and particular biological processes that are targeted by a particular motif. One way of visualizing the functional structure of the predicted targets of a motif, is to represent these as a network, with links between pairs of genes that are known to be functionally related. The STRING database [312] maintains a curated collection of functional links between proteins, where 'functional link' can range from direct physical interaction, to over-representation of the protein pair within abstracts of scientific articles. For any set of proteins, STRING provides visualizations of the network of known functional interactions between these proteins, which visually brings out groups of proteins known to be functionally related. ISMARA provides such a STRING network picture for the targets of each motif (for visibility at most the top 200 targets are shown). Supplementary Fig. B.12 shows the STRING network for the predicted targets of HNF1A. Note that the picture

Search:
Show 20 entries
Showing 1 to 20 of 200 entries

Promoter	Score	Refseq	Gene Symbol	Gene Name
chr4_-_72649734	161.545	NM_000583	GC	group-specific component (vitamin D binding protein)
chr1_-_159684274	158.324	NM_000567	CRP	C-reactive protein, pentraxin-related
chr4_+_74347461	145.745	NM_001133	AFM	afamin
chr4_-_155511838	131.965	NM_000508 NM_021871	FGA	fibrinogen alpha chain
chr4_+_155484131	129.741	NM_001184741 NM_005141	FGB	fibrinogen beta chain
chr17_-_64225496	119.840	NM_000042	APOH	apolipoprotein H (beta-2-glycoprotein I)
chr4_+_74269971	115.688	NM_000477	ALB	albumin
chr1_+_159557615	106.774	NM_001639	APCS	amyloid P component, serum
chr17_+_41052813	95.277	NM_000151	G6PC	glucose-6-phosphatase, catalytic subunit
chr2_+_234668914	93.719	NM_000463	UGT1A1	UDP glucuronosyltransferase 1 family, polypeptide A1
chr5_-_147211141	91.777		SPINK1	serine peptidase inhibitor, Kazal type 1
chr2_+_234601511	89.928	NM_001072	UGT1A6	UDP glucuronosyltransferase 1 family, polypeptide A6
chr19_-_36303766	89.070		PRODH2	proline dehydrogenase (oxidase) 2
chrX_-_105282714	88.849	NM_000354	SERPINA7	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 7
chr5_-_147211235	88.367	NM_003122	SPINK1	serine peptidase inhibitor, Kazal type 1
chr8_-_17752847	88.288	NM_147203 NM_201553 NM_004467 NM_201552	FGL1	fibrinogen-like 1
chr14_-_94854910	86.867		SERPINA1	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1
chr6_+_161123224	86.262	NM_000301 NM_001168338	PLG	plasminogen
chr14_-_94789641	82.717	NM_001756	SERPINA6	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 6
chr14_-_94855120	81.964	NM_000295	SERPINA1	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1

Figure B.10: Top target promoters of the HNF1A motif for the Illumina Body Map 2. Targets are sorted by the log-likelihood score S_{pm}. Shown for each target promoter are the promoter ID (a link to the SwissRegulon web-browser page showing the promoter on the genome), the target score S_{pm}, associated RefSeq transcripts, associated gene symbols (links to NCBI pages), and gene names (which typically provide a short description of the gene's function). By default the top 20 targets are shown, but this can be changed using the drop-down menu at the top of the table. A search box allows users to search for genes or transcripts within the entire target list.

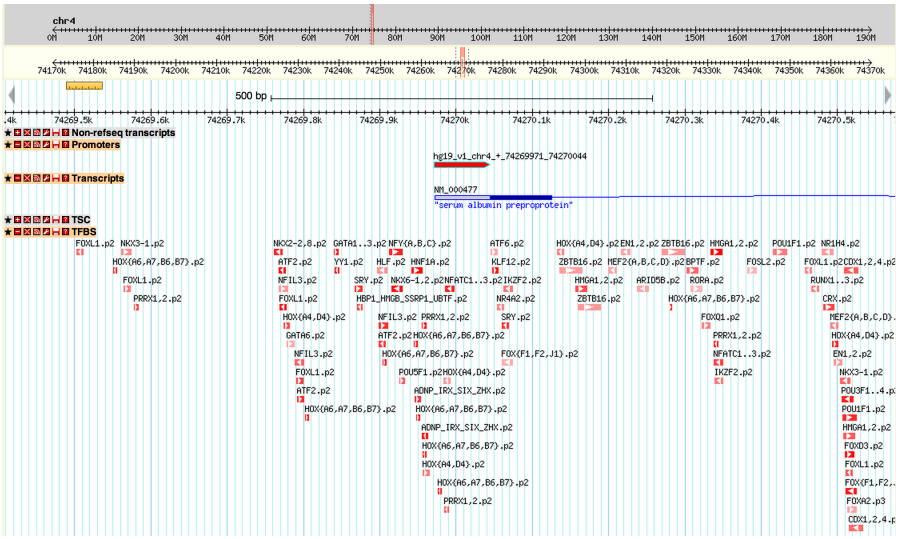


Figure B.11: Example of a promoter region as displayed in the SwissRegulon genome browser. The region shown corresponds to the proximal promoter of the Albumin gene (the 7th highest target of the HNF1A motif) and this is the region that will be displayed when following the link to the promoter displayed in Suppl. Fig. B.10. The genome browser shows the RefSeq transcript, the promoter, the associated annotated transcript start cluster (TSC) based on the CAGE data, and all the predicted TFBSs. Here the intensity of the color indicates the posterior probability assigned to each site, and the name of the cognate motif is written above each side. The arrows inside the TFBSs indicate on which strand the motif occurs. Note that an HNF1A site occurs just upstream of the TSC.

is itself a link to the STRING database, where the figure is interactive and allows the user more detailed information on each of the proteins in the network and each functional link between the proteins.

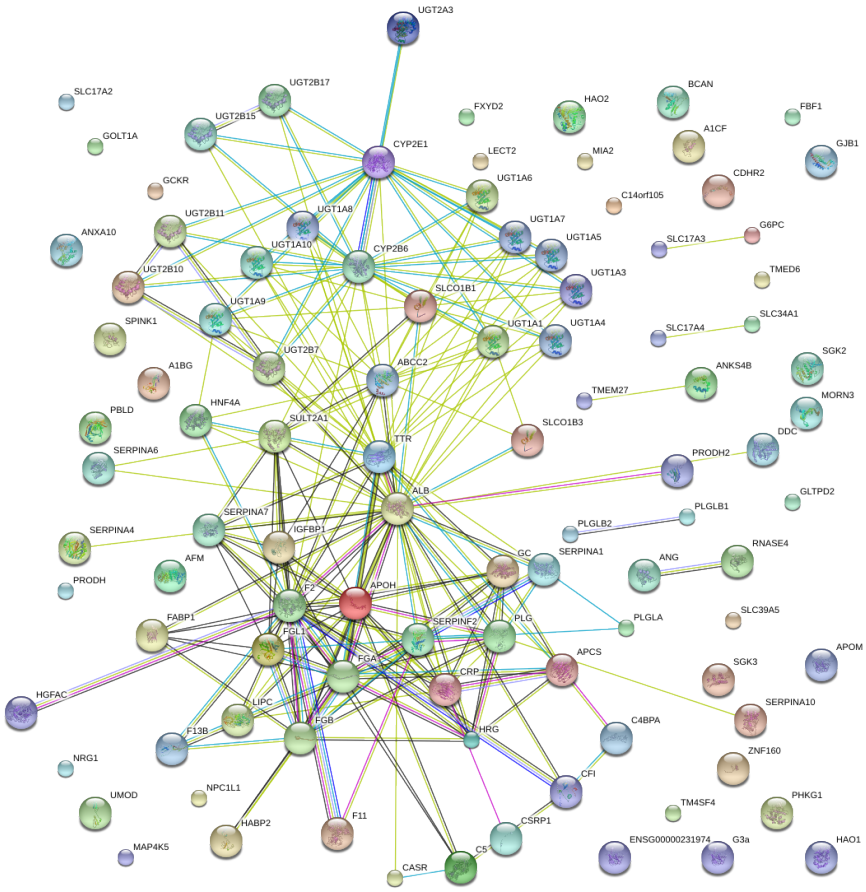


Figure B.12: Network of target genes of the HNF1A motif as displayed by the STRING database [312]. Each node corresponds to a predicted target gene of the HNF1A motif (in the Illumina Body Map 2, i.e. data set 1). Links are drawn by STRING whenever there is any evidence that the two genes may interact or be functionally linked, where evidence may range from measured direct protein-protein interaction to significant co-occurrence of the gene names within abstracts of articles.

Apart from the STRING network, ISMARA also provides list of Gene Ontology categories that are enriched among the predicted targets of a motif. Lists are provided for the 'biological process', 'cellular component', and 'molecular function' hierarchies. A p-value for enrichment is calculated using a simple hypergeometric test and only categories with a p-value below 0.05 are shown. The categories can be sorted either by the fold-enrichment of targets relative to what would be expected by chance or by the p-value of the enrichment. As an example, Suppl. Fig. B.13 shows the most significantly enriched categories of the biological process hierarchy for the HNF1A motif.
One of our our aims is to understand the causal structure of the transcription regulatory network, and a first step in that direction are predictions of direct

Gene overrepresentation in process category:

Search:
Show 10 entries
Showing 1 to 20 of 67 entries

Enrichment	P-value	GO Accession	GO Term
12.10	6.48e-15	GO:0006805	xenobiotic metabolic process
12.10	6.48e-15	GO:0071466	cellular response to xenobiotic stimulus
12.01	7.66e-15	GO:0009410	response to xenobiotic stimulus
8.02	3.15e-12	GO:0008202	steroid metabolic process
4.22	8.94e-11	GO:0006082	organic acid metabolic process
3.40	7.82e-09	GO:0006629	lipid metabolic process
8.73	4.04e-08	GO:0006820	anion transport
3.77	4.66e-08	GO:0042180	cellular ketone metabolic process
3.74	1.18e-07	GO:0019752	carboxylic acid metabolic process
3.74	1.18e-07	GO:0043436	oxoacid metabolic process
5.23	6.40e-07	GO:0032787	monocarboxylic acid metabolic process
2.20	4.08e-06	GO:0065008	regulation of biological quality
2.30	1.04e-05	GO:0044281	small molecule metabolic process
7.49	1.88e-05	GO:0006814	sodium ion transport
21.33	2.19e-05	GO:0017144	drug metabolic process
10.92	2.22e-05	GO:0015711	organic anion transport
4.09	2.34e-05	GO:0071702	organic substance transport
3.07	2.50e-05	GO:0055085	transmembrane transport
2.09	2.70e-05	GO:0042221	response to chemical stimulus
15.00	4.37e-05	GO:0030193	regulation of blood coagulation

Figure B.13: Top over-represented categories from the Gene Ontology hierarchy of biological processes among the predicted targets of the HNF1A motif. The categories are sorted by the significance of their enrichment (second column), and the first column shows the fold-enrichment relative to random expectation. The third and fourth columns in the table show the GO identifier and a description of the categories and these are again links to pages with more extensive information on the GO category. Finally, the user can interactively change the number of top categories shown using the drop-down menu or search for keywords.

regulatory interactions between the motifs. For each motif, we check its list of predicted targets for promoters of TFs that are associated with other motifs. Using this we build a regulatory network where nodes correspond to motifs and a directed edge from motif m to motif m' occurs whenever a promoter of at least one of the TFs associated with motif m' is a predicted target of motif m. On the page with results of a given motif, a part of this regulatory network centered around the motif in question is shown, i.e. all edges from or to the motif in question as well as edges between the direct neighbors of the motif. Supplementary Fig. B.14 shows the most significant interactions of this network for the HNF1A motif. Note that a slider on the left-hand side of the network allows the user to vary a cut-off on the target score S_{pm}, i.e. showing only nodes and edges over the cut-off. In addition, placing the mouse pointer over a node brings up a pop-up with the z-value of the motif, and placing the mouse pointer on an edge will bring up a pop-up with the target score of the link.

Note that ISMARA predicts that HNF1A targets HNF4A, FOXA2, NR5A2, and its own promoter. In addition, HNF4A and FOXA2 are predicted to target the HNF1A promoter as well. A literature search shows that, in fact, all these direct regulatory interactions have independent experimental support [314–319], demonstrating that the top predicted direct regulatory interactions between regulators can be highly accurate.

Finally, as described in section B.1.7.3, we also fit the average expression level $\tilde{E}_p$ of each promoter in terms of mean motif activities $\bar{A}_m$. For each motif, a z-score $\tilde{z}_m$ quantifies the significance of the motif in explaining the

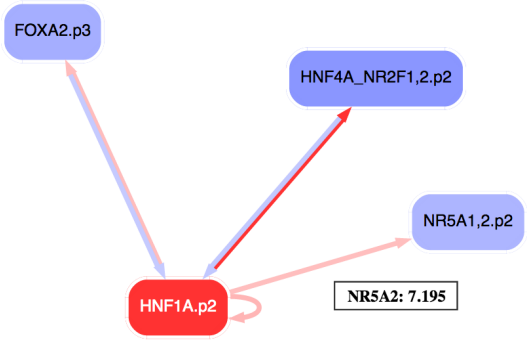


Figure B.14: The top predicted direct regulatory interactions between HNF1A and other motifs. An edge from motif m to m' is drawn whenever a promoter p, associated with motif m', is a predicted target of motif m, with target score S_{pm} larger than a given cut-off c. In the web browser, the user can interactively change the cut-off c using the slider on the left of the figure. In this example the cut-off was set at 7.195. When the cursor is placed on an edge the target score S_{pm} is shown, e.g. in this example the score of HNF1A targeting the NR5A2 promoter is shown. The intensity of the color of each motif corresponds to its z-score. Finally, only the direct network neighborhood of the motif in question (HNF1A) is shown, i.e. edges that are directly linked to HNF1A, or that link between motifs that directly link to HNF1A.

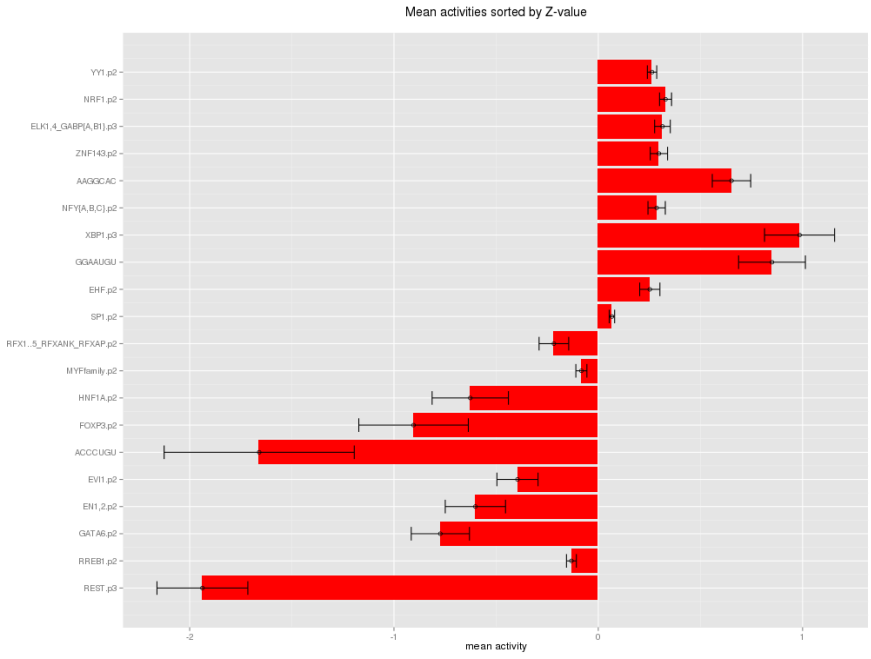


Figure B.15: Regulatory motifs most predictive for high or low average absolute expression across the IBM2 samples. For each promoter an average expression was calculated and for each motif m a mean activity $\bar{A}_m$ (red bar), and its standard-error $\delta\bar{A}_m$ (error-bar) was calculated. The z-value of a motif's mean activity is defined as the ratio $\tilde{z}_m = \bar{A}_m/\delta\bar{A}_m$ and the table shows the motifs with the most positive and most negative z-values.

mean expression level of the promoter, i.e. highly positive $\tilde{z}_m$ indicates that the occurrence of the motif is predictive for a high average expression level of the promoter, whereas a highly negative $\tilde{z}_m$ indicates that the occurrence of the motif is predictive for low average expression of the promoter.

B.4 REPRODUCIBILITY OF MOTIF ACTIVITIES

The inferred motif activities depend both on our binding site predictions, and on the assumed simple linear relationship between predicted numbers of sites and mRNA expression. As explained in the main text, there are many reasons why such a 'cartoon' model is very unlikely to produce an accurate quantitative model of genome-wide expression profiles. As a consequence, one may wonder how robust the inferred motif activities are. However, as shown in Suppl. Fig. B.16, the motif activities inferred from the two replicates of the human GNF atlas are typically more reproducible across these replicates than the expression levels of the individual promoters which are used to infer the motif activities. The reason for this is that the motif activity is inferred from the behavior of the hundreds to thousands of predicted targets of the motif. Thus, although at each individual promoter the expression is likely a complex function of the regulatory sites and the linear model is likely a poor approximation, these complications are effectively averaged out when inferring motif activities from the joint behavior of all targets.

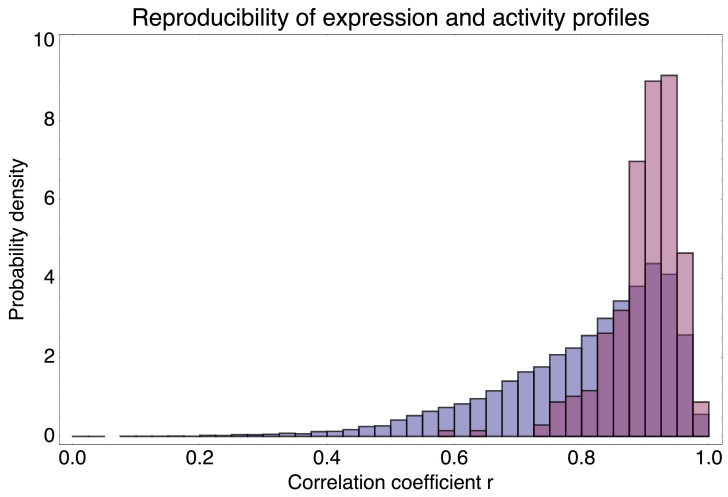


Figure B.16: Reproducibility of the inferred motif activities and the expression profiles of promoters. For each motif, and each promoter, we calculated the Pearson correlation coefficient of the activity/expression profiles for the two replicates of the samples in the human GNF atlas [334]. The figure shows the distribution of observed correlation coefficients for the motif activities (red) and promoter expression profiles (blue). The motif activities are generally considerably more reproducible than the expression profiles of the promoters from which they are inferred.

B.5 MOTIFS DIS-REGULATED IN TUMOR CELLS

To identify motifs whose motif activities are consistently dis-regulated in tumors, we first separated all samples s from the GNF and NCI-60 data sets into the set of tumor samples T and non-tumor samples N. Next, we used the replicate averaging described in section B.1.8 to calculate, for each motif, an average activity $\langle \bar{A}^T \rangle$ in tumor samples, an associated error-bar $\delta \bar{A}^T$, an av-

erage activity in non-tumor samples $\langle \bar{A}^N \rangle$, and an error-bar $\delta \bar{A}^N$ associated with the average activity in non-tumor samples. From these, we calculate a z-value z_m for each motif m that quantifies the significance of the difference in the average activities in tumor and non-tumor samples. Tables B.2 and B.3 show the motifs with highest and lowest z-values, respectively. That is, these are the motifs most significantly dis-regulated in tumor cells.

Motif	z**-values**
blah_family.p2	2.398858
HIF1A.p2	2.230493
E2F1..5.p2	2.140652
ARNT_ARNT2_BHLHB2_MAX_MYC_USF1.p2	2.071274
BPTF.p2	1.977484
NFY{A,B,C}.p2	1.920594
FOXD3.p2	1.915846
TFDP1.p2	1.901083
ELF1,2,4.p2	1.874818
ZNF143.p2	1.802732
ATF4.p2	1.786143
YY1.p2	1.735238
EHF.p2	1.718308
NRF1.p2	1.674024
ELK1,4_GABP{A,B1}.p3	1.667680
CCUUCAU (hsa-miR-205)	1.525379
PAX5.p2	1.500615
UCAAGUA (hsa-miR-26a, hsa-miR-26b, hsa-miR-1297, hsa-miR-4465)	1.404557
BACH2.p2	1.371868
GUAACAG (hsa-miR-194)	1.349047
HES1.p2	1.317505

Table B.2: Motifs that are most consistently upregulated in tumor samples of the NCI-60 and GNF data sets, relative to healthy (non-tumor) tissues in the GNF data set. The motifs are sorted by their z-value (shown in the second column).

Motif	z-values
SMAD1..7,9.p2	-2.194113
HAND1,2.p2	-2.185943
TGIF1.p2	-2.117814
MAZ.p2	-2.076224
TFCP2.p2	-2.071225
KLF12.p2	-1.958392
GGCUCAG (hsa-miR-24)	-1.918863
FOX{D1,D2}.p2	-1.839199
TBX4,5.p2	-1.805228
FOXP3.p2	-1.740035
EVI1.p2	-1.701934
HBP1_HMGB_SSRP1_UBTF.p2	-1.688854
AAAGUGC (hsa-miR-17, hsa-miR-20a, hsa-miR-20b, hsa-miR-93, hsa-miR-106a, hsa-miR-106b, hsa-miR-519d)	-1.628037
GAGAUGA (hsa-miR-143, hsa-miR-4770)	-1.619611
HIC1.p2	-1.607936
NANOG{mouse}.p2	-1.576193
FEV.p2	-1.574951
MYOD1.p2	-1.565920
NR1H4.p2	-1.562673
POU1F1.p2	-1.556216
TCF4_dimer.p2	-1.536692
MYFfamily.p2	-1.514719
TAL1_TCF{3,4,12}.p2	-1.499900
POU5F1.p2	-1.480033
NR3C1.p2	-1.473553
HOX{A5,B5}.p2	-1.440485
STAT1,3.p3	-1.417964
GTF2A1,2.p2	-1.416557
RORA.p2	-1.391819
CAGCAGG (hsa-miR-214, hsa-miR-761, hsa-miR-3619-5p)	-1.356781
ETS1,2.p2	-1.337667
EN1,2.p2	-1.337051
AR.p2	-1.330996
RREB1.p2	-1.330444
CUCCCAA (hsa-miR-150)	-1.318296
CACAGUG (hsa-miR-128)	-1.318135
JUN.p2	-1.313498

Table B.3: Motifs that are most consistently down-regulated in tumor samples of the NCI-60 and GNF data sets, relative to healthy (non-tumor) tissues in the GNF data set. The motifs are sorted by their z-value (shown in the second column).

Many of the TFs that ISMARA identifies as disregulated in cancer are well-known in cancer biology, including HIF1A[574] (Suppl. Fig. B.17), MYC[575], E2F1..5[576], NF-Y[577], YY1[578], TFCP2[579], and the SMAD TFs[580]. However, our brief survey of the literature also suggests that several other TFs that ISMARA identifies as consistently disregulated in cancers are currently not recognized as major players in cancer biology, although there is some evidence in the literature that these TFs may play a role in cancer. These TFs include HAND1,2[581], KLF12[582], BPTF[583], FOXD3[584], and ZNF143[585].

ISMARA also identifies a number of miRNAs whose targets are either consistently upregulated in tumors, e.g. hsa-miR-205 (Suppl. Fig. B.17) and hsa-miR-26, or consistently down-regulated, e.g. hsa-miR-24 and the hsa-miR-17/93/106 seed family. Indeed, multiple studies have found hsa-miR-205 to be down-regulated in a number of different cancers, and hsa-miR-205 has been shown to have tumor suppressor function[586–590]. It has also been

shown that hsa-miR-26a delivery suppresses hepatic tumors in mouse[591], supporting the downregulation of this miRNA in cancer. Conversely, hsa-miR-17 is a known oncogene[592], consistent with the downregulation of its targets in cancer. The literature on hsa-miR-24 function in cancer is more ambiguous[593]. Some evidence has been provided that hsa-miR-24 acts as repressor of apoptosis and is upregulated in certain cancers[594]. On the other hand, another study found that hsa-miR-24 can inhibit proliferation[595]. Notably, the latter study suggested that hsa-miR-24 acts through seedless target sites, which by construction are not detected by TargetScan. In summary, in this system ISMARA successfully identified oncogenes and tumor suppressors *ab initio*.

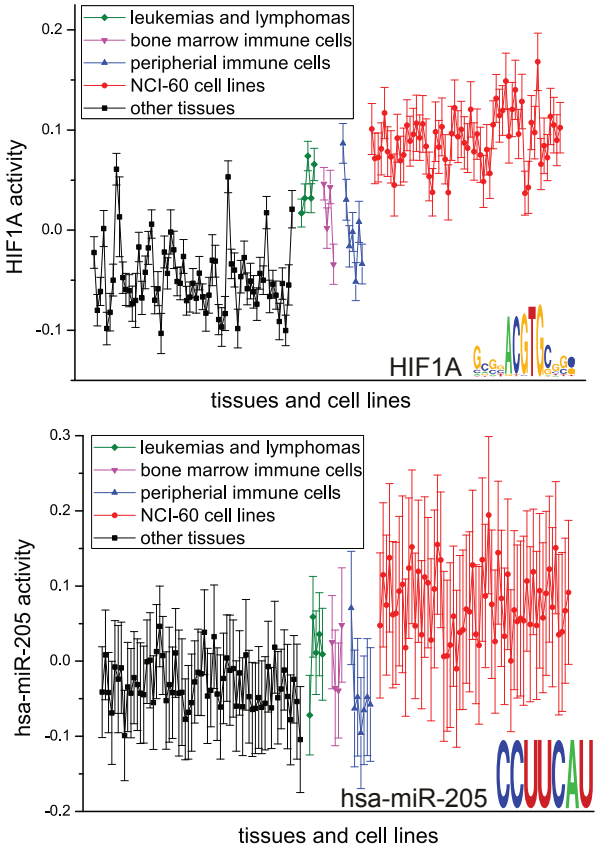


Figure B.17: Motif activities (with error-bars) across the human GNF and NCI-60 samples for an example TF (HIF1A, left panel) and miRNA motif (hsa-miR-205, right panel) that are disregulated in cancer. Note that different subsets of samples are colored differently as indicated in the legend.

B.6 EXAMPLE OF SPECIES-SPECIFIC TARGETING

The MotEvo algorithm that we use for predicting TFBSs in all promoters operates on multiple alignments and incorporates information on binding site conservation using an explicit model of TFBS evolution. This does not mean, however, that MotEvo only predicts binding sites that are well-conserved

across orthologous promoters in mammals. Although evidence of conservation increases the posterior probability assigned to a given TFBS, species-specific TFBSs that are predicted to have high-affinity for the regulator can also attain high posterior probability. Consequently, ISMARA will typically also identify targets that are species-specific or specific to a subclade of closely-related species, e.g. primate-specific targets. An example of a primate-specific target is ISMARA's prediction that, in the innate immune response time course in HUVEC cells, the IRF motif targets the promoter of the ATF5 transcription factor. As shown in Suppl. Fig. B.18, the corresponding TFBS for IRF in the ATF5 promoter is primate-specific, i.e. only conserved in Rhesus Macaque.

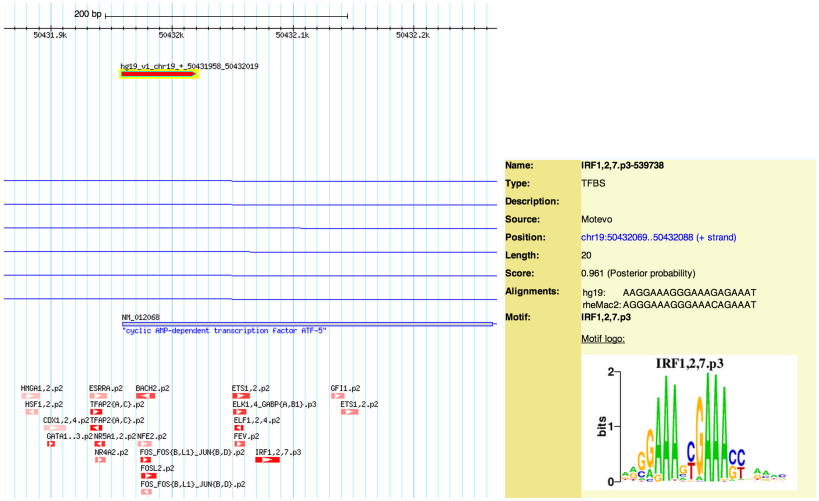


Figure B.18: Example of a primate-specific target prediction of ISMARA. ISMARA predicts that the IRF motif targets the ATF5 promoter in the innate immune response time course of HUVEC cells. **Left panel:** Close-up of the predicted TFBSs in the ATF5 promoter, as displayed in the SwissRegulon genome browser [425]. The predicted IRF site occurs roughly 60 base pairs downstream of the promoter. **Right panel:** Detailed information on the IRF site in the ATF5 promoter. Besides human, an orthologous IRF site is only found in Rhesus Macaque. However, because of the site's strong match to the IRF motif, the site is still assigned a high posterior probability.

B.7 VALIDATION OF PREDICTED NFκB TARGETS USING CHIP-SEQ DATA

To assess the accuracy of the genome-wide targets that ISMARA predicts we compared the predicted targets of NFκB in the innate immune response time course in which HUVEC cells were treated with TNFα with NFκB targets based on ChIP-seq experiments.

We collected data on NFκB binding sites in 10 lymphoblastoid cell lines derived from 10 different individuals of African, European, and Asian ancestry [343]. From this study we obtained predicted peaks for 33 ChIP-seq samples (10 different individuals with between 2 and 5 replicates per individual),

with each peak's significance quantified by a z-value. Because ISMARA's predictions are exclusively associated with promoters, we focused on ChIP-seq peaks associated with each of the promoters in our promoter set. For each human promoter, and each of the 33 ChIP-seq data-sets, we calculated a binding score by summing the z-values of all peaks whose center fall within 1 kilobase of the center of the promoter. Then, for each human promoter, we calculated a final binding score by averaging the binding scores across the 33 ChIP-seq data-sets. Using a cut-off score of $z = 4.5$, a little over 8% of promoters (2969 of 35821) are then classified as showing significant evidence of binding.

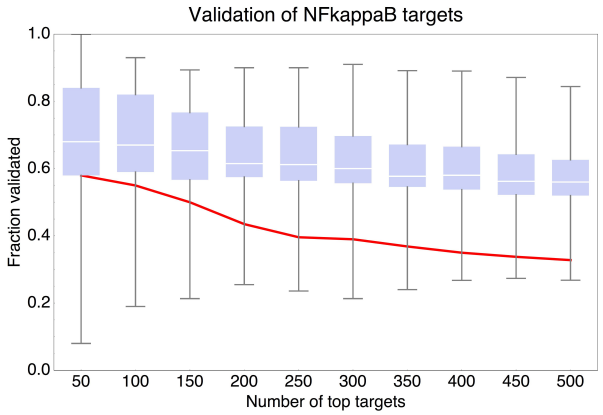


Figure B.19: Validation of ISMARA's predicted NFκB targets from the innate immune response time course in HUVEC cells using ChIP-seq data from lymphoblastoid cell lines in 10 different individuals [343]. The red line shows the percentage of top predicted target promoters that have a ChIP-seq binding peak as a function of the number of top predicted promoters. The box-plot indicates the variation in ChIP-seq binding across samples from the different individuals. In particular, for each of 33 ChIP-seq samples, its target promoters are 'validated' by comparison with the other 32 ChIP-seq samples exactly in the same way as for the ISMARA targets. The box-plot shows the 5, 25, 50, 75, and 95 percentiles of the distribution of percentages of validated targets across the 33 samples.

There we 2636 promoters that had a predicted regulatory site for NFκB. Sorting these 2636 human promoters by their ISMARA target score for NFκB, we then calculated the fraction of the top 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500 targets that show binding according to the ChIP-seq data (red line in Fig. B.19). Almost two-thirds of the top 50 targets are validated by ChIP-seq binding, more than half of the top 150 targets, and about 40% of the top 300 targets.

To compare this validation of predicted targets with the reproducibility of the ChIP-seq data themselves across replicates and individuals, we 'validated' the binding promoters as measured by each ChIP-seq data-set by the average of all other ChIP-seq data-sets. Specifically, for each ChIP-seq data-set, we sorted all promoters by its binding score, and then calculated what fraction of top targets have an average binding score over the cut-off according to all *other* ChIP-seq data-sets. Doing this for all 33 samples we obtained a distribution of the fraction of validated top x targets and calcu-

lated median, inter-quartile range, and 5 and 95 percentile for each value of $x \in \{50, 100, 150, 200, 250, 300, 350, 400, 450, 500\}$ (the box-whisker plots in Suppl. Fig. B.19). We observed that the validation rate for ChIP-seq targets is higher, i.e. typically between 60 and 70 percent, than for the ISMARA targets. To some extent this may result from the fact that all ChIP-seq data were obtained in the same cell type, which was different from the HUVEC cells used in the innate immune response time course. However, there was considerable variability in the validation rates of ChIP-seq samples, and some samples had lower validation rates than ISMARA targets. This result shows that the accuracy of ISMARA's target predictions are comparable to targets obtained through ChIP-seq.

B.8 XBP1 MOTIF ACTIVITY AND MRNA EXPRESSION

The XPB1 motif is the third most significant motif in the innate immune response time course in which HUVEC cells were treated with TNFα. The motif is upregulated during the time course. However, as shown in Suppl. Fig. B.20, the mRNA expression of the XPB1 gene is almost constant across the time course, and not significantly correlated with the motif's activity. In fact, it has been established that XBP1's activity is regulated post-transcriptionally, i.e. through alternative splicing [350, 351].

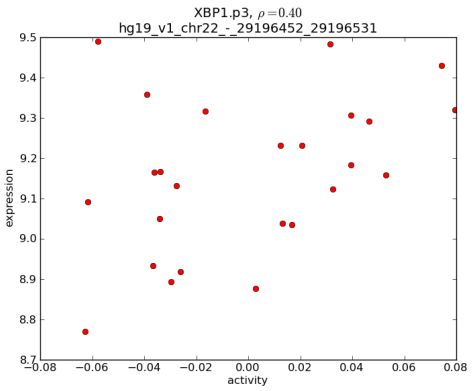


Figure B.20: Scatter plot showing the correlation between the inferred activity of the XBP1 motif and the mRNA expression of the XBP1 gene for the innate immune response time course. The mRNA expression is shown on a logarithmic scale (base 2) along the vertical axis. Note the small range in expression variation.

B.9 EPITHELIAL-MESENCHYMAL TRANSITION: INCLUDING MICRORNAS IN CORE REGULATORY NETWORKS

To illustrate ISMARA's ability to integrate the role of both TFs and miRNAs in the gene regulatory network, we took advantage of data from a system in which miRNAs are known to play important regulatory roles: the epithelial-to-mesenchymal transition (EMT). Recently, mRNA expression measurements were performed in duplicate on epithelial and 3 independently-

isolated mesenchymal subpopulations within immortalized mammary epithelial cells[355]. After running ISMARA on this data (results at http://ismara.unibas.ch/supp/dataset5/ismara_report/), we used replicate-averaging to identify regulators that most consistently and significantly explain the mRNA expression differences between epithelial and mesenchymal cells (results at http://ismara.unibas.ch/supp/dataset5/averaged_report/).

Remarkably, much of what is known about EMT (reviewed by Polyak and Weinberg[356]) is inferred automatically by ISMARA using only the gene expression data. In particular, among the top regulators that ISMARA infers in this system are SNAI1..3, ZEB1, and a family of miRNAs consisting of hsa-miR-141 and hsa-miR-200a (sharing the same seed sequence), that have been shown to form a regulatory network essential for EMT. The predicted activity changes of these regulators are consistent with the extant literature. Namely, the decrease in SNAI1..3 and ZEB1 activity (which indicates a reduced level of their predicted targets) in mesenchymal subpopulations is consistent with the fact that both of them are mainly acting as repressors and are transcriptionally up-regulated in the mesenchymal state. The hsa-miR-141 and hsa-miR-200a miRNAs are known to be down-regulated in the mesenchymal state, causing the mRNA levels of their targets to increase, which matches the positive change in activity predicted by ISMARA. Known regulatory interactions between these factors are also uncovered by ISMARA. For instance, *ZEB1* is the top predicted target of the hsa-miR-141/200a miRNAs and existing literature confirms that the direct regulation of *ZEB1* by hsa-miR-200 is critical in EMT[596–598]. Similarly, promoters of E-cadherin (*CDH1*) gene are the 3rd and 4th top target promoters of the ZEB1 and SNAI1..3 motifs, respectively, and indeed this gene is an epithelial marker known to be targeted by both SNAIL transcription factors[599] and by ZEB1[600]. These key predictions by ISMARA are summarized in Fig. B.21.

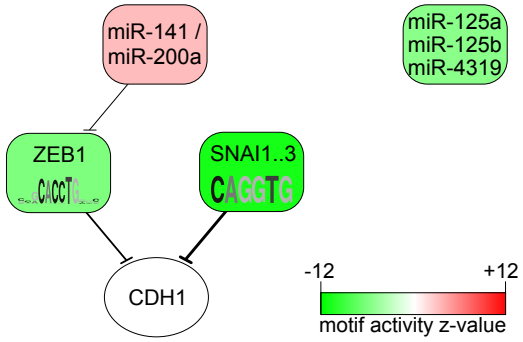


Figure B.21: Core TF and miRNA regulatory interactions in the epithelial-to-mesenchymal transition, as predicted by ISMARA. Each rectangular node corresponds to a regulatory motif with its color indicating the significance of the change in activity when going from the epithelial to mesenchymal state (z-value defined as $z = (A_{m,\mathrm{mes}} - A_{m,\mathrm{epi}}) / \sqrt{\delta A^2_{m,\mathrm{mes}} + \delta A^2_{m,\mathrm{epi}}}$). Green/Red indicates targets of the motif are down/up-regulated in the mesenchymal state. Both Zeb1 and Snail are predicted to target the E-cadherin (CDH1) promoter. Note that all interactions shown are repressive.

Cell	Description
GM12878	B-lymphocyte, lymphoblastoid
HepG2	hepatocellular carcinoma
HMEC	mammary epithelial cells
HSMM	skeletal muscle myoblasts
Huvec	umbilical vein endothelial cells
K562	chronic myelogenous leukemia
NHEK	epidermal keratinocytes
NHLF	lung fibroblasts

Table B.4: Human tissues and cell lines used as the source of experimental material in the ENCODE data sets for which we analyzed ChIP-seq data of chromatin marks. We used all available samples for which a consistent measurement platform was used.

The activity of the family containing the hsa-miR-125a/b and hsa-miR-4319 miRNAs is the most significantly reduced miRNA family in EMT. This suggests that these miRNAs play a role in mesenchymal cells, consistent with observations that hsa-miR-125b promotes invasive tumor characteristics[601].

B.10 ANALYSIS OF THE ENCODE CHIP-SEQ DATA

To illustrate ISMARA's performance on ChIP-seq data we used data from the ENCODE Project in which expression and 9 different chromatin modifications were measured across 8 different cell types[357]. Supplementary table B.4 shows the list of cell types used together with their description and Suppl. table B.5 shows a list of all the signals that were measured. For simplicity, we will refer to all 10 signals (which include expression and the binding of the CTCF transcription factor) as 'marks' in our description below.
We first ran ISMARA separately on the data sets for each of the 10 signals. For all the ChIP-seq data we thus modeled the occurrence of each of the marks at promoters in terms of the predicted TFBSs at the promoters. Supplementary table B.6 lists all the data sets that were analyzed in this paper and shows, including references to the original publications, and lists for each data set the URL at which ISMARA's results for the corresponding data set can be found. Note that, for data sets 1, 2, and 5, there are also replicate averaged results available. These can be found by replacing 'ismara_report' at the end of the URL with 'averaged_report'.

Profiling	Platform
expression	Affymetrix HT Human Genome U133A Array
H3K4me3	Illumina Genome Analyzer II
H3K27me3	Illumina Genome Analyzer II
H3K27ac	Illumina Genome Analyzer II
H3K9ac	Illumina Genome Analyzer II
H3K36me3	Illumina Genome Analyzer II
H3K4me1	Illumina Genome Analyzer II
CTCF	Illumina Genome Analyzer II
H3K4me2	Illumina Genome Analyzer II
H4K20me1	Illumina Genome Analyzer II

Table B.5: List of the signals (i.e. expression, histone modifications, and the binding of one TF) and corresponding measurement platforms from the ENCODE data sets, that we used to demonstrate ISMARA's performance on ChIP-seq data sets. We used available BED and CEL files from the GSE26386 and GSE26312 GEO series.

Data Set	ISMARA URL
Illumina body map 2	ismara.unibas.ch/supp/dataset1_IBM/ismara_report
GNF SymAtlas + NCI-60 cancer cell lines, human [334, 335]	ismara.unibas.ch/supp/dataset2/ismara_report
Inflammatory response time course, HUVEC [336]	ismara.unibas.ch/supp/dataset3/ismara_report
Mucociliary differentiation, bronchial epithelial cells, human [352]	ismara.unibas.ch/supp/dataset4/ismara_report
Epithelial-Mesenchymal Transition, human [355]	ismara.unibas.ch/supp/dataset5/ismara_report
ENCODE cell lines, expression [357]	ismara.unibas.ch/supp/dataset6.1_ENCODE_expression/ismara_report
ENCODE cell lines, H3K4me3 [357]	ismara.unibas.ch/supp/dataset6.2_ENCODE_H3K4me3/ismara_report
ENCODE cell lines, H3K27me3 [357]	ismara.unibas.ch/supp/dataset6.3_ENCODE_H3K27me3/ismara_report
ENCODE cell lines, H3K27ac [357]	ismara.unibas.ch/supp/dataset6.4_ENCODE_H3K27ac/ismara_report
ENCODE cell lines, H3K9ac [357]	ismara.unibas.ch/supp/dataset6.5_ENCODE_H3K9ac/ismara_report
ENCODE cell lines, H3K36me3 [357]	ismara.unibas.ch/supp/dataset6.6_ENCODE_H3K36me3/ismara_report
ENCODE cell lines, H3K4me1 [357]	ismara.unibas.ch/supp/dataset6.7_ENCODE_H3K4me1/ismara_report
ENCODE cell lines, CTCF [357]	ismara.unibas.ch/supp/dataset6.8_ENCODE_CTCF/ismara_report
ENCODE cell lines, H3K4me2 [357]	ismara.unibas.ch/supp/dataset6.9_ENCODE_H3K4me2/ismara_report
ENCODE cell lines, H4K20me1 [357]	ismara.unibas.ch/supp/dataset6.10_ENCODE_H4K20me1/ismara_report

Table B.6: URLs with the results of ISMARA's analyses of the data sets discussed in this paper.

B.10.1 *PCA analysis*

We first performed principal component analysis of the 10 marks across all promoters genome-wide, separately for each of the 8 cell types, as described in section B.1.10. As shown in Suppl. Fig. B.23, we find that the first principal component explains approximately 60% of the variation in each of the 8 cell types. In addition, the first principal component is almost identical in each of the cell types. This strongly suggests that this first principal component is a general feature of the distribution of chromatin marks. Moreover, the fact that this component aligns positively with expression and activity-associated chromatin marks, suggests that this first component reflects general promoter activity. We then pooled the data from all samples and performed principal component analysis on this complete data set, i.e. treating each promoter sam-

ple combination (p, s) as if it were a separate promoter. The resulting first principal component is shown in Fig. 6B of the main article.

Next, as described in section B.1.10, we took the inferred motif activities for all marks and removed the component along the first principal component. That is, we removed the contribution to the motif activities that comes from the general 'promoter activity'. As an illustration, Suppl. Fig. B.22 shows the inferred motif activities for 5 motifs (SNAI, IRF, HNF4a_NR2F1, TEAD1, and GATA6) both before (left panels) and after (right panels) the contribution from general promoter activity has been removed, for expression and the activation associated marks H3K4me3, H3K4me2, H3K9ac, H3K27ac, and H3K36me3. As the figure shows, before removal of the first PCA component, the activities for all marks are highly correlated, but this correlation disappears when the first PCA component is removed. This confirms that the highly correlated motif activities and the activation-associated chromatin marks is accounted for by the first PCA component that captures the relative chromatin mark levels associated with the general activity of a promoter. The remaining activities (right panels) thus provide a clearer insight in the specific role of a motif for specific marks across the cell-types. For example, for the SNAI motif the two acetylation marks are highly positive in HepG2 cells, whereas expression and H3K36me3 are clearly negative. Thus, promoters carrying SNAI sites tend to have higher histone acetylation levels than expected based on their general activity, and lower gene expression and H3K36me3 levels than expected based on the general activity.

As described in section B.1.10, after removing the contribution of the first principal component to the motif activities, we re-calculated significance z-values z_m^i for each motif m and each mark i. In addition, we calculated a specificity s_m^i which measures the fraction of the overall that is associated with mark i. That is, a motif m will be highly specific for mark i if it has a high z-value z_m^i, and low z-values for all other marks. To identify motifs that are either most significant or highly specific for particular marks, we plotted scatter plots showing the significance and specificity for each motif (Suppl. Fig. B.24). In each of the scatters we have indicated in red those motifs that had either very high significance or high specificity for the motif. Interestingly, we often find that the motifs with highest significance for a particular mark also have high specificity. For example, HNF1a is both most significant and most specific for H3K4me2 levels in promoters. Not surprisingly, the occurrence of CTCF motifs is the most significant determinant of the observed levels of bound CTCF.

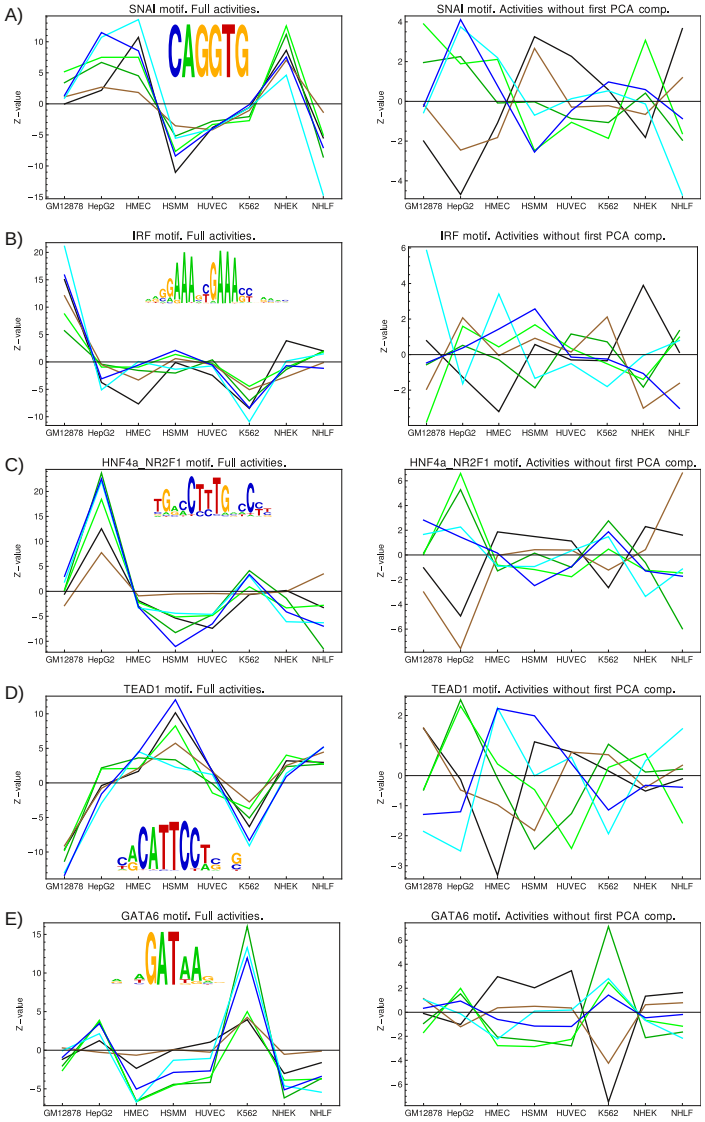


Figure B.22: Inferred motif activities for 5 example motifs on the ENCODE ChIP-seq data sets measuring chromatin [357]. Each row (labeled A through E) shows the activities for explaining expression (black), H3K4me3 (dark green), H3K4me2 (light green), H3K9ac (dark blue), H3K27ac (light blue), and H3K36me3 (brown) levels, for one motif. The left panels show the motif activities as inferred from the original data the right panel the motif activities after the contribution along the first principal component has been subtracted. The names of the motifs are indicated above each panel and sequence logos are shown as insets. Note that the motif activities for the different marks go from highly correlated to essentially uncorrelated as the first principal component is removed.

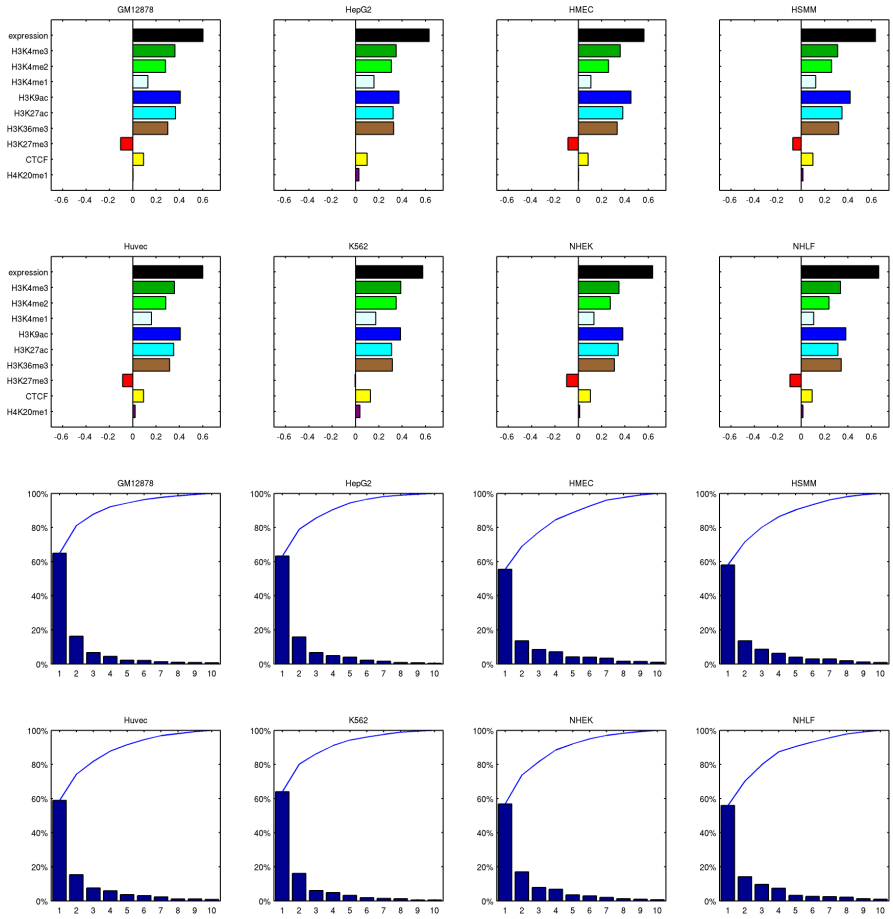


Figure B.23: First principal component explaining the largest amount of chromatin mark and expression levels associated with each promoter, separately for each of the 8 **cell types (top** 8 **panels).** The bars indicate the relative contributions of expression and each of the chromatin marks to the first principal component. Note that the first principal component is virtually identical in each cell type. The bottom 8 panels show the fraction of the total variance explained by each subsequent principal component (bars) and the cumulative fraction of variance explained by consecutive components. Note that, for each cell type, close to 60% of the variance in expression and the 9 chromatin marks is explained by the first component.

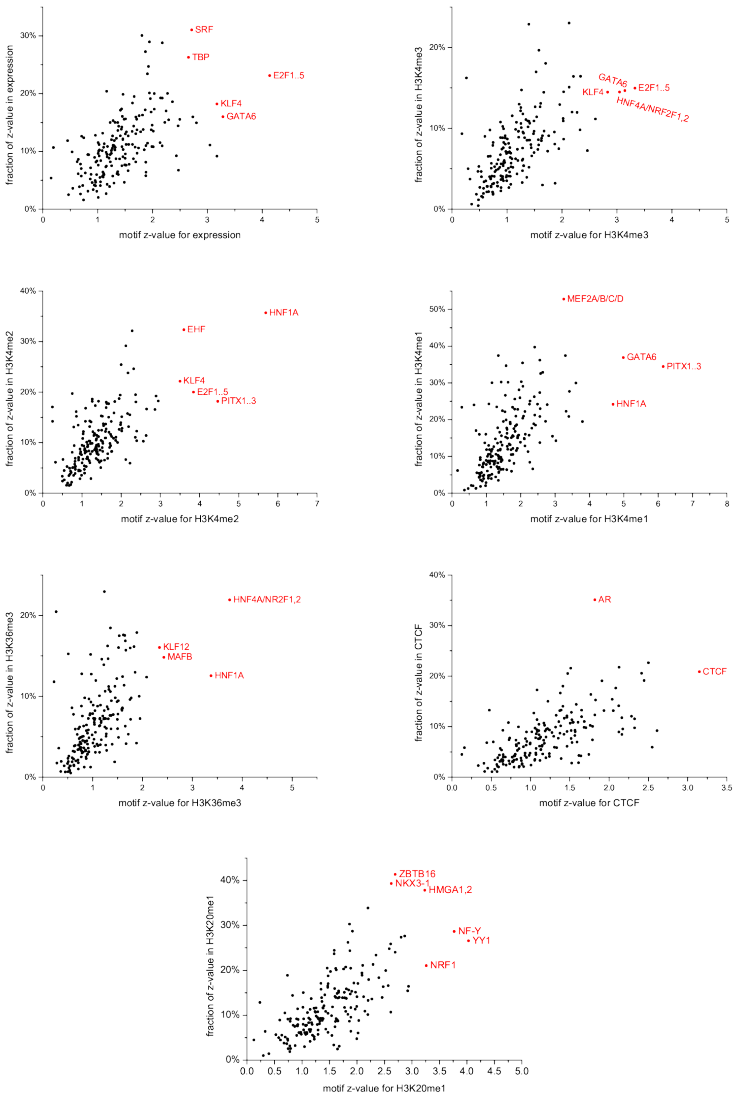


Figure B.24: Significances and specificities of the motifs for explaining variations in different chromatin marks. Each panel corresponds to one mark (as indicated on the axes) and each dot corresponds to one motif. The significance of each motif is quantified by a z-value of the motif's activity for a given mark, after motif activities along the first principal component have been removed (see section B.1.10). The specificity of a motif for a given mark is the fraction of all significance associated with a given mark (its z-value squared relative to the sum of all z-values squared, see section B.1.10). the most significant and/or specific motifs for each mark are indicated in red.

C

SUPPLEMENTARY MATERIAL TO CHAPTER 5

C.1 SUPPLEMENTARY FIGURES

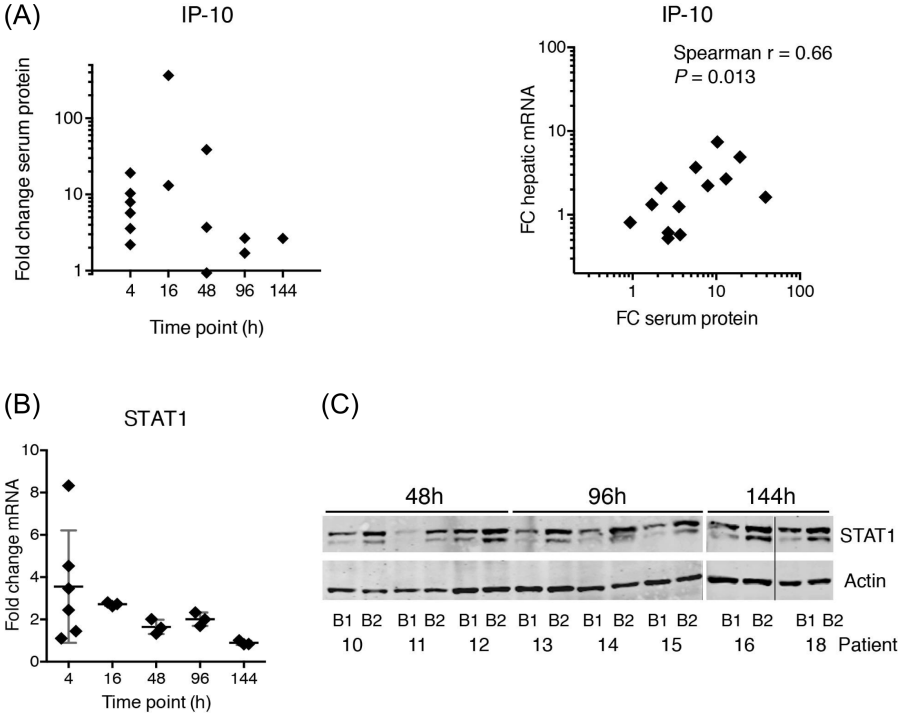


Figure C.1: Gene expression induction is confirmed on protein level. **(A)** IP-10 serum protein detected by ELISA significantly correlates with paired hepatic IP-10 mRNA. Shown are fold changes (FC) over time. Spearman correlation analysis was performed. Significance is indicated on the plot. **(B)** Fold induction of STAT1 mRNA in paired liver samples at indicated time points. Indicated are additionally mean with SEM. **(C)** STAT1 WB of paired liver samples shows increased protein levels in B2. Protein expression of the 4h time point was shown previously [389]. The black line separates non-contiguous bands on the same gel.

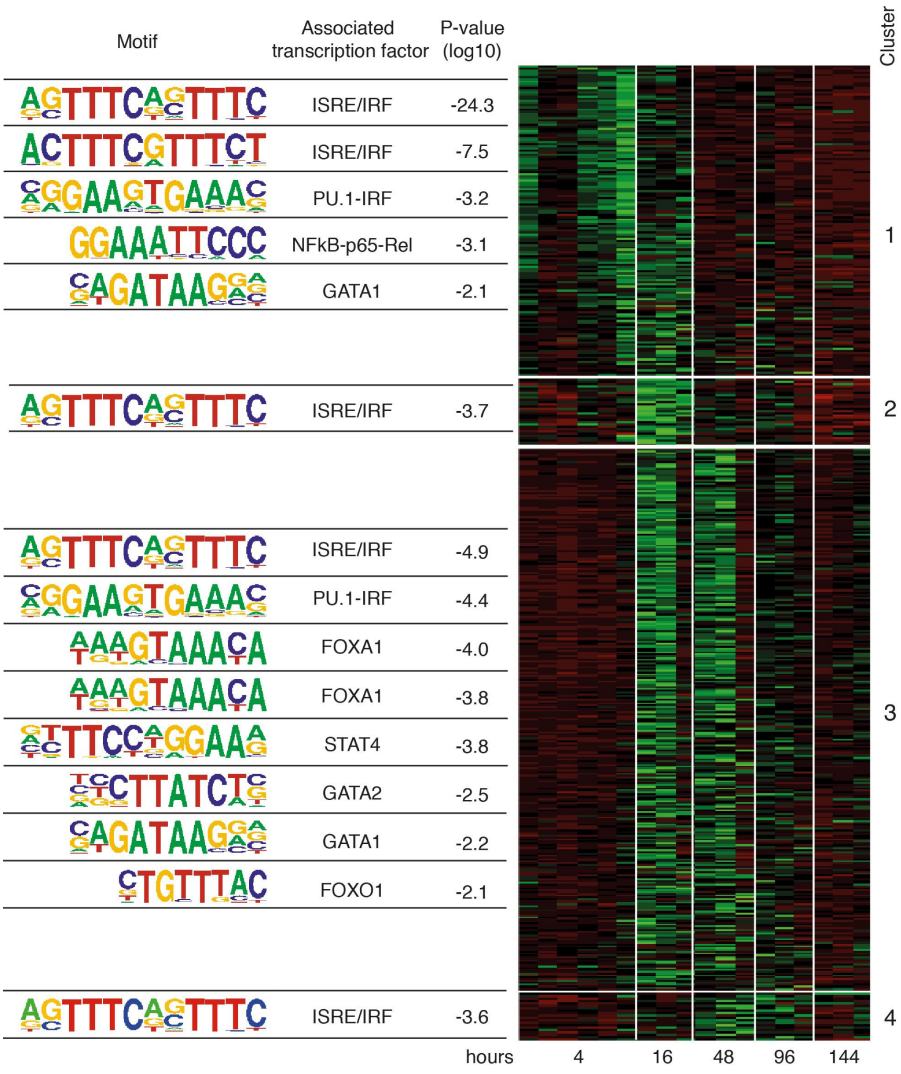


Figure C.2: *In silico* transcription factor binding analysis reveals ISRE as the main promoter binding site in every gene expression cluster. Known TFBS in promoter regions (2 kbp upstream and 500 bp downstream of the transcription start site) of every gene in each cluster were analysed by HOMER software. Indicated are the motif, the name and the significance of enrichment of TFBS in each cluster. The heatmap represents all genes in the cluster normalized to the mean expression of all samples for this gene, with green indicating higher than average and red below average expression.

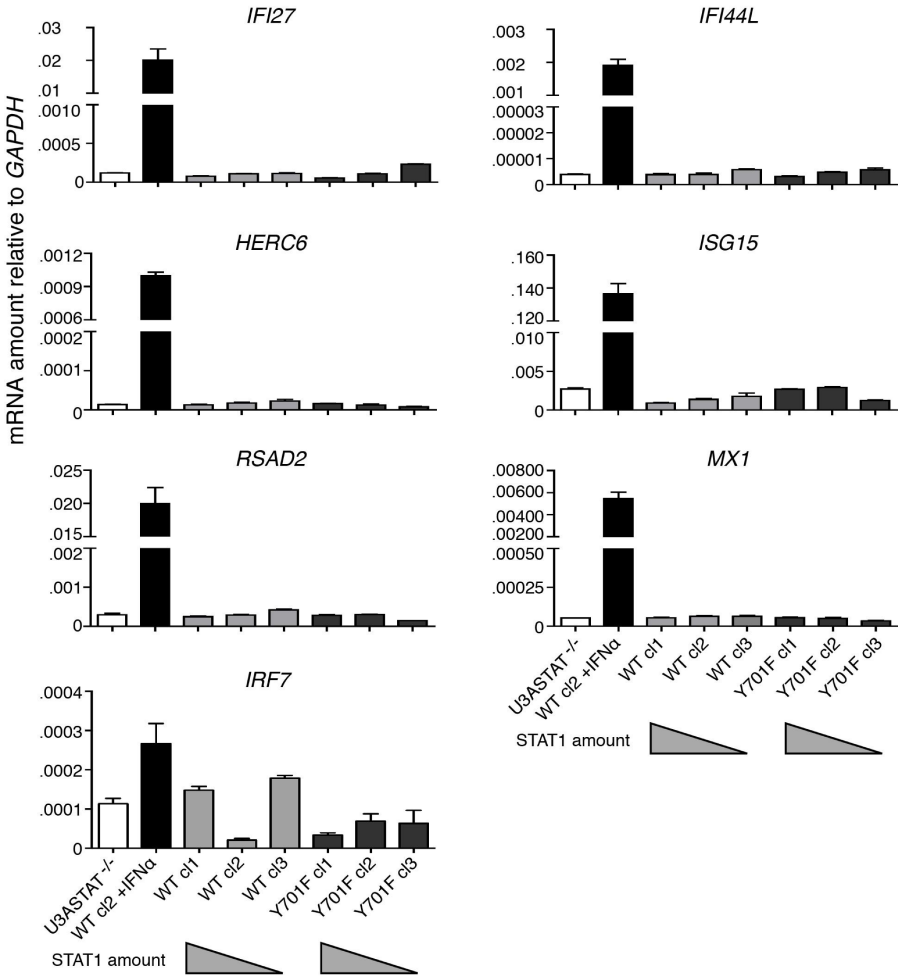


Figure C.3: mRNA expression assessed by quantitative RT-PCR of six representative ISGs. mRNA expression assessed by quantitative RT-PCR of six representative ISGs (IFI27, HERC6, RSAD2, IFI44L, ISG15, MX1), and IRF7 without IFN-α treatment not showing any upregulation in the different clones (WT cl1-3, Y701F cl1-3) compared to U3A STAT1$^{-/-}$ cells. As a positive control WT cl2 was treated with 1000 U/ml IFN-α for 8 hours (lane 2). The relative amount of STAT1 protein per clone is illustrated graphically below. Shown are mean values of 3 replicates with SEM.

C.2 SUPPLEMENTARY TABLES

Due to its overlength, the table is not printed here.

Please request it from the author or access it online at:

http://dm5migu4zj3pb.cloudfront.net/manuscripts/70000/70408/JCI70408sd1.xls

Table C.1: Four robust clusters of upregulated genes, termed early (144 genes), intermediate (31 genes), late (299 genes), and very late ISGs (20 genes).

Due to its overlength, the table is not printed here.
Please request it from the author or access it online at:
http://dm5migu4zj3pb.cloudfront.net/manuscripts/70000/70408/JCI70408sd2.xls

Table C.2: Genes that were downregulated upon pegIFN-α2b treatment.

Due to its overlength, the table is not printed here.
Please request it from the author or access it online at:
http://dm5migu4zj3pb.cloudfront.net/manuscripts/70000/70408/JCI70408sd3.xls

Table C.3: Genes that are up-regulated 144 hours post injection of PegIntron (pegIFN-α2b) or Pegasys (pegIFN-α2a).

Due to its overlength, the table is not printed here.
Please request it from the author or access it online at:
http://dm5migu4zj3pb.cloudfront.net/manuscripts/70000/70408/JCI70408sd4.xls

Table C.4: Gene ontology terms and corresponding enrichment scores and significance.

Gene	Forward primer sequence	Reverse primer sequence
GAPDH	5' GCTCCTCCTGTTCGACAGTCA 3'	5' ACCTTCCCCATGGTGTCTGA 3'
IFI44L	5' GCTGCGGGCTGCAGAT 3'	5' CTCTCTCAATTGCACCAGTTTCC 3'
RSAD2	5' CTTTGTGCTGCCCCTTGAG 3'	5' TCCATACCAGCTTCCTTAAGCAA 3'
IFI27	5' GGCAGCCTTGTGGCTACTCT 3'	5' CCCAGGATGAACTTGGTCAATC 3'
USP18	5' CTCAGTCCCGACGTGGAACT 3'	5' ATCTCTCAAGCGCCATGCA 3'
HERC6	5' CACTACCACTCCCTGGCATT 3'	5' TGTTACTTCCCCAGCCAAAV 3'
MX1	5' GTGCATTGCAGAAGGTCAGA 3'	5' TCAGGAGCCAGCTTAGGTGT 3'
ISG15	5' TCCTGCTGGTGGTGGACAA 3'	5' TTGTTATTCCTCACCAGGATGCT 3'
OAS1	5' TGATGCCCTGGGTCAGTTG 3'	5' TCGGTGCACTCCTCGATGA 3'
OAS2	5' ACAGCTGAAAGCCTTTTGGA 3'	5' AAGTTTCGCTGCAGGACTGT 3'
IRF7	5' CTTGGCTCCTGAGAGGGCAG 3'	5' CGAAGTGCTTCCAGGGCA 3'
LGALS3BP	5' GGCTGGCTGAAGAGCAACTG 3'	5' GTGGGTGCTCCTGGTTTCAT 3'

Table C.5: Primer sequences used for real-time RT-PCR analysis.

D

SUPPLEMENTARY MATERIAL TO CHAPTER 6

D.1 SUPPLEMENTARY MATERIALS

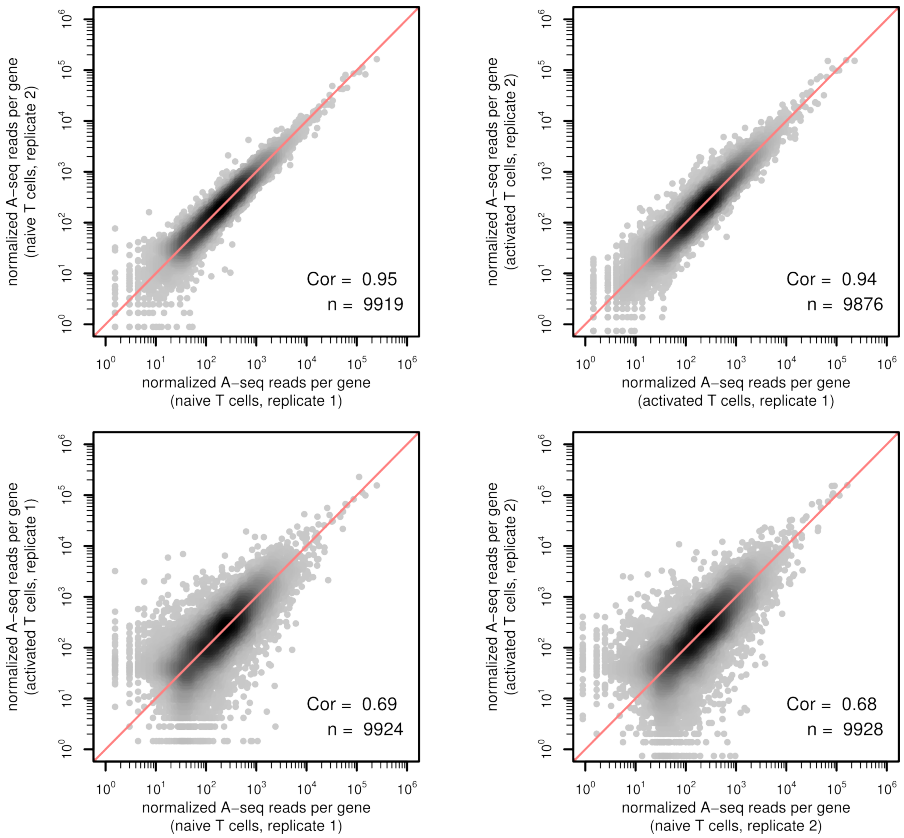


Figure D.1: Correlation of gene expression among pairs of naive and activated murine T cell samples. The number of A-seq reads that mapped to terminal exons of transcripts assigned to a particular gene was used as an estimate of the expression level of a gene.

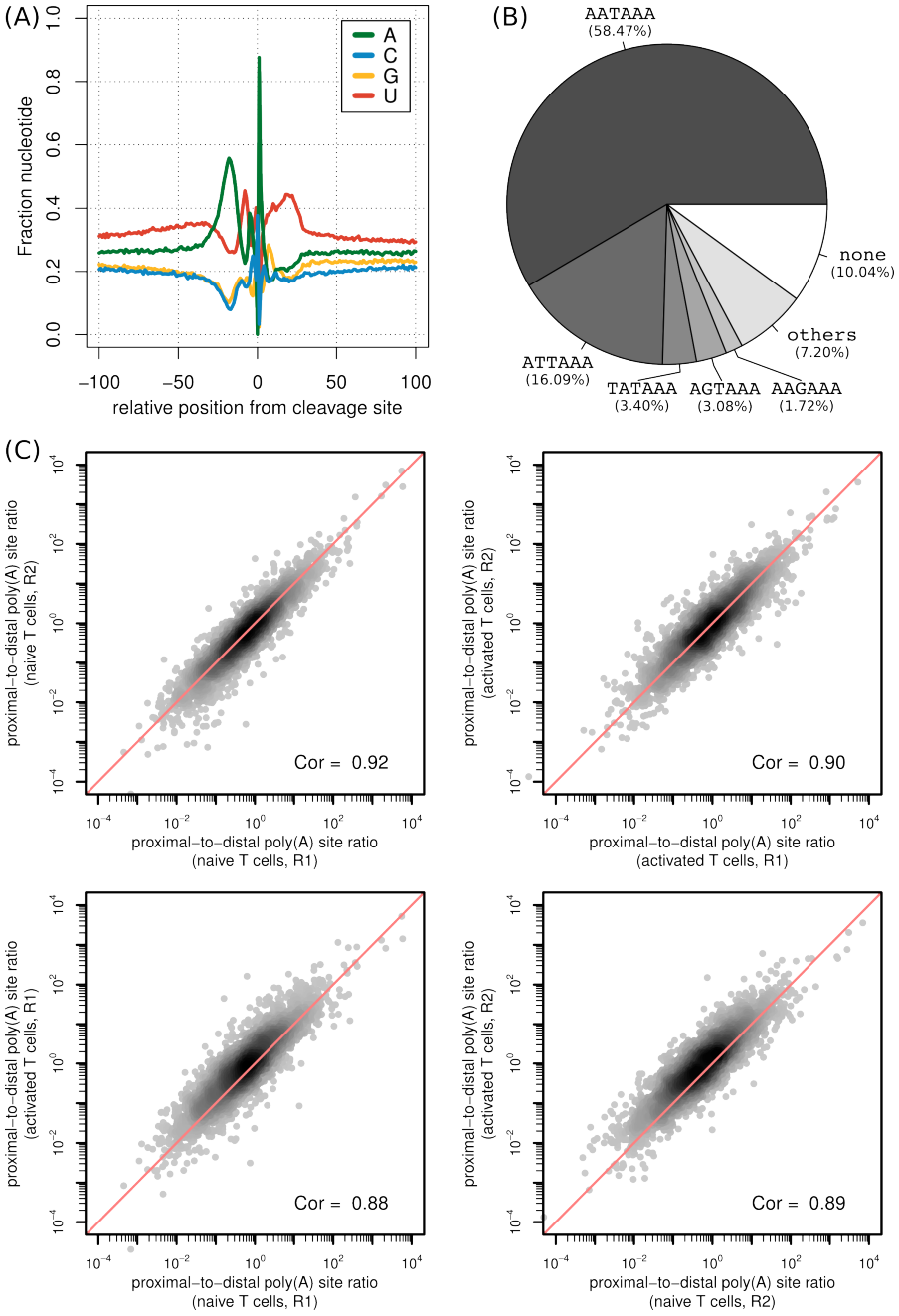


Figure D.2: Basic features of the inferred poly(A) sites in murine T cells. (A) Nucleotide composition around the inferred poly(A) sites. **(B)** Relative frequency of polyadenylation motifs detected in the 40 nucleotide region upstream of the poly(A) sites. **(C)** Comparison of the ratio of A-seq reads assigned to proximal and distal poly(A) sites between replicates of 3' end sequencing libraries obtained from naive and activated murine T cells.

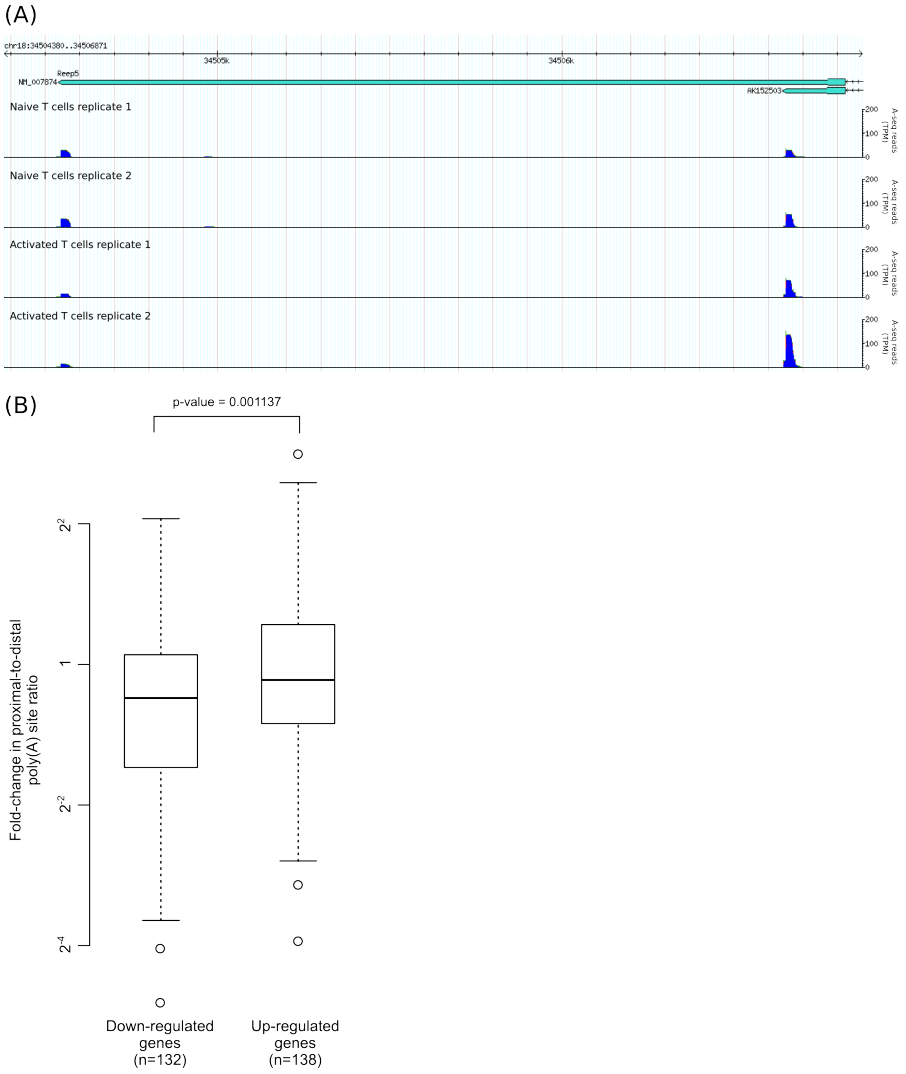


Figure D.3: Reep5 read densities and fold-change in proximal-to-distal poly(A) site ratio. (A) CLIPZ genome browser screen shot of the gene locus of Reep5. Upon T cell activation proximal poly(A) sites are used more frequently compared to the naive state. **(B)** Boxplots of the fold-change in proximal-to-distal poly(A) site ratio upon T cell activation. Downregulated genes show a more pronounced 3' UTR shortening than up-regulated genes (p-value obtained from a two-sided t-test).

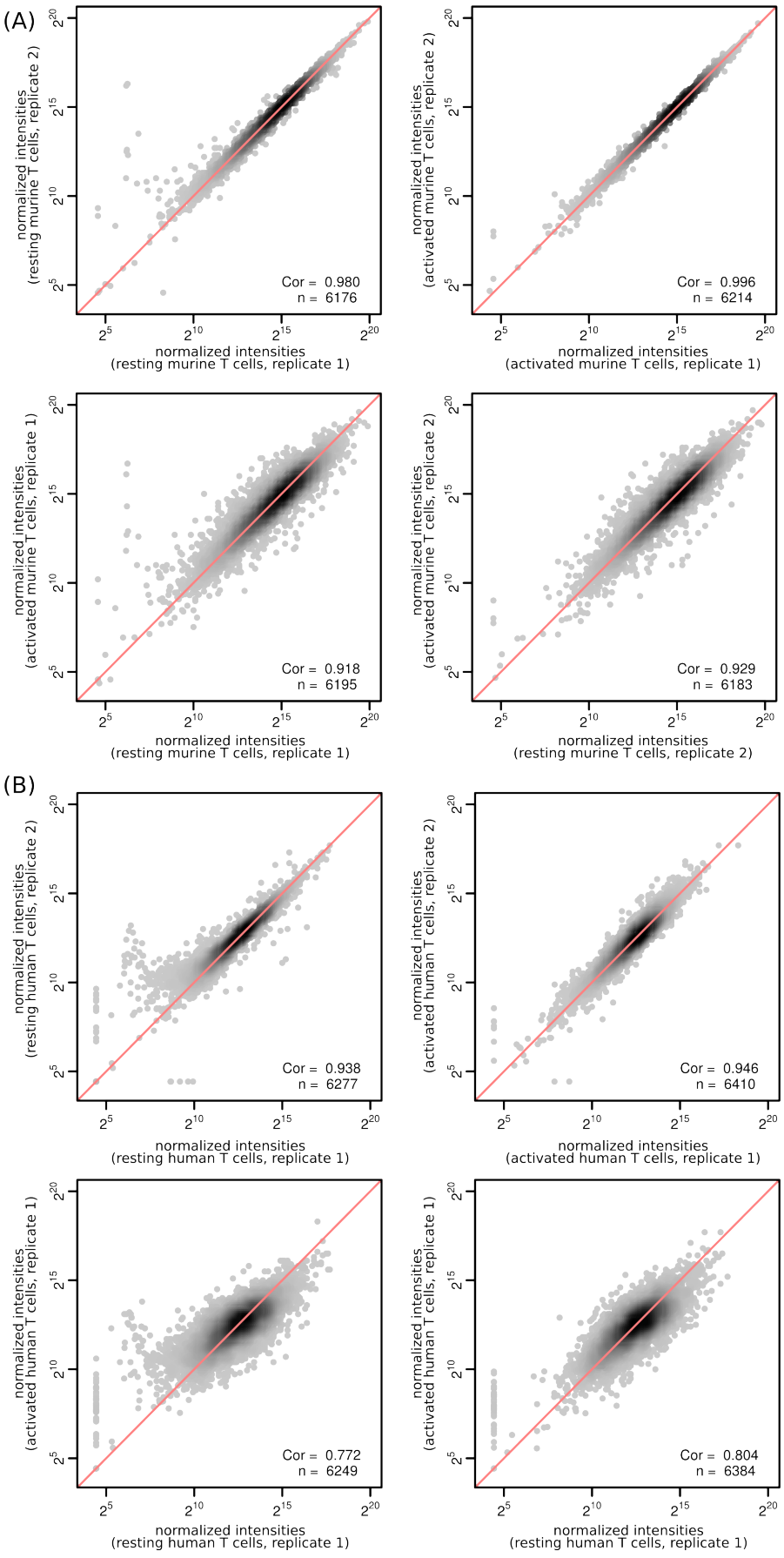


Figure D.4: Correlation of protein expression levels among pairs of naive and activated T cell samples. (A) Murine T cells. **(B)** Human T cells.

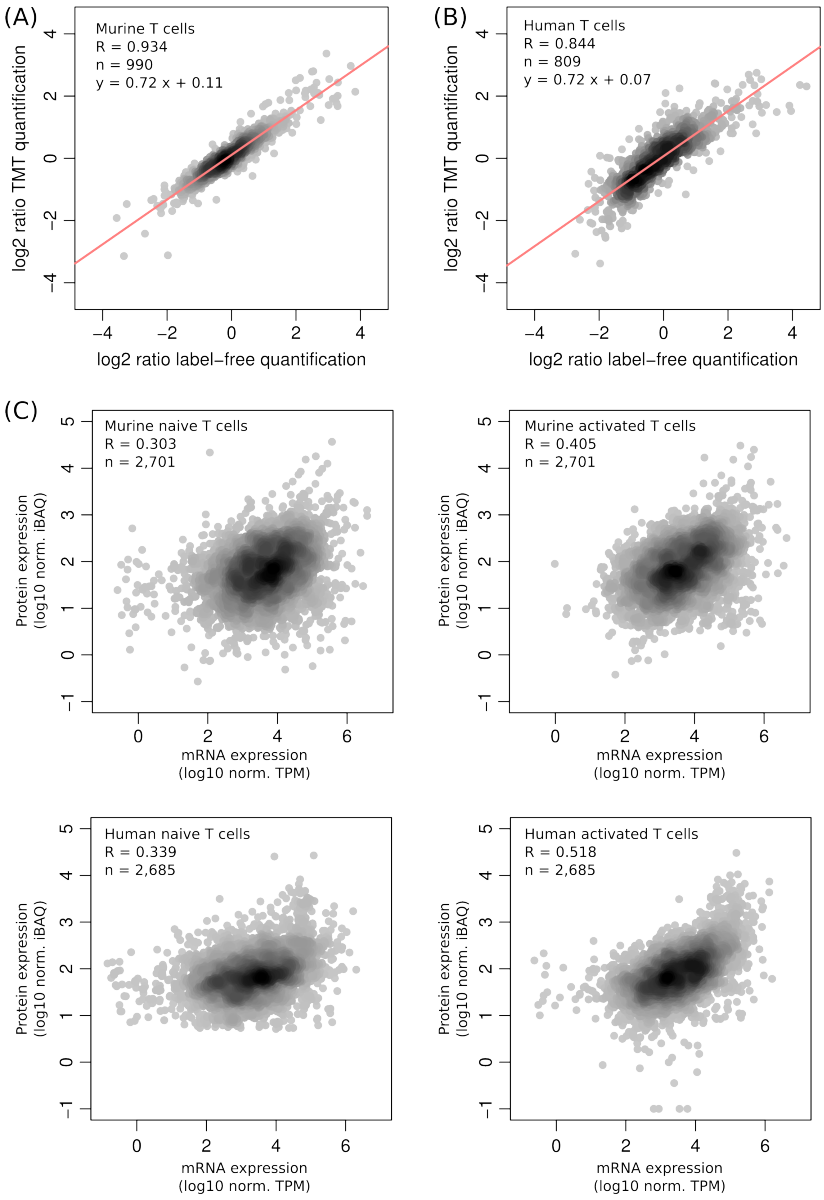


Figure D.5: Quality assessment of protein quantification. Tandem mass tagging (TMT) and label-free quantification (LFQ) protein ratios for the mouse **(A)** and human **(B)** data set. **(C)** Comparison of mRNA expression levels (A-seq, tags per million (TPM)) to protein expression levels (LFQ, iBAQ).

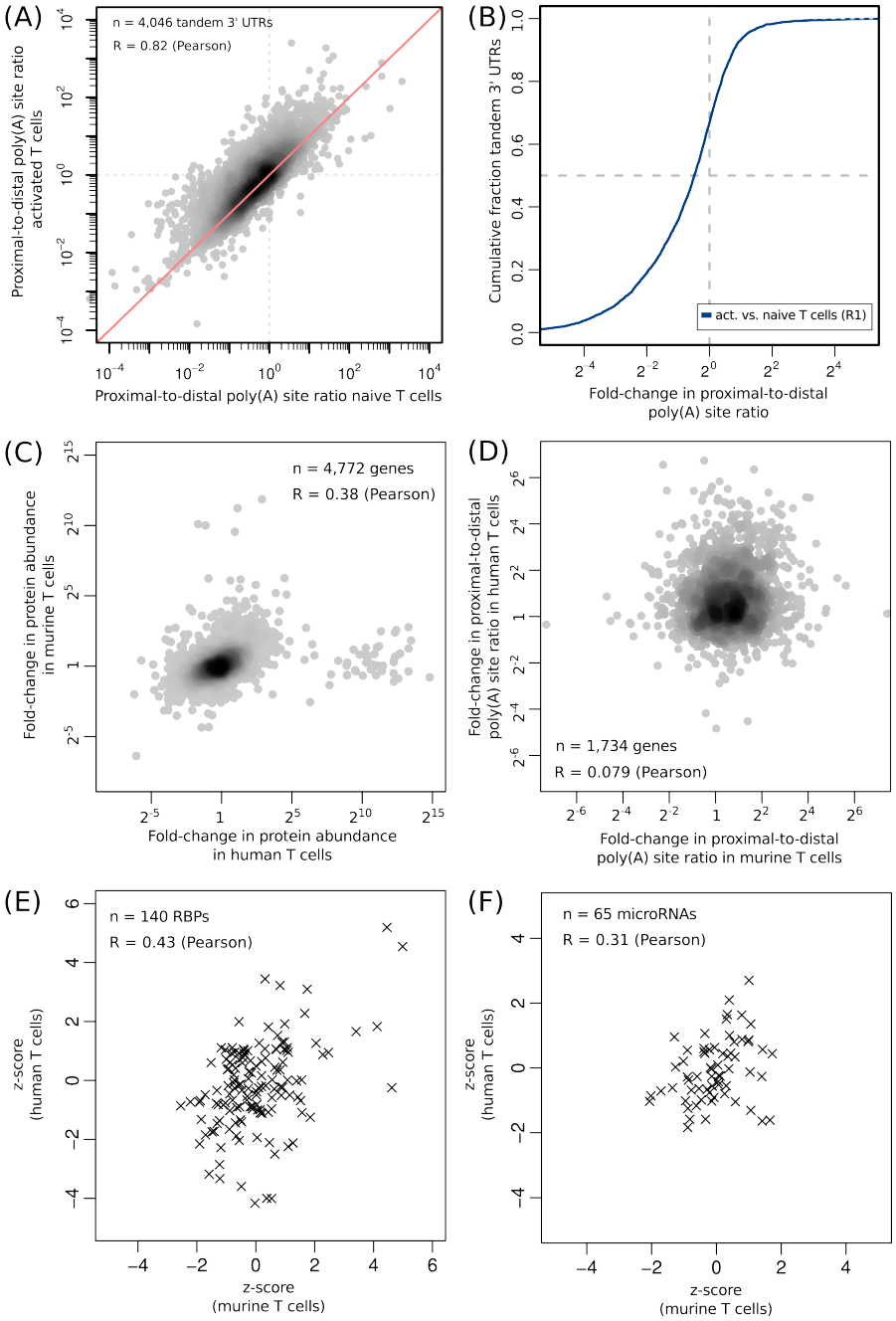


Figure D.6: Assessment of human A-seq library quality and comparison to results obtained in murine T cells. **(A)** Scatter plot comparing poly(A) site use in naive and activated human T cells for genes with tandem poly(A) sites. **(B)** Cumulative distribution function of the change in tandem poly(A) site use of data shown in A. For comparison to murine T cells see Figure 6.1D. **(C)** Comparison of changes in protein levels upon T cell activation in murine and human T cells. **(D)** Comparison of changes in poly(A) site use upon T cell activation in murine and human T cells. **(E-F)** Comparison of z-values quantifying the loss of binding sites for orthologous RBPs and miRNAs in murine and human 3' UTRs.

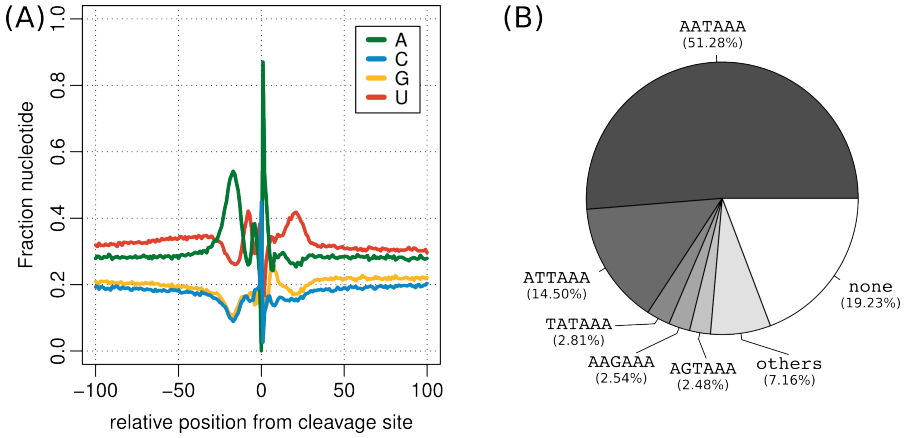


Figure D.7: Basic features of the inferred poly(A) sites in human T cells. (A) Nucleotide composition around the inferred poly(A) sites. **(B)** Relative frequency of polyadenylation motifs detected in the 40 nucleotide region upstream of the poly(A) site.

Please request the table from the author or access it online at: http://www.nature.com/ncomms/2014/141121/ncomms6465/extref/ncomms6465-s1.pdf

Table D.1: Summary statistics for Aseq samples of murine naive and activated T cells.

Please request the table from the author or access it online at: http://www.nature.com/ncomms/2014/141121/ncomms6465/extref/ncomms6465-s1.pdf

Table D.2: Results of GO term enrichment analysis on differentially expressed genes.

Please request the table from the author or access it online at: http://www.nature.com/ncomms/2014/141121/ncomms6465/extref/ncomms6465-s1.pdf

Table D.3: Number of genes with inferred tandem poly(A) sites in murine naive and activated T cells.

Please request the table from the author or access it online at: http://www.nature.com/ncomms/2014/141121/ncomms6465/extref/ncomms6465-s1.pdf

Table D.4: Genes that show significant changes in proximal-to-distal poly(A) site usage in murine naive and activated T cells. P/D = proximal-to-distal

Please request the table from the author or access it online at:
http://www.nature.com/ncomms/2014/141121/ncomms6465/extref/ncomms6465-s1.pdf

Table D.5: Evaluation of miRNA target sites in the common and alternative parts of the 3' UTRs of murine genes with tandem poly(A) sites. TS = target sites, FC = fold-change. The number in parenthesis is the number of mature miRNAs that share the seed sequence with the indicated miRNA.

Please request the table from the author or access it online at:
http://www.nature.com/ncomms/2014/141121/ncomms6465/extref/ncomms6465-s1.pdf

Table D.6: Comparison of changes in protein expression levels measured with quantitative Western blots and mass spectrometry for a set of randomly chosen genes. Western blot signals were quantified with LICOR-software.

Please request the table from the author or access it online at:
http://www.nature.com/ncomms/2014/141121/ncomms6465/extref/ncomms6465-s1.pdf

Table D.7: Summary statistics for Aseq samples of human naive and activated T cells.

Please request the table from the author or access it online at:
http://www.nature.com/ncomms/2014/141121/ncomms6465/extref/ncomms6465-s1.pdf

Table D.8: Number of genes with inferred tandem poly(A) sites in human naive and activated T cells.

Please request the supplementary dataset from the author or access it online at:
http://www.nature.com/ncomms/2014/141121/ncomms6465/extref/ncomms6465-s2.xls

Table D.9: mRNAs that are predicted to be regulated by regulators whose impact on the transcriptome is most affected by the systematic change in poly(A) site use.

Please request the supplementary dataset from the author or access it online at:
http://www.nature.com/ncomms/2014/141121/ncomms6465/extref/ncomms6465-s3.xls

Table D.10: Protein level changes upon activation of mouse T cells

Please request the supplementary dataset from the author or access it online at:
http://www.nature.com/ncomms/2014/141121/ncomms6465/extref/ncomms6465-s4.xls

Table D.11: Protein level changes upon activation of human T cells

E

SUPPLEMENTARY MATERIAL TO CHAPTER 7

E.1 3' END SEQUENCING PROTOCOLS

E.1.1 *2P-Seq*

In the 2P-Seq protocol, reverse transcription is accomplished by an anchored oligo(dT) primer. The products of reverse transcription and PCR amplification are expected to have 20 As preceding the 3' adapter. Libraries are sequenced in anti-sense direction with a custom primer. Reads should be reverse complemented [103, 485].

E.1.2 *3'-Seq*

In the 3'-Seq protocol of Mayr and colleagues, reverse transcription is accomplished by an anchored oligo(dT) primer. The products of reverse transcription and PCR amplification are expected to have 17 As preceding the 3' adapter. Libraries are sequenced in sense direction requiring removal of the 3' adapter sequence and preceding As to pinpoint the 3' end [500].

E.1.3 *3P-Seq*

In the 3P-seq protocol, a biotinylated adapter is ligated to the end of the poly(A) tail via splint-ligation. After partial digestion, poly(A) regions are captured with streptavidin and reverse transcription is carried out only with dTTP. Most of the poly(A) tail is then removed through RNase H digestion. Adapter ligation, reverse transcription and PCR amplification follow before the library is sequenced in anti-sense direction. Consequently, pinpointing the 3' end requires the reads to be reverse complemented [472, 602].

E.1.4 *3'READS*

3' region extraction and deep sequencing (3'READS) is a protocol that utilizes a special primer (45 thymidines followed by 5 uridines) to capture poly(A) containing RNA fragments. RNase H digestion releases transcripts 3' ends from the most of the poly(A) tail. Subsequently, the fragments are subjected to adapter ligation, reverse transcription, and PCR amplification before they are sequenced in anti-sense direction. The cleavage site is inferred as the first non-A of the 3' end of the read's reverse complement [484, 504].

E.1.5 *A-seq*

In the A-seq protocol, reverse transcription is accomplished by an anchored oligo-dT primer. The products of reverse transcription and PCR amplification are expected to have six As preceding the 3' adapter. Libraries are sequenced in sense direction requiring removal of the 3' adapter sequence and preceding As to pinpoint the 3' end [82].

E.1.6 *A-seq (version 2)*

The second version of the A-seq protocol has the following changes: (1) The steps of the protocol are conducted such that the generation of adapter dimers is minimized. (2) Libraries are sequenced in anti-sense direction and the mRNA cleavage site is inferred as the first nucleotide after a stretch of 4 random nucleotides and 3 Ts [429].

E.1.7 *DRS*

In the direct RNA sequencing (DRS) protocol, 3' ends of transcripts are hybridized to poly(dT)-coated flow cell surfaces where antisense strand synthesis is initiated. This has the advantage that no prior reverse transcription or cDNA amplification is needed [93, 486, 603, 604].

E.1.8 *PAS-seq*

In the PAS-Seq protocol, reverse transcription is accomplished with an anchored oligo-dT primer. The products of reverse transcription and PCR amplification are expected to have 20 As preceding the 3' adapter. Libraries are sequenced in anti-sense direction with a custom primer requiring the reverse complement of the reads to pinpoint the 3' end [479].

E.1.9 *PolyA-seq*

Library preparation for the PolyA-seq protocol includes the following steps: (1) Reverse transcription, primed with an oligo-dT sequence, (2) second strand synthesis with random hexamers linked to a second PCR primer, and (3) PCR amplification. Sequencing is accomplished in anti-sense orientation with a primer ending in 10 Ts and the resulting reads need to be reverse complemented to pinpoint the pre-mRNA cleavage site [98, 605].

E.1.10 *SAPAS*

In the SAPAS protocol, reverse transcription is accomplished by an anchored oligo-dT primer. The products of reverse transcription and PCR amplification are expected to have the sequence AAAAAAGAAAAAAGAAAAAA

preceding the 3' adapter. Libraries are sequenced in anti-sense direction with a regular primer requiring to trim 20 nucleotides from the 5' end of reads and to reverse complement reads to pinpoint the 3' end [481, 606].

E.2 SUPPLEMENTARY FIGURES

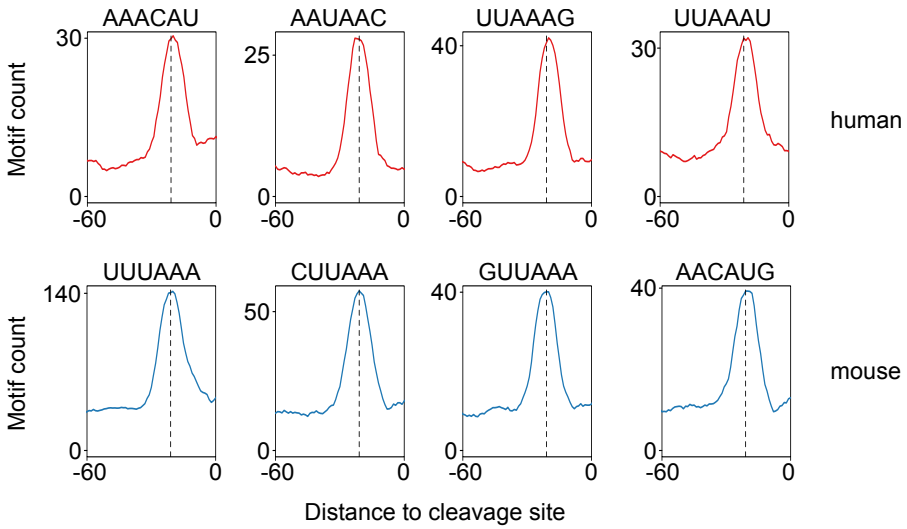


Figure E.1: Frequency profiles of the poly(A) signals that have been identified only in human (red) or mouse (blue).

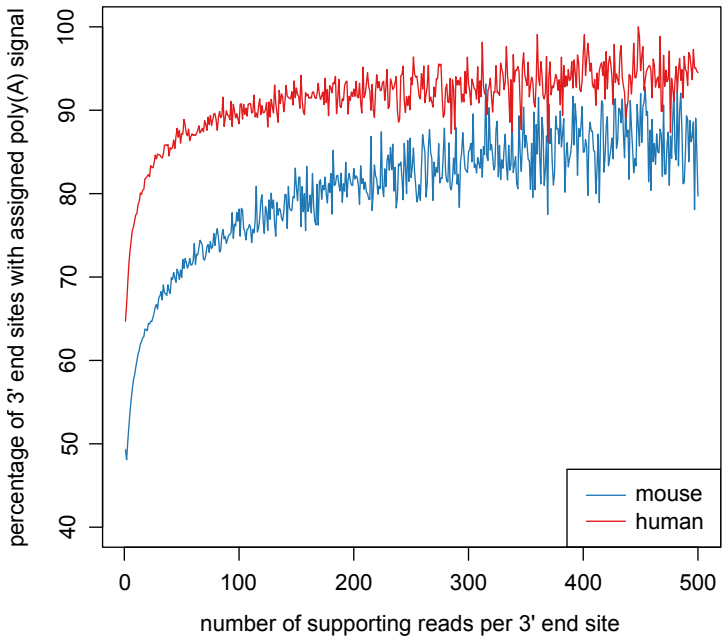


Figure E.2: Fraction of the putative 3' end sites with an assigned poly(A) signal in their upstream region (60 to 10 nucleotides upstream) as a function of the number of supporting reads per site (summed reads over all considered samples).

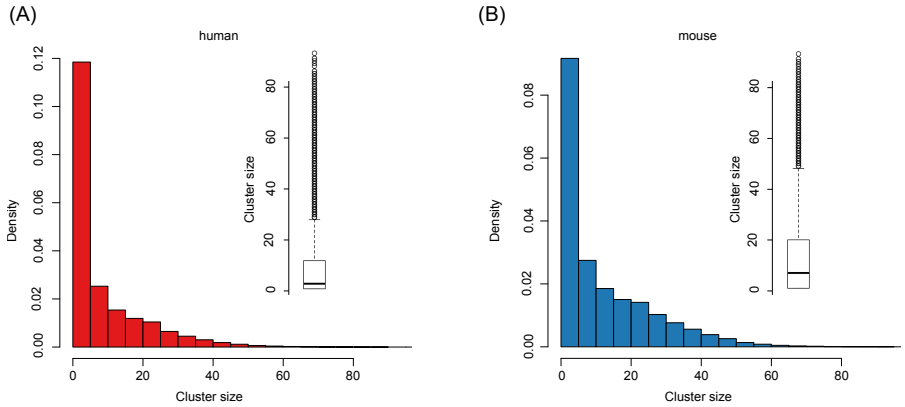


Figure E.3: Distribution of cluster sizes (A) human catalog (B) mouse catalog. The large majority of clusters has a short span (less than 20 nt) in both human and mouse.

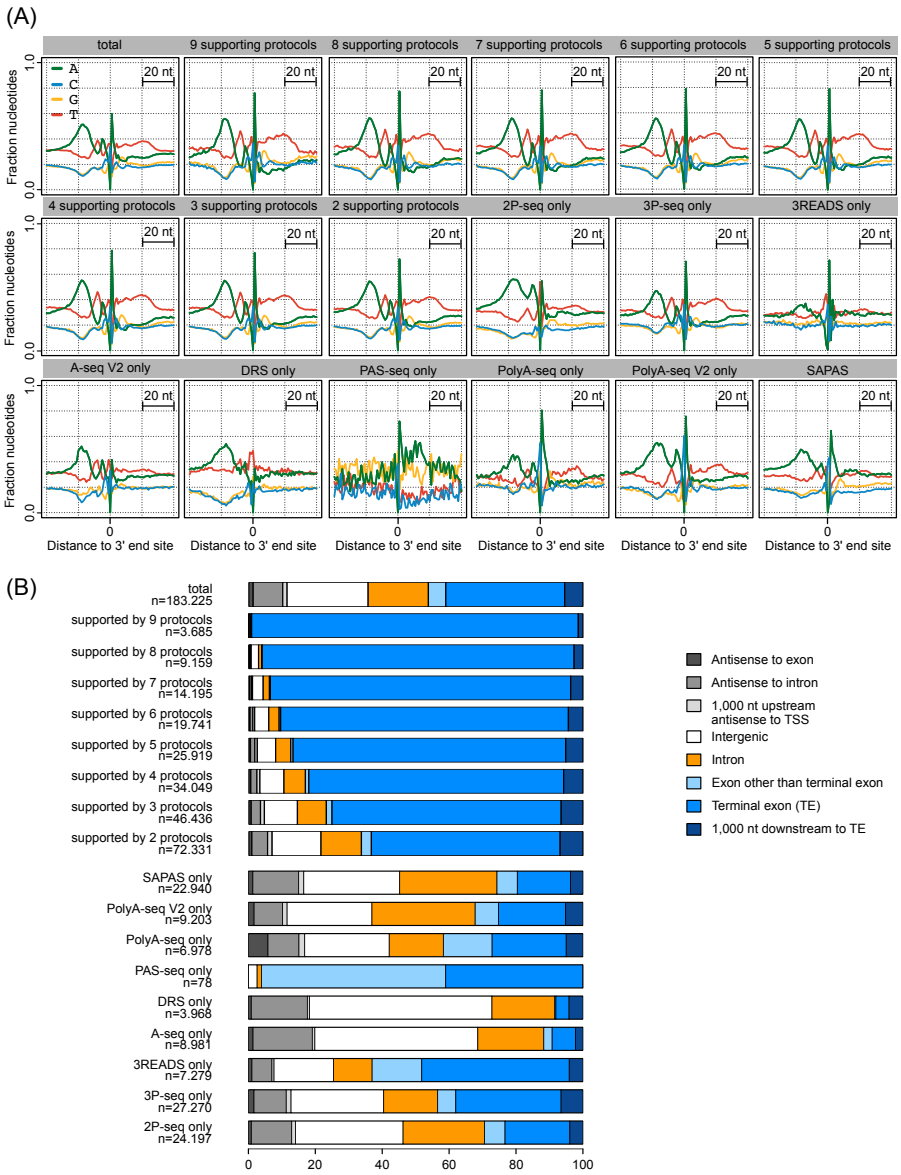


Figure E.4: Characteristics of mouse poly(A) clusters. (A) Nucleotide composition around cleavage sites supported by the indicated number of protocols or the name of the protocol for clusters that had a single protocol support. **(B)** Annotation of clusters supported by various types of protocols (n - number of poly(A) clusters in the indicated category).

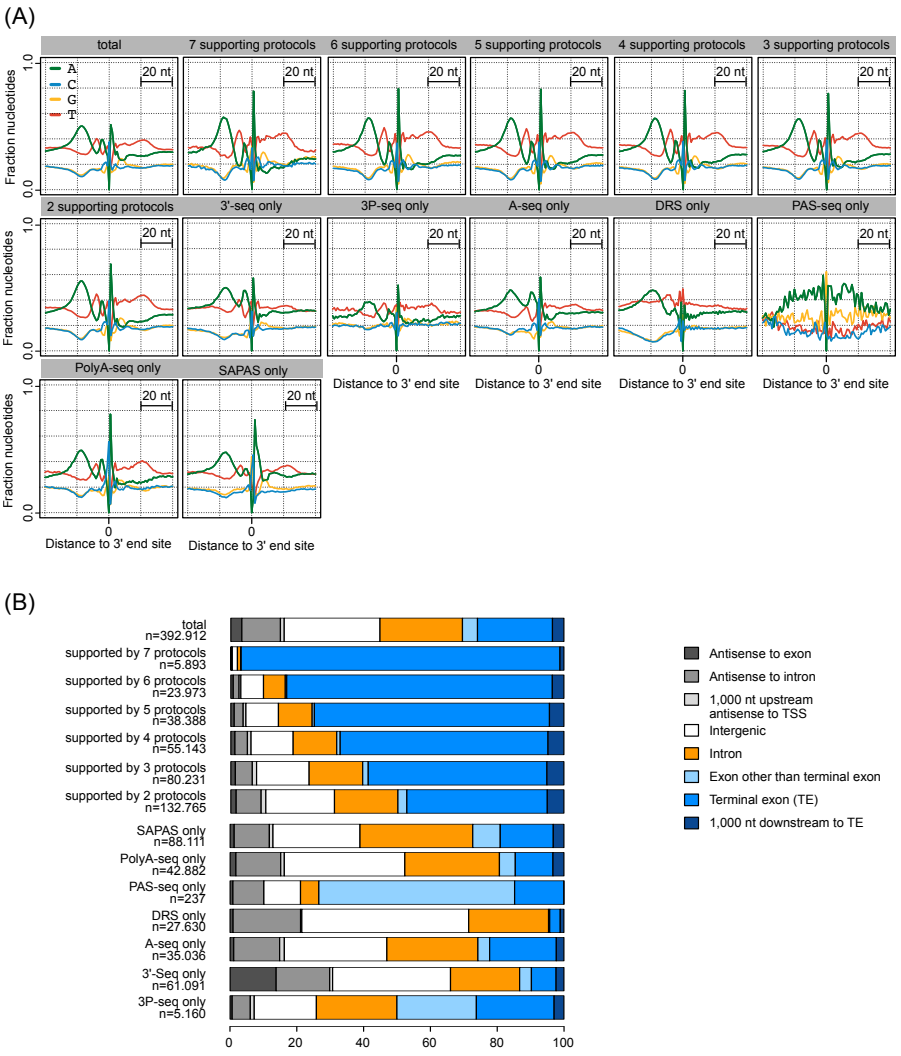


Figure E.5: Characteristics of human poly(A) clusters. (A) Nucleotide composition around cleavage sites supported by the indicated number of protocols or the name of the protocol for clusters that had a single protocol support. **(B)** Annotation of clusters supported by various types of protocols (n - number of poly(A) clusters in the indicated category).

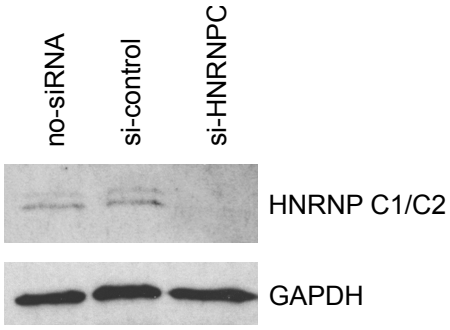


Figure E.6: Western blot showing the expression levels of HNRNP C1/C2 and GAPDH in cells that were either untreated, or treated with either a control siRNA or with si-HNRNPC (50 picomoles siRNA per well of a 6-well plate).

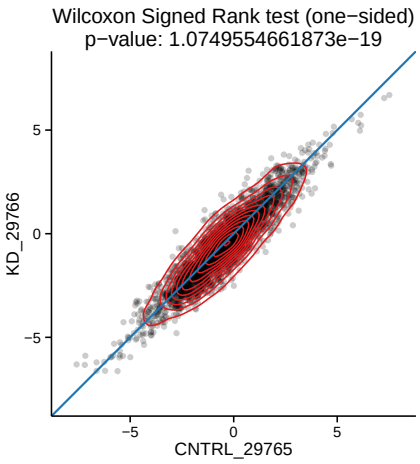


Figure E.7: Contour plot of the proximal-to-distal poly(A) site usage ratios in si-HNRNPC transfected versus si-Control transfected HEK 293 cells in replicate 1. For each plot only exons having exactly two expressed poly(A) sites were considered (2607 exons in total). The proximal-to-distal ratio is significantly higher in cells treated with the control siRNA indicating that on average 3' UTRs tend to be elongated, rather than shortened, upon knockdown of HNRNPC.

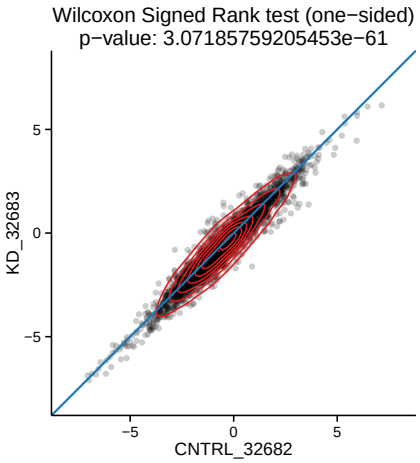


Figure E.8: Contour plot of the proximal-to-distal poly(A) site usage ratios in si-HNRNPC transfected versus si-Control transfected HEK 293 cells in replicate 2. For each plot only exons having exactly two expressed poly(A) sites were considered (2607 exons in total). The proximal-to-distal ratio is significantly higher in cells treated with the control siRNA indicating that on average 3' UTRs tend to be elongated, rather than shortened, upon knockdown of HNRNPC.

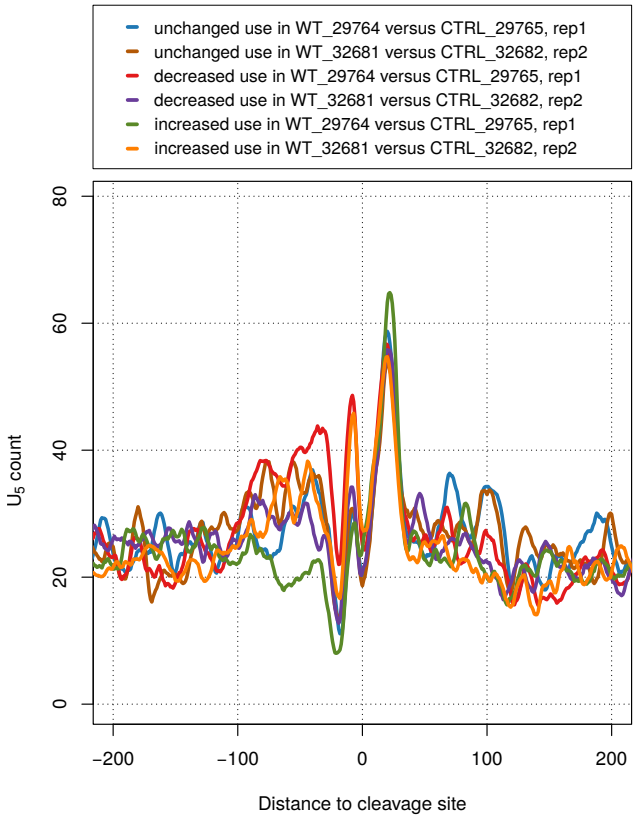


Figure E.9: Smoothened ($\pm$5 nt) density of non-overlapping $(U)_5$ tracts in the vicinity of sites with a consistent behavior (increased, unchanged, decreased use) in untransfected (wild type, WT) compared to the si-Control transfected (CTRL) HEK 293 cells.

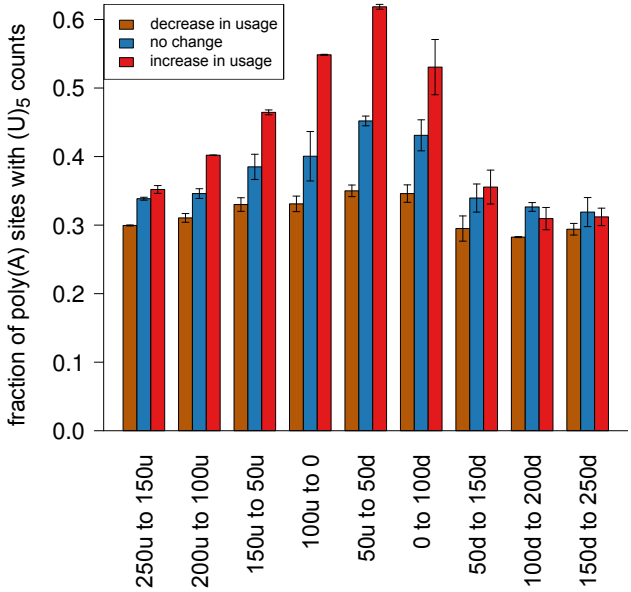


Figure E.10: Relationship between the $(U)_5$ content around poly(A) sites and their behavior upon HNRNPC knock-down. 1000 poly(A) sites that increased most, decreased most or changed least (and reproducibly, between the two replicate experiments) in usage upon HNRNPC knock-down were extracted, and the fractions of each of these types of sites that had at least one occurrence of the $(U)_5$ motif at the indicated distance from the poly(A) site were calculated. 'u' and 'd' indicate upstream and downstream of poly(A) sites and the numbers indicate the boundaries (in nt) of the windows relative to poly(A) sites.

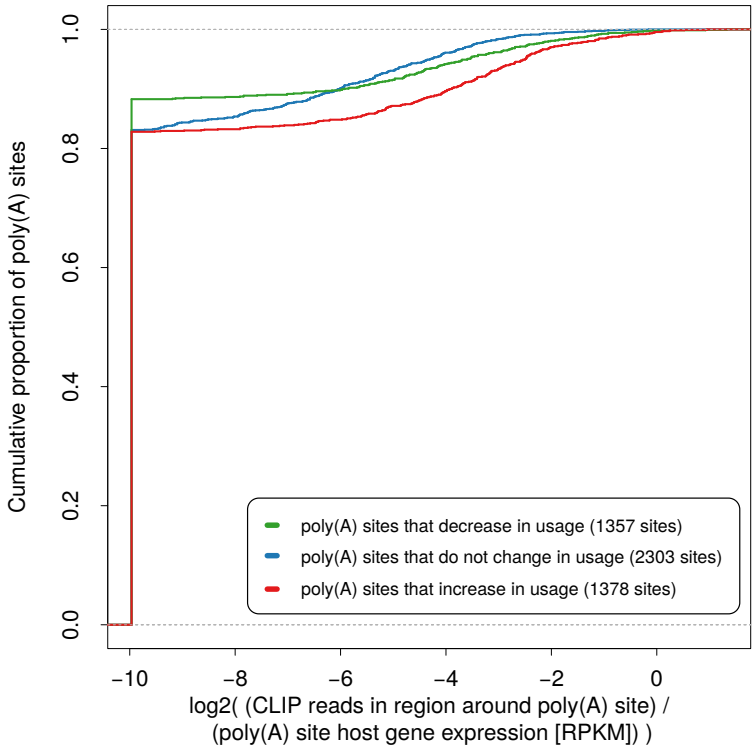


Figure E.11: Number of HNRNPC CLIP reads that intersect with a region of $\pm$50 nucleotides around poly(A) sites belonging to different categories (consistently decreased/unchanged/increased poly(A) site usage upon HNRNPC knock-down). The number of HNRNPC CLIP reads was normalized by the expression ([RPKM]) of each poly(A) site's host gene. Poly(A) sites that increase in usage have a significantly higher CLIP read support compared to poly(A) sites that do not change in usage upon HNRNPC knock-down (p-value $<$0.0007, two-sided Kolmogorov-Smirnov test).

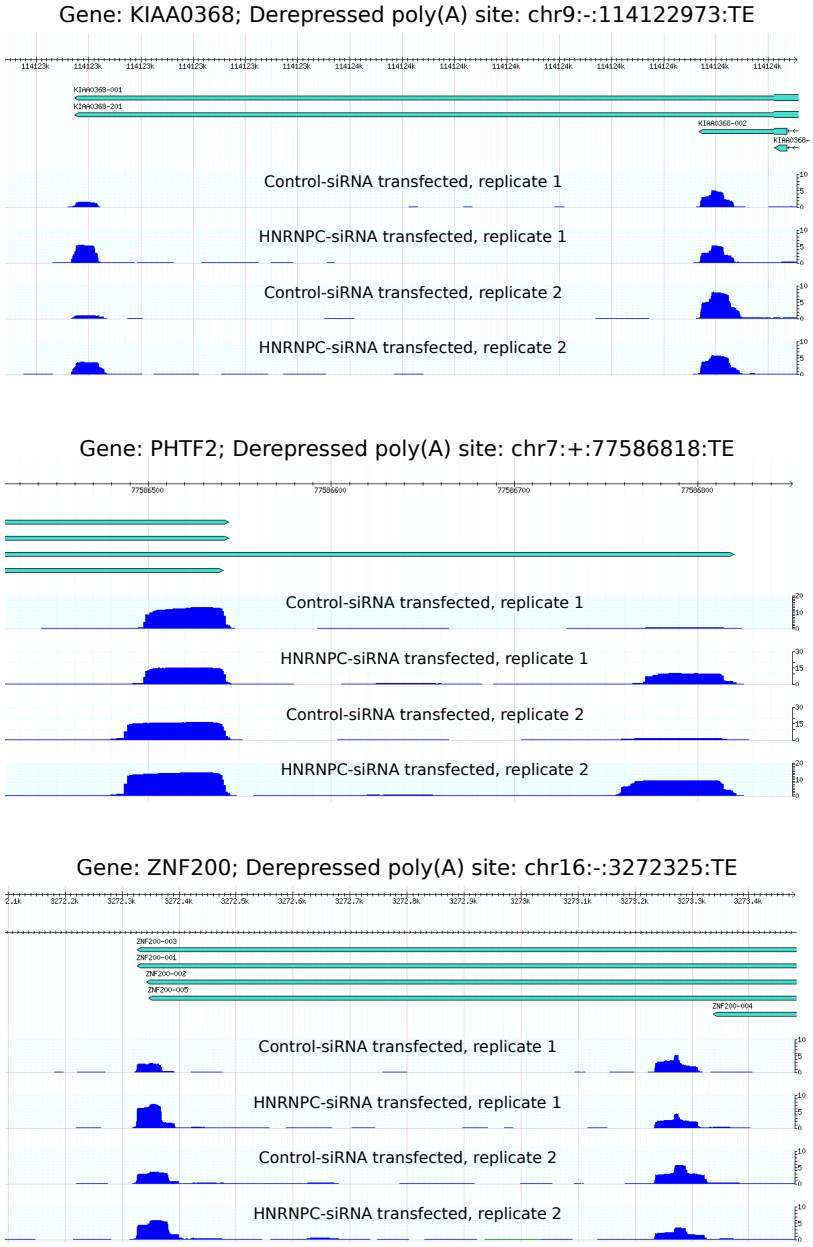


Figure E.12: Browser shots of A-Seq read densities within 3' UTRs with **distal poly(A) sites that are derepressed upon knock-down of HNRNPC**. The y-axis shows library size normalized read counts per nucleotide.

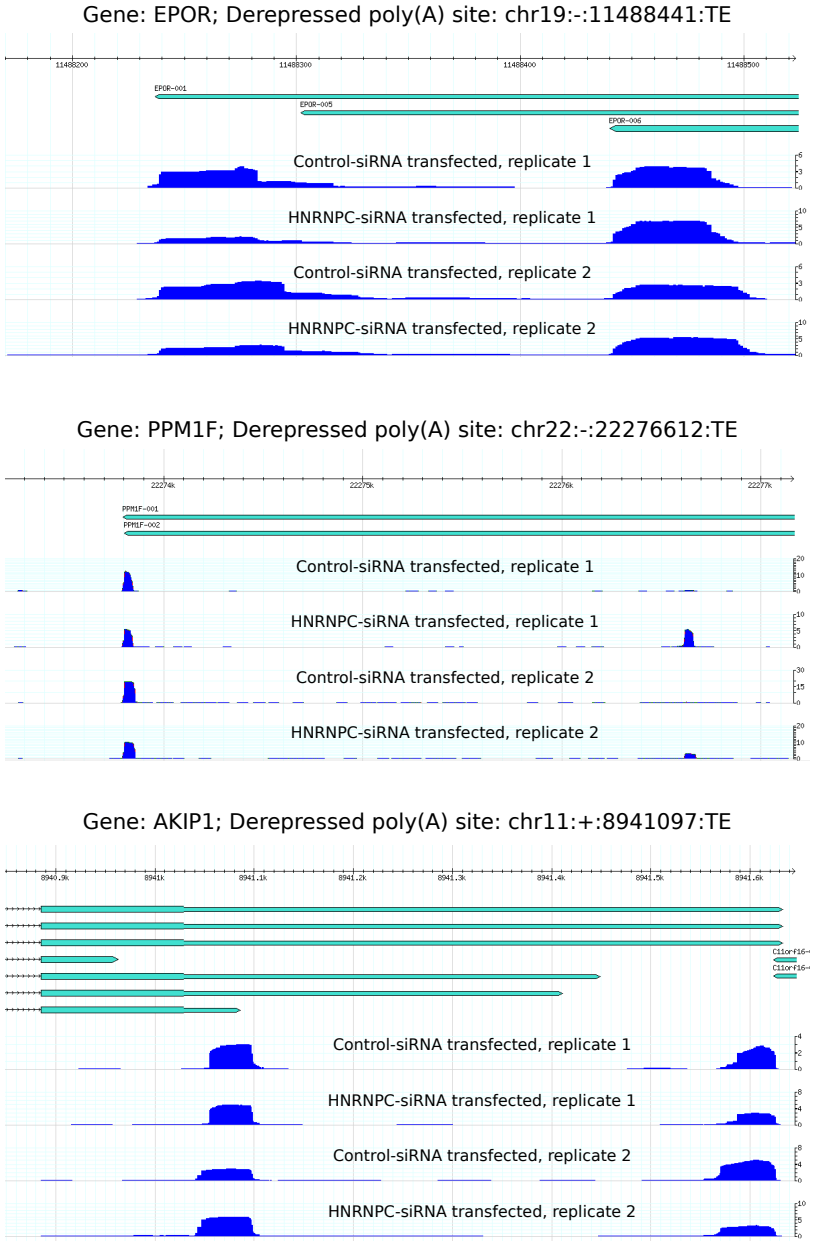


Figure E.13: Browser shots of A-Seq read densities within 3' UTRs with **proximal poly(A) sites that are derepressed upon knock-down of HNRNPC**. The y-axis shows library size normalized read counts per nucleotide.

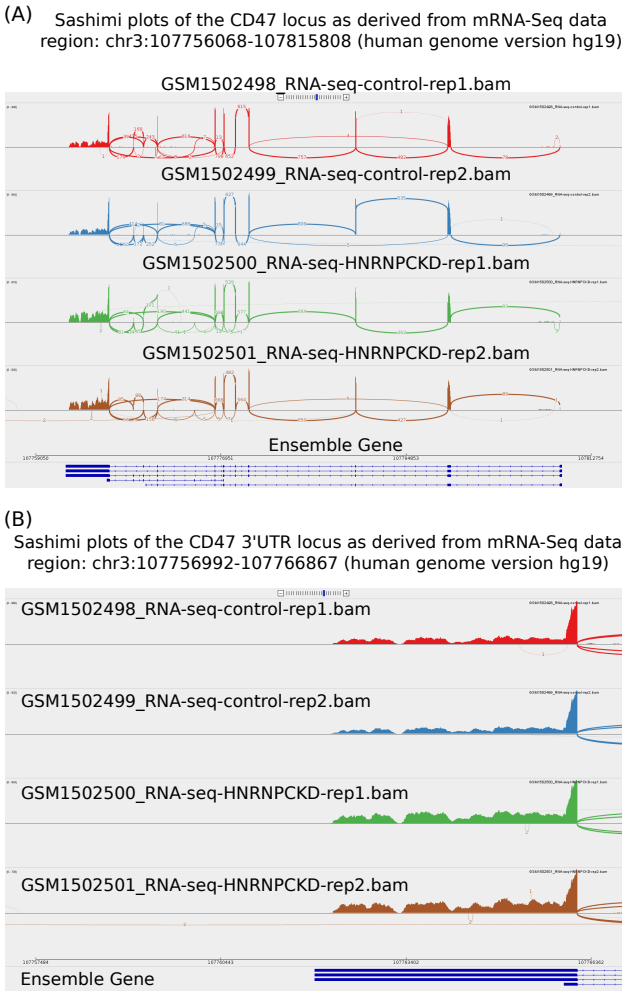


Figure E.14: "Sashimi" plots of **(A)** the CD47 gene locus and **(B)** the CD47 3' UTR locus. Plots were constructed from previously published (see [508]) mRNA-Seq data (2 replicates of 2 experiments) obtained from HEK 293 cells that have been transfected with si-Control or si-HNRNPC, respectively. After adaptor removal, paired-end reads were mapped applying the STAR aligner with default settings [534]. The mappings were visualized (Sashimi plots) using the Integrative Genomics Viewer (IGV) software [607].

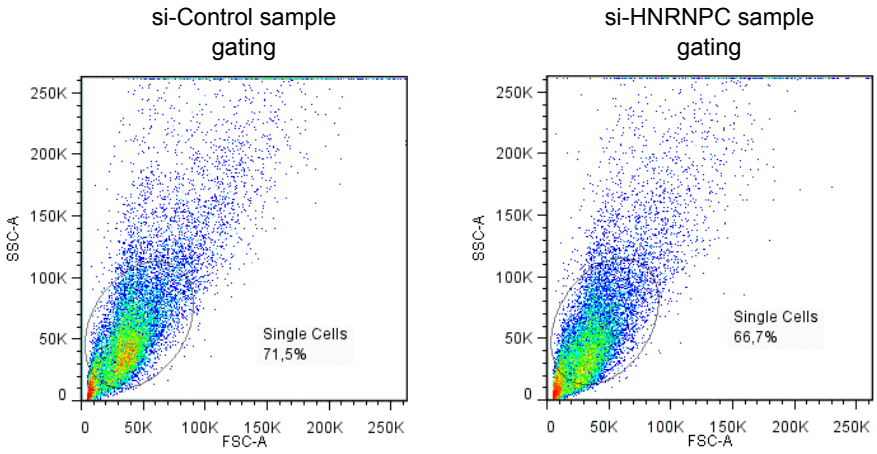


Figure E.15: For indirect immunophenotyping of membrane CD47 levels in HEK 293 cells that were either treated with a control siRNA (left panel) or with si-HNRNPC (right panel) a minimum of 10000 gated events was considered. The gate is indicated.

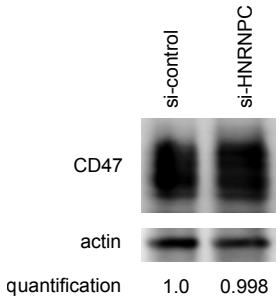


Figure E.16: Western blots of CD47 and Actin proteins in cells treated with either a control siRNA or with si-HNRNPC for 72 hrs. Signals were quantified with the ImageJ software and relative CD47 levels are reported with respect to Actin and control siRNA (= 1.0).

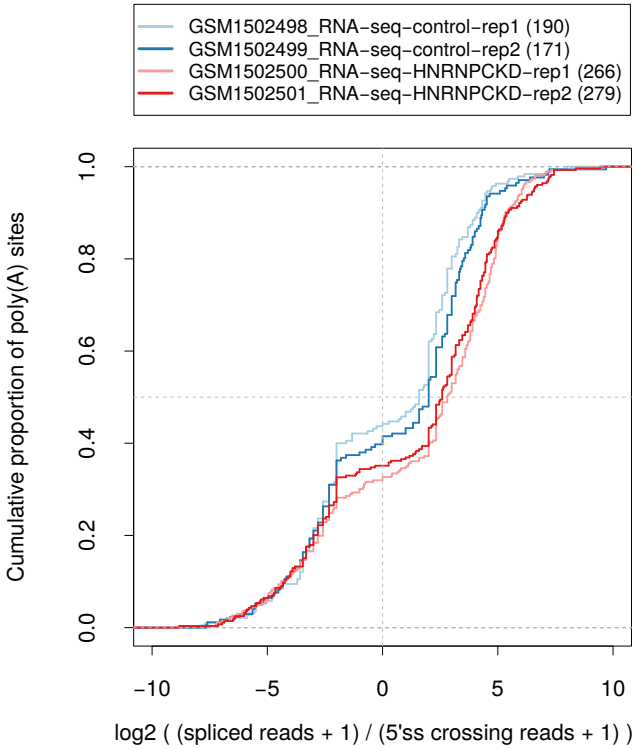


Figure E.17: Cumulative distribution functions of the log2 ratios of spliced reads to reads that map beyond the 5' splice site (5'ss) of the closest, upstream located exon of each consistenly derepressed, intronic poly(A) site. Intronic poly(A) sites are associated predominantly with the emergence of new exons relative to the extension of internal exons, in both si-Control and si-HNRNPC transfected cells. The HNRNPC knock-down causes a further significant shift towards novel terminal exons created by splicing rather than by internal exon extension (replicate 1 p-value: 4.0e-06, replicate 2 p-value: 8.6e-03, two-sided Mann-Whitney U test). The numbers shown in the legend (written in brackets) indicate the number of intronic poly(A) sites that were used to construct this plot (for more details, see the Methods section).

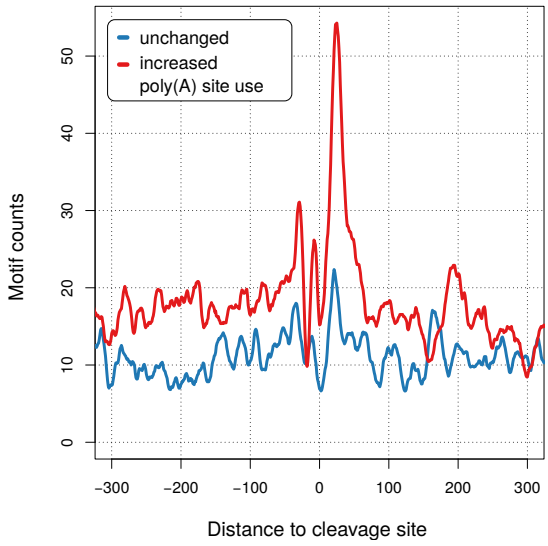


Figure E.18: Smoothened ($\pm$5 nt) density of non-overlapping $(U)_5$ tracts in the vicinity of intronic poly(A) sites with a consistent behavior (increased or unchanged use) in the two HNRNPC knock-down A-seq2 experiments.

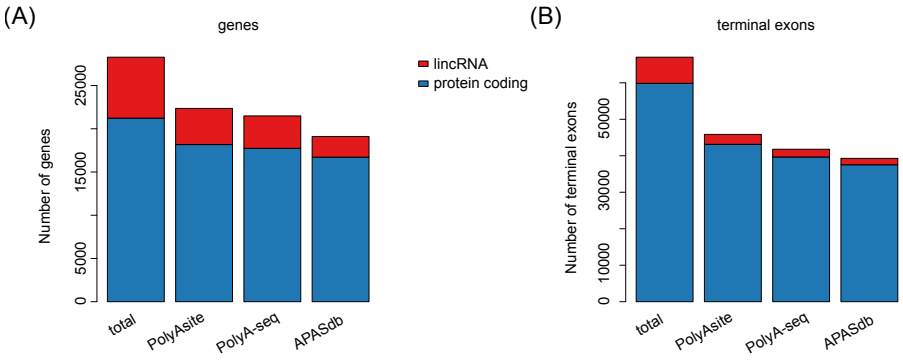


Figure E.19: Number of annotated features (based on the UCSC Basic Table of the GENCODE v19 human (hg19) annotation) that are covered by sites from different atlases. (A) Coverage of genes by sites from PolyAsite (present manuscript), PolyA-seq [98] and APASdb [481]. A gene was considered covered if the genomic position of at least one poly(A) site was within the genomic range of the gene. **(B)** Same as **(A)** but for the terminal exons from the annotation.

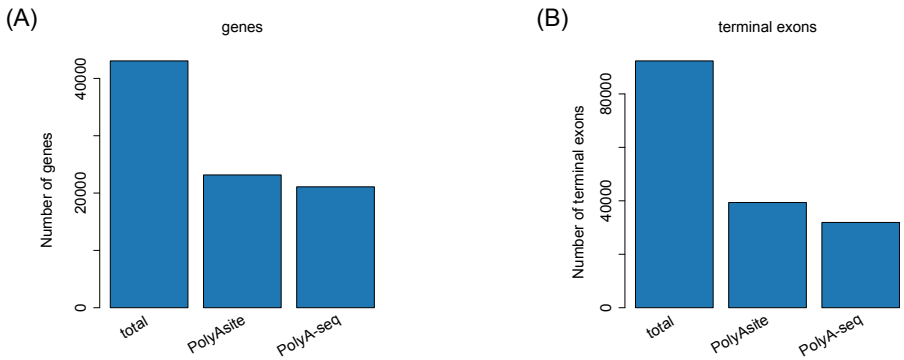


Figure E.20: Number of annotated features (based on the ENSEMBL mouse (mm10) annotation from UCSC) that are covered by sites from different atlases. **(A)** Coverage of genes by sites from PolyAsite and PolyA-seq [98]. A gene was considered to be covered if the genomic position of at least one poly(A) site was within the genomic range of the gene. **(B)** Same as **(A)** but for the terminal exons from the annotation.

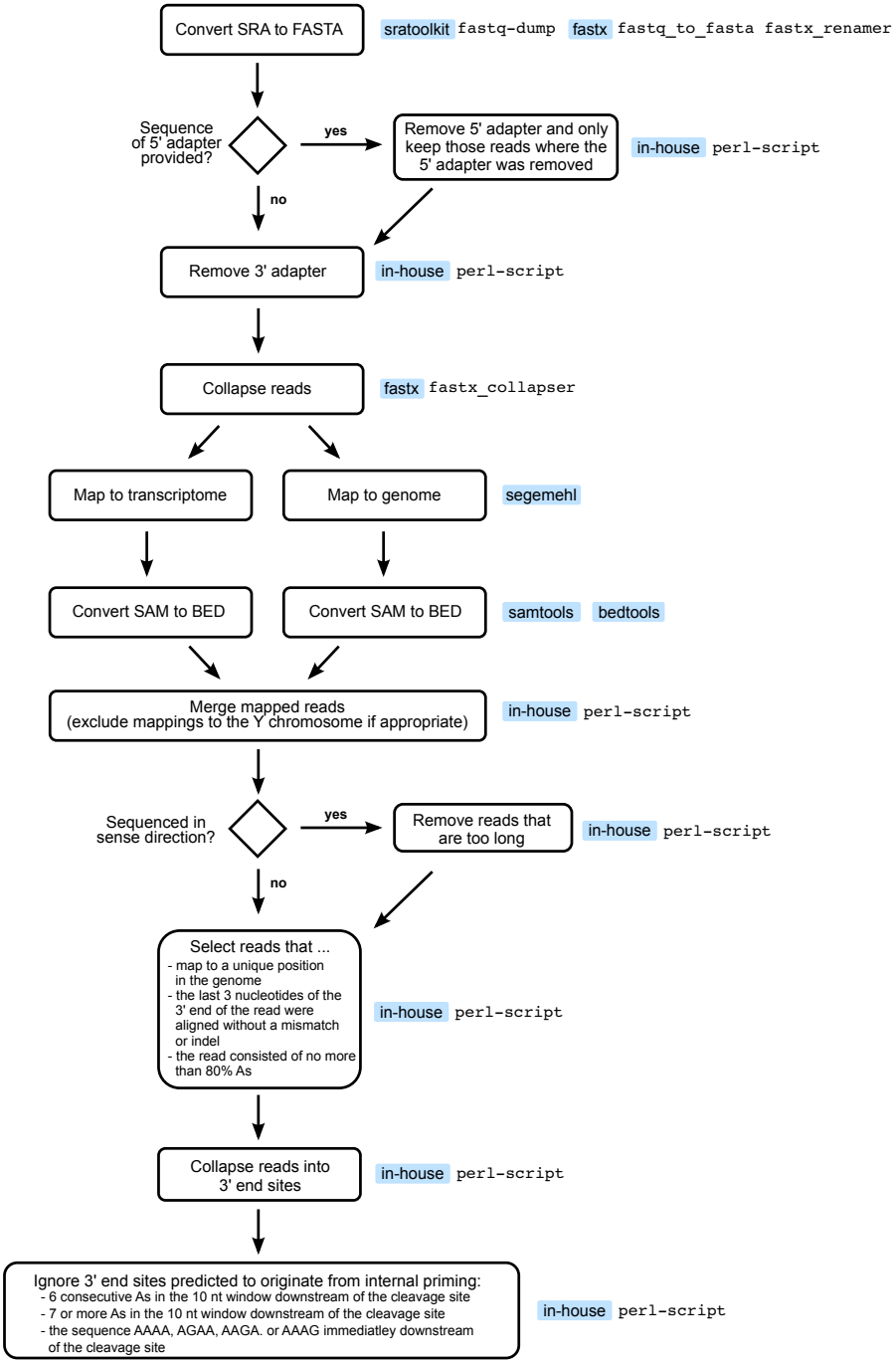


Figure E.21: Outline of the computational pipeline for processing 3' end sequencing data.

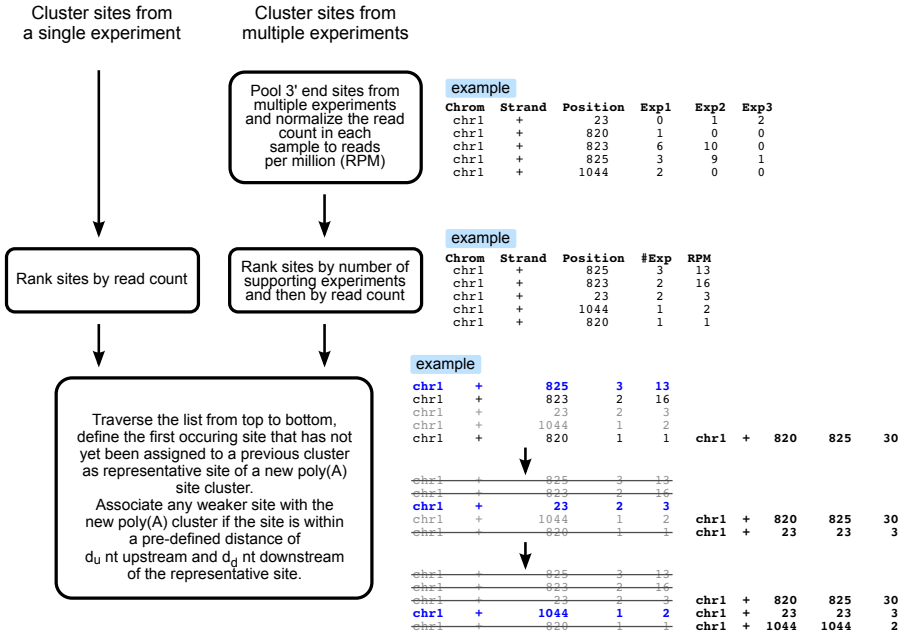
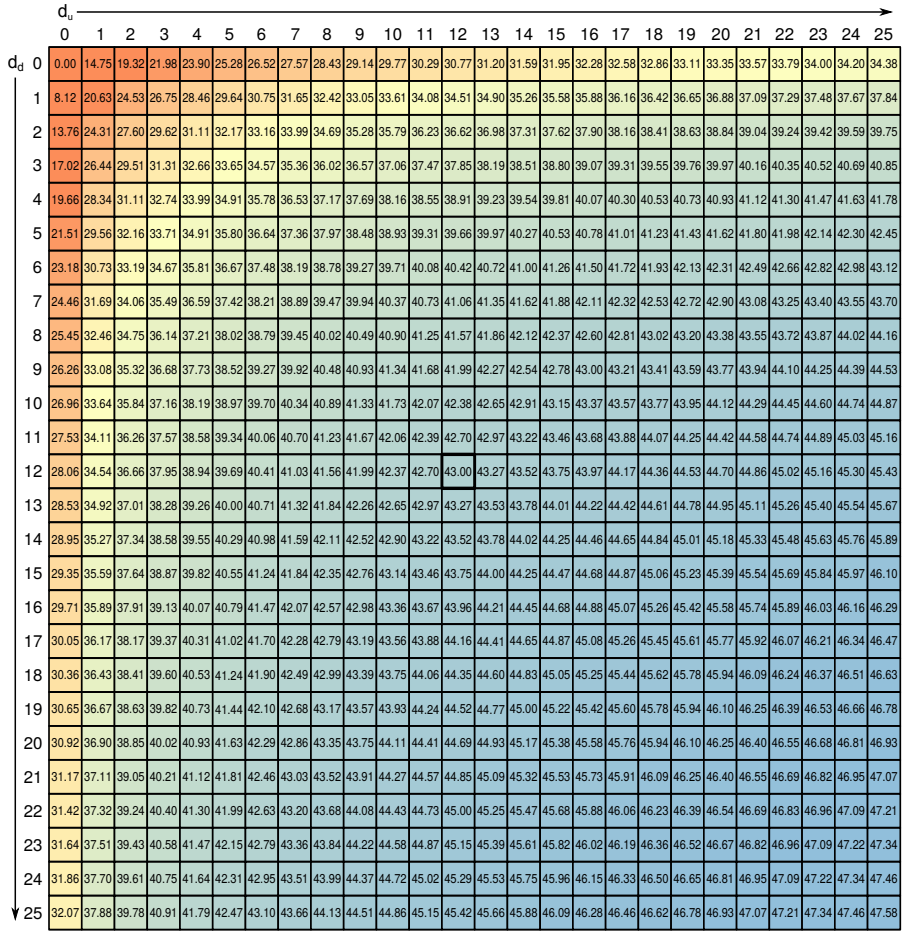


Figure E.22: Outline of the computational pipeline for clustering closely spaced 3' end sites into 3' end processing regions. A toy example data set is used to illustrate the procedure.



d_d \ d_u	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
0	0.00	14.75	19.32	21.98	23.90	25.28	26.52	27.57	28.43	29.14	29.77	30.29	30.77	31.20	31.59	31.95	32.28	32.58	32.86	33.11	33.35	33.57	33.79	34.00	34.20	34.38
1	8.12	20.63	24.53	26.75	28.46	29.64	30.75	31.65	32.42	33.05	33.61	34.08	34.51	34.90	35.26	35.58	35.88	36.16	36.42	36.65	36.88	37.09	37.29	37.48	37.67	37.84
2	13.76	24.31	27.60	29.62	31.11	32.17	33.16	33.99	34.69	35.28	35.79	36.23	36.62	36.98	37.31	37.62	37.90	38.16	38.41	38.63	38.84	39.04	39.24	39.42	39.59	39.75
3	17.02	26.44	29.51	31.31	32.66	33.65	34.57	35.36	36.02	36.57	37.06	37.47	37.85	38.19	38.51	38.80	39.07	39.31	39.55	39.76	39.97	40.16	40.35	40.52	40.69	40.85
4	19.66	28.34	31.11	32.74	33.99	34.91	35.78	36.53	37.17	37.69	38.16	38.55	38.91	39.23	39.54	39.81	40.07	40.30	40.53	40.73	40.93	41.12	41.30	41.47	41.63	41.78
5	21.51	29.56	32.16	33.71	34.91	35.80	36.64	37.36	37.97	38.48	38.93	39.31	39.66	39.97	40.27	40.53	40.78	41.01	41.23	41.43	41.62	41.80	41.98	42.14	42.30	42.45
6	23.18	30.73	33.19	34.67	35.81	36.67	37.48	38.19	38.78	39.27	39.71	40.08	40.42	40.72	41.00	41.26	41.50	41.72	41.93	42.13	42.31	42.49	42.66	42.82	42.98	43.12
7	24.46	31.69	34.06	35.49	36.59	37.42	38.21	38.89	39.47	39.94	40.37	40.73	41.06	41.35	41.62	41.88	42.11	42.32	42.53	42.72	42.90	43.08	43.25	43.40	43.55	43.70
8	25.45	32.46	34.75	36.14	37.21	38.02	38.79	39.45	40.02	40.49	40.90	41.25	41.57	41.86	42.12	42.37	42.60	42.81	43.02	43.20	43.38	43.55	43.72	43.87	44.02	44.16
9	26.26	33.08	35.32	36.68	37.73	38.52	39.27	39.92	40.48	40.93	41.34	41.68	41.99	42.27	42.54	42.78	43.00	43.21	43.41	43.59	43.77	43.94	44.10	44.25	44.39	44.53
10	26.96	33.64	35.84	37.16	38.19	38.97	39.70	40.34	40.89	41.33	41.73	42.07	42.38	42.65	42.91	43.15	43.37	43.57	43.77	43.95	44.12	44.29	44.45	44.60	44.74	44.87
11	27.53	34.11	36.26	37.57	38.58	39.34	40.06	40.70	41.23	41.67	42.06	42.39	42.70	42.97	43.22	43.46	43.68	43.88	44.07	44.25	44.42	44.58	44.74	44.89	45.03	45.16
12	28.06	34.54	36.66	37.95	38.94	39.69	40.41	41.03	41.56	41.99	42.37	42.70	43.00	43.27	43.52	43.75	43.97	44.17	44.36	44.53	44.70	44.86	45.02	45.16	45.30	45.43
13	28.53	34.92	37.01	38.28	39.26	40.00	40.71	41.32	41.84	42.26	42.65	42.97	43.27	43.53	43.78	44.01	44.22	44.42	44.61	44.78	44.95	45.11	45.26	45.40	45.54	45.67
14	28.95	35.27	37.34	38.58	39.55	40.29	40.98	41.59	42.11	42.52	42.90	43.22	43.52	43.78	44.02	44.25	44.46	44.65	44.84	45.01	45.18	45.33	45.48	45.63	45.76	45.89
15	29.35	35.59	37.64	38.87	39.82	40.55	41.24	41.84	42.35	42.76	43.14	43.46	43.75	44.00	44.25	44.47	44.68	44.87	45.06	45.23	45.39	45.54	45.69	45.84	45.97	46.10
16	29.71	35.89	37.91	39.13	40.07	40.79	41.47	42.07	42.57	42.98	43.36	43.67	43.96	44.21	44.45	44.68	44.88	45.07	45.26	45.42	45.58	45.74	45.89	46.03	46.16	46.29
17	30.05	36.17	38.17	39.37	40.31	41.02	41.70	42.28	42.79	43.19	43.56	43.88	44.16	44.41	44.65	44.87	45.08	45.26	45.45	45.61	45.77	45.92	46.07	46.21	46.34	46.47
18	30.36	36.43	38.41	39.60	40.53	41.24	41.90	42.49	42.99	43.39	43.75	44.06	44.35	44.60	44.83	45.05	45.25	45.44	45.62	45.78	45.94	46.09	46.24	46.37	46.51	46.63
19	30.65	36.67	38.63	39.82	40.73	41.44	42.10	42.68	43.17	43.57	43.93	44.24	44.52	44.77	45.00	45.22	45.42	45.60	45.78	45.94	46.10	46.25	46.39	46.53	46.66	46.78
20	30.92	36.90	38.85	40.02	40.93	41.63	42.29	42.86	43.35	43.75	44.11	44.41	44.69	44.93	45.17	45.38	45.58	45.76	45.94	46.10	46.25	46.40	46.55	46.68	46.81	46.93
21	31.17	37.11	39.05	40.21	41.12	41.81	42.46	43.03	43.52	43.91	44.27	44.57	44.85	45.09	45.32	45.53	45.73	45.91	46.09	46.25	46.40	46.55	46.69	46.82	46.95	47.07
22	31.42	37.32	39.24	40.40	41.30	41.99	42.63	43.20	43.68	44.08	44.43	44.73	45.00	45.25	45.47	45.68	45.88	46.06	46.23	46.39	46.54	46.69	46.83	46.96	47.09	47.21
23	31.64	37.51	39.43	40.58	41.47	42.15	42.79	43.36	43.84	44.22	44.58	44.87	45.15	45.39	45.61	45.82	46.02	46.19	46.36	46.52	46.67	46.82	46.96	47.09	47.22	47.34
24	31.86	37.70	39.61	40.75	41.64	42.31	42.95	43.51	43.99	44.37	44.72	45.02	45.29	45.53	45.75	45.96	46.15	46.33	46.50	46.65	46.81	46.95	47.09	47.22	47.34	47.46
25	32.07	37.88	39.78	40.91	41.79	42.47	43.10	43.66	44.13	44.51	44.86	45.15	45.42	45.66	45.88	46.09	46.28	46.46	46.62	46.78	46.93	47.07	47.21	47.34	47.46	47.58

Figure E.23: Evaluation of the distance parameters for clustering closely spaced, putative 3' end processing sites. d_u and d_d refer to the distance upstream and downstream of the representative site, respectively. Values in the plot denote the percentage of 3' end processing sites that were part of a multi-site cluster when a particular set of distance parameters was applied to cluster individual sites. While initially there is a steep increase in the proportion of reads in clusters, a plateau is soon reached. Distances $d_u = 12$ and $d_d = 12$ were chosen in this study.

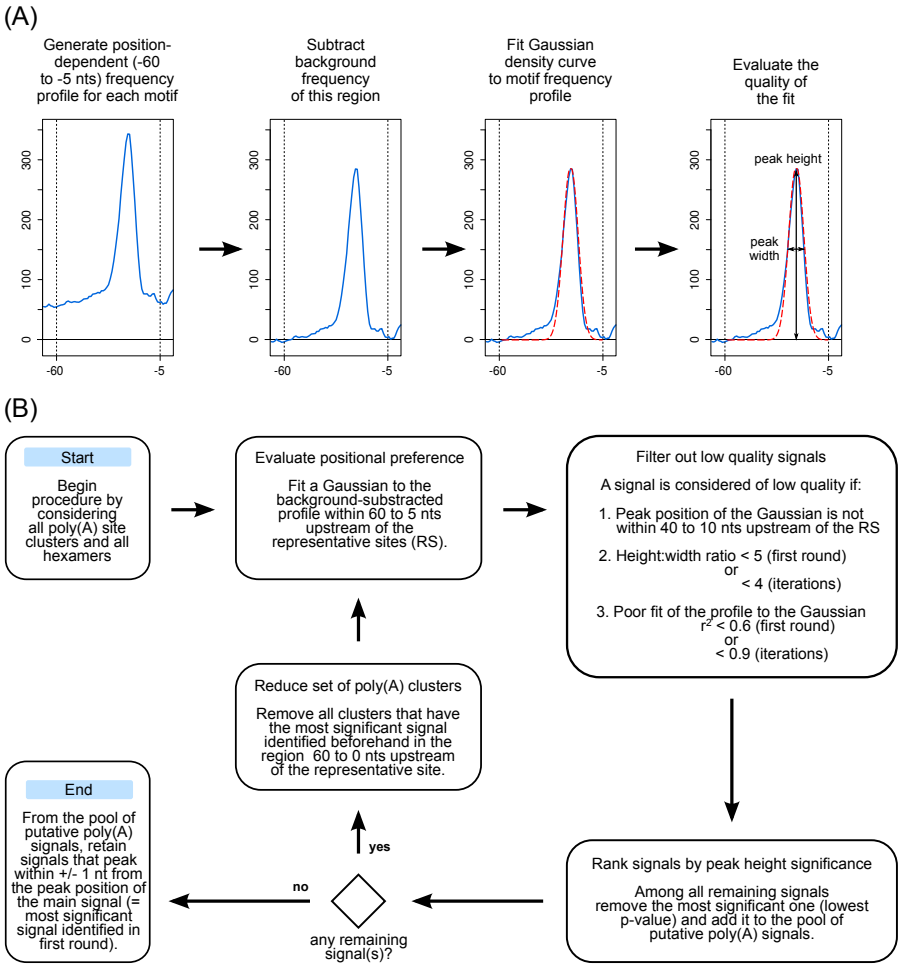


Figure E.24: Outline of the computational procedure used to identify poly(A) signals from poly(A) site clusters obtained from high-throughput sequencing of pre-mRNA 3' ends.

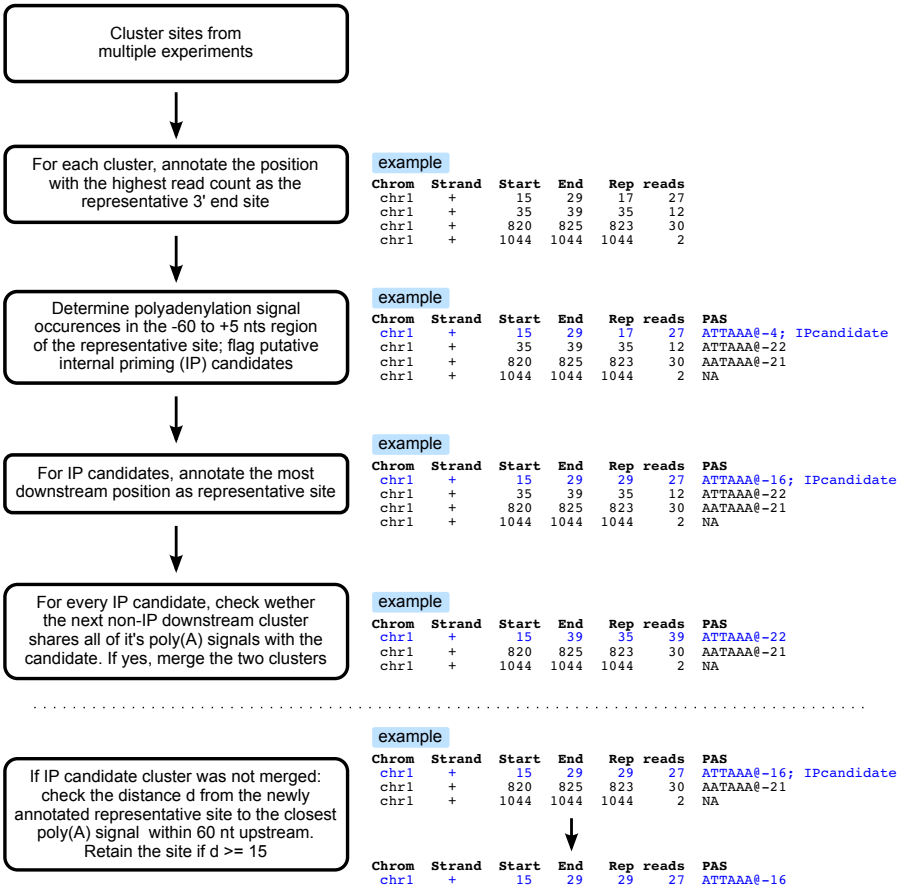


Figure E.25: Outline of the strategy to evaluate poly(A) clusters potentially originating from internal priming.

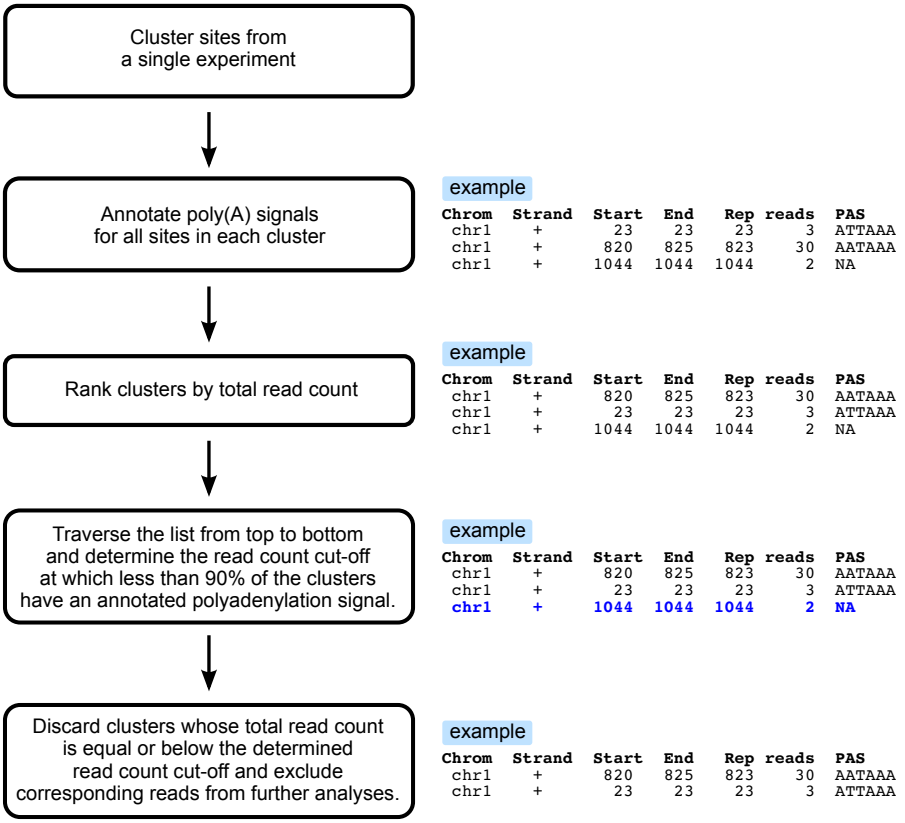


Figure E.26: Outline of the procedure that we used to filter out clusters that do not have sufficient experimental support (sample-specific cut-off of read counts).

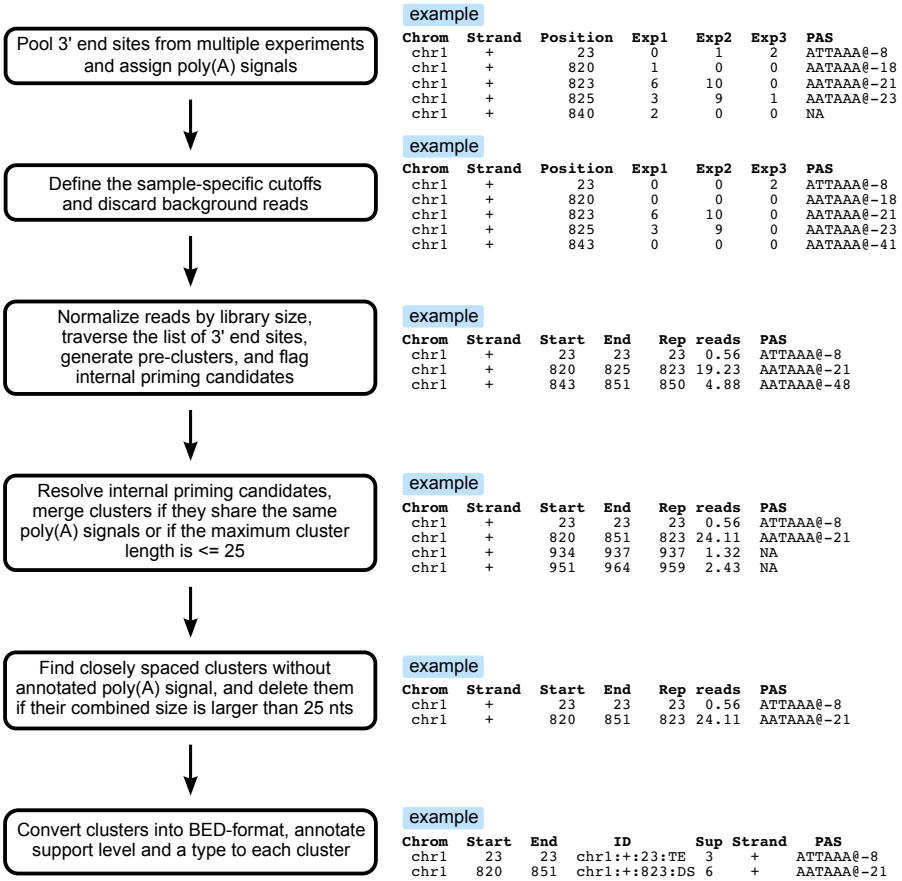


Figure E.27: Outline of the computational procedure that we used to combine 3' end processing sites from multiple experiments into a comprehensive catalog of 3' end processing clusters.

E.3 SUPPLEMENTARY TABLES

Table E.1: Comparison of poly(A) sites that were reported by Derti et al. [98] and You et al. [481] for different human tissues. Both of these studies reported only one genomic position per poly(A) site cluster. To be more permissive in evaluating the overlap of these data sets, we first extended the poly(A) sites from these data sets by 25 nt up- and downstream. A poly(A) site from one study was considered to overlap if there was at least one cluster in the other data set such that both clusters overlapped each other by at least one nucleotide. For each tissue we report both the number of poly(A) site clusters that overlapped as well as those that were unique to a specific data set. In parentheses, the average number of reported reads for the underlying poly(A) sites of the corresponding set of clusters is indicated.

	PolyA-seq clusters overlapping with APASdb clusters	**APASdb clusters overlapping with PolyA-seq clusters**	**PolyA-seq unique clusters**	**APASdb unique clusters**
brain	31,356 (58.47)	30,856 (90.04)	57,754 (19.25)	23,827 (10.83)
kidney	23,793 (104.27)	23,090 (121.53)	71,152 (29.39)	12,006 (19.78)
liver	25,923 (175.45)	25,152 (116.98)	62,317 (16.23)	10,741 (7.26)
muscle	21,910 (151.16)	21,227 (123.36)	90,888 (17.03)	10,743 (37.56)
testes	34,810 (117.72)	34,057 (66.84)	80,258 (11.61)	34,860 (18.47)

Table E.2: Overview of the samples used to build the genome-wide catalog of 3' end processing site in human

series ID	**sample ID**	**protocol**	**tissue/cell type**	**gender**	**publication**
GSE40859	GSM1003590	"DRS"	"HeLa"	F	[603]
GSE40859	GSM1003591	"DRS"	"HeLa"	F	[603]
GSE40859	GSM1003592	"DRS"	"HeLa"	F	[603]
SRP025988	SRX388391	"DRS"	"HeLa"	F	[93]
SRP022151	SRX275752	"DRS"	"K562"	F	[486]
SRP022151	SRX275753	"DRS"	"K562"	F	[486]
SRP022151	SRX275806	"DRS"	"K562"	F	[486]
SRP022151	SRX275827	"DRS"	"K562"	F	[486]
SRP003483	SRX026582	"SAPAS"	"MDA-MB-231"	F	[606]
SRP003483	SRX026583	"SAPAS"	"MCF-10A"	F	[606]
SRP003483	SRX026584	"SAPAS"	"MCF-7"	F	[606]
GSE25450	GSM624686	"PAS-Seq"	"HeLa"	F	[479]
GSE30198	GSM747470	"PolyA-seq"	"Brain"	NA	[98]
GSE30198	GSM747471	"PolyA-seq"	"Kidney"	NA	[98]

Continued on next page

Table E.2 – continued from previous page

series ID	sample ID	protocol	tissue/cell type	gender	publication
GSE30198	GSM747472	"PolyA-seq"	"Liver"	NA	[98]
GSE30198	GSM747473	"PolyA-seq"	"MAQC Brain"	NA	[98]
GSE30198	GSM747474	"PolyA-seq"	"MAQC Brain"	NA	[98]
GSE30198	GSM747475	"PolyA-seq"	"MAQC UHR"	NA	[98]
GSE30198	GSM747476	"PolyA-seq"	"MAQC UHR"	NA	[98]
GSE30198	GSM747477	"PolyA-seq"	"Muscle"	NA	[98]
GSE30198	GSM747479	"PolyA-seq"	"Testis"	NA	[98]
GSE30198	GSM747480	"PolyA-seq"	"UHR"	NA	[98]
GSE37037	GSM909242	"A-seq"	"HEK293"	F	[81]
GSE37037	GSM909243	"A-seq"	"HEK293"	F	[81]
GSE37037	GSM909244	"A-seq"	"HEK293"	F	[81]
GSE37037	GSM909245	"A-seq"	"HEK293"	F	[81]
GSE40137	GSM986133	"A-seq"	"HEK293"	F	[82]
GSE40137	GSM986134	"A-seq"	"HEK293"	F	[82]
GSE40137	GSM986135	"A-seq"	"HEK293"	F	[82]
GSE40137	GSM986136	"A-seq"	"HEK293"	F	[82]
GSE40137	GSM986137	"A-seq"	"HEK293"	F	[82]
GSE40137	GSM986138	"A-seq"	"HEK293"	F	[82]
SRP029953	SRX351949	"3'-Seq"	"native B cells"	NA	[500]
SRP029953	SRX351950	"3'-Seq"	"native B cells"	NA	[500]
SRP029953	SRX351952	"3'-Seq"	"brain"	NA	[500]
SRP029953	SRX351953	"3'-Seq"	"breast"	F	[500]
SRP029953	SRX359328	"3'-Seq"	"embryonic stem cells (H9)"	F	[500]
SRP029953	SRX359329	"3'-Seq"	"ovary"	F	[500]
SRP029953	SRX359330	"3'-Seq"	"skeletal muscle"	NA	[500]
SRP029953	SRX359331	"3'-Seq"	"testis"	NA	[500]
SRP029953	SRX359332	"3'-Seq"	"MCF10A"	F	[500]
SRP029953	SRX359333	"3'-Seq"	"MCF10A"	F	[500]
SRP029953	SRX359334	"3'-Seq"	"MCF7"	F	[500]
SRP029953	SRX359335	"3'-Seq"	"HeLa"	F	[500]
SRP029953	SRX359336	"3'-Seq"	"HEK293"	F	[500]
SRP029953	SRX359337	"3'-Seq"	"NTERA2"	M	[500]
SRP029953	SRX359339	"3'-Seq"	"B-LCL cells"	NA	[500]
SRP029953	SRX359340	"3'-Seq"	"MCF10A"	F	[500]
SRP029953	SRX359341	"3'-Seq"	"MCF10A"	F	[500]
GSE52527	GSM1268942	"3P-Seq"	"HeLa"	F	[472]
GSE52527	GSM1268943	"3P-Seq"	"HEK293"	F	[472]
GSE52527	GSM1268944	"3P-Seq"	"Huh7"	NA	[472]
GSE52527	GSM1268945	"3P-Seq"	"IMR90"	F	[472]
GSE56657	GSM1366428	"DRS"	"neuroendocrine tumor"	F	[604]
GSE56657	GSM1366429	"DRS"	"neuroendocrine tumor"	M	[604]

Continued on next page

Table E.2 – continued from previous page

series ID	sample ID	protocol	tissue/cell type	gender	publication
GSE56657	GSM1366430	"DRS"	"Pituitary"	M	[604]
SRP041182	SRX517334	"SAPAS"	"testis"	M	[481]
SRP041182	SRX517333	"SAPAS"	"ovary"	F	[481]
SRP041182	SRX517332	"SAPAS"	"skeletal muscle"	M	[481]
SRP041182	SRX517331	"SAPAS"	"adipose"	M	[481]
SRP041182	SRX517330	"SAPAS"	"thymus"	M	[481]
SRP041182	SRX517329	"SAPAS"	"small intestine"	M	[481]
SRP041182	SRX517328	"SAPAS"	"pancreas"	F	[481]
SRP041182	SRX517327	"SAPAS"	"liver"	M	[481]
SRP041182	SRX517326	"SAPAS"	"prostate"	M	[481]
SRP041182	SRX517325	"SAPAS"	"breast"	F	[481]
SRP041182	SRX517324	"SAPAS"	"bladder"	F	[481]
SRP041182	SRX517323	"SAPAS"	"uterus"	F	[481]
SRP041182	SRX517322	"SAPAS"	"lung"	M	[481]
SRP041182	SRX517321	"SAPAS"	"placenta"	F	[481]
SRP041182	SRX517320	"SAPAS"	"lymph node"	M	[481]
SRP041182	SRX517319	"SAPAS"	"heart"	M	[481]
SRP041182	SRX517318	"SAPAS"	"cervix"	F	[481]
SRP041182	SRX517317	"SAPAS"	"kidney"	M	[481]
SRP041182	SRX517316	"SAPAS"	"stomach"	M	[481]
SRP041182	SRX517315	"SAPAS"	"spleen"	M	[481]
SRP041182	SRX517314	"SAPAS"	"thyroid"	F	[481]
SRP041182	SRX517313	"SAPAS"	"brain"	F	[481]

Table E.3: Overview of the samples used to build the genome-wide catalog of 3' end processing site in mouse

series ID	sample ID	protocol	tissue/cell type	gender	publication
GSE30198	GSM747481	"PolyA-seq"	"Brain"	NA	[98]
GSE30198	GSM747482	"PolyA-seq"	"Kidney"	NA	[98]
GSE30198	GSM747483	"PolyA-seq"	"Liver"	NA	[98]
GSE30198	GSM747484	"PolyA-seq"	"Muscle"	NA	[98]
GSE30198	GSM747485	"PolyA-seq"	"Testis"	NA	[98]
GSE54950	GSM1327166	"A-seq V2"	"T cells"	NA	[429]
GSE54950	GSM1327167	"A-seq V2"	"T cells"	NA	[429]
GSE54950	GSM1327168	"A-seq V2"	"T cells"	NA	[429]
GSE54950	GSM1327169	"A-seq V2"	"T cells"	NA	[429]
GSE46433	GSM1130096	"2P-Seq"	"embryonic stem cells"	NA	[485]
GSE46433	GSM1130097	"2P-Seq"	"embryonic stem cells"	NA	[485]
GSE46433	GSM1130098	"2P-Seq"	"embryonic stem cells"	NA	[485]
GSE46433	GSM1130099	"2P-Seq"	"embryonic stem cells"	NA	[485]

Continued on next page

Table E.3 – continued from previous page

series ID	sample ID	protocol	tissue/cell type	gender	publication
GSE46433	GSM1130100	"2P-Seq"	"embryonic stem cells"	NA	[485]
GSE46433	GSM1130101	"2P-Seq"	"embryonic stem cells"	NA	[485]
SRP025988	SRX304982	"DRS"	"embryonic stem cell line E14Tg2a"	M	[93]
SRP025988	SRX304983	"DRS"	"embryonic stem cell line E14Tg2a"	M	[93]
GSE44698	GSM1089085	"2P-Seq"	"3T3"	NA	[103]
GSE44698	GSM1089086	"2P-Seq"	"3T3"	NA	[103]
GSE44698	GSM1089087	"2P-Seq"	"3T3"	NA	[103]
GSE44698	GSM1089088	"2P-Seq"	"3T3"	NA	[103]
GSE44698	GSM1089089	"2P-Seq"	"3T3"	NA	[103]
GSE44698	GSM1089090	"2P-Seq"	"3T3"	NA	[103]
GSE44698	GSM1089091	"2P-Seq"	"3T3"	NA	[103]
GSE44698	GSM1089092	"2P-Seq"	"3T3"	NA	[103]
GSE44698	GSM1089093	"2P-Seq"	"3T3"	NA	[103]
GSE44698	GSM1089094	"2P-Seq"	"3T3"	NA	[103]
GSE44698	GSM1089095	"2P-Seq"	"3T3"	NA	[103]
GSE44698	GSM1089096	"2P-Seq"	"3T3"	NA	[103]
GSE52528	GSM1268946	"3P-seq"	"heart"	NA	[472]
GSE52528	GSM1268947	"3P-seq"	"muscle"	NA	[472]
GSE52528	GSM1268948	"3P-seq"	"liver"	NA	[472]
GSE52528	GSM1268949	"3P-seq"	"lung"	NA	[472]
GSE52528	GSM1268950	"3P-seq"	"wat"	NA	[472]
GSE52528	GSM1268951	"3P-seq"	"kidney"	NA	[472]
GSE52528	GSM1268952	"3P-seq"	"heart"	NA	[472]
GSE52528	GSM1268953	"3P-seq"	"muscle"	NA	[472]
GSE52528	GSM1268954	"3P-seq"	"liver"	NA	[472]
GSE52528	GSM1268955	"3P-seq"	"lung"	NA	[472]
GSE52528	GSM1268956	"3P-seq"	"wat"	NA	[472]
GSE52528	GSM1268957	"3P-seq"	"kidney"	NA	[472]
GSE52528	GSM1268958	"3P-seq"	"embryonic stem cells"	NA	[472]
GSE25450	GSM624687	"PAS-Seq"	"ES"	NA	[479]
GSE60487	GSM1480973	"PolyA-seq V2"	"MEF"	NA	[605]
GSE60487	GSM1480974	"PolyA-seq V2"	"MEF"	NA	[605]
GSE60487	GSM1480975	"PolyA-seq V2"	"MEF"	NA	[605]
GSE60487	GSM1480976	"PolyA-seq V2"	"MEF"	NA	[605]
GSE60487	GSM1480977	"PolyA-seq V2"	"MEF"	NA	[605]
GSE60487	GSM1480978	"PolyA-seq V2"	"MEF"	NA	[605]
GSE60487	GSM1480979	"PolyA-seq V2"	"MEF"	NA	[605]
GSE60487	GSM1480980	"PolyA-seq V2"	"MEF"	NA	[605]

Continued on next page

Table E.3 – continued from previous page

series ID	sample ID	protocol	tissue/cell type	gender	publication
GSE62001	GSM1518105	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518106	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518107	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518108	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518109	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518110	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518111	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518112	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518113	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518082	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518089	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518090	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518102	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518103	"3READS"	"NA"	NA	[504]
GSE62001	GSM1586365	"3READS"	"NA"	NA	[504]
GSE62001	GSM1586366	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518096	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518097	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518098	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518072	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518073	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518074	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518075	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518076	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518077	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518078	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518079	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518080	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518081	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518083	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518084	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518085	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518086	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518087	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518088	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518091	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518092	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518093	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518094	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518095	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518099	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518101	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518104	"3READS"	"NA"	NA	[504]
GSE62001	GSM1586367	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518071	"3READS"	"NA"	NA	[504]
GSE62001	GSM1518114	"3READS"	"NA"	NA	[504]
GSE62001	GSM1586368	"3READS"	"NA"	NA	[504]

Continued on next page

Table E.3 – continued from previous page

series ID	sample ID	protocol	tissue/cell type	gender	publication
GSE62001	GSM1518100	"3READS"	"NA"	NA	[504]
GSE62001	GSM1586363	"3READS"	"NA"	NA	[504]
GSE62001	GSM1586364	"3READS"	"NA"	NA	[504]
SRP039327	SRX480169	"SAPAS"	"thymus"	NA	[481]
SRP039327	SRX480179	"SAPAS"	"thymus"	NA	[481]
SRP039327	SRX480205	"SAPAS"	"thymus"	NA	[481]
SRP039327	SRX480212	"SAPAS"	"thymus"	NA	[481]
SRP039327	SRX480221	"SAPAS"	"thymus"	NA	[481]
SRP039327	SRX480227	"SAPAS"	"thymus"	NA	[481]
SRP039327	SRX480229	"SAPAS"	"thymus"	NA	[481]
SRP039327	SRX480250	"SAPAS"	"thymus"	NA	[481]
SRP039327	SRX480287	"SAPAS"	"thymus"	NA	[481]

Table E.4: The 100 most significantly enriched hexamers (binomial test relative to what is expected given the mononucleotide composition of the region from -60 to 0 nt relative to poly(A) site) in the human poly(A) site catalog

hexamer	-log p-value
AATAAA	122788.1
AAATAA	42670.49
AAAAAA	33960.3
ATAAAA	33379.19
TAAAAA	24249.76
AAAATA	21755.03
AAAAAT	19162.31
TTAAAA	16451.96
ATAAAT	14493.43
AAAAAG	14079.72
TTTTTT	13455.43
ATTAAA	12302.28
TAAAAT	11913.92
GCCTGG	11751.91
ATAAAG	11628.45
CCTGGG	11165.77
TTTTCT	10964.83
TGTTTT	10879.94
CCAGCC	10729.18
AAAATG	9002.596
CAGCCT	8279.236
CTTTTT	8043.175
AGAAAA	7959.7
TTTCTT	7707.476

Continued on next page

Table E.4 – continued from previous page

hexamer	-log p-value
CTGGGC	7594.283
AAAGAA	7535.008
AAGAAA	7519.484
AAATGT	7297.44
GAAAAA	7156.527
AGCCTG	7106.297
TTTAAA	7019.924
TTTTTC	6929.253
TTTTGT	6754.398
CCTCCC	6622.351
TTGTTT	6515.799
TTCTTT	6484.465
TTTTAA	6444.964
TTTCTG	6351.61
CAATAA	6137.289
TAAATG	5913.602
TTTTTG	5750.779
AAAAAC	5741.94
TAAATA	5719.061
TCTTTT	5691.07
ATTTTT	5690.314
CTCCAG	5609.213
CAAAAA	5564.294
TTTGTT	5252.513
TTTTTA	5163.368
CTGTCT	5128.945
TGTGTG	5124.415
AAAACA	5094.2
CCCAGC	5042.282
TTCTGT	5016.795
CTCTGT	4984.282
ATAAAC	4984.15
CTCCCC	4866.824
TATTTT	4738.292
AAAAGA	4679.872
TTTCCT	4662.104
CTGCTG	4550.984
TTTTCC	4286.656
CCTGGC	4259.37
CCTGCC	4236.644
CTGCCT	4207.258
CTGTTT	4086.569

Continued on next page

Table E.4 – continued from previous page

hexamer	-log p-value
CCCTCC	4082.152
GGAAAA	4078.892
ACAGAG	4074.031
CTGTGT	4001.796
TCTGTG	3969.594
GTTTTT	3911.444
CCCAGG	3869.135
TGTCTC	3865.269
GCCTCC	3851.923
TGCTTT	3843.789
TGCCTG	3713.514
CTTCCC	3708.302
CCCCAG	3686.223
TAATAA	3629.887
TTTCTC	3577.619
TGGAAA	3574.17
TAAAAG	3557.743
TGCTGT	3532.84
TTTATT	3526.132
CCCCCA	3524.531
TCCAGC	3520.258
GAATAA	3458.727
GCTGTG	3405.909
TCTCTG	3392.311
CCACTG	3378.823
CCTCTG	3304.089
TTTCCC	3297.584
GGGAGG	3271.045
CATTTT	3270.061
TTCCTG	3266.088
CTGCCC	3236.691
CTTTCT	3230.07
CAGAGC	3226.857
CTGTGG	3207.589

Table E.5: The 100 most significantly enriched hexamers (binomial test relative to what is expected given the mononucleotide composition of the region from -60 to 0 nt relative to poly(A) site) in the mouse poly(A) site catalog

hexamer	-log p-value
AATAAA	78344.66

Continued on next page

Table E.5 – continued from previous page

hexamer	-log p-value
AAAAAA	33032.07
AAATAA	28932.12
ATAAAT	17302.62
ATAAAA	14803.36
TAAAAA	12938.72
TTAAAA	10366.85
TAAATA	10122.15
AAAAAG	8097.119
ATTAAA	7668.254
CAGTGT	6974.536
ATAAAG	6855.813
AAAATA	6839.607
ACAGTG	6185.573
CTGCCT	5763.978
TGTTTT	5692.668
TGTCTG	5583.763
CCTCCC	5520.302
TTTAAA	5008.553
GTGTAC	4968.018
GTGTGT	4958.019
GACAGC	4933.256
TAAAAT	4914.887
AAAAAT	4852.199
CCTCTG	4693.22
TAATAA	4460.155
CTTCTG	4436.615
TGTGTG	4411.729
CTGAAG	4159.753
TGTACT	4135.415
TTGTTT	3858.373
TTTTGT	3721.03
ATAAAC	3683.916
CCTGCC	3667.125
GTGTCT	3663.924
TTTTCT	3652.31
TGCCTC	3617.359
CTACAG	3575.848
AAAGAA	3570.49
GCTACA	3527.289
TTCTGG	3512.262
CTGTCT	3499.525
TTTGTT	3488.113

Continued on next page

Table E.5 – continued from previous page

hexamer	-log p-value
CTCCCC	3386.621
AGACAG	3353.467
TCTGAA	3231.828
ACAGCT	3161.227
CTGGTG	3148.898
AAATCT	3076.442
TCTGCC	3032.614
AAATGT	3023.56
CTGTGT	2979.327
CTCTGC	2974.548
AGTGTA	2935.839
CAATAA	2867.629
TTTCCT	2843.454
GGTGTG	2836.151
TGTGTC	2810.496
CCTGTC	2803.988
TTTTTT	2748.095
CCCTGT	2719.253
TGAAGA	2718.407
CTTCCT	2690.973
AAGAAA	2651.799
AAAAGA	2636.556
CCCTCC	2573.799
CTGCTG	2560.113
TTTCTT	2559.386
GCTGGG	2522.802
AAAAAC	2519.491
TCTCTG	2486.791
TCTGTG	2482.156
TTTCTG	2480.577
AAACCC	2460.335
AGCTAC	2456.855
TTTTAA	2438.885
TGCTGG	2436.94
CCTGGG	2436.371
GTCTGA	2414.336
TGCTGT	2412.297
CTCTGT	2361.324
TTCTGT	2360.056
GTGCTG	2358.721
AAAATG	2341.729
CAGCTA	2295.836

Continued on next page

Table E.5 – continued from previous page

hexamer	-log p-value
CCCTCT	2275.77
TACAGT	2265.152
TGTCTC	2255.793
TAAATG	2252.428
CTCCTG	2230.726
TTCTTT	2206.821
AAAACA	2176.917
CTGGGA	2176.094
TGCCTG	2171.784
CTCTTC	2161.823
GCCTCC	2150.538
GCTGTG	2141.131
TAAATC	2138.624
ACCCTG	2131.258
CCTGTG	2111.563

Table E.6: Summary statistics of 3' end sequencing libraries (A-Seq2 protocol [429]) for control-siRNA and HNRNPC-siRNA transfected HEK 293 cells.

	control-siRNA replicate 1 (ID: 29765)	**HNRNPC-siRNA replicate 1 (ID: 29766)**	**control-siRNA replicate 2 (ID: 32682)**	**HNRNPC-siRNA replicate 2 (ID: 32683)**
Number of reads sequenced	55,274,416	47,917,208	68,650,218	78,065,144
considered high-confidence reads that mapped to a unique position in the genome	6,836,446	9,265,965	13,818,252	15,319,388
Number of reads assigned to tandem poly(A) site clusters having >1 protocol support	2,991,716	4,115,507	6,989,361	8,601,510
Number of reads assigned to sample-specific clusters	2,976,577	4,107,667	6,893,361	8,529,512

Table E.7: Overview of the number and the proportion of features annotated in the human genome that are covered by poly(A) sites from different atlases.

		total	PolyAsite		PolyA-seq		APASdb	
			covered sites	percentage	covered sites	percentage	covered sites	percentage
genes	protein coding	21,232	18,139	85.43 %	17,742	83.56 %	16,724	78.77 %
	lincRNA	7,048	4,160	59.02 %	3,745	53.14 %	2,387	33.87 %
terminal exons	protein coding	59,869	42,579	71.12 %	39,670	66.26 %	37,533	62.69 %
	lincRNA	7,153	2,689	37.59 %	2,115	29.57 %	1,753	24.51 %

Table E.8: Overview of the number and the proportion of features annotated in the mouse genome that are covered by poly(A) sites from different atlases.

	total	PolyAsite		PolyA-seq	
		covered sites	percentage	covered sites	percentage
genes	43,054	22,988	53.39 %	21,088	48.98 %
terminal exons	92,351	38,529	41.72 %	31,903	34.55 %

E.4 SUPPLEMENTARY DATA

Please request the data from the author or access it online at: http://genome.cshlp.org/content/suppl/2016/07/04/gr.202432.115.DC1/Supplementary_Data_S1.bed

Table E.9: Supplemental Data Human.

Please request the data from the author or access it online at: http://genome.cshlp.org/content/suppl/2016/07/04/gr.202432.115.DC1/Supplementary_Data_S2.bed

Table E.10: Supplemental Data Mouse.

BIBLIOGRAPHY

[1] G. Manhes, C. J. Allègre, B. Dupré, and B. Hamelin. Lead isotope study of basic-ultrabasic layered complexes: Speculations about the age of the earth and primitive mantle characteristics. *Earth Planet. Sci. Lett.*, 47(3):370–382, 1980.

[2] G. B. Dalrymple. The age of the earth in the twentieth century: a problem (mostly) solved. *Geological Society, London, Special Publications*, 190(1):205–221, 2001.

[3] N. Noffke, D. Christian, D. Wacey, and R. M. Hazen. Microbially induced sedimentary structures recording an ancient ecosystem in the ca. 3.48 billion-year-old dresser formation, pilbara, western australia. *Astrobiology*, 13(12):1103–1124, 2013.

[4] S. L. Miller. A production of amino acids under possible primitive earth conditions. *Science*, 117(3046):528–529, 1953.

[5] S. L. Miller and H. C. Urey. Organic compound synthesis on the primitive earth: Several questions about the origin of life have been answered, but much remains to be studied. *Science*, 130(3370):245–251, 1959.

[6] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*, volume 4th edition. Garland Science, 2002.

[7] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Essential Cell Biology*, volume 3rd edition. Garland Science, 2010.

[8] G. Li and D. Reinberg. Chromatin higher-order structures and gene regulation. *Curr. Opin. Genet. Dev.*, 21(2):175–186, 2011.

[9] O. Bell, V. K. Tiwari, N. H. Thomä, and D. Schübeler. Determinants and dynamics of genome accessibility. *Nat. Rev. Genet.*, 12(8):554–564, 2011.

[10] F. Spitz and E. E. M. Furlong. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, 13(9):613–626, 2012.

[11] Z. Wang, C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, W. Peng, M. Q. Zhang, and K. Zhao. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet*, 40(7):897–903, 2008.

[12] R. Karlic, H. R. Chung, J. Lasserre, K. Vlahovicek, and M. Vingron. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A*, 107(7):2926–2931, 2010.

[13] S. C. Tippmann, R. Ivanek, D. Gaidatzis, A. Scholer, L. Hoerner, E. van Nimwegen, P. F. Stadler, M. B. Stadler, and D. Schubeler. Chromatin measurements reveal contributions of synthesis and decay to steady-state mRNA levels. *Mol Syst Biol*, 8:593, 2012.

[14] X. Dong, M. C. Greven, A. Kundaje, S. Djebali, J. B. Brown, C. Cheng, T. R. Gingeras, M. Gerstein, R. Guigo, E. Birney, and Z. Weng. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol*, 13(9):R53, 2012.

[15] A. Eccleston, N. DeWitt, C. Gunter, B. Marte, and D. Nath. Epigenetics. *Nature*, 447(7143):395–395, 2007.

[16] H. Ledford. Language: Disputed definitions. *Nature*, 455(7216):1023–1028, 2008.

[17] J. Lan, S. Hua, X. He, and Y. Zhang. DNA methyltransferases and methyl-binding proteins of mammals. *Acta Biochim. Biophys. Sin.*, 42 (4):243–252, 2010.

[18] M. Weber, I. Hellmann, M. B. Stadler, L. Ramos, S. Pääbo, M. Rebhan, and D. Schübeler. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.*, 39(4):457–466, 2007.

[19] F. Watt and P. L. Molloy. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes Dev.*, 2(9):1136–1143, 1988.

[20] R. Santoro and I. Grummt. Molecular mechanisms mediating methylation-dependent silencing of ribosomal gene transcription. *Mol. Cell*, 8(3):719–725, 2001.

[21] M. Wiench, S. John, S. Baek, T. A. Johnson, M.-H. Sung, T. Escobar, C. A. Simmons, K. H. Pearce, S. C. Biddie, P. J. Sabo, R. E. Thurman, J. A. Stamatoyannopoulos, and G. L. Hager. DNA methylation status predicts cell type-specific enhancer activity. *EMBO J.*, 30(15):3028–3039, 2011.

[22] P. L. Jones, G. J. Veenstra, P. A. Wade, D. Vermaak, S. U. Kass, N. Landsberger, J. Strouboulis, and A. P. Wolffe. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat. Genet.*, 19(2):187–191, 1998.

[23] O. Bogdanović and G. J. C. Veenstra. DNA methylation and methyl-CpG binding proteins: developmental requirements and function. *Chromosoma*, 118(5):549–565, 2009.

[24] M. Ku, R. P. Koche, E. Rheinbay, E. M. Mendenhall, M. Endoh, T. S. Mikkelsen, A. Presser, C. Nusbaum, X. Xie, A. S. Chi, M. Adli, S. Kasif, L. M. Ptaszek, C. A. Cowan, E. S. Lander, H. Koseki, and B. E. Bernstein. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.*, 4(10): e1000242, 2008.

[25] E. Viré, C. Brenner, R. Deplus, L. Blanchon, M. Fraga, C. Didelot, L. Morey, A. Van Eynde, D. Bernard, J.-M. Vanderwinden, M. Bollen, M. Esteller, L. Di Croce, Y. de Launoit, and F. Fuks. The polycomb group protein EZH2 directly controls DNA methylation. *Nature*, 439 (7078):871–874, 2006.

[26] E. M. Mendenhall, R. P. Koche, T. Truong, V. W. Zhou, B. Issac, A. S. Chi, M. Ku, and B. E. Bernstein. GC-rich sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genet.*, 6(12):e1001244, 2010.

[27] A. Sing, D. Pannell, A. Karaiskakis, K. Sturgeon, M. Djabali, J. Ellis, H. D. Lipshitz, and S. P. Cordes. A vertebrate polycomb response element governs segmentation of the posterior hindbrain. *Cell*, 138(5): 885–897, 2009.

[28] C. J. Woo, P. V. Kharchenko, L. Daheron, P. J. Park, and R. E. Kingston. A region of the human HOXD cluster that confers polycomb-group responsiveness. *Cell*, 140(1):99–110, 2010.

[29] R. A. Gupta, N. Shah, K. C. Wang, J. Kim, H. M. Horlings, D. J. Wong, M.-C. Tsai, T. Hung, P. Argani, J. L. Rinn, Y. Wang, P. Brzoska, B. Kong, R. Li, R. B. West, M. J. van de Vijver, S. Sukumar, and H. Y. Chang. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, 464(7291):1071–1076, 2010.

[30] M.-C. Tsai, O. Manor, Y. Wan, N. Mosammaparast, J. K. Wang, F. Lan, Y. Shi, E. Segal, and H. Y. Chang. Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, 329(5992):689–693, 2010.

[31] K. Plath, J. Fang, S. K. Mlynarczyk-Evans, R. Cao, K. A. Worringer, H. Wang, C. C. de la Cruz, A. P. Otte, B. Panning, and Y. Zhang. Role of histone H3 lysine 27 methylation in X inactivation. *Science*, 300 (5616):131–135, 2003.

[32] J. Zhao, B. K. Sun, J. A. Erwin, J.-J. Song, and J. T. Lee. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science*, 322(5902):750–756, 2008.

[33] S. I. S. Grewal and D. Moazed. Heterochromatin and epigenetic control of gene expression. *Science*, 301(5634):798–802, 2003.

[34] X.-J. Yang and E. Seto. HATs and HDACs: from structure, function and regulation to novel strategies for therapy and prevention. *Oncogene*, 26(37):5310–5318, 2007.

[35] B. G. Wilson and C. W. M. Roberts. SWI/SNF nucleosome remodellers and cancer. *Nat. Rev. Cancer*, 11(7):481–492, 2011.

[36] L. Ho, E. L. Miller, J. L. Ronan, W. Q. Ho, R. Jothi, and G. R. Crabtree. esBAF facilitates pluripotency by conditioning the genome for LIF/STAT3 signalling and by regulating polycomb function. *Nat. Cell Biol.*, 13(8):903–913, 2011.

[37] Z.-K. Zhang, K. P. Davies, J. Allen, L. Zhu, R. G. Pestell, D. Zagzag, and G. V. Kalpana. Cell cycle arrest and repression of cyclin D1 transcription by INI1/hSNF5. *Mol. Cell. Biol.*, 22(16):5975–5988, 2002.

[38] D. S. Latchman. Transcription factors: an overview. *Int. J. Biochem. Cell Biol.*, 29(12):1305–1312, 1997.

[39] S. T. Whiteside and S. Goodbourn. Signal transduction and nuclear targeting: regulation of transcription factor activity by subcellular localisation. *J. Cell Sci.*, 104 (Pt 4):949–955, 1993.

[40] P. Cartwright and K. Helin. Nucleocytoplasmic shuttling of transcription factors. *Cell. Mol. Life Sci.*, 57(8-9):1193–1206, 2000.

[41] L. Xu and J. Massagué. Nucleocytoplasmic shuttling of signal transducers. *Nat. Rev. Mol. Cell Biol.*, 5(3):209–219, 2004.

[42] J. M. Olefsky. Nuclear receptor minireview series. *J. Biol. Chem.*, 276 (40):36863–36864, 2001.

[43] R. M. Evans and D. J. Mangelsdorf. Nuclear receptors, RXR, and the big bang. *Cell*, 157(1):255–266, 2014.

[44] J. D. Nardozzi, K. Lott, and G. Cingolani. Phosphorylation meets nuclear import: a review. *Cell Commun. Signal.*, 8:32, 2010.

[45] C. M. Horvath, G. R. Stark, I. M. Kerr, and J. E. Darnell, Jr. Interactions between STAT and non-STAT proteins in the interferon-stimulated gene factor 3 transcription complex. *Mol. Cell. Biol.*, 16 (12):6957–6964, 1996.

[46] M. Martinez-Moczygemba, M. J. Gutch, D. L. French, and N. C. Reich. Distinct STAT structure promotes interaction of STAT2 with the p48 subunit of the interferon-alpha-stimulated transcription factor ISGF3. *J. Biol. Chem.*, 272(32):20070–20076, 1997.

[47] C. Schindler, D. E. Levy, and T. Decker. JAK-STAT signaling: from interferons to cytokines. *J. Biol. Chem.*, 282(28):20059–20063, 2007.

[48] L. C. Platanias. Mechanisms of type-i- and type-II-interferon-mediated signalling. *Nat. Rev. Immunol.*, 5(5):375–386, 2005.

[49] J. E. Darnell, Jr, I. M. Kerr, and G. R. Stark. Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science*, 264(5164):1415–1421, 1994.

[50] FANTOM Consortium, H. Suzuki, A. R. R. Forrest, E. van Nimwegen, C. O. Daub, P. J. Balwierz, K. M. Irvine, T. Lassmann, T. Ravasi, Y. Hasegawa, M. J. L. de Hoon, S. Katayama, K. Schroder, P. Carninci, Y. Tomaru, M. Kanamori-Katayama, A. Kubosaki, A. Akalin, Y. Ando, E. Arner, M. Asada, H. Asahara, T. Bailey, V. B. Bajic, D. Bauer, A. G. Beckhouse, N. Bertin, J. Björkegren, F. Brombacher, E. Bulger, A. M. Chalk, J. Chiba, N. Cloonan, A. Dawe, J. Dostie, P. G. Engström, M. Essack, G. J. Faulkner, J. L. Fink, D. Fredman, K. Fujimori, M. Furuno, T. Gojobori, J. Gough, S. M. Grimmond, M. Gustafsson, M. Hashimoto, T. Hashimoto, M. Hatakeyama, S. Heinzel, W. Hide, O. Hofmann, M. Hörnquist, L. Huminiecki, K. Ikeo, N. Imamoto, S. Inoue, Y. Inoue, R. Ishihara, T. Iwayanagi, A. Jacobsen, M. Kaur, H. Kawaji, M. C. Kerr, R. Kimura, S. Kimura, Y. Kimura, H. Kitano, H. Koga, T. Kojima, S. Kondo, T. Konno, A. Krogh, A. Kruger, A. Kumar, B. Lenhard, A. Lennartsson, M. Lindow, M. Lizio, C. Macpherson, N. Maeda, C. A. Maher, M. Maqungo, J. Mar, N. A. Matigian, H. Matsuda, J. S. Mattick, S. Meier, S. Miyamoto, E. Miyamoto-Sato, K. Nakabayashi, Y. Nakachi, M. Nakano, S. Nygaard, T. Okayama, Y. Okazaki, H. Okuda-Yabukami, V. Orlando, J. Otomo, M. Pachkov, N. Petrovsky, C. Plessy, J. Quackenbush, A. Radovanovic, M. Rehli, R. Saito, A. Sandelin, S. Schmeier, C. Schönbach, A. S. Schwartz, C. A. Semple, M. Sera, J. Severin, K. Shirahige, C. Simons, G. St Laurent, M. Suzuki, T. Suzuki, M. J. Sweet, R. J. Taft, S. Takeda, Y. Takenaka, K. Tan, M. S. Taylor, R. D. Teasdale, J. Tegnér, S. Teichmann, E. Valen, C. Wahlestedt, K. Waki, A. Waterhouse, C. A. Wells, O. Winther, L. Wu, K. Yamaguchi, H. Yanagawa, J. Yasuda, M. Zavolan, D. A. Hume, Riken Omics Science Center, T. Arakawa, S. Fukuda, K. Imamura, C. Kai, A. Kaiho, T. Kawashima, C. Kawazu, Y. Kitazume, M. Kojima, H. Miura, K. Murakami, M. Murata, N. Ninomiya, H. Nishiyori, S. Noma, C. Ogawa, T. Sano, C. Simon, M. Tagami, Y. Takahashi, J. Kawai, and Y. Hayashizaki. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet.*, 41(5):553–562, 2009.

[51] T. Glisovic, J. L. Bachorik, J. Yong, and G. Dreyfuss. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.*, 582(14): 1977–1986, 2008.

[52] N. Proudfoot. Connecting transcription to messenger RNA processing. *Trends Biochem. Sci.*, 25(6):290–293, 2000.

[53] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[54] S. Gerstberger, M. Hafner, and T. Tuschl. A census of human RNA-binding proteins. *Nat. Rev. Genet.*, 15(12):829–845, 2014.

[55] V. H. Cowling. Regulation of mRNA cap methylation. *Biochem. J*, 425(2):295–302, 2010.

[56] Y. Lee and D. C. Rio. Mechanisms and regulation of alternative Pre-mRNA splicing. *Annu. Rev. Biochem.*, 84:291–323, 2015.

[57] D. J. Smith, C. C. Query, and M. M. Konarska. "nought may endure but mutability": spliceosome dynamics and the regulation of splicing. *Mol. Cell*, 30(6):657–666, 2008.

[58] A. J. Matlin, F. Clark, and C. W. J. Smith. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, 6(5):386–398, 2005.

[59] E. de Klerk and P. A. C. 't Hoen. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends Genet.*, 31(3):128–139, 2015.

[60] T. W. Nilsen and B. R. Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463, 2010.

[61] G. Ast. How did alternative splicing evolve? *Nat. Rev. Genet.*, 5(10): 773–782, 2004.

[62] Y. Xing and C. Lee. Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes. *Nat. Rev. Genet.*, 7(7):499–509, 2006.

[63] L. Fedorova and A. Fedorov. Introns in gene evolution. *Genetica*, 118 (2-3):123–131, 2003.

[64] M. Chen and J. L. Manley. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat. Rev. Mol. Cell Biol.*, 10(11):741–754, 2009.

[65] M. C. Wahl, C. L. Will, and R. Lührmann. The spliceosome: design principles of a dynamic RNP machine. *Cell*, 136(4):701–718, 2009.

[66] A. Busch and K. J. Hertel. Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdiscip. Rev. RNA*, 3(1):1–12, 2012.

[67] A. G. Matera and Z. Wang. A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.*, 15(2):108–121, 2014.

[68] H. Sun and L. A. Chasin. Multiple splicing defects in an intronic false exon. *Mol. Cell. Biol.*, 20(17):6414–6425, 2000.

[69] X.-D. Fu and M. Ares, Jr. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet.*, 15(10):689–701, 2014.

[70] J. F. Cáceres, S. Stamm, D. M. Helfman, and A. R. Krainer. Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors. *Science*, 265(5179):1706–1709, 1994.

[71] J. Zhu, A. Mayeda, and A. R. Krainer. Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Mol. Cell*, 8(6): 1351–1361, 2001.

[72] D. L. Black. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, 72:291–336, 2003.

[73] O. Porrua and D. Libri. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat. Rev. Mol. Cell Biol.*, 16(3):190–202, 2015.

[74] Y. Shi, D. C. Di Giammartino, D. Taylor, A. Sarkeshik, W. J. Rice, J. R. Yates, 3rd, J. Frank, and J. L. Manley. Molecular architecture of the human pre-mRNA 3' processing complex. *Mol. Cell*, 33(3):365–376, 2009.

[75] D. Zheng and B. Tian. RNA-binding proteins in regulation of alternative cleavage and polyadenylation. *Adv. Exp. Med. Biol.*, 825:97–127, 2014.

[76] A. R. Gruber, G. Martin, W. Keller, and M. Zavolan. Means to an end: mechanisms of alternative polyadenylation of messenger RNA precursors. *Wiley Interdiscip. Rev. RNA*, 5(2):183–196, 2014.

[77] Y. Takagaki and J. L. Manley. RNA recognition by the human polyadenylation factor CstF. *Mol. Cell. Biol.*, 17(7):3907–3914, 1997.

[78] C. Yao, E.-A. Choi, L. Weng, X. Xie, J. Wan, Y. Xing, J. J. Moresco, P. G. Tu, J. R. Yates, 3rd, and Y. Shi. Overlapping and distinct functions of CstF64 and cstf64τ in mammalian mRNA 3' processing. *RNA*, 19(12):1781–1790, 2013.

[79] Y. Shi and J. L. Manley. The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site. *Genes Dev.*, 29(9):889–897, 2015.

[80] K. M. Brown and G. M. Gilmartin. A mechanism for the regulation of pre-mRNA 3' processing by human cleavage factor im. *Mol. Cell*, 12 (6):1467–1476, 2003.

[81] G. Martin, A. R. Gruber, W. Keller, and M. Zavolan. Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Rep.*, 1 (6):753–763, 2012.

[82] A. R. Gruber, G. Martin, W. Keller, and M. Zavolan. Cleavage factor im is a key regulator of 3' UTR length. *RNA Biol.*, 9(12):1405–1412, 2012.

[83] H. de Vries, U. Rüegsegger, W. Hübner, A. Friedlein, H. Langen, and W. Keller. Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors. *EMBO J.*, 19 (21):5895–5904, 2000.

[84] S. West and N. J. Proudfoot. Human pcf11 enhances degradation of RNA polymerase II-associated nascent RNA and transcriptional termination. *Nucleic Acids Res.*, 36(3):905–914, 2008.

[85] S. Bienroth, E. Wahle, C. Suter-Crazzolara, and W. Keller. Purification of the cleavage and polyadenylation factor involved in the 3'-processing of messenger RNA precursors. *J. Biol. Chem.*, 266(29): 19768–19776, 1991.

[86] K. G. Murthy and J. L. Manley. Characterization of the multisubunit cleavage-polyadenylation specificity factor from calf thymus. *J. Biol. Chem.*, 267(21):14804–14811, 1992.

[87] I. Kaufmann, G. Martin, A. Friedlein, H. Langen, and W. Keller. Human fip1 is a subunit of CPSF that binds to u-rich RNA elements and stimulates poly(a) polymerase. *EMBO J.*, 23(3):616–626, 2004.

[88] C. R. Mandel, S. Kaneko, H. Zhang, D. Gebauer, V. Vethantham, J. L. Manley, and L. Tong. Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature*, 444(7121):953–956, 2006.

[89] E. Pauws, A. H. van Kampen, S. A. van de Graaf, J. J. de Vijlder, and C. Ris-Stalpers. Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. *Nucleic Acids Res.*, 29(8):1690–1694, 2001.

[90] E. Beaudoing, S. Freier, J. R. Wyatt, J. M. Claverie, and D. Gautheret. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.*, 10(7):1001–1010, 2000.

[91] S. L. Chan, I. Huppertz, C. Yao, L. Weng, J. J. Moresco, J. R. Yates, 3rd, J. Ule, J. L. Manley, and Y. Shi. CPSF30 and wdr33 directly bind to AAUAAA in mammalian mRNA 3' processing. *Genes Dev.*, 28(21): 2370–2380, 2014.

[92] L. Schönemann, U. Kühn, G. Martin, P. Schäfer, A. R. Gruber, W. Keller, M. Zavolan, and E. Wahle. Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. *Genes Dev.*, 28(21):2381–2393, 2014.

[93] B. Lackford, C. Yao, G. M. Charles, L. Weng, X. Zheng, E.-A. Choi, X. Xie, J. Wan, Y. Xing, J. M. Freudenberg, P. Yang, R. Jothi, G. Hu, and Y. Shi. Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal. *EMBO J.*, 33(8):878–889, 2014.

[94] K. G. Murthy and J. L. Manley. The 160-kd subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation. *Genes Dev.*, 9(21):2672–2683, 1995.

[95] E. Wahle. A novel poly(a)-binding protein acts as a specificity factor in the second phase of messenger RNA polyadenylation. *Cell*, 66(4): 759–768, 1991.

[96] U. Kühn and E. Wahle. Structure and function of poly(a) binding proteins. *Biochim. Biophys. Acta*, 1678(2-3):67–84, 2004.

[97] M. Hoque, Z. Ji, D. Zheng, W. Luo, W. Li, B. You, J. Y. Park, G. Yehia, and B. Tian. Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat. Methods*, 10(2):133–139, 2013.

[98] A. Derti, P. Garrett-Engele, K. D. Macisaac, R. C. Stevens, S. Sriram, R. Chen, C. A. Rohl, J. M. Johnson, and T. Babak. A quantitative atlas of polyadenylation in five mammals. *Genome Res.*, 22(6):1173–1183, 2012.

[99] R. Elkon, A. P. Ugalde, and R. Agami. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.*, 14 (7):496–506, 2013.

[100] R. Elkon, J. Drost, G. van Haaften, M. Jenal, M. Schrier, J. A. F. Oude Vrielink, and R. Agami. E2F mediates enhanced alternative polyadenylation in proliferation. *Genome Biol.*, 13(7):R59, 2012.

[101] C. Mayr and D. P. Bartel. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138(4):673–684, 2009.

[102] R. Sandberg, J. R. Neilson, A. Sarma, P. A. Sharp, and C. B. Burge. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science*, 320(5883):1643–1647, 2008.

[103] N. Spies, C. B. Burge, and D. P. Bartel. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.*, 23(12):2078–2090, 2013.

[104] B. Tian and J. L. Manley. Alternative cleavage and polyadenylation: the long and short of it. *Trends Biochem. Sci.*, 38(6):312–320, 2013.

[105] B. D. Berkovits and C. Mayr. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature*, 522(7556):363–367, 2015.

[106] L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027):769–773, 2005.

[107] A. A. Bazzini, M. T. Lee, and A. J. Giraldez. Ribosome profiling shows that mir-430 reduces translation before causing mRNA decay in zebrafish. *Science*, 336(6078):233–237, 2012.

[108] S. Djuranovic, A. Nahvi, and R. Green. miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. *Science*, 336(6078):237–240, 2012.

[109] C.-Y. A. Chen, D. Zheng, Z. Xia, and A.-B. Shyu. Ago-TNRC6 triggers microRNA-mediated decay by promoting two deadenylation steps. *Nat. Struct. Mol. Biol.*, 16(11):1160–1166, 2009.

[110] Y. Lee, M. Kim, J. Han, K.-H. Yeom, S. Lee, S. H. Baek, and V. N. Kim. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.*, 23(20):4051–4060, 2004.

[111] W. Filipowicz, S. N. Bhattacharyya, and N. Sonenberg. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.*, 9(2):102–114, 2008.

[112] D. P. Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233, 2009.

[113] E. P. Murchison, J. F. Partridge, O. H. Tam, S. Cheloufi, and G. J. Hannon. Characterization of dicer-deficient murine embryonic stem cells. *Proc. Natl. Acad. Sci. U. S. A.*, 102(34):12135–12140, 2005.

[114] Y. Wang, R. Medvid, C. Melton, R. Jaenisch, and R. Blelloch. DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. *Nat. Genet.*, 39(3):380–385, 2007.

[115] A. J. Gruber and M. Zavolan. Modulation of epigenetic regulators and cell fate decisions by miRNAs. *Epigenomics*, 5(6):671–683, 2013.

[116] S. Griffiths-Jones. miRBase: microRNA sequences and annotation. *Curr. Protoc. Bioinformatics*, Chapter 12:Unit 12.9.1–10, 2010.

[117] T. Vergoulis, I. S. Vlachos, P. Alexiou, G. Georgakilas, M. Maragkakis, M. Reczko, S. Gerangelos, N. Koziris, T. Dalamagas, and A. G. Hatzigeorgiou. TarBase 6.0: capturing the exponential growth of miRNA

targets with experimental support. *Nucleic Acids Res.*, 40(Database issue):D222–9, 2012.

[118] B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, 2005.

[119] J. Hausser, A. P. Syed, B. Bilen, and M. Zavolan. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res.*, 23(4):604–615, 2013.

[120] B. Wightman, I. Ha, and G. Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in c. elegans. *Cell*, 75(5):855–862, 1993.

[121] R. C. Lee, R. L. Feinbaum, and V. Ambros. The c. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, 1993.

[122] M. Chalfie, H. R. Horvitz, and J. E. Sulston. Mutations that lead to reiterations in the cell lineages of c. elegans. *Cell*, 24(1):59–69, 1981.

[123] M. S. Ebert and P. A. Sharp. Roles for microRNAs in conferring robustness to biological processes. *Cell*, 149(3):515–524, 2012.

[124] C. M. Croce and G. A. Calin. miRNAs, cancer, and stem cell division. *Cell*, 122(1):6–7, 2005.

[125] S. Udali, P. Guarini, S. Moruzzi, S.-W. Choi, and S. Friso. Cardiovascular epigenetics: from DNA methylation to microRNAs. *Mol. Aspects Med.*, 34(4):883–901, 2013.

[126] R. J. Perera and A. Ray. Epigenetic regulation of miRNA genes and their role in human melanomas. *Epigenomics*, 4(1):81–90, 2012.

[127] A. Rouhi, D. L. Mager, R. K. Humphries, and F. Kuchenbauer. MiRNAs, epigenetics, and cancer. *Mamm. Genome*, 19(7-8):517–525, 2008.

[128] Z. Wang, H. Yao, S. Lin, X. Zhu, Z. Shen, G. Lu, W. S. Poon, D. Xie, M. C.-M. Lin, and H.-F. Kung. Transcriptional and epigenetic regulation of human microRNAs. *Cancer Lett.*, 331(1):1–10, 2013.

[129] R. I. Gregory, K.-P. Yan, G. Amuthan, T. Chendrimada, B. Doratotaj, N. Cooch, and R. Shiekhattar. The microprocessor complex mediates the genesis of microRNAs. *Nature*, 432(7014):235–240, 2004.

[130] T. P. Chendrimada, R. I. Gregory, E. Kumaraswamy, J. Norman, N. Cooch, K. Nishikura, and R. Shiekhattar. TRBP recruits the dicer complex to ago2 for microRNA processing and gene silencing. *Nature*, 436(7051):740–744, 2005.

[131] H. Siomi and M. C. Siomi. Posttranscriptional regulation of microRNA biogenesis in animals. *Mol. Cell*, 38(3):323–332, 2010.

[132] E. P. Ricci, T. Limousin, R. Soto-Rifo, P. S. Rubilar, D. Decimo, and T. Ohlmann. miRNA repression of translation in vitro takes place during 43S ribosomal scanning. *Nucleic Acids Res.*, 41(1):586–598, 2013.

[133] T. Fukaya and Y. Tomari. PABP is not essential for microRNA-mediated translational repression and deadenylation in vitro. *EMBO J.*, 30(24):4998–5009, 2011.

[134] G. Mathonnet, M. R. Fabian, Y. V. Svitkin, A. Parsyan, L. Huck, T. Murata, S. Biffo, W. C. Merrick, E. Darzynkiewicz, R. S. Pillai, W. Filipowicz, T. F. Duchaine, and N. Sonenberg. MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F. *Science*, 317(5845):1764–1767, 2007.

[135] J. E. Braun, E. Huntzinger, and E. Izaurralde. A molecular link between miRISCs and deadenylases provides new insight into the mechanism of gene silencing by microRNAs. *Cold Spring Harb. Perspect. Biol.*, 4(12), 2012.

[136] P. S. Linsley, J. Schelter, J. Burchard, M. Kibukawa, M. M. Martin, S. R. Bartz, J. M. Johnson, J. M. Cummins, C. K. Raymond, H. Dai, N. Chau, M. Cleary, A. L. Jackson, M. Carleton, and L. Lim. Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Mol. Cell. Biol.*, 27(6):2240–2252, 2007.

[137] M. Selbach, B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, 2008.

[138] D. Baek, J. Villén, C. Shin, F. D. Camargo, S. P. Gygi, and D. P. Bartel. The impact of microRNAs on protein output. *Nature*, 455(7209):64–71, 2008.

[139] H. Guo, N. T. Ingolia, J. S. Weissman, and D. P. Bartel. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–840, 2010.

[140] S. W. Chi, J. B. Zang, A. Mele, and R. B. Darnell. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460 (7254):479–486, 2009.

[141] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano, A.-C. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, and T. Tuschl. PAR-CliP–a method to identify transcriptome-wide the binding sites of RNA binding proteins. *J. Vis. Exp.*, (41), 2010.

[142] M. Khorshid, J. Hausser, M. Zavolan, and E. van Nimwegen. A biophysical miRNA-mRNA interaction model infers canonical and non-canonical targets. *Nat. Methods*, 10(3):253–255, 2013.

[143] R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, 19(1):92–105, 2009.

[144] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks. MicroRNA targets in drosophila. *Genome Biol.*, 5(1):R1, 2003.

[145] E. Hornstein and N. Shomron. Canalization of development by microRNAs. *Nat. Genet.*, 38 Suppl:S20–4, 2006.

[146] H. Herranz and S. M. Cohen. MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems. *Genes Dev.*, 24(13):1339–1344, 2010.

[147] A. Re, D. Corá, D. Taverna, and M. Caselle. Genome-wide survey of microRNA-transcription factor feed-forward regulatory circuits in human. *Mol. Biosyst.*, 5(8):854–867, 2009.

[148] M. Osella, C. Bosia, D. Corá, and M. Caselle. The role of incoherent microRNA-mediated feedforward loops in noise buffering. *PLoS Comput. Biol.*, 7(3):e1001101, 2011.

[149] D. Wilson, V. Charoensawan, S. K. Kummerfeld, and S. A. Teichmann. DBD–taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.*, 36(Database issue):D88–92, 2008.

[150] S. Carbon, A. Ireland, C. J. Mungall, S. Shu, B. Marshall, S. Lewis, AmiGO Hub, and Web Presence Working Group. AmiGO: online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289, 2009.

[151] L. Serrano, B. N. Vazquez, and J. Tischfield. Chromatin structure, pluripotency and differentiation. *Exp. Biol. Med.*, 238(3):259–270, 2013.

[152] R. M. Rivera and J. W. Ross. Epigenetics in fertilization and preimplantation embryo development. *Prog. Biophys. Mol. Biol.*, 113(3): 423–432, 2013.

[153] H. Hao. Genome-wide occupancy analysis by ChIP-chip and ChIP-Seq. *Adv. Exp. Med. Biol.*, 723:753–759, 2012.

[154] P. G. Giresi, J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb. FAIRE (Formaldehyde-Assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res.*, 17(6): 877–885, 2007.

[155] R. K. Auerbach, G. Euskirchen, J. Rozowsky, N. Lamarre-Vincent, Z. Moqtaderi, P. Lefrançois, K. Struhl, M. Gerstein, and M. Snyder. Mapping accessible chromatin regions using Sono-Seq. *Proc. Natl. Acad. Sci. U. S. A.*, 106(35):14926–14931, 2009.

[156] J. R. Hesselberth, X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, W. S. Noble, S. Fields, and J. A. Stamatoyannopoulos. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods*, 6(4):283–289, 2009.

[157] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132 (2):311–322, 2008.

[158] D. E. Schones, K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, 2008.

[159] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. LeProust, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236):362–366, 2009.

[160] R. B. Deal, J. G. Henikoff, and S. Henikoff. Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones. *Science*, 328(5982):1161–1164, 2010.

[161] A. J. Bannister and T. Kouzarides. Regulation of chromatin by histone modifications. *Cell Res.*, 21(3):381–395, 2011.

[162] R. Z. Jurkowska, T. P. Jurkowski, and A. Jeltsch. Structure and function of mammalian DNA methyltransferases. *Chembiochem*, 12(2): 206–222, 2011.

[163] A. M. Deaton and A. Bird. CpG islands and the regulation of transcription. *Genes Dev.*, 25(10):1010–1022, 2011.

[164] Y. Bergman and H. Cedar. DNA methylation dynamics in health and disease. *Nat. Struct. Mol. Biol.*, 20(3):274–281, 2013.

[165] R. I. Verona, M. R. W. Mann, and M. S. Bartolomei. Genomic imprinting: intricacies of epigenetic regulation in clusters. *Annu. Rev. Cell Dev. Biol.*, 19:237–259, 2003.

[166] A. M. Cotton, L. Lam, J. G. Affleck, I. M. Wilson, M. S. Peñaherrera, D. E. McFadden, M. S. Kobor, W. L. Lam, W. P. Robinson, and C. J. Brown. Chromosome-wide DNA methylation analysis predicts human tissue-specific X inactivation. *Hum. Genet.*, 130(2):187–201, 2011.

[167] A. Malousi, N. Maglaveras, and S. Kouidou. Intronic CpG content and alternative splicing in human genes containing a single cassette exon. *Epigenetics*, 3(2):69–73, 2008.

[168] S. J. Brown, P. Stoilov, and Y. Xing. Chromatin and epigenetic regulation of pre-mRNA processing. *Hum. Mol. Genet.*, 21(R1):R90–6, 2012.

[169] L. I. Gómez Acuña, A. Fiszbein, M. Alló, I. E. Schor, and A. R. Kornblihtt. Connections between chromatin signatures and splicing. *Wiley Interdiscip. Rev. RNA*, 4(1):77–91, 2013.

[170] G. V. Avvakumov, J. R. Walker, S. Xue, Y. Li, S. Duan, C. Bronner, C. H. Arrowsmith, and S. Dhe-Paganon. Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. *Nature*, 455(7214):822–825, 2008.

[171] J. Sharif, M. Muto, S.-I. Takebayashi, I. Suetake, A. Iwamatsu, T. A. Endo, J. Shinga, Y. Mizutani-Koseki, T. Toyoda, K. Okamura, S. Tajima, K. Mitsuya, M. Okano, and H. Koseki. The SRA protein np95 mediates epigenetic inheritance by recruiting dnmt1 to methylated DNA. *Nature*, 450(7171):908–912, 2007.

[172] L. S. Chuang, H. I. Ian, T. W. Koh, H. H. Ng, G. Xu, and B. F. Li. Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. *Science*, 277(5334):1996–2000, 1997.

[173] P. J. Lee, L. L. Washer, D. J. Law, C. R. Boland, I. L. Horon, and A. P. Feinberg. Limited up-regulation of DNA methyltransferase in human colon cancer reflecting increased cell proliferation. *Proc. Natl. Acad. Sci. U. S. A.*, 93(19):10366–10370, 1996.

[174] K. D. Robertson, K. Keyomarsi, F. A. Gonzales, M. Velicescu, and P. A. Jones. Differential mRNA expression of the human DNA methyltransferases (DNMTs) 1, 3a and 3b during the G(0)/G(1) to S phase transition in normal and tumor cells. *Nucleic Acids Res.*, 28(10):2108–2113, 2000.

[175] D. Watanabe, I. Suetake, T. Tada, and S. Tajima. Stage- and cell-specific expression of dnmt3a and dnmt3b during embryogenesis. *Mech. Dev.*, 118(1-2):187–190, 2002.

[176] M. Okano, D. W. Bell, D. A. Haber, and E. Li. DNA methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.

[177] W. A. Pastor, L. Aravind, and A. Rao. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nat. Rev. Mol. Cell Biol.*, 14(6):341–356, 2013.

[178] S.-L. Lin, D. C. Chang, C.-H. Lin, S.-Y. Ying, D. Leu, and D. T. S. Wu. Regulation of somatic cell reprogramming through inducible mir-302 expression. *Nucleic Acids Res.*, 39(3):1054–1065, 2011.

[179] A. M. Duursma, M. Kedde, M. Schrier, C. le Sage, and R. Agami. mir-148 targets human DNMT3b protein coding region. *RNA*, 14(5): 872–877, 2008.

[180] L. Sinkkonen, T. Hugenschmidt, P. Berninger, D. Gaidatzis, F. Mohn, C. G. Artus-Revel, M. Zavolan, P. Svoboda, and W. Filipowicz. MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. *Nat. Struct. Mol. Biol.*, 15(3):259–267, 2008.

[181] R. Benetti, S. Gonzalo, I. Jaco, P. Muñoz, S. Gonzalez, S. Schoeftner, E. Murchison, T. Andl, T. Chen, P. Klatt, E. Li, M. Serrano, S. Millar, G. Hannon, and M. A. Blasco. A mammalian microRNA cluster controls DNA methylation and telomere recombination via rbl2-dependent regulation of DNA methyltransferases. *Nat. Struct. Mol. Biol.*, 15(9): 998, 2008.

[182] M. Fabbri, R. Garzon, A. Cimmino, Z. Liu, N. Zanesi, E. Callegari, S. Liu, H. Alder, S. Costinean, C. Fernandez-Cymering, S. Volinia, G. Guler, C. D. Morrison, K. K. Chan, G. Marcucci, G. A. Calin, K. Huebner, and C. M. Croce. MicroRNA-29 family reverts aberrant methylation in lung cancer by targeting DNA methyltransferases 3A and 3B. *Proc. Natl. Acad. Sci. U. S. A.*, 104(40):15805–15810, 2007.

[183] R. Garzon, S. Liu, M. Fabbri, Z. Liu, C. E. A. Heaphy, E. Callegari, S. Schwind, J. Pang, J. Yu, N. Muthusamy, V. Havelange, S. Volinia, W. Blum, L. J. Rush, D. Perrotti, M. Andreeff, C. D. Bloomfield, J. C. Byrd, K. Chan, L.-C. Wu, C. M. Croce, and G. Marcucci. MicroRNA-29b induces global DNA hypomethylation and tumor suppressor gene reexpression in acute myeloid leukemia by targeting directly DNMT3A and 3B and indirectly DNMT1. *Blood*, 113(25):6411–6418, 2009.

[184] P. Zhang, B. Huang, X. Xu, and W. C. Sessa. Ten-eleven translocation (tet) and thymine DNA glycosylase (TDG), components of the demethylation pathway, are direct targets of miRNA-29a. *Biochem. Biophys. Res. Commun.*, 437(3):368–373, 2013.

[185] C. Braconi, N. Huang, and T. Patel. MicroRNA-dependent regulation of DNA methyltransferase-1 and tumor suppressor gene expression by interleukin-6 in human malignant cholangiocytes. *Hepatology*, 51(3): 881–890, 2010.

[186] S. Sander, L. Bullinger, K. Klapproth, K. Fiedler, H. A. Kestler, T. F. E. Barth, P. Möller, S. Stilgenbauer, J. R. Pollack, and T. Wirth. MYC

stimulates EZH2 expression by repression of its negative regulator mir-26a. *Blood*, 112(10):4202–4212, 2008.

[187] S. Varambally, Q. Cao, R.-S. Mani, S. Shankar, X. Wang, B. Ateeq, B. Laxman, X. Cao, X. Jing, K. Ramnarayanan, J. C. Brenner, J. Yu, J. H. Kim, B. Han, P. Tan, C. Kumar-Sinha, R. J. Lonigro, N. Palanisamy, C. A. Maher, and A. M. Chinnaiyan. Genomic loss of microRNA-101 leads to overexpression of histone methyltransferase EZH2 in cancer. *Science*, 322(5908):1695–1699, 2008.

[188] J. M. Friedman, P. A. Jones, and G. Liang. The tumor suppressor microRNA-101 becomes an epigenetic player by targeting the polycomb group protein EZH2 in cancer. *Cell Cycle*, 8(15):2313–2314, 2009.

[189] K. E. Szulwach, X. Li, R. D. Smrt, Y. Li, Y. Luo, L. Lin, N. J. Santistevan, W. Li, X. Zhao, and P. Jin. Cross talk between microRNA and epigenetic regulation in adult neurogenesis. *J. Cell Biol.*, 189(1): 127–141, 2010.

[190] A. H. Juan, R. M. Kumar, J. G. Marx, R. A. Young, and V. Sartorelli. Mir-214-dependent regulation of the polycomb protein ezh2 in skeletal muscle and embryonic stem cells. *Mol. Cell*, 36(1):61–74, 2009.

[191] J. Godlewski, M. O. Nowicki, A. Bronisz, S. Williams, A. Otsuki, G. Nuovo, A. Raychaudhury, H. B. Newton, E. A. Chiocca, and S. Lawler. Targeting of the bmi-1 oncogene/stem cell renewal factor by microRNA-128 inhibits glioma proliferation and self-renewal. *Cancer Res.*, 68(22):9125–9130, 2008.

[192] U. Wellner, J. Schubert, U. C. Burk, O. Schmalhofer, F. Zhu, A. Sontag, B. Waldvogel, C. Vannier, D. Darling, A. zur Hausen, V. G. Brunton, J. Morton, O. Sansom, J. Schüler, M. P. Stemmler, C. Herzberger, U. Hopt, T. Keck, S. Brabletz, and T. Brabletz. The EMT-activator ZEB1 promotes tumorigenicity by repressing stemness-inhibiting microRNAs. *Nat. Cell Biol.*, 11(12):1487–1495, 2009.

[193] A. J. Gruber, W. A. Grandy, P. J. Balwierz, Y. A. Dimitrova, M. Pachkov, C. Ciaudo, E. v. Nimwegen, and M. Zavolan. Embryonic stem cell-specific microRNAs contribute to pluripotency by inhibiting regulators of multiple differentiation pathways. *Nucleic Acids Res.*, 42 (14):9313–9326, 2014.

[194] K. Sakurai, C. Furukawa, T. Haraguchi, K.-I. Inada, K. Shiogama, T. Tagawa, S. Fujita, Y. Ueno, A. Ogata, M. Ito, Y. Tsutsumi, and H. Iba. MicroRNAs mir-199a-5p and -3p target the brm subunit of SWI/SNF to generate a double-negative feedback loop in a variety of human cancers. *Cancer Res.*, 71(5):1680–1689, 2011.

[195] A. S. Yoo, B. T. Staahl, L. Chen, and G. R. Crabtree. MicroRNA-mediated switching of chromatin-remodelling complexes in neural development. *Nature*, 460(7255):642–646, 2009.

[196] E. J. Noonan, R. F. Place, D. Pookot, S. Basak, J. M. Whitson, H. Hirata, C. Giardina, and R. Dahiya. mir-449a targets HDAC-1 and induces growth arrest in prostate cancer. *Oncogene*, 28(14):1714–1724, 2009.

[197] J.-F. Chen, E. M. Mandel, J. M. Thomson, Q. Wu, T. E. Callis, S. M. Hammond, F. L. Conlon, and D.-Z. Wang. The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. *Nat. Genet.*, 38(2):228–233, 2006.

[198] C. E. Winbanks, B. Wang, C. Beyer, P. Koh, L. White, P. Kantharidis, and P. Gregorevic. TGF-beta regulates mir-206 and mir-29 to control myogenic differentiation through regulation of HDAC4. *J. Biol. Chem.*, 286(16):13805–13814, 2011.

[199] Z. Li, M. Q. Hassan, M. Jafferji, R. I. Aqeilan, R. Garzon, C. M. Croce, A. J. van Wijnen, J. L. Stein, G. S. Stein, and J. B. Lian. Biological functions of mir-29b contribute to positive regulation of osteoblast differentiation. *J. Biol. Chem.*, 284(23):15676–15684, 2009.

[200] A. H. Williams, G. Valdez, V. Moresi, X. Qi, J. McAnally, J. L. Elliott, R. Bassel-Duby, J. R. Sanes, and E. N. Olson. MicroRNA-206 delays ALS progression and promotes regeneration of neuromuscular synapses in mice. *Science*, 326(5959):1549–1554, 2009.

[201] D. Simon, B. Laloo, M. Barillot, T. Barnetche, C. Blanchard, C. Rooryck, M. Marche, I. Burgelin, I. Coupry, N. Chassaing, B. Gilbert-Dussardier, D. Lacombe, C. Grosset, and B. Arveiler. A mutation in the 3'-UTR of the HDAC6 gene abolishing the post-transcriptional regulation mediated by hsa-mir-433 is linked to a new form of dominant x-linked chondrodysplasia. *Hum. Mol. Genet.*, 19 (10):2015–2027, 2010.

[202] J. A. Simon and R. E. Kingston. Occupying chromatin: Polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put. *Mol. Cell*, 49(5):808–824, 2013.

[203] T. I. Lee, R. G. Jenner, L. A. Boyer, M. G. Guenther, S. S. Levine, R. M. Kumar, B. Chevalier, S. E. Johnstone, M. F. Cole, K.-I. Isono, H. Koseki, T. Fuchikami, K. Abe, H. L. Murray, J. P. Zucker, B. Yuan, G. W. Bell, E. Herbolsheimer, N. M. Hannett, K. Sun, D. T. Odom, A. P. Otte, T. L. Volkert, D. P. Bartel, D. A. Melton, D. K. Gifford, R. Jaenisch, and R. A. Young. Control of developmental regulators by polycomb in human embryonic stem cells. *Cell*, 125(2):301–313, 2006.

[204] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.-K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, 2007.

[205] G. Pan, S. Tian, J. Nie, C. Yang, V. Ruotti, H. Wei, G. A. Jonsdottir, R. Stewart, and J. A. Thomson. Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell*, 1(3):299–312, 2007.

[206] X. D. Zhao, X. Han, J. L. Chew, J. Liu, K. P. Chiu, A. Choo, Y. L. Orlov, W.-K. Sung, A. Shahab, V. A. Kuznetsov, G. Bourque, S. Oh, Y. Ruan, H.-H. Ng, and C.-L. Wei. Whole-genome mapping of histone H3 lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell*, 1(3):286–298, 2007.

[207] P. Arnold, A. Schöler, M. Pachkov, P. J. Balwierz, H. Jørgensen, M. B. Stadler, E. van Nimwegen, and D. Schübeler. Modeling of epigenome dynamics identifies transcription factors that mediate polycomb targeting. *Genome Res.*, 23(1):60–73, 2013.

[208] H. Wang, L. Wang, H. Erdjument-Bromage, M. Vidal, P. Tempst, R. S. Jones, and Y. Zhang. Role of histone H2A ubiquitination in polycomb silencing. *Nature*, 431(7010):873–878, 2004.

[209] Z. Gao, J. Zhang, R. Bonasio, F. Strino, A. Sawai, F. Parisi, Y. Kluger, and D. Reinberg. PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. *Mol. Cell*, 45(3):344–356, 2012.

[210] L. Tavares, E. Dimitrova, D. Oxley, J. Webster, R. Poot, J. Demmers, K. Bezstarosti, S. Taylor, H. Ura, H. Koide, A. Wutz, M. Vidal, S. Elderkin, and N. Brockdorff. RYBP-PRC1 complexes mediate H2A ubiquitylation at polycomb target sites independently of PRC2 and H3K27me3. *Cell*, 148(4):664–678, 2012.

[211] K. L. Yap, S. Li, A. M. Muñoz Cabello, S. Raguz, L. Zeng, S. Mujtaba, J. Gil, M. J. Walsh, and M.-M. Zhou. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell*, 38 (5):662–674, 2010.

[212] J. M. Polo, E. Anderssen, R. M. Walsh, B. A. Schwarz, C. M. Nefzger, S. M. Lim, M. Borkent, E. Apostolou, S. Alaei, J. Cloutier, O. Bar-Nur, S. Cheloufi, M. Stadtfeld, M. E. Figueroa, D. Robinton, S. Natesan, A. Melnick, J. Zhu, S. Ramaswamy, and K. Hochedlinger. A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell*, 151(7): 1617–1632, 2012.

[213] H. Lickert, J. K. Takeuchi, I. Von Both, J. R. Walls, F. McAuliffe, S. L. Adamson, R. M. Henkelman, J. L. Wrana, J. Rossant, and B. G. Bruneau. Baf60c is essential for function of BAF chromatin remodelling complexes in heart development. *Nature*, 432(7013):107–112, 2004.

[214] J. Lessard, J. I. Wu, J. A. Ranish, M. Wan, M. M. Winslow, B. T. Staahl, H. Wu, R. Aebersold, I. A. Graef, and G. R. Crabtree. An essential switch in subunit composition of a chromatin remodeling complex during neural development. *Neuron*, 55(2):201–215, 2007.

[215] L. Ho, J. L. Ronan, J. Wu, B. T. Staahl, L. Chen, A. Kuo, J. Lessard, A. I. Nesvizhskii, J. Ranish, and G. R. Crabtree. An embryonic stem cell chromatin remodeling complex, esBAF, is essential for embryonic stem cell self-renewal and pluripotency. *Proc. Natl. Acad. Sci. U. S. A.*, 106(13):5181–5186, 2009.

[216] L. Ho, R. Jothi, J. L. Ronan, K. Cui, K. Zhao, and G. R. Crabtree. An embryonic stem cell chromatin remodeling complex, esBAF, is an essential component of the core pluripotency transcriptional network. *Proc. Natl. Acad. Sci. U. S. A.*, 106(13):5187–5191, 2009.

[217] N. Singhal, J. Graumann, G. Wu, M. J. Araúzo-Bravo, D. W. Han, B. Greber, L. Gentile, M. Mann, and H. R. Schöler. Chromatin-Remodeling components of the BAF complex facilitate reprogramming. *Cell*, 141(6):943–955, 2010.

[218] J. Hausser, P. Berninger, C. Rodak, Y. Jantscher, S. Wirth, and M. Zavolan. MirZ: an integrated microRNA expression atlas and target prediction resource. *Nucleic Acids Res.*, 37(Web Server issue):W266–72, 2009.

[219] V. Azuara, P. Perry, S. Sauer, M. Spivakov, H. F. Jørgensen, R. M. John, M. Gouti, M. Casanova, G. Warnes, M. Merkenschlager, and A. G. Fisher. Chromatin signatures of pluripotent cell lines. *Nat. Cell Biol.*, 8(5):532–538, 2006.

[220] E. Bernstein, S. Y. Kim, M. A. Carmell, E. P. Murchison, H. Alcorn, M. Z. Li, A. A. Mills, S. J. Elledge, K. V. Anderson, and G. J. Hannon. Dicer is essential for mouse development. *Nat. Genet.*, 35(3):215–217, 2003.

[221] C. Kanellopoulou, S. A. Muljo, A. L. Kung, S. Ganesan, R. Drapkin, T. Jenuwein, D. M. Livingston, and K. Rajewsky. Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes Dev.*, 19(4):489–501, 2005.

[222] A. Marson, S. S. Levine, M. F. Cole, G. M. Frampton, T. Brambrink, S. Johnstone, M. G. Guenther, W. K. Johnston, M. Wernig, J. Newman, J. M. Calabrese, L. M. Dennis, T. L. Volkert, S. Gupta, J. Love,

N. Hannett, P. A. Sharp, D. P. Bartel, R. Jaenisch, and R. A. Young. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, 134(3):521–533, 2008.

[223] H. B. Houbaviy, M. F. Murray, and P. A. Sharp. Embryonic stem cell-specific MicroRNAs. *Dev. Cell*, 5(2):351–358, 2003.

[224] Y. Wang, S. Baskerville, A. Shenoy, J. E. Babiarz, L. Baehner, and R. Blelloch. Embryonic stem cell-specific microRNAs regulate the G1-S transition and promote rapid proliferation. *Nat. Genet.*, 40(12): 1478–1483, 2008.

[225] R. L. Judson, J. E. Babiarz, M. Venere, and R. Blelloch. Embryonic stem cell-specific microRNAs promote induced pluripotency. *Nat. Biotechnol.*, 27(5):459–461, 2009.

[226] F. Anokye-Danso, C. M. Trivedi, D. Juhr, M. Gupta, Z. Cui, Y. Tian, Y. Zhang, W. Yang, P. J. Gruber, J. A. Epstein, and E. E. Morrisey. Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. *Cell Stem Cell*, 8(4):376–388, 2011.

[227] S. Hu, K. D. Wilson, Z. Ghosh, L. Han, Y. Wang, F. Lan, K. J. Ransohoff, P. Burridge, and J. C. Wu. MicroRNA-302 increases reprogramming efficiency via repression of NR2F2. *Stem Cells*, 31(2):259–268, 2013.

[228] D. Subramanyam, S. Lamouille, R. L. Judson, J. Y. Liu, N. Bucay, R. Derynck, and R. Blelloch. Multiple targets of mir-302 and mir-372 promote reprogramming of human fibroblasts to induced pluripotent stem cells. *Nat. Biotechnol.*, 29(5):443–448, 2011.

[229] B. Liao, X. Bao, L. Liu, S. Feng, A. Zovoilis, W. Liu, Y. Xue, J. Cai, X. Guo, B. Qin, R. Zhang, J. Wu, L. Lai, M. Teng, L. Niu, B. Zhang, M. A. Esteban, and D. Pei. MicroRNA cluster 302-367 enhances somatic cell reprogramming by accelerating a mesenchymal-to-epithelial transition. *J. Biol. Chem.*, 286(19):17359–17364, 2011.

[230] R. Li, J. Liang, S. Ni, T. Zhou, X. Qing, H. Li, W. He, J. Chen, F. Li, Q. Zhuang, B. Qin, J. Xu, W. Li, J. Yang, Y. Gan, D. Qin, S. Feng, H. Song, D. Yang, B. Zhang, L. Zeng, L. Lai, M. A. Esteban, and D. Pei. A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts. *Cell Stem Cell*, 7 (1):51–63, 2010.

[231] A. M. Singh and S. Dalton. The cell cycle and myc intersect with mechanisms that regulate pluripotency and reprogramming. *Cell Stem Cell*, 5(2):141–149, 2009.

[232] L. Litovchick, S. Sadasivam, L. Florens, X. Zhu, S. K. Swanson, S. Velmurugan, R. Chen, M. P. Washburn, X. S. Liu, and J. A. DeCaprio.

Evolutionarily conserved multisubunit RBL2/p130 and E2F4 protein complex represses human cell cycle-dependent genes in quiescence. *Mol. Cell*, 26(4):539–551, 2007.

[233] M. T. McCabe, J. N. Davis, and M. L. Day. Regulation of DNA methyltransferase 1 by the pRb/E2F1 pathway. *Cancer Res.*, 65(9):3624–3632, 2005.

[234] L. He, H. Liu, and L. Tang. SWI/SNF chromatin remodeling complex: a new cofactor in reprogramming. *Stem Cell Rev.*, 8(1):128–136, 2012.

[235] C. Melton, R. L. Judson, and R. Blelloch. Opposing microRNA families regulate self-renewal in mouse embryonic stem cells. *Nature*, 463 (7281):621–626, 2010.

[236] P. Landgraf, M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, S. Pfeffer, A. Rice, A. O. Kamphorst, M. Landthaler, C. Lin, N. D. Socci, L. Hermida, V. Fulci, S. Chiaretti, R. Foà, J. Schliwka, U. Fuchs, A. Novosel, R.-U. Müller, B. Schermer, U. Bissels, J. Inman, Q. Phan, M. Chien, D. B. Weir, R. Choksi, G. De Vita, D. Frezzetti, H.-I. Trompeter, V. Hornung, G. Teng, G. Hartmann, M. Palkovits, R. Di Lauro, P. Wernet, G. Macino, C. E. Rogler, J. W. Nagle, J. Ju, F. N. Papavasiliou, T. Benzing, P. Lichter, W. Tam, M. J. Brownstein, A. Bosio, A. Borkhardt, J. J. Russo, C. Sander, M. Zavolan, and T. Tuschl. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, 129(7):1401–1414, 2007.

[237] I. Heo, C. Joo, J. Cho, M. Ha, J. Han, and V. N. Kim. Lin28 mediates the terminal uridylation of let-7 precursor MicroRNA. *Mol. Cell*, 32 (2):276–284, 2008.

[238] A. Rybak, H. Fuchs, L. Smirnova, C. Brandt, E. E. Pohl, R. Nitsch, and F. G. Wulczyn. A feedback loop comprising lin-28 and let-7 controls pre-let-7 maturation during neural stem-cell commitment. *Nat. Cell Biol.*, 10(8):987–993, 2008.

[239] S. R. Viswanathan, G. Q. Daley, and R. I. Gregory. Selective blockade of microRNA processing by lin28. *Science*, 320(5872):97–100, 2008.

[240] M. S. Kumar, J. Lu, K. L. Mercer, T. R. Golub, and T. Jacks. Impaired microRNA processing enhances cellular transformation and tumorigenesis. *Nat. Genet.*, 39(5):673–677, 2007.

[241] X. Zhong, N. Li, S. Liang, Q. Huang, G. Coukos, and L. Zhang. Identification of microRNAs regulating reprogramming factor LIN28 in embryonic stem cells and cancer cells. *J. Biol. Chem.*, 285(53):41961–41971, 2010.

[242] A. S. Yoo, A. X. Sun, L. Li, A. Shcheglovitov, T. Portmann, Y. Li, C. Lee-Messer, R. E. Dolmetsch, R. W. Tsien, and G. R. Crabtree.

MicroRNA-mediated conversion of human fibroblasts to neurons. *Nature*, 476(7359):228–231, 2011.

[243] H. Tang, L. Yao, X. Tao, Y. Yu, M. Chen, R. Zhang, and C. Xu. mir-9 functions as a tumor suppressor in ovarian serous carcinoma by targeting TLN1. *Int. J. Mol. Med.*, 32(2):381–388, 2013.

[244] X. Chen, D. He, X. D. Dong, F. Dong, J. Wang, L. Wang, J. Tang, D.-N. Hu, D. Yan, and L. Tu. MicroRNA-124a is epigenetically regulated and acts as a tumor suppressor by controlling multiple targets in uveal melanoma. *Invest. Ophthalmol. Vis. Sci.*, 54(3):2248–2256, 2013.

[245] M. Karsy, E. Arslan, and F. Moy. Current progress on understanding MicroRNAs in glioblastoma multiforme. *Genes Cancer*, 3(1):3–15, 2012.

[246] D. P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, 2004.

[247] J. E. Babiarz, J. G. Ruby, Y. Wang, D. P. Bartel, and R. Blelloch. Mouse ES cells express endogenous shRNAs, siRNAs, and other microprocessor-independent, dicer-dependent small RNAs. *Genes Dev.*, 22(20):2773–2785, 2008.

[248] C. Ciaudo, N. Servant, V. Cognat, A. Sarazin, E. Kieffer, S. Viville, V. Colot, E. Barillot, E. Heard, and O. Voinnet. Highly dynamic and sex-specific expression of microRNAs during early ES cell differentiation. *PLoS Genet.*, 5(8):e1000620, 2009.

[249] A. K. L. Leung, A. G. Young, A. Bhutkar, G. X. Zheng, A. D. Bosson, C. B. Nielsen, and P. A. Sharp. Genome-wide identification of ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs. *Nat. Struct. Mol. Biol.*, 18(2):237–244, 2011.

[250] G. X. Y. Zheng, A. Ravi, J. M. Calabrese, L. A. Medeiros, O. Kirak, L. M. Dennis, R. Jaenisch, C. B. Burge, and P. A. Sharp. A latent pro-survival function for the mir-290-295 cluster in mouse embryonic stem cells. *PLoS Genet.*, 7(5):e1002054, 2011.

[251] C.-H. Kuo, J. H. Deng, Q. Deng, and S.-Y. Ying. A novel role of mir-302/367 in reprogramming. *Biochem. Biophys. Res. Commun.*, 417(1): 11–16, 2012.

[252] S. A. Hanina, W. Mifsud, T. A. Down, K. Hayashi, D. O'Carroll, K. Lao, E. A. Miska, and M. A. Surani. Genome-wide identification of targets and function of individual MicroRNAs in mouse embryonic stem cells. *PLoS Genet.*, 6(10):e1001163, 2010.

[253] P. J. Balwierz, M. Pachkov, P. Arnold, A. J. Gruber, M. Zavolan, and E. van Nimwegen. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res.*, 24(5):869–884, 2014.

[254] J. S. Tsang, M. S. Ebert, and A. van Oudenaarden. Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures. *Mol. Cell*, 38(1):140–153, 2010.

[255] D. Gaidatzis, E. van Nimwegen, J. Hausser, and M. Zavolan. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, 8:69, 2007.

[256] B. Granadino, C. Arias-de-la Fuente, C. Pérez-Sánchez, M. Párraga, L. A. López-Fernández, J. del Mazo, and J. Rey-Campos. Fhx (foxj2) expression is activated during spermatogenesis and very early in embryonic development. *Mech. Dev.*, 97(1-2):157–160, 2000.

[257] F. Martín-de Lara, P. Sánchez-Aparicio, C. Arias de la Fuente, and J. Rey-Campos. Biological effects of FoxJ2 over-expression. *Transgenic Res.*, 17(6):1131–1141, 2008.

[258] M. K. Skinner, A. Rawls, J. Wilson-Rawls, and E. H. Roalson. Basic helix-loop-helix transcription factor gene family phylogenetics and nomenclature. *Differentiation*, 80(1):1–8, 2010.

[259] K. Yamamizu, Y. Piao, A. A. Sharov, V. Zsiros, H. Yu, K. Nakazawa, D. Schlessinger, and M. S. H. Ko. Identification of transcription factors for lineage-specific ESC differentiation. *Nat. Rep. Stem Cells*, 1(6): 545–559, 2013.

[260] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, 18(20):6097–6100, 1990.

[261] J. Torres and F. M. Watt. Nanog maintains pluripotency of mouse embryonic stem cells by inhibiting NFkappaB and cooperating with stat3. *Nat. Cell Biol.*, 10(2):194–201, 2008.

[262] P. Lüningschrör, B. Stöcker, B. Kaltschmidt, and C. Kaltschmidt. mir-290 cluster modulates pluripotency by repressing canonical nf-κb signaling. *Stem Cells*, 30(4):655–664, 2012.

[263] M. Chae, K. Kim, S.-M. Park, I.-S. Jang, T. Seo, D.-M. Kim, I.-C. Kim, J.-H. Lee, and J. Park. IRF-2 regulates NF-kappaB activity by modulating the subcellular localization of NF-kappaB. *Biochem. Biophys. Res. Commun.*, 370(3):519–524, 2008.

[264] P. J. Hurlin and J. Huang. The MAX-interacting transcription factor network. *Semin. Cancer Biol.*, 16(4):265–274, 2006.

[265] D. E. Ayer and R. N. Eisenman. A switch from Myc:Max to Mad:Max heterocomplexes accompanies monocyte/macrophage differentiation. *Genes Dev.*, 7(11):2110–2119, 1993.

[266] P. J. Hurlin, C. Quéva, P. J. Koskinen, E. Steingrímsson, D. E. Ayer, N. G. Copeland, N. A. Jenkins, and R. N. Eisenman. Mad3 and mad4: novel max-interacting transcriptional repressors that suppress c-myc dependent transformation and are expressed during neural and epidermal differentiation. *EMBO J.*, 14(22):5646–5659, 1995.

[267] C. Quéva, G. A. McArthur, B. M. Iritani, and R. N. Eisenman. Targeted deletion of the s-phase-specific myc antagonist mad3 sensitizes neuronal and lymphoid cells to radiation-induced apoptosis. *Mol. Cell. Biol.*, 21(3):703–712, 2001.

[268] H. Cam and B. D. Dynlacht. Emerging roles for E2F: beyond the G1/S transition and DNA replication. *Cancer Cell*, 3(4):311–316, 2003.

[269] H. Sekine, J. Mimura, M. Yamamoto, and Y. Fujii-Kuriyama. Unique and overlapping transcriptional roles of arylhydrocarbon receptor nuclear translocator (arnt) and arnt2 in xenobiotic and hypoxic responses. *J. Biol. Chem.*, 281(49):37507–37516, 2006.

[270] Y. Han, S.-Z. Kuang, A. Gomer, and D. L. Ramirez-Bergeron. Hypoxia influences the vascular expansion and differentiation of embryonic stem cell cultures through the temporal expression of vascular endothelial growth factor receptors in an ARNT-dependent manner. *Stem Cells*, 28(4):799–809, 2010.

[271] K. Smith and S. Dalton. Myc transcription factors: key regulators behind establishment and maintenance of pluripotency. *Regen. Med.*, 5 (6):947–959, 2010.

[272] H. J. Bussemaker, H. Li, and E. D. Siggia. Regulatory element detection using correlation with expression. *Nat. Genet.*, 27(2):167–171, 2001.

[273] M. Setty, K. Helmy, A. A. Khan, J. Silber, A. Arvey, F. Neezen, P. Agius, J. T. Huse, E. C. Holland, and C. S. Leslie. Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma. *Mol. Syst. Biol.*, 8:605, 2012.

[274] K. Yagita, K. Horie, S. Koinuma, W. Nakamura, I. Yamanaka, A. Urasaki, Y. Shigeyoshi, K. Kawakami, S. Shimada, J. Takeda, and Y. Uchiyama. Development of the circadian oscillator during differentiation of mouse embryonic stem cells in vitro. *Proc. Natl. Acad. Sci. U. S. A.*, 107(8):3846–3851, 2010.

[275] P. Mu, Y.-C. Han, D. Betel, E. Yao, M. Squatrito, P. Ogrodowski, E. de Stanchina, A. D'Andrea, C. Sander, and A. Ventura. Genetic dissection of the mir-17 92 cluster of microRNAs in myc-induced b-cell lymphomas. *Genes Dev.*, 23(24):2806–2811, 2009.

[276] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy–analysis of affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3): 307–315, 2004.

[277] Z. Wu, R. A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. A Model-Based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.*, 99(468):909–917, 2004.

[278] C. Fraley and A. E. Raftery. MCLUST version 3: an R package for normal mixture modeling and model-based clustering. Technical report, DTIC Document, 2006.

[279] B. S. Carvalho and R. A. Irizarry. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, 26(19):2363–2367, 2010.

[280] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[281] C. Ciaudo, F. Jay, I. Okamoto, C.-J. Chen, A. Sarazin, N. Servant, E. Barillot, E. Heard, and O. Voinnet. RNAi-dependent and independent control of LINE1 accumulation and mobility in mouse embryonic stem cells. *PLoS Genet.*, 9(11):e1003791, 2013.

[282] K. J. Livak and T. D. Schmittgen. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-delta delta C(T)) method. *Methods*, 25(4):402–408, 2001.

[283] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3:318–356, 1961.

[284] M. R. Fabian, N. Sonenberg, and W. Filipowicz. Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.*, 79:351–379, 2010.

[285] W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, 5(4):276–287, 2004.

[286] V. Matys, E. Fricke, R. Geffers, E. Gössling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D.-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Münch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, 31(1):374–378, 2003.

[287] M. Pachkov, I. Erb, N. Molina, and E. van Nimwegen. SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res*, 35(Database issue):D127–D131, 2007.

[288] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, 10(4):252–263, 2009.

[289] E. van Nimwegen. Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics*, 8 Suppl 6:S4, 2007.

[290] P. Arnold, I. Erb, M. Pachkov, N. Molina, and E. van Nimwegen. MotEvo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics*, 28(4):487–494, 2012.

[291] P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V. B. Bajic, S. E. Brenner, S. Batalov, A. R. R. Forrest, M. Zavolan, M. J. Davis, L. G. Wilming, V. Aidinis, J. E. Allen, A. Ambesi-Impiombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L. Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christoffels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C. F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojobori, R. E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T. K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P. T. Krishnan, A. Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schönbach, K. Sekiguchi, C. A. M. Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A. Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S. M. Grimmond, R. D. Teasdale, E. T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J. S. Mattick, D. A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shi-

raki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai, Y. Hayashizaki, FANTOM Consortium, and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563, 2005.

[292] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, and T. R. Gingeras. Landscape of transcription in human cells. *Nature*, 489 (7414):101–108, 2012.

[293] P. J. Balwierz, P. Carninci, C. O. Daub, J. Kawai, Y. Hayashizaki, W. V. Belle, C. Beisel, and E. van Nimwegen. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol*, 10(7):R79, 2009.

[294] N. Novershtern, A. Subramanian, L. N. Lawton, R. H. Mak, W. N. Haining, M. E. McConkey, N. Habib, N. Yosef, C. Y. Chang, T. Shay, G. M. Frampton, A. C. Drake, I. Leskov, B. Nilsson, F. Preffer, D. Dombkowski, J. W. Evans, T. Liefeld, J. S. Smutko, J. Chen, N. Friedman, R. A. Young, T. R. Golub, A. Regev, and B. L. Ebert. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, 144(2):296–309, 2011.

[295] N. Yosef, A. K. Shalek, J. T. Gaublomme, H. Jin, Y. Lee, A. Awasthi, C. Wu, K. Karwacz, S. Xiao, M. Jorgolli, D. Gennert, R. Satija, A. Shakya, D. Y. Lu, J. J. Trombetta, M. R. Pillai, P. J. Ratcliffe, M. L. Coleman, M. Bix, D. Tantin, H. Park, V. K. Kuchroo, and A. Regev. Dynamic regulatory network controlling TH17 cell differentiation. *Nature*, 496(7446):461–468, 2013.

[296] K. M. Summers, S. Raza, E. van Nimwegen, T. C. Freeman, and D. A. Hume. Co-expression of FBN1 with mesenchyme-specific genes in mouse cell lines: implications for phenotypic variability in Marfan syndrome. *Eur J Hum Genet*, 18(11):1209–1215, 2010.

[297] N. Aceto, N. Sausgruber, H. Brinkhaus, D. Gaidatzis, G. Martiny-Baron, G. Mazzarol, S. Confalonieri, M. Quarto, G. Hu, P. J. Balwierz, M. Pachkov, S. J. Elledge, E. van Nimwegen, M. B. Stadler, and M. Bentires-Alj. Tyrosine phosphatase SHP2 promotes breast cancer progression and maintains tumor-initiating cells via activation of key transcription factors and a positive feedback signaling loop. *Nat Med*, 18(4):529–537, 2012.

[298] E. Arner, N. Mejhert, A. Kulyté, P. J. Balwierz, M. Pachkov, M. Cormont, S. Lorente-Cebrián, A. Ehrlund, J. Laurencikiene, P. Hedén, K. Dahlman-Wright, J.-F. Tanti, Y. Hayashizaki, M. Rydén, I. Dahlman, E. van Nimwegen, C. O. Daub, and P. Arner. Adipose tissue microRNAs as regulators of CCL2 production in human obesity. *Diabetes*, 61(8):1986–1993, 2012.

[299] J. Pérez-Schindler, S. Summermatter, S. Salatino, F. Zorzato, M. Beer, P. J. Balwierz, E. van Nimwegen, J. N. Feige, J. Auwerx, and C. Handschin. The corepressor NCoR1 antagonizes pgc-1α and estrogen-related receptor α in the regulation of skeletal muscle function and oxidative metabolism. *Mol. Cell. Biol.*, 32(24):4913–4924, 2012.

[300] N. Tiwari, N. Meyer-Schaller, P. Arnold, H. Antoniadis, M. Pachkov, E. van Nimwegen, and G. Christofori. Klf4 is a transcriptional regulator of genes critical for EMT, including Jnk1 (Mapk8). *PLoS One*, 8 (2):e57329, 2013.

[301] S. J. Vervoort, A. R. Lourenço, R. van Boxtel, and P. J. Coffer. SOX4 mediates tgf-β-induced expression of mesenchymal markers during mammary cell epithelial to mesenchymal transition. *PLoS One*, 8(1): e53238, 2013.

[302] P. S. Eisele, S. Salatino, J. Sobek, M. O. Hottiger, and C. Handschin. The peroxisome proliferator-activated receptor γ coactivator $1\alpha/\beta$ (PGC-1) coactivators repress the transcriptional activity of NF-κB in skeletal muscle cells. *J Biol Chem*, 288(4):2246–2260, 2013.

[303] T. Suzuki, M. Nakano-Ikegaya, H. Yabukami-Okuda, M. de Hoon, J. Severin, S. Saga-Hatano, J. W. Shin, A. Kubosaki, C. Simon, Y. Hasegawa, Y. Hayashizaki, and H. Suzuki. Reconstruction of monocyte transcriptional regulatory network accompanies monocytic functions in human fibroblasts. *PLoS One*, 7(3):e33474, 2012.

[304] R. Hasegawa, Y. Tomaru, M. de Hoon, H. Suzuki, Y. Hayashizaki, and J. W. Shin. Identification of ZNF395 as a novel modulator of adipogenesis. *Exp Cell Res*, 2012.

[305] N. Tiwari, V. K. Tiwari, L. Waldmeier, P. J. Balwierz, P. Arnold, M. Pachkov, N. Meyer-Schaller, D. Schubeler, E. van Nimwegen, and G. Christofori. Sox4 is a master regulator of epithelial-mesenchymal

transition by controlling Ezh2 expression and epigenetic reprogramming. *Cancer Cell*, 23(6):768–783, 2013.

[306] F. Meier-Abt, E. Milani, T. Roloff, H. Brinkhaus, S. Duss, D. S. Meyer, I. Klebba, P. J. Balwierz, E. van Nimwegen, and M. Bentires-Alj. Parity induces differentiation and reduces Wnt/Notch signaling ratio and proliferation potential of basal stem/progenitor cells isolated from mouse mammary epithelium. *Breast Cancer Res*, 15(2):R36, 2013.

[307] C. J. Kuo, P. B. Conley, C. L. Hsieh, U. Francke, and G. R. Crabtree. Molecular cloning, functional expression, and chromosomal localization of mouse hepatocyte nuclear factor 1. *Proc Natl Acad Sci U S A*, 87:9838–9842, 1990.

[308] M. S. Serfas and A. L. Tyner. HNF-1 alpha and HNF-1 beta expression in mouse intestinal crypts. *Am J Physiol*, 265:G506–513, 1993.

[309] M. Pontoglio, J. Barra, M. Hadchouel, A. Doyen, C. Kress, J. P. Bach, C. Babinet, and M. Yaniv. Hepatocyte nuclear factor 1 inactivation results in hepatic dysfunction, phenylketonuria, and renal Fanconi syndrome. *Cell*, 84:575–585, 1996.

[310] H. Kawata, K. Yamada, Z. Shou, T. Mizutani, T. Yazawa, M. Yoshino, T. Sekiguchi, T. Kajitani, and K. Miyamoto. Zinc-fingers and homeoboxes (ZHX) 2, a novel member of the ZHX family, functions as a transcriptional repressor. *Biochem J*, 373(Pt 3):747–757, 2003.

[311] G. Courtois, S. Baumhueter, and G. R. Crabtree. Purified hepatocyte nuclear factor 1 interacts with a family of hepatocyte-specific promoters. *Proc Natl Acad Sci U S A*, 85(21):7937–7941, 1988.

[312] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering. STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, 37(Database issue):D412–D416, 2009.

[313] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.*, 25(1):25–29, 2000.

[314] G. Piaggio, L. Tomei, C. Toniatti, R. De Francesco, J. Gerstner, and R. Cortese. LFB1/HNF1 acts as a repressor of its own transcription. *Nucleic Acids Res*, 22(20):4284–4290, 1994.

[315] S. F. Boj, M. Parrizas, M. A. Maestro, and J. Ferrer. A transcription factor regulatory circuit in differentiated pancreatic cells. *Proc Natl Acad Sci U S A*, 98(25):14481–14486, 2001.

[316] R. Bartoov-Shifman, R. Hertz, H. Wang, C. B. Wollheim, J. Bar-Tana, and M. D. Walker. Activation of the insulin gene promoter through a direct effect of hepatocyte nuclear factor 4 alpha. *J Biol Chem*, 277 (29):25914–25919, 2002.

[317] Y. Tomaru, M. Nakanishi, H. Miura, Y. Kimura, H. Ohkawa, Y. Ohta, Y. Hayashizaki, and M. Suzuki. Identification of an inter-transcription factor regulatory network in human hepatoma cells by Matrix RNAi. *Nucleic Acids Res*, 37(4):1049–1060, 2009.

[318] I. M. Bochkis, J. Schug, D. Z. Ye, S. Kurinna, S. A. Stratton, M. C. Barton, and K. H. Kaestner. Genome-wide location analysis reveals distinct transcriptional circuitry by paralogous regulators Foxa1 and Foxa2. *PLoS Genet*, 8(6):e1002770, 2012.

[319] X. Molero, E. C. Vaquero, M. Flandez, A. M. Gonzalez, M. A. Ortiz, E. Cibrian-Uhalte, J. M. Servitja, A. Merlos, N. Juanpere, M. Massumi, A. Skoudy, R. Macdonald, J. Ferrer, and F. X. Real. Gene expression dynamics after murine pancreatitis unveils novel roles for Hnf1 α in acinar cell homeostasis. *Gut*, 61(8):1187–1196, 2012.

[320] E. Bolcun-Filas, L. A. Bannister, A. Barash, K. J. Schimenti, S. A. Hartford, J. J. Eppig, M. A. Handel, L. Shen, and J. C. Schimenti. A-MYB (MYBL1) transcription factor is a master regulator of male meiosis. *Development*, 138(15):3319–3330, 2011.

[321] A. Toscani, R. V. Mettus, R. Coupland, H. Simpkins, J. Litvin, J. Orth, K. S. Hatton, and E. P. Reddy. Arrest of spermatogenesis and defective breast development in mice lacking A-myb. *Nature*, 386(6626):713–717, 1997.

[322] G. C. Horvath, M. K. Kistler, and W. S. Kistler. RFX2 is a candidate downstream amplifier of A-MYB regulation in mouse spermatogenesis. *BMC Dev Biol*, 9:63, 2009.

[323] M. R. Campanero, M. Armstrong, and E. Flemington. Distinct cellular factors regulate the c-myb promoter through its E2F element. *Mol Cell Biol*, 19(12):8442–8450, 1999.

[324] E. D. Ponomarev, T. Veremeyko, N. Barteneva, A. M. Krichevsky, and H. L. Weiner. MicroRNA-124 promotes microglia quiescence and suppresses EAE by deactivating macrophages via the C/EBP-α-PU.1 pathway. *Nat Med*, 17(1):64–70, 2011.

[325] X. B. Shi, L. Xue, A. H. Ma, C. G. Tepper, R. Gandour-Edwards, H. J. Kung, and R. W. deVere White. Tumor suppressive miR-124 targets androgen receptor and inhibits proliferation of prostate cancer cells. *Oncogene*, 32(35):4130–4138, 2013.

[326] Y. J. Liang, Q. Y. Wang, C. X. Zhou, Q. Q. Yin, M. He, X. T. Yu, D. X. Cao, G. Q. Chen, J. R. He, and Q. Zhao. MiR-124 targets Slug

to regulate epithelial-mesenchymal transition and metastasis of breast cancer. *Carcinogenesis*, 34(3):713–722, 2013.

[327] A. M. Cuervo and J. F. Dice. When lysosomes get old. *Exp Gerontol*, 35(2):119–131, 2000.

[328] D. J. Kurz, S. Decary, Y. Hong, and J. D. Erusalimsky. Senescence-associated (beta)-galactosidase reflects an increase in lysosomal mass during replicative ageing of human endothelial cells. *J Cell Sci*, 113 (Pt 20):3613–3622, 2000.

[329] D. C. Rubinsztein, G. Marino, and G. Kroemer. Autophagy and aging. *Cell*, 146(5):682–695, 2011.

[330] K. Okamoto, T. Kakuma, S. Fukuchi, T. Masaki, T. Sakata, and H. Yoshimatsu. Sterol regulatory element binding protein (SREBP)-1 expression in brain is affected by age but not by hormones or metabolic changes. *Brain Res*, 1081(1):19–27, 2006.

[331] Y. M. Kim, H. T. Shin, Y. H. Seo, H. O. Byun, S. H. Yoon, I. K. Lee, D. H. Hyun, H. Y. Chung, and G. Yoon. Sterol regulatory element-binding protein (SREBP)-1-mediated lipogenesis is involved in cell senescence. *J Biol Chem*, 285(38):29069–29077, 2010.

[332] Y. K. Seo, T. I. Jeon, H. K. Chong, J. Biesinger, X. Xie, and T. F. Osborne. Genome-wide localization of SREBP-2 in hepatic chromatin predicts a role in autophagy. *Cell Metab*, 13(4):367–375, 2011.

[333] T. R. Peterson, S. S. Sengupta, T. E. Harris, A. E. Carmack, S. A. Kang, E. Balderas, D. A. Guertin, K. L. Madden, A. E. Carpenter, B. N. Finck, and D. M. Sabatini. mTOR complex 1 regulates lipin 1 localization to control the SREBP pathway. *Cell*, 146(3):408–420, 2011.

[334] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101(16): 6062–6067, 2004.

[335] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. V. de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet*, 24(3):227–235, 2000.

[336] Y. Wada, Y. Ohta, M. Xu, S. Tsutsumi, T. Minami, K. Inoue, D. Komura, J. Kitakami, N. Oshida, A. Papantonis, A. Izumi, M. Kobayashi, H. Meguro, Y. Kanki, I. Mimura, K. Yamamoto, C. Mataki, T. Hamakubo, K. Shirahige, H. Aburatani, H. Kimura,

T. Kodama, P. R. Cook, and S. Ihara. A wave of nascent transcription on activated human genes. *Proc Natl Acad Sci U S A*, 106(43): 18357–18361, 2009.

[337] K. Inoue, M. Kobayashi, K. Yano, M. Miura, A. Izumi, C. Mataki, T. Doi, T. Hamakubo, P. C. Reid, D. A. Hume, M. Yoshida, W. C. Aird, T. Kodama, and T. Minami. Histone deacetylase inhibitor reduces monocyte adhesion to endothelium through the suppression of vascular cell adhesion molecule-1 expression. *Arterioscler Thromb Vasc Biol*, 26(12):2652–2659, 2006.

[338] S. Kempe, H. Kestler, A. Lasar, and T. Wirth. NF-kappaB controls the global pro-inflammatory response in endothelial cells: evidence for the regulation of a pro-atherogenic program. *Nucleic Acids Res*, 33(16): 5308–5319, 2005.

[339] H. Harada, E. Takahashi, S. Itoh, K. Harada, T. A. Hori, and T. Taniguchi. Structure and regulation of the human interferon regulatory factor 1 (IRF-1) and IRF-2 genes: implications for a gene network in the interferon system. *Mol Cell Biol*, 14(2):1500–1509, 1994.

[340] R. M. Ten, V. Blank, O. Le Bail, P. Kourilsky, and A. IsraÃ«l. Two factors, IRF1 and KBF1/NF-kappa B, cooperate during induction of MHC class I gene expression by interferon alpha beta or Newcastle disease virus. *C R Acad Sci III*, 316(5):496–501, 1993.

[341] G. Martins and K. Calame. Regulation and functions of Blimp-1 in T and B lymphocytes. *Annu Rev Immunol*, 26:133–169, 2008.

[342] U. Seifert, L. P. Bialy, F. Ebstein, D. Bech-Otschir, A. Voigt, F. Schröter, T. Prozorovski, N. Lange, J. Steffen, M. Rieger, U. Kuckelkorn, O. Aktas, P.-M. Kloetzel, and E. Krüger. Immunoproteasomes preserve protein homeostasis upon interferon-induced oxidative stress. *Cell*, 142(4):613–624, 2010.

[343] M. Kasowski, F. Grubert, C. Heffelfinger, M. Hariharan, A. Asabere, S. M. Waszak, L. Habegger, J. Rozowsky, M. Shi, A. E. Urban, M. Y. Hong, K. J. Karczewski, W. Huber, S. M. Weissman, M. B. Gerstein, J. O. Korbel, and M. Snyder. Variation in transcription factor binding among humans. *Science*, 328(5975):232–235, 2010.

[344] L. H. Glimcher. XBP1: the last two decades. *Ann Rheum Dis*, 69 Suppl 1:i67–i71, 2010.

[345] P. S. Gargalovic, N. M. Gharavi, M. J. Clark, J. Pagnon, W.-P. Yang, A. He, A. Truong, T. Baruch-Oren, J. A. Berliner, T. G. Kirchgessner, and A. J. Lusis. The unfolded protein response is an important regulator of inflammatory genes in endothelial cells. *Arterioscler Thromb Vasc Biol*, 26(11):2490–2496, 2006.

[346] M. Civelek, E. Manduchi, R. J. Riley, C. J. Stoeckert, Jr, and P. F. Davies. Chronic endoplasmic reticulum stress activates unfolded protein response in arterial endothelium in regions of susceptibility to atherosclerosis. *Circ Res*, 105(5):453–461, 2009.

[347] M. Kitamura. Control of NF-kappaB and inflammation by the unfolded protein response. *Int. Rev. Immunol.*, 30:4–15, 2011.

[348] A. Kaser, A.-H. Lee, A. Franke, J. N. Glickman, S. Zeissig, H. Tilg, E. E. S. Nieuwenhuis, D. E. Higgins, S. Schreiber, L. H. Glimcher, and R. S. Blumberg. XBP1 links ER stress to intestinal inflammation and confers genetic risk for human inflammatory bowel disease. *Cell*, 134(5):743–756, 2008.

[349] J. Li, J. J. Wang, and S. X. Zhang. Preconditioning with endoplasmic reticulum stress mitigates retinal endothelial inflammation via activation of X-box binding protein 1. *J Biol Chem*, 286(6):4912–4921, 2011.

[350] H. Yoshida, T. Matsui, A. Yamamoto, T. Okada, and K. Mori. XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. *Cell*, 107(7):881–891, 2001.

[351] M. Calfon, H. Zeng, F. Urano, J. H. Till, S. R. Hubbard, H. P. Harding, S. G. Clark, and D. Ron. IRE1 couples endoplasmic reticulum load to secretory capacity by processing the XBP-1 mRNA. *Nature*, 415 (6867):92–96, 2002.

[352] A. J. Ross, L. A. Dailey, L. E. Brighton, and R. B. Devlin. Transcriptional profiling of mucociliary differentiation in human airway epithelial cells. *Am J Respir Cell Mol Biol*, 37(2):169–185, 2007.

[353] E. Bonnafe, M. Touka, A. AitLounis, D. Baas, E. Barras, C. Ucla, A. Moreau, F. Flamant, R. Dubruille, P. Couble, J. Collignon, B. Durand, and W. Reith. The transcription factor RFX3 directs nodal cilium development and left-right asymmetry specification. *Mol Cell Biol*, 24 (10):4417–4427, 2004.

[354] L. El Zein, A. Ait-Lounis, L. Morlé, J. Thomas, B. Chhin, N. Spassky, W. Reith, and B. Durand. RFX3 governs growth and beating efficiency of motile cilia in mouse and controls the expression of genes involved in human ciliopathies. *J. Cell Sci.*, 122(Pt 17):3180–3189, 2009.

[355] C. Scheel, E. N. Eaton, S. H.-J. Li, C. L. Chaffer, F. Reinhardt, K.-J. Kah, G. Bell, W. Guo, J. Rubin, A. L. Richardson, and R. A. Weinberg. Paracrine and autocrine signals induce and maintain mesenchymal and stem cell states in the breast. *Cell*, 145(6):926–940, 2011.

[356] K. Polyak and R. A. Weinberg. Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nat Rev Cancer*, 9(4):265–273, 2009.

[357] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011.

[358] L. W. Yuan and J. E. Gambee. Histone acetylation by p300 is involved in CREB-mediated transcription on chromatin. *Biochim Biophys Acta*, 1541(3):161–169, 2001.

[359] K. Masternak, N. Peyraud, M. Krawczyk, E. Barras, and W. Reith. Chromatin remodeling and extragenic transcription at the MHC class II locus control region. *Nat Immunol*, 4(2):132–137, 2003.

[360] Q. Gan, P. Thiebaud, N. Theze, L. Jin, G. Xu, P. Grant, and G. K. Owens. WD repeat-containing protein 5, a ubiquitously expressed histone methyltransferase adaptor protein, regulates smooth muscle cell-selective gene activation through interaction with pituitary homeobox 2. *J Biol Chem*, 286(24):21853–21864, 2011.

[361] K. O. Kizer, H. P. Phatnani, Y. Shibata, H. Hall, A. L. Greenleaf, and B. D. Strahl. A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation. *Mol Cell Biol*, 25(8):3305–3316, 2005.

[362] W. Yuan, J. Xie, C. Long, H. Erdjument-Bromage, X. Ding, Y. Zheng, P. Tempst, S. Chen, B. Zhu, and D. Reinberg. 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 (23):15701–15707, 2009.

[363] Q. Cui, Z. Yu, E. Purisima, and E. Wang. Principles of microRNA regulation of a human cellular signaling network. *Mol Syst Biol*, 2:46, 2006.

[364] Y. Zhou, J. Ferguson, J. Chang, and Y. Kluger. Inter- and intra-combinatorial regulation by transcription factors and microRNAs. *BMC Genomics*, 8:396, 2007.

[365] F. Gao, B. C. Foat, and H. J. Bussemaker. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, 5:31, 2004.

[366] D. C. Bauer, F. A. Buske, and T. L. Bailey. Dual-functioning transcription factors in the developmental gene network of Drosophila melanogaster. *BMC Bioinformatics*, 11(1):366, 2010.

[367] M. L. Bulyk. DNA microarray technologies for measuring protein-DNA interactions. *Curr Opin Biotechnol*, 17(4):422–430, 2006.

[368] N. D. Heintzman, G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. K. Lee, R. K. Stuart, C. W. Ching, K. A. Ching, J. E. Antosiewicz-Bourget, H. Liu, X. Zhang, R. D. Green, V. V. Lobanenkov, R. Stewart, J. A. Thomson, G. E. Crawford, M. Kellis, and B. Ren. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, 2009.

[369] M. B. Stadler, R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Scholer, E. van Nimwegen, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, V. K. Tiwari, and D. Schubeler. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480(7378):490–495, 2011.

[370] Y. Shen, F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, and B. Ren. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–120, 2012.

[371] M. de Hoon and Y. Hayashizaki. Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. *Biotechniques*, 44(5):627–8, 630, 632, 2008.

[372] C. Notredame, D. G. Higgins, and J. Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–217, 2000.

[373] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol.*, 11(10):R106, 2010.

[374] D. H. Nguyen and P. D'haeseleer. Deciphering principles of transcription regulation in eukaryotic genomes. *Mol Syst Biol*, 2:2006.0012, 2006.

[375] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, pages 267–288, 1996.

[376] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2): 301–320, 2005.

[377] M. T. Dill, Z. Makowska, G. Trincucci, A. J. Gruber, J. E. Vogt, M. Filipowicz, D. Calabrese, I. Krol, D. T. Lau, L. Terracciano, E. van Nimwegen, V. Roth, and M. H. Heim. Pegylated ifn-α regulates hepatic gene expression through transient Jak/STAT activation. *J. Clin. Invest.*, 124(4):1568–1581, 2014.

[378] D. B. Stetson and R. Medzhitov. Type I interferons in host defense. *Immunity*, 25(3):373–381, 2006.

[379] A. J. Sadler and B. R. G. Williams. Interferon-inducible antiviral effectors. *Nat. Rev. Immunol.*, 8(7):559–568, 2008.

[380] M. M. Song and K. Shuai. The suppressor of cytokine signaling (SOCS) 1 and SOCS3 but not SOCS2 proteins inhibit interferon-mediated antiviral and antiproliferative activities. *J. Biol. Chem.*, 273 (52):35056–35062, 1998.

[381] A. C. Larner, A. Chaudhuri, and J. E. Darnell, Jr. Transcriptional induction by interferon. new protein(s) determine the extent and length of the induction. *J. Biol. Chem.*, 261(1):453–459, 1986.

[382] M. Sarasin-Filipowicz, X. Wang, M. Yan, F. H. T. Duong, V. Poli, D. J. Hilton, D.-E. Zhang, and M. H. Heim. Alpha interferon induces long-lasting refractoriness of JAK-STAT signaling in the mouse liver through induction of USP18/UBP43. *Mol. Cell. Biol.*, 29(17):4841–4851, 2009.

[383] J. H. Hoofnagle, K. D. Mullen, D. B. Jones, V. Rustgi, A. Di Bisceglie, M. Peters, J. G. Waggoner, Y. Park, and E. A. Jones. Treatment of chronic non-A,non-B hepatitis with recombinant human alpha interferon. a preliminary report. *N. Engl. J. Med.*, 315(25):1575–1578, 1986.

[384] G. M. Lauer and B. D. Walker. Hepatitis C virus infection. *N. Engl. J. Med.*, 345(1):41–52, 2001.

[385] M. W. Fried, M. L. Shiffman, K. R. Reddy, C. Smith, G. Marinos, F. L. Gonçales, Jr, D. Häussinger, M. Diago, G. Carosi, D. Dhumeaux, A. Craxi, A. Lin, J. Hoffman, and J. Yu. Peginterferon alfa-2a plus ribavirin for chronic hepatitis C virus infection. *N. Engl. J. Med.*, 347 (13):975–982, 2002.

[386] M. P. Manns, J. G. McHutchison, S. C. Gordon, V. K. Rustgi, M. Shiffman, R. Reindollar, Z. D. Goodman, K. Koury, M. Ling, and J. K. Albrecht. Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis c: a randomised trial. *Lancet*, 358(9286):958–965, 2001.

[387] S. Zeuzem, S. V. Feinman, J. Rasenack, E. J. Heathcote, M. Y. Lai, E. Gane, J. O'Grady, J. Reichen, M. Diago, A. Lin, J. Hoffman, and M. J. Brunda. Peginterferon alfa-2a in patients with chronic hepatitis C. *N. Engl. J. Med.*, 343(23):1666–1672, 2000.

[388] K. L. Lindsay, C. Trepo, T. Heintges, M. L. Shiffman, S. C. Gordon, J. C. Hoefs, E. R. Schiff, Z. D. Goodman, M. Laughlin, R. Yao, J. K. Albrecht, and Hepatitis Interventional Therapy Group. A randomized, double-blind trial comparing pegylated interferon alfa-2b to interferon alfa-2b as initial treatment for chronic hepatitis C. *Hepatology*, 34(2): 395–403, 2001.

[389] M. Sarasin-Filipowicz, E. J. Oakeley, F. H. T. Duong, V. Christen, L. Terracciano, W. Filipowicz, and M. H. Heim. Interferon signaling and treatment outcome in chronic hepatitis C. *Proc. Natl. Acad. Sci. U. S. A.*, 105(19):7034–7039, 2008.

[390] L. Chen, I. Borozan, J. Feld, J. Sun, L.-L. Tannis, C. Coltescu, J. Heathcote, A. M. Edwards, and I. D. McGilvray. Hepatic gene expression discriminates responders and nonresponders in treatment of chronic hepatitis C viral infection. *Gastroenterology*, 128(5):1437–1444, 2005.

[391] T. Asselah, I. Bieche, S. Narguet, A. Sabbagh, I. Laurendeau, M.-P. Ripault, N. Boyer, M. Martinot-Peignoux, D. Valla, M. Vidaud, and P. Marcellin. Liver gene expression signature to predict response to pegylated interferon plus ribavirin combination therapy in patients with chronic hepatitis C. *Gut*, 57(4):516–524, 2008.

[392] J. J. Feld, S. Nanda, Y. Huang, W. Chen, M. Cam, S. N. Pusek, L. M. Schweigler, D. Theodore, S. L. Zacks, T. J. Liang, and M. W. Fried. Hepatic gene expression during treatment with peginterferon and ribavirin: Identifying molecular pathways for treatment response. *Hepatology*, 46(5):1548–1563, 2007.

[393] M. T. Dill, F. H. T. Duong, J. E. Vogt, S. Bibert, P.-Y. Bochud, L. Terracciano, A. Papassotiropoulos, V. Roth, and M. H. Heim. Interferon-induced gene expression is a stronger predictor of treatment response than IL28B genotype in patients with hepatitis C. *Gastroenterology*, 140(3):1021–1031, 2011.

[394] D. T.-Y. Lau, A. Negash, J. Chen, N. Crochet, M. Sinha, Y. Zhang, J. Guedj, S. Holder, T. Saito, S. M. Lemon, B. A. Luxon, A. S. Perelson, and M. Gale, Jr. Innate immune tolerance and the role of kupffer cells in differential responses to interferon therapy among patients with HCV genotype 1 infection. *Gastroenterology*, 144(2):402–413.e12, 2013.

[395] G. R. Foster. Pegylated interferons for the treatment of chronic hepatitis c: pharmacological and clinical differences between peginterferon-alpha-2a and peginterferon-alpha-2b. *Drugs*, 70(2):147–165, 2010.

[396] M. Silva, J. Poo, F. Wagner, M. Jackson, D. Cutler, M. Grace, R. Bordens, C. Cullen, J. Harvey, and M. Laughlin. A randomised trial to compare the pharmacokinetic, pharmacodynamic, and antiviral effects of peginterferon alfa-2b and peginterferon alfa-2a in patients with chronic hepatitis C (COMPARE). *J. Hepatol.*, 45(2):204–213, 2006.

[397] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple combinations

of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, 38(4): 576–589, 2010.

[398] T. Tamura, H. Yanai, D. Savitsky, and T. Taniguchi. The IRF family transcription factors in immunity and oncogenesis. *Annu. Rev. Immunol.*, 26:535–584, 2008.

[399] K. Hoshino, T. Sugiyama, M. Matsumoto, T. Tanaka, M. Saito, H. Hemmi, O. Ohara, S. Akira, and T. Kaisho. IkappaB kinase-alpha is critical for interferon-alpha production induced by toll-like receptors 7 and 9. *Nature*, 440(7086):949–953, 2006.

[400] H. Cheon and G. R. Stark. Unphosphorylated STAT1 prolongs the expression of interferon-induced immune regulatory genes. *Proc. Natl. Acad. Sci. U. S. A.*, 106(23):9373–9378, 2009.

[401] M. Müller, C. Laxton, J. Briscoe, C. Schindler, T. Improta, J. E. Darnell, Jr, G. R. Stark, and I. M. Kerr. Complementation of a mutant cell line: central role of the 91 kda polypeptide of ISGF3 in the interferon-alpha and -gamma signal transduction pathways. *EMBO J.*, 12(11): 4221–4228, 1993.

[402] S. Hao and D. Baltimore. The stability of mRNA influences the temporal order of the induction of genes encoding inflammatory molecules. *Nat. Immunol.*, 10(3):281–288, 2009.

[403] X. Wu and G. Brewer. The regulation of mRNA stability in mammalian cells: 2.0. *Gene*, 500(1):10–21, 2012.

[404] M. E. Cramp, S. Rossol, S. Chokshi, P. Carucci, R. Williams, and N. V. Naoumov. Hepatitis C virus-specific t-cell reactivity during interferon and ribavirin treatment in chronic hepatitis C. *Gastroenterology*, 118 (2):346–355, 2000.

[405] S. M. Kamal, J. Fehr, B. Roesler, T. Peters, and J. W. Rasenack. Peginterferon alone or with ribavirin enhances HCV-specific CD4 t-helper 1 responses in patients with chronic hepatitis C. *Gastroenterology*, 123 (4):1070–1083, 2002.

[406] M. A. A. Claassen, R. J. de Knegt, D. Turgut, Z. M. A. Groothuismink, H. L. A. Janssen, and A. Boonstra. Negative regulation of hepatitis C virus specific immunity is highly heterogeneous and modulated by pegylated interferon-alpha/ribavirin therapy. *PLoS One*, 7(11):e49389, 2012.

[407] E. Barnes, G. Harcourt, D. Brown, M. Lucas, R. Phillips, G. Dusheiko, and P. Klenerman. The dynamics of t-lymphocyte responses during combination therapy for chronic hepatitis C virus infection. *Hepatology*, 36(3):743–754, 2002.

[408] J. H. Aberle, G. Perstinger, L. Weseslindtner, U. Sinzinger, C. Gurguta, P. Steindl-Munda, M. Kundi, T. Popow-Kraupp, P. Ferenci, and H. Holzmann. CD4+ T cell responses in patients with chronic hepatitis C undergoing peginterferon/ribavirin therapy correlate with faster, but not sustained, viral clearance. *J. Infect. Dis.*, 195(9):1315–1319, 2007.

[409] R. E. Lanford, B. Guerra, H. Lee, D. Chavez, K. M. Brasky, and C. B. Bigger. Genomic response to interferon-alpha in chimpanzees: implications of rapid downregulation for hepatitis C kinetics. *Hepatology*, 43(5):961–972, 2006.

[410] O. A. Malakhova, K. I. Kim, J.-K. Luo, W. Zou, K. G. S. Kumar, S. Y. Fuchs, K. Shuai, and D.-E. Zhang. UBP43 is a novel regulator of interferon signaling independent of its ISG15 isopeptidase activity. *EMBO J.*, 25(11):2358–2367, 2006.

[411] J. E. Fenner, R. Starr, A. L. Cornish, J.-G. Zhang, D. Metcalf, R. D. Schreiber, K. Sheehan, D. J. Hilton, W. S. Alexander, and P. J. Hertzog. Suppressor of cytokine signaling 1 regulates the immune response to infection by a unique inhibition of type I interferon activity. *Nat. Immunol.*, 7(1):33–39, 2006.

[412] J. G. McHutchison, E. J. Lawitz, M. L. Shiffman, A. J. Muir, G. W. Galler, J. McCone, L. M. Nyberg, W. M. Lee, R. H. Ghalib, E. R. Schiff, J. S. Galati, B. R. Bacon, M. N. Davis, P. Mukhopadhyay, K. Koury, S. Noviello, L. D. Pedicone, C. A. Brass, J. K. Albrecht, M. S. Sulkowski, and IDEAL Study Team. Peginterferon alfa-2b or alfa-2a with ribavirin for treatment of hepatitis C infection. *N. Engl. J. Med.*, 361(6):580–593, 2009.

[413] M. H. Heim. Interferon signaling. In *Signaling Pathways in Liver Diseases*, pages 189–200. Springer Berlin Heidelberg, 2010.

[414] J. W. Schoggins, S. J. Wilson, M. Panis, M. Y. Murphy, C. T. Jones, P. Bieniasz, and C. M. Rice. A diverse range of gene products are effectors of the type I interferon antiviral response. *Nature*, 472(7344): 481–485, 2011.

[415] P. Metz, E. Dazert, A. Ruggieri, J. Mazur, L. Kaderali, A. Kaul, U. Zeuge, M. P. Windisch, M. Trippler, V. Lohmann, M. Binder, M. Frese, and R. Bartenschlager. Identification of type I and type II interferon-induced effectors controlling hepatitis C virus replication. *Hepatology*, 56(6):2082–2093, 2012.

[416] X.-Y. Liu, W. Chen, B. Wei, Y.-F. Shan, and C. Wang. IFN-induced TPR protein IFIT3 potentiates antiviral signaling by bridging MAVS and TBK1. *J. Immunol.*, 187(5):2559–2568, 2011.

[417] M. Yoneyama, M. Kikuchi, K. Matsumoto, T. Imaizumi, M. Miyagishi, K. Taira, E. Foy, Y.-M. Loo, M. Gale, Jr, S. Akira, S. Yonehara,

A. Kato, and T. Fujita. Shared and unique functions of the DExD/H-box helicases RIG-I, MDA5, and LGP2 in antiviral innate immunity. *J. Immunol.*, 175(5):2851–2858, 2005.

[418] G. A. Versteeg, R. Rajsbaum, M. T. Sánchez-Aparicio, A. M. Maestre, J. Valdiviezo, M. Shi, K.-S. Inn, A. Fernandez-Sesma, J. Jung, and A. García-Sastre. The e3-ligase TRIM family of proteins regulates signaling pathways triggered by innate immune pattern-recognition receptors. *Immunity*, 38(2):384–398, 2013.

[419] C. Wilkins, J. Woodward, D. T.-Y. Lau, A. Barnes, M. Joyce, N. McFarlane, J. A. McKeating, D. L. Tyrrell, and M. Gale, Jr. IFITM1 is a tight junction protein that inhibits hepatitis C virus entry. *Hepatology*, 57(2):461–469, 2013.

[420] A. R. Everitt, S. Clare, T. Pertel, S. P. John, R. S. Wash, S. E. Smith, C. R. Chin, E. M. Feeley, J. S. Sims, D. J. Adams, H. M. Wise, L. Kane, D. Goulding, P. Digard, V. Anttila, J. K. Baillie, T. S. Walsh, D. A. Hume, A. Palotie, Y. Xue, V. Colonna, C. Tyler-Smith, J. Dunning, S. B. Gordon, GenISIS Investigators, MOSAIC Investigators, R. L. Smyth, P. J. Openshaw, G. Dougan, A. L. Brass, and P. Kellam. IFITM3 restricts the morbidity and mortality associated with influenza. *Nature*, 484(7395):519–523, 2012.

[421] F. Terenzi, D. J. Hui, W. C. Merrick, and G. C. Sen. Distinct induction patterns and functions of two closely related interferon-inducible human genes, ISG54 and ISG56. *J. Biol. Chem.*, 281(45):34064–34071, 2006.

[422] I. Marié, J. E. Durbin, and D. E. Levy. Differential viral induction of distinct interferon-alpha genes by positive feedback through interferon regulatory factor-7. *EMBO J.*, 17(22):6660–6669, 1998.

[423] I. Marié, E. Smith, A. Prakash, and D. E. Levy. Phosphorylation-induced dimerization of interferon regulatory factor 7 unmasks DNA binding and a bipartite transactivation domain. *Mol. Cell. Biol.*, 20(23): 8803–8814, 2000.

[424] S. Wieland, Z. Makowska, B. Campana, D. Calabrese, M. T. Dill, J. Chung, F. V. Chisari, and M. H. Heim. Simultaneous detection of hepatitis C virus and interferon stimulated gene expression in infected human liver. *Hepatology*, 59(6):2121–2130, 2014.

[425] M. Pachkov, P. J. Balwierz, P. Arnold, E. Ozonov, and E. van Nimwegen. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res.*, 41(Database issue): D214–20, 2013.

[426] I. Mikaelian and A. Sergeant. A general and fast method to generate multiple site directed mutations. *Nucleic Acids Res.*, 20(2):376, 1992.

[427] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5(10):R80, 2004.

[428] W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[429] A. R. Gruber, G. Martin, P. Müller, A. Schmidt, A. J. Gruber, R. Gumienny, N. Mittal, R. Jayachandran, J. Pieters, W. Keller, E. van Nimwegen, and M. Zavolan. Global 3' UTR shortening has a limited effect on protein abundance in proliferating T cells. *Nat. Commun.*, 5:5465, 2014.

[430] Z. Ji, J. Y. Lee, Z. Pan, B. Jiang, and B. Tian. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc. Natl. Acad. Sci. U. S. A.*, 106(17):7028–7033, 2009.

[431] Z. Ji and B. Tian. Reprogramming of 3' untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PLoS One*, 4(12):e8419, 2009.

[432] E. Huntzinger and E. Izaurralde. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat. Rev. Genet.*, 12(2):99–110, 2011.

[433] A. Castello, B. Fischer, K. Eichelbaum, R. Horos, B. M. Beckmann, C. Strein, N. E. Davey, D. T. Humphreys, T. Preiss, L. M. Steinmetz, J. Krijgsveld, and M. W. Hentze. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, 149(6):1393–1406, 2012.

[434] A. G. Baltz, M. Munschauer, B. Schwanhäusser, A. Vasile, Y. Murakawa, M. Schueler, N. Youngs, D. Penfold-Brown, K. Drew, M. Milek, E. Wyler, R. Bonneau, M. Selbach, C. Dieterich, and M. Landthaler. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell*, 46(5):674–690, 2012.

[435] C. M. Brennan and J. A. Steitz. HuR and mRNA stability. *Cell. Mol. Life Sci.*, 58(2):266–277, 2001.

[436] M. Baou, J. D. Norton, and J. J. Murphy. AU-rich RNA binding proteins in hematopoiesis and leukemogenesis. *Blood*, 118(22):5732–5740, 2011.

[437] I. Gupta, S. Clauder-Münster, B. Klaus, A. I. Järvelin, R. S. Aiyar, V. Benes, S. Wilkening, W. Huber, V. Pelechano, and L. M. Steinmetz.

Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA-protein interactions. *Mol. Syst. Biol.*, 10:719, 2014.

[438] S. Anders, A. Reyes, and W. Huber. Detecting differential usage of exons from RNA-seq data. *Genome Res.*, 22(10):2008–2017, 2012.

[439] N. Bonnefoy-Berard, A. Aouacheria, C. Verschelde, L. Quemeneur, A. Marçais, and J. Marvel. Control of proliferation by bcl-2 family members. *Biochim. Biophys. Acta*, 1644(2-3):159–168, 2004.

[440] A. Y. Wen, K. M. Sakamoto, and L. S. Miller. The role of the transcription factor CREB in immune function. *J. Immunol.*, 185(11):6413–6419, 2010.

[441] D. Y. Lee, B. K. Choi, D. G. Lee, Y. H. Kim, C. H. Kim, S. J. Lee, and B. S. Kwon. 4-1BB signaling activates the t cell factor 1 effector/β-catenin pathway with delayed kinetics via ERK signaling and delayed PI3K/AKT activation to promote the proliferation of CD8+ T cells. *PLoS One*, 8(7):e69677, 2013.

[442] M. Khorshid, C. Rodak, and M. Zavolan. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res.*, 39(Database issue):D245–52, 2011.

[443] D. D. Billadeau, J. C. Nolz, and T. S. Gomez. Regulation of t-cell activation by the cytoskeleton. *Nat. Rev. Immunol.*, 7(2):131–143, 2007.

[444] J. C. Nolz, T. S. Gomez, P. Zhu, S. Li, R. B. Medeiros, Y. Shimizu, J. K. Burkhardt, B. D. Freedman, and D. D. Billadeau. The WAVE2 complex regulates actin cytoskeletal reorganization and CRAC-mediated calcium entry during T cell activation. *Curr. Biol.*, 16(1):24–34, 2006.

[445] J. H. Hartwig, M. Thelen, A. Rosen, P. A. Janmey, A. C. Nairn, and A. Aderem. MARCKS is an actin filament crosslinking protein regulated by protein kinase C and calcium-calmodulin. *Nature*, 356(6370): 618–622, 1992.

[446] J. B. Zuchero, A. S. Coutts, M. E. Quinlan, N. B. L. Thangue, and R. D. Mullins. p53-cofactor JMY is a multifunctional actin nucleation factor. *Nat. Cell Biol.*, 11(4):451–459, 2009.

[447] L. Li, J.-Y. Shi, G.-Q. Zhu, and B. Shi. MiR-17-92 cluster regulates cell proliferation and collagen synthesis by targeting TGFB pathway in mouse palatal mesenchymal cells. *J. Cell. Biochem.*, 113(4):1235–1244, 2012.

[448] P. Zhang, C. Zheng, H. Ye, Y. Teng, B. Zheng, X. Yang, and J. Zhang. MicroRNA-365 inhibits vascular smooth muscle cell proliferation through targeting cyclin D1. *Int. J. Med. Sci.*, 11(8):765–770, 2014.

[449] H. Wu, M. Huang, P. Cao, T. Wang, Y. Shu, and P. Liu. MiR-135a targets JAK2 and inhibits gastric cancer cell proliferation. *Cancer Biol. Ther.*, 13(5):281–288, 2012.

[450] W. Shen, M. Song, J. Liu, G. Qiu, T. Li, Y. Hu, and H. Liu. MiR-26a promotes ovarian cancer proliferation and tumorigenesis. *PLoS One*, 9(1):e86871, 2014.

[451] Y. Liao and B. Lönnerdal. Global microRNA characterization reveals that mir-103 is involved in IGF-1 stimulated mouse intestinal cell proliferation. *PLoS One*, 5(9):e12976, 2010.

[452] D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A. Smith, C. A. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. F. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, and T. R. Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, 2013.

[453] M. Jenal, R. Elkon, F. Loayza-Puch, G. van Haaften, U. Kühn, F. M. Menzies, J. A. F. Oude Vrielink, A. J. Bos, J. Drost, K. Rooijers, D. C. Rubinsztein, and R. Agami. The poly(a)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. *Cell*, 149 (3):538–553, 2012.

[454] M. V. Lee, S. E. Topper, S. L. Hubler, J. Hose, C. D. Wenger, J. J. Coon, and A. P. Gasch. A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Mol. Syst. Biol.*, 7:514, 2011.

[455] M. Beck, A. Schmidt, J. Malmstroem, M. Claassen, A. Ori, A. Szymborska, F. Herzog, O. Rinner, J. Ellenberg, and R. Aebersold. The quantitative proteome of a human cell line. *Mol. Syst. Biol.*, 7:549, 2011.

[456] L. Ting, R. Rad, S. P. Gygi, and W. Haas. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods*, 8 (11):937–940, 2011.

[457] P. Mertins, N. D. Udeshi, K. R. Clauser, D. R. Mani, J. Patel, S.-E. Ong, J. D. Jaffe, and S. A. Carr. iTRAQ labeling is superior to mTRAQ for quantitative global proteomics and phosphoproteomics. *Mol. Cell. Proteomics*, 11(6):M111.014423, 2012.

[458] A. E. Merrill, A. S. Hebert, M. E. MacGilvray, C. M. Rose, D. J. Bailey, J. C. Bradley, W. W. Wood, M. El Masri, M. S. Westphall, A. P. Gasch, and J. J. Coon. NeuCode labels for relative protein quantification. *Mol. Cell. Proteomics*, 13(9):2503–2512, 2014.

[459] M. Wilhelm, J. Schlegl, H. Hahne, A. Moghaddas Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J.-H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, and B. Kuster. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502): 582–587, 2014.

[460] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, 42(Database issue):D7–17, 2014.

[461] Q. Pan, M. A. Bakowski, Q. Morris, W. Zhang, B. J. Frey, T. R. Hughes, and B. J. Blencowe. Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends Genet.*, 21(2): 73–77, 2005.

[462] D. Kaida, M. G. Berg, I. Younis, M. Kasim, L. N. Singh, L. Wan, and G. Dreyfuss. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*, 468(7324):664–668, 2010.

[463] M. G. Berg, L. N. Singh, I. Younis, Q. Liu, A. M. Pinto, D. Kaida, Z. Zhang, S. Cho, S. Sherrill-Mix, L. Wan, and G. Dreyfuss. U1 snRNP determines mRNA length and regulates isoform expression. *Cell*, 150 (1):53–64, 2012.

[464] N. Kedersha and P. Anderson. Mammalian stress granules and processing bodies. *Methods Enzymol.*, 431:61–81, 2007.

[465] H. M. Burgess and N. K. Gray. An integrated model for the nucleocytoplasmic transport of cytoplasmic poly(a)-binding proteins. *Commun. Integr. Biol.*, 5(3):243–247, 2012.

[466] P. Perez-Pinera, D. D. Kocak, C. M. Vockley, A. F. Adler, A. M. Kabadi, L. R. Polstein, P. I. Thakore, K. A. Glass, D. G. Ousterout, K. W. Leong, F. Guilak, G. E. Crawford, T. E. Reddy, and C. A. Gersbach. RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nat. Methods*, 10(10):973–976, 2013.

[467] S. Bauer, S. Grossmann, M. Vingron, and P. N. Robinson. Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, 24(14):1650–1651, 2008.

[468] T. D. Wu and C. K. Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9): 1859–1875, 2005.

[469] A. J. Gruber, R. Schmidt, A. R. Gruber, G. Martin, S. Ghosh, M. Belmadani, W. Keller, and M. Zavolan. A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the

repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Res.*, 26(8):1145–1159, 2016.

[470] A. O. Subtelny, S. W. Eichhorn, G. R. Chen, H. Sive, and D. P. Bartel. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature*, 508(7494):66–71, 2014.

[471] C. P. Masamha, Z. Xia, J. Yang, T. R. Albrecht, M. Li, A.-B. Shyu, W. Li, and E. J. Wagner. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature*, 510(7505):412–416, 2014.

[472] J.-W. Nam, O. S. Rissland, D. Koppstein, C. Abreu-Goodger, C. H. Jan, V. Agarwal, M. A. Yildirim, A. Rodriguez, and D. P. Bartel. Global analyses of the effect of different cellular contexts on microRNA targeting. *Mol. Cell*, 53(6):1031–1043, 2014.

[473] H. Zhang, J. Hu, M. Recce, and B. Tian. PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res.*, 33(Database issue):D116–20, 2005.

[474] J. Y. Lee, I. Yeh, J. Y. Park, and B. Tian. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res.*, 35 (Database issue):D165–8, 2007.

[475] J. Kawai, A. Shinagawa, K. Shibata, M. Yoshino, M. Itoh, Y. Ishii, T. Arakawa, A. Hara, Y. Fukunishi, H. Konno, J. Adachi, S. Fukuda, K. Aizawa, M. Izawa, K. Nishi, H. Kiyosawa, S. Kondo, I. Yamanaka, T. Saito, Y. Okazaki, T. Gojobori, H. Bono, T. Kasukawa, R. Saito, K. Kadota, H. Matsuda, M. Ashburner, S. Batalov, T. Casavant, W. Fleischmann, T. Gaasterland, C. Gissi, B. King, H. Kochiwa, P. Kuehl, S. Lewis, Y. Matsuo, I. Nikaido, G. Pesole, J. Quackenbush, L. M. Schriml, F. Staubli, R. Suzuki, M. Tomita, L. Wagner, T. Washio, K. Sakai, T. Okido, M. Furuno, H. Aono, R. Baldarelli, G. Barsh, J. Blake, D. Boffelli, N. Bojunga, P. Carninci, M. F. de Bonaldo, M. J. Brownstein, C. Bult, C. Fletcher, M. Fujita, M. Gariboldi, S. Gustincich, D. Hill, M. Hofmann, D. A. Hume, M. Kamiya, N. H. Lee, P. Lyons, L. Marchionni, J. Mashima, J. Mazzarelli, P. Mombaerts, P. Nordone, B. Ring, M. Ringwald, I. Rodriguez, N. Sakamoto, H. Sasaki, K. Sato, C. Schönbach, T. Seya, Y. Shibata, K. F. Storch, H. Suzuki, K. Toyo-oka, K. H. Wang, C. Weitz, C. Whittaker, L. Wilming, A. Wynshaw-Boris, K. Yoshida, Y. Hasegawa, H. Kawaji, S. Kohtsuki, Y. Hayashizaki, and RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium. Functional annotation of a full-length mouse cDNA collection. *Nature*, 409(6821): 685–690, 2001.

[476] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, 7(12):1009–1015, 2010.

[477] F. Ozsolak, A. R. Platt, D. R. Jones, J. G. Reifenberger, L. E. Sass, P. McInerney, J. F. Thompson, J. Bowers, M. Jarosz, and P. M. Milos. Direct RNA sequencing. *Nature*, 461(7265):814–818, 2009.

[478] A. H. Beck, Z. Weng, D. M. Witten, S. Zhu, J. W. Foley, P. Lacroute, C. L. Smith, R. Tibshirani, M. van de Rijn, A. Sidow, and R. B. West. 3'-end sequencing for expression quantification (3SEQ) from archival tumor samples. *PLoS One*, 5(1):e8768, 2010.

[479] P. J. Shepard, E.-A. Choi, J. Lu, L. A. Flanagan, K. J. Hertel, and Y. Shi. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*, 17(4):761–772, 2011.

[480] Y. Lin, Z. Li, F. Ozsolak, S. W. Kim, G. Arango-Argoty, T. T. Liu, S. A. Tenenbaum, T. Bailey, A. P. Monaghan, P. M. Milos, and B. John. An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Res.*, 40(17):8460–8471, 2012.

[481] L. You, J. Wu, Y. Feng, Y. Fu, Y. Guo, L. Long, H. Zhang, Y. Luan, P. Tian, L. Chen, G. Huang, S. Huang, Y. Li, J. Li, C. Chen, Y. Zhang, S. Chen, and A. Xu. APASdb: a database describing alternative poly(a) sites and selection of heterogeneous cleavage sites downstream of poly(a) signals. *Nucleic Acids Res.*, 43(Database issue):D59–67, 2015.

[482] I. Ulitsky, A. Shkumatava, C. H. Jan, A. O. Subtelny, D. Koppstein, G. W. Bell, H. Sive, and D. P. Bartel. Extensive alternative polyadenylation during zebrafish development. *Genome Res.*, 22(10):2054–2066, 2012.

[483] Y. Li, Y. Sun, Y. Fu, M. Li, G. Huang, C. Zhang, J. Liang, S. Huang, G. Shen, S. Yuan, L. Chen, S. Chen, and A. Xu. Dynamic landscape of tandem 3' UTRs during zebrafish development. *Genome Res.*, 22(10): 1899–1906, 2012.

[484] M. Hoque, Z. Ji, D. Zheng, W. Luo, W. Li, B. You, J. Y. Park, G. Yehia, and B. Tian. Analysis of alternative cleavage and polyadenylation by 3 [prime] region extraction and deep sequencing. *Nat. Methods*, 10(2): 133–139, 2013.

[485] A. E. Almada, X. Wu, A. J. Kriz, C. B. Burge, and P. A. Sharp. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature*, 499(7458):360–363, 2013.

[486] X. Ji, J. Wan, M. Vishnu, Y. Xing, and S. A. Liebhaber. αcp Poly(C) binding proteins act as global regulators of alternative polyadenylation. *Mol. Cell. Biol.*, 33(13):2560–2573, 2013.

[487] N. J. Proudfoot. Ending the message: poly(a) signals then and now. *Genes Dev.*, 25(17):1770–1782, 2011.

[488] N. J. Proudfoot and G. G. Brownlee. 3' non-coding region sequences in eukaryotic messenger RNA. *Nature*, 263(5574):211–214, 1976.

[489] B. Tian, J. Hu, H. Zhang, and C. S. Lutz. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.*, 33(1):201–212, 2005.

[490] B. Tian and J. H. Graber. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip. Rev. RNA*, 3(3):385–396, 2012.

[491] J. H. Graber, C. R. Cantor, S. C. Mohr, and T. F. Smith. In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species. *Proc. Natl. Acad. Sci. U. S. A.*, 96(24):14055–14060, 1999.

[492] C. C. MacDonald and J.-L. Redondo. Reexamining the polyadenylation signal: were we wrong about AAUAAA? *Mol. Cell. Endocrinol.*, 190(1-2):1–8, 2002.

[493] J. Wilusz, D. I. Feig, and T. Shenk. The C proteins of heterogeneous nuclear ribonucleoprotein complexes interact with RNA sequences downstream of polyadenylation cleavage sites. *Mol. Cell. Biol.*, 8(10):4477–4483, 1988.

[494] X. Zhao, D. Oberg, M. Rush, J. Fay, H. Lambkin, and S. Schwartz. A 57-nucleotide upstream early polyadenylation element in human papillomavirus type 16 interacts with hfip1, CstF-64, hnRNP C1/C2, and polypyrimidine tract binding protein. *J. Virol.*, 79(7):4270–4288, 2005.

[495] P. Castelo-Branco, A. Furger, M. Wollerton, C. Smith, A. Moreira, and N. Proudfoot. Polypyrimidine tract binding protein modulates efficiency of polyadenylation. *Mol. Cell. Biol.*, 24(10):4174–4183, 2004.

[496] S. A. Alkan, K. Martincic, and C. Milcarek. The hnRNPs F and H2 bind to similar sequences to influence gene expression. *Biochem. J*, 393(Pt 1):361–371, 2006.

[497] G. K. Arhin, M. Boots, P. S. Bagga, C. Milcarek, and J. Wilusz. Downstream sequence elements with different affinities for the hnRNP H/H' protein influence the processing efficiency of mammalian polyadenylation signals. *Nucleic Acids Res.*, 30(8):1842–1850, 2002.

[498] S. Millevoi, A. Decorsière, C. Loulergue, J. Iacovoni, S. Bernat, M. Antoniou, and S. Vagner. A physical and functional link between splicing factors promotes pre-mRNA 3' end processing. *Nucleic Acids Res.*, 37 (14):4672–4683, 2009.

[499] FANTOM Consortium and the RIKEN PMI and CLST (DGT), A. R. R. Forrest, H. Kawaji, M. Rehli, J. K. Baillie, M. J. L. de Hoon, V. Haberle, T. Lassmann, I. V. Kulakovskiy, M. Lizio, M. Itoh, R. Andersson, C. J. Mungall, T. F. Meehan, S. Schmeier, N. Bertin,

M. Jørgensen, E. Dimont, E. Arner, C. Schmidl, U. Schaefer, Y. A. Medvedeva, C. Plessy, M. Vitezic, J. Severin, C. A. Semple, Y. Ishizu, R. S. Young, M. Francescatto, I. Alam, D. Albanese, G. M. Altschuler, T. Arakawa, J. A. C. Archer, P. Arner, M. Babina, S. Rennie, P. J. Balwierz, A. G. Beckhouse, S. Pradhan-Bhatt, J. A. Blake, A. Blumenthal, B. Bodega, A. Bonetti, J. Briggs, F. Brombacher, A. M. Burroughs, A. Califano, C. V. Cannistraci, D. Carbajo, Y. Chen, M. Chierici, Y. Ciani, H. C. Clevers, E. Dalla, C. A. Davis, M. Detmar, A. D. Diehl, T. Dohi, F. Drabløs, A. S. B. Edge, M. Edinger, K. Ekwall, M. Endoh, H. Enomoto, M. Fagiolini, L. Fairbairn, H. Fang, M. C. Farach-Carson, G. J. Faulkner, A. V. Favorov, M. E. Fisher, M. C. Frith, R. Fujita, S. Fukuda, C. Furlanello, M. Furino, J.-I. Furusawa, T. B. Geijtenbeek, A. P. Gibson, T. Gingeras, D. Goldowitz, J. Gough, S. Guhl, R. Guler, S. Gustincich, T. J. Ha, M. Hamaguchi, M. Hara, M. Harbers, J. Harshbarger, A. Hasegawa, Y. Hasegawa, T. Hashimoto, M. Herlyn, K. J. Hitchens, S. J. Ho Sui, O. M. Hofmann, I. Hoof, F. Hori, L. Huminiecki, K. Iida, T. Ikawa, B. R. Jankovic, H. Jia, A. Joshi, G. Jurman, B. Kaczkowski, C. Kai, K. Kaida, A. Kaiho, K. Kajiyama, M. Kanamori-Katayama, A. S. Kasianov, T. Kasukawa, S. Katayama, S. Kato, S. Kawaguchi, H. Kawamoto, Y. I. Kawamura, T. Kawashima, J. S. Kempfle, T. J. Kenna, J. Kere, L. M. Khachigian, T. Kitamura, S. P. Klinken, A. J. Knox, M. Kojima, S. Kojima, N. Kondo, H. Koseki, S. Koyasu, S. Krampitz, A. Kubosaki, A. T. Kwon, J. F. J. Laros, W. Lee, A. Lennartsson, K. Li, B. Lilje, L. Lipovich, A. Mackay-Sim, R.-I. Manabe, J. C. Mar, B. Marchand, A. Mathelier, N. Mejhert, A. Meynert, Y. Mizuno, D. A. de Lima Morais, H. Morikawa, M. Morimoto, K. Moro, E. Motakis, H. Motohashi, C. L. Mummery, M. Murata, S. Nagao-Sato, Y. Nakachi, F. Nakahara, T. Nakamura, Y. Nakamura, K. Nakazato, E. van Nimwegen, N. Ninomiya, H. Nishiyori, S. Noma, S. Noma, T. Noazaki, S. Ogishima, N. Ohkura, H. Ohimiya, H. Ohno, M. Ohshima, M. Okada-Hatakeyama, Y. Okazaki, V. Orlando, D. A. Ovchinnikov, A. Pain, R. Passier, M. Patrikakis, H. Persson, S. Piazza, J. G. D. Prendergast, O. J. L. Rackham, J. A. Ramilowski, M. Rashid, T. Ravasi, P. Rizzu, M. Roncador, S. Roy, M. B. Rye, E. Saijyo, A. Sajantila, A. Saka, S. Sakaguchi, M. Sakai, H. Sato, S. Savvi, A. Saxena, C. Schneider, E. A. Schultes, G. G. Schulze-Tanzil, A. Schwegmann, T. Sengstag, G. Sheng, H. Shimoji, Y. Shimoni, J. W. Shin, C. Simon, D. Sugiyama, T. Sugiyama, M. Suzuki, N. Suzuki, R. K. Swoboda, P. A. C. 't Hoen, M. Tagami, N. Takahashi, J. Takai, H. Tanaka, H. Tatsukawa, Z. Tatum, M. Thompson, H. Toyodo, T. Toyoda, E. Valen, M. van de Wetering, L. M. van den Berg, R. Verado, D. Vijayan, I. E. Vorontsov, W. W. Wasserman, S. Watanabe, C. A. Wells, L. N. Winteringham, E. Wolvetang, E. J. Wood, Y. Yamaguchi, M. Yamamoto, M. Yoneda, Y. Yonekura, S. Yoshida, S. E. Zabierowski, P. G. Zhang, X. Zhao, S. Zucchelli, K. M. Summers, H. Suzuki, C. O. Daub, J. Kawai, P. Heutink, W. Hide,

T. C. Freeman, B. Lenhard, V. B. Bajic, M. S. Taylor, V. J. Makeev, A. Sandelin, D. A. Hume, P. Carninci, and Y. Hayashizaki. A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470, 2014.

[500] S. Lianoglou, V. Garg, J. L. Yang, C. S. Leslie, and C. Mayr. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.*, 27(21):2380–2396, 2013.

[501] M. D. Sheets, S. C. Ogg, and M. P. Wickens. Point mutations in AAUAAA and the poly (a) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res.*, 18(19):5799–5805, 1990.

[502] F. Ozsolak, P. Kapranov, S. Foissac, S. W. Kim, E. Fishilevich, A. P. Monaghan, B. John, and P. M. Milos. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell*, 143(6):1018–1029, 2010.

[503] K. Venkataraman, K. M. Brown, and G. M. Gilmartin. Analysis of a noncanonical poly(a) site reveals a tripartite mechanism for vertebrate poly(a) site recognition. *Genes Dev.*, 19(11):1315–1327, 2005.

[504] W. Li, B. You, M. Hoque, D. Zheng, W. Luo, Z. Ji, J. Y. Park, S. I. Gunderson, A. Kalsotra, J. L. Manley, and B. Tian. Systematic profiling of poly(a)+ transcripts modulated by core 3' end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. *PLoS Genet.*, 11(4):e1005166, 2015.

[505] J. König, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D. J. Turner, N. M. Luscombe, and J. Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.*, 17(7):909–915, 2010.

[506] K. Zarnack, J. König, M. Tajnik, I. n. Martincorena, S. Eustermann, I. Stévant, A. Reyes, S. Anders, N. M. Luscombe, and J. Ule. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of alu elements. *Cell*, 152(3):453–466, 2013.

[507] Z. Cieniková, F. F. Damberger, J. Hall, F. H.-T. Allain, and C. Maris. Structural and mechanistic insights into poly(uridine) tract recognition by the hnRNP C RNA recognition motif. *J. Am. Chem. Soc.*, 136(41): 14536–14544, 2014.

[508] N. Liu, Q. Dai, G. Zheng, C. He, M. Parisien, and T. Pan. N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature*, 518(7540):560–564, 2015.

[509] M. Tajnik, A. Vigilante, S. Braun, H. Hänel, N. M. Luscombe, J. Ule, K. Zarnack, and J. König. Intergenic alu exonisation facilitates the

evolution of tissue-specific transcript ends. *Nucleic Acids Res.*, 43(21): 10492–10505, 2015.

[510] S. Kishore, L. Jaskiewicz, L. Burger, J. Hausser, M. Khorshid, and M. Zavolan. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods*, 8(7):559–564, 2011.

[511] J. P. ten Klooster, I. v. Leeuwen, N. Scheres, E. C. Anthony, and P. L. Hordijk. Rac1-induced cell migration requires membrane recruitment of the nuclear oncogene SET. *EMBO J.*, 26(2):336–345, 2007.

[512] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, and T. J. Hubbard. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.*, 22(9): 1760–1774, 2012.

[513] R. Davis and Y. Shi. The polyadenylation code: a unified model for the regulation of mRNA alternative polyadenylation. *J. Zhejiang Univ. Sci. B*, 15(5):429–437, 2014.

[514] P. Miura, S. Shenker, C. Andreu-Agullo, J. O. Westholm, and E. C. Lai. Widespread and extensive lengthening of 3' UTRs in the mammalian brain. *Genome Res.*, 23(5):812–825, 2013.

[515] Z. Xia, L. A. Donehower, T. A. Cooper, J. R. Neilson, D. A. Wheeler, E. J. Wagner, and W. Li. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat. Commun.*, 5:5274, 2014.

[516] D. C. Di Giammartino, W. Li, K. Ogami, J. J. Yashinskie, M. Hoque, B. Tian, and J. L. Manley. RBBP6 isoforms regulate the human polyadenylation machinery and modulate expression of mRNAs with AU-rich 3' UTRs. *Genes Dev.*, 28(20):2248–2260, 2014.

[517] E. de Klerk, J. T. den Dunnen, and P. A. C. 't Hoen. RNA sequencing: from tag-based profiling to resolving complete transcript structure. *Cell. Mol. Life Sci.*, 71(18):3537–3551, 2014.

[518] Z. Wu, X. Liu, L. Liu, H. Deng, J. Zhang, Q. Xu, B. Cen, and A. Ji. Regulation of lncRNA expression. *Cell. Mol. Biol. Lett.*, 19(4):561–575, 2014.

[519] M. J. Hangauer, I. W. Vaughn, and M. T. McManus. Pervasive transcription of the human genome produces thousands of previously

unidentified long intergenic noncoding RNAs. *PLoS Genet.*, 9(6): e1003569, 2013.

[520] A. McCloskey, I. Taniguchi, K. Shinmyozu, and M. Ohno. hnRNP C tetramer measures RNA length to classify RNA polymerase II transcripts for export. *Science*, 335(6076):1643–1646, 2012.

[521] A. L. Beyer, M. E. Christensen, B. W. Walker, and W. M. LeStourgeon. Identification and characterization of the packaging proteins of core 40S hnRNP particles. *Cell*, 11(1):127–138, 1977.

[522] Y. D. Choi and G. Dreyfuss. Isolation of the heterogeneous nuclear RNA-ribonucleoprotein complex (hnRNP): a unique supramolecular assembly. *Proc. Natl. Acad. Sci. U. S. A.*, 81(23):7471–7475, 1984.

[523] S. R. Whitson, W. M. LeStourgeon, and A. M. Krezel. Solution structure of the symmetric coiled coil tetramer formed by the oligomerization domain of hnRNP c: implications for biological function. *J. Mol. Biol.*, 350(2):319–337, 2005.

[524] M. Görlach, C. G. Burd, and G. Dreyfuss. The determinants of RNA-binding specificity of the heterogeneous nuclear ribonucleoprotein C proteins. *J. Biol. Chem.*, 269(37):23074–23078, 1994.

[525] Y. Shi. Alternative polyadenylation: new insights from global analyses. *RNA*, 18(12):2105–2117, 2012.

[526] P. Flicek, I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. García-Girón, L. Gordon, T. Hourlier, S. Hunt, T. Juettemann, A. K. Kähäri, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W. M. McLaren, M. Muffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sheppard, D. Sobral, K. Taylor, A. Thormann, S. Trevanion, S. White, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T. J. P. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa, and S. M. J. Searle. Ensembl 2013. *Nucleic Acids Res.*, 41(Database issue):D48–55, 2013.

[527] L. R. Meyer, A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, B. J. Raney, A. Pohl, V. S. Malladi, C. H. Li, B. T. Lee, K. Learned, V. Kirkup, F. Hsu, S. Heitner, R. A. Harte, M. Haeussler, L. Guruvadoo, M. Goldman, B. M. Giardine, P. A. Fujita, T. R. Dreszer, M. Diekhans, M. S. Cline, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent. The UCSC genome browser database: extensions and updates 2013. *Nucleic Acids Res.*, 41(Database issue):D64–9, 2013.

[528] S. Hoffmann, C. Otto, S. Kurtz, C. M. Sharma, P. Khaitovich, J. Vogel, P. F. Stadler, and J. Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput. Biol.*, 5(9):e1000502, 2009.

[529] R. C. Team. R: A language and environment for statistical computing. R foundation for statistical computing, vienna, austria., 2014.

[530] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Res.*, 14(6):1188–1190, 2004.

[531] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC genome browser database: update 2006. *Nucleic Acids Res.*, 34(Database issue):D590–8, 2006.

[532] L. Jaskiewicz, B. Bilen, J. Hausser, and M. Zavolan. Argonaute CLIP–a method to identify in vivo targets of miRNAs. *Methods*, 58(2):106–112, 2012.

[533] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[534] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[535] N. E. Pagliaccetti, R. Eduardo, S. H. Kleinstein, X. J. Mu, P. Bandi, and M. D. Robek. Interleukin-29 functions cooperatively with interferon to induce antiviral gene expression and inhibit hepatitis C virus replication. *J. Biol. Chem.*, 283(44):30079–30089, 2008.

[536] M. Pai, R. Prabhu, A. Panebra, S. Nangle, S. Haque, F. Bastian, R. Garry, K. Agrawal, S. Goodbourn, and S. Dash. Activation of interferon-stimulated response element in huh-7 cells replicating hepatitis C virus subgenomic RNA. *Intervirology*, 48(5):301–311, 2005.

[537] J.-R. Jheng, J.-Y. Ho, and J.-T. Horng. ER stress, autophagy, and RNA viruses. *Front. Microbiol.*, 5:388, 2014.

[538] J. Wang, R. Kang, H. Huang, X. Xi, B. Wang, J. Wang, and Z. Zhao. Hepatitis C virus core protein activates autophagy through EIF2AK3 and ATF6 UPR pathway-mediated MAP1LC3B and ATG12 expression. *Autophagy*, 10(5):766–784, 2014.

[539] E. Merquiol, D. Uzi, T. Mueller, D. Goldenberg, Y. Nahmias, R. J. Xavier, B. Tirosh, and O. Shibolet. HCV causes chronic endoplasmic reticulum stress leading to adaptation and interference with the unfolded protein response. *PLoS One*, 6(9):e24660, 2011.

[540] T. Asselah, I. Bièche, A. Mansouri, I. Laurendeau, D. Cazals-Hatem, G. Feldmann, P. Bedossa, V. Paradis, M. Martinot-Peignoux, D. Lebrec, C. Guichard, E. Ogier-Denis, M. Vidaud, Z. Tellier, V. Soumelis, P. Marcellin, and R. Moreau. In vivo hepatic endoplasmic reticulum stress in patients with chronic hepatitis C. *J. Pathol.*, 221(3):264–274, 2010.

[541] U. Alon. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.*, 8(6):450–461, 2007.

[542] H.-M. Zhang, S. Kuang, X. Xiong, T. Gao, C. Liu, and A.-Y. Guo. Transcription factor and microRNA co-regulatory loops: important regulatory motifs in biological processes and diseases. *Brief. Bioinform.*, 16(1):45–58, 2015.

[543] W.-T. Hsieh, K.-R. Tzeng, J.-S. Ciou, J. J. Tsai, N. Kurubanjerdjit, C.-H. Huang, and K.-L. Ng. Transcription factor and microRNA-regulated network motifs for cancer and signal transduction networks. *BMC Syst. Biol.*, 9 Suppl 1:S5, 2015.

[544] K. Li, Z. Li, N. Zhao, Y. Xu, Y. Liu, Y. Zhou, D. Shang, F. Qiu, R. Zhang, Z. Chang, and Y. Xu. Functional analysis of microRNA and transcription factor synergistic regulatory network based on identifying regulatory motifs in non-small cell lung cancer. *BMC Syst. Biol.*, 7:122, 2013.

[545] K. Poos, J. Smida, M. Nathrath, D. Maugg, D. Baumhoer, and E. Korsching. How microRNA and transcription factor co-regulatory networks affect osteosarcoma cell proliferation. *PLoS Comput. Biol.*, 9 (8):e1003210, 2013.

[546] K. Polyak. Heterogeneity in breast cancer. *J. Clin. Invest.*, 121(10): 3786–3788, 2011.

[547] D. Melisi, L. Calvetti, M. Frizziero, and G. Tortora. Pancreatic cancer: systemic combination therapies for a heterogeneous disease. *Curr. Pharm. Des.*, 20(42):6660–6669, 2014.

[548] Z. Chen, C. M. Fillmore, P. S. Hammerman, C. F. Kim, and K.-K. Wong. Non-small-cell lung cancers: a heterogeneous set of diseases. *Nat. Rev. Cancer*, 14(8):535–546, 2014.

[549] K. H. Allison and G. W. Sledge. Heterogeneity and cancer. *Oncology*, 28(9):772–778, 2014.

[550] E. Shapiro, T. Biezuner, and S. Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.*, 14(9):618–630, 2013.

[551] A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann. The technology and biology of single-cell RNA sequencing. *Mol. Cell*, 58(4):610–620, 2015.

[552] N. D. Marjanovic, R. A. Weinberg, and C. L. Chaffer. Cell plasticity and heterogeneity in cancer. *Clin. Chem.*, 59(1):168–179, 2013.

[553] C. E. Meacham and S. J. Morrison. Tumour heterogeneity and cancer cell plasticity. *Nature*, 501(7467):328–337, 2013.

[554] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, and R. Edgar. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, 37(Database issue):D885–90, 2009.

[555] F. Jay and C. Ciaudo. An RNA tool kit to study the status of mouse ES cells: sex determination and stemness. *Methods*, 63(1):85–92, 2013.

[556] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engstrom, M. C. Frith, A. R. Forrst, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Taber, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. GRimmend, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume, and Y. Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6):626–635, 2006.

[557] R. Kodzius, M. Kojima, H. Nishiyori, M. Nakamura, S. Fukuda, M. Tagami, D. Sasaki, K. Imamura, C. Kai, M. Harbers, Y. Hayashizaki, and P. Carninci. CAGE: cap analysis of gene expression. *Nat. Methods*, 3(3):211–222, 2006.

[558] W. J. Kent. BLAT–the BLAST-like alignment tool. *Genome Res.*, 12 (4):656–664, 2002.

[559] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, W. J. Kent, and U. of California Santa Cruz. The UCSC Genome Browser Database. *Nucleic Acids Res*, 31(1):51–54, 2003.

[560] E. Portales-Casamar, S. Thongjuea, A. T. Kwon, D. Arenillas, X. Zhao, E. Valen, D. Yusuf, B. Lenhard, W. W. Wasserman, and A. Sandelin. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 38:D105–D110, 2010.

[561] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, J. S. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. Sonnhammer, and A. Bateman. The Pfam protein families database. *Nucleic Acids Res*, 36: D281–D288, 2008.

[562] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755–763, 1998.

[563] R. Siddharthan, E. D. Siggia, and E. van Nimwegen. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol*, 1(7):e67, 2005.

[564] J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, J. E. Richardson, M. Airey, A. Anagnostopoulos, R. Babiuk, R. Baldarelli, J. Beal, S. Bello, N. Butler, J. Campbell, L. Corbani, S. Giannatto, H. Dene, M. Dolan, H. Drabkin, K. Forthofer, M. Knowlton, J. Lewis, M. McAndrews-Hill, S. McClatchy, D. Miers, L. Ni, H. Onda, J. E. Ormsby, J. Recla, D. Reed, B. Richards-Smith, R. Shaw, R. Sinclair, D. Sitnikov, C. Smith, L. Washburn, and Y. Zhu. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res*, 40(Database issue):D881–886, 2012.

[565] S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller. Human-mouse alignments with BLASTZ. *Genome Res*, 13(1):103–107, 2003.

[566] N. Molina and E. van Nimwegen. Universal patterns of purifying selection at noncoding positions in bacteria. *Genome Res*, 18(1):148–160, 2008.

[567] W. Miller, K. Rosenbloom, R. C. Hardison, M. Hou, J. Taylor, B. Raney, R. Burhans, D. C. King, R. Baertsch, D. Blankenberg, S. L. K. Pond, A. Nekrutenko, B. Giardine, R. S. Harris, S. Tyekucheva, M. Diekhans, T. H. Pringle, W. J. Murphy, A. Lesk, G. M. Weinstock, K. Lindblad-Toh, R. A. Gibbs, E. S. Lander, A. Siepel, D. Haussler, and W. J. Kent. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res*, 17(12):1797–1808, 2007.

[568] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3): 307–315, 2004.

[569] C. Fraley and A. E. Raftery. Model-Based Clustering, Discriminant Analysis and Density Estimation. *J Am Stat Assoc*, 97:611–631, 2002.

[570] C. Fraley and A. E. Raftery. MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. Technical Report 504, University of Washington, Department of Statistics, 2006, revised in 2009.

[571] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[572] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock. GO::TermFinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18): 3710–3715, 2004.

[573] Bodymap2. Illumina human body map 2.0 project. Geo Accession GSE30611, 2011. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30611.

[574] G. L. Semenza. Defining the role of hypoxia-inducible factor 1 in cancer biology and therapeutics. *Oncogene*, 29(5):625–634, 2010.

[575] N. Meyer and L. Z. Penn. Reflecting on 25 years with MYC. *Nat Rev Cancer*, 8(12):976–990, 2008.

[576] H. Z. Chen, S. Y. Tsai, and G. Leone. Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nat Rev Cancer*, 9(11):785–797, 2009.

[577] D. Dolfini and R. Mantovani. Targeting the Y/CCAAT box in cancer: YB-1 (YBX1) or NF-Y? *Cell Death Differ*, 20(5):676–685, 2013.

[578] G. Castellano, E. Torrisi, G. Ligresti, G. Malaponte, L. Militello, A. E. Russo, J. A. McCubrey, S. Canevari, and M. Libra. The involvement of the transcription factor Yin Yang 1 in cancer development and progression. *Cell Cycle*, 8(9):1367–1372, 2009.

[579] B. K. Yoo, L. Emdad, R. Gredler, C. Fuller, C. I. Dumur, K. H. Jones, C. Jackson-Cook, Z. Z. Su, D. Chen, U. H. Saxena, U. Hansen, P. B. Fisher, and D. Sarkar. Transcription factor Late SV40 Factor (LSF) functions as an oncogene in hepatocellular carcinoma. *Proc Natl Acad Sci U S A*, 107(18):8357–8362, 2010.

[580] D. Samanta and P. K. Datta. Alterations in the Smad pathway in human cancers. *Front Biosci (Landmark Ed)*, 17:1281–1293, 2012.

[581] J. Martinez Hoyos, A. Ferraro, S. Sacchetti, S. Keller, I. De Martino, E. Borbone, P. Pallante, M. Fedele, D. Montanaro, F. Esposito, P. Cserjesi, L. Chiariotti, G. Troncone, and A. Fusco. HAND1 gene expression is negatively regulated by the High Mobility Group A1 proteins and is drastically reduced in human thyroid carcinomas. *Oncogene*, 28 (6):876–885, 2009.

[582] Y. Nakamura, T. Migita, F. Hosoda, N. Okada, M. Gotoh, Y. Arai, M. Fukushima, M. Ohki, S. Miyata, K. Takeuchi, I. Imoto, H. Katai, T. Yamaguchi, J. Inazawa, S. Hirohashi, Y. Ishikawa, and T. Shibata. Krüppel-like factor 12 plays a significant role in poorly differentiated gastric cancer progression. *Int. J. Cancer*, 125(8):1859–1867, 2009.

[583] Y. Buganim, I. Goldstein, D. Lipson, M. Milyavsky, S. Polak-Charcon, C. Mardoukh, H. Solomon, E. Kalo, S. Madar, R. Brosh, M. Perelman,

R. Navon, N. Goldfinger, I. Barshack, Z. Yakhini, and V. Rotter. A novel translocation breakpoint within the BPTF gene is associated with a pre-malignant phenotype. *PLoS One*, 5(3):e9657, 2010.

[584] K. J. Basile, E. V. Abel, and A. E. Aplin. Adaptive upregulation of FOXD3 and resistance to PLX4032/4720-induced cell death in mutant B-RAF melanoma cells. *Oncogene*, 31(19):2471–2479, 2012.

[585] H. Izumi, T. Wakasugi, S. Shimajiri, A. Tanimoto, Y. Sasaguri, E. Kashiwagi, Y. Yasuniwa, M. Akiyama, B. Han, Y. Wu, T. Uchiumi, T. Arao, K. Nishio, R. Yamazaki, and K. Kohno. Role of ZNF143 in tumor growth through transcriptional regulation of DNA replication and cell-cycle-associated genes. *Cancer Sci*, 101(12):2538–2545, 2010.

[586] P. Gandellini, M. Folini, N. Longoni, M. Pennati, M. Binda, M. Colecchia, R. Salvioni, R. Supino, R. Moretti, P. Limonta, R. Valdagni, M. G. Daidone, and N. Zaffaroni. miR-205 Exerts tumor-suppressive functions in human prostate through down-regulation of protein kinase Cepsilon. *Cancer Res*, 69(6):2287–2295, 2009.

[587] S. Majid, A. A. Dar, S. Saini, S. Yamamura, H. Hirata, Y. Tanaka, G. Deng, and R. Dahiya. MicroRNA-205-directed transcriptional activation of tumor suppressor genes in prostate cancer. *Cancer*, 116(24): 5637–5649, 2010.

[588] A. A. Dar, S. Majid, D. de Semir, M. Nosrati, V. Bezrookove, and M. Kashani-Sabet. miRNA-205 suppresses melanoma cell proliferation and induces senescence via regulation of E2F1 protein. *J Biol Chem*, 286(19):16606–16614, 2011.

[589] H. Wu, S. Zhu, and Y. Y. Mo. Suppression of cell growth and invasion by miR-205 in breast cancer. *Cell Res*, 19(4):439–448, 2009.

[590] S. Liu, M. T. Tetzlaff, A. Liu, B. Liegl-Atzwanger, J. Guo, and X. Xu. Loss of microRNA-205 expression is associated with melanoma progression. *Lab Invest*, 92(7):1084–1096, 2012.

[591] J. Kota, R. R. Chivukula, K. A. O'Donnell, E. A. Wentzel, C. L. Montgomery, H. W. Hwang, T. C. Chang, P. Vivekanandan, M. Torbenson, K. R. Clark, J. R. Mendell, and J. T. Mendell. Therapeutic microRNA delivery suppresses tumorigenesis in a murine liver cancer model. *Cell*, 137(6):1005–1017, 2009.

[592] L. He, J. M. Thomson, M. T. Hemann, E. Hernando-Monge, D. Mu, S. Goodson, S. Powers, C. Cordon-Cardo, S. W. Lowe, G. J. Hannon, and S. M. Hammond. A microRNA polycistron as a potential human oncogene. *Nature*, 435(7043):828–833, 2005.

[593] R. Chhabra, R. Dubey, and N. Saini. Cooperative and individualistic functions of the microRNAs in the miR-23a 27a 24-2 cluster and its implication in human diseases. *Mol Cancer*, 9:232, 2010.

[594] K. H. To, S. Pajovic, B. L. Gallie, and B. L. Theriault. Regulation of p14ARF expression by miR-24: a potential mechanism compromising the p53 response during retinoblastoma development. *BMC Cancer*, 12:69, 2012.

[595] A. Lal, F. Navarro, C. A. Maher, L. E. Maliszewski, N. Yan, E. O'Day, D. Chowdhury, D. M. Dykxhoorn, P. Tsai, O. Hofmann, K. G. Becker, M. Gorospe, W. Hide, and J. Lieberman. miR-24 Inhibits cell proliferation by targeting E2F2, MYC, and other cell-cycle genes via binding to "seedless" 3'UTR microRNA recognition elements. *Mol Cell*, 35 (5):610–625, 2009.

[596] M. Xiong, L. Jiang, Y. Zhou, W. Qiu, L. Fang, R. Tan, P. Wen, and J. Yang. The miR-200 family regulates TGF-beta1-induced renal tubular epithelial to mesenchymal transition through Smad pathway by targeting ZEB1 and ZEB2 expression. *Am J Physiol Renal Physiol*, 302 (3):F369–F379, 2012.

[597] U. Burk, J. Schubert, U. Wellner, O. Schmalhofer, E. Vincan, S. Spaderna, and T. Brabletz. A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells. *EMBO Rep*, 9(6):582–589, 2008.

[598] P. A. Gregory, A. G. Bert, E. L. Paterson, S. C. Barry, A. Tsykin, G. Farshid, M. A. Vadas, Y. Khew-Goodall, and G. J. Goodall. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat Cell Biol*, 10(5):593–601, 2008.

[599] K. M. Hajra, D. Y.-S. Chen, and E. R. Fearon. The SLUG zinc-finger protein represses E-cadherin in breast cancer. *Cancer Res*, 62(6):1613–1618, 2002.

[600] M. L. Grooteclaes and S. M. Frisch. Evidence for a function of CtBP in epithelial gene regulation and anoikis. *Oncogene*, 19(33):3823–3828, 2000.

[601] F. Tang, R. Zhang, Y. He, M. Zou, L. Guo, and et al. MicroRNA-125b Induces Metastasis by Targeting STARD13 in MCF-7 and MDA-MB-231 Breast Cancer Cells. *PLoS One*, 7(5):e35435, 2012.

[602] C. H. Jan, R. C. Friedman, J. G. Ruby, and D. P. Bartel. Formation, regulation and evolution of caenorhabditis elegans 3'UTRs. *Nature*, 469(7328):97–101, 2011.

[603] C. Yao, J. Biesinger, J. Wan, L. Weng, Y. Xing, X. Xie, and Y. Shi. Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation. *Proc. Natl. Acad. Sci. U. S. A.*, 109(46):18773–18778, 2012.

[604] A. Rehfeld, M. Plass, K. Døssing, U. Knigge, A. Kjær, A. Krogh, and L. Friis-Hansen. Alternative polyadenylation of tumor suppressor genes in small intestinal neuroendocrine tumors. *Front. Endocrinol.*, 5:46, 2014.

[605] R. Batra, K. Charizanis, M. Manchanda, A. Mohan, M. Li, D. J. Finn, M. Goodwin, C. Zhang, K. Sobczak, C. A. Thornton, and M. S. Swanson. Loss of MBNL leads to disruption of developmentally regulated alternative polyadenylation in RNA-mediated disease. *Mol. Cell*, 56 (2):311–322, 2014.

[606] Y. Fu, Y. Sun, Y. Li, J. Li, X. Rao, C. Chen, and A. Xu. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res.*, 21 (5):741–747, 2011.

[607] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nat. Biotechnol.*, 29(1):24–26, 2011.