Influence of histone modifications on mRNA abundance and structure

Rosa Karlić

September 2011

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Gutachter:
Prof. Dr. Martin Vingron
Prof. Dr. Kristian Vlahoviček

1. Gutachter: Prof. Dr. Martin Vingron
2. Gutachter: Prof. Dr. Kristian Vlahoviček

Tag der Disputation: 16. Dezember 2011

Preface

Publications The results presented in the first part of Chapter 2 were published in the *Proceedings of the National Academy of Sciences* [111]. I was involved in designing the study, and implemented all the analyses presented in this thesis. I would like to acknowledge *Ho-Ryun Chung* for his contribution to the design of the study and his help with analyzing the data, as well as his development of the preliminary "Histone code for transcription". The results presented in Sections 2.5 and 2.6, as well as Chapter 3 have not been published yet, but manuscripts for both of these analyses are in preparation.

Acknowledgments I would like to thank *Martin Vingron* and *Kristian Vlahoviček* for the support and guidance that they provided by co-supervising my thesis, and all the ideas that they contributed during my PhD studies. I would especially like to thank them for giving me the opportunity and the freedom to pursue my scientific interests, which helped me develop my independence and, I believe, made me a better scientist.

I would especially like to thank *Ho-Ryun Chung*, who directly supervised my work during most of my PhD. Our daily interactions contributed greatly to my knowledge of epigenetics and computational biology, and also helped me realize what kind of scientist I want to be. Most of the results presented in this thesis are a product of our fruitful scientific discussions.

My special thanks go to *Julia Lasserre*, who influenced considerably my knowledge of statistics and machine learning, and contributed significantly to the design of analyses described in this thesis. I appreciated greatly her patience and her readiness to help me with any problems I encountered during my PhD, as well as her support and friendship outside of work.

I would furthermore like to thank all past and present members of the Computational Molecular Biology Department at the Max Planck Institute for Molecular Genetics, as well the members of the Bioinformatics Group at Zagreb University, for their help and the wonderful working environment they created. It made the time which I spent working on my PhD thesis a thoroughly enjoyable one.

To *Brian Cusack* and *Paz Polak*, thank you for the many scientific discussions we had during my PhD studies. Most of all, thank you for your friendship and your help in making me realize that there is a funny side to everything. It made even the most stressful days seem better.

I would like to thank *Ho-Ryun Chung*, *Brian Cusack* and *Julia Lasserre* for their critical reading of this thesis and all of their helpful comments. I would also like to acknowledge *Hannes Luz* and *Kirsten Kelleher* for the wonderful job they did as coordinators of the International Max Planck Research School for Computational Biology and Scientific Computing. I greatly appreciated their help and support.

During my PhD I was funded by the International Max Planck Research School for Computational Biology and Scientific Computing and the Croatian Ministry of Science, for which I am very grateful. I would also like to acknowledge the L'Oreal Adria-UNESCO National Fellowship "For Women in Science", which I was awarded during the last year of my PhD studies.

Lastly, I would like to thank all of my friends and family in both Berlin and Croatia. Your support made all of this possible, and your friendship made my experience in Berlin an unforgettable one. I would especially like to thank my parents for always believing in me and teaching me that no goal is unattainable, if you set your mind to it. You are the ones I have to thank for everything I achieved. Finally, to Igor, thank you for always being at my side and supporting me in following my dreams. I could not have done it without you.

Berlin, September 2011 Rosa Karlić

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The term epigenetics describes heritable changes in the genome that are not transmitted through the DNA sequence. Covalent mechanisms of storing epigenetic information include histone modifications and DNA methylation, both of which have been linked to regulation of various cellular processes. The goal of this thesis is to investigate the relationships between histone modifications and two different processes involved in the production of a mature mRNA, namely mRNA transcription and splicing. In this chapter we will give a brief overview of the current knowledge about histone modifications and their mechanisms of action. We will also describe mechanisms involved in mRNA transcription, splicing and degradation. Finally, we will describe some of the main experimental methods used to generate publicly available data used in our analysis.

1.1 Histone modifications

The DNA in the nucleus of eukaryotic organisms is packaged into a compact structure called chromatin. The basic repeating unit of chromatin is the nucleosome, formed by wrapping 147 base pairs (bp) of DNA around an octamer of four core histones, H2A, H2B, H3, and H4 [55, 118, 119, 140, 210]. The fact that tails of histone proteins, which protrude from the nucleosome, can be post-translationally modified by the introduction of acetyl and methyl groups was discovered almost fifty years ago, accompanied by the observation that these modifications can influence RNA synthesis [7]. It is now known that many amino acid residues in both the tails and the core of histones are subject to these and additional post-translational covalent modifications [14].

Currently, at least nine different types of modifications of histone proteins have been identified: acetylation, methylation, phosphorylation, deimination, sumoylation, ubiquitination, ADP ribosylation, proline isomerisation and addition of a β-N-acetylglucosamine sugar residue. In addition, histone N-terminal tails can be clipped, a process that was identified in both lower and higher eukaryotes (reviewed in [14]). The three most studied types of histone modifications are acetylation, phosphorylation and methylation. Acetylation is restricted to ϵ-amino groups of lysine side chains [242], while phosphorylation occurs on serines, threonines and tyrosines [159].

Among histone residues, methylation is mostly found on lysines and arginines, and in contrast to the two previously mentioned modifications, can be present in several different states. Namely, lysines can be modified by the addition of one, two or three methyl groups, while arginines can be monomethylated, as well as asymmetrically or symmetrically methylated [154]. A common nomenclature is used for the description of histone modifications starting with the abbreviated name of the modified histone, followed by a single letter denoting the type of the amino acid residue, a number indicating the position of the residue in the sequence, and ending with the type of modification [213]. Thus, three methyl groups added to a lysine at position 4 of histone H3 would be denoted as H3K4me3.

The hypothesis that post-translational acetylation and methylation of histones can influence transcription was proposed by Allfrey *et al.*, in one of the earliest studies of histone modifications [7]. Since then, histone modifications have been linked to a number of additional chromatin-dependent processes, including replication, DNA-repair and recombination [121]. The observation that a single nucleosome can be decorated by multiple histone modifications has led to the "histone code hypothesis", which states that "distinct histone modifications, on one or more tails, act sequentially or in combination to bring about distinct downstream events" [106, 203].

There are two main mechanisms by which histone modifications affect cellular processes. First is the alteration of chromatin structure and function by inducing a change in the charge of the nucleosome particle. For example, addition of an acetyl group neutralizes the positive charge of the lysine residues, and can potentially disturb electrostatic interactions between the histone and the negatively charged DNA strand, presumably resulting in a less compact chromatin structure. Indeed it has been experimentally shown that acetylation of histones regulates chromatin structure during interphase, with hyperacetylation causing increased chromatin accessibility [84]. Changes in chromatin accessibility can influence other processes, such as transcription, as supported by the observation that histone acetylations are frequently enriched in promoters of active genes, possibly facilitating binding of transcription factors to the promoter [228]. However, not all negatively charged modifications will necessarily induce the decondensation of chromatin. Instead, they can exert their effect by regulating the binding of different protein complexes, the other main mechanism of function of histone modifications. This is exemplified by the phosphorylation of serine 10 of histone H3, a modification that is correlated with chromatin condensation, although the phosphate adds a negative charge to the histone [230]. This somewhat counterintuitive result is explained by the fact that H3S10ph disturbs the binding of HP1 protein, enabling proper chromatin condensation and chromosome segregation during mitosis [70]. In the absence of H3S10ph, HP1 binds to H3K9me3, providing another example of how histone modifications can regulate protein recruitment.

Histone modifications, along with DNA methylation, are considered one of the two main types of epigenetic modifications [26]. Epigenetics involves the study of "heritable phenotypes resulting from changes in a chromosome without alternations in the

DNA sequence" [24]. Imprinting, X chromosome inactivation and heterochromatin formation are just a few of the processes transmitted to the next generation by epigenetic inheritance, and although it is clear that histone modifications are involved in epigenetic inheritance, the exact mechanism by which the heritable information is transmitted from mother to daughter cell still remains to be elucidated [121]. Recent discoveries of enzymes which can remove histone modifications showed that most of these modifications are dynamic [38, 103, 191, 242], and raised the question of whether histone modifications are more likely to maintain rather than actually transmit epigenetic information. A recent study in mice showed that an epigenetic phenotype can be inherited through a small RNA molecule present in sperm [170] and raised the possibility that small RNAs could also be involved in the inheritance of chromatin states, since they have been shown to be able to direct chromatin-modifying complexes to DNA [121, 218]. In summary, the mechanism of transmission of epigenetic information to the next generation is one of the most fundamental open questions in the field of epigenetics, and remains to be answered by further studies.

1.2 Transcription of mRNA

1.2.1 The transcription cycle

Transcription is a process of synthesizing a complementary RNA copy of a DNA sequence. In eukaryotes, all protein-coding genes are transcribed by RNA polymerase II (PolII), a protein complex composed of 12 subunits which catalyzes the formation of the phosphodiester bond between incoming ribonucleotides to enable the formation of the nascent RNA chain [192].

Transcription proceeds in a series of steps, also referred to as transcription cycle. The transcription cycle starts with the assembly of the preinitiation complex (PIC) at the promoter, after which PolII initiates transcription. PolII then leaves the promoter in a process called promoter clearance, in order to proceed with efficient transcript elongation. Finally, transcription terminates upon reaching the end of the gene, and PolII is released from the gene, ready to start another transcription cycle (reviewed in [63, 206]; Fig. 1.1).

The preinitiation complex consists of PolII and a set of basal transcription factors, namely TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH, whose main role is to facilitate promoter recognition and unwinding of DNA at the start of the transcription cycle [195]. Early studies showed that assembly of PIC at the promoter was mediated by the recognition of TATA-box by the TATA-binding protein (TBP) subunit of TFIID ([93], and references therein). The TATA-box is a common element of the core promoter, the DNA sequence which drives transcription initiation [109]. It was long believed that the TATA-box motif is essential for successful promoter recognition. However, the annotation and characterization of more and more genes revealed that only a fraction of vertebrate core promoters contain a TATA-box [115].

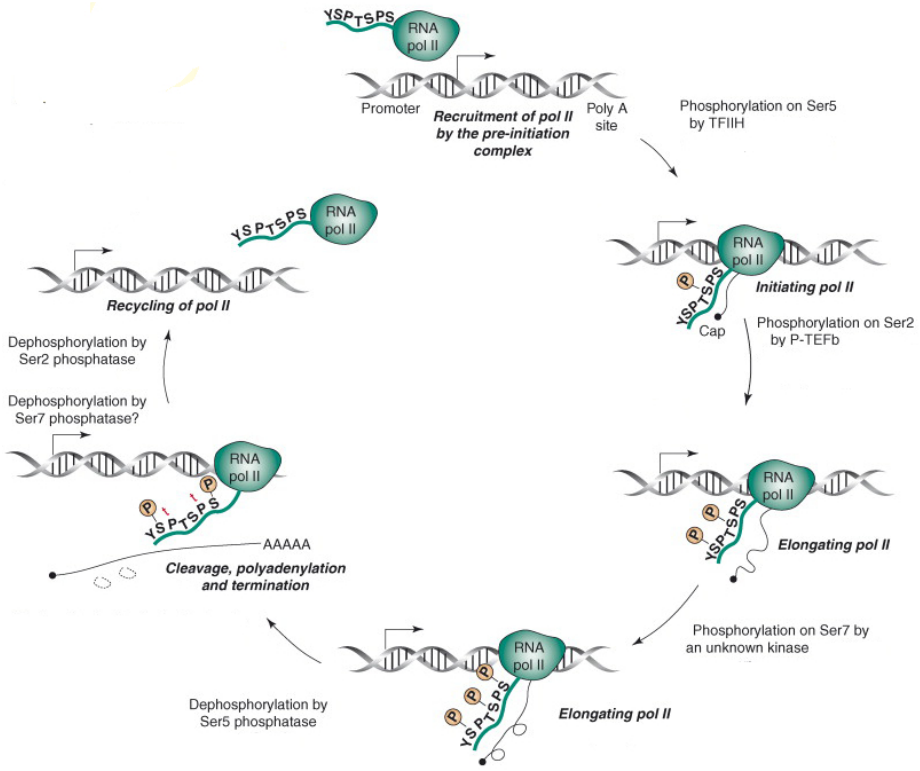

Figure 1.1: The transcription cycle. Schematic representation of the transcription cycle (adapted from Egloff and Murphy [63])

Furthermore, later studies identified many additional core promoter elements, such as the downstream promoter element (DPE), TFIIB recognition elements (BREs), the initiator element (INR) or the downstream core element (DCE). These elements can be recognized by different subunits of PIC, such as the recognition of INR and DPE by subunits of TFIID called TATA-associated factors (TAFs) or recognition of BRE by TFIIB (reviewed in [109]). Computational studies revealed that there is an enrichment of certain combinations of core promoter elements, although it is very rare that they all appear together in the same promoter [71, 80, 158]. Taken together, these findings imply that there are several different modes of recognition of the core promoter, that depend on its sequence features. Furthermore, the finding that components of TFIID are differentially expressed during myogenesis suggests that the mode of recognition of the core promoter could depend on the type and developmental stage of the cell [58].

Once PIC is assembled at the promoter, PolII can proceed with transcription initiation. For initiation to occur, the DNA at the promoter has to be unwound, in a process called promoter melting, catalyzed by the helicase activity of a subunit of TFIIH [89]. After DNA is unwound, PolII can catalyze the formation of the first phosphodiester bond of the nascent RNA molecule. The largest subunit of eukaryotic RNA polymerase II, Rpb1, has a C-terminal domain (CTD) which consists of multiple tandem repeats of the consensus sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser (YSPTSPS).

The important role of CTD is emphasized by the finding that its deletion is lethal in both lower and higher eukaryotes. Phosphorylation of different residues of CTD has a role in transcriptional regulation, although the finding that transcription can proceed *in vitro* even after deletion of CTD points to the fact that the role is an auxiliary one (reviewed in [63]). During initiation, Ser-5 of the CTD is phosphorylated by Cdk7 kinase, a subunit of the general transcription factor TFIIH. Phosphorylation of Ser-5 has a role in the recruitment of capping enzymes to the 5' end of the nascent RNA chain.

After transcription is initiated, PolII will transcribe a short stretch of nucleotides from the template DNA strand and then pause at a position of around 25 to 50 nucleotides downstream of the transcription start site (TSS) [46]. This phenomenon, known as promoter-proximal pausing, is induced by the association of PolII with the DRB-sensitivity inducing factor (DSIF) and negative elongation factor (NELF) [219, 241, 252]. Additionally, since the pause site is in the vicinity of the first nucleosome downstream of the TSS, it is possible that this nucleosome also facilitates promoter-proximal pausing by slowing down PolII progression. However, nucleosomes do not seem to be essential in this process, since PolII pausing can be induced *in vitro* on DNA templates lacking nucleosomes by addition of DSIF and NELF [241]. Phosphorylation of the Ser-2 of CTD by the positive transcription elongation factor b (P-TEFb) complex causes dissociation of NELF from PolII and transition to elongation of RNA.

Promoter-proximal pausing could have several different physiological functions [46]. Firstly, it may result in lower transcript levels of genes. Secondly, it could be involved in keeping certain genes poised for rapid transcription, and possibly in the maintenance of an open chromatin state. A good example of this is the regulation of transcription of "immediate early" genes, genes that are involved in various rapid responses of the cell to external stimuli, such as heat-shock proteins. For example, in normal cellular conditions Hsp70 is kept in a poised state by NELF and DSIF. Then, after the induction of heat shock, P-TEFb is recruited to the gene, the block is released and Hsp70 is rapidly transcribed [10, 236]. Recent genome-wide studies of PolII distributions of human and *Drosophila* genes showed that almost 30% of genes, mostly involved in immediate-early responses and developmental processes, exist in a "poised" state, confirming that PolII pausing is a significant factor in RNA transcription regulation [87, 151]. Conversely, PolII pausing could also be connected to early termination of elongation. In yeast, an early termination pathway causes termination within the first few hundred nucleotides of elongation (for review, see [32]). The pathway is mediated by a termination complex targeted to 5' ends of genes through sequence-specific binding to RNA and association with phosphorylated Ser-5 of PolII CTD. Ser-2 phosphorylation suppresses the early termination pathway and enables efficient elongation. Although early termination has not been explicitly confirmed in higher eukaryotes, the existence of many promoter-associated short RNAs in mammals implies that at least some of the paused polymerases could be involved in such a pathway, resulting in production of truncated transcripts (reviewed in [35]).

Lastly, pausing could serve as a checkpoint for other processes functionally coupled to transcription, such as RNA processing.

Once PolII has cleared the promoter, it can proceed with the elongation of the nascent transcript. PolII catalyzes the formation of the phosphodiester bond between the 3' end of the nascent transcript and an incoming nucleotide triphosphate (NTP). Inside the catalytic center of PolII the nascent RNA 3' end is bound to the *i* site, while the incoming NTP occupies the *i+1* site. After the phosphodiester bond is formed, the newly formed 3' end is translocated from the *i+1* site to the *i* site [156]. A recent model of the mechanism of elongation postulates that this movement is probably induced by the bending of the F bridge helix of PolII. Since the F bridge helix can oscillate between bent and straight conformations, the model proposes that the loading of the required NTP into the *i+1* site prevents the RNA 3' end from occupying this site. In this way Pol II would be stabilized in a forward translocated state and on average backward motion or "backtracking" of PolII would be inhibited. However, binding of an incorrect NTP would favor movement of PolII in the opposite direction, allowing backtracking in the case of incorrectly incorporated nucleotides, which could have an influence in ensuring transcriptional fidelity [15].

During elongation, the forward movement of PolII can be obstructed, resulting in either transcriptional pausing or transcriptional arrest [192]. Transcriptional pausing is a self-reversible process which occurs when PolII stops nascent RNA synthesis for a certain period of time. In contrast, transcriptional arrest is an irreversible process, and cannot be resolved without involvement of accessory factors. The exact mechanisms of inducing transcriptional pause and arrest are still debated, but are likely to include backtracking of PolII and structural rearrangements inside the enzyme [197]. Many general elongation factors function by helping PolII to overcome pausing and arrest, and resume synthesis of nascent RNA. TFIIF, Elongin and ELL have all been shown to be capable of increasing the rate of PolII elongation by decreasing the time of PolII pausing ([192], and references therein). On the other hand, TFIIS helps PolII to overcome transcriptional arrest by stimulating PolII-mediated cleavage of the nascent transcript [171, 194].

Mechanisms of transcript termination are not as well understood as other parts of the transcription cycle. In mammals, termination does not take place at conserved sites, but can occur at a distance of a few kilobases downstream of the end of the mature transcript [167]. Regulation of transcription termination is very important, since defective termination can interfere with transcription from downstream promoters, splicing and degradation of mRNA [143, 232]. It is also important that PolII is correctly released so that it can be utilized in further rounds of the transcription cycle. Termination of most mammalian protein-coding genes is functionally coupled with processing of the 3' end of nascent RNA [174].

In mammals, cleavage and polyadenylation of nascent RNA is mediated by a highly complex 3' end processing machinery. Briefly, a conserved hexanucleotide, AAUAAA, is recognized by the cleavage and polyadenylation specificity factor (CPSF), which is recruited to the body of elongating PolII. Upon transcription of the conserved

hexanucleotide, CPSF binds to this site and induces pausing of PolII. This is followed by the binding of the cleavage stimulation factor (CstF) to a downstream U/GU-rich element. CPSF then binds to CstF, and mediates cleavage of the nascent transcript upstream of the conserved hexanucleotide and downstream of the U/GU-rich region, followed by the release of PolII (reviewed in [174]).

Although pausing of PolII seems to be an important step in achieving termination, it does not by itself necessarily lead to release of PolII. In humans, efficient termination requires the presence of an $5' \rightarrow 3'$ exoribonuclease Xrn2 [231]. According to one model this exoribonuclease degrades the downstream RNA product until it reaches PolII. The collision of Xrn2 with PolII, perhaps helped by the action of the enzyme sentaxin, which possesses a helicase activity, would then cause the release of PolII from the template DNA and termination of transcription [123, 141]. It was recently shown that Ser-7 of PolII CTD can also be phosphorylated during transcription of some protein coding genes. The function of this phosphorylation is not yet known, although the fact that this mark is usually observed at the 3' ends of genes could imply that it is somehow involved in 3' processing and/or transcription termination [40, 64].

1.2.2 Chromatin and transcription regulation

The different steps of the transcription cycle have to be tightly regulated to achieve a precise control of gene expression. Activation of transcription depends largely on the action of sequence-specific transcription factors, which can either promote or repress the recruitment of PolII and/or general transcription factors to the promoter, mainly through interactions with transcriptional coactivators and corepressors [110]. However, since transcription takes place in a chromatin context, its progression is also dependent on the chromatin structure of the gene. This section briefly summarizes the current knowledge on different mechanisms of regulating chromatin dynamics, as well as their influence on regulation of transcription.

All the steps of the transcription cycle, from binding of transcription factors and preinitiation complex assembly to the progression of PolII through the transcript, are affected by the packaging of genomic DNA into the nucleosome [235], a barrier that can be traversed in several different ways to ensure efficient transcription. A study investigating nucleosome positioning in the yeast genome showed that promoters of most yeast genes contain a nucleosome-free region at around 200 bp upstream of the transcription start site [246]. The nucleosome-free regions were shown to be enriched in poly-A and poly-T sequences, suggesting that the loss of nucleosomes is governed by the sequence composition of the promoter region. Furthermore, most functional (occupied) transcription factor binding sites were devoid of nucleosomes, suggesting that the lower nucleosome occupancy somehow facilitates access of transcription factors to the promoter. Alternatively, the binding of transcription factors could cause the exclusion of nucleosomes from the binding sites. Although it seems clear that initial nucleosome positioning is mainly governed by the sequence composition of the

genome, chromatin structure can subsequently be altered through changes of nucleosome position and structure, mediated by ATP-dependent chromatin remodelers and histone chaperone proteins [214]. Chromatin remodelers mainly act by moving nucleosomes along the DNA, evicting nucleosomes or altering the composition of the nucleosome, exerted by either removing certain histone proteins from the nucleosome or replacing the histone protein with a histone variant [49]. Histone chaperones, which can also work in combination with chromatin remodelers, can influence nucleosome stability or serve as a storage for histones [161]. Changes in chromatin structure resulting from the action of chromatin remodelers and histone chaperones can be very important for progressing through the transcription cycle. For instance, histone removal from the yeast PHO5 promoter mediated by the histone chaperone Asf-1 is essential for activation of transcription and recruitment of general transcription factors ([4], and references therein).

Histone modifications are another mechanism of regulation of chromatin dynamics, that has been intensively studied in the context of transcription. It has been found that individual modifications can be associated with transcriptional activation or repression. Acetylation and phosphorylation generally accompany transcription; sumoylation, deimination, and proline isomerization are usually found in transcriptionally silent regions; methylation and ubiquitination are implicated in both activation and repression of transcription [121]. Apart from the type of covalent modification, the final effect on transcription regulation also depends on the residue that is modified and the region of the gene where the modification occurs [131]. Although in general little is known about the exact relationship between individual histone modifications and different steps of the transcription cycle, recent studies have increased our understanding of the mechanisms involved in this regulation, some of which are summarized below.

First of all, histone modifications can function by promoting recruitment of transcription factors to the promoter. This was confirmed by a study of the binding of Myc transcription factor in the human genome, which discovered that only a fraction of binding sites that matched the consensus sequence for Myc binding were actually bound by the protein *in vivo* [86]. Furthermore, the binding was dependent on a specific chromatin signature around the binding site, consisting of high levels of H3K4 and H3K79 methylations and H3 acetylation. This was one of the first indications that transcription factor binding could be regulated by histone modifications, and that sequence-specific recognition of binding sites is dependent on chromatin recognition.

Histone modifications can also influence recruitment of the preinitiation complex (PIC) to the promoter. For example, it was observed that Spt-Ada-Gcn5-acetyltransferase (SAGA) is required for the transcription of about 10% of yeast genes [126]. SAGA is a large multi-protein complex that possesses histone acetyltransferase (HAT) activity. Shukla *et al.* [193] showed that the expression of some SAGA-dependent genes directly depends on its HAT activity and H3 acetylation levels. More specifically, in the PHO84 promoter, H3 acetylation regulates PIC formation,

possibly by decondensation of chromatin, thus making the DNA more accessible to binding of PolII and general transcription factors.

In addition to inducing changes in chromatin structure, histone modifications can also act by recruiting additional histone-modifying or chromatin-remodeling complexes. An example of this is H3K4me3, a modification enriched in promoters of active genes that is usually considered to be a hallmark of active transcription [131]. This modification is established at the nucleosome in the intricate *trans*-tail pathway that involves interactions with PolII and cross-talk with other histone modifications (reviewed in [76]). In yeast, a prerequisite for trimethylation of H3K4 is the monoubiquitination of residue H2BK120, mediated by the PAF protein, an elongation factor associated with the phosphorylated Ser-5 of the C-terminal domain of PolII. Monoubiquitination is required for efficient di- and trimethylation, but not monomethylation, of H3K4 by the methyltransferase Set1, a part of the COMPASS complex ([125], and references therein). An analogous pathway has been identified in human cells, implying that the *trans*-tail pathway is conserved in eukaryotes [114]. It was also recently discovered that some proteins specifically recognize H3K4me3, but not dimethylated or monomethylated H3K4 (reviewed in [249]). This includes different chromatin remodeling factors, such as NURF, which can bind to H3K4me3 and mediate transcription initiation, possibly by inducing nucleosome sliding [133, 238]. A recent study investigating the genome-wide distribution of H3K4me3 showed that this modification, along with PolII, occupies the promoters of most protein coding genes in human embryonic stem cells [87]. However, not all of the genes enriched in H3K4me3 were transcriptionally active. This finding implies that, even though H3K4me3 is usually associated with active genes, it might be involved in regulating initiation, but it is probably not regulating subsequent steps of the transcription cycle which are needed to ensure efficient progression to elongation.

One of the well studied examples of how histone modifications can influence promoter clearance and transition to elongation is H3K27me3, a modification involved in repression of transcription. Trimethylation of H3K27 is catalyzed by the Polycomb repressive complex 2 (PRC2) [33]. H3K27me3 is in turn specifically recognized by a subunit of the Polycomb repressive complex 1 (PRC1) [33, 185]. PRC1 also contains a subunit with a ubiquitin ligase activity, which mono-ubiqutinylates histone H2A at lysine 119 [57, 222], an event which leads to repression of promoter clearance by preventing the recruitment of FACT [202, 250]. FACT is a histone chaperone that normally enables PolII, once stalled at the promoter, to proceed to elongation, by mediating efficient displacement of one H2A/H2B dimer from the nucleosome [163]. If recruitment of FACT is blocked, PolII can not overcome the barrier imposed by the first nucleosome and stays stalled at the promoter.

Several histone modifications, such as H3K36me3, H4K20me1 or H3K27me1, apart from having an influence on transcriptional events in the promoter region (PIC assembly, initiation and promoter clearance), have been found to be enriched in gene bodies of actively transcribed genes [18]. This implies that methylation of histones might have a function in elongation itself. On the other hand, even though coding

regions of transcribed genes are highly enriched in histone acetyltransferases and histone deacetylases, histone acetylation shows minimal enrichment in these regions (reviewed in [14]). One possible explanation for this observation is that acetylation of coding regions is dynamic and coupled with transcribing PolII. Acetylation could function by relaxing the chromatin structure to facilitate PolII progression, before subsequent deacetylation. It has been shown that elongating PolII is associated with Set2, a methyltransferase that catalyzes the establishment of trimethylation of H3K36 ([239], and references therein). H3K36me3 can in turn recruit histone deacetylases and mediate deacetylation of histones in coding regions, in order to reestablish a more condensed chromatin structure and prevent the initiation of cryptic transcription within coding regions [36, 108, 112].

In conclusion, different processes that regulate chromatin structure, such as removal/ replacement of nucleosomes and changes in the nucleosome structure, as well as covalent modifications of histone proteins, can have an influence on regulation of transcription. These different mechanisms can work in combination, as in the case of the *trans*-tail pathway, which provides evidence in support of the "histone code" hypothesis. However, although some mechanisms by which histone modifications influence the transcriptional process have been elucidated, in general little is known about this relationship, making further studies necessary to fully understand it.

1.3 mRNA splicing

1.3.1 Spliceosome assembly and splicing regulation

The final product of the transcriptional process is a complementary RNA copy of the DNA sequence of the transcribed gene, called precursor mRNA (pre-mRNA) [175]. The production of a mature mRNA transcript which can be translated into a protein requires additional processing steps of the pre-mRNA, one of which is the excision of intron sequences during the splicing process. Splicing consists of two transesterification reactions involving three specific sites in the intronic sequence: the 5' splice site (5'SS), the 3' splice site (3'SS) and the branch point sequence (BPS), also called core splicing signals [224]. The splicing process is catalysed by the spliceosome, a large complex composed of five small nuclear ribonucleoprotein particles (snRNPs) and additional auxiliary non-snRNP proteins [175, 220]. Spliceosome assembly begins with the binding of the U1 snRNP to the 5'SS of the intron, binding of splicing factor 1 (SF1) to the BPS and binding of the U2 auxiliary factor (U2AF) to the 3'SS and to a pyrimidine rich region upstream of the 3'SS called the polypyrimidine tract (PPT), forming the E complex [44, 175, 220]. Subsequent steps are executed either through interactions across the intron between factors which recognize the 3'SS and the upstream 5'SS (intron definition) or by interactions across the exon among factors recognizing the 3'SS and the downstream 5'SS (exon definition), a mode of splice site recognition predominant in mammals [25]. SF1 is subsequently displaced from the

BPS by the binding of the U2 snRNP, leading to the formation of the prespliceosome (A complex). The A complex is then converted to the precatalytic spliceosome B complex by the addition of U4/U6 and U5 snRNPs. A catalytically active spliceosome (C complex) is formed after additional conformational changes and loss of U1 and U4 snRNPs ([44, 175, 220]; Fig. 1.2).

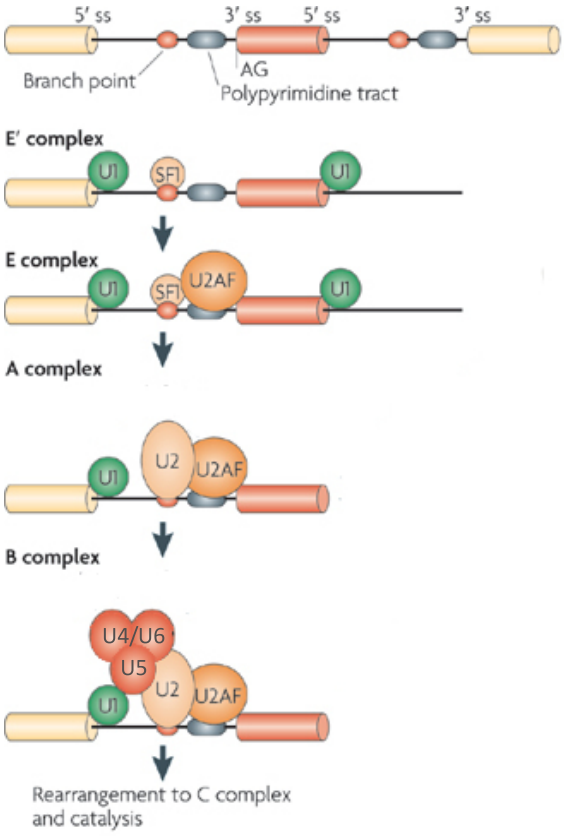

Figure 1.2: Spliceosome assembly. Schematic representation of the splicing process and spliceosome assembly (adapted from Chen and Manley, [44])

However, not all of a gene's exons are necessarily included in the final mature mRNA. The process of selecting different splice sites during splicing is called alternative splicing, and it results in the production of various transcript isoforms containing distinct combinations of exons [224]. It is currently estimated that over 90% of human genes undergo alternative splicing, with different isoforms usually being expressed in a tissue-specific manner, influencing numerous differentiation and developmental processes [221]. Fig. 1.3 depicts the most frequently observed forms of alternative splicing, whose combinations can then lead to more complex alternative splicing outcomes. Mutations that disrupt splicing patterns have been connected with the onset of various genetic diseases, such as spinal muscular atrophy, Duchenne muscular dystrophy and cystic fibrosis [50], emphasizing the need for tight regulation of the splicing process. Splice site selection and the final structure of the spliced transcript is influenced by various factors.

Recognition of splice sites depends on their affinity for splicing factors U1 and U2AF, involved in the formation of the E complex [44]. Genomic annotation of 5' splice site and 3' splice site positions in the genome permitted the alignment of splice site sequences and identification of a consensus sequence [248]. Moreover, this has also led to the development of more complicated statistical models for identification of novel splice sites [164, 243]. In humans, binding of U1 and U2AF occurs through complementary base pairing, and disruption of the consensus sequence leads to reduced binding of the splicing factors to pre-mRNA [177, 253]. The consensus sequence of the branch point sites of human introns, bound by SF1 and later U2, is degenerate, suggesting that additional *cis*- and *trans*-acting elements are needed to enhance its recognition [77]. Although the splice site consensus sequences are better defined, even the presence of strong splice sites, matching the consensus sequence to a high degree, is not a definite predictor of whether the exon will be included in the final transcript. This is demonstrated by the existence of pseudoexons, intronic sequences surrounded by splice sites closely matching the consensus sequence. However, pseudoexons are never or very rarely spliced [204], implying that only part of the information for splice site selection is encoded in the core splicing signals themselves, with additional signals being required for accurate splicing of pre-mRNAs.

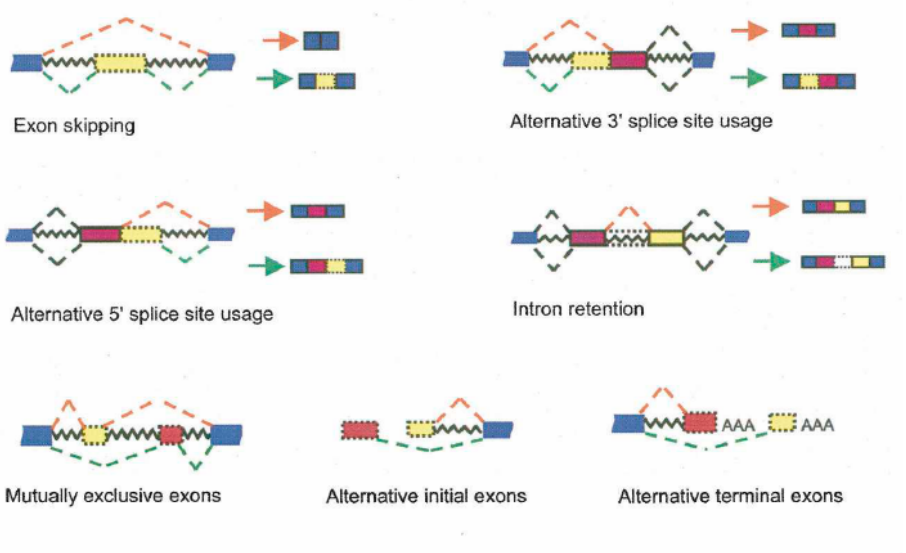

Figure 1.3: Most frequent forms of alternative splicing. Schematic representation of the most frequently occurring forms of alternative splicing (adapted from Wang et al. [224])

Apart from core splicing signals, there are also numerous additional *cis*-regulatory elements which can either promote or suppress splice site recognition, called splicing regulatory elements (SREs). These elements are classified in four different categories, depending on their position and influence on splice site recognition: exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs) and intronic splicing silencers (ISSs) [44, 224]. Various SREs have been identified by molecular genetics techniques, mainly by analyzing mutations which lead to a change

in splicing patterns of specific genes [224]. Apart from these individual approaches, different methods for genome-wide identification of SREs have been developed. These methods generally rely on two different approaches: (1) computational searches for (possibly conserved) overrepresented motifs in a particular group of intronic or exonic sequences and (2) experimental methods based on SELEX experiments used to find sequences that either influence splicing or bind to particular known splicing factors (reviewed in [41]).

SREs can be bound by various proteins, which can then either promote or repress splice site recognition. ESEs are usually bound by members of the SR protein family [44]. The binding of SR proteins then facilitates splice site recognition by recruitment of components of the splicing machinery, such as U1, U2AF and U2, to 5' and 3' splice sites [47, 82, 142]. Although splicing regulation mediated by ISEs is not well understood, various proteins binding to these elements have been identified, such as the T-cell restricted intracellular antigen I (TIA1), which promotes the recognition of weak splice sites by recruiting the U1 splicing factor to the nascent mRNA [72, 73, 79]. ESSs and ISSs are usually bound by heterologous nuclear ribonucleoprotein particles (hnRNPs) [44]. Splicing inhibition can then be achieved by steric hindrance of the binding of snRNPs. A well known example of this mechanism is the binding of the polypyrimidine tract binding protein (PTB, also called hnRNP I) to the polypyrimidine tract, an event that represses the binding of U2AF and splicing of the downstream 3' splice site [198]. hnRNPs can also function by inhibiting the binding of activator proteins to splicing enhancers, as in the case of hnRNP A1 repressing the splicing of HIV-1 *tat* exon 3, which is normally promoted by binding of SR protein SC35 [251]. Another model, supported by NMR structural analysis of the four RNA binding domains of PTB in complex with RNA, proposes that binding of PTB to several different PPTs can bring distal PPTs into close proximity and induce RNA looping [157]. Positioning of an exon or a branch point sequence inside the RNA loop could then prevent binding of splicing factors or spliceosomal assembly. Some SRE-binding proteins are expressed in a tissue-specific manner, among them FOX1 and FOX2 proteins, which are preferentially expressed in muscle, heart and neurons [107, 217], and brain specific factors nPTB, Nova1 and Nova2 [105, 144]. Even the core proteins involved in spliceosome assembly, such as snRNPs, show differential expression among various human tissues, further contributing to tissue-specific regulation of alternative splicing [85].

The aforementioned example of repression of binding of SC35 by hnRNP A1 shows that alternative splicing can be regulated by combinatorial effects of different SREs. There is also accumulating evidence that the specific effect of some SRE-binding proteins depends on the position of the SRE in the sequence. For instance, PTB, Nova and FOX2 proteins all repress exon inclusion when bound to SREs upstream of the 3' splice site, but enhance exon inclusion if they bind to elements downstream of the 5' splice site [136, 216, 244], thus adding a further level of complexity to alternative splicing regulation. A recent study used a set of 1,014 genomic variables, comprised of various RNA motifs and other transcript features, to model tissue-specific regulation of 3,665 cassette-type alternative splicing events in 27 different mouse tissues, also

taking into account the position of the various features in the genomic sequence [16]. This model, which achieved quite high prediction accuracy among some sets of exons, was a first attempt at establishing a code for splicing determined by combinations of different features that could govern splice site selection.

1.3.2 Transcription-coupled splicing

There is substantial evidence that transcription and splicing can occur simultaneously in the nucleus. The first report of co-transcriptional splicing came from a study of *Drosophila melanogaster* embryos, where electron microscopy was used to show that spliceosome formation can occur while the transcript is still in contact with the template strand [28], and co-transcriptional splicing has since then been observed in many additional genes of various species [23, 208, 237]. The discovery of the simultaneous occurrence of the two processes points to the possibility of their functional coupling. The existence of functional coupling was confirmed by the observation that the type of promoter used to drive transcription influences the outcome of alternative splicing [52, 53]. Two different models of how transcription could influence splicing have been proposed [120]. The "recruitment model" proposes that PolII and/or various transcription factors influence the efficiency of splicing through their interactions with splicing factors. This model is supported by the observation that the C-terminal domain (CTD) of PolII has a role in targeting splicing factors to transcripts *in vivo* and that the truncation of the CTD inhibits pre-mRNA splicing [150]. A second model, called "the kinetic model", postulates that the rate of transcription elongation influences splice site selection. The kinetic model assumes that in the case of mutually exclusive exons with competing 3' splice sites, a weak one followed by a strong one, both 3' splice sites will be presented simultaneously to the splicing machinery and splicing will occur at the stronger splice site if the rate of PolII elongation is high. However, if PolII elongation is slow, the weaker splice site will be recognized first and the intron bounded by this splice site will be spliced out [120]. One of the studies showing strongest evidence in support of this kinetic model used a slow mutant of PolII to inhibit skipping of the fibronectin EDI exon in human cells [56].

1.3.3 Chromatin and alternative splicing

In addition to various RNA-binding motifs, tissue-specific expression of splicing factors and PolII elongation rate, chromatin structure was also recently identified as a potential regulator of alternative splicing outcomes. This relationship was first confirmed by a study that used mammalian cells transfected with minigenes carrying the fibronectin EDI exon to investigate mechanisms of alternative splicing regulation. The study showed that treatment of the cells with trichostatin A, an inhibitor of histone deacetylation, decreases the inclusion of the EDI exon in mature mRNA

transcripts. This effect was believed to be a consequence of increased PolII processivity, caused in turn by increased acetylation levels of chromatin due to inhibition of histone deacetylases [155].

The recent development of methods for genome-wide mapping of chromatin structure provides further evidence in support of its role in the regulation of alternative splicing. Digestion of genomic DNA using micrococcal nuclease (MNase) produces mononucleosome-sized DNA, which can then be sequenced using next-generation sequencing techniques and mapped back to the genome to produce high-resolution nucleosome positioning maps. Several independent studies analyzed genome-wide nucleosome positioning data for human CD4+ T-cells and *C.elegans* mixed-tissue population of cells. These studies described an unusual pattern of nucleosome enrichment in internal exons of both human and *C. elegans* genes [9, 101, 186, 200, 211]. This effect was shown not to be a consequence of GC content or protein-coding potential of exons, and is mostly not dependent on gene expression level. Later studies confirmed exonic nucleosome enrichment in additional cell types, as well as other species, such as *A. thaliana* and *O. latipes*, further supporting the notion that this is a stable and evolutionarily conserved phenomenon [48, 59, 153].

This finding raised the question of the functional significance of the enrichment of nucleosomes in exons. One possible explanation is that nucleosomes play a role in exon definition. This hypothesis was mainly inspired the observation that the average size of mammalian exons and the length of the DNA sequence which is wrapped around the nucleosome are quite similar [137]. An additional possibility is that nucleosomes are involved in the regulation of alternative splicing, supported by results of several analyses of nucleosome enrichment patterns at intron-exon boundaries of alternative and constitutive exons, which showed that constitutive exons and alternative exons that are included in the transcript have higher nucleosome occupancy levels than alternative exons with low frequency of inclusion [186]. The enrichment of nucleosomes is negatively correlated with splice site strength. Included alternative exons with weak splice sites have high levels of nucleosome occupancy. On the other hand, pseudoexons, DNA sequences that are surrounded by splice sites but show no evidence of being efficiently spliced, show a depletion of nucleosomes. Taken together, these observations could indicate that high nucleosome occupancy promotes the inclusion of alternative exons with weaker splice sites and possibly represses splicing of pseudoexons with strong splice sites [200, 211]. Nucleosomes could influence splicing by altering PolII elongation rate, causing it to pause at sites of nucleosome accumulation, an effect that was previously observed in various studies [98, 104], and is supported by the fact that the levels of PolII are indeed higher in exons [48, 59, 186]. Paused PolII could subsequently influence alternative splicing either through mechanisms described by the recruitment model or by the kinetic model (see Section 1.3.2).

In addition to nucleosomes, the levels of some histone modifications seem to be related to the intron-exon structure of a gene. This was first shown in a study of highly synchronized populations of *C. elegans* larval stages, where the authors studied the genome-wide distribution of three histone modifications and noticed that levels of

H3K36me3, but not H3K4me3 or H3K9me3, are highly enriched in exons compared to introns, a result that was further corroborated by identification of an analogous pattern of H3K36me3 enrichment in exons of human and mouse genes [117]. Additional studies identified further modifications whose levels differ between introns and exons [9, 59, 100, 101, 186, 200, 211], although in some cases this enrichment could merely be a reflection of the underlying nucleosome density. However, some histone modifications exhibit a significant difference in distribution between introns and exons even after normalization for nucleosome levels. For example, exonic enrichment of H3K36me3 and depletion of H3K9me3 was observed by two independent studies [59, 200]. The finding that levels of H3K36me3 (normalized for nucleosome occupancy) differ between alternative and constitutive exons [59, 117], implies that not only nucleosomes, but also histone modifications, could have a prominent role in regulation of alternative splicing.

One possible way in which histone modifications could regulate alternative splicing is an influence on PolII processivity, exerted by modulating the local chromatin condensation, a mechanism proposed in the initial study which recognized the detrimental effect of inhibition of histone deacetylases (HDACs) on the inclusion levels of fibronectin EDI exon [155]. In support of this mechanism, it was recently shown that inhibition of HDACs causes increased acetylation levels of histone H4 and enhanced PolII processivity in the vicinity of the EDI exon, causing it to be excluded from the mature fibronectin transcript [97]. Furthermore, a recent study showed that membrane depolarization of neuronal cells causes skipping of exon 18 of the neural cell adhesion molecule (NCAM) pre-mRNA, and that the skipping of the exon was associated with increased levels of H3K9ac and H3K36me3, as well as increased chromatin accessibility [184]. The finding that enrichment of H3K36me3 is connected with exon skipping is opposite to observations from recent genome-wide studies (see above) and implies that the same histone modification can have various impacts on alternative splicing, possibly depending on additional features of the exon. The alternative splicing outcome of exon 18 is dependent on PolII elongation rate, confirmed by the fact that transcription of NCAM by a slow mutant polymerase results in the inclusion of exon 18. These results are consistent with the theory that histone modifications could influence alternative splicing through modulation of PolII processivity, supported by a previous observation that hyperacetylation of histones leads to an increase in PolII elongation rate and a decrease in PolII pausing [166].

Another potential mechanism of regulation of alternative splicing is the effect of direct or indirect interactions between histone modifications and various components of the splicing machinery [137]. A recent study investigated the histone modification profiles of exons IIIb and IIIc of the human fibroblast growth factor receptor 2 (FGFR2) gene using chromatin immunoprecipitation (ChIP) followed by real-time qPCR [139]. These two exons undergo mutually exclusive alternative splicing in a tissue-specific manner, regulated by binding of PTB to regions surrounding exon IIIb, resulting in its exclusion from the transcript. Analysis of the distribution of histone modifications in the FGFR2 gene showed an enrichment of H3K36me3 and H3K4me1 and depletion of H3K27me3, H3K4me3 and H3K9me1 in the region surrounding the IIIb exon in

the cases where it is excluded from the transcript. This observation was confirmed on several other PTB-dependent exons, while PTB-independent exons did not exhibit such a profile. The authors also showed that modulation of expression levels of H3K36 methyltransferase SET2 and H3K4 methyltransferase ASH2 influences the inclusion of exon IIIb. This observation confirms a causal relationship between histone modifications and alternative splicing, which is apparently mediated by MRG15, a protein which specifically binds H3K36me3, leading to subsequent recruitment of PTB to the nascent RNA and repression of exon IIIb inclusion [139]. Additional studies report recruitment of components of the splicing machinery by histone modifications mediated through interactions with chromatin-binding proteins (e.g. the recruitment of U2snRNP by H3K4me3 through an interaction with CHD1 complex [196] or the recruitment of hnRNPs by HP1 protein bound to H3K9me3 [165]). These findings suggest the existence of further combinations of interactions through which histone modifications could participate in the regulation of the splicing process [137].

Taken together, the results of these recent studies imply that chromatin structure, more specifically positioning of nucleosomes and enrichment of certain histone modifications, does play a role in the regulation of alternative splicing. However, even though some insight into possible mechanisms of regulation has been gained, we are still far away from fully understanding them. For example, the influence of H3K36me3 on alternative splicing of exons, mediated by its interaction with MRG15 and recruitment of PTB, seems to be much stronger for exons regulated by weak PTB binding sites than for strongly PTB-dependent exons [139]. This finding indicates that alternative splicing regulation depends not only on histone modifications but also on the sequence features of the exons and their surrounding regions. This suggests that histone modifications and/nucleosome occupancy work in combination with other means of splicing regulation, such as RNA-binding motifs, tissue-specific expression of splicing factors and the effect of PolII. It is possible that different combinations of these mechanisms are responsible for alternative splicing of distinct subgroups of exons, implying that further studies, using both experimental and theoretical approaches, will be needed to achieve a better understanding of these complex processes.

1.4 mRNA stability and degradation

Levels of transcripts present in the cell are regulated not only by the rate of mRNA transcription, but also by post-transcriptional mechanisms controlling mRNA stability and degradation. These mechanisms involve various *cis*-acting elements, such as conserved sequence elements and secondary structures, and *trans*-acting factors, including mRNA binding proteins and microRNAs [20, 88]. In this section we give a brief overview of some of the best studied mechanisms of post-transcriptional regulation of physiological mRNA levels.

1.4.1 AU-rich elements

AU-rich elements (AREs) are sequence elements located in 3' untranslated regions (UTRs) of some mRNAs rich in A and U nucleotides. These elements play a role in the regulation of mRNA stability and translation by recruiting various ARE-binding proteins. Association of AREs with ARE-binding proteins commonly results either in a reduction of stability and rapid degradation of mRNA or in decreased efficiency of translation of the transcribed mRNA, although there are some cases where an opposite effect of ARE-binding proteins on mRNA stability and translation has been described [17]. AREs were first discovered in 3' UTRs of genes involved in inflammatory response [34], and were initially believed to be restricted to a small subset of mRNAs. However, later studies estimated that up to 8% of human genes encode mRNAs that contain ARE sequences [13]. The expression of many of these genes requires precise spatial and temporal control, as in the case of genes whose functions are related to cell growth or response to external stimuli [13, 113].

AREs have been roughly classified into three different classes (I, II and III). Although a real consensus sequence has not been defined, the different classes of AREs are usually characterized by a uridine rich region, which can be accompanied by the presence of a number of (often overlapping) pentamer (AUUUA) or nonamer (UUAUUUAWW; W = A or U) motifs [43].

1.4.2 MicroRNAs

MicroRNAs (miRNAs) are an endogenous class of short non-coding RNAs, approximately 21-22 nucleotides long, expressed in both plant and animal species. miRNAs are involved in post-transcriptional regulation of gene expression, since they can pair to mRNAs of protein-coding genes and induce their repression [20, 69].

miRNA genes are transcribed by PolII and processed to produce a precursor molecule called primary-miRNA (pri-miRNA) [129]. The pri-miRNAs are folded into hairpin structures and can often include sequences for several mature miRNAs. Pri-miRNA precursors are processed in two steps, catalyzed by enzymes Drosha and Dicer, to produce miRNA duplexes [128]. Typically one strand of the duplex is bound by an Argonaute protein, to form a part of the RNA-induced silencing complex (RISC) (reviewed in [20]). Once the mature miRNA is loaded into RISC, it can pair with a target mRNA and direct post-transcriptional repression. miRNA target sites are most often located in 3' UTRs of transcripts, and are believed to be primarily recognized by complementary base pairing between the mRNA sequence and the nucleotides in positions 2-8 of the miRNA, called the miRNA "seed". miRNAs bound to the target transcripts most commonly induce either translational repression or a reduction of the stability of the transcribed mRNA [69].

Many different methods for predicting functional targets of metazoan miRNAs have been developed. While all of them are based on identifying matches between the

3' UTR regions of transcripts and the seed region of the miRNA, they also utilize various additional features, such as target site conservation and accessibility, pairing of miRNA and mRNA outside of the seed region or the context in which the target site occurs, to reduce the number of false positive predictions (reviewed in [21]).

1.4.3 Nonsense-mediated decay

Nonsense-mediated decay (NMD) is a quality control mechanism for gene expression that is coupled with translation [39]. NMD degrades aberrant transcripts containing premature termination codons (PTCs), which are most commonly created by random nonsense and frameshift mutations, programmed DNA rearrangements or splicing errors [75, 134, 146]. The translation of such transcripts could lead to truncated proteins, potentially producing deleterious dominant-negative or gain-of-function effects [39].

In mammalian cells, NMD is triggered by the exon junction complex (EJC), a protein complex which is deposited upstream of exon-exon junctions following pre-mRNA splicing [94]. In normal transcripts EJCs are located upstream of the stop codon and displaced by the ribosome. However, in transcripts harboring premature stop codons, the ribosome will dissociate from the mRNA before it reaches the EJC. EJC will then recruit other core factors of the NMD pathway, thereby triggering steps which will lead to mRNA decay (reviewed in [39]).

1.5 Experimental methods for determining transcript levels and chromatin structure

This section describes some of the main experimental methods used to generate publicly available data analyzed in this thesis.

1.5.1 DNA microarrays

Microarray technology is used to measure quantities of target molecules in biological extracts, based on the strength of hybridization between a target molecule and a probe attached to a solid surface [65, 181]. In DNA microarrays the probes are short DNA oligonucleotide fragments whose sequence and position on the array is precisely defined. Probes can be designed to match the sequences of known genes, promoters, exons or other regions of interest. The targets, which can be either a genomic DNA or mRNA (cDNA) sample, are labeled, most commonly with a fluorescent probe. The array is then incubated with the sample until hybridization occurs. After hybridization, the fluorescence intensity of each probe-target pair is quantified, giving an estimate of the strength of hybridization. The strength of hybridization should be

proportional to the concentrations of the probe and the target, therefore this signal can be used to determine the relative abundance of the target molecule [229].

DNA microarrays initially became popular because they enabled measurements of expression levels of many genes at the same time [182]. In gene-expression studies, RNAs harvested from two or more samples are converted to cDNA and labeled with different fluorescent dyes. In two-color systems, both samples are incubated on the array simultaneously. Then for each probe-target pair the intensities of fluorescence of the two dyes are measured and compared, in order to determine differences in expression between the two samples. In one-color systems, each sample is incubated on a separate array, and fluorescence intensities are compared to quantify relative expression levels of genes in different samples [162]. Apart from their widespread use in gene expression measurements, microarrays have been used for various additional purposes, such as genotyping, alternative splicing detection or DNA resequencing [99]. They can also be used in combination with chromatin immunoprecipitation (ChIP, see below) to conduct genome-wide studies of DNA-protein binding [172].

1.5.2 Next-generation sequencing

In the past few years various high-throughput sequencing methods have been developed, collectively termed next-generation sequencing (NGS) approaches. These methods rely on massively parallel sequencing, enabling them to produce a large amount of sequencing reads in a relatively short time, when compared to conventional Sanger sequencing [190]. There are various commercially available NGS platforms, such as the 454 FLX (Roche Applied Science), Solexa Genome Analyzer (Illumina), SOLiD (Applied Biosystems), HeliScope (Helicos Biosciences) and PacBio RS (Pacific BioSciences) [147]. Although the exact protocols differ between platforms, they all follow the same general steps, starting with the preparation of the template for sequencing, produced by random fragmentation of genomic DNA. The fragmented DNA is then usually amplified (clonally amplified templates), although recently methods that do not require amplification of templates have been developed (e.g. single-molecule templates, used by HeliScope and PacBio RS). Methods using single-molecule templates have the advantage of requiring smaller amounts of starting material and not being susceptible to amplification bias. In both approaches, templates are immobilized on a solid surface prior to sequencing. The sequence is then determined in cycles, where in each cycle the complementary strand is extended using either incorporation of dye-labeled nucleotides or ligation of dye-labeled probes. Before entering the next cycle fluorescence is used to determine the identity of the nucleotide or the probe, usually followed by cleavage of the fluorescent dye. The length of sequenced reads ranges from 25 bp to over 900 bp, depending on the platform used, with up to 50 Gb of read data generated in a single sequencing run (reviewed in [147]).

One of the many applications of NGS methods is the study of transcriptomes, entire sets of RNA molecules produced in the cell, using RNA sequencing (RNA-Seq) [225]. In RNA-Seq a library of cDNA fragments is prepared from an RNA sample and

sequenced using NGS technologies, and the sequenced reads are then either assembled *de novo* or aligned to the reference genome or transcriptome. The mapped reads can then be used for precise determination of gene and exon boundaries, detection of alternative splicing events or discovery of novel transcripts. Furthermore, since the number of reads mapped to a given transcript is proportional to the abundance of the transcript in the original RNA population, RNA-Seq can be used for quantification of levels of various transcripts, including mRNAs, small RNAs and non-coding RNAs (reviewed in [225]). Other applications of NGS methods include SNP discovery, whole genome resequencing and genome-wide profiling of chromatin structure (ChIP-Seq, see below) [92, 160].

1.5.3 High-throughput methods for chromatin structure determination

In recent years, two high-throughput methods for mapping genome-wide nucleosome occupancy have been developed. Both of these methods are based on digestion of genomic DNA by micrococcal nuclease (MNase), a bacterial enzyme that degrades the linker DNA connecting two nucleosomes, in order to produce a sample containing mononucleosomes. Mononucleosomal DNA is then extracted and either sequenced using NGS technologies (MNase-Seq) or hybridized to a DNA microarray (MNase-Chip) to reveal the positions of nucleosomes in the genome [5, 127, 246].

High-throughput methods for localization of histone modifications in the genome are based on chromatin immunoprecipitation (ChIP), a method for mapping DNA-protein interactions *in vivo* [199]. In ChIP, mononucleosomes are produced either by crosslinking the histone proteins to the DNA and subsequent sonication or by digesting the linker sequence using MNase. Mononucleosomes are then immunoprecipitated using antibodies against specific histone modifications, histone proteins or histone variants. The DNA is extracted from purified DNA-protein complexes by proteolytic digestion. The extracted DNA then serves to construct a library that can either be hybridized to an array (ChIP-Chip) or sequenced using NGS methods (ChIP-Seq) to determine the sequences of isolated fragments [96].

1.6 Thesis overview

Histone modifications are associated with regulation of various cellular processes. Although many genome-wide studies of localization and function of histone modifications have recently been conducted, our knowledge of the exact mechanisms of their functions is still somewhat limited. The objective of this thesis is to model the relationships of histone modifications with mRNA transcription and mRNA splicing, in order to gain a better understanding of their effect on these two processes.

The relationship between histone modifications and transcription is studied in Chapter 2. We will develop linear models to show that the nature of this relationship is quantitative. We will furthermore show that the relationships between histone modifications and transcription hold universally across cell types. We will then identify histone modifications that are most informative of the expression levels of genes, and study these modifications in different types of promoters. This approach will enable us to further investigate their possible role in the regulation of the transcription cycle. In the end, we will investigate the influence of different degradation rates of transcripts on the results of our analysis.

Chapter 3 focuses on the recently discovered link between histone modifications and alternative splicing. We will develop models to predict the outcome of alternative splicing from features connected to the chromatin structure of the alternative exons. We will also present results showing that alternative splicing is influenced by expression levels of analyzed transcripts. Finally, we will investigate the relationship between various mechanisms of splicing regulation, including chromatin structure and different sequence elements known to have an influence on splicing.

In Chapter 4 we will discuss novel findings presented in Chapters 2 and 3 in the context of the current knowledge about the relationship of chromatin structure and various cellular processes, and propose future improvements of our analysis which could provide a better understanding of these relationships.

Chapter 2

Modeling gene expression levels using histone modifications

The presence or absence of certain histone modifications has been shown to correlate with the expression status of human genes [131]. In this chapter we present the results of an analysis in which we used machine learning methods to address four major questions: (i) Is there a quantitative relationship between histone modifications levels and transcription? (ii) Are there histone modifications that are more important than others to predict transcript levels? (iii) Are there different requirements for different promoter types? (iv) Are the relationships general? We furthermore studied how the different degradation rates of transcripts influence the results of our analysis.

2.1 Histone modifications are highly predictive of gene expression

Various recent studies showed that the levels of different histone modifications are correlated with the expression status of human genes [131]. However, in general, little is known about the exact nature of the relationship between histone modifications and transcription. We considered that there are two possible mechanisms of action of histone modifications: (1) The levels of modifications have to exceed a certain threshold to determine the on/off status of a gene, or (2) the levels of the modifications encode the expression level and are therefore quantitatively related to transcript abundance. To investigate whether the relationship between histone modifications and gene expression is indeed quantitative, we analyzed publicly available genome-wide localization data for 38 histone modifications and one histone variant in human CD4+ T-cells (henceforth "modifications"), which was produced by chromatin immunoprecipitation followed by next-generation sequencing (ChIP-Seq) [18, 228]. We derived linear models to quantitatively relate the levels of histone modifications at the promoter to the expression levels of genes [183].

We downloaded the RefSeq Genes annotation track for the human genome sequence (hg18, March 2006) from the University of California, Santa Cruz Genome Bioinfor-

matics web site [1]. We also acquired the coordinates of uniquely mapped ChIP-Seq tags for 19 lysine or arginine histone methylations, one histone variant, and 19 histone acetylations in CD4+ T-cells [18, 228]. In the ChIP-Seq experiments conducted in these studies, only the DNA corresponding to the ends of immunoprecipitated mononucleosomes is sequenced. In our analysis, the coordinates of ChIP-Seq tags were transformed by adding or subtracting 73 base pairs (for tags mapping to the + or - strand, respectively) to center the tags on the nucleosome. For 27,212 RefSeq genes, we counted the number of tags in a 4,001 base pair region surrounding the transcription start site (TSS). The tags in this region were summed and each gene was represented by 39 values (one per modification). In order to control for the bias of open chromatin regions preferentially having higher tag counts than closed chromatin regions [11, 209], we also obtained ChIP-Seq data for unspecific ChIPs using goat and rabbit IgG antibodies in CD4+ T-cells [227] and counted these tags in the aforementioned regions. Since these antibodies do not specifically bind to mononucleosomes, the coordinates of the tags were not transformed by adding or subtracting base pairs. Expression microarray data for resting T-cells performed on Affymetrix Human Genome U133 Plus 2.0 GeneChips was taken from Schones *et al.* [183]. Raw expression values were averaged over all replicates and only the RefSeq genes that could be uniquely mapped to an Affymetrix probe identifier were used in further analysis.

The fact that some RefSeq genes correspond to alternative transcripts of the same gene could unjustly bias the results of our analysis. For instance, if many cases of alternative transcripts with similar expression levels and histone modification profiles appear in the model, the fit of the final model could reflect the relationships in these subsets of closely related transcripts, rather than a general mechanism present in all genes. To avoid this possible bias we restricted our analysis to only one transcript per gene. We therefore mapped all RefSeq genes to their corresponding Unigene clusters, which represent sequences that appear to come from the same transcriptional locus [233]. We divided values for each of the 39 modifications by the highest sum of tag counts over all promoters, so that modifications with a globally lower number of tags also had an influence in deciding which promoter to use for analysis. We averaged the values for 39 different modifications for each promoter. In each Unigene cluster, we kept only the promoter that had the highest average amount of ChIP-Seq tags in a 4,001 base pair region surrounding the annotated transcription start site for further analysis. We further reduced the gene set by removing all Affymetrix probes which could be mapped to more than one Unigene cluster and by removing all Unigene clusters which contained RefSeq genes whose exons overlapped by more than 20%. The final set comprised 14,802 RefSeq genes.

We used the levels of histone modifications in the promoter region of the gene as predictor variables in a linear regression model. Each modification i and promoter j was represented by the sum of tag counts N_{ij} in the 4,001 base pair region surrounding the TSS. We first transformed the sum of tag counts N_{ij} for each modification i and promoter j to a logarithmic scale. We added pseudocounts α_i to each N_{ij}, to avoid undefined values of the logarithm when N_{ij} equals zero. The pseudocount α_i was

determined by the following procedure. The original set D of 14,802 promoters was divided into two random sets *D1* and *D2*. We then optimized the pseudocounts α_i on the random set *D1* (4,934 promoters). For each modification i, the search space for α_i ranged from zero to the maximal value of N_i. Each α_i was used to transform N_{ij} into $N'_{ij} = log(N_{ij} + \alpha_i)$ and the correlation between N'_{ij} and the logarithm of measured expression values was determined on *D1*. The value α_i which maximized this correlation was then chosen to compute N'_{ij} in *D2* for all the remaining analyses. For the remaining 9,868 promoters we built a linear regression model where the entire set of modifications and the control IgG data served as predictor variables, and the logarithm of expression values served as a response variable Y (Eq. 2.1, referred to as the "full model").

$$Y = \beta_0 + \sum_{i=1}^{41} \beta_i N'_i \tag{2.1}$$

We estimated the prediction accuracy for the linear regression model using a 10-fold cross-validation setting to ensure that a possible quantitative relationship is of general nature and not limited to a certain subset of genes. We implemented the 10-fold cross-validation procedure as follows. We used set *D2* to train 10 linear models, with 90% of the data used for training and 10% for testing the predictions of the linear model. The test sets for each linear model were non-overlapping. The predicted expression value for each promoter corresponds to the value predicted by the linear model in which this promoter was used for testing. The prediction accuracy of the linear regression was calculated as the Pearson correlation coefficient r between predicted and measured values of expression.

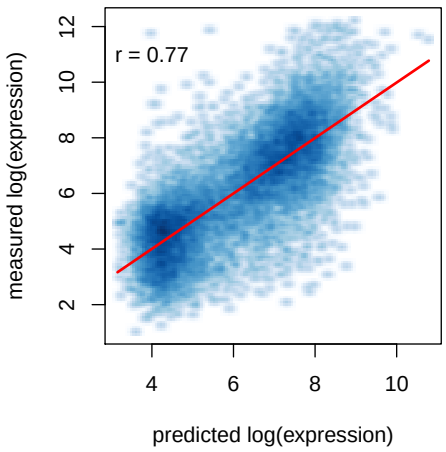

Figure 2.1: Prediction of gene expression levels measured by microarrays in CD4+ T-cells. Scatterplot of the measured expression values against expression values predicted using the full linear model. The color shading encodes the local densities at each point in the scatterplot, with darker shades indicating higher densities. The red line indicates the linear fit between predicted and measured expression values (y = x + 0.02).

The expression values predicted by the linear regression model are very well correlated to expression measured by microarray experiments ($r = 0.77$, p-value of t-test $< 2.2 \cdot 10^{-16}$; Fig. 2.1). This finding clearly demonstrates that the amounts of histone

modifications at the promoter are predictive of the expression levels of genes and that the relationship between levels of histone modifications in the promoter region and gene expression is a quantitative one.

Probes used in microarray experiments can differ in their affinity [132], introducing a bias in measurements of absolute expression levels of different genes. To ensure that this bias does not influence our results, we repeated the analysis using expression levels of genes in CD4+ T-cells measured by next-generation sequencing of mRNA (RNA-Seq) [45]. We mapped the RNA-Seq tags to 14,802 genes used in the analysis and summed the number of tags mapping to transcribed regions. In order to produce a measure of expression value normalized for the length of the gene, we divided the sums by the number of base pairs in exonic regions. We added a pseudocount of 1 to these length-normalized sums. The logarithm of normalized tag counts was taken as a measure of expression level. Pseudocounts optimized for these expression values were added to the histone modification levels, which were then transformed to a logarithmic scale (as described above). The prediction of expression using histone modifications was conducted in the same way as described for the analysis using microarray data. The prediction accuracy for modeling RNA-Seq derived expression values is even higher ($r = 0.81$; Fig. 2.2) than the one using microarray expression data ($r = 0.77$; Fig. 2.1), confirming that the results of our analysis are not significantly influenced by possible measurement biases due to the microarray technology.

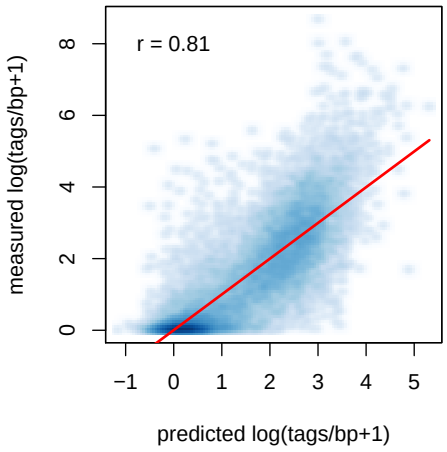

Figure 2.2: Prediction of gene expression levels measured by RNA-Seq in CD4+ T-cells. Smoothed color density representation of the scatterplot of the measured expression values against expression values predicted using the full linear model. Darker shades indicate higher densities. The red line indicates the linear fit between predicted and measured expression values ($y = x + 0.003$).

2.2 Identifying most informative histone modifications

A recent study of genome-wide localization of histone modifications showed that many histone modifications, referred to as "backbone modifications", frequently co-occur in the genome and their levels are highly correlated [228]. This finding could imply

that the functions of some of these modifications are redundant. Alternatively, a subset of modifications could be sufficient for regulation of transcription, recruiting further modifications and thus explaining the high correlation between them. We decided to test whether all histone modifications are equally important for successful modeling of gene expression using linear regression and use feature selection to identify modifications whose levels harbor most of the information about gene expression levels.

Feature selection is often conducted by using stepwise procedures to identify subsets of relevant predictor variables [91]. Some feature selection algorithms start by training the model on all individual predictor variables and identifying the variable which results in the most predictive model. Then, at each round of feature selection the feature which provides the highest gain in prediction accuracy is added to the model (forward-stepwise selection). Alternatively, one can train a model on the complete set of predictor variables, and remove the least informative variable at each step of feature selection (backward-stepwise selection). However, since the levels of histone modifications at the promoter are highly correlated, neither forward- nor backward-stepwise selection are appropriate for our analysis. Namely, in the case of correlated predictor variables the contribution of individual variables to the variance explained by the model will depend on the selection order, because of the overlap of information contained in the correlated variables [30]. This implies that using stepwise selection approaches could result in omitting histone modifications which are important for prediction of gene expression levels, but are also highly correlated to other predictor variables, from the final model.

We instead built linear models where all possible combinations of one to three histone modifications were used to model gene expression levels. The different models, henceforth referred to as "one-modification models", "two-modifications models" and "three-modifications models", are described by Eq. 2.2, 2.3 and 2.4, respectively, where Y is the logarithm of gene expression, and N'_i is the optimized value of ChIP-Seq tag counts for a histone modification, histone variant or unspecific control antibody used in the model. We then compared the prediction accuracy of linear models using different combinations of modifications in order to determine which are the most informative ones. All possible one-modification (41 models), two-modifications (820 models), and three-modifications models (10,660 models) were produced and their performance assessed as described in Section 2.1.

$$Y = \beta_0 + \beta_i N'_i \qquad i \in \{1, ..., 41\} \tag{2.2}$$

$$Y = \beta_0 + \beta_i N'_i + \beta_j N'_j \qquad i, j \in \{1, ..., 41\}\,;\; i \neq j \tag{2.3}$$

$$Y = \beta_0 + \beta_i N'_i + \beta_j N'_j + \beta_k N'_k \qquad i, j, k \in \{1, ..., 41\}\,;\; i \neq j \neq k \tag{2.4}$$

Analysis of the prediction accuracy of one-modification, two-modifications and three-modifications models shows that predictions of some of these models are very well correlated to gene expression (Fig. 2.3). For example, we determined that the top one-modification ($r_{max} = 0.72$, H3K27ac), two-modifications ($r_{max} = 0.74$, H3K27ac + H4K20me1) and three-modifications models ($r_{max} = 0.75$, H3K27ac + H3K4me1 + H4K20me1) almost reach the prediction accuracy of the full model ($r_{full} = 0.77$). On the other hand, the lowest ranking models have almost no predictive power. These results establish that not all modifications are equally important for prediction of gene expression, possibly because of a high degree of redundancy. Moreover, the levels of a single modification (H3K27ac) can be used to faithfully model gene expression.

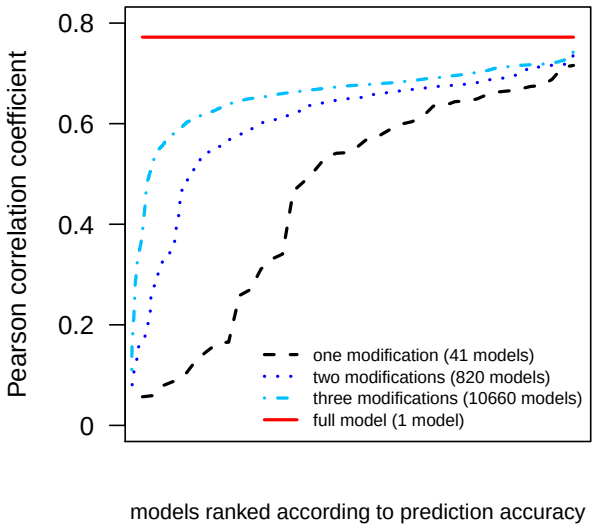

Figure 2.3: Comparison of the performance of models using combinations of different numbers of modifications. Prediction accuracy of all possible one-modification, two-modifications and three-modifications models compared to the prediction accuracy of the full model for CD4+ T-cells. Models are sorted by ascending prediction accuracy along the x axis.

Even though some of the models using one histone modification can be used to successfully predict gene expression levels, the prediction accuracy nevertheless increases as one goes from the best one-modification model to the full model. Using more modifications increases the number of free parameters in the model, ie. model complexity. An increase in model complexity generally leads to a better fit of the model to the training data, but can also lead to reduction in generalization and poor predictive power on unseen data [91]. However, the increase in accuracy observed in our analysis is not simply due to higher model complexity, because the prediction accuracy of the models is computed on test data. To confirm this, we used the Bayesian information criterion (BIC) [188], a model selection criterion derived within a Bayesian framework which takes into account the maximum likelihood of the model L, the number of free parameters k and the sample size n (Eq. 2.5). BIC tries to identify the optimal model among a set of candidate models by penalizing the increase in the number of parameters, with the optimal model being the one which minimizes the BIC value. We produced linear models using all possible combinations of 1-5 and 37-41 modifications (there are too many 6-36 modification models, so we excluded them). In each group of models (corresponding to the number of modifications used in combination),

the model with highest prediction accuracy was identified, and the trade-off between model complexity and prediction accuracy assessed using the BIC. The BIC value keeps decreasing continuously for models using 1-5 modifications, suggesting that it is not the model complexity which governs the increase in prediction accuracy. However, the BIC values decrease only slightly after using more than four modifications (Fig. 2.4). Furthermore, the BIC value shows a slight increase when increasing the number of modifications from 37 to 41. Taken together, our results suggest that the levels of as few as three modifications at the promoter are enough to faithfully model expression of the associated gene, and the addition of further modifications does not significantly improve the goodness of fit of the model.

$$BIC = -2 \cdot log(L) + k \cdot log(n) \tag{2.5}$$

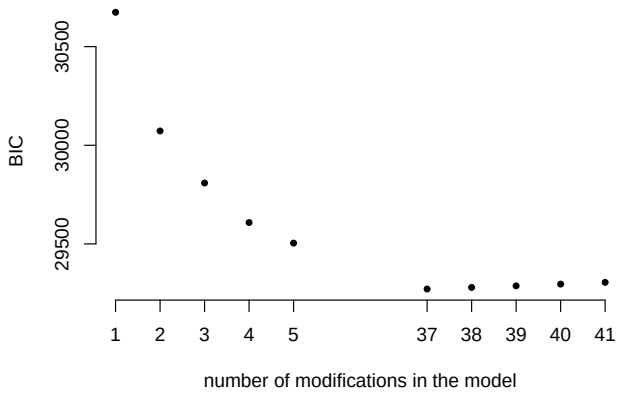

Figure 2.4: Comparison of BIC values for models using combinations of different numbers of modifications. Plot of the BIC of the best linear models using combinations of 1-5 and 37-41 variables (histone modifications and IgG control). Best model is chosen as the one with the highest prediction accuracy, measured by the Pearson correlation coefficient *r*.

To identify modifications whose levels harbor most of the information about gene expression we focused on the three-modifications models. We produced all 10,660 possible three-modifications models and assessed their prediction accuracy as described in Section 2.1. We determined all three-modifications models where the Pearson correlation coefficient r between measured and predicted expression values reached at least 95% of the one obtained by the full model ($r_{full} = 0.77$). There were 142 models that satisfied this criterion, which is a sufficiently high number to justify an overrepresentation analysis by computing the probability of observing a particular modification in that many subsets due to chance alone. The number of times each modification appears among this set of models was divided by the number of best scoring models to determine the fraction of appearance of each histone modification. Our results show that four histone modifications, H4K20me1, H3K27ac, H3K79me1, and H2BK5ac (Fig. 2.5), are significantly overrepresented in the set of models (p-values of the hypergeometric test $7.58 \cdot 10^{-50}$, $8.95 \cdot 10^{-46}$, $7.83 \cdot 10^{-30}$, and $2.88 \cdot 10^{-27}$, respectively), each of them appearing in roughly half of the studied models (57.7%, 54.9%, 42.9%, and 40.8%, respectively). The remaining histone modifications appear in at most 7% of the models, a frequency expected from random sampling (p-value

of the hypergeometric test 0.47). Goat and rabbit IgG were found in only a small number of the best models (2.11% and 3.52%, p-values of the hypergeometric test 0.99 and 0.95, respectively), which shows that they do not contribute significantly to the prediction accuracy.

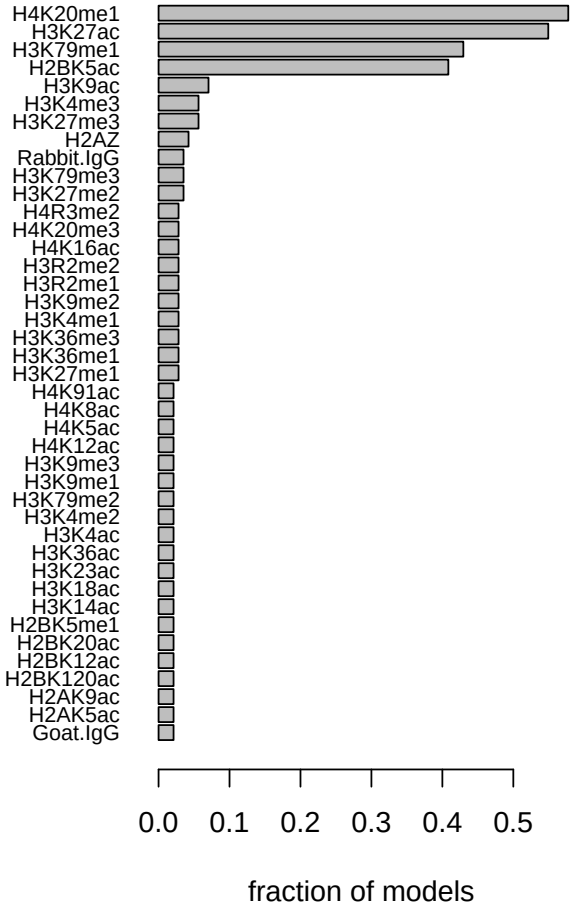

Figure 2.5: Overrepresentation analysis for all promoters. Bar plot showing the frequency of appearance of different histone modifications in best scoring three-modifications models (142 models) for CD4+ T-cells.

The appearance of goat and rabbit IgG in a small number of best models, along with the fact that the prediction accuracy of one-modification models trained on these variables is low ($r_{goat\_IgG} = 0.15$, $r_{rabbit\_IgG} = 0.09$; Fig. 2.6), shows that the high prediction accuracy of linear models using histone modifications as predictors is not merely a consequence of higher accessibility of open chromatin. The result of the overrepresentation analysis is robust to variations of the threshold used to define best scoring models, presuming that the set of best scoring models does not exceed 20% of the total number of models, which then naturally leads to random inclusion of other histone modifications (Fig. 2.7). Thus, H4K20me1, H3K27ac, H3K79me1, and H2BK5ac appear to be the most important modifications associated with gene expression levels.

Interestingly, the prediction accuracies of the one-modification models, based on the

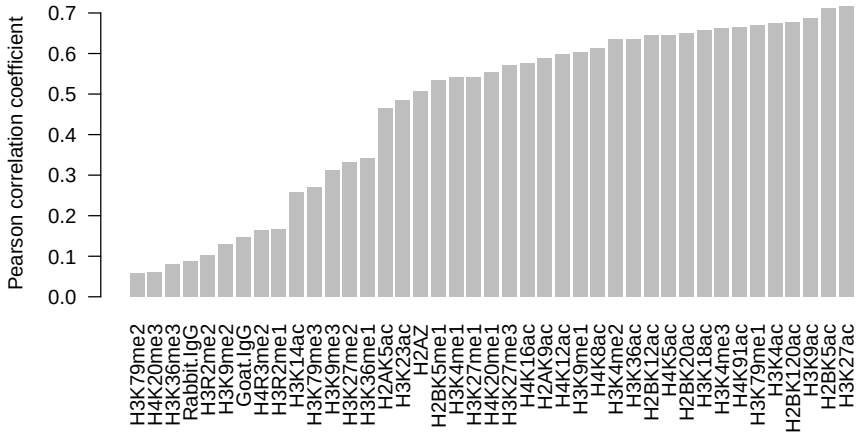

Figure 2.6: Prediction accuracy of one-modification models. Bar plot showing the Pearson correlation coefficient *r* of measured expression levels and expression levels predicted by one-modification linear models.

overrepresented modifications only, greatly vary ($r_{H3K27ac} = 0.72$, $r_{H2BK5ac} = 0.71$, $r_{H3K79me1} = 0.67$, and $r_{H4K20me1} = 0.55$; Fig. 2.6). Furthermore, the two modifications with the highest individual information content, H3K27ac and H2BK5ac, appear only twice together in the set of best scoring models (Fig. 2.8), suggesting that the information they provide is redundant, which is supported by the finding that their levels are highly correlated ($r = 0.97$). H4K20me1 and H3K79me1 occur together in only three of the 142 models, indicating that they are at least partially redundant. Moreover, we found that in almost all 142 models (92.95%), H3K27ac or H2BK5ac occur together with either H4K20me1 or H3K79me1.

Several histone modifications, such as H3K4me3, H3K27me3 or H3K36me3, have frequently been associated with transcriptional regulation [131]. These histone modifications achieve different results when used to predict gene expression ($r_{H3K4me3} = 0.66$, $r_{H3K27me3} = 0.57$, and $r_{H3K36me3} = 0.08$; Fig. 2.6). H3K36me3, although it was previously shown to be correlated with active transcription, has almost no predictive power. The most probable explanation of this result is that H3K36me3 usually enriched across coding regions of the genes [18], and not the promoter region, which we studied in this analysis. On the other hand, H3K4me3 and H3K27me3, associated with activation and repression of transcription, respectively [131], are predictive of gene expression. However, they do not appear among the most informative modifications, suggesting that despite their individual correlation with gene expression, combinations of other histone modifications provide more information on the outcome of the transcriptional process.

Using linear models, we showed that levels of histone modifications at the promoter are highly predictive of the expression level of human genes (Fig. 2.1), a result which supports the existence of a quantitative relationship between gene expression and histone modifications. However, other studies investigated the relationship between

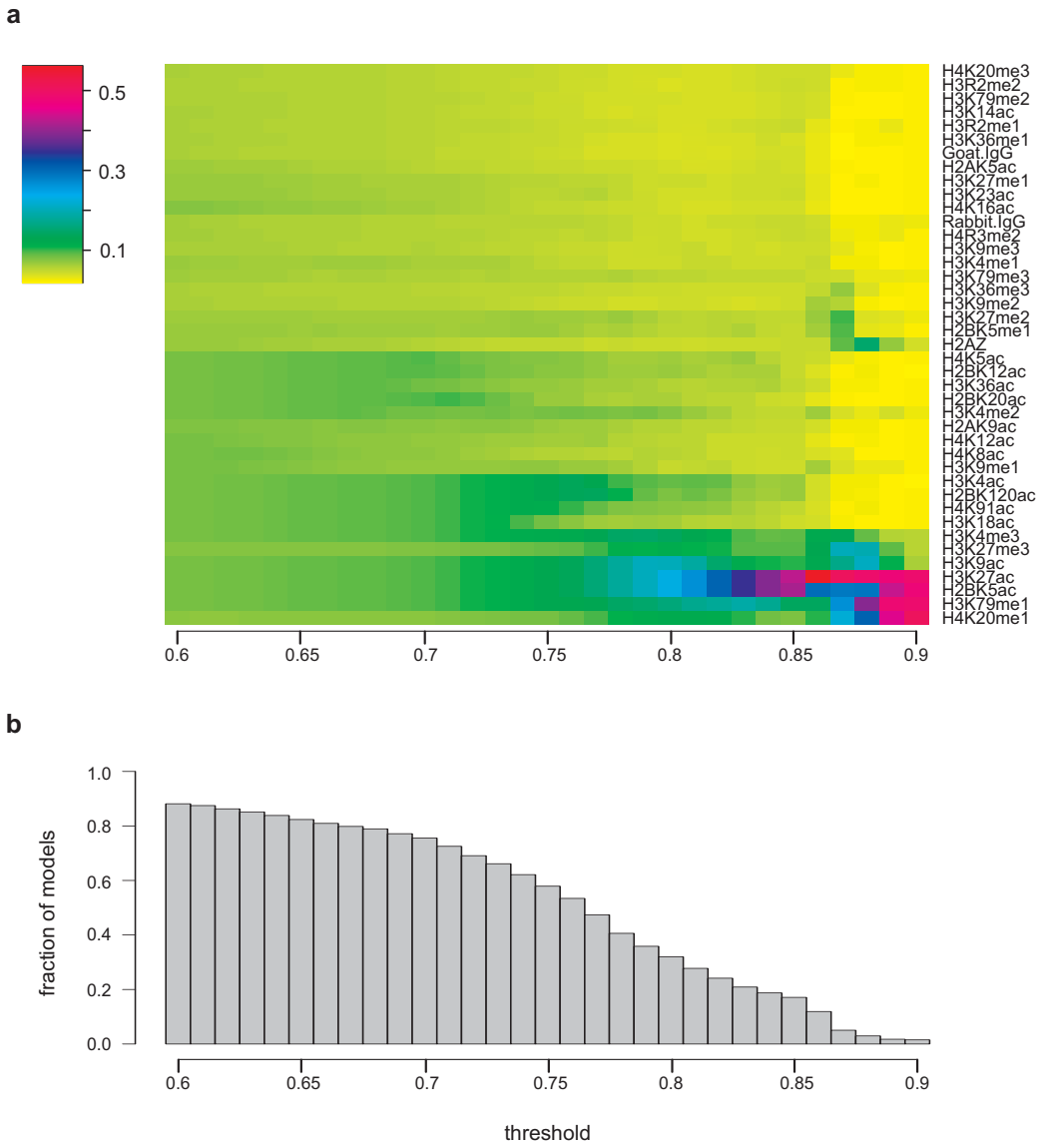

Figure 2.7: Overrepresentation analysis is robust to variations of the threshold used to define best scoring models. (a) Frequency of appearance of histone modifications in best scoring models depending on the threshold of prediction accuracy, defined as the ratio of the correlation obtained with the three-modifications model and the correlation obtained with the full model.(b) Fraction of three-modifications models (10,660 models) with prediction accuracy higher or equal to the threshold.

histone modifications and gene expression by discretizing the levels of histone modifications and classifying the promoters for each modification into groups [228, 245], e.g., modification X is present or absent at the promoters of certain genes. This approach implies that histone modifications influence the on/off status of a gene, rather then determining the exact level of gene expression. If this is indeed the case, discretization should be beneficial for modeling gene expression. To test this assumption, we decided to train linear models using discretized values of histone modifications, and compare their performance with models trained on the measured modification

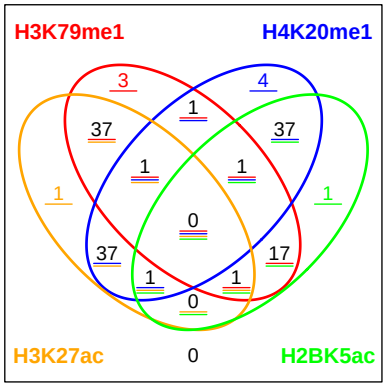

Figure 2.8: Co-occurrence of the important histone modifications in the best scoring models. Venn diagram showing the co-occurrence of the four important histone modifications in the best scoring three-modifications models (142 models). Best scoring models are defined as reaching at least 95% of Pearson correlation coefficient obtained by the full model.

levels.

We conducted the discretization of modifications using the procedure described by Wang *et al.* [228]. A particular modification was considered enriched at the promoter when the ChIP-Seq tag count was higher than a threshold determined with Benjamini-Hochberg's method (using the default false detection rate of the R function p.adjust [168]). The background model, used to compute the p-values necessary for the threshold estimation, follows a Poisson distribution parameterized by the genome-wide tag density. For each histone modification and each promoter, we considered the modification to be enriched if the p-value was lower than 0.01 and assigned it a value of one. For all promoters, all modifications that were not determined to be enriched were assigned a value of zero. We trained the full linear model and all possible three-modifications models using this discretized data in the same way as described before for raw (continuous) tag counts of histone modifications (Section 2.1).

We compared full models incorporating either the levels directly (continuous model) or a binary classification of them (discrete model). Although the difference in the Pearson correlation coefficients obtained by the models is not very large ($r_{full\_continuous}$ = 0.77 and $r_{full\_discrete}$ = 0.74; Fig. 2.1 and Fig. 2.9 (a), respectively), the mean squared error (MSE) increased from 1.54 for the continuous model to 1.71 for the discrete model. The same is true for the best three-modifications continuous and discrete models. Here, the discrete model is only able to reproduce the general trend in expression values and thus has a higher MSE (MSE = 1.84; Fig. 2.9 (b)) than the continuous model (MSE = 1.68, which is even lower than the MSE for the full discrete model; Fig. 2.9 (c)).

Since the discrete models do not perform better than the corresponding continuous models, we conclude that discretization has no beneficial effect on the prediction accuracy and argue that in our modeling framework discretization is not necessary and is even reducing the predictive power at the cost of increasing the number of parameters.

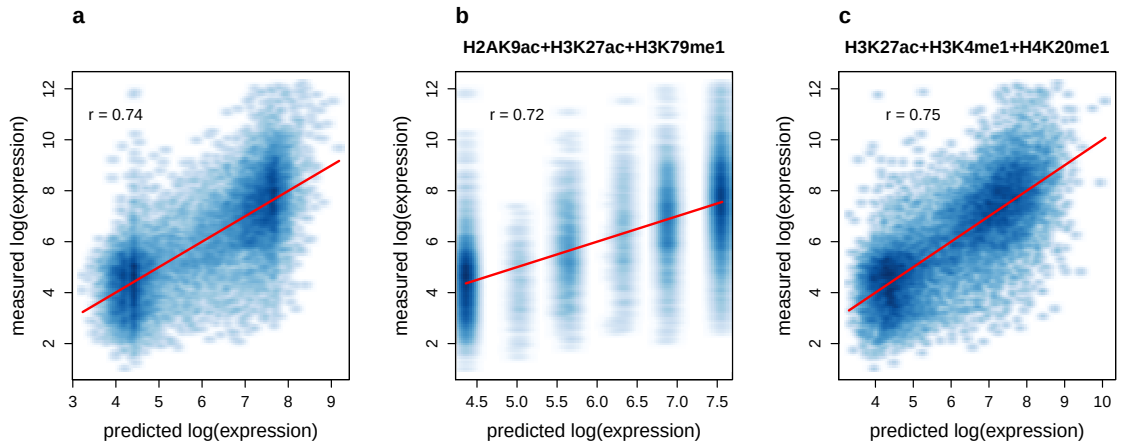

Figure 2.9: Comparison of the performance of models using continuous and discretized data. (a) Scatterplot of measured and predicted log(expression) values using a full linear model trained on discretized tag counts for 39 histone modifications and two control unspecific IgG antibodies in CD4+ T-cells. (b, c) Scatterplots of measured and predicted log(expression) values using the three-modifications model with highest prediction accuracy trained on discretized (b) or continuous (c) tag counts in CD4+ T-cells. The modifications used to train the model are indicated on each plot.

2.3 Requirement of histone modifications depends on the CpG content of the promoter

Given the good agreement between modeled and measured expression values, we proceeded with further analysis of our models to infer the relationships between distinct histone modifications and different groups of promoters. More specifically, we wanted to investigate whether the predictive power of different histone modifications depends on the CpG content of the promoter.

The human genome is in general depleted of CpG dinucleotides. The exceptions are regions of high CpG dinucleotide frequency, called CpG islands, which are often associated with gene promoters [78, 207]. However, not all promoters of human genes overlap CpG islands. It was recently discovered that human genes can be classified into two distinct groups according to the CpG content in the promoter region, high-CpG content promoters (HCP) and low-CpG content promoters (LCP) [180]. Genes whose expression is regulated by HCPs are usually ubiquitously expressed and connected to "housekeeping" or developmental functions, while LPCs are in general associated with highly tissue-specific genes [66, 148, 180]. Moreover, it was shown that histone proteins of nucleosomes in HCPs and LCPs are differently modified [148]. Nucleosomes in HCPs are almost always decorated with H3K4me3, and an additional H3K27me3 mark is added when they are in the repressed state. On the other hand, nucleosomes in LCPs carry the H3K4me3 modification only when they are expressed. We reasoned that if HCPs and LCPs are differently marked by histone modifications then the predictive power of histone modifications should also differ between these

two groups of promoters.

In order to divide promoters according to their CpG content we first calculated the normalized CpG content in the region of 3,000 base pairs surrounding the TSS as defined by Saxonov *et al.* [180]. We classified the promoters with a normalized CpG content greater than 0.4 as HCP, and the others as LCP. We then proceeded to train the full linear model on both groups separately in a 10-fold cross-validation setting, as described in Section 2.1. Since the set *D1* was used for the optimization of pseudocounts of histone modification levels, only promoters in set *D2* were used for determining the regression parameters of the full models for HCPs (7,089 promoters) and LCPs (2,779 promoters). The prediction accuracy for LCPs ($r = 0.72$) is comparable to HCPs ($r = 0.75$), indicating that the accuracy of the model does not depend on the CpG content of the promoters (Fig. 2.10).

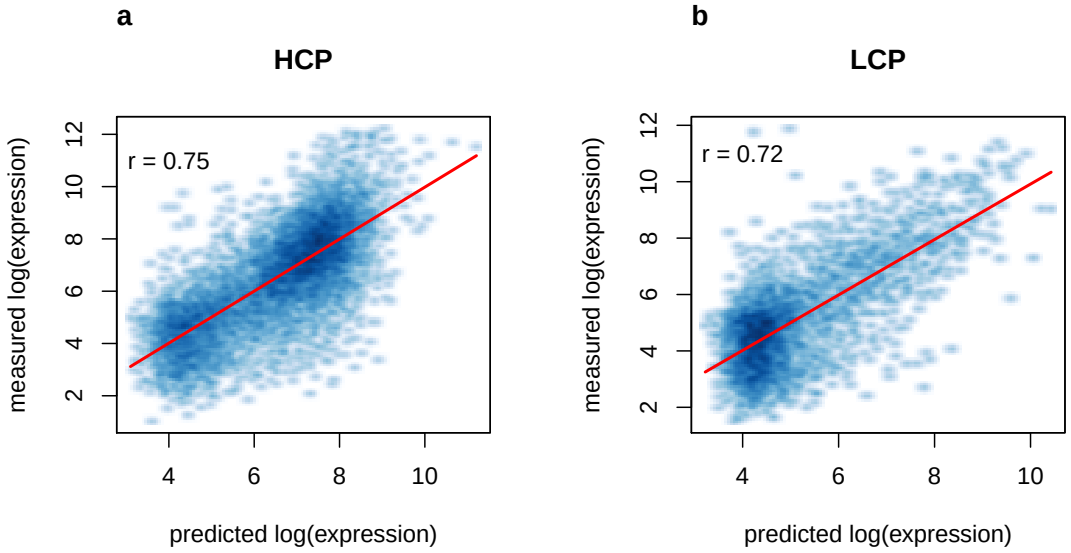

Figure 2.10: Prediction of gene expression for HCPs and LCPs. Smoothed color density representation of the scatterplot of measured expression values against expression values predicted by the full linear model for genes divided according to the promoter CpG content. The prediction accuracy is comparable for HCPs (a) ($r = 0.75$, equation obtained after linear regression: $y = 0.99x + 0.03$) and LCPs (b) ($r = 0.72$, equation obtained after linear regression: $y = 0.98x + 0.09$). The color shading encodes the local densities at each point in the scatterplot, with darker shades indicating higher densities.

We next wanted to determine if the predictive power of histone modifications differs between LCPs and HCPs. We therefore built linear models with all combinations of one, two, and three modifications, for both sets of promoters separately, and determined the overrepresented modifications in best scoring three-modifications models, as described in Section 2.2.

For one-modification models, we found that the best prediction accuracy for both HCPs and LCPs was achieved when H3K27ac was used as the predictor variable ($r_{H3K27ac,LCP} = 0.65$, $r_{H3K27ac,HCP} = 0.68$). This result is in agreement with the

results obtained using all promoters, where the model trained on H3K27ac was also determined to be the most predictive one ($r_{H3K27ac,all} = 0.72$).

We proceeded to examine models using combinations of two histone modifications for training. For HCPs the overall ranking of models remained very similar to the ranking of models determined for all promoters. The best performance, for both HCPs and all promoters, was obtained using the combination of H3K27ac and H4K20me1 ($r_{H3K27ac+H4K20me1,HCP} = 0.71$, $r_{H3K27ac+H4K20me1,all} = 0.74$). These results are hardly surprising because HCPs constitute 72% of all analyzed promoters, suggesting that the results for all promoters were dominated by HCPs. However, for LCPs, the ranking of the models changed compared to all promoters. Strikingly, although the most accurate one-modification model for LCPs was the one trained on H3K27ac, the best performing two-modifications model did not contain this modification. Instead, the model with the combination of H3K4me3 and H3K79me1 performed best ($r = 0.69$, compared to H3K27ac and H3K79me1 $r = 0.67$). We take this result as evidence that our modeling approach is able to identify combinatorial actions of histone modifications that are not just additive.

Next, we determined the overrepresented modifications in the best performing three-modifications models. H4K20me1 and H3K27ac (and possibly H2BK5ac) are significantly overrepresented among the best scoring models for HCPs (p-values of the hypergeometric test $9.97 \cdot 10^{-43}$, $2.58 \cdot 10^{-31}$, and 0.003, respectively), and H3K4me3 and H3K79me1 are significantly overrepresented in the LCPs (p-values of the hypergeometric test $9.71 \cdot 10^{-36}$ and $2.1 \cdot 10^{-34}$, respectively), demonstrating that different modifications are important for the prediction of expression of genes in these two groups (Fig. 2.11).

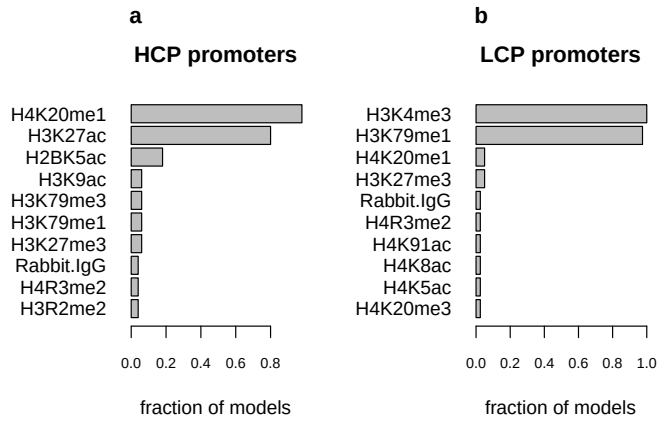

Figure 2.11: Overrepresentation analysis for HCPs and LCPs. Bar plots showing the frequency of appearance of different histone modifications in best scoring three-modifications models for (a) HCPs (50 models) and (b) LCPs (40 models) in CD4+ T-cells. Best scoring models are defined as reaching at least 95% of prediction accuracy of the full model trained on HCPs and LCPs, respectively. Only the top ten modifications are depicted.

To gain further insight into the possible functions of the histone modifications that were highly correlated with gene expression for HCPs and LCPs, we examined the average tag densities for these five modifications in the region surrounding the transcription start site (TSS) (Fig. 2.12), referred to as "localization analysis". For each

modification, we first calculated the average number of tags mapped to each position in the region of -500 to +3,000 base pairs around the TSS, taking into account all 14,802 analyzed genes. We then normalized the profiles for all modifications by dividing the average number of tags at each position by the maximum average number of tags over all positions. This allowed the direct comparison of modification profiles, since the average tag values for all modifications, which normally vary greatly in magnitude, were scaled to a range between 0 and 1. We found that H3K4me3, H3K27ac, and H2BK5ac have the highest levels at the promoter, with the highest peaks around 100 base pairs downstream of the TSS. H3K79me1 is enriched along the gene body, and H4K20me1 shows two distinct patterns: a peak close to the promoter at a similar position to H3K4me3 and H3K27ac, and a further enrichment across the gene body region.

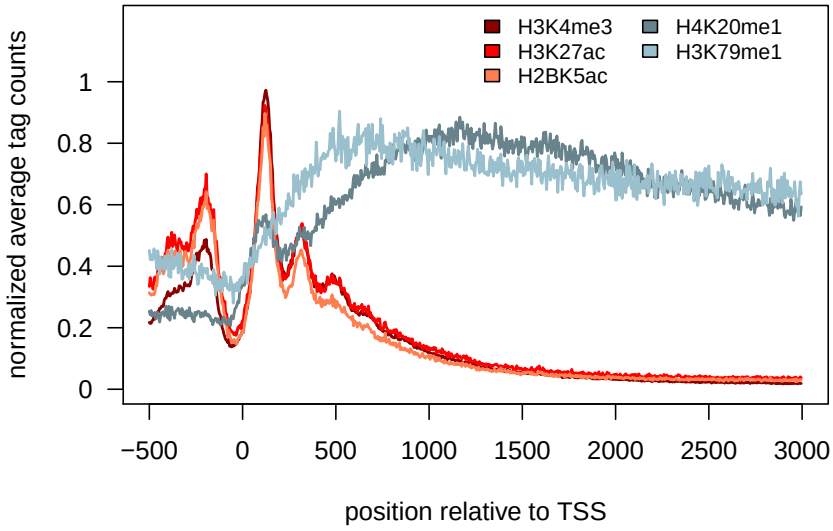

Figure 2.12: Localization analysis of important modifications. Normalized cumulative tag counts in the region of -500 to +3,000 base pairs surrounding the transcription start site of RefSeq genes in CD4+ T-cells for the five important modifications identified by our analysis.

A possible explanation of the observation that the predictive power of histone modifications differs for genes with high-CpG and low-CpG content promoters is that these two groups of genes are regulated at different transcriptional steps. This hypothesis is supported by the finding that H3K4me3 is more informative of the expression levels of LCP than HCP genes. Since H3K4me3 is thought to be a mark of transcriptional initiation ([87], and references therein), this implies that LCPs are regulated at the initiation step. On the other hand, this step does not seem to be rate limiting for HCPs, suggesting that they are regulated at later steps of the transcription cycle, ie. transition from initiation to elongation. A recent study indeed showed that most protein-coding genes in human embryonic stem cells undergo transcription initiation and have nucleosomes marked by H3K4me3, even though only a subset of them later

proceeds to elongation and the production of the full transcript [87], confirming that the expression of at least a subset of human genes is regulated at steps following transcription initiation.

If expression driven by LCPs and HCPs is indeed regulated at different steps of the transcription cycle then there should be a difference in accumulation of RNA polymerase II (PolII) enzymes along the genes associated with these promoters. More specifically, we reasoned that HCPs should almost always harbor an initiating PolII, irrespective of their expression level. On the other hand, if LCPs are regulated at initiation or the steps preceding it PolII should accumulate only at the promoters of actively transcribed genes. To challenge this idea, we analyzed the profile of PolII occupancy [18], both in the promoter and the gene body of HCP and LCP genes used in our analysis. Both groups were further divided into either highly or lowly expressed genes, where lowly expressed genes were the ones whose expression value was lower than the median expression value of the corresponding group. For each gene we mapped the ChIP-Seq tags for PolII to the transcribed region, as well as to upstream and downstream regions corresponding to 20% of the length of the transcribed region. We then counted the number of tags in windows corresponding to 2% of the length of the transcribed unit and calculated the average profile of PolII occupancy in the four groups divided according to the promoter type (HCP versus LCP) and expression level (highly expressed versus lowly expressed).

We found that highly expressed genes with HCPs have very high levels of PolII at their promoters compared to corresponding genes with LCPs, although the PolII densities along the gene body are indistinguishable (Fig. 2.13). In the case of lowly expressed genes, we observed that the level of PolII in HCPs is much higher than in corresponding LCPs, almost approaching the level for highly expressed LCPs. Thus, we conclude that PolII found in HCPs is preferentially regulated at the transition to elongation, while neither preinitiation complex formation, PolII recruitment, nor initiation seem to be rate limiting. LCPs however, are preferentially regulated at the steps preceding the transition to elongation, since PolII enrichment at the promoter is only observed in highly expressed LCP genes. We conclude that different histone modifications are important for prediction of expression of genes with HCPs and LCPs and that these histone modifications are probably related to different transcriptional steps. Involvement in distinct transcriptional steps could furthermore explain the difference in average enrichment profiles of analyzed modifications, observed during localization analysis (Fig. 2.12).

We previously used expression values measured by RNA-Seq to show that the prediction accuracy of the linear model is not unjustly influenced by possibles biases due to microarray technology (Section 2.1). We proceeded to investigate whether the results of the overrepresentation analysis can also be reproduced using RNA-Seq [45] instead of microarray data. Briefly, we trained all possible three-modifications models for all promoters, HCPs and LCPs using expression values measured by RNA-Seq as response variables. We then identified the best scoring models and repeated the overrepresentation analysis, as described in Section 2.2. The overrepresentation analysis

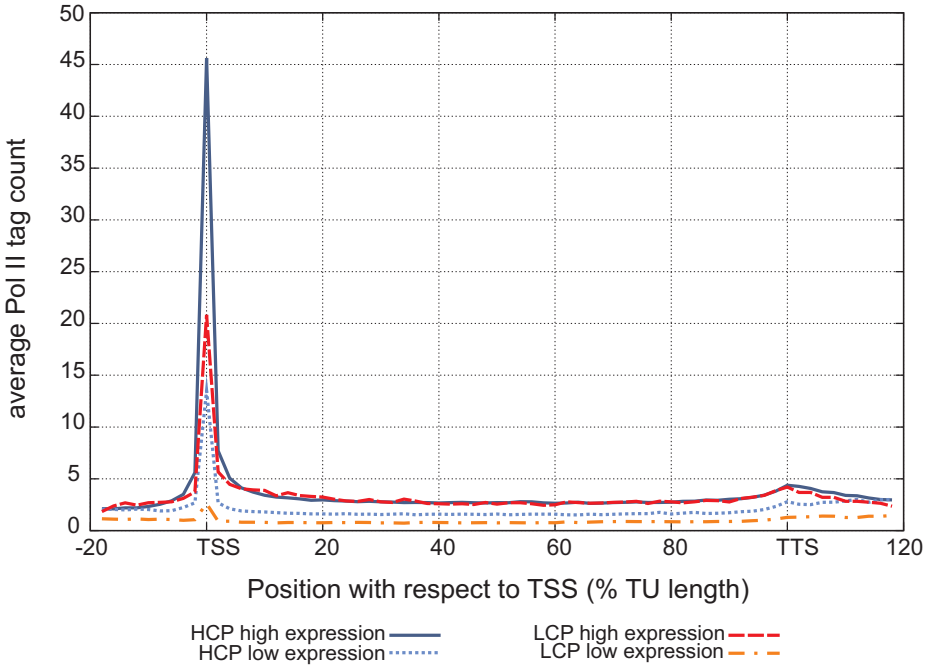

Figure 2.13: Average profile of PolII occupancy in groups of genes divided according to the promoter type and expression level. The average number of PolII ChIP-Seq tags is calculated in windows corresponding to 2% of the length of the transcribed unit. TSS - transcription start site, TTS - transcription termination site, TU - transcribed unit.

showed that the sets of modifications identified as important for all promoters, HCPs and LCPs are comparable between the RNA-Seq and microarray-derived expression values (Fig. 2.14). The only difference was that only H4K20me1, H3K27ac, and H2BK5ac, but not H3K79me1, are identified as being overrepresented in best scoring linear models for all promoters (p-values of the hypergeometric test $4.07 \cdot 10^{-54}$, $7.91 \cdot 10^{-13}$, $4.09 \cdot 10^{-24}$ and $8.35 \cdot 10^{-01}$, respectively). However, when analyzing best scoring models for LCPs, H3K79me1 clearly comes up as overrepresented (p-value of the hypergeometric test $2.17 \cdot 10^{-29}$). We conclude that a possible measurement bias due to microarray technology is not a major factor in identifying modifications important for prediction of expression of different groups of genes.

2.4 Histone modifications are predictive of gene expression across different cell types

Results presented in Sections 2.1 and 2.2 show that histone modifications are predictive of gene expression in CD4+ cells and furthermore that models incorporating only the information of four histone modifications can accurately predict gene expression levels. Next, we wanted to check whether models trained on the data of one cell type can be used to predict gene expression in another cell type, ie. whether

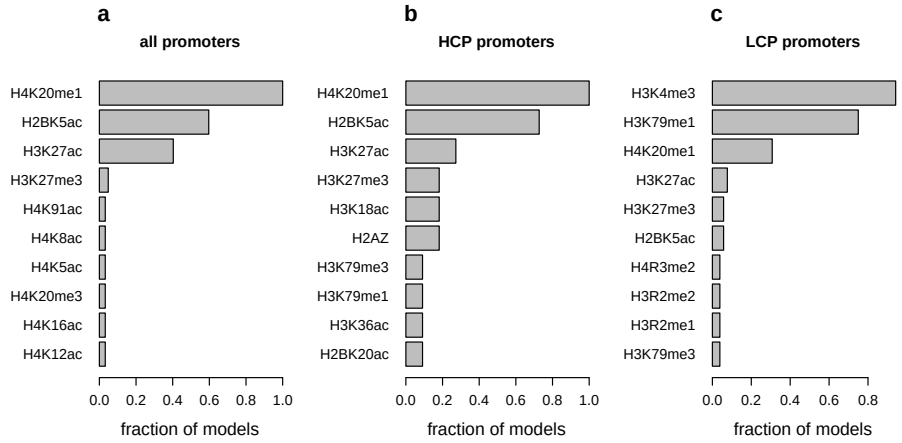

Figure 2.14: Overrepresentation analysis for linear models using RNA-Seq expression data. Overrepresentation of histone modifications in best scoring three-modifications linear models. Best scoring linear models achieve at least 95% of the prediction accuracy (measured as the Pearson correlation coefficient r between predicted and measured expression values) of the full linear model. (a) H4K20me1, H2BK5ac, and H3K27ac are overrepresented in best scoring models (62 models) for all promoters. (b) H4K20me1, H2BK5ac, and H3K27ac are overrepresented in best scoring models (11 models) for HCP promoters. (c) H3K4me3, H3K79me1, and H4K20me1 are overrepresented in best scoring models (52 models) for LCP promoters.

the relationship between histone modifications and gene expression holds universally across different cell types. We obtained genome-wide ChIP-Seq localization data for histone modifications in three different cell types: CD36+ and CD133+ cells [54] and IMR90 cells [2]. We also obtained microarray gene expression measurements for CD36+ and CD133+ cells [54] and RNA-Seq gene expression measurements for IMR90 cells [135]. To allow comparison of expression levels of genes in different cells types, expression values from all four cell types were normalized, in order to remove any effects which arise from differences in laboratory procedures. Expression values measured by microarrays for 14,802 RefSeq genes in CD4+, CD36+ and CD133+ cells were normalized using quantile normalization. The same approach was also used to normalize RNA-Seq expression values for CD4+ and IMR90 cells. Genes where the average number of RNA-Seq tags per base pair was equal to zero in either CD4+ or IMR90 cells were omitted, leaving 11,941 RefSeq genes for further analysis. Expression of genes in CD4+, CD36+ and CD133+ cells is correlated, which is expected since all of these cell types belong to the hematopoetic cell lineage. Measurements from IMR90 cells, a more distant fetal fibroblast cell line, show more differentially expressed genes when compared to the CD4+ cells (Fig. 2.15).

In order to determine if the relationships between histone modifications and gene expression are conserved in different cell types we used parameters of a linear model trained on histone modification data measured in CD4+ cells to predict gene expression in CD36+, CD133+ and IMR90 cells. Since gene expression is measured by two different experimental techniques (microarrays for CD36+ and CD133+ cells

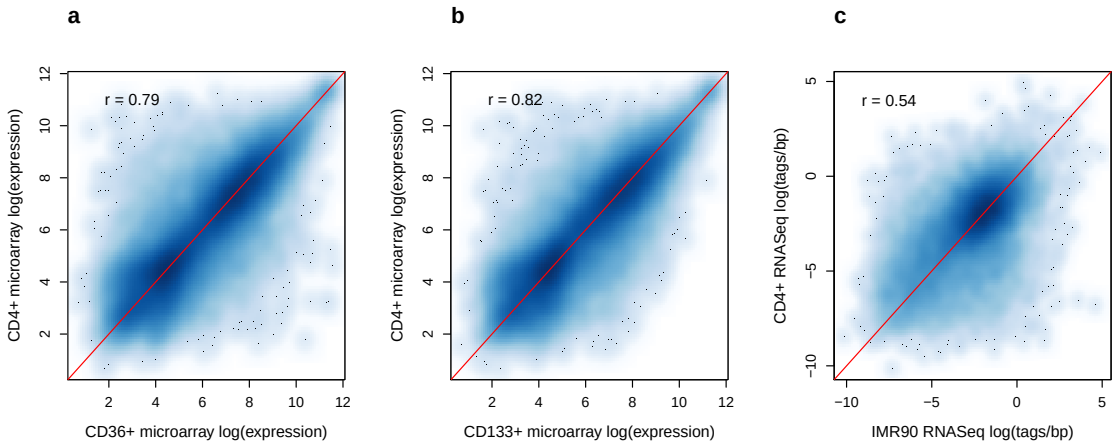

Figure 2.15: Comparison of gene expression levels in different cell types. Correlation of gene expression values measured in CD36+ (a), CD133+ (b) and IMR90 (c) cells with gene expression values measured in CD4+ cells. Expression levels in different cell types were normalized using quantile normalization.

and RNA-Seq for IMR90 cells), we trained two linear models on histone modification data from CD4+ cells, using expression measured by microarrays and RNA-Seq as response variables. For all four cell types, we determined the levels of histone modifications in the regions surrounding the TSS of RefSeq genes, as described in Section 2.1. Each of these values was transformed to a logarithmic scale, after adding a pseudocount of 1, since the optimized pseudocounts are specific for each cell type. The levels of histone modifications in the promoter regions were then used as predictor variables in linear models used to predict microarray and RNA-Seq expression values. Both linear models were trained only on histone modifications common to all four datasets (H3K4me1/3, H3K27me3, H4K20me1, H3K9me3 and H3K36me3).

Since the gene expression profiles of CD36+ and CD133+ cells are highly correlated to CD4+ T-cells ($r = 0.79$ and $r = 0.82$, respectively; Fig. 2.15 (a) and (b)), we restricted the prediction to genes with a fold change higher than five (2,622 genes for CD36+ cells and 2,304 genes for CD133+ cells). The correlation of predicted and measured expression values is high for both CD36+ and CD133+ cells ($r = 0.7$ and $r = 0.69$, respectively; Fig. 2.16 (a) and (b)). Since the gene expression levels are much less correlated between CD4+ and IMR90 cells ($r = 0.54$; Fig. 2.15 (c)), the analysis in IMR90 cells was not restricted only to highly differentially expressed genes. The linear model trained in CD4+ cells again achieves a high prediction accuracy ($r = 0.71$; Fig. 2.16 (c)), despite very different expression profiles of genes in these two cell types.

High prediction accuracy of both models on independent test data from different cell types strongly suggests that the relationship between histone modifications and gene expression is general and does not depend on the cellular context. Histone modifications are equally successful in predicting differentially expressed genes in closely related cell types where the expression of most genes is highly correlated

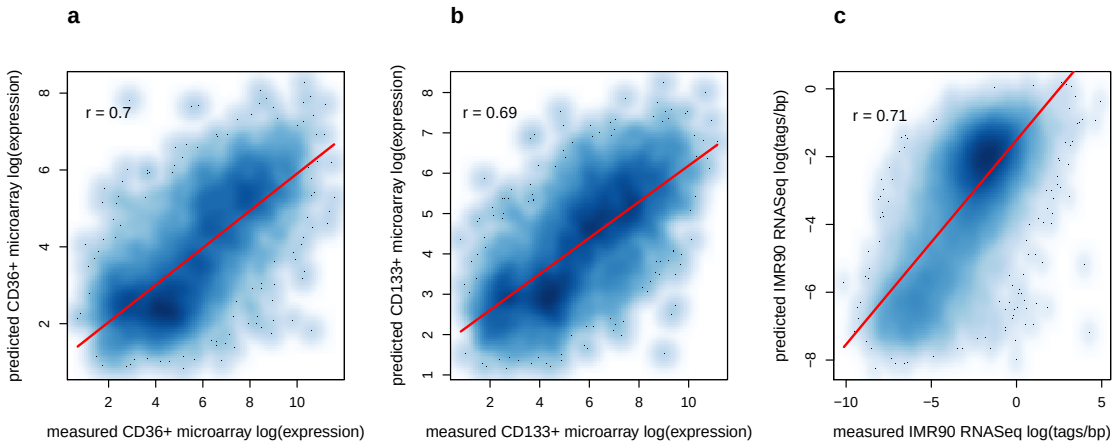

Figure 2.16: Prediction of gene expression levels in different cell types. Expression values of genes in CD36+ (a), CD133+ (b) and IMR90 (c) cells predicted using model parameters trained on data from CD4+ T-cells. The equations of the regression line for CD36+, CD133+ and IMR90 cells (y = 0.48x + 1.07, y = 0.45x + 1.71 and y = 0.60x - 1.51, respectively) show a high value of the intercept and a slope different from one due to the fact that the levels of the histone modifications were not normalized across cell types.

(CD4+, CD36+ and CD133+ cells), as well as changes in gene expression between highly divergent cell types (CD4+ and IMR90 cells).

2.5 Unexplained variance could be a result of mRNA degradation

In Section 2.1 we showed that there is a quantitative relationship between levels of histone modifications in promoter regions of human genes and the expression level of these genes. This relationship can be modeled quite accurately using a simple linear model, with a Pearson correlation coefficient r between measured and predicted expression levels of transcripts in CD4+ cells of 0.77. This result shows that the linear model performs well, explaining 59% of variance of the measured expression values. However, there is still a large amount of variance which can not be accounted for by this simple model, some of which is a consequence of noise in experimental measurements. In addition, at least some fraction of the unexplained variance could be attributed to the fact that expression values measured either by microarrays or by RNA-Seq represent steady-state levels of mRNA molecules in the cell. These steady-state levels are determined not only by the production of mRNAs but also by their degradation rate. Since histone modifications affect the chromatin structure of DNA we assumed that they influence mostly the production level of mRNA, and cannot be used to model the different degradation rates of transcripts.

If the unexplained variance of the linear model is indeed a consequence of the degradation rate of mRNA molecules, then the residuals of this model (residual = measured

expression value − fitted expression value) should be correlated with a measure of the degradation rate of mRNAs. In order to determine if this is really the case we downloaded measurements of levels of nascent transcripts in IMR90 cells measured by a global run-on-sequencing (GRO-Seq) assay [51], a method based on nuclear run-on (NRO) assays. In NRO assays nuclei are isolated, endogenous nucleotides washed away and inhibitors of transcription initiation are often added to block new initiation events and ensure that elongation can proceed unhindered. Transcripts which are associated with transcriptionally engaged RNA polymerase II (PolII) are then extended using labeled nucleotides [95]. Core *et al.* [51] used 5-bromouridine 5'-triphosphate (BrUTP) to label nascent mRNAs, which were subsequently isolated and sequenced using next-generation sequencing methods. The transcript pool isolated from the nuclei consists almost exclusively of nascent transcripts, while previously accumulated mRNAs are excluded, and GRO-Seq therefore gives a measure of the production level of mRNAs in the cell.

We hypothesized that histone modifications, given that they influence chromatin structure of the DNA, are more predictive of the production level than the overall expression level of mRNA. We therefore trained two linear models, where the response variables were mRNA production levels measured by GRO-Seq [51] and mRNA expression levels measured by RNA-Seq [135] in IMR90 cells. For both methods we mapped the sequenced tags to the 14,802 transcripts used in previous analyses and calculated the average number of GRO-Seq or RNA-Seq tags per base pair (tags/bp), as described in Section 2.1. We used the logarithm of the average number of GRO-Seq and RNA-Seq tags per base pair as a measure of production rate and steady-state levels of transcripts, respectively. In order to avoid using a pseudocount, we removed all genes the average number of either GRO-Seq or RNA-Seq tags in the transcript was equal to zero, leaving 11,869 transcripts for further analysis. In both models we used the levels of 24 histone modifications in promoter regions of RefSeq genes in IMR90 cells [2] as predictor variables. Histone modification data was processed as described in Section 2.1. Data for GRO-Seq and RNA-Seq measurements was quantile normalized to ensure that prediction accuracies for both models are comparable. The model used to predict mRNA production levels measured by GRO-Seq achieves a Pearson correlation coefficient of $r = 0.84$ (Fig. 2.17 (a)) between measured and fitted values, compared to $r = 0.75$ for the model where the response variable was mRNA expression level measured by RNA-Seq (Fig. 2.17 (b)). This result shows that histone modifications are indeed more informative of the production levels of mRNA than the overall steady-state expression levels of mRNAs in the cell.

We determined that there is a strong correlation ($r = 0.64$; p-value$< 2.2 \cdot 10^{-16}$) between the residuals of the RNA-Seq model (henceforth "RNA-Seq model residuals") and the difference between RNA-Seq and GRO-Seq measurements (henceforth "measurement difference"). However, both of these quantities are dependent on the actual RNA-Seq measurements, which could unjustly inflate the correlation coefficient. We therefore used partial correlation to control for the influence of RNA-Seq measurements. Using the average numbers of RNA-Seq tags/bp in analyzed transcripts, we trained two linear models to predict either the RNA-Seq model residuals

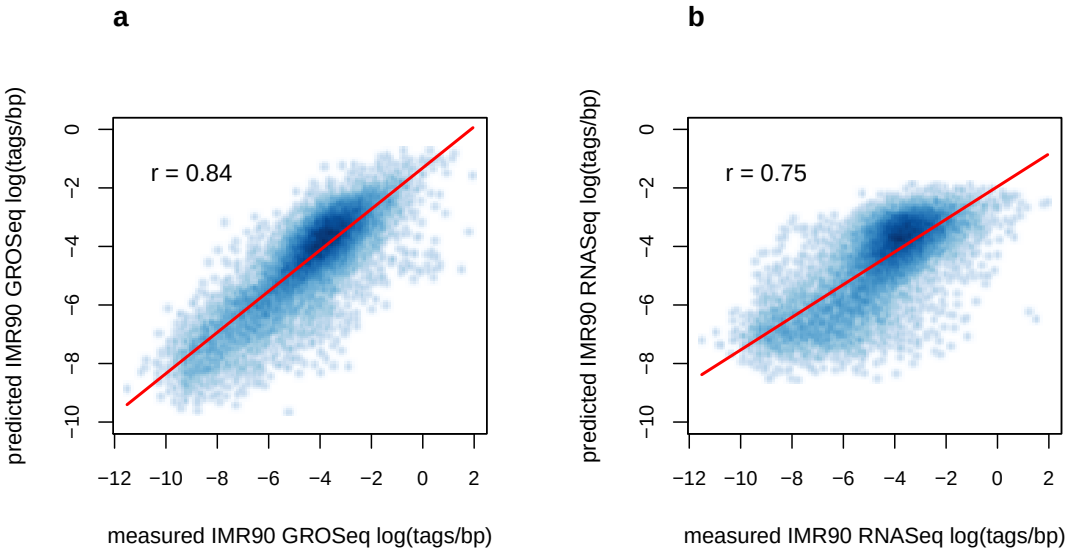

Figure 2.17: Prediction of expression values measured using GRO-Seq and RNA-Seq. Smoothed color density representation of the scatterplot of measured and predicted expression values for models in which levels of mRNA measured by GRO-Seq (a) and RNA-Seq (b) in IMR90 cells were used as response variables. The equations of the regression line for GRO-Seq (y = 0.7x - 1.32) and RNA-Seq models (y = 0.56x - 1.95) show a high value of the intercept and a slope different from one due to the fact that pseudocounts added to histone modification levels were not optimized for individual expression measurements.

or the measurement difference. We then calculated the Pearson correlation coefficient r between the residuals of these two linear models. Since the residuals of the two linear models are no longer correlated to the average RNA-Seq tag counts, the effect of these measurements is removed from the analysis.

However, even after removing this effect the RNA-Seq model residuals and the measurement difference are still highly correlated ($r = 0.56$; Fig. 2.18). This finding implies that the model used to predict transcript expression levels could be improved if we succeeded in modeling the difference between production and steady-state levels of mRNA, ie. mRNA degradation.

2.6 Modeling mRNA degradation rate

We showed that linear models trained on levels on histone modifications at the promoter achieve better accuracy in predicting expression levels of genes measured by GRO-Seq compared to RNA-Seq technology (Section 2.5). A possible explanation of this finding is that levels of histone modifications at the promoter are more predictive of the production rate of mRNA than its steady-state level, under the assumption that the difference in GRO-Seq and RNA-Seq measurements is caused by different degradation rates of individual transcripts. If this assumption is true, the inclusion of

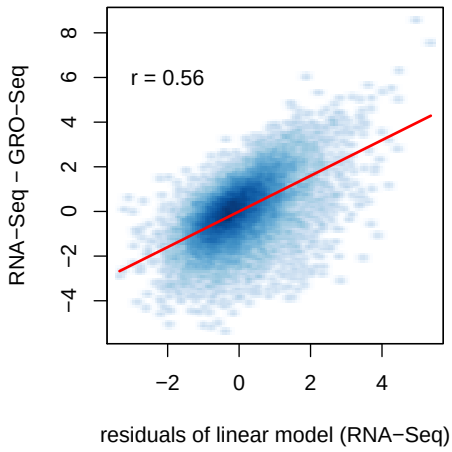

Figure 2.18: Correlation of the RNA-Seq model residuals and the difference between RNA-Seq and GRO-Seq measurements. The effect of RNA-Seq has been removed using partial correlation. The equation of the regression line is y = 0.8x. The slope of the regression line is influenced by outliers.

additional features which provide information on the degradation rates of transcripts should be beneficial for the prediction of RNA-Seq measured gene expression.

We decided to expand the set of predictor variables by adding features corresponding to two different types of sequence elements known to influence mRNA degradation rates, namely AU-rich elements (AREs) and miRNA target sites, both of which are usually enriched in 3' untranslated regions (3' UTRs) of transcripts [20, 88]. In order to determine the number of occurrences of these sequence features in the analyzed transcripts we downloaded the sequences of 3' UTRs of human transcripts from the UCSC Genome Browser (hg18, March 2006) [1]. Out of the 14,802 Refseq transcripts used to train and test linear models for prediction of gene expression in CD4+ cells, 13,513 had annotated 3' UTRs and were used in further analysis.

We first attempted to model the influence of AREs on the steady state level of mRNA in the cell. The appearance of AREs in the 3' UTR is often connected to a reduction in mRNA stability [17]. AREs are commonly enriched in uridine nucleotides, and are furthermore characterized by the presence of pentamer (AUUUA) or nonamer (UUAUUUAWW; W = A or U) motifs [43]. Pentamer and nonamer motifs can occur together and can sometimes also overlap in a single AU-rich element. We counted the number of occurences of the pentamer and the nonamer motif in the 3' UTRs of the 13,513 transcripts. We then used the number of occurrences of these motifs as two additional predictor variables, henceforth called "ARE features".

We used ARE features, in addition to levels of histone modifications at the promoter, to train linear models and predict gene expression measured by RNA-Seq in both IMR90 and CD4+ cells. To test whether inclusion of ARE features has an influence on the performance of linear regression we compared the prediction accuracy of these models to the prediction accuracy of the corresponding models trained on histone modification levels only.

For 13,513 transcripts with annotated 3' UTRs we calculated the number of RNA-Seq tags mapped to the transcript in either IMR90 or CD4+ cells [45, 135]. The number

of RNA-Seq tags was then normalized for exon length and used as a measure of the steady-state expression levels of the transcripts in the respective cell type. We also mapped ChIP-Seq tags for 24 histone modifications (measured in IMR90 cells; [2]) and 39 histone modifications (measured in CD4+ cells; [18, 228]) to the regions of 4,001 base pairs surrounding the transcription start sites (TSS) of the transcripts. Both histone modifications and expression levels were transformed to a logarithmic scale, after removing transcripts where any of these values were equal to zero. We therefore avoided the need to define cell-type-specific pseudocounts, but the number of transcripts was reduced to 11,029 (IMR90 cells) and 7,278 (CD4+ cells). For both cell types, we trained linear models to predict steady-state expression levels, using histone modification levels and ARE features as predictor variables. The models were trained in a 10-fold cross-validation setting and the predicted expression level was taken to be the value predicted when the transcript was part of the test set, as described in Section 2.1. The prediction accuracy of the model was calculated as the Pearson correlation coefficient r between measured and predicted expression levels.

We then compared the models which incorporated the ARE features to the corresponding models using only histone modification levels. For both IMR90 and CD4+ cells, the models which included ARE features showed a slight increase in prediction accuracy ($r_{HM+ARE} = 0.748$ compared to $r_{HM} = 0.746$ for IMR90 cells; $r_{HM+ARE} = 0.783$ compared to $r_{HM} = 0.774$ for CD4+ cells; HM - histone modification features, ARE - ARE features). To determine if the difference in prediction accuracy is significant we calculated the values of the Bayesian information criterion (BIC) on training data in each round of the 10-fold cross-validation. We then calculated the average BIC values for linear models trained in IMR90 and CD4+ cells. For both cell types the average BIC values are lower for models which include ARE features than models trained only on histone modifications ($\mathrm{BIC}_{HM+ARE} = 31589.50$ compared to $\mathrm{BIC}_{HM} = 31650.70$ for IMR90 cells; $\mathrm{BIC}_{HM+ARE} = 22000.37$ compared to $\mathrm{BIC}_{HM} = 22222.37$ for CD4+ cells). This shows that although the increase in prediction accuracy is modest, it is not merely a consequence of increased model complexity.

We further expanded the linear models by incorporating features related to the number of target sites for different miRNAs in the 3' UTR regions of analyzed transcripts. We downloaded information on miRNA families from TargetScan Human Release 5.1 [130] and extracted sequences matching positions 2-8 of the mature sequence for 545 miRNA families which contained previously annotated human miRNAs (comprising in total 677 miRNAs). We proceeded to predict target sites for the miRNAs belonging to the 545 different miRNA families, using a adapted version of the TargetScan algorithm which did not take into account target site conservation. Most miRNA target site prediction programs require perfect complementarity between the 3' UTR and the 5' region of miRNA corresponding to nucleotides 2-7, called the miRNA "seed" [21]. However, complementarity can also be extended to adjacent nucleotides, giving rise to several different types of matches between the 3' UTR and the miRNA sequence. Most commonly occurring types of matches are 6-mer (seed match), 7mer-A1 (seed match + A at position 1), 7mer-m8 (seed match + match at position 8) and 8mer (seed match + match at position 8 + A at position 1) [21]. For all analyzed tran-

scripts and all 545 miRNA families, we extracted predicted 7mer-A1, 7mer-m8 and 8mer target sites, since these type of matches were shown to have a larger influence on expression levels of transcripts [83].

Recently, Grimson *et al.* [83] developed a linear model to predict the effect of occurrence of single miRNA target sites on transcript expression levels. In independent experiments, the authors transfected HeLa cells with 11 different miRNA, and used microarrays to measure expression levels of transcripts which contain predicted single target sites for these miRNAs in both transfected and not transfected cells. The authors determined that the extent of downregulation of the transcript depends not only on the type of mRNA-miRNA match, but also on the context in which the target site appears in the 3' UTR. They identified several context features with an influence on miRNA target site efficiency, such as the local AU content, pairing of the 3' UTR to the 3' region of the miRNA and distance of the target site to the end of the 3' UTR, and developed a method of scoring these features. For each type of mRNA-miRNA match, they used linear regression to relate the scores of context features to the measured amount of downregulation of specific transcripts. The final prediction of the appropriate linear model, called the context score, should then correspond to the efficiency of the predicted miRNA target site.

For each of the 13,513 analyzed transcripts and all predicted 7mer-A1, 7mer-m8 and 8mer target sites for 545 miRNA families we calculated the scores for AU contribution, 3' pairing and distance from the end of the 3' UTR as described by Grimson *et al.* [83]. We then used the coefficients of the linear models trained in their study to calculate the context score for each predicted miRNA target site. In the case where multiple target sites for a particular miRNA family were predicted in one transcript, we calculated the final context score as the sum of all single context scores, since we assumed that multiple target sites would contribute to mRNA degradation in an additive manner. Each transcript was therefore described by 545 additional features, corresponding to context scores of predicted target sites for different miRNA families, henceforth called "miRNA features".

miRNAs are believed to regulate gene expression in a tissue-specific manner [124]. We therefore restricted the miRNA features to the miRNAs which are highly expressed in CD4+ and IMR90 cells, respectively. For CD4+ cells, we obtained data produced by deep sequencing of stem-loop sequences of human miRNAs [19]. The expression level of the corresponding mature miRNA in CD4+ cells was calculated as the total number of tags mapped to all associated stem-loop sequences. We also obtained microarray measurements of miRNA expression in IMR90 cells [234], and calculated the expression levels for individual miRNAs as the average signal intensity over all replicates. In total, 543 and 535 out of 677 miRNAs had measured expression levels in CD4+ and IMR90 cells, respectively. For each cell type we then ordered all the miRNAs according to their expression level and reduced the set of miRNA features to the miRNAs whose expression was in the top 10%. In this way the set of miRNA features was reduced to features corresponding to 38 miRNA families for CD4+ cells

and 34 miRNA families for IMR90 cells. These miRNA families comprise of 54 and 53 mature miRNAs, respectively.

For each cell type, we predicted expression levels of analyzed transcripts using linear models trained on a set of predictor variables which included histone modification features, ARE features and miRNA features (for highly expressed miRNAs only). In order to test whether miRNA target site context scores provide additional information on expression levels of transcripts we compared the prediction accuracy of these models with the accuracy of models trained only on histone modification and ARE features. All the models were trained in a 10-fold cross-validation setting and the prediction accuracy of the model determined as described above.

For both cell types, models which included information on miRNA target sites performed slightly worse than models trained only on histone modification and ARE features ($r_{HM+ARE+miRNA} = 0.747$ compared to $r_{HM+ARE} = 0.748$ for IMR90 cells; $r_{HM+ARE+miRNA} = 0.782$ compared to $r_{HM+ARE} = 0.783$ for CD4+ cells; HM - histone modification features, ARE - ARE features, miRNA - miRNA features). Comparison of average BIC values from 10-fold cross-validation further confirms that inclusion of additional miRNA features only leads to an increase in model complexity without providing an improvement in prediction accuracy $\mathrm{BIC}_{HM+ARE+miRNA} = 31854.68$ compared to $\mathrm{BIC}_{HM+ARE} = 31589.01$ for IMR90 cells; $\mathrm{BIC}_{HM+ARE+miRNA} = 22302.04$ compared to $\mathrm{BIC}_{HM+ARE} = 21999.86$ for CD4+ cells).

We wanted to examine the contribution of different features to the prediction accuracy of the linear models trained on different sets of predictor variables both for CD4+ and IMR90 cells. We focused on models which incorporated, in addition to histone modification features, either ARE and miRNA features or only ARE features. We therefore calculated the average values of the regression coefficients for all predictor variables in 10-fold cross-validation. ARE and miRNA features were scaled by subtracting the mean and dividing by the standard deviation of the variable. Scaling of these variables before training of the model ensures that their regression coefficients are not unjustly influenced by the different magnitudes of the features. For each feature we also calculated the median p-value of the coefficient in 10-fold cross-validation. To calculate the p-value, one must first determine the t-statistic of the estimated coefficient, by dividing the coefficient with its standard error. The p-value of the coefficient gives the probability of observing a t-statistic greater or equal in magnitude under the null hypothesis that the true coefficient value is zero, ie. it has no influence on the response variable [61]. We considered regression coefficients with a median p-value lower than 0.05 to have a statistically significant contribution to model accuracy.

In both cell types, several histone modifications contribute significantly to prediction accuracy of the linear models, and their influence on expression levels (determined by the average regression coefficient) is in most cases similar in CD4+ and IMR90 cells. However, the levels of several histone modifications in promoter regions are highly correlated, in some cases reaching almost perfect correlation [228]. Since the appearance of correlated variables can influence the sign and magnitude of their

regression coefficients and their presumed contribution to the accuracy of the model [61], we did not study their regression coefficients in more detail.

On the other hand, ARE features and miRNA features do not display this high degree of correlation, either within themselves, or with histone modification features (see Appendix, Fig. A.1 and Fig. A.2), and we therefore placed more emphasis on the analysis of their contribution to the model. We found that regression coefficients for the feature corresponding to the number of pentamer AU-rich motifs are statistically significant in all four linear models studied (Fig. 2.19 and Fig. 2.20). Furthermore, the regression coefficients were always negative, in agreement with previous studies which showed that the presence of AU-rich elements has a negative influence on mRNA stability [17]. In contrast, the feature corresponding to nonamer AU-rich motifs was not significant in any of the studied models. A possible explanation for this finding is that nonamer AU-rich motifs appear in the 3' UTR regions of analyzed transcripts much less frequently than pentamer AU-rich motifs. We also examined the coefficients of linear models which included information on miRNA target site context scores. Only two miRNAs, hsa-miR-125b and hsa-miR-145, had significant regression coefficients for IMR90 cells, while in the model trained on data from CD4+ cells none of the miRNA features contributed significantly to prediction accuracy (Fig. 2.20). Furthermore, the significant regression coefficients of the two miRNA features in the model for IMR90 cells were much smaller than the coefficients for either histone modification or ARE features. In addition, the regression coefficient for hsa-miR-125b was positive, which is opposite to the commonly accepted role of miRNAs in reducing transcript stability [21]. While it is possible that this miRNA exerts its influence by inhibiting positive regulators of transcription, we believe that a more likely explanation is that context scores are not able to successfully model miRNA-mediated degradation in our modeling framework. This finding, along with the fact that the models do not perform better than the ones which incorporate only ARE features, shows that the context scores of miRNAs do not contribute significantly to the accuracy of the prediction of transcript expression levels.

2.7 Conclusions

In summary, we found that the levels of histone modifications are well correlated to gene expression and that this relationship can be generalized across different cell types. Moreover, our analysis revealed that the number of important modifications can be reduced from 39 to four, indicating that these four modifications may play a crucial role in the transcriptional process, either by reinforcing each other or in a combinatorial manner. Upon separating promoters into LCPs and HCPs, different sets of modifications were found to be important for the prediction of expression levels, which indicates that these promoters are regulated differently. We furthermore showed that models trained on levels of histone modifications at the promoter are more successful in predicting the production rate of mRNA than its steady-state level.

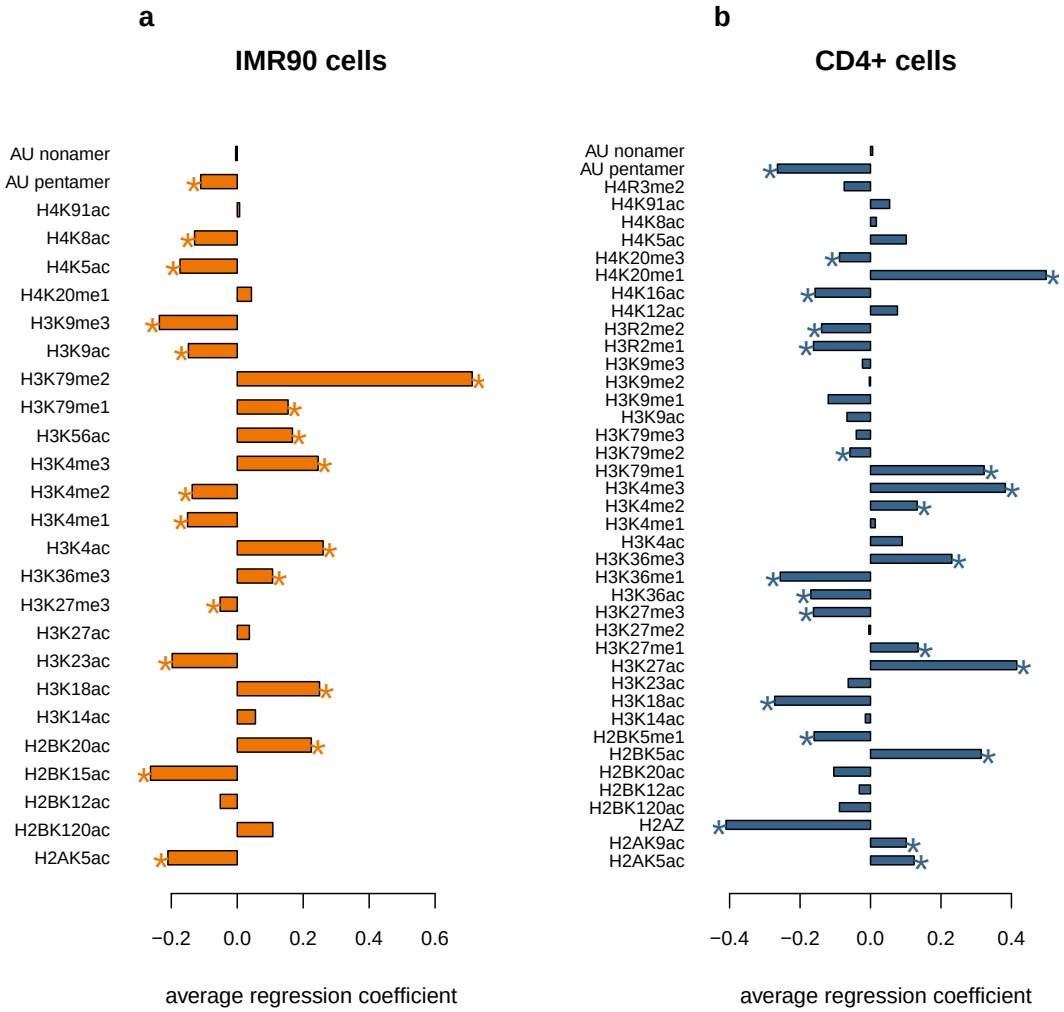

Figure 2.19: Regression coefficients of models trained on histone modifications and AU-rich elements. Bar plots showing average regression coefficients of linear models incorporating histone modification features and ARE features, used to predict gene expression levels in IMR90 (a) and CD4+ (b) cells. Stars mark the regression coefficients whose median p-value is lower than 0.05 in 10-fold cross-validation.

Since histone modifications do not provide information on the degradation rate of the transcript, we extended the model to include features associated with regulation of mRNA degradation. Although the prediction accuracy did not drastically improve, some features contributed significantly to model performance and their contribution corresponded to previous knowledge of their influence on mRNA stability. We conclude that a future improvement in modeling the degradation rate of transcripts could lead to even more accurate models of gene expression.

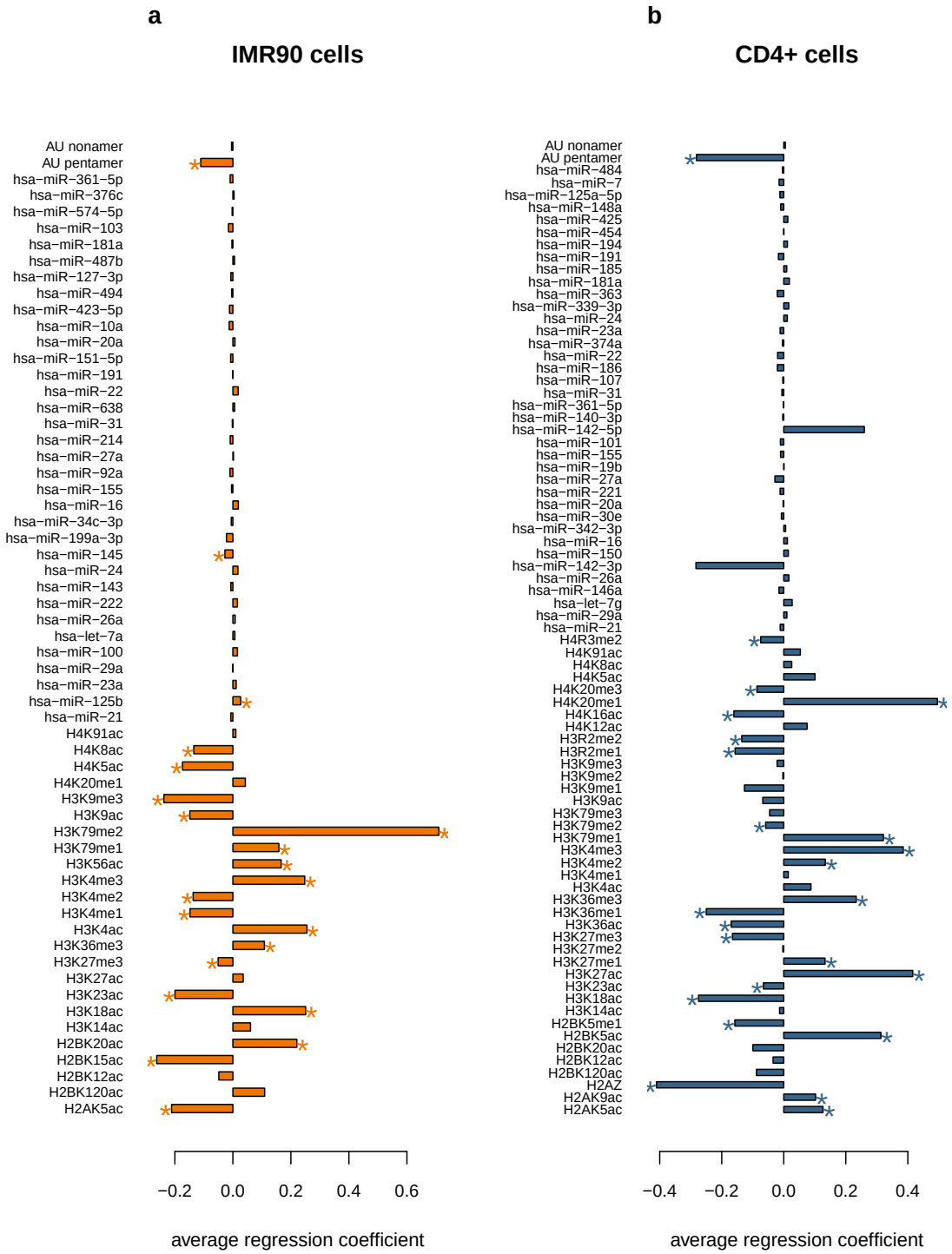

Figure 2.20: Regression coefficients of models trained on histone modifications, AU-rich elements and miRNA context scores. Bar plots showing average regression coefficients of linear models incorporating histone modification features, ARE features and miRNA features, used to predict gene expression levels in IMR90 (a) and CD4+ (b) cells. Stars mark the regression coefficients whose median p-value is lower than 0.05 in 10-fold cross-validation.

Chapter 3

Influence of histone modifications on regulation of alternative splicing

Results of several recent studies suggested the existence of a link between histone modifications and gene structure, ie. that some histone modifications are enriched in exonic vs. intronic regions of genes. It has been furthermore shown that the levels of several histone modifications in exons which undergo alternative splicing events differ depending on whether the exon is included or included in the transcript [137]. In this chapter we present the results of an analysis in which we used logistic regression models to investigate the nature of the relationship between histone modifications and alternative splicing.

3.1 Compiling a dataset of exon skipping events

If histone modifications indeed have a significant influence on the outcome of alternative splicing events it should be possible to infer this outcome using the levels of histone modifications in the alternative exon as predictors. To test this hypothesis we focused on exon skipping, the predominant type of alternative splicing event in human cells [152, 178]. We downloaded annotations for all transcripts in the human genome from Ensembl version 50 (hg18) [3] and used the AStalavista web server [74] to extract all annotated alternative splicing events between all pairs of overlapping protein-coding transcripts (except for the transcripts mapping to haplotype chromosomes c6_COX and c6_QBL).

We obtained publicly available genome-wide RNA-Seq expression data [45] for CD4+ T-cells. We used ELAND software (Illumina) to map RNA-Seq tags to all exons of transcript pairs where an exon skipping event has been reported by AStalavista, after removing all first and last exons and exons shorter than 50 base pairs (bp). To ensure that we only include unique splicing events in our analysis, we mapped all Ensembl transcripts to their respective genes and used only one alternative splicing event per gene for further analysis. In the case when two or more transcript pairs with a reported exon skipping event belonged to the same gene, we chose the event associated with the transcript pair which had the highest average number of RNA-Seq tags per exon in the longer transcript. In order to restrict the analysis to transcripts

which are expressed in CD4+ T-cells, we removed all the transcripts for which the median number of tags mapping to constitutive exons was equal to zero. We further pruned the set of transcripts by removing all transcripts containing less than three exons, leading to a final set of 2130 transcripts where an exon skipping event has been previously annotated.

In order to determine whether an alternative exon is included in the transcript or not we defined a measure called the "inclusion ratio". The inclusion ratio reflects the frequency of inclusion of the alternative exon in the transcript and is calculated as

$$\text{inclusion ratio} = \frac{\frac{tags_E}{L_E} - \frac{tags_T}{L_T}}{\frac{tags_E}{L_E} + \frac{tags_T}{L_T}} \tag{3.1}$$

where $tags_E$ and $tags_T$ denote the number of RNA-Seq tags mapped to the alternative exon or all constitutive exons in the transcript, respectively. L_E and L_T denote the lengths of the alternative exon and all constitutive exons in the transcript, respectively. For each of the 2130 alternative exons in our dataset we calculated the inclusion ratio in CD4+ T-cells. The inclusion ratio shows a bimodal distribution, as seen in Fig. 3.1. Value of an inclusion ratio of -0.9 was chosen to distinguish alternative exons that are skipped (inclusion ratio < -0.9) from those that are included (inclusion ratio ≥ -0.9) in the transcript. This led to 424 alternative exons being annotated as skipped and 1706 as included in our transcripts.

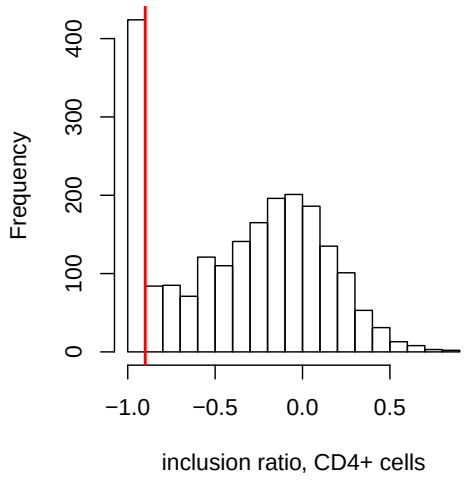

Figure 3.1: Distribution of inclusion ratio for alternative exons chosen for the analysis in CD4+ T-cells. The red line marks the boundary between skipped (inclusion ratio < -0.9) and included (inclusion ratio ≥ -0.9) alternative exons.

We wanted to ensure that the inclusion ratio really reflects the inclusion/exclusion of an alternative exon from the transcript. We therefore checked the number of mapped RNA-Seq tags per base pair of the alternative exon in groups of exons defined to be skipped or included in the transcript according to the inclusion ratio. The number of RNA-Seq tags per base pair is significantly lower for alternative exons defined as "skipped" (Fig. 3.2 (a); p-value of the Wilcoxon rank sum test $= 7.11 \cdot 10^{-202}$), indicating that these exons are genuinely excluded from the transcript. As

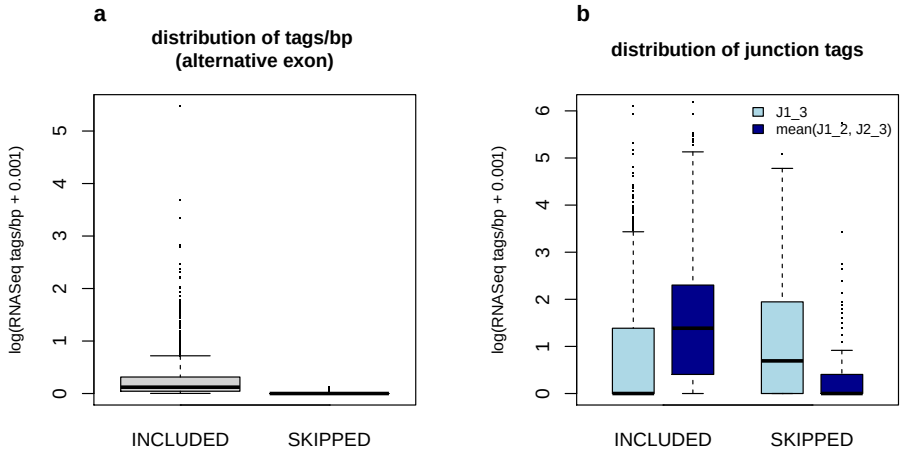

Figure 3.2: Distribution of RNA-Seq tags/bp and junction tags in included and skipped alternative exons. (a) Distribution of RNA-Seq tags per base pair in alternative exons defined to be skipped or included according to the inclusion ratio. (b) Distribution of junction tags mapped to junctions J1_3, J1_2 and J2_3 of exons defined to be skipped or included according to the inclusion ratio.

a further test we examined the distribution of junction tags in the groups of exons characterized as included or skipped according to the inclusion ratio. Junction tags mapped to all possible exon-exon junctions retrieved from Ensembl version 50 were taken from the study by Chepelev *et al.* [45]. For each of the 2130 alternative exons we mapped the junction tags to one of three possible exon-exon junctions: junction between the alternative exon and the upstream exon (J1_2), junction between the alternative exon and the downstream exon (J2_3) and junction between upstream and downstream exons (J1_3). Junction tags mapped to the J1_3 exon-exon junction are indicative of an exon skipping event, while tags mapped to J1_2 and J2_3 support the inclusion of the alternative exon in the transcript. The number of junction tags mapped to J1_3 is significantly higher in the group of alternative exons defined to be skipped by the inclusion ratio than in the group of exons defined to be included in the transcript (Fig. 3.2 (b); p-value of the Wilcoxon rank sum test $= 3.751 \cdot 10^{-8}$), while the mean number of junction tags mapped to either J1_2 or J2_3 is significantly lower for the skipped exons (Fig. 3.2 (b); p-value of the Wilcoxon rank sum test $= 6.404 \cdot 10^{-114}$). Taken together, these results confirm that the inclusion ratio is an appropriate measure of the outcome of alternative splicing events.

3.2 Logistic regression model for prediction of alternative splicing

Each alternative exon j was assigned a class label S (skipped exons, inclusion ratio < -0.9) or I (included exons, inclusion ratio ≥ -0.9). We downloaded publicly available genome-wide ChIP-Seq data for 38 histone modifications, 1 histone variant

(H2A.Z) and RNA polymerase II (PolII) in CD4+ T-cells [18, 183, 228]. PolII binding was measured by three separate experiments, using antibodies for total PolII, Ser-5 phosphorylated PolII (PolIIS5P) and unphosphorylated PolII (PolIIUP). We also obtained MNase-Seq data for nucleosome positioning in the same cell type [183]. We mapped the ChIP-Seq/MNase-Seq tags for the 43 variables to 2130 alternative exons chosen for the analysis. The number of tags N_{ij} for each variable i and exon j was transformed to the logarithmic scale after adding a pseudocount of 10^{-3} to ensure the logarithm is always defined ($N'_{ij} = log(N_{ij} + 10^{-3})$). We then used logistic regression to estimate the posterior probability p_j of exon j belonging to class I given the feature vector N'_{ij} (Eq. 3.2).

$$p_j(\text{class} = I \mid N'j) = \frac{1}{1 + e^{(\beta_0 + \sum_{i=1}^{43} \beta_i N'_{ij})}} \tag{3.2}$$

Since our dataset consists of two unbalanced classes, with around 80% of the exons belonging to class I, we optimized the cut-off probability for assigning an exon to class I in a nested 5-fold cross-validation setting. The prediction accuracy of the logistic regression model in each cross-validation loop was estimated using the g-mean measure proposed by Kubat *et al.* ([122]; Eq. 3.3), which is not unduly influenced by unequal class sizes. TPR and TNR denote the true positive rate and true negative rate, respectively. TP, FP, TN and FN, denote the number of true positives, false positives, true negatives and false negatives, respectively.

$$\begin{aligned} g\text{-}mean &= \sqrt{TPR \cdot TNR} \\ TPR &= \frac{TP}{TP + FN} \\ TNR &= \frac{TN}{TN + FP} \end{aligned} \tag{3.3}$$

The nested 5-fold cross-validation, where the outer 5-fold cross-validation was used to obtain an estimate of prediction accuracy, while the inner 5-fold cross-validation was used to optimize the cut-off probability, was implemented as follows. In the outer 5-fold cross-validation the entire dataset D of 2170 alternative exons was randomly divided into five disjoint sets D_1,..., D_5 of equal size. In each cross-validation loop one set D_i was used as a test set while the remaining four sets were joined to form a calibration set $C_i = D \backslash D_i$, which was then used in an inner 5-fold cross-validation. The calibration set C_i was randomly divided into five disjoint sets C_{i1},..., C_{i5} of roughly equal sizes. In each loop of the inner 5-fold cross-validation one of the sets C_{ij} was left out to serve as a validation set. The remaining four sets were joined to form a training set $T_{ij} = C_i \backslash C_{ij}$, which was then used to learn the unknown parameters β of the logistic regression model M_{ij}^{inner}. The trained model M_{ij}^{inner}

was then used to estimate the posterior probability of belonging to class I for exons belonging to the validation set C_{ij}. Each exon was assigned to the class I if the posterior probability of belonging to class I was greater than or equal to a threshold t_{ij}, and to class S otherwise. t_{ij} was varied from 0 to 1, in intervals of 0.01, and the g-mean was computed on C_{ij} for every t_{ij}. We determined which t_{ij} gave the highest average prediction accuracy across the inner 5-fold cross-validation, and used this value as the optimal threshold t_i^{opt} in the outer 5-fold cross-validation. A logistic regression model M_i^{outer} was then trained on the entire calibration set C_i and used to estimate the posterior probability p_k of exon k in the test set D_i. The exon k was assigned to class I if p_k was greater than or equal to the optimized threshold t_i^{opt}, and to class S otherwise. The prediction for each exon k in dataset D was the class label assigned to the exon when it was part of the test set D_i.

To ensure that our model is robust to random perturbations of the data used to train the model, we repeated the nested 5-fold cross-validation 100 times. The final prediction for each exon was the class label assigned by at least 50% of the 100 models trained using nested 5-fold cross-validation. The final accuracy was calculated as the proportion of correctly classified exons (Eq. 3.4).

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.4}$$

3.3 Predicting exon skipping using histone modifications

The results of prediction of the logistic regression model using 43 predictor variables are shown in Table 3.1. Using data for exon levels of histone modifications, nucleosomes and PolII as predictors, we are able to accurately predict inclusion/skipping for 64.04% of alternative exons. The distributions of g-mean, TPR and TNR obtained in each of the 100 models trained using 5-fold nested cross-validation shows that the predictions are not influenced by unequal class sizes (Fig. 3.3). Even though this result shows that chromatin structure does incorporate information on the outcome of alternative splicing events, the number of missclassified exons was surprising, given that many recent studies reported a significant difference in enrichment of different chromatin modifications between alternative exons skipped or included in the transcript [9, 59, 100, 117, 186].

One possible explanation is that potential complicated relationships between predictor variables cannot be captured using a simple logistic regression model. We therefore repeated the classification using random forests, an ensemble tree-based algorithm which can exploit non-linear relationships between variables [29]. The classification results of random forests depend on the distribution of class labels, with the prediction accuracy being biased toward the majority class in case of unbalanced

		Class labels S	Class labels I
Prediction	S	268	610
	I	156	1096

accuracy = 64.04%

Table 3.1: Prediction of exon skipping using a logistic regression model trained on levels of histone modifications, PolII and nucleosomes in alternative exons. The predicted class label for each exon is the final prediction obtained by 100 repeats of 5-fold nested cross-validation.

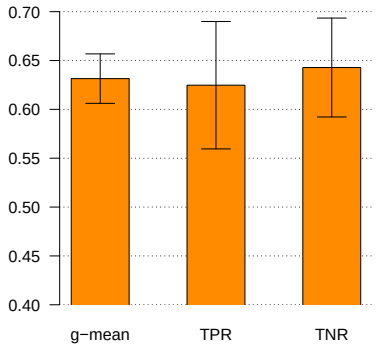

Figure 3.3: Distribution of the prediction accuracy for 100 logistic regression models trained using nested 5-fold cross-validation. The g-mean, TPR and TNR correspond to values obtained in each round of the outer 5-fold cross-validation.

class sizes. This obstacle can be overcome by under-sampling the majority class in order to produce more balanced prediction results [42]. In random forests, the test set error is estimated internally during the run of the algorithm, suppressing the need for cross-validation. For each tree, N samples are drawn with replacement from the dataset and represent the training set. The rest of the data ("out-of-bag" data) constitutes the test set for this tree. The prediction for each sample is made by taking the average of predictions over all trees for which the sample was part of the out-of-bag data. As a result, the prediction accuracy calculated on the final predictions gives an unbiased estimate of the out-of-bag error rate. When combining random forests with under-sampling, each tree is grown using a random sample from the dataset consisting of a fixed number of cases from both classes. In order to determine an optimal sample size, we fixed the number of cases from the minority class S to 100% and varied the number of cases drawn from the majority class I from 5% to 100% in intervals of 1%. For each combination of parameters we trained a random forest with 1000 trees on 80% of the data. We then determined the prediction accuracy of the random forest using the remaining 20% of the data as an independent set. We obtained the best prediction accuracy (accuracy = 64.14%) with the random forest where the majority sample size was 15%. The prediction accuracy was balanced in both classes (g-mean = 0.64, TPR = 0.65, TNR = 0.63). The prediction accuracy of logistic regression (accuracy = 64.04%) is comparable to the one obtained using random forest (accuracy = 64.14%). Furthermore, the predicted class labels overlapped in 83.7% of cases. This highly significant overlap (p-value of the chi-squared test = $9.113 \cdot 10^{-206}$) demonstrates that the results of the analysis are stable and do not depend on the classification method used.

Since the results of logistic regression showed that histone modifications are predictive

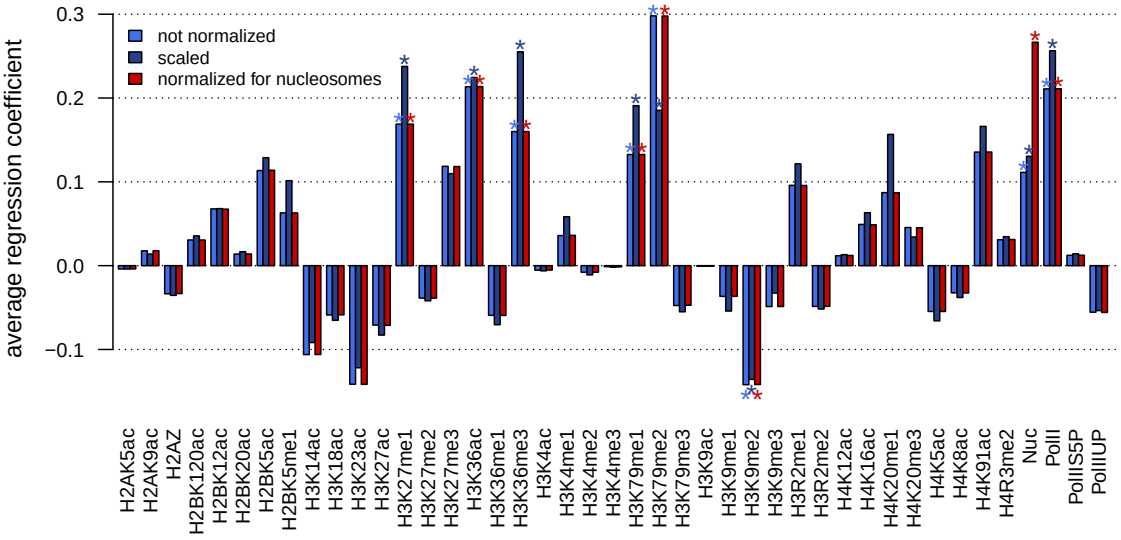

Figure 3.4: Regression coefficients of the logistic regression model. Bar plot showing the average regression coefficients of variables used in logistic regression to predict alternative splicing for models using not normalized variables, scaled variables and variables normalized for nucleosome occupancy. Stars mark the coefficients whose median p-values are less than 0.05 in 100 repeats of nested 5-fold cross-validation.

of alternative splicing outcomes, we were interested in discovering which variables contributed most to prediction accuracy. In order to identify the most informative features, we extracted the coefficients of logistic regression models used to predict the labels of the test set data it the outer loop of nested 5-fold cross-validation (100 repeats). We also recorded the p-value for each coefficient. The p-value of the coefficient β_i gives the probability of observing coefficient β_i under the null hypothesis that there is no relationship between predictor variable i and the response variable. For each variable i we calculated the median p-value of the coefficient β_i over the 100 repeats of nested 5-fold cross-validation, and considered the variables whose median p-value was lower than 0.05 to contribute significantly to prediction accuracy. Although the levels of different histone modifications in the promoter regions of genes are highly correlated [228], they show a much lesser degree of correlation in the exonic regions (see Appendix, Fig. A.3). For this reason, we presume that the analysis of regression coefficients of variables used to train the logistic regression model provides a reliable estimate of their contribution to prediction accuracy. According to this criterion the most informative features are the levels of nucleosomes (Nuc), PolII, H3K27me1, H3K36ac, H3K36me3, H3K79me1, H3K79me2 and H3K9me2 (Fig. 3.4). This result is in agreement with previous studies, since most of these variables, except H3K36ac, were shown to be differentially distributed between alternatively and constitutively spliced exons [9, 59, 100, 117].

The magnitude of regression coefficients depends on the scale of predictor variables. Since the overall number of mapped ChIP-Seq reads differs between predictor variables, possibly due to different antibody specificity or experimental procedures, we

decided to scale the predictor variables in order to ensure that the results we observe are biologically meaningful. We scaled each predictor variable i by subtracting the mean and dividing by the standard deviation, and then repeated the prediction using logistic regression, as described in Section 3.2. We observed that the prediction accuracy of the model (accuracy = 63.8%; Fig. 3.5) is similar to the one achieved by the model using not scaled variables (accuracy = 64.04%). We then reexamined the median p-values of the regression coefficients. We determined that the coefficients of all previously mentioned variables were still statistically significant (median p-value < 0.05), implying that the fact that they were identified as the most informative features is not a consequence of the different scales of the variables (Fig. 3.4). For details of the performance of the model using scaled predictor variables see Appendix, Table A.1.

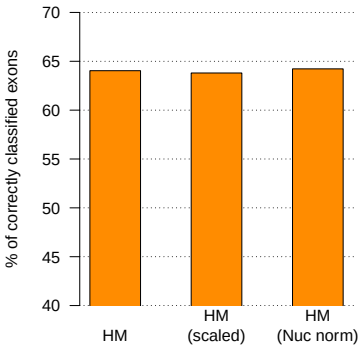

Figure 3.5: Comparison of the prediction accuracy of logistic regression models using not normalized variables, scaled variables and variables normalized for nucleosome occupancy. HM - not normalized variables, HM scaled - scaled variables, HM Nuc norm - variables normalized for nucleosome occupancy. Accuracy is calculated as the percentage of correctly classified exons.

Authors of several recent studies suggested that the enrichment of histone modifications in exons is merely a reflection of the underlying nucleosome density, and claimed that this enrichment vanishes if the number of ChIP-Seq reads for a certain modification is normalized by the nucleosome density in the same region [200, 211]. In order to test this hypothesis we decided to uncorrelate the levels of histone modifications and nucleosomes. We fitted the values for each histone modification i to the nucleosome levels in exons, represented by the vector of log transformed values of the number of MNase-Seq reads, using a linear regression model. We then repeated the logistic regression with 43 predictor variables, where each histone modification was represented by the residuals of the corresponding linear model, as described in Section 3.2. In this way we removed the effect of nucleosomes on histone modification levels, since the residuals represent the information contained in histone modifications that cannot be explained by nucleosome occupancy. We also identified significant features, as described above. The model using features normalized for nucleosome occupancy correctly predicted the outcome of alternative splicing for 64.23% of analyzed exons, a result which is comparable to the one obtained by not normalized data (accuracy = 64.04%; Fig. 3.5), and the same features were identified as significant (Fig. 3.4). This result shows that even though histone modifications might be correlated to nucleosome levels, they nevertheless contain additional information about the outcome of the splicing process. For details of the performance of the model using variables normalized for nucleosome occupancy see Appendix, Table A.2.

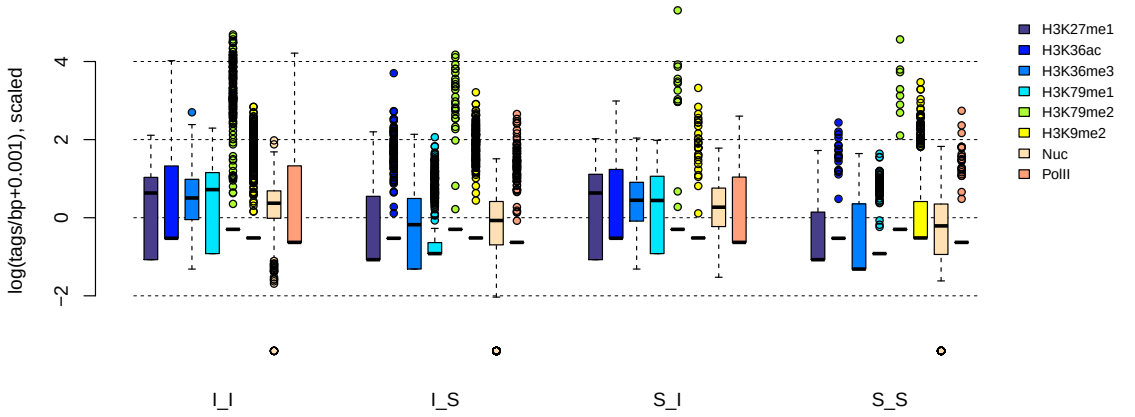

Figure 3.6: Distribution of the average number of tags/bp for significant variables. Boxplot showing the distribution the average number of ChIP-Seq or MNase-Seq tags per base pair in different groups of exons according to the results of the predictions of the logistic regression model. Only variables which have statistically significant regression coefficients are depicted. Since the magnitudes of variables differ, they were scaled by subtracting the mean and dividing by the standard deviation, in order to depict them in a single plot.

We next wanted to see how the levels of the significant features are distributed in different groups of exons according to prediction. We therefore divided the 2130 alternative exons used in the analysis into four groups: true positives (skipped exons correctly classified as skipped), false positives (skipped exons incorrectly classified as included), true negatives (included exons correctly classified as included) and false negatives (included exons incorrectly classified as skipped), henceforth simply referred to as *S_S*, *S_I*, *I_I* and *I_S*, respectively, where the first letter indicates the true class label and the second letter the class predicted by logistic regression. We then compared the distributions of the levels of the eight significant variables in the four groups (Fig. 3.6).

We first observed that the levels of H3K36me3, H3K36ac, H3K27me1 and H3K79me1/2, as well as PolII and nucleosomes were much higher in exons predicted to belong to class *I*, irrespective of the true class label. For the exons belonging to group *I_I* the observed pattern could be explained by the fact that the accumulation of nucleosomes can slow down the progression of PolII [98, 104], and in this way possibly facilitate the inclusion of the exon, an effect which has been observed before [56]. Nucleosomes and PolII have both been observed to be enriched in constitutive vs. alternative exons, and levels of nucleosome occupancy were usually negatively correlated with splice site strength, supporting their possible role in enhancing the inclusion of exons surrounded by weak splice sites [186, 200, 211]. Some studies also showed the connection between the levels of H3K36me3, H3K27me1 and H3K79me1/2 and splicing, with higher levels of these modifications being correlated with exon inclusion [9, 59, 117]. H3K9me2 was the only histone modification which exhibited higher levels in correctly predicted skipped exons (*S_S*). Negative correlation of H3K9me2 and exon inclusion was previously observed in a study by Dhami *et al.* [59].

Remarkably, the exons that are skipped from the transcript, but are incorrectly predicted to be included (S_I) have almost exactly the same average modification profile as the exons that are actually included (I_I). Such a result could be explicable if the initial assignment of alternative exons to classes S and I had been wrong, possibly because inclusion score -0.9 does not represent the optimal boundary between classes. However, varying the decision boundary does not significantly change the results of the prediction, rendering such an explanation unlikely.

Alternatively, this observation could reflect a real biological phenomenon. For example, it is possible that the same histone modification can have a different effect on alternative splicing of different groups of exons, depending on some additional features, possibly encoded within the sequence of the exon. This would imply that such features would then have to be included in the model in order to correctly predict the outcome of alternative splicing.

3.4 Influence of experimental artifacts on the prediction accuracy of the model

A number of different experimental artifacts resulting from the use of RNA-Seq and ChIP-Seq protocols could influence the calculation of predictor and response variables, and thereby bias the prediction. The inclusion ratio is calculated using the number of RNA-Seq tags mapping to the alternative exon and the other exons of the transcript that contains it, and is therefore probably susceptible to the same biases as the RNA-Seq method itself, namely the number of base pairs in the exon which can be uniquely mapped [187] and the GC content of the mapped region [60]. Each of these artifacts could potentially influence the calculation of the inclusion ratio and lead to a false assignment of an exon to either the S (skipped) or I (included) class, or alternatively lead to biases in determining the levels of nucleosomes, histone modifications and PolII in alternative exons. To investigate whether any of these experimental artifacts influence the prediction accuracy of logistic regression, we divided the 2130 alternative exons into four previously described groups according to the outcome of binary classification (S_S, S_I, I_I and I_S). A significant difference in any of the features known to introduce bias to RNA-Seq/ChIP-Seq measurements between the four groups could also influence the prediction method.

We first examined the number of base pairs of the alternative exon which can be uniquely mapped and analyzed whether this property influences the prediction accuracy of the method. To test this we generated all possible reads of length 30 for each of the 2130 alternative exons. We mapped these reads against the reference human genome (hg18) using ELAND software (Illumina) and for each alternative exon calculated the proportion of reads which can be uniquely mapped. Fig. 3.7 shows the distribution of the proportion of uniquely mappable reads in different groups of

exons. Exons belonging to group *S_S* have a significantly lower proportion of mappable reads, compared to the other groups of exons (71.2% of exons in group *S_S* have at least 80% of mappable reads, compared to 96.07%, 90.32% and 87.8% for groups *I_I*, *I_S* and *S_I*, respectively). This could imply that the absence of mapped RNA-Seq and ChIP-Seq tags for skipped exons which are correctly predicted to be skipped could be a consequence of the mappability of exons and not a real biological phenomenon.

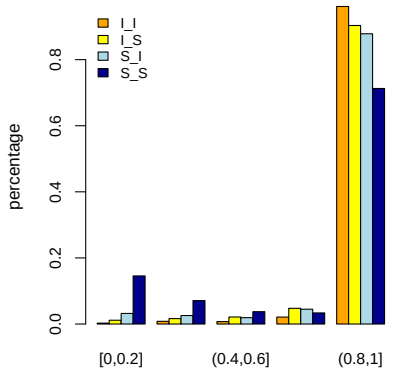

Figure 3.7: Proportion of uniquely mappable reads in different groups of alternative exons according to the prediction.

To eliminate the bias arising from different exon mappability, we removed from the analysis all the alternative exons where less than 80% of reads could be uniquely mapped against the reference human genome (196 exons). We predicted the outcome of the alternative splicing event for the remaining 1934 exons using logistic regression as described in Section 3.2. The number of accurately classified exons decreased slightly from 64.04% (using all 2130 exons) exons to 62.05% (using exons where at least 80% of the reads are uniquely mappable; Fig. 3.8). This indicates that the mappability of exons has a slight influence on the prediction accuracy of our model, but the results of the analysis are not solely a consequence of experimental biases.

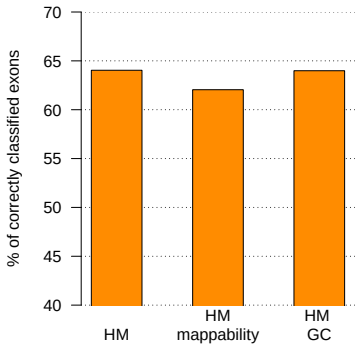

Figure 3.8: Comparison of the prediction accuracy of logistic regression models using not normalized data and data normalized for experimental biases. HM - model using not normalized variables, HM mappability - model trained on 1934 exons where at least 80% of the reads are uniquely mappable, HM GC - model trained on variables normalized for GC content. Accuracy is calculated as the percentage of correctly classified exons.

Another possible origin of bias in the analysis is the GC content of the exon itself. It has been shown that the RNA-Seq protocol shows a significant bias towards regions with higher GC content, ie. that regions with higher GC content generally have a higher read coverage than regions with low GC content [60]. Analysis of the distribution of exonic GC content in the four different groups according to the predictions

shows that there is a significant difference in GC content between exons belonging to these groups. Exons in groups *I_I* and *S_I* in general have a significantly higher GC content than exons in groups *I_S* and *S_S* (Fig. 3.9 and Table 3.2). There are two possible ways in which this finding could influence our analysis. Firstly, we use the number of ChIP-Seq tags for histone modifications, nucleosomes and PolII mapped to the alternative exon as predictor variables in the logistic regression model. If there is a strong GC bias than the measured value for each variable could actually reflect the GC content of the exon, and not the real chromatin state. If this is the case, then the prediction obtained by logistic regression would be a result of classifying together exons of similar GC content. Secondly, the difference in GC content could influence the original labeling of the exons, where some included exons could wrongly be labeled as skipped because of a lower number of RNA-Seq tags originating from lower GC content and vice versa.

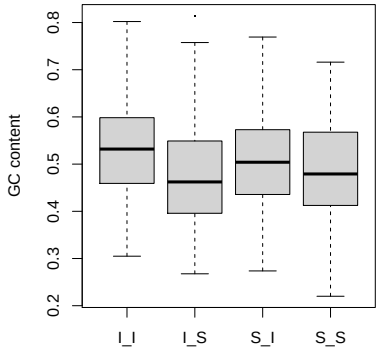

Figure 3.9: Distribution of GC content of exons belonging to different groups according to the results of the predictions.

	I_I	*I_S*	*S_I*	*S_S*
I_I	1	2.1e-26	2.6e-02	1.1e-10
I_S	2.1e-26	1	2.7e-04	1.4e-01
S_I	2.6e-02	2.7e-04	1	3.2e-02
S_S	1.1e-10	1.4e-01	3.2e-02	1

Table 3.2: p-values of the Wilcoxon rank sum test for the difference in the distribution of GC content in different groups of exons according to the results of the predictions. p-values lower than 0.05 are indicated in red.

In order to remove the possible bias arising from the sequence composition of the analyzed exons we normalized all variables for GC content. For each variable i we trained a linear model where the GC content of the exon was used as the predictor variable and the variable i as the response. We then trained a logistic regression model, as described in Section 3.2, where each original variable i was represented by the residuals of the corresponding linear model. In this way we removed the correlation between the GC content of the exon and the number of ChIP-Seq or MNase-Seq tags mapped to it. Since the results of prediction using logistic regression with variables normalized for GC content (accuracy = 63.99%; Fig. 3.8) are similar to the ones obtained using not normalized variables we conclude that the predictive power of histone modifications, nucleosomes and PolII is not merely a consequence of differences in GC content of the analyzed exons. Details of the performance of the

two models examining the influence of experimental biases on our analysis can be found in the Appendix, Tables A.3 and A.4.

3.5 Effect of transcript expression levels on the relationship between chromatin structure and alternative splicing

Examination of the coefficients of different predictor variables used for logistic regression showed that only 8 out of the original 43 variables have coefficients with statistically significant p-values, namely H3K27me1, H3K36ac, H3K36me3, H3K79me1/2, H3K9me2, nucleosomes and total PolII (Fig. 3.4). H3K27me1, H3K36me3 and H3K79me1/2 are histone modifications which have previously been shown to be correlated with the expression level of the gene and are presumably involved in transcriptional regulation [18, 189, 201]. H3K9me2 has been shown to be related to co-transcriptional splicing [6]. Since it has been shown that the level of enrichment of several of these modifications in the gene body region is highly correlated to the expression level of the gene, the enrichment of these modifications in certain alternative exons could be related to the overall expression level of the transcript and not specifically connected to splicing.

To further investigate this possible effect of transcript expression on prediction accuracy of the logistic regression model we examined the distribution of transcript expression levels in the four groups of exons divided according to the classification outcome. The results of this analysis show that the transcript expression levels are significantly higher in alternative exons predicted to be included (groups *I_I* and *S_I*) compared to the exons predicted to be skipped in the transcript (groups *I_S* and *S_S*), irrespective of their true class label (Fig. 3.10 and Table 3.3). These results indicate that the enrichment of histone modifications in the exon could simply be a consequence of the transcript expression level. If this were the case, it is possible that splicing is somehow regulated by transcript expression levels, rather than being under the influence of histone modifications.

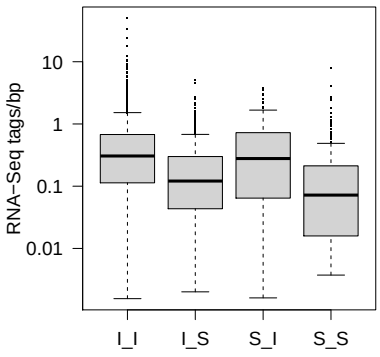

Figure 3.10: Distribution of the number of RNA-Seq tags/bp for transcripts associated with different groups of alternative exons according to the results of the predictions of the model trained on not normalized variables. The y axis is plotted on a logarithmic scale.

	I_I	*I_S*	*S_I*	*S_S*
I_I	1	4.7e-36	7.3e-02	3.6e-38
I_S	4.7e-36	1	4.8e-06	3.9e-07
S_I	7.3e-02	4.8e-06	1	1.8e-11
S_S	3.6e-38	3.9e-07	1.8e-11	1

Table 3.3: p-values of the Wilcoxon rank sum test for the difference in the distribution of the number of RNA-Seq tags/bp for transcripts associated with different groups of alternative exons according to the results of the predictions of the model trained on not normalized variables. p-values lower than 0.05 are indicated in red.

To check if the information about splicing outcome contained in chromatin structure is really dependent on transcript expression level we decided to uncorrelate values for histone modifications, nucleosomes and Pol II from the transcript expression level. To do this, we employed the strategy that we previously used to normalize the variables for effects of nucleosome occupancy and GC content (see Sections 3.3 and 3.4). Namely, for each variable i, we trained a linear model where the logarithm of the transcript expression level served as a predictor and the variable i as the response variable. Then, for each variable i we extracted the residuals of the corresponding linear model. The residuals do not contain any information which can be explained by transcript expression levels, and can therefore be used to represent the variable i normalized for expression. We then repeated the prediction of alternative splicing outcome using logistic regression (as described in Section 3.2), using the predictor variables normalized for transcript expression levels.

If the prediction accuracy of the method indeed depends only on the expression level of the transcript, then the prediction of logistic regression using normalized variables should be no better than random guessing. However, the logistic regression model still has predictive power (Table 3.4). This shows that chromatin structure really contains information on alternative splicing events, although the prediction accuracy using normalized variables (accuracy = 60.75%) is lower than for not normalized data (accuracy = 64.04%; Fig. 3.11). The distribution of transcript expression values of exons grouped by their true and predicted labels shows that the expression levels are significantly higher in groups where the true class label is *I*, compared to groups where the true class label is *S* (Fig. 3.12 and Table 3.5). However, the prediction accuracy within the groups no longer depends on the expression level of the transcript, confirmed by the fact that there is no significant difference in the distribution of expression levels of transcripts associated with exons in groups *I_I* and *I_S* or *S_I* and *S_S*. This implies that any remaining difference in histone modifications, nucleosomes or PolII enrichment observed between included and skipped exons could really be directly associated to splicing, instead of being merely correlated with the expression level of the transcript or exon.

After training the logistic regression model with variables normalized for transcript expression levels, we again examined the median p-values of regression coefficients to determine which variables contribute the most to prediction accuracy, as described

		Class labels S	Class labels I
Prediction	S	247	659
	I	177	1047

accuracy = 60.75%

Table 3.4: Prediction of exon skipping using logistic regression with variables normalized for transcript expression levels. The predicted class label for each exon is the final prediction obtained by 100 repeats of 5-fold nested cross-validation.

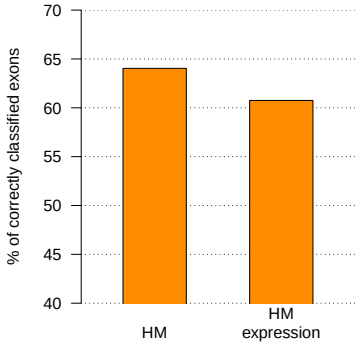

Figure 3.11: Comparison of the prediction accuracy of logistic regression using not normalized variables and variables normalized for transcript expression levels. HM - model using not normalized variables, HM expression - model using variables normalized for transcript expression levels. Accuracy is calculated as the percentage of correctly classified exons.

in Section 3.3. After controlling for the effect of expression levels, only six variables had significant regression coefficients (Fig. 3.13). Both nucleosomes and PolII are once again identified as significant features, and the fact that their regression coefficients are positive supports their possible role in enhancing exon inclusion, by kinetic coupling of transcription and splicing [120]. In the analysis conducted using not normalized data six histone modifications were determined to contribute significantly to prediction accuracy (Fig. 3.4). After normalizing for expression values, only four of these six variables (H3K36me3, H3K36ac, H3K79me2 and H3K27me1) still have significant regression coefficients. This finding shows that even though these histone modifications are connected to the elongation of transcripts, they could also be involved in regulation of splicing. On the other hand, regression coefficients for H3K9me2 and H3K79me1 are no longer significant, making them more likely to be connected to regulation of elongation rather than having a direct influence on splicing.

Once again, the distributions of histone modification levels in groups of exons according to prediction accuracy show that the levels of all significant variables are higher in groups of exons predicted to be included (I_I and S_I) than in the groups of exons predicted to be skipped (S_I and S_S; Fig. 3.14). The exons predicted to be included in the transcript (I_I and S_I) are characterized by very similar chromatin profiles. One explanation is that even though the chromatin structure is similar, the exons differ in some other respect, possibly encoded in the sequence, causing them to have different splicing outcomes. The profiles of exons predicted to be skipped (groups I_S and S_S) are also highly similar. The exceptions are the levels of H3K36me3, which are higher in included exons (I_S) compared to skipped exons (S_S). This could imply that enrichment of this modification facilitates the inclusion of exons, even though the difference in the level of H3K36me3 is too small to be detected by the method

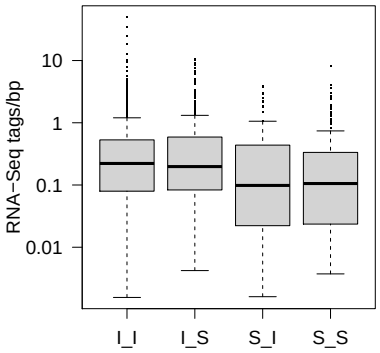

Figure 3.12: Distribution of the number of RNA-Seq tags/bp for transcripts associated with different groups of alternative exons according to the results of the predictions of the model trained on variables normalized for transcript expression levels. The y axis is plotted on a logarithmic scale.

	I_I	*I_S*	*S_I*	*S_S*
I_I	1	7.6e-01	6.6e-06	2.8e-10
I_S	7.6e-01	1	2.4e-06	2.9e-10
S_I	6.6e-06	2.4e-06	1	6.7e-01
S_S	2.8e-10	2.9e-10	6.7e-01	1

Table 3.5: p-values of the Wilcoxon rank sum test for the difference in the distribution of the number of RNA-Seq tags/bp for transcripts associated with different groups of alternative exons according to the results of the predictions of the logistic regression model trained on variables normalized for transcript expression levels. p-values lower than 0.05 are indicated in red.

used.

We conclude that, although histone modifications are correlated to transcriptional activity, they also seem to possess additional information which can be used to predict the outcome of splicing. This could imply that certain histone modifications could indeed be involved in splicing regulation, not only through their connection with transcription elongation, but also by directly influencing the splicing process.

3.6 Alternative splicing outcome is correlated to transcript expression level

The observation that transcripts in which the alternative exons are included generally have higher expression levels than transcripts where the exons are skipped is in itself interesting (p-value of the Wilcoxon rank sum test $= 2.86 \cdot 10^{-15}$; Fig. 3.15 (a)) and could imply that the level of transcription itself somehow regulates the outcome of alternative splicing. However, it is possible that this observation is due to some bias in the inclusion ratio, ie. that there is a higher probability of an alternative exon being labeled as included if it is part of a transcript with a high expression level. We therefore conducted a simulation experiment to check for a possible dependency of the inclusion ratio on the transcript expression levels. We started the simulation by determining the distributions of all possible values of exon and transcript expression

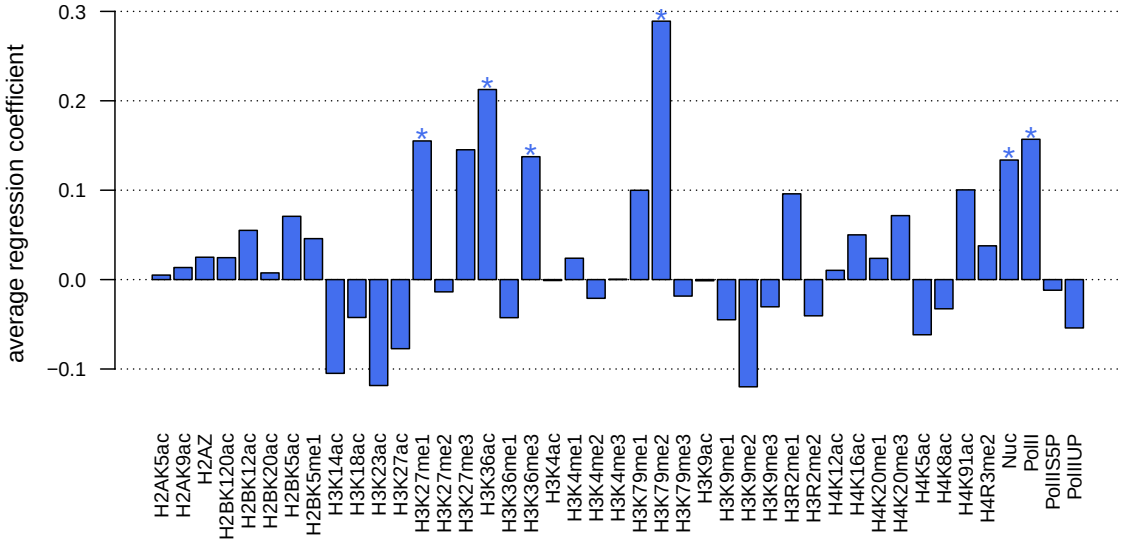

Figure 3.13: Regression coefficients of the logistic regression model trained on variables normalized for transcript expression levels. Bar plot showing the average regression coefficients of the logistic regression model. Stars mark the coefficients whose median p-values are less than 0.05 in 100 repeats of nested 5-fold cross-validation.

level in the sample of alternative exons which we used in our analysis. We then drew 100000 random pairs of values from these two distributions, and for each pair of exon and transcript expression levels calculated an inclusion ratio. Each exon was then assigned a class label, with exons being assigned to class I if the inclusion ratio was greater or equal to -0.9, and to class S otherwise. We then compared the distributions of transcript expression levels in groups of exons labeled as skipped or included using both real data and simulated data. The analysis showed that in simulated data, transcripts in which the alternative exons are skipped in general have significantly higher expression levels than the transcripts in which the exons are included (p-value of the Wilcoxon rank sum test $< 2.2 \cdot 10^{-16}$), an effect which is exactly the opposite of the one observed for real data (Fig. 3.15 (a) and (c)). Furthermore, when examining the relationship between simulated inclusion ratio values and transcript expression level, we observed that they are in general negatively correlated (Fig. 3.15 (d)). The exception is the case when the inclusion ratio is exactly equal to -1, indicating that there are no RNA-Seq tags mapped to the exons. This value is equally likely to appear across the entire range of transcript expression values. The negative correlation between the inclusion ratio and transcript expression levels implies that the probability to detect a skipped exon using this measure is higher for highly expressed transcripts than for lowly expressed transcripts. However, even though the negative relationship is also present in real data, we still observe higher expression levels for transcripts with included exons (Fig. 3.15 (a) and (b)). This is not a consequence of different transcript expression levels in real and simulated data, since there is no significant difference in their distributions (p-value of the t-test = 0.969). Based on these results, we conclude that the observed difference in expression levels of transcripts

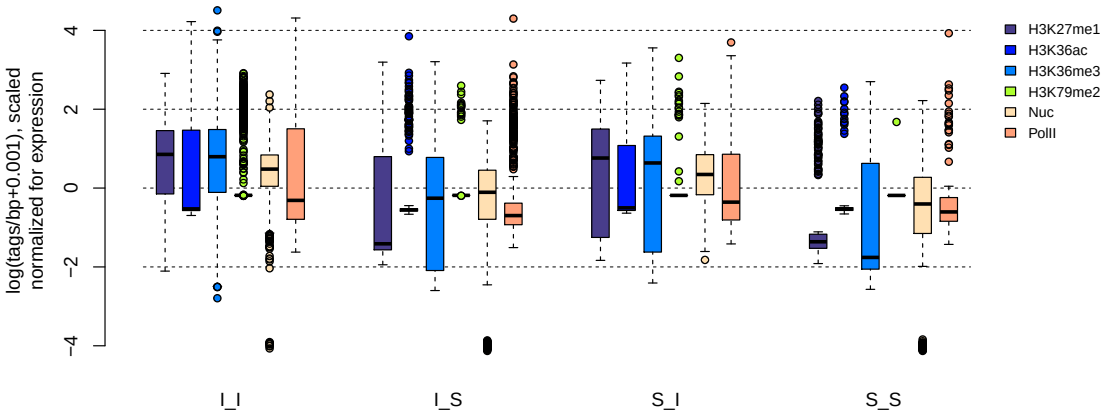

Figure 3.14: Distribution of the average number of tags/bp for significant variables of the logistic regression model trained on variables normalized for transcript expression levels. Boxplot showing the distribution the average number of ChIP-Seq or MNase-Seq tags per base pair in different groups of exons according to the results of the predictions of the logistic regression model. Only variables which have statistically significant regression coefficients are depicted. Since the magnitudes of variables differ, they were scaled by subtracting the mean and dividing by the standard deviation, in order to depict them in a single plot.

with included or skipped exons is not a bias inherent to the measure used to label the alternatively spliced exons but rather a true biological phenomenon, implying that transcript expression level could influence alternative splicing.

We hypothesized that transcript expression level influences the outcome of the splicing process, since we observed that transcripts in which the alternative exon is included in general have a higher expression level then the ones in which the alternative exon is skipped. However, this observation could also be explained inversely, ie. that the outcome of the splicing process either affects or is correlated to the measured expression level of the transcript. Although a recent study has shown that the assembly of spliceosome components does not influence PolII elongation kinetics [31], there are several additional mechanisms by which splicing could affect measured expression levels of genes.

First of all, the inclusion or skipping of the alternative exon could frameshift the downstream sequence and/or introduce a premature termination codon (PTC). Transcripts containing PTCs can be degraded by the nonsense mediated decay (NMD) mechanism, a quality control mechanism for gene expression that is coupled with translation [39]. If such scenarios indeed occur in the analyzed set of transcripts, they might provide an explanation for the difference in transcript expression levels observed between transcripts with included or skipped exons (Fig. 3.15 (a)), because NMD would lead to lower transcript levels . However, analysis of the coding sequences of 2130 transcripts used in the analysis discovered only one case in which an additional termination codon was introduced, showing that the difference in transcript

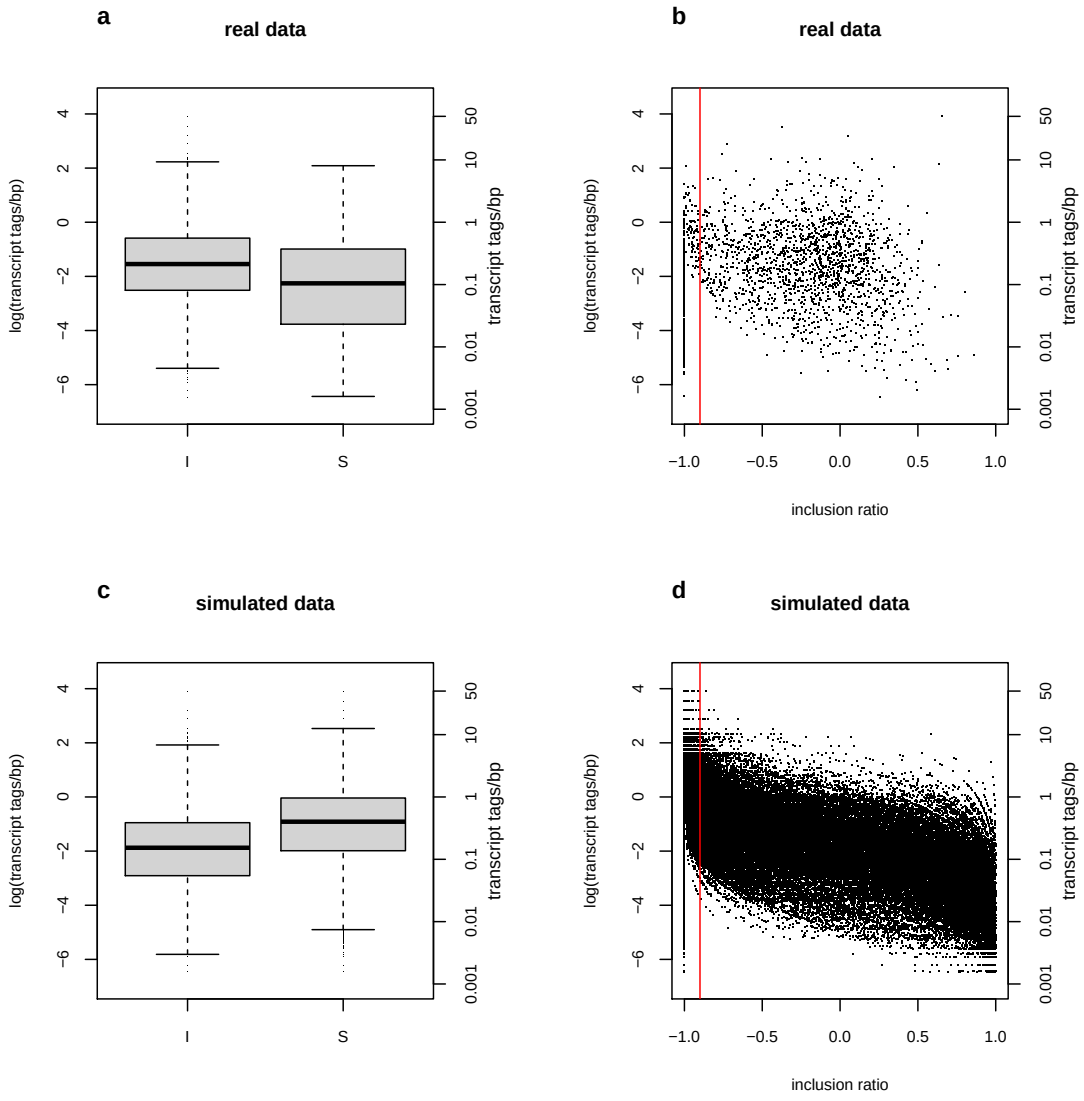

Figure 3.15: Comparison of expression levels of transcripts associated with either included or skipped exons in real and simulated data. (a) and (c) Distribution of transcript expression levels for exons labeled as included (I) or skipped (S) in real and simulated data, respectively. (b) and (d) Relationship between transcript expression levels and inclusion ratio in real and simulated data, respectively. The red line denotes the decision boundary used to label the alternative exons as either skipped or included.

expression levels cannot be explained by NMD.

Secondly, the alternative splicing outcome could be correlated with the degradation rate of the transcript and in this way connected to the measured expression level. Alternative exons used in our analysis were chosen by pairwise comparison of Ensembl transcript annotations. Although we restricted the analysis to exon skipping events, we searched only for events occurring in the coding region of the transcript and therefore the possibility that the two transcripts differ in some other respect, for

example have different 5' and/or 3' untranslated regions (UTRs), cannot be ruled out. There are at least two mechanisms, namely microRNA (miRNA) and AU-rich element (ARE) mediated degradation, where the appearance of sequence motifs in the 3' UTR has previously been connected with the degradation rate of the transcripts. Comparison of 3' UTR regions of 2130 pairs of transcripts corresponding to alternative splicing events used in our analysis showed that in 44.08% of cases both transcripts have the same 3' UTR, in 1.17% of cases neither transcript has a 3' UTR and in 54.74% of cases either only one of the transcripts has a 3' UTR or they have different 3' UTRs, suggesting that the difference in expression level could indeed arise from differences in degradation rate caused by the structure of the 3' UTR region.

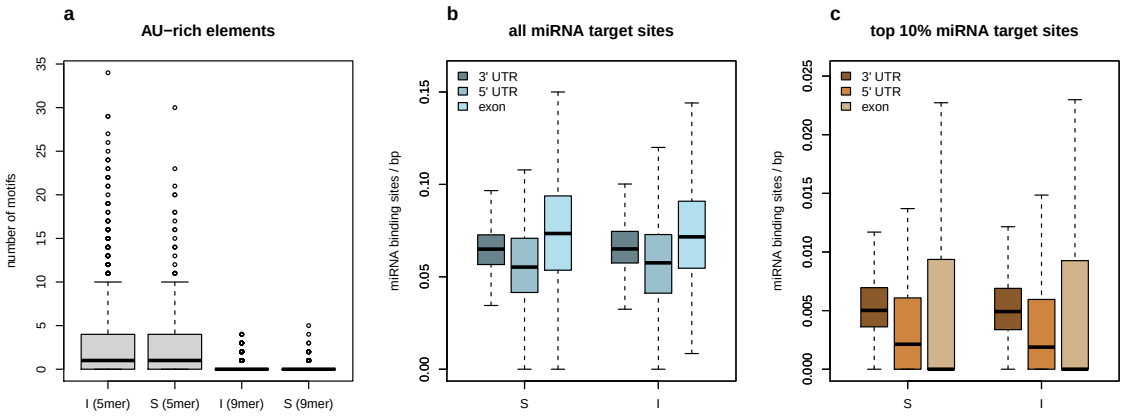

Figure 3.16: Distribution of different sequence features known to regulate mRNA degradation in groups of transcripts associated with either skipped or included exons. I - transcripts where the alternative exon is included, S - transcripts where the alternative exon is skipped. Outliers are not depicted in plots (b) and (c).

We first investigated whether the difference in the expression levels of transcripts in which the alternative exon is included or skipped could be explained by the presence of AU-rich elements (AREs) in the 3' UTR region. AREs are usually characterized by a uridin rich region and the presence of a number of (often overlapping) pentamer (AUUUA) or nonamer (UUAUUUAWW; W = A or U) motifs [43]. Each alternative exon in our analysis is associated with two transcripts, one where the exon is annotated to be included and one where the exon is annotated to be skipped from the transcript. For each alternative exon, we chose one of these two transcripts according to the label assigned using the inclusion ratio (ie. for the exons labeled as included we chose the transcript where the exon was also annotated as included and vice versa) and retrieved the 3' UTR sequence of this transcript. We then counted the number of overlapping occurrences of either the pentamer or the nonamer motif in the transcripts where the 3' UTR length was greater or equal to the motif length. Out of the 2009 transcripts whose 3' UTR was at least 5 bp long, 64.31% contained at least one pentamer motif. However, there was only a slight difference in the distribution of the number of pentamer motifs in the 3' UTR region of transcripts where the exon is included compared to transcripts where the exon is skipped (Fig. 3.16 (a); p-value

of the Wilcoxon rank sum test = 0.047). At least one nonamer motif was present in 3' UTRs of 15.14% of the 2008 transcripts whose 3' UTR was at least 9 bp long, but there is again no significant difference in the number of nonamer motifs between different groups of transcripts (Fig. 3.16 (a); p-value of the Wilcoxon rank sum test = 0.756). We conclude that the observed differences in expression levels of transcripts used in our analysis are not a consequence of the distribution of AREs in the 3' UTR region.

We next studied the presence of miRNA target sites in 3' UTR regions of transcripts associated with either included or skipped alternative exons. We downloaded information on miRNA families from TargetScan Human Release 5.1 [130]. We removed all miRNA families which did not include any previously annotated human miRNA, leaving us with 545 miRNA families comprising altogether 677 different miRNAs. For each miRNA family we extracted sequences matching positions 2-8 of the mature sequence. The 5' region of miRNA corresponding to nucleotides 2-7 is called the miRNA "seed" and is shared between all members of one miRNA family. We then used an adapted version of the TargetScan algorithm, not taking conservation into account, to scan 3' UTRs of chosen transcripts and identify 7mer-A1 (seed match + A at position 1), 7mer-m8 (seed match + match at position 8) and 8mer (seed match + match at position 8 + A at position 1) matches to the seed sequence for each miRNA family [21]. For each analyzed transcript we counted the total number of 7mer-A1, 7mer-m8 and 8mer matches for all miRNA families in the 3' UTR region. For each transcript we normalized the total number of predicted miRNA target sites by the length of the 3' UTR region to account for the higher probability of random matches in longer 3' UTRs. All transcripts without a 3' UTR were removed from the analysis. We observed no significant difference in the distribution of the total number of miRNA target sites in 3' UTRs of transcripts where the alternative exon is included compared to transcripts where the exon is skipped (Fig. 3.16 (b); p-value of the Wilcoxon rank sum test = 0.1973).

Since miRNAs are believed to be involved in tissue-specific expression regulation [124], not all miRNAs will be equally expressed in a certain cell type. This implies that even though there is no difference in the total number of miRNA target sites in 3' UTRs of analyzed transcripts, they could still be differently targeted by miRNAs highly expressed in CD4+ cells. We therefore obtained data for miRNA expression in CD4+ cells produced by deep sequencing of stem-loop sequences of human miRNAs [19]. We mapped the measured number of tags for stem-loop sequences to the corresponding mature miRNA. In the cases where a single mature miRNA can be produced from more than one stem-loop sequence we summed the total number of tags measured for all associated stem-loop sequences. In total, 543 out of 677 miRNAs had measured expression levels in CD4+ cells. To check if there is a difference in the number of target sites for highly expressed miRNAs in the 3' UTRs of the analyzed transcripts, we ordered the 543 miRNAs according to their expression level in CD4+ cells and kept only the ones whose expression was in the top 10% (54 mature miRNAs belonging to 38 different miRNA families). We calculated the total number of target sites for 38 miRNA families in 3' UTRs of transcripts where the alternative exon was either

included or skipped, but saw no significant difference in the distribution of target sites for highly expressed miRNAs between these two groups of transcripts (Fig. 3.16 (c); p-value of the Wilcoxon rank sum test = 0.2977).

Despite the fact that miRNAs predominantly bind to 3' UTR regions of transcripts [21], there are reports of binding of miRNAs to targets in 5' UTRs or coding regions, although the efficiency of targeting and reduction of mRNA stability seems to be much lower than for targets in 3' UTR regions [62, 90, 130]. We therefore decided to investigate the distributions of the total number of target sites for all human miRNA families and highly expressed miRNAs in the 5' UTRs of analyzed transcripts, as well as in the alternative exon itself. The analysis was conducted analogously to the analysis of miRNA target sites in the 3' UTRs and a number of potential miRNA binding sites was identified. However, there was no significant difference in the distribution of the total number of miRNA target sites either in the 5' UTR or alternative exons between the two groups of transcripts (Fig. 3.16 (b); p-values of the Wilcoxon rank sum test: 0.321 and 0.578, respectively). The distribution of target sites for highly expressed miRNAs also did not show any significant difference between transcripts with included or skipped exons, either for 5' UTRs or for alternative exons themselves (Fig. 3.16 (c); p-values of the Wilcoxon rank sum test: 0.694 and 0.112, respectively).

Our analysis does not show any obvious evidence that the difference in transcript expression between groups of transcripts in which the alternative exon is either skipped or included is a consequence of targeting by microRNAs. However, we cannot completely exclude this possibility. First of all, although many methods for miRNA target prediction exist, accurately predicting true biological targets and their influence on the degradation of mRNA is still an open challenge. In addition to the number of target sites, other features, such as target site conservation, type and length of the match to the seed region or other characteristics of the 3'UTR, have been used to model the experimentally measured degradation rate of mRNAs containing target sites for specific miRNAs. However, even though this approach is more complex than the one used in this analysis, it has resulted in predictions of limited accuracy [83]. Furthermore, additional effects, such as overlaps between miRNA target sites or the existence of feedback loops between miRNAs, will probably have to be taken into account to successfully model miRNA-mediated degradation, a task which is beyond the scope of this analysis.

In conclusion, we observed that transcripts in which alternative exons are included in general have higher expression levels than transcripts in which alternative exons are skipped. This observation seems to reflect a real biological phenomenon, since we could find no evidence of its being a consequence of either a measurement bias (eg. a bias inherent to the inclusion ratio used to label exons as either included or skipped) or of the different degradation mechanisms used to control mRNA levels in the cell. This finding implies that further studies will be needed to completely understand the interplay of transcription and splicing.

3.7 Expanding the model using sequence features

The analysis of the results of the prediction of alternative splicing using histone modifications revealed that subsets of exons which are included in the transcript have similar chromatin profiles with subsets of exons which are skipped. However, even though the chromatin profile is similar, the splicing process has a different outcome (Sections 3.3 and 3.5). We were interested to see what could be the cause of this observation.

The 5' splice site (5'SS) and 3' splice site (3'SS) contain invariant dinucleotides GT at the 5' end of the intron and AG at the 3' end of the intron, respectively. The alignment of known splice sites has also enabled the identification of consensus sequences of the 5'SS and 3'SS which can then be used to predict splice site locations [248]. In general, splice sites whose sequences are highly similar to the consensus sequence are considered to be strong splice sites, and therefore more efficient in promoting splicing of the exons, in contrast to weak splice sites highly divergent from the consensus sequence. The outcome of the splicing process can also be influenced by the presence of various splicing regulatory elements (SREs) which can function as either enhancers or silencers of splicing. These elements can be present both in the exon and the intron sequence, and are called exonic splicing enhancers (ESE) or silencers (ESS) and intronic splicing enhancers (ISE) and silencers (ISS) [44, 224]. A possible explanation of different splicing outcomes of exons with similar chromatin profiles would be if chromatin modifications have different influences on splicing, depending on the underlying sequence of the exon. To test this hypothesis we proceeded to build additional logistic regression models which include different sequence features among predictor variables.

We first included information on splice site strength in the model. For each of the 2130 alternative exons used in the analysis we scored the strength of the 5' and 3' splice site using four different methods: counting the number of matches of the splice site and the consensus sequence of human 5' and 3' splice sites [248], using a score based on the position specific scoring matrix (PSSM) of the 5' and 3' splice site based on 118,000 constitutive exons in non-redundant genes [248], MaxEntScan score which models the sequences of 5' and 3' splice sites using the Maximum Entropy Principle and tries to capture dependencies between adjacent and non-adjacent positions in the pattern [243], and a score produced by GeneSplicer, a method which combines decision trees with Markov models to characterize patterns of sequences around 5' and 3' splice sites [164]. We then trained four different logistic regression models. In each of the models we used the 43 variables corresponding to average numbers of ChIP-Seq or MNase-Seq tags/bp in alternative exons for histone modifications, PolII and nucleosomes (henceforth "chromatin features"). These variables were normalized for transcript expression levels, as described in Section 3.5. In each model we additionally included the strengths of the 5'SS and 3'SS, calculated by one of the four methods described, as predictor variables. Each model was trained using 100 repeats of nested 5-fold cross-validation, as described in Section 3.2. The results of the prediction show that the

scores obtained by the GeneSplicer algorithm have the highest influence on prediction accuracy (accuracy = 62.54%, compared to the model using only chromatin features accuracy = 60.75%; Fig. 3.17). We then repeated the analysis using scaled data, to account for differences in magnitude of scores obtained using different methods of calculating splice site strength. The variables corresponding to different scores were scaled by subtracting the mean and dividing by the standard deviation of the variable. The results of prediction using scaled data are comparable to the ones using not scaled data (Fig. 3.17), implying that the higher contribution of scores calculated by the GeneSplicer algorithm is not a consequence of the differences in magnitudes of individual predictor variables. Detailed descriptions of the performances of all the models trained on chromatin features and variables corresponding to splice site strength are available in the Appendix, Table A.5.

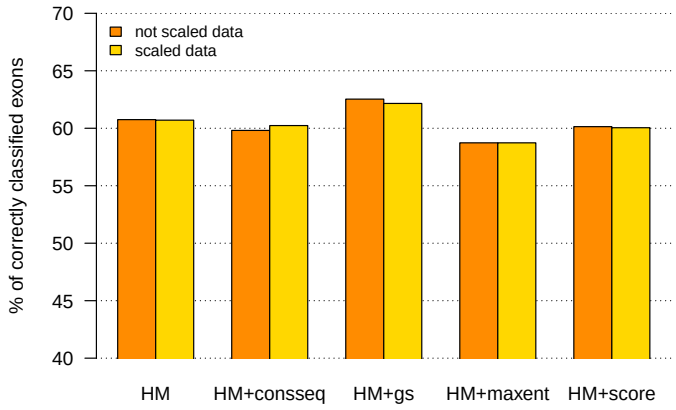

Figure 3.17: Comparison of the prediction accuracy of logistic regression models trained on chromatin features and different measures of splice site strength. The labels denote the features used to train the individual models. HM - chromatin features. Four different measures of splice site strength were used: consseq - matches to the consensus sequence, maxent - maximum entropy score, gs - GeneSplicer scores, and score - scores based on a PSSM of the splice site. Accuracy is calculated as the percentage of correctly classified exons.

We further expanded our model to include the occurrence of different splicing regulatory elements in the intronic regions surrounding the analyzed exon and the exon itself. We counted the occurrence of binding motifs for four different SR proteins (SF2/ASF, SC35, SRp40 and SRp55) which have previously been shown to function primarily by enhancing the inclusion of the exon to which they are bound [81]. We used position specific scoring matrices which describe binding motifs for these proteins produced from SELEX experiments [37] to identify potential binding sites of SR proteins in the exons. For each exon and each SR protein we calculated the score for binding of the protein in non-overlapping windows of length n, where n is the length of the binding motif according to the PSSM. For each exon we then counted the number of windows where the score is significant. The significance of the score was determined by comparing it with a threshold value T for each protein

(T_{SC35} = 2.383; T_{SRp40} = 2.670; T_{SRp55} = 2.676; SF2/ASF was represented by two different matrices $T_{SF2/ASF}$ = 1.956 and $T_{SF2/ASF(IgM-BRCA1)}$ = 1.867), where the threshold values were the ones defined by Cartegni *et al.* [37]. We also downloaded the sequences of 238 hexamers (RESCUE-ESE sequences; [67]), 133 decamers [226] and 1019 octamers [247] which were identified as potential exonic splicing enhancers or silencers in different computational studies and counted the number of overlapping occurrences of these sequences in the exon region.

We also counted the occurrences of motifs for binding of SRE-binding proteins FOX1/2 ([U]GCAUG [149]), Nova (YCAY; Y = U or C [215]) and PTB (UYUYU; Y = U or C [240]) in the exon, as well as the intronic regions of 200 bp upstream and downstream of the exon. We checked the same three regions for the presence of G-runs and CA-repeats, since both of these elements were shown to be involved in splicing regulation [102, 145]. For each region we searched for the longest stretch of the motif G_n or CA_n, $n \geq 3$ and counted the number of occurrences of these motifs in the region. We furthermore calculated the percentage of U nucleotides in the region of 200 bp downstream of the exons, since it has been shown that splicing factors T-cell restricted intracellular antigen I (TIA1) and TIA1-like 1 (TIAL1) can bind to U-rich sequences downstream of 5' splice site and in this way enhance the inclusion of exons [12].

We compared the performance of five logistic regression models trained on an increasing number of variables, in order to determine if sequence features indeed contribute to the accuracy of the prediction of alternative splicing. The first model was trained using only variables for histone modifications, PolII and nucleosomes normalized for expression levels (chromatin features; 43 variables). In the second model we also included the information on splice site strength determined by the GeneSplicer algorithm (gs; 2 variables). The third model contains additional information of the occurrence of sequence regulatory elements in the exon region (ESE/S; 13 variables). The fourth model contains variables describing the distribution of intronic splicing regulatory elements (ISE/S; 11 variables), instead of exonic splicing regulatory elements. The fifth model includes information on the occurrence of both exonic and intronic splicing regulatory elements. All five models were trained using 100 repeats of nested 5-fold cross validation, as described in Section 3.2.

The results of this analysis discovered that including increasing amounts of information on the sequence features of the exon indeed contributes to prediction accuracy of the model (Fig. 3.18). The best prediction accuracy was achieved by the model which includes information on splice site strength and intronic splicing regulatory elements in addition to histone modifications, PolII and nucleosomes achieved the best prediction accuracy (HM + gs + ESE/S; accuracy = 64.55%). Although the difference in accuracy is not very high, this model still performs better than the model containing only information on chromatin structure (HM; accuracy = 60.75%). The results of the analysis using scaled data (to account for differences in the magnitudes of variables) are comparable to the ones using not scaled data. Detailed performances

of all models which include information about different exonic and intronic splicing regulatory elements are available in the Appendix, Table A.6.

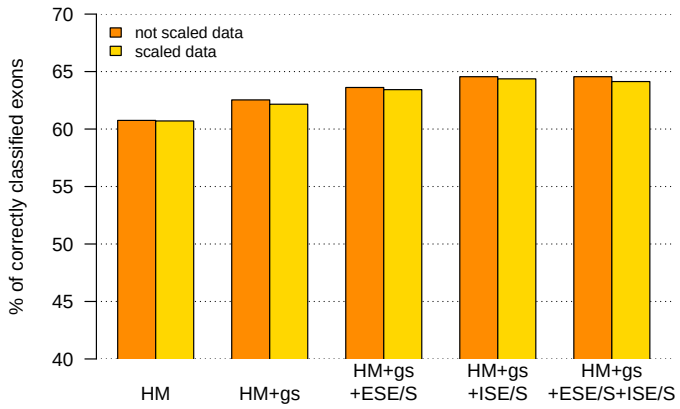

Figure 3.18: Comparison of the prediction accuracy of logistic regression models trained on chromatin features, splice site strength and splicing regulatory elements. The labels denote the features used to train individual models. HM - chromatin features, gs - 3' and 5' splice site strength calculated by GeneSplicer, ESE/S - exonic splicing enhancers and silencers, ISE/S - intronic splicing enhancers and silencers. Accuracy is calculated as the percentage of correctly classified exons.

We proceeded by studying the average regression coefficients of the variables in the model containing variables describing chromatin structure and all the sequence features (HM + gs + ESE/S + ISE/S), in order to determine which variables contribute the most to prediction accuracy. We analyzed the model where the variables were scaled before training, to ensure that the coefficients are not influenced by the magnitude of the variable. We determined that 14 variables had a significant contribution to prediction accuracy. The significant variables were identified as the ones whose regression coefficients had median p-values lower than 0.05 in 100 repeats of the nested 5-fold cross-validation, as described in Section 3.3. Fig. 3.19 shows the average regression coefficients of the significant variables. Out of the variables describing chromatin structure, H3K27me1, H3K36ac, H3K36me3, H3K79me1 and H3K79me2 were identified as having a significant influence on prediction accuracy. These variables, as well as PolII, are all positively associated with exon inclusion. These results are in agreement with analyses of linear models trained only on features representing the chromatin structure of alternative exons (Sections 3.3 and 3.5). Eight variables containing information on sequence features were identified as important in our analysis. Four of these variables, namely the strength of the 3'SS and 5'SS, as well as the number of RESCUE-ESE sequences and the U-content of the downstream region, have a positive influence on exon inclusion. The regression coefficients of features describing the occurrences of PTB and FOX binding sites in both the upstream and the downstream region were negative, implying that these elements are likely to promote exon skipping.

In conclusion, we determined that the inclusion of features associated with sequence

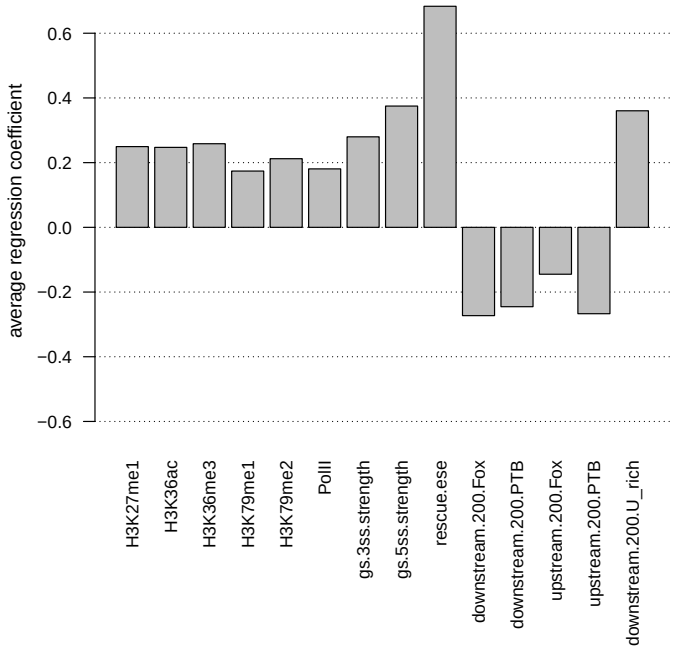

Figure 3.19: Regression coefficients of the logistic regression model trained on chromatin features, splice site strength and splicing regulatory elements. Bar plot showing the average regression coefficients of variables used in logistic regression to predict alternative splicing. Only the variables with significant regression coefficients are depicted (median p-value < 0.05 in 100 repeats of 5-fold cross-validation).

elements known to influence alternative splicing significantly improved the accuracy of predictions obtained using logistic regression. This result shows that both chromatin structure and sequence properties of alternative exons contain information of the outcome of alternative splicing. We propose that the relationship between these different properties of alternative exons should be studied further to better understand the interplay of various mechanisms of splicing regulation.

3.8 Conclusions

In this chapter, we showed that levels of histone modifications, nucleosomes and PolII are predictive of the splicing outcomes of alternative exons. These results are not dependent on experimental artifacts, such as the mappability or GC content of alternative exons. Furthermore, we identified several histone modifications which have a significant contribution to prediction accuracy, implying that they could be directly related to the splicing process. We also determined that splicing is dependent on expression levels of transcripts, confirming the existence of functional coupling of transcription and splicing. Finally, we observed that chromatin structure and sequence features both contain information about alternative splicing, suggesting that many different mechanisms are involved in the regulation of this process.

Chapter 4

Discussion

This thesis presented results of analyses in which publicly available measurements of localization of histone modifications were used to elucidate the relationship between histone modifications and cellular processes, namely transcription and splicing. In this chapter, we discuss the main findings of our analysis, put them in the context of the current knowledge of these topics and present possible future improvements.

4.1 Relationship between histone modifications and transcription

Histone modifications are quantitatively related to gene expression levels. Histone modifications have been linked to various chromatin dependent processes, including transcription [131]. We used linear models to study the nature of this relationship and showed that the levels of histone modifications at the promoter are quantitatively related to the expression level of human genes (Section 2.1). Other studies classified the promoters for each modification into groups [228, 245], e.g., modification X is present or absent. Discretization ought to have two beneficial effects, namely the reduction of noise and parameters. Although discretization is necessary in some modeling approaches to reduce the number of parameters, e.g., learning a Bayesian network [228], in our approach, it increases the number of parameters, because one has to choose at least one threshold for each modification in addition to the slopes in the linear regression model. If discretization is indeed beneficial for modeling gene expression, we expect that the results of a discrete model should be better than a corresponding continuous model. However, we found that although models using discretized features are able to reproduce the general trend in expression values, the mean squared error of prediction is significantly higher than for continuous models trained on the same features. This is true both for the full linear model and models trained on combinations of three histone modifications (Section 2.2). We conclude that discretization has no beneficial effect on the prediction accuracy and argue that in our modeling framework discretization is not necessary and is even reducing the predictive power at the cost of increasing the number of parameters.

Histone modifications have different contributions to the prediction accuracy of the model of gene expression. We demonstrated that only a few histone modifications are necessary to faithfully model gene expression (Section 2.2). This finding can be understood if one assumes that the histone modifications belong to different groups, whose members are either involved in transcription or not. The modifications within the transcription related groups provide almost the same information and our approach selects one representative modification. Alternatively, the selected histone modifications are involved in distinct steps during the transcription cycle.

We used feature selection to identify modifications which are "important" for predicting gene expression levels. Upon analyzing all promoters, we found that H2BK5ac, H3K27ac, H3K79me1, and H4K20me1 are overrepresented in models giving rise to the highest prediction accuracy in CD4+ T-cells. A recent study identified a common set of 17 modifications (mainly acetylations), referred to as the backbone. These modifications colocalize and their levels are well correlated [228]. Genes with all of these backbone modifications present tend to be expressed, suggesting that either all or a subset of them are involved in transcription. Our analysis revealed that only two of these modifications, H3K27ac and H2BK5ac, are important for modeling gene expression. This indicates that the remaining backbone modifications carry either redundant information or are less important for gene expression. Furthermore, the other two important modifications, H3K79me1 and H4K20me1, have been shown to be enriched in highly expressed genes, along with the modification backbone [228]. This observation is in line with the idea that H3K79me1 and H4K20me1 are also involved in transcription. Thus, we conclude that our approach identified histone modifications which are likely to be key players in the transcriptional process.

Our ability to identify important modifications, out of a highly correlated set, stems from our approach of using actual levels of histone modifications, rather than discretizing the data into enriched and depleted regions as employed by previous studies (e.g. [228]), which implicitly assumes that the modifications encode the on/off status of genes. The importance of these modifications could imply that they are critically required to recruit protein complexes necessary for transcription, while the enrichment of additional modifications could be necessary, but not rate limiting. Since the importance of these modifications is a reflection of their combinatorial influence, not their individual correlation to expression, we reason that the analysis of importance of modifications is a more suitable approach to drawing conclusions about their possible functions than a simple analysis of the occurrence of modifications at active or repressed genes.

Differential requirement of histone modifications in high Vs. low CpG content promoters. We tested whether the identified "important" modifications depend on the promoter structure of the analyzed genes. Indeed, we identified different sets of modifications important for modeling gene expression driven by low-CpG content promoters (LCPs) or high-CpG content promoters (HCPs). In LCPs, we found that H3K4me3 and H3K79me1, while in HCPs H3K27ac and H4K20me1,

were identified (Section 2.3). These assignments can be reproduced using RNA-Seq instead of the microarray data, suggesting that a possible measurement bias due to the microarray technology is not a major factor in our analysis.

The reason for the difference in the important histone modifications in LCPs and HCPs is unclear, but indicates that different regulatory mechanisms act on these two promoter types. A possible clue for the function of the selected modifications is provided by the analysis of their average normalized tag densities in the region surrounding the transcription start site (TSS; see Fig. 2.12). H3K4me3, H3K27ac, and H2BK5ac have the highest levels at the promoter, with the highest peak around 100 base pairs downstream of the TSS. H3K79me1 is enriched along the gene body, and H4K20me1 shows two distinct patterns: a peak close to the promoter at a similar position to H3K4me3 and H3K27ac, and an enrichment across the gene body region. The localization of these histone modifications suggests that H3K27ac, H2BK5ac, H3K4me3, and H4K20me1 function during transcription initiation and/or promoter clearance, whereas H3K79me1 and H4K20me1 are involved in transcription elongation.

Important histone modifications could be related to distinct steps of the transcription cycle. The observation that different modifications were identified as important in HCPs and LCPs indicates that these subsets of genes are regulated at different steps of the transcription cycle. To investigate this further, we studied the average RNA polymerase II (PolII) density at the promoter and gene body regions of genes with HCPs and LCPs, depending on their expression level (Section 2.3). We found that in LCPs PolII is only present when they drive expression, while HCPs tend to accumulate PolII regardless of the expression status of the corresponding genes (see Fig. 2.13). Thus, either preinitiation complex (PIC) formation, Pol II recruitment, or initiation of transcription can be a rate limiting step in LCPs, whereas HCPs are most probably regulated at the transition from initiation to elongation. We combined these findings with the important modifications identified for different groups of genes to relate the possible functions of different histone modifications to distinct steps of the transcription cycle.

Transcription starts with the assembly of the PIC, which is followed by PolII recruitment and transcription initiation [63, 206], steps that are likely to be regulated in LCPs. We found that the levels of H3K4me3 are very important for the prediction of gene expression driven by LCPs, suggesting that H3K4me3 is involved in transcription initiation (and steps preceding it). Furthermore, since the localization analysis revealed that H3K4me3 has its highest levels around 100 base pairs downstream of the TSS, it is enriched at the right place to be involved in initiation, in line with previous results (e.g. [87] and references therein). Transcription proceeds with the transition from an initiating PolII to an elongating PolII [63, 206], which is a likely target of regulation in HCPs. H3K27ac (and to a lesser degree H2BK5ac, which are highly correlated and likely either to act redundantly or to occur only together) and H4K20me1 were found to be important for the quantitative modeling of gene expression driven by HCPs, suggesting that they are connected to the transition from

initiation to elongation. The localization analysis supported this view, as all three modifications exhibit distinct peaks at around 100 base pairs downstream of the TSS at a similar location as H3K4me3.

After initiation, PolII transcribes roughly 50 nucleotides before it stops for capping and possibly for loading additional components required for elongation and RNA processing [169, 176, 179]. Thus, all three modifications are located at the right place to recruit the activity that is required for the transition to elongation, as just upstream of this location both transcription initiation and the transition to elongation are taking place. A possible action of H3K27ac might be to prevent the repressive tri-methylation of the same residue, since H3K27ac and H3K27me3 are mutually exclusive. H3K27me3 is regarded as a hallmark of repression, which is facilitated by the recruitment of the PRC1 complex, leading to H2A monoubiquitination and repression of elongation. H2A monoubiquitination prevents the recruitment of FACT, which is required for elongation through nucleosomes ([250], and references therein). Alternatively, H3K27ac itself could be recognized by a protein complex required for the transition to elongation, taking an active role in the regulation of transcription. The functions of H2BK5ac and H4K20me1 in general and in particular in relation to transcription are not well understood.

In LCPs the one-modification model incorporating the levels of H3K27ac performs best, suggesting that the transition to elongation is also regulated in LCPs. In support of this idea we found that LCPs show a substantial albeit lesser accumulation of PolII at the promoter when they drive expression (Fig. 2.13). If all recruited PolII were to be progressing into elongation, such an enrichment at the promoter would not be expected. However, the models using combinations of H3K4me3 and H3K79me1 outperform any combination involving H3K27ac. This finding can be explained assuming the following scenario: H3K27ac is added contingent on the presence of an initiating PolII (possibly determined by H3K4me3) as well as on external stimuli, which exert the regulatory input. Subsequently, H3K27ac facilitates the transition to elongation. Therefore, the levels of H3K27ac is correlated to both the rate of transcription initiation and the transition to elongation. Thus, H3K27ac is the single most correlated modification. However, by moving from the one- to models incorporating more modifications in LCPs, the best model contains H3K4me3 and H3K79me1. These two modifications contain information about transcription initiation (H3K4me3; see above) and the successful transition to an elongating Pol II (H3K79me1; see below), while H3K27ac harbors only partial information about the progression through the transcription cycle.

Finally, PolII is elongating, which enables it to escape the promoter region and to transcribe the main portion of the transcriptional unit [63, 206]. Localization analysis suggests that H4K20me1 and H3K79me1 are involved in transcription elongation (Fig. 2.12). H4K20me1 exhibits a peak near the TSS as well as an increase in the gene body, indicating that it has two roles, one for the transition to elongation and one for elongation. H3K79me1 is almost absent at the TSS and its levels rising

more downstream in the gene body, indicating that it is involved in transcription elongation, in line with previous observations [189, 201].

Based on our analysis, we propose a preliminary "Histone code of transcription", which relates histone modifications to the progression through the transcription cycle. First, H3K4me3 signals the completion of initiation. The level of H3K4me3 directly correlates with the rate of PIC formation, PolII recruitment and transcription initiation, and in LCPs these steps are rate limiting. For HCPs, these steps are not rate limiting. H3K27ac, H2BK5ac and H4K20me1 are likely to be involved in the transition to elongation. This step is critically regulated in HCPs and to a lesser extent also in LCPs. Finally, H4K20me1 and H3K79me1 are correlated to transcription elongation. Our model awaits critical testing by experimental means to reveal whether identified modifications are actually a cause for the indicated processes, and whether the important modifications themselves are rate limiting for these steps or they are merely representatives of groups of related modifications.

Histone modification levels are predictive of gene expression across different cell types. Because we showed that histone modification levels are predictive of the gene expression levels in CD4+ T-cells, we further investigated whether this is a universal property which holds true for other cell types. We were able to successfully predict expression of genes in CD36+, CD133+ and IMR90 cells, using histone modification data measured in these cells and model parameters trained on CD4+ data (Section 2.4). Significantly, the prediction accuracy does not depend strongly on the level of change in expression in different cell types. Thus, our results establish the idea that the relationships between histone modification and gene expression are general, since the model performed successfully both in closely related CD36+ and CD133+ cells and highly divergent IMR90 cells. Furthermore, they underscore that the histone modifications and the transcriptional process are tightly connected to each other. We want to emphasize that our analysis as well as the data do not allow for deciding whether the histone modifications are cause or consequence of transcription, because the uncovered relationships are correlative in nature and therefore inherently undirected. However, our results imply that the histone modifications are very close to PolII in the regulatory network controlling its activity. Whether they are upstream and/or downstream has to be elucidated in further experimental studies.

Information of transcript stability could improve models of gene expression. Although histone modification levels at the promoter are highly predictive of the expression levels of genes, we found that they cannot explain all of the variance of measured expression levels. We determined that at least some amount of unexplained variance could be caused by different degradation rates of analyzed transcripts (Section 2.5). We therefore expanded our model to include features which provide information on mRNA stability, namely the number of AU-rich elements and miRNA target sites (Section 2.6). We observed that the inclusion of features connected to AU-rich elements increases the prediction accuracy of the model, and that these features are negatively correlated to expression levels of genes, confirming the experimentally determined negative effect of AU-rich elements on mRNA stability [17]. However,

features connected to miRNA targets did not cause any significant increase in prediction accuracy. We conclude that, although miRNAs are presumably involved in regulating the degradation rate of transcripts used in this analysis, the approach that we used is inappropriate to model such a relationship. First of all, we used context scores of miRNA target sites to model the influence of miRNA on degradation rates of transcripts. These context scores were based on a simple linear model relating the context features of the 3' untranslated region to the level of down-regulation of transcripts [83]. It has recently been proposed that the predictive power of such context features could be improved using support vector regression [27]. This implies that using a more complicated model to predict the extent of transcript downregulation mediated by miRNAs might be beneficial for our analysis. Secondly, the context scores were trained on HeLa cells transfected with different miRNAs. We cannot exclude the possibility that, although the context scores might be able to predict differences in expression of transcripts caused by overexpression of different miRNAs, the endogenous miRNAs might cause smaller differences in transcript expression levels, which might be too subtle to be captured by such a model. Furthermore, context scores were originally used to model down-regulation caused by single miRNA target sites, while the transcripts we analyzed contained multiple target sites. We believe that more advanced models will have to be developed to model the complicated relationships between multiple target sites, including their possible overlap and mutual influence of different miRNAs. We conclude that a future improvement in modeling the degradation rate of transcripts could lead to even more accurate models of gene expression.

4.2 Connection of histone modifications to the regulation of alternative splicing

Chromatin structure and PolII density are predictive of alternative splicing events. It was recently proposed that chromatin structure plays a role in the regulation of alternative splicing, in addition to other mechanisms [137, 138]. To study this relationship, we used logistic regression to predict splicing outcomes of alternative exons based on the levels of various histone modifications, as well as PolII and nucleosome density in exonic regions. Our analysis showed that features associated with PolII density and chromatin structure of the exons are indeed predictive of exon inclusion or skipping, and that this result is not a consequence of experimental biases related to mappability or GC content of the alternative exons (Sections 3.3 and 3.4). Furthermore, although the levels of histone modifications are correlated to transcript expression levels, they also contain additional information about the splicing process, that cannot be explained by expression levels alone (Section 3.5).

Our analysis identified several variables with a significant influence on prediction of the alternative splicing outcome of the exon (Sections 3.3 and 3.5). First of all, in most cases, increased levels of PolII were associated with increased inclusion of exons. A

link between PolII and the splicing process has been observed before, and two different models were proposed to explain this phenomenon [120]. The first model postulates that PolII facilitates recruitment of splicing factors to transcripts, an effect which has previously been observed *in vivo* [150]. The second, kinetic, model is based on the finding that slowing down the procession of PolII along the transcript facilitates the recognition of splice sites by the splicing machinery [56]. If higher PolII density in fact corresponds to PolII enzymes stalled at intron-exon boundaries, this model could also provide an explanation for the positive association of PolII and exon inclusion observed in our analysis. We also determined that higher nucleosome density has a beneficial influence on splicing of alternative exons, a finding which corresponds to results of previous analyses of nucleosome enrichment of alternative and constitutive exons [9, 48, 59, 101, 153, 186, 200, 211]. A likely explanation is that nucleosomes pose an obstacle for the procession of PolII along the transcript. Since PolII would in this way be stalled at the exon, it could then either recruit splicing factors to the transcript more efficiently (recruitment model) or cause more efficient recognition of the splice site by the components of the splicing machinery (kinetic model).

Four histone modifications, H3K27me1, H3K36me3, H3K36ac and H3K79me2, were consistently identified as significant for prediction of alternative splicing, even after controlling for effects of transcript expression levels and nucleosome density (see Fig. 3.4 and Fig. 3.13). These four histone modifications mostly exhibit a positive association with exon inclusion. The exact function of these modifications in the regulation of alternative splicing is not yet known. However, with the exception of H3K36ac, all of them were identified by several independent studies as enriched in constitutive vs. alternative exons [9, 59, 117], suggesting a positive influence on exon inclusion and thus supporting the validity of our analysis. These histone modifications could influence the splicing process either by compacting the structure of chromatin and slowing down PolII elongation rate or by direct recruitment of splicing factors.

Regulation of alternative splicing by chromatin structure and splicing regulatory elements. Although the logistic model we developed is predictive of alternative splicing, there is still a set of exons where the alternative splicing outcome cannot be correctly predicted based only on the information conveyed by chromatin structure. More specifically, different sets of exons sometimes exhibit similar chromatin profiles, but have different splicing outcomes (see Fig. 3.6 and Fig. 3.14). One possible explanation is that in some exons chromatin structure does not have an influence on alternative splicing, but that the splicing process is rather governed by other factors. Alternatively, a particular histone modification could be connected to different splicing outcomes, depending on the context in which it appears.

Opposing effects of individual histone modifications on splicing have been observed before. For example, although most genome-wide studies of histone modification levels in exons found that increased levels of H3K36me3 are positively correlated with exon inclusion [9, 117] two recent independent studies showed that high levels H3K36me3 can actually repress the inclusion of alternatively spliced exons [139, 184]. Furthermore, studies investigating the relationship of alternative splicing and

H3K9me2, the only histone mark displaying higher levels in skipped when compared to included exons, also showed inconsistent results. A genome-wide study of H3K9me2 enrichment in exons showed it to be negatively correlated to exon inclusion [59], while an experimental study of the human fibronectin gene connected increased levels of H3K9me2 with enhanced inclusion of the alternative exon 33, also called EDI exon [6]. A possible explanation of these observations is that the same histone modification could have a different effect on alternative splicing of different groups of exons, possibly depending on the sequence characteristics of the exons themselves.

To test this hypothesis, we added features corresponding to known splicing regulatory elements to the set of predictor variables, and used them to train a logistic regression model and predict the outcome of alternative splicing. We observed that the inclusion of these features improves the prediction accuracy of our model of alternative splicing (Section 3.7). Furthermore, we identified features with the highest contribution to the prediction accuracy of the model. These features comprised of several histone modifications, PolII and a set of eight sequence regulatory elements (Fig. 3.19). This finding shows that the information contained within chromatin structure differs from the one encoded in the exon sequence, suggesting that alternative splicing is regulated by an intricate network of various cellular mechanisms. This suggests that histone modifications indeed have a role in the splicing process, and are not merely correlated to underlying sequence features of the exons, known to be involved in alternative splicing regulation. However, since the set of known sequence features used in this analysis was quite limited, we cannot fully exclude this possibility.

We studied in more detail the average regression coefficients of the eight significant sequence features in order to determine their possible influence on splicing of alternative exons (Fig. 3.19). Results of this analysis were mostly in accordance with existing knowledge about the effect of both splice site strength and different splicing regulatory elements on alternative splicing. As expected, the strength of the 3'SS and 5'SS both have a positive influence on exon inclusion. The number of RESCUE-ESE sequences and the U-content of the downstream region are also positively associated with exon inclusion, consistent with the findings that the appearance of these elements promotes the inclusion of alternative exons [12, 67]. The regression coefficient of the feature describing the appearance of binding sites for polypyrimidine tract binding protein (PTB) in the regions upstream of exons is negative, a result which is in agreement with the established function of PTB in inhibition of splicing [198]. According to our analysis, the appearance of PTB binding sites downstream of exons also has a negative influence on exon inclusion. These results are opposite to the ones obtained by study of alternative splicing in HeLa cells, which showed that the occurrence of PTB binding sites downstream of the exons in some cases promotes exon inclusion [136]. However, the same study showed that in the case of the occurrence of PTB binding sites both in the upstream and the downstream region, the upstream repressive function is dominant over downstream activating elements. The regions used in our analysis exhibit a positive correlation of the occurrence of upstream and downstream PTB binding sites (Pearson correlation coefficient $r = 0.21$, p-value $< 2.2 \cdot 10^{-16}$). This finding could explain the negative regression coefficient, if the

positive influence of a downstream activating element is antagonized by the influence of the upstream element in many exons. The last two significant variables connected to regulatory elements in the intron are binding sites for FOX1/2 proteins upstream and downstream of the exons. The negative coefficient of the variable which contains the number of FOX1/2 binding sites upstream of exons agrees with the finding that FOX proteins act as repressors of splicing when they bind to upstream regions [244]. However, FOX proteins can also act to enhance splicing, if bound downstream of the exon [244], but our analysis shows a negative association between the number of FOX1/2 binding sites in this region and exon inclusion. This could be explained by the fact that FOX1 and FOX2 proteins are preferentially expressed in brain, heart and muscle tissues [107, 217]. We therefore presume that the exons which are under the influence of these splicing factors will be spliced less efficiently in CD4+ T-cells.

We are aware of only one previous study which endeavored to predict genome-wide splicing patterns of transcripts [16]. In this study the authors used an information theory approach to develop a "splicing code" based on 1,014 different sequence features of 3,665 cassette-type alternative exons, which was then used to predict tissue-dependent changes in alternative splicing of exons in 27 different mouse tissues. The "splicing code" achieved a high accuracy in predicting alternative splicing patterns between pairs of tissues. The predicted direction of change was correct for 82.4% of 346 alternative exons evaluated by microarrays. The authors also used RT-PCR to confirm the predicted direction of change for 93.3% of 14 exons analyzed in 14 different tissues.

There are a few possible reasons why the accuracies of our models are lower than those obtained in the aforementioned study by Barash *et al.* [16]. First of all, we used a much smaller number of features (69 features for the most complex model incorporating information on chromatin structure and sequence features of exons) compared to the "splicing code" (1,014 features). However, the model incorporating only 69 features still managed to accurately predict the splicing patterns of 64.37% of analyzed alternative exons, confirming that histone modifications, along with sequence features, indeed harbor information on the outcome of the splicing process. In evaluating the results of the "splicing code" predictions, Barash *et al.* focused on cases where the predicted difference in splicing patterns was large, while in our analysis we made no such distinction. It is possible that small changes in expression levels of alternative exons are accompanied by small changes in chromatin structure, which might not necessarily be identifiable by a simple logistic regression model.

Furthermore, our model was trained to predict splicing patterns in only one tissue. We believe that the performance of the model could be improved by incorporating chromatin profiles across different tissues. Apart from the fact that inclusion of data from different tissues could reduce the effect of experimental noise on the analysis, it could also provide a better understanding of the relationship between histone modifications and different sequence features which regulate splicing in a tissue-dependent manner. We believe that the prediction of splicing patterns across tissues should be

restricted to transcripts which exhibit similar transcript expression levels, but different exon expression levels, in the respective tissues, in order to remove the effects of transcription on alternative splicing. However, such an approach would reduce the set of exons used to learn the parameters of the logistic regression model. Our model was trained only on the exons for which alternative splicing events have previously been annotated. While this decreases the chance of falsely labeling exons which do not undergo alternative splicing, it also reduces the size of the training set. We believe that a further reduction in the size of the training set, necessary for models trained on multiple tissues, would negatively influence the prediction accuracy of logistic regression.

We believe that our analysis contributes to a better understanding of the mechanisms regulating alternative splicing and that the models used here could be further improved in the future. Recently, several methods for direct determination of alternative splicing events from next-generation sequencing data have been developed [8, 173, 212, 223]. Employing such an approach would increase number of exons used to train the model, providing more information on the relationship of chromatin structure and splicing. Furthermore, it would be interesting to see if such relationships hold universally for different cell types, as in the case of the relationship between histone modifications and transcription (Section 2.4).

Influence of transcript expression levels on alternative splicing events. During the course of our analysis we observed that transcripts in which alternative exons are included in general have higher expression levels than the transcripts in which the exons are skipped (Section 3.6). This observation seems to be a reflection of a real biological phenomenon, since we found no evidence that it is caused either by measurement biases or different mechanisms of regulation of transcript stability. Although this finding confirms the possible functional coupling of transcription and splicing, the question of how this elevated expression level could influence splicing remains open.

The notion that transcription and splicing could be functionally coupled was first motivated by the discovery that most splicing events occur co-transcriptionally [28]. Since then, several different models of how transcription could influence splicing have been proposed. One such model, called the "kinetic model", postulates that the efficiency of splicing of competing alternative exons depends on the elongation rate of PolII [120]. Such a relationship could explain our finding that transcript expression levels seem to have an influence on the outcome of alternative splicing events.

Most of the studies which investigated the relationship of transcript elongation and alternative splicing showed that slower elongation rates favor the inclusion of alternative exons, while faster elongation rates favor the skipping of exons [22, 56, 205]. If we presume that the elongation rate is highly correlated with the expression level of a transcript, we would expect transcripts where the exons are skipped to exhibit higher expression levels, exactly the opposite of what we observe in our analysis. However, the kinetic model proposed that slower elongation rates would primarily enhance the inclusion of exons with weak splice sites [120]. If the slower elongation

favors the inclusion of only a subset of analyzed exons, this type of influence could not be identified in this analysis, since we made no distinction between exons based on splice site strength.

Another possible explanation is that although the elongation rate itself is potentially lower in the transcripts where the exons are included, favoring the inclusion of the exon, either PolII recruitment or transition from initiation to elongation could actually be more efficient in these transcripts. Studies of transcription kinetics showed that while many mammalian genes are associated with only one polymerase at a time, some of them can be transcribed by multiple PolII enzymes simultaneously [68, 116]. A higher number of PolII complexes associated with the transcript could result in higher expression levels, even if the elongation rate is slower, because the expression measurements correspond to the amount of mRNA accumulated in the cell. The number of PolII enzymes present at the genes could be influenced by the structure of the promoter of the gene in question. Since examples of genes where the promoter structure is connected to the alternative splicing outcome have been observed [53], this could potentially explain the discrepancies between our observations and results of previous studies.

We conclude that our results confirm the existence of functional coupling of transcription and splicing, proposed by previous studies ([120], and references therein). However, further studies, possibly focused on different subsets of alternative exons according to their sequence characteristics, will be needed to fully understand the mechanisms by which transcription influences the outcome of alternative splicing.

Summary

Histones are frequently decorated with covalent modifications. These histone modifications are thought to be involved in various chromatin-dependent processes including transcription and splicing. To elucidate the relationship between histone modifications and these two processes, we derived models to predict the expression level of genes and the structure of transcribed mRNA from histone modification levels.

We found that histone modification levels and gene expression are very well correlated. Moreover, we show that only a small number of histone modifications are necessary to accurately predict gene expression. We show that different sets of histone modifications are necessary to predict gene expression driven by high CpG content promoters (HCPs) or low CpG content promoters (LCPs). Quantitative models involving H3K4me3 and H3K79me1 are the most predictive of the expression levels in LCPs, whereas HCPs require H3K27ac and H4K20me1. We propose a preliminary "Histone Code of Transcription", where H3K4me3 is involved in RNA polymerase II (PolII) recruitment and/or initiation, the combinatorial action of H3K27ac and H4K20me1 leads to the transition to elongation, and finally H3K79me1 and H4K20me1 signal the transition to an elongating PolII. The preliminary "Histone Code of Transcription" awaits confirmation by further experimental studies. We furthermore show that the connections between histone modifications and gene expression seem to be general, as we were able to predict gene expression levels of one cell type using a model trained on another one. We propose that our model could be further improved by including information about different mechanisms of regulation of mRNA stability and degradation, giving rise to more accurate predictions of gene expression levels.

Using logistic models, we showed that levels of histone modifications, nucleosomes and PolII are predictive of the splicing outcomes of alternative exons, and that this result is not a consequence of experimental artifacts. Furthermore, we identified four histone modifications, namely H3K27me1, H3K36ac, H3K36me3 and H3K79me2, which consistently have a significant contribution to prediction accuracy of our models. This finding implies that they could be directly related to the splicing process, in agreement with recent analyses of the relationship of chromatin structure and the splicing process. We also established that histone modifications convey information about alternative splicing different from the one encoded in the DNA sequence of exons and surrounding regions, suggesting a possible interplay between these two mechanisms of splicing regulation. Finally, we confirmed the existence of functional coupling of transcription and splicing, by studying the dependence of structure of transcripts on their expression levels. The exact mechanisms behind these observations will have to be studied further.

Bibliography

[1] "`http://genome.ucsc.edu/`".

[2] "`http://nihroadmap.nih.gov/epigenomics/`".

[3] "`ftp://ftp.ensembl.org/pub/release-50/gtf/homo_sapiens`".

[4] M. W. Adkins, S. K. Williams, J. Linger, and J. K. Tyler. Chromatin disassembly from the PHO5 promoter is essential for the recruitment of the general transcription machinery and coactivators. *Mol Cell Biol*, 27(18):6372–6382, Sep 2007.

[5] I. Albert, T. N. Mavrich, L. P. Tomsho, J. Qi, S. J. Zanton, S. C. Schuster, and B. F. Pugh. Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature*, 446(7135):572–576, Mar 2007.

[6] M. All, V. Buggiano, J. P. Fededa, E. Petrillo, I. Schor, M. de la Mata, E. Agirre, M. Plass, E. Eyras, S. A. Elela, R. Klinck, B. Chabot, and A. R. Kornblihtt. Control of alternative splicing through siRNA-mediated transcriptional gene silencing. *Nat Struct Mol Biol*, 16(7):717–724, Jul 2009.

[7] V. G. Allfrey, R. Faulkner, and A. E. Mirsky. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc Natl Acad Sci U S A*, 51:786–94, 1964.

[8] A. Ameur, A. Wetterbom, L. Feuk, and U. Gyllensten. Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol*, 11(3):R34, 2010.

[9] R. Andersson, S. Enroth, A. Rada-Iglesias, C. Wadelius, and J. Komorowski. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res*, 19(10):1732–1741, Oct 2009.

[10] E. D. Andrulis, E. Guzmn, P. Dring, J. Werner, and J. T. Lis. High-resolution localization of Drosophila Spt5 and Spt6 at heat shock genes in vivo: roles in promoter proximal pausing and transcription elongation. *Genes Dev*, 14(20):2635–2649, Oct 2000.

[11] R. K. Auerbach, G. Euskirchen, J. Rozowsky, N. Lamarre-Vincent, Z. Moqtaderi, P. Lefranois, 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, Sep 2009.

[12] I. Aznarez, Y. Barash, O. Shai, D. He, J. Zielenski, L.-C. Tsui, J. Parkinson, B. J. Frey, J. M. Rommens, and B. J. Blencowe. A systematic analysis of intronic sequences downstream of 5' splice sites reveals a widespread role for U-rich motifs and TIA1/TIAL1 proteins in alternative splicing regulation. *Genome Res*, 18(8):1247–1258, Aug 2008.

[13] T. Bakheet, B. R. G. Williams, and K. S. A. Khabar. ARED 3.0: the large and diverse AU-rich transcriptome. *Nucleic Acids Res*, 34(Database issue):D111–D114, Jan 2006.

[14] A. J. Bannister and T. Kouzarides. Regulation of chromatin by histone modifications. *Cell Res*, 21(3):381–395, Mar 2011.

[15] G. Bar-Nahum, V. Epshtein, A. E. Ruckenstein, R. Rafikov, A. Mustaev, and E. Nudler. A ratchet mechanism of transcription elongation and its control. *Cell*, 120(2):183–193, Jan 2005.

[16] Y. Barash, J. A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B. J. Blencowe, and B. J. Frey. Deciphering the splicing code. *Nature*, 465(7294):53–59, May 2010.

[17] C. Barreau, L. Paillard, and H. B. Osborne. AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Res*, 33(22):7138–7150, 2005.

[18] A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, 2007.

[19] A. Barski, R. Jothi, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, and K. Zhao. Chromatin poises miRNA- and protein-coding genes for expression. *Genome Res*, 19(10):1742–51, 2009.

[20] D. P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, Jan 2004.

[21] D. P. Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233, Jan 2009.

[22] E. Batsch, M. Yaniv, and C. Muchardt. The human SWI/SNF subunit Brm is a regulator of alternative splicing. *Nat Struct Mol Biol*, 13(1):22–29, Jan 2006.

[23] G. Baurn and L. Wieslander. Splicing of Balbiani ring 1 gene pre-mRNA occurs simultaneously with transcription. *Cell*, 76(1):183–192, Jan 1994.

[24] S. L. Berger, T. Kouzarides, R. Shiekhattar, and A. Shilatifard. An operational definition of epigenetics. *Genes Dev*, 23(7):781–783, Apr 2009.

[25] S. M. Berget. Exon recognition in vertebrate splicing. *J Biol Chem*, 270(6):2411–2414, Feb 1995.

[26] B. E. Bernstein, A. Meissner, and E. S. Lander. The mammalian epigenome. *Cell*, 128(4):669–81, 2007.

[27] D. Betel, A. Koppal, P. Agius, C. Sander, and C. Leslie. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol*, 11(8):R90, 2010.

[28] A. L. Beyer and Y. N. Osheim. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev*, 2(6):754–765, Jun 1988.

[29] L. Breiman. Random Forests. *Mach Learn*, 45(1):5–32, 2001.

[30] J. Bring. A geometric approach to compare variables in a regression model. *The American Statistician*, 50(1):57–62, Feb 1996.

[31] Y. Brody, N. Neufeld, N. Bieberstein, S. Z. Causse, E.-M. Bhnlein, K. M. Neugebauer, X. Darzacq, and Y. Shav-Tal. The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing. *PLoS Biol*, 9(1):e1000573, 2011.

[32] S. Buratowski. Progression through the RNA polymerase II CTD cycle. *Mol Cell*, 36(4):541–546, Nov 2009.

[33] R. Cao and Y. Zhang. The functions of E(Z)/EZH2-mediated methylation of lysine 27 in histone H3. *Curr Opin Genet Dev*, 14(2):155–164, Apr 2004.

[34] D. Caput, B. Beutler, K. Hartog, R. Thayer, S. Brown-Shimer, and A. Cerami. Identification of a common nucleotide sequence in the 3'-untranslated region of mRNA molecules specifying inflammatory mediators. *Proc Natl Acad Sci U S A*, 83(6):1670–1674, Mar 1986.

[35] P. Carninci. Molecular biology: The long and short of RNAs. *Nature*, 457(7232):974–975, Feb 2009.

[36] M. J. Carrozza, B. Li, L. Florens, T. Suganuma, S. K. Swanson, K. K. Lee, W.-J. Shia, S. Anderson, J. Yates, M. P. Washburn, and J. L. Workman. Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell*, 123(4):581–592, Nov 2005.

[37] L. Cartegni, J. Wang, Z. Zhu, M. Q. Zhang, and A. R. Krainer. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res*, 31(13):3568–71, 2003.

[38] B. Chang, Y. Chen, Y. Zhao, and R. K. Bruick. JMJD6 is a histone arginine demethylase. *Science*, 318(5849):444–447, Oct 2007.

[39] Y.-F. Chang, J. S. Imam, and M. F. Wilkinson. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem*, 76:51–74, 2007.

[40] R. D. Chapman, M. Heidemann, T. K. Albert, R. Mailhammer, A. Flatley, M. Meisterernst, E. Kremmer, and D. Eick. Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. *Science*, 318(5857):1780–1782, Dec 2007.

[41] L. A. Chasin. Searching for splicing motifs. *Adv Exp Med Biol*, 623:85–106, 2007.

[42] C. Chen, A. Liaw, and L. Breiman. Using random forest to learn unbalanced data. Technical Report 666, Department of Statistics, University of California at Berkeley, 2004.

[43] C. Y. Chen and A. B. Shyu. AU-rich elements: characterization and importance in mRNA degradation. *Trends Biochem Sci*, 20(11):465–470, Nov 1995.

[44] 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, Nov 2009.

[45] I. Chepelev, G. Wei, Q. Tang, and K. Zhao. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res*, 37(16):e106, 2009.

[46] K. Chiba, J. Yamamoto, Y. Yamaguchi, and H. Handa. Promoter-proximal pausing and its release: molecular mechanisms and physiological functions. *Exp Cell Res*, 316(17):2723–2730, Oct 2010.

[47] S. Cho, A. Hoang, R. Sinha, X.-Y. Zhong, X.-D. Fu, A. R. Krainer, and G. Ghosh. Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly. *Proc Natl Acad Sci U S A*, 108(20):8233–8238, May 2011.

[48] R. K. Chodavarapu, S. Feng, Y. V. Bernatavichute, P.-Y. Chen, H. Stroud, Y. Yu, J. A. Hetzel, F. Kuo, J. Kim, S. J. Cokus, D. Casero, M. Bernal, P. Huijser, A. T. Clark, U. Krmer, S. S. Merchant, X. Zhang, S. E. Jacobsen, and M. Pellegrini. Relationship between nucleosome positioning and DNA methylation. *Nature*, 466(7304):388–392, Jul 2010.

[49] C. R. Clapier and B. R. Cairns. The biology of chromatin remodeling complexes. *Annu Rev Biochem*, 78:273–304, 2009.

[50] T. A. Cooper, L. Wan, and G. Dreyfuss. RNA and disease. *Cell*, 136(4):777–793, Feb 2009.

[51] L. J. Core, J. J. Waterfall, and J. T. Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–1848, Dec 2008.

[52] P. Cramer, J. F. Cceres, D. Cazalla, S. Kadener, A. F. Muro, F. E. Baralle, and A. R. Kornblihtt. Coupling of transcription with alternative splicing: RNA pol II promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer. *Mol Cell*, 4(2):251–258, Aug 1999.

[53] P. Cramer, C. G. Pesce, F. E. Baralle, and A. R. Kornblihtt. Functional association between promoter structure and transcript alternative splicing. *Proc Natl Acad Sci U S A*, 94(21):11456–11460, Oct 1997.

[54] K. Cui, C. Zang, T. Y. Roh, D. E. Schones, R. W. Childs, W. Peng, and K. Zhao. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell*, 4(1):80–93, 2009.

[55] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 A resolution. *J Mol Biol*, 319(5):1097–113, 2002.

[56] M. de la Mata, C. R. Alonso, S. Kadener, J. P. Fededa, M. Blaustein, F. Pelisch, P. Cramer, D. Bentley, and A. R. Kornblihtt. A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell*, 12(2):525–532, Aug 2003.

[57] M. de Napoles, J. E. Mermoud, R. Wakao, Y. A. Tang, M. Endoh, R. Appanah, T. B. Nesterova, J. Silva, A. P. Otte, M. Vidal, H. Koseki, and N. Brockdorff. Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation. *Dev Cell*, 7(5):663–676, Nov 2004.

[58] M. D. E. Deato and R. Tjian. Switching of the core transcription machinery during myogenesis. *Genes Dev*, 21(17):2137–2149, Sep 2007.

[59] P. Dhami, P. Saffrey, A. W. Bruce, S. C. Dillon, K. Chiang, N. Bonhoure, C. M. Koch, J. Bye, K. James, N. S. Foad, P. Ellis, N. A. Watkins, W. H. Ouwehand, C. Langford, R. M. Andrews, I. Dunham, and D. Vetrie. Complex exon-intron marking by histone modifications is not determined solely by nucleosome distribution. *PLoS One*, 5(8):e12339, 2010.

[60] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*, 36(16):e105, 2008.

[61] N. Draper and H. Smith. *Applied regression analysis (3rd ed.)*. John Wiley & Sons, 1998.

[62] G. Easow, A. A. Teleman, and S. M. Cohen. Isolation of microRNA targets by miRNP immunopurification. *RNA*, 13(8):1198–1204, Aug 2007.

[63] S. Egloff and S. Murphy. Cracking the RNA polymerase II CTD code. *Trends Genet*, 24(6):280–8, 2008.

[64] S. Egloff, D. O'Reilly, R. D. Chapman, A. Taylor, K. Tanzhaus, L. Pitts, D. Eick, and S. Murphy. Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. *Science*, 318(5857):1777–1779, Dec 2007.

[65] R. Ekins and F. W. Chu. Microarrays: their origins and applications. *Trends Biotechnol*, 17(6):217–218, Jun 1999.

[66] N. Elango and S. V. Yi. DNA methylation and structural and functional bimodality of vertebrate promoters. *Mol Biol Evol*, 25(8):1602–1608, Aug 2008.

[67] W. G. Fairbrother, G. W. Yeo, R. Yeh, P. Goldstein, M. Mawson, P. A. Sharp, and C. B. Burge. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res*, 32(Web Server issue):W187–W190, Jul 2004.

[68] A. M. Femino, F. S. Fay, K. Fogarty, and R. H. Singer. Visualization of single RNA transcripts in situ. *Science*, 280(5363):585–590, Apr 1998.

[69] 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, Feb 2008.

[70] W. Fischle, B. S. Tseng, H. L. Dormann, B. M. Ueberheide, B. A. Garcia, J. Shabanowitz, D. F. Hunt, H. Funabiki, and C. D. Allis. Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. *Nature*, 438(7071):1116–1122, Dec 2005.

[71] P. C. FitzGerald, D. Sturgill, A. Shyakhtenko, B. Oliver, and C. Vinson. Comparative genomics of Drosophila and human core promoters. *Genome Biol*, 7(7):R53, 2006.

[72] P. Foerch, O. Puig, N. Kedersha, C. Martinez, S. Granneman, B. Seraphin, P. Anderson, and J. Valcarcel. The apoptosis-promoting factor TIA-1 is a regulator of alternative pre-mRNA splicing. *Mol Cell*, 6(5):1089–1098, Nov 2000.

[73] P. Foerch, O. Puig, C. Martinez, B. Seraphin, and J. Valcarcel. The splicing regulator TIA-1 interacts with U1-C to promote U1 snRNP recruitment to 5' splice sites. *EMBO J*, 21(24):6882–6892, Dec 2002.

[74] S. Foissac and M. Sammeth. ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res*, 35(Web Server issue):W297–9, 2007.

[75] P. A. Frischmeyer and H. C. Dietz. Nonsense-mediated mrna decay in health and disease. *Hum Mol Genet*, 8(10):1893–1900, 1999.

[76] S. M. Fuchs, R. N. Laribee, and B. D. Strahl. Protein modifications in transcription elongation. *Biochim Biophys Acta*, 1789(1):26–36, 2009.

[77] K. Gao, A. Masuda, T. Matsuura, and K. Ohno. Human branch point consensus sequence is yUnAy. *Nucleic Acids Res*, 36(7):2257–2267, Apr 2008.

[78] M. Gardiner-Garden and M. Frommer. CpG islands in vertebrate genomes. *J Mol Biol*, 196(2):261–282, Jul 1987.

[79] F. D. Gatto-Konczak, C. F. Bourgeois, C. L. Guiner, L. Kister, M. C. Gesnel, J. Stvenin, and R. Breathnach. The RNA-binding protein TIA-1 is a novel mammalian splicing regulator acting through intron sequences adjacent to a 5' splice site. *Mol Cell Biol*, 20(17):6287–6299, Sep 2000.

[80] N. I. Gershenzon and I. P. Ioshikhes. Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics*, 21(8):1295–1300, Apr 2005.

[81] B. R. Graveley. Sorting out the complexity of SR protein functions. *RNA*, 6(9):1197–1211, Sep 2000.

[82] B. R. Graveley, K. J. Hertel, and T. Maniatis. The role of U2AF35 and U2AF65 in enhancer-dependent splicing. *RNA*, 7(6):806–818, Jun 2001.

[83] A. Grimson, K. K.-H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, 27(1):91–105, Jul 2007.

[84] S. M. Grisch, M. Wachsmuth, K. F. Tth, P. Lichter, and K. Rippe. Histone acetylation increases chromatin accessibility. *J Cell Sci*, 118(Pt 24):5825–5834, Dec 2005.

[85] A. R. Grosso, A. Q. Gomes, N. L. Barbosa-Morais, S. Caldeira, N. P. Thorne, G. Grech, M. von Lindern, and M. Carmo-Fonseca. Tissue-specific splicing factor gene expression signatures. *Nucleic Acids Res*, 36(15):4823–4832, Sep 2008.

[86] E. Guccione, F. Martinato, G. Finocchiaro, L. Luzi, L. Tizzoni, V. D. Olio, G. Zardo, C. Nervi, L. Bernard, and B. Amati. Myc-binding-site recognition in the human genome is determined by chromatin context. *Nat Cell Biol*, 8(7):764–770, Jul 2006.

[87] M. G. Guenther, S. S. Levine, L. A. Boyer, R. Jaenisch, and R. A. Young. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1):77–88, 2007.

[88] J. Guhaniyogi and G. Brewer. Regulation of mRNA stability in mammalian cells. *Gene*, 265(1-2):11–23, Mar 2001.

[89] E. Guzman and J. T. Lis. Transcription factor TFIIH is required for promoter melting in vivo. *Mol Cell Biol*, 19(8):5652–5658, Aug 1999.

[90] 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. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, Apr 2010.

[91] T. Hastie, R. Tibshirani, and T. Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer-Verlag, New York, New York, 2001.

[92] R. D. Hawkins, G. C. Hon, and B. Ren. Next-generation genomics: an integrative approach. *Nat Rev Genet*, 11(7):476–486, Jul 2010.

[93] N. Hernandez. TBP, a universal eukaryotic transcription factor? *Genes Dev*, 7(7B):1291–1308, Jul 1993.

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

[95] K. Hirayoshi and J. T. Lis. Nuclear run-on assays: assessing transcription by measuring density of engaged RNA polymerases. *Methods Enzymol*, 304:351–362, 1999.

[96] M. Hirst and M. A. Marra. Next generation sequencing based approaches to epigenomics. *Brief Funct Genomics*, 9(5-6):455–465, Dec 2010.

[97] J. Hnilicova, S. Hozeifi, E. Duskova, J. Icha, T. Tomankova, and D. Stanek. Histone deacetylase activity modulates alternative splicing. *PLoS One*, 6(2):e16727, 2011.

[98] C. Hodges, L. Bintu, L. Lubkowska, M. Kashlev, and C. Bustamante. Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. *Science*, 325(5940):626–628, Jul 2009.

[99] J. D. Hoheisel. Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet*, 7(3):200–210, Mar 2006.

[100] G. Hon, W. Wang, and B. Ren. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput Biol*, 5(11):e1000566, Nov 2009.

[101] J. T. Huff, A. M. Plocik, C. Guthrie, and K. R. Yamamoto. Reciprocal intronic and exonic histone modification regions in humans. *Nat Struct Mol Biol*, 17(12):1495–1499, Dec 2010.

[102] J. Hui, L.-H. Hung, M. Heiner, S. Schreiner, N. Neumller, G. Reither, S. A. Haas, and A. Bindereif. Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. *EMBO J*, 24(11):1988–1998, Jun 2005.

[103] Y. ichi Tsukada, J. Fang, H. Erdjument-Bromage, M. E. Warren, C. H. Borchers, P. Tempst, and Y. Zhang. Histone demethylation by a family of JmjC domain-containing proteins. *Nature*, 439(7078):811–816, Feb 2006.

[104] M. G. Izban and D. S. Luse. Transcription on nucleosomal templates by RNA polymerase II in vitro: inhibition of elongation with enhancement of sequence-specific pausing. *Genes Dev*, 5(4):683–696, Apr 1991.

[105] K. B. Jensen, B. K. Dredge, G. Stefani, R. Zhong, R. J. Buckanovich, H. J. Okano, Y. Y. Yang, and R. B. Darnell. Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron*, 25(2):359–371, Feb 2000.

[106] T. Jenuwein and C. D. Allis. Translating the histone code. *Science*, 293(5532):1074–80, 2001.

[107] Y. Jin, H. Suzuki, S. Maegawa, H. Endo, S. Sugano, K. Hashimoto, K. Yasuda, and K. Inoue. A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. *EMBO J*, 22(4):905–912, Feb 2003.

[108] A. A. Joshi and K. Struhl. Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to Pol II elongation. *Mol Cell*, 20(6):971–978, Dec 2005.

[109] T. Juven-Gershon, J.-Y. Hsu, J. W. Theisen, and J. T. Kadonaga. The RNA polymerase II core promoter - the gateway to transcription. *Curr Opin Cell Biol*, 20(3):253–259, Jun 2008.

[110] J. T. Kadonaga. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell*, 116(2):247–257, Jan 2004.

[111] 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, Feb 2010.

[112] M.-C. Keogh, S. K. Kurdistani, S. A. Morris, S. H. Ahn, V. Podolny, S. R. Collins, M. Schuldiner, K. Chin, T. Punna, N. J. Thompson, C. Boone, A. Emili, J. S. Weissman, T. R. Hughes, B. D. Strahl, M. Grunstein, J. F. Greenblatt, S. Buratowski, and N. J. Krogan. Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. *Cell*, 123(4):593–605, Nov 2005.

[113] K. S. A. Khabar. The AU-rich transcriptome: more than interferons and cytokines, and its role in disease. *J Interferon Cytokine Res*, 25(1):1–10, Jan 2005.

[114] J. Kim, M. Guermah, R. K. McGinty, J.-S. Lee, Z. Tang, T. A. Milne, A. Shilatifard, T. W. Muir, and R. G. Roeder. RAD6-Mediated transcription-coupled

H2B ubiquitylation directly stimulates H3K4 methylation in human cells. *Cell*, 137(3):459–471, May 2009.

[115] T. H. Kim, L. O. Barrera, M. Zheng, C. Qu, M. A. Singer, T. A. Richmond, Y. Wu, R. D. Green, and B. Ren. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–880, Aug 2005.

[116] H. Kimura, K. Sugaya, and P. R. Cook. The transcription cycle of RNA polymerase II in living cells. *J Cell Biol*, 159(5):777–782, Dec 2002.

[117] P. Kolasinska-Zwierz, T. Down, I. Latorre, T. Liu, X. S. Liu, and J. Ahringer. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet*, 41(3):376–381, Mar 2009.

[118] R. D. Kornberg. Chromatin structure: a repeating unit of histones and DNA. *Science*, 184(139):868–71, 1974.

[119] R. D. Kornberg and J. O. Thomas. Chromatin structure; oligomers of the histones. *Science*, 184(139):865–8, 1974.

[120] A. R. Kornblihtt. Chromatin, transcript elongation and alternative splicing. *Nat Struct Mol Biol*, 13(1):5–7, Jan 2006.

[121] T. Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.

[122] M. Kubat, R. C. Holte, and S. Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30(2-3):195–215, 1998.

[123] J. N. Kuehner, E. L. Pearson, and C. Moore. Unravelling the means to an end: RNA polymerase II transcription termination. *Nat Rev Mol Cell Biol*, 12(5):283–294, May 2011.

[124] M. Lagos-Quintana, R. Rauhut, A. Yalcin, J. Meyer, W. Lendeckel, and T. Tuschl. Identification of tissue-specific microRNAs from mouse. *Curr Biol*, 12(9):735–739, Apr 2002.

[125] J. S. Lee, A. Shukla, J. Schneider, S. K. Swanson, M. P. Washburn, L. Florens, S. R. Bhaumik, and A. Shilatifard. Histone crosstalk between H2B monoubiquitination and H3 methylation mediated by COMPASS. *Cell*, 131(6):1084–96, 2007.

[126] T. I. Lee, H. C. Causton, F. C. Holstege, W. C. Shen, N. Hannett, E. G. Jennings, F. Winston, M. R. Green, and R. A. Young. Redundant roles for the TFIID and SAGA complexes in global transcription. *Nature*, 405(6787):701–704, Jun 2000.

[127] W. Lee, D. Tillo, N. Bray, R. H. Morse, R. W. Davis, T. R. Hughes, and C. Nislow. A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet*, 39(10):1235–1244, Oct 2007.

[128] Y. Lee, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Rdmark, S. Kim, and V. N. Kim. The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425(6956):415–419, Sep 2003.

[129] Y. Lee, K. Jeon, J.-T. Lee, S. Kim, and V. N. Kim. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J*, 21(17):4663–4670, Sep 2002.

[130] 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, Jan 2005.

[131] B. Li, M. Carey, and J. L. Workman. The role of chromatin during transcription. *Cell*, 128(4):707–719, Feb 2007.

[132] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98(1):31–36, 2001.

[133] H. Li, S. Ilin, W. Wang, E. M. Duncan, J. Wysocka, C. D. Allis, and D. J. Patel. Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature*, 442(7098):91–5, 2006.

[134] S. Li and M. F. Wilkinson. Nonsense surveillance in lymphocytes? *Immunity*, 8(2):135–141, Feb 1998.

[135] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, Nov 2009.

[136] M. Llorian, S. Schwartz, T. A. Clark, D. Hollander, L.-Y. Tan, R. Spellman, A. Gordon, A. C. Schweitzer, P. de la Grange, G. Ast, and C. W. J. Smith. Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. *Nat Struct Mol Biol*, 17(9):1114–1123, Sep 2010.

[137] R. F. Luco, M. Allo, I. E. Schor, A. R. Kornblihtt, and T. Misteli. Epigenetics in alternative pre-mRNA splicing. *Cell*, 144(1):16–26, Jan 2011.

[138] R. F. Luco and T. Misteli. More than a splicing code: integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation. *Curr Opin Genet Dev*, 21(4):366–372, Aug 2011.

[139] R. F. Luco, Q. Pan, K. Tominaga, B. J. Blencowe, O. M. Pereira-Smith, and T. Misteli. Regulation of alternative splicing by histone modifications. *Science*, 327(5968):996–1000, Feb 2010.

[140] K. Luger, A. W. Mader, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, 389(6648):251–60, 1997.

[141] W. Luo, A. W. Johnson, and D. L. Bentley. The role of Rat1 in coupling mRNA 3'-end processing to transcription termination: implications for a unified allosteric-torpedo model. *Genes Dev*, 20(8):954–965, Apr 2006.

[142] A. M. MacMillan, P. S. McCaw, J. D. Crispino, and P. A. Sharp. SC35-mediated reconstitution of splicing in U2AF-depleted nuclear extract. *Proc Natl Acad Sci U S A*, 94(1):133–136, Jan 1997.

[143] C. K. Mapendano, S. Lykke-Andersen, J. Kjems, E. Bertrand, and T. H. Jensen. Crosstalk between mRNA 3' end processing and transcription initiation. *Mol Cell*, 40(3):410–422, Nov 2010.

[144] V. Markovtsov, J. M. Nikolic, J. A. Goldman, C. W. Turck, M. Y. Chou, and D. L. Black. Cooperative assembly of an hnRNP complex induced by a tissue-specific homolog of polypyrimidine tract binding protein. *Mol Cell Biol*, 20(20):7463–7479, Oct 2000.

[145] A. J. McCullough and S. M. Berget. G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol Cell Biol*, 17(8):4562–4571, Aug 1997.

[146] N. J. McGlincy and C. W. J. Smith. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends Biochem Sci*, 33(8):385–393, Aug 2008.

[147] M. L. Metzker. Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46, Jan 2010.

[148] 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.[see comment]. *Nature*, 448(7153):553–60, 2007.

[149] S. Minovitsky, S. L. Gee, S. Schokrpur, I. Dubchak, and J. G. Conboy. The splicing regulatory element, UGCAUG, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons. *Nucleic Acids Res*, 33(2):714–724, 2005.

[150] T. Misteli and D. L. Spector. RNA polymerase II targets pre-mRNA splicing factors to transcription sites in vivo. *Mol Cell*, 3(6):697–705, Jun 1999.

[151] G. W. Muse, D. A. Gilchrist, S. Nechaev, R. Shah, J. S. Parker, S. F. Grissom, J. Zeitlinger, and K. Adelman. RNA polymerase is poised for activation across the genome. *Nat Genet*, 39(12):1507–11, 2007.

[152] H. Nagasaki, M. Arita, T. Nishizawa, M. Suwa, and O. Gotoh. Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. *Gene*, 364:53–62, Dec 2005.

[153] S. Nahkuri, R. J. Taft, and J. S. Mattick. Nucleosomes are preferentially positioned at exons in somatic and sperm cells. *Cell Cycle*, 8(20):3420–3424, Oct 2009.

[154] S. S. Ng, W. W. Yue, U. Oppermann, and R. J. Klose. Dynamic protein methylation in chromatin biology. *Cell Mol Life Sci*, 66(3):407–422, Feb 2009.

[155] G. Nogues, S. Kadener, P. Cramer, D. Bentley, and A. R. Kornblihtt. Transcriptional activators differ in their abilities to control alternative splicing. *J Biol Chem*, 277(45):43110–43114, Nov 2002.

[156] E. Nudler. RNA polymerase active center: the molecular engine of transcription. *Annu Rev Biochem*, 78:335–361, 2009.

[157] F. C. Oberstrass, S. D. Auweter, M. Erat, Y. Hargous, A. Henning, P. Wenter, L. Reymond, B. Amir-Ahmady, S. Pitsch, D. L. Black, and F. H.-T. Allain. Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science*, 309(5743):2054–2057, Sep 2005.

[158] U. Ohler. Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. *Nucleic Acids Res*, 34(20):5943–5950, 2006.

[159] M. Oki, H. Aihara, and T. Ito. Role of histone phosphorylation in chromatin dynamics and its implications in diseases. *Subcell Biochem*, 41:319–336, 2007.

[160] P. J. Park. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–680, Oct 2009.

[161] Y.-J. Park and K. Luger. Structure and function of nucleosome assembly proteins. *Biochem Cell Biol*, 84(4):549–558, Aug 2006.

[162] T. A. Patterson, E. K. Lobenhofer, S. B. Fulmer-Smentek, P. J. Collins, T.-M. Chu, W. Bao, H. Fang, E. S. Kawasaki, J. Hager, I. R. Tikhonova, S. J. Walker, L. Zhang, P. Hurban, F. de Longueville, J. C. Fuscoe, W. Tong, L. Shi, and R. D. Wolfinger. Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nat Biotechnol*, 24(9):1140–1150, Sep 2006.

[163] R. Pavri, B. Zhu, G. Li, P. Trojer, S. Mandal, A. Shilatifard, and D. Reinberg. Histone H2B monoubiquitination functions cooperatively with FACT to regulate elongation by RNA polymerase II. *Cell*, 125(4):703–717, May 2006.

[164] M. Pertea, X. Lin, and S. L. Salzberg. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res*, 29(5):1185–1190, Mar 2001.

[165] L. Piacentini, L. Fanti, R. Negri, V. D. Vescovo, A. Fatica, F. Altieri, and S. Pimpinelli. Heterochromatin protein 1 (HP1a) positively regulates euchromatic gene expression through RNA transcript association and interaction with hnRNPs in Drosophila. *PLoS Genet*, 5(10):e1000670, Oct 2009.

[166] R. U. Protacio, G. Li, P. T. Lowary, and J. Widom. Effects of histone tail domains on the rate of transcriptional elongation through a nucleosome. *Mol Cell Biol*, 20(23):8866–8878, Dec 2000.

[167] N. J. Proudfoot. How RNA polymerase II terminates transcription in higher eukaryotes. *Trends Biochem Sci*, 14(3):105–110, Mar 1989.

[168] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[169] E. B. Rasmussen and J. T. Lis. In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proc Natl Acad Sci U S A*, 90(17):7923–7, 1993.

[170] M. Rassoulzadegan, V. Grandjean, P. Gounon, and F. Cuzin. Inheritance of an epigenetic change in the mouse: a new role for RNA. *Biochem Soc Trans*, 35(Pt 3):623–625, Jun 2007.

[171] D. Reines, M. J. Chamberlin, and C. M. Kane. Transcription elongation factor SII (TFIIS) enables RNA polymerase II to elongate through a block to transcription in a human gene in vitro. *J Biol Chem*, 264(18):10799–10809, Jun 1989.

[172] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–2309, Dec 2000.

[173] H. Richard, M. H. Schulz, M. Sultan, A. Nrnberger, S. Schrinner, D. Balzereit, E. Dagand, A. Rasche, H. Lehrach, M. Vingron, S. A. Haas, and M.-L. Yaspo. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Res*, 38(10):e112, Jun 2010.

[174] P. Richard and J. L. Manley. Transcription termination by nuclear RNA polymerases. *Genes Dev*, 23(11):1247–1269, Jun 2009.

[175] J. Rino and M. Carmo-Fonseca. The spliceosome: a self-organized macromolecular machine in the nucleus? *Trends Cell Biol*, 19(8):375–384, Aug 2009.

[176] A. Roguev, D. Schaft, A. Shevchenko, W. W. Pijnappel, M. Wilm, R. Aasland, and A. F. Stewart. The Saccharomyces cerevisiae Set1 complex includes an Ash2 homologue and methylates histone 3 lysine 4. *Embo J*, 20(24):7137–48, 2001.

[177] B. Ruskin, P. D. Zamore, and M. R. Green. A factor, U2AF, is required for U2 snRNP binding and splicing complex assembly. *Cell*, 52(2):207–219, Jan 1988.

[178] M. Sammeth, S. Foissac, and R. Guig. A general definition and nomenclature for alternative splicing events. *PLoS Comput Biol*, 4(8):e1000147, 2008.

[179] A. Saunders, L. J. Core, and J. T. Lis. Breaking barriers to transcription elongation. *Nat Rev Mol Cell Biol*, 7(8):557–67, 2006.

[180] S. Saxonov, P. Berg, and D. L. Brutlag. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A*, 103(5):1412–7, 2006.

[181] M. Schena, R. A. Heller, T. P. Theriault, K. Konrad, E. Lachenmeier, and R. W. Davis. Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol*, 16(7):301–306, Jul 1998.

[182] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, Oct 1995.

[183] 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):887–98, 2008.

[184] I. E. Schor, N. Rascovan, F. Pelisch, M. All, and A. R. Kornblihtt. Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. *Proc Natl Acad Sci U S A*, 106(11):4325–4330, Mar 2009.

[185] B. Schuettengruber, D. Chourrout, M. Vervoort, B. Leblanc, and G. Cavalli. Genome regulation by polycomb and trithorax proteins. *Cell*, 128(4):735–45, 2007.

[186] S. Schwartz, E. Meshorer, and G. Ast. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol*, 16(9):990–995, Sep 2009.

[187] S. Schwartz, R. Oren, and G. Ast. Detection and removal of biases in the analysis of next-generation sequencing reads. *PLoS One*, 6(1):e16685, 2011.

[188] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):pp. 461–464, 1978.

[189] A. C. Seila, J. M. Calabrese, S. S. Levine, G. W. Yeo, P. B. Rahl, R. A. Flynn, R. A. Young, and P. A. Sharp. Divergent transcription from active promoters. *Science*, 2008.

[190] J. Shendure and H. Ji. Next-generation DNA sequencing. *Nat Biotechnol*, 26(10):1135–1145, Oct 2008.

[191] Y. Shi, F. Lan, C. Matson, P. Mulligan, J. R. Whetstine, P. A. Cole, R. A. Casero, and Y. Shi. Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell*, 119(7):941–953, Dec 2004.

[192] A. Shilatifard, R. C. Conaway, and J. W. Conaway. The RNA polymerase II elongation complex. *Annu Rev Biochem*, 72:693–715, 2003.

[193] A. Shukla, P. Bajwa, and S. R. Bhaumik. SAGA-associated Sgf73p facilitates formation of the preinitiation complex assembly at the promoters either in a HAT-dependent or independent manner in vivo. *Nucleic Acids Res*, 34(21):6225–6232, 2006.

[194] S. Sigurdsson, A. B. Dirac-Svejstrup, and J. Q. Svejstrup. Evidence that transcript cleavage is essential for RNA polymerase II transcription and cell viability. *Mol Cell*, 38(2):202–210, Apr 2010.

[195] T. W. Sikorski and S. Buratowski. The basal initiation machinery: beyond the general transcription factors. *Curr Opin Cell Biol*, 21(3):344–351, Jun 2009.

[196] r. Sims, R. J., S. Millhouse, C. F. Chen, B. A. Lewis, H. Erdjument-Bromage, P. Tempst, J. L. Manley, and D. Reinberg. Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. *Mol Cell*, 28(4):665–76, 2007.

[197] R. J. Sims, R. Belotserkovskaya, and D. Reinberg. Elongation by RNA polymerase II: the short and long of it. *Genes Dev*, 18(20):2437–2468, Oct 2004.

[198] R. Singh, J. Valcrcel, and M. R. Green. Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. *Science*, 268(5214):1173–1176, May 1995.

[199] M. J. Solomon, P. L. Larsen, and A. Varshavsky. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell*, 53(6):937–947, Jun 1988.

[200] N. Spies, C. B. Nielsen, R. A. Padgett, and C. B. Burge. Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell*, 36(2):245–254, Oct 2009.

[201] D. J. Steger, M. I. Lefterova, L. Ying, A. J. Stonestrom, M. Schupp, D. Zhuo, A. L. Vakoc, J. E. Kim, J. Chen, M. A. Lazar, G. A. Blobel, and C. R. Vakoc. DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells. *Mol Cell Biol*, 28(8):2825–39, 2008.

[202] J. K. Stock, S. Giadrossi, M. Casanova, E. Brookes, M. Vidal, H. Koseki, N. Brockdorff, A. G. Fisher, and A. Pombo. Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. *Nat Cell Biol*, 9(12):1428–1435, Dec 2007.

[203] B. D. Strahl and C. D. Allis. The language of covalent histone modifications. *Nature*, 403(6765):41–5, 2000.

[204] H. Sun and L. A. Chasin. Multiple splicing defects in an intronic false exon. *Mol Cell Biol*, 20(17):6414–6425, Sep 2000.

[205] J. Sun, A. L. Blair, S. E. Aiyar, and R. Li. Cofactor of BRCA1 modulates androgen-dependent transcription and alternative splicing. *J Steroid Biochem Mol Biol*, 107(3-5):131–139, 2007.

[206] J. Q. Svejstrup. The RNA polymerase II transcription cycle: cycling through chromatin. *Biochim Biophys Acta*, 1677(1-3):64–73, 2004.

[207] D. Takai and P. A. Jones. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A*, 99(6):3740–3745, Mar 2002.

[208] C. N. Tennyson, H. J. Klamut, and R. G. Worton. The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nat Genet*, 9(2):184–190, Feb 1995.

[209] L. Teytelman, B. Ozaydin, O. Zill, P. Lefranois, M. Snyder, J. Rine, and M. B. Eisen. Impact of chromatin structures on DNA processing for genomic analyses. *PLoS One*, 4(8):e6700, 2009.

[210] J. O. Thomas and R. D. Kornberg. An octamer of histones in chromatin and free in solution. *Proc Natl Acad Sci U S A*, 72(7):2626–30, 1975.

[211] H. Tilgner, C. Nikolaou, S. Althammer, M. Sammeth, M. Beato, J. Valcrcel, and R. Guig. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol*, 16(9):996–1001, Sep 2009.

[212] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, May 2009.

[213] B. M. Turner. Reading signals on the nucleosome with a new nomenclature for modified histones. *Nat Struct Mol Biol*, 12(2):110–112, Feb 2005.

[214] J. K. Tyler. Chromatin assembly. Cooperation between histone chaperones and ATP-dependent nucleosome remodeling machines. *Eur J Biochem*, 269(9):2268–2274, May 2002.

[215] J. Ule, K. B. Jensen, M. Ruggiu, A. Mele, A. Ule, and R. B. Darnell. CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 302(5648):1212–1215, Nov 2003.

[216] J. Ule, G. Stefani, A. Mele, M. Ruggiu, X. Wang, B. Taneri, T. Gaasterland, B. J. Blencowe, and R. B. Darnell. An RNA map predicting Nova-dependent splicing regulation. *Nature*, 444(7119):580–586, Nov 2006.

[217] J. G. Underwood, P. L. Boutz, J. D. Dougherty, P. Stoilov, and D. L. Black. Homologues of the Caenorhabditis elegans Fox-1 protein are neuronal splicing regulators in mammals. *Mol Cell Biol*, 25(22):10005–10016, Nov 2005.

[218] A. Verdel, S. Jia, S. Gerber, T. Sugiyama, S. Gygi, S. I. S. Grewal, and D. Moazed. RNAi-mediated targeting of heterochromatin by the RITS complex. *Science*, 303(5658):672–676, Jan 2004.

[219] T. Wada, T. Takagi, Y. Yamaguchi, A. Ferdous, T. Imai, S. Hirose, S. Sugimoto, K. Yano, G. A. Hartzog, F. Winston, S. Buratowski, and H. Handa. DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes Dev*, 12(3):343–356, Feb 1998.

[220] M. C. Wahl, C. L. Will, and R. Lhrmann. The spliceosome: design principles of a dynamic RNP machine. *Cell*, 136(4):701–718, Feb 2009.

[221] 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, Nov 2008.

[222] 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, Oct 2004.

[223] L. Wang, Y. Xi, J. Yu, L. Dong, L. Yen, and W. Li. A statistical method for the detection of alternative splicing using RNA-seq. *PLoS One*, 5(1):e8529, 2010.

[224] Z. Wang and C. B. Burge. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, 14(5):802–813, May 2008.

[225] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009.

[226] Z. Wang, M. E. Rolish, G. Yeo, V. Tung, M. Mawson, and C. B. Burge. Systematic identification and analysis of exonic splicing silencers. *Cell*, 119(6):831–845, Dec 2004.

[227] Z. Wang, C. Zang, K. Cui, D. E. Schones, A. Barski, W. Peng, and K. Zhao. Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell*, 138(5):1019–31, 2009.

[228] 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.

[229] A. Watson, A. Mazumder, M. Stewart, and S. Balasubramanian. Technology for microarray analysis of gene expression. *Curr Opin Biotechnol*, 9(6):609–614, Dec 1998.

[230] Y. Wei, C. A. Mizzen, R. G. Cook, M. A. Gorovsky, and C. D. Allis. Phosphorylation of histone H3 at serine 10 is correlated with chromosome condensation during mitosis and meiosis in Tetrahymena. *Proc Natl Acad Sci U S A*, 95(13):7480–7484, Jun 1998.

[231] S. West, N. Gromak, and N. J. Proudfoot. Human 5' –¿ 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature*, 432(7016):522–525, Nov 2004.

[232] S. West and N. J. Proudfoot. Transcriptional termination enhances protein expression in human cells. *Mol Cell*, 33(3):354–364, Feb 2009.

[233] D. L. Wheeler, D. M. Church, S. Federhen, A. E. Lash, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, E. Sequeira, T. A. Tatusova, and L. Wagner. Database resources of the National Center for Biotechnology. *Nucleic Acids Res*, 31(1):28–33, 2003.

[234] K. D. Wilson, S. Venkatasubrahmanyam, F. Jia, N. Sun, A. J. Butte, and J. C. Wu. MicroRNA profiling of human-induced pluripotent stem cells. *Stem Cells Dev*, 18(5):749–758, Jun 2009.

[235] J. L. Workman and R. E. Kingston. Nucleosome core displacement in vitro via a metastable transcription factor-nucleosome complex. *Science*, 258(5089):1780–1784, Dec 1992.

[236] C.-H. Wu, Y. Yamaguchi, L. R. Benjamin, M. Horvat-Gordon, J. Washinsky, E. Enerly, J. Larsson, A. Lambertsson, H. Handa, and D. Gilmour. NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila. *Genes Dev*, 17(11):1402–1414, Jun 2003.

[237] J. Wuarin and U. Schibler. Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Mol Cell Biol*, 14(11):7219–7225, Nov 1994.

[238] J. Wysocka, T. Swigut, H. Xiao, T. A. Milne, S. Y. Kwon, J. Landry, M. Kauer, A. J. Tackett, B. T. Chait, P. Badenhorst, C. Wu, and C. D. Allis. A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature*, 442(7098):86–90, 2006.

[239] T. Xiao, H. Hall, K. O. Kizer, Y. Shibata, M. C. Hall, C. H. Borchers, and B. D. Strahl. Phosphorylation of RNA polymerase II CTD regulates H3 methylation in yeast. *Genes Dev*, 17(5):654–663, Mar 2003.

[240] Y. Xue, Y. Zhou, T. Wu, T. Zhu, X. Ji, Y.-S. Kwon, C. Zhang, G. Yeo, D. L. Black, H. Sun, X.-D. Fu, and Y. Zhang. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol Cell*, 36(6):996–1006, Dec 2009.

[241] Y. Yamaguchi, T. Takagi, T. Wada, K. Yano, A. Furuya, S. Sugimoto, J. Hasegawa, and H. Handa. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell*, 97(1):41–51, Apr 1999.

[242] 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, Aug 2007.

[243] G. Yeo and C. B. Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*, 11(2-3):377–394, 2004.

[244] G. W. Yeo, N. G. Coufal, T. Y. Liang, G. E. Peng, X.-D. Fu, and F. H. Gage. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol*, 16(2):130–137, Feb 2009.

[245] H. Yu, S. Zhu, B. Zhou, H. Xue, and J. D. Han. Inferring causal relationships among different histone modifications and gene expression. *Genome Res*, 18(8):1314–24, 2008.

[246] G.-C. Yuan, Y.-J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*, 309(5734):626–630, Jul 2005.

[247] X. H.-F. Zhang and L. A. Chasin. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev*, 18(11):1241–1250, Jun 2004.

[248] X. H.-F. Zhang, C. S. Leslie, and L. A. Chasin. Computational searches for splicing signals. *Methods*, 37(4):292–305, Dec 2005.

[249] Y. Zhang. It takes a PHD to interpret histone methylation. *Nat Struct Mol Biol*, 13(7):572–574, Jul 2006.

[250] W. Zhou, P. Zhu, J. Wang, G. Pascual, K. A. Ohgi, J. Lozach, C. K. Glass, and M. G. Rosenfeld. Histone H2A monoubiquitination represses transcription by inhibiting RNA polymerase II transcriptional elongation. *Mol Cell*, 29(1):69–80, 2008.

[251] 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, Dec 2001.

[252] Y. Zhu, T. Pe'ery, J. Peng, Y. Ramanathan, N. Marshall, T. Marshall, B. Amendt, M. B. Mathews, and D. H. Price. Transcription elongation factor P-TEFb is required for HIV-1 tat transactivation in vitro. *Genes Dev*, 11(20):2622–2632, Oct 1997.

[253] Y. Zhuang and A. M. Weiner. A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. *Cell*, 46(6):827–835, Sep 1986.

Appendix

A.1 Supplementary figures

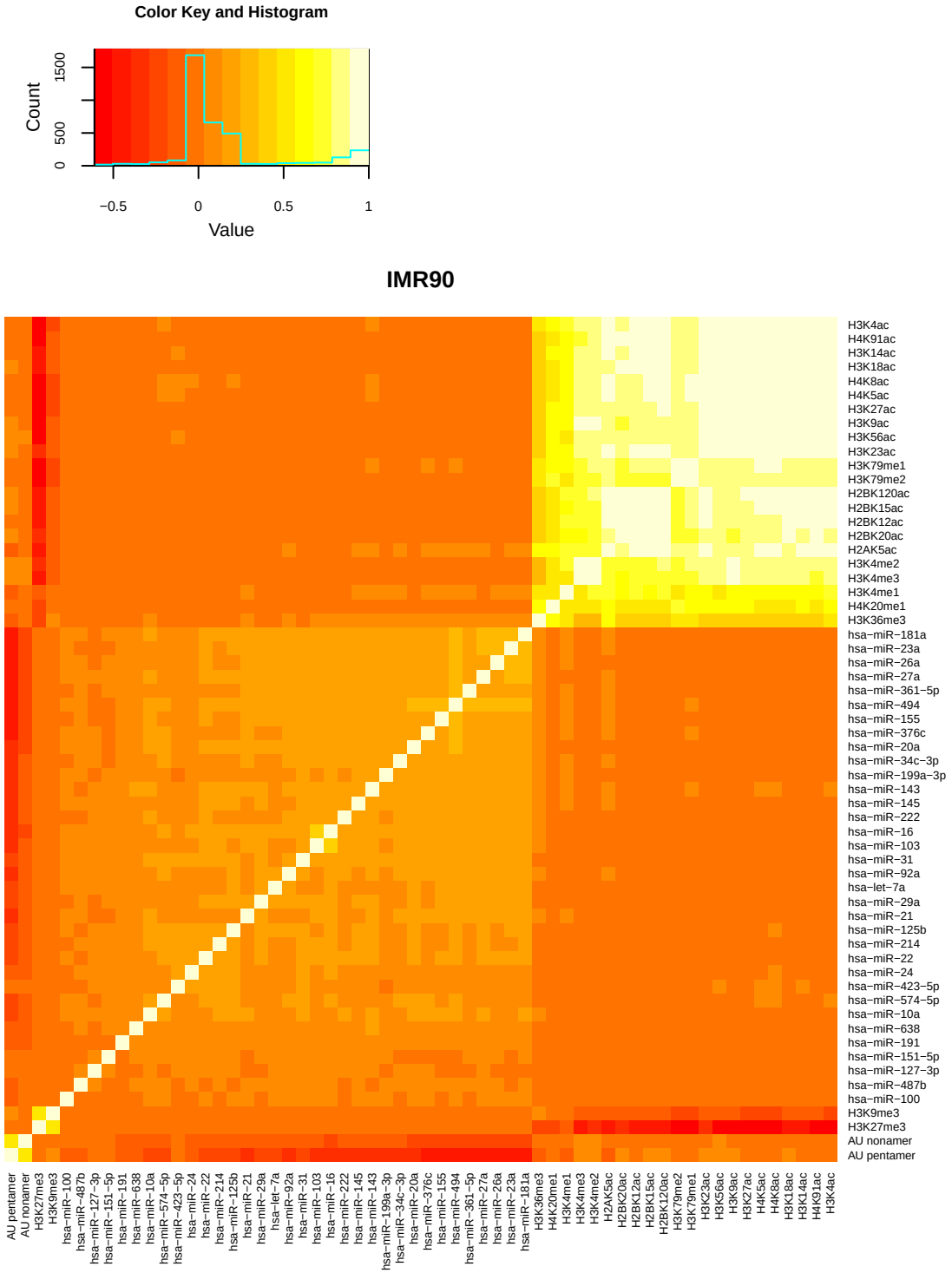

Figure A.1: Correlation of histone modifications, AU-rich elements and miRNA target sites in IMR90 cells. Heatmap showing the Pearson correlation coefficients between levels of histone modification levels in the promoter region, the number of pentamer and nonamer motifs for AU-rich elements in the 3' UTR and miRNA target sites in the 3' UTR. Only target sites for miRNAs whose expression is in the top 10% in IMR90 cells were used for the analysis.

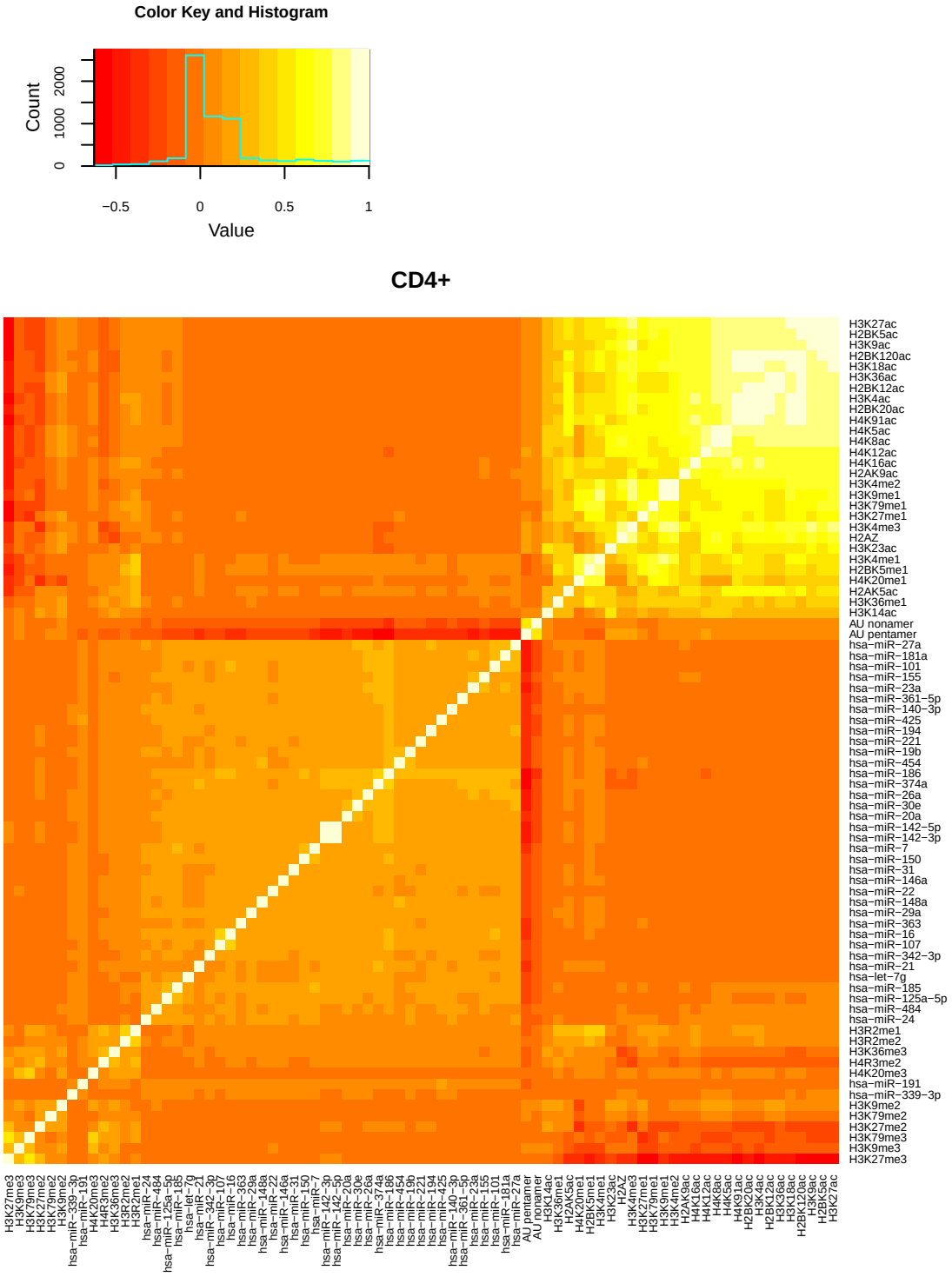

Figure A.2: Correlation of histone modifications, AU-rich elements and miRNA target sites in CD4+ cells. Heatmap showing the Pearson correlation coefficients between levels of histone modification levels in the promoter region, the number of pentamer and nonamer motifs for AU-rich elements in the 3' UTR and miRNA target sites in the 3' UTR. Only target sites for miRNAs whose expression is in the top 10% in CD4+ cells were used for the analysis.

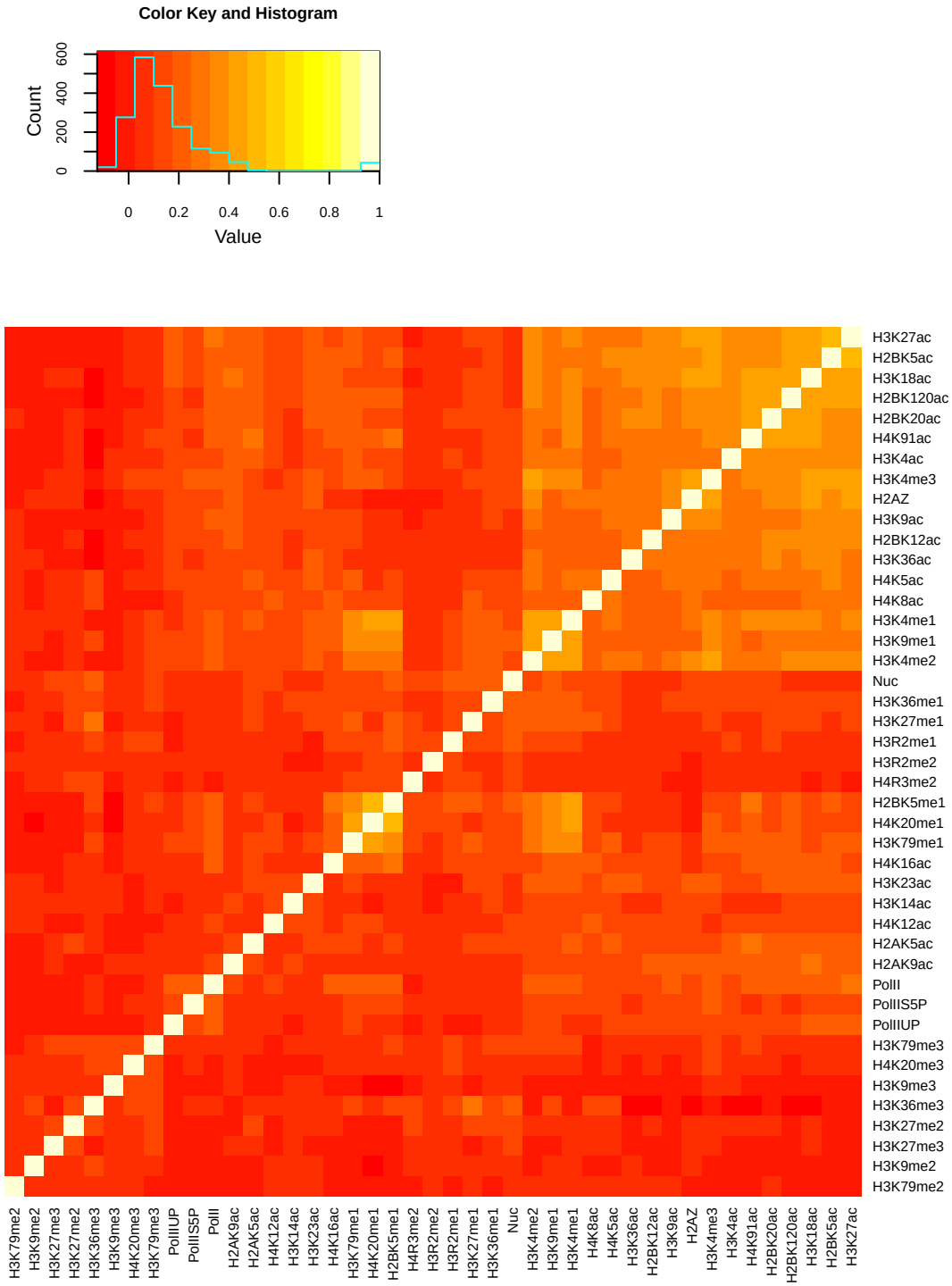

Figure A.3: Correlation of histone modifications, PolII and nucleosomes in alternative exons. Heatmap showing the Pearson correlation coefficients between the average number of ChIP-Seq or MNase-Seq tags/bp for histone modifications, PolII and nucleosomes in alternative exons.

A.2 Supplementary tables

		Class labels S	I
Prediction	S	268	615
	I	156	1091

accuracy = 63.08%

Table A.1: Prediction of exon skipping using logistic regression with scaled predictor variables. The predicted class label for each exon is the final prediction obtained by 100 repeats of 5-fold nested cross-validation.

		Class labels S	I
Prediction	S	268	606
	I	156	1100

accuracy = 64.23%

Table A.2: Prediction of exon skipping using logistic regression with variables corresponding to histone modifications normalized for nucleosome occupancy. The predicted class label for each exon is the final prediction obtained by 100 repeats of 5-fold nested cross-validation.

		Class labels S	I
Prediction	S	200	606
	I	128	1000

accuracy = 62.05%

Table A.3: Prediction of exon skipping using logistic regression for exons where at least 80% of the reads could be uniquely mapped (1934 exons). The predicted class label for each exon is the final prediction obtained by 100 repeats of 5-fold nested cross-validation.

		Class labels S	I
Prediction	S	260	603
	I	164	1103

accuracy = 63.99%

Table A.4: Prediction of exon skipping using logistic regression with predictor variables normalized for GC content of the exons. The predicted class label for each exon is the final prediction obtained by 100 repeats of 5-fold nested cross-validation.

HM+consseq	Not scaled data			Scaled data		
		Class labels			Class labels	
		S	I		S	I
Prediction	S	246	678	S	248	671
	I	178	1028	I	176	1035
	accuracy = 59.81%			accuracy = 60.23%		
HM+gs	Not scaled data			Scaled data		
		Class labels			Class labels	
		S	I		S	I
Prediction	S	262	636	S	263	645
	I	162	1070	I	161	1061
	accuracy = 62.54%			accuracy = 62.16%		
HM+maxent	Not scaled data			Scaled data		
		Class labels			Class labels	
		S	I		S	I
Prediction	S	248	703	S	251	706
	I	176	1003	I	173	1000
	accuracy = 58.73%			accuracy = 58.73%		
HM+score	Not scaled data			Scaled data		
		Class labels			Class labels	
		S	I		S	I
Prediction	S	251	676	S	250	677
	I	173	1030	I	174	1029
	accuracy = 60.14%			accuracy = 60.05%		

Table A.5: Prediction of exon skipping using a logistic regression model trained on histone modifications and splice site strength. Measures of splice site strength used in each model are indicated in the table. HM - histone modifications, consseq - matches to the consensus sequence, gs - GeneSplicer score, maxent - maximum entropy score, score - scores based on a PSSM of the splice site. The predicted class label for each exon is the final prediction obtained by 100 repeats of 5-fold nested cross-validation.

HM+gs+ESE/S	Not scaled data			Scaled data		
		Class labels			Class labels	
		S	I		S	I
Prediction	S	259	610	S	257	612
	I	165	1096	I	167	1094
	accuracy = 63.62%			accuracy = 63.43%		
HM+gs+ISE/S	Not scaled data			Scaled data		
		Class labels			Class labels	
		S	I		S	I
Prediction	S	280	611	S	278	613
	I	144	1095	I	146	1093
	accuracy = 64.55%			accuracy = 64.37%		
HM+gs+ESE/S+ISE/S	Not scaled data			Scaled data		
		Class labels			Class labels	
		S	I		S	I
Prediction	S	273	608	S	273	613
	I	151	1098	I	151	1093
	accuracy = 64.37%			accuracy = 64.13%		

Table A.6: Prediction of exon skipping using a logistic regression model trained on histone modifications, splice site strength and splicing regulatory elements. The features used in each model are indicated in the table. HM - histone modifications, gs - splice site strength measured by GeneSplicer, ESE/S - exonic splicing enhancers/silencers, ISE/S - intronic splicing enhancers/silencers. The predicted class label for each exon is the final prediction obtained by 100 repeats of 5-fold nested cross-validation.

Notation and abbreviations

3'SS	3' splice site
3' UTR	3' untranslated region
5'SS	5' splice site
5' UTR	5' untranslated region
A	adenine
ADP	adenosine diphosphate
ARE	AU-rich element
ATP	adenosine triphosphate
BIC	Bayesian information criterion
bp	base pair
BPS	branch point sequence
BRE	TFIIB recognition element
BrUTP	5-bromouridine 5'-triphosphate
C	cytosine
cDNA	complementary DNA
ChIP	chromatin immunoprecipitation
ChIP-Seq	chromatin immunoprecipitation followed by next-generation sequencing
CPSF	cleavage and polyadenylation specificity factor
CstF	cleavage stimulation factor
CTD	C-terminal domain of RNA polymerase II
DCE	downstream core element
DNA	deoxyribonucleic acid
DPE	downstream promoter element
DSIF	DRB-sensitivity inducing factor
EJC	exon junction complex
ESE	exonic splicing enhancer
ESS	exonic splicing silencer
FGFR2	fibroblast growth factor receptor 2
G	guanine
GRO-Seq	global run-on sequencing
HAT	histone acetyltransferase
HDAC	histone deacetylase
HCP	high-CpG content promoter
hnRNP	heterologous nuclear ribonucleoprotein particle
INR	initiator element
ISE	intronic splicing enhancer
ISS	intronic splicing silencer

kb	kilobase
LCP	low-CpG content promoter
miRNA	microRNA
MNase	micrococcal nuclease
mRNA	messenger RNA
MSE	mean squared error
NELF	negative elongation factor
NGS	next-generation sequencing
NMD	nonsense-mediated decay
NMR	nuclear magnetic resonance
NRO assay	nuclear run-on assay
NTP	nucleoside triphosphate
PIC	preinitiation complex
PolII	RNA polymerase II
PolIIS5P	Ser-5 phosphorylated PolII
PolIIUP	unphosphorylated PolII
PPT	polypyrimidine tract
PRC1	Polycomb repressive complex 1
PRC2	Polycomb repressive complex 2
pre-mRNA	precursor mRNA
pri-miRNA	primary microRNA
Pro	proline
PSSM	position specific scoring matrix
P-TEFb	positive transcription elongation factor b
PTB	polypirimidine tract binding protein
PTC	premature termination codon
qPCR	quantitative polymerase chain reaction
RISC	RNA-induced silencing complex
RNA	ribonucleic acid
RNA-Seq	sequencing of RNA using next-generation sequencing methods
r	Pearson correlation coefficient
SAGA	Spt-Ada-Gcn5-acetyltransferase
Ser	serine
SF1	splicing factor 1
snRNP	small nuclear ribonucleoprotein particle
T	thymine
TBP	TATA-binding protein
tags/bp	tags per base pair
TAF	TATA-associated factor
Thr	threonine
TIA1	T-cell restricted intracellular antigen I
TIAL1	TIA1-like 1
TSS	transcription start site
Tyr	tyrosine
U	uridine

U2AF	U2 auxiliary factor
UTR	untranslated region

Zusammenfassung

Histonproteinen liegen häufig chemisch modifiziert vor. Diese Modifikationen sind an vielen chromatinabhängigen Prozessen beteiligt. Diese Arbeit untersucht den Zusammenhang zwischen Histonmodifikationen und Transkription bzw. Splicing anhand von Modellen, die den Expressionslevel von Genen bzw. die Struktur von mRNAs mit Hilfe der Histonmodifikationen vorhersagen.

Unsere Ergebnisse zeigen, dass die Häufigkeit von Histonmodifikationen am Promotor und der Genexpressionslevel stark miteinander korreliert sind. Die Vorhersage der Genexpression hängt dabei nur von wenigen Histonmodifikationen ab. Dabei haben wir unterschiedliche Gruppen von Modifikationen identifiziert, die für eine gute Vorhersagequalität in Promotoren mit hohem bzw. niedrigen CpG Gehalt notwendig sind. Quantitative Modelle, die die Information von H4K4me3 und H3K79me1 beinhalten, haben die beste Qualität in Promotoren mit niedrigem CpG Gehalt, während Modelle, die H3K27ac und H4K20me1 verwenden, am Besten sind fr Promotoren mit hohem CpG Gehalt. Basierend auf diesen Ergebnissen schlagen wir einen vorläufigen "Histoncode für die Transkription" vor, in dem H3K4me3 an der Rekrutierung und/oder Initiation von RNA Polymerase II (Pol II) beteiligt ist, H3K27ac und H4K20me1 den Übergang zur Elongation ermöglicht und H3K79me1 und H4K20me1 den erfolgreichen Übergang anzeigt. Dieser vorläufige Histoncode für die Transkription muss in zukünftigen experimentellen Studien kritisch überprüft werden. Die gefundenen Zusammenhänge zwischen Histonmodifikationen und Genexpression sind von allgemeiner Natur, da es möglich war, die Genexpression von Zellen mit Modellen vorherzusagen, die in einem anderen Zelltyp erstellt worden sind. Unsere Ergebnisse zeigen weiterhin, dass unsere Modelle durch die Berücksichtigung von Prozessen, die die mRNA Stabilität beeinflussen, weiter verbessert werden können.

Unsere Untersuchungen zeigen, dass die Häufigkeit von Histonmodifikationen, Nukleosomen und Pol II verwendet werden können, um alternatives Splicing vorherzusagen. Wir haben vier Histonmodifikationen identifiziert (H3K27me1, H3K36ac, H3K36me3 und H3K79me2), die signifikant zur Vorhersagequalität unserer Modelle beitragen. Dieses Ergebnis legt nahe, dass diese Modifikationen im direkten Zusammenhang mit dem Splicingprozess stehen könnten. Histonmodifikationen tragen Information über alternatives Splicing, die z.T. komplementär zu den Sequenzinformationen in Exons und den umgebenden Regionen sind, was auf ein Zusammenspiel von Histonmodifikations- und Sequenzabhängigen Prozessen in der Regulation von Splicing hindeutet. Unsere Ergbnisse zeigen eine Abhängigkeit zwischen der Struktur von Transkripten und deren Expressionslevel, was eine funktionelle Kopplung zwischen Transkription

und Splicing bestätigt. Die Mechanismen, die diesen Beobachtungen zu Grunde liegen, müssen in Zukunft weiter untersucht werden.

Curriculum vitae

For reasons of data protection, the Curriculum vitae is not published in the online version.

For reasons of data protection, the Curriculum vitae is not published in the online version.

For reasons of data protection, the Curriculum vitae is not published in the online version.

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Berlin, September 2011 Rosa Karlić