Quantitative modeling and statistical analysis of protein-DNA binding sites

Matthias Siebert

München 2015

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig–Maximilians–Universität München

Quantitative modeling and statistical analysis of protein-DNA binding sites

Matthias Siebert
aus Frankfurt am Main, Deutschland

2015

Erklärung:

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Dr. Johannes Söding betreut.

Eidesstattliche Versicherung:

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, am 18. Dezember 2015

Matthias Siebert

Dissertation eingereicht am 18. Dezember 2015

1. Gutachter: Dr. Johannes Söding
2. Gutachter: Prof. Dr. Patrick Cramer

Mündliche Prüfung am 18. Februar 2016

Acknowledgments

I would like to express my deep gratitude to Johannes Söding for giving me the opportunity to work in his group. He was an amazing supervisor and mentor. His love for Bayesian modeling was compelling and I am particularly grateful for the confidence he placed in me and my capabilities. I especially thank Patrick Cramer for being the second assessor of my dissertation and for offering me to be part of a fantastic collaboration that resulted in his first functional genomics paper. His love to biology is catching. I am also very grateful to Dietmar Martin, Ulrike Gaul, Klaus Förstemann, and Roland Beckmann for spending their time as members of my dissertation committee.

I am much obliged to Andreas Mayer and Michael Lidschreiber. I think our collaboration was teamwork at its best, manifested as mutual appreciation, trust, and fun. The diligence of Andreas Mayer in establishing the ChIP-chip protocol laid the foundation for elaborate data processing and a sound statistical analysis. I am very grateful to Michi for all the precious moments we had in Munich's concert venues and on European festivals. We will always be connected by our shared love to music! I would also like to thank Katja Sträßer for a fruitful collaboration, and especially Dominik Meinel for being a diligent scholar in computational and statistical skills. I thank Sabine Brunner for conducting preliminary data analyses.

I am very thankful to Holger Hartmann for his continuous support of his XXmotif software and all the scientific and personal exchange we had in the early days of this thesis. I would like to thank Phillipp Torkler for providing PAR-CLIP data sets and Mark Heron for core promoter sequences, and both for numerous scientific discussions. I would also like to thank Katharina Hembach for analyzing genomic-context PBM data sets. I am very pleased that Anja Kiesel continues to develop GIMMEmotif in order to bring iIMMs to full bloom. Good luck!

I am particularly grateful for the very pleasant atmosphere in the computational biology hall of the Gene Center, which was established by the former Söding and the Tresch group members and could be preserved by the present Söding and the Gagneur group members. Thank you all for scientific exchange and lots of fun! In this context, I also thank the Graduate School of Quantitative Biosciences Munich for sponsoring the table soccer, which appeared to be indispensable for occasional non-scientific distraction. I would also like to thank the family of Johannes. I really enjoyed the Christmas dinners at your home and still appreciate the baby carrier we received as a gift for Ophelia. It is again in daily use for Lavinia. I am most thankful to all my dear friends who supported, influenced, or simply amused me along the way.

My eternal gratitude goes to my beloved family. First and foremost to Andrea Hildebrand, who I got to know in Johannes group when she was completing her studies with her diploma thesis. Thank you for all the time in intimate togetherness and together with our wonderful daughters Ophelia and Lavinia! Last but not least, this thesis was only possible with the unlimited love and backup of my parents.

Summary

This thesis consists of three parts. First, I extend the popular position weight matrix (PWM) model for describing DNA and RNA regulatory motifs such as transcription factor binding sites. The PWM model assumes that nucleotide positions contribute independently to the binding energy. Various models that take account of nucleotide dependencies have been shown to perform better in specific settings, but none has become widely used. As they contain many more parameters, they are more difficult to learn and more prone to overfitting. Here, I propose to describe regulatory motifs using inhomogeneous interpolated Markov models (iIMMs). iIMMs are similar to inhomogeneous Markov models of order k, but they automatically adapt context size k position-specifically for each $(k + 1)$-mer depending on its frequency, thus effectively preventing overtraining. I derive an iIMM expectation maximization algorithm, GIMMEmotif (General Interpolated Markov Model learning for the Elicitation of motifs), for de novo discovery of enriched motifs. On 446 ChIP-seq data sets from ENCODE, iIMMs achieve 41% mean and 26% median improvements in the partial area under the ROC curve, that is in predicting binding instances on held-out data. iIMMs also excel in learning complex regulatory regions, improving the fraction of correctly predicted transcription start sites, polyadenylation sites and bacterial pause sites over PWMs by between 26% and 101%. These results argue in favor of generally replacing PWMs by iIMMs.

Second, I reanalyze DNA binding regions of TFIIB and TBP measured by Venters and Pugh (2013) in human cells using ChIP-exo (chromatin immunoprecipitation with lambda exonuclease digestion followed by high-throughput sequencing). I show that the claimed universality of four degenerate core promoter elements (CPEs)—TATA box, BRE_u, BRE_d, and INR—is explained by the low specificities of the patterns used and that the same match frequencies are obtained with two negative controls (randomized sequences and scrambled patterns). CPE patterns are not positionally enriched around TFIIB locations, except for TATA elements with zero or one mismatched positions around mRNA-associated TFIIB peaks. I also cast doubt on the proposed biological significance of most of the non-mRNA-associated ChIP-exo peaks, 72% of which lie within repetitive regions. This reanalysis was published as Brief Communication Arising (Siebert and Söding, 2014) and led to the retraction of the original study (Venters and Pugh, 2014).

Third, I present high-resolution genome-wide occupancy profiling by chromatin immunoprecipitation (ChIP) coupled to tiling microarrays (ChIP-chip) of RNA polymerase (RNAP) II, its phosphorylated isoforms, its elongation factors and components of the RNAP II initiation and termination machinery in proliferating yeast cells, as published in Mayer et al. (2010). To eliminate all experimental biases, we developed a normalization procedure that corrects for nonspecific antibody binding by using input measurements as well as mock immunoprecipitations. Singular value decomposition analysis provides strong evidence for a general elongation complex—that is, one composed of all elongation factors—that mediates chromatin transcription and mRNA processing at all RNAP II genes. While RNAP II exchanges initiation factors for elongation factors during a single $5'$ transition just downstream of the transcription start site, the general elongation complex disassembles in a two-step $3'$ transition around the polyadenylation site. Transitions are uniform and independent of gene length, type and expression. General elongation complexes are active, as their gene occupancy predicts mRNA expression levels. The results also show that RNAP II C-terminal repeat domain (CTD) phosphorylation patterns previously observed at individual genes occur globally and that levels of CTD phosphorylation do not correlate with the *in vivo* occupancy of Spt6 and Pcf11 that bind the phosphorylated CTD *in vitro*. This indicates CTD-independent recruitment mechanisms and CTD masking *in vivo*.

Publications

Parts of this work have been published or are in the process of publication.

Part I

2016 (est.) **Higher-order models consistently outperform PWMs at predicting regulatory motifs in nucleotide sequences**
M. Siebert and J. Söding
Manuscript in preparation.

2013 ***P*-value-based regulatory motif discovery using positional weight matrices**
H. Hartmann, E. W. Guthöhrlein, **M. Siebert**, S. Luehr, and J. Söding
Genome Research 23(1):181–194.

Part II

2014 **Universality of core promoter elements?**
M. Siebert and J. Söding
Nature 511(7510):E11–E12.

Part III

2013 **Recruitment of TREX to the transcription machinery by its direct binding to the phospho-CTD of RNA polymerase II**
D. M. Meinel, C. Burkert-Kautzsch, A. Kieser, E. O'Duibhir, **M. Siebert**, A. Mayer, P. Cramer, J. Söding, F. C. P. Holstege, and K. Sträßer
PLoS Genetics 9(11):e1003914.

2010 **Uniform transitions of the general RNA polymerase II transcription complex**
A. Mayer*, M. Lidschreiber*, **M. Siebert***, K. Leike, J. Söding, and P. Cramer
(* contributed equally)
Nature Structural & Molecular Biology 17(10):1272–1278.

2009 **A guideline for ChIP-chip data quality control and normalization**
M. Siebert, M. Lidschreiber, H. Hartmann, and J. Söding
EpiGeneSys protocol 47.

Table of Contents

List of Figures

List of Tables

Part I.

Higher-order models consistently outperform PWMs at predicting regulatory motifs in nucleotide sequences

1. Introduction

The precise control of gene expression allows the cell to adapt its protein inventory in response to developmental and environmental cues. At the heart of gene expression regulation lies the binding of proteins to DNA and RNA sequences. Specifically, DNA-binding transcription factors play a fundamental role in regulating gene transcription, the major contributor to protein variance in the cell (Section 1.1). The progress in experimental technology to measure the binding specificity of DNA-binding proteins *in vitro* and determine their genomic binding locations *in vivo* (Section 1.2) advances the biological understanding of protein-DNA binding determinants (Section 1.3). To better predict targets of regulatory binding proteins and to enable predictive modeling of gene regulatory networks in health and disease, accurate quantitative models of the binding specificities of the involved factors are of critical importance. After reviewing existing computational approaches to model protein-DNA binding specificity (Section 1.4), I conclude the introduction by outlining my contribution to learning complex binding models and resultant biological insights into protein-nucleic acid interactions (Section 1.5).

1.1. Protein-DNA binding-mediated transcriptional control determines protein abundance

To highlight the central role that transcription factors play in the regulation of gene expression and protein abundance, I discuss two central dogmas that have been postulated, challenged, and, recently, reestablished. According to the first dogma, the variance in protein levels is mainly determined by the rate of gene transcription, as opposed to post-transcriptional processes such as translation and degradation of mRNAs and proteins. Schwanhäusser et al. (2011) challenged this view by claiming that differences in translation rates dominate, with transcription rates explaining only 34% of the variance in protein abundances. Recent studies reestablished transcription rates to account for the larger contribution (summarized by Li and Biggin (2015)). For instance, by calculating error-corrected estimates, Li et al. (2014a) ascribe 73% of the variance in protein levels to changes in mRNA synthesis rates. Similar estimates were ascertained from independent experimental measurements by Battle et al. (2015) and Jovanovic et al. (2015). Hence, transcriptional control is the principal contributing factor to protein abundance variation.

The second dogma can be traced back to Jacob and Monod (1961), who discovered that transcription factors control the synthesis of proteins by binding to regulatory DNA elements, thus ascribing the main determinant of transcription rates to transcription factors (see Ptashne (2014) for a historical reflection). In the last decade, this view was challenged by the experimental discovery of correlations between genome-wide chromatin modification patterns and gene expression (e.g. Barski et al. (2007)), leading to the formulation of the epigenetic code (Turner, 2007). Computational approaches, predicting gene expression from histone modification states (Karlić et al. (2010), Dong et al. (2012)) and human epigenomic marks from DNA motifs (Whitaker et al., 2015), as well as epigenome-wide association studies, designed to study human diseases (reviewed by Rakyan et al. (2011)), followed. However, this correlative effect turned out to be indirect instead of causally determining. For instance, the origin of DNA methylation patterns was found to lie in the genotype (Gertz et al., 2011). Population genetics (Kilpinen et al., 2013) and functional genomics (Kwasnieski et al., 2014) attributed the causal role of gene expression regulation to sequence-specific transcription factors, with histone modifications frequently reflecting the primary regulatory event. Similarly, histone modification patterns can be accurately predicted from transcription factor binding (Benveniste et al., 2014). Finally, transcription factors have the capacity to solely mediate the reprogramming of differentiated cells into a pluripotent state (Takahashi and Yamanaka, 2006).

Hence, transcriptional control and, consequently, protein abundance, is primarily mediated by the binding of transcription factors to specific sequences in promoter-proximal and distal DNA elements. In order to unravel gene regulatory networks it is therefore crucial to measure and model the binding specificity and genome-wide binding locations of transcription factors.

1.2. Experimental methods for determining protein-DNA interactions

The *in vitro* DNA binding specificities of proteins, such as transcription factors, can be experimentally determined using high-throughput technology (Stormo and Zhao, 2010). For instance, large-scale binding assays have been established on microarray platforms, such as universal (Berger et al. (2006), Badis et al. (2009)) and genomic-context (Gordân et al., 2013) protein binding microarrays (PBMs) or cognate site identifier (CSI) arrays (Warren et al., 2006), as well as sequencing machines, as realized in HiTS-FLIP (Nutiu et al., 2011). Other high-throughput approaches are based on the systematic evolution of ligands by exponential enrichment (SELEX) (Jolma et al. (2010), Jolma et al. (2013)) or microfluidic devices that mechanically induce trapping of molecular interactions (MITOMI) (Maerkl and Quake (2007), Geertz et al. (2012)).

In order to determine protein-DNA interactions *in vivo*, Gilmour and Lis (1984) introduced chromatin immunoprecipitation (ChIP) of proteins crosslinked to DNA. ChIP only uncovers specific binding of proteins to DNA (Poorey et al. (2013), Toth and Biggin (2000)). However, binding can be either direct or indirect (via other directly binding proteins). To enable the elucidation of genome-wide binding locations, the ChIP procedure was later coupled to microarray analysis (ChIP-chip) (Ren et al., 2000) and high-throughput sequencing (ChIP-seq) (Johnson et al., 2007). Subsequently, genomic binding locations can be used to build binding specificity models. Compared to DNA binding specificity models learned from *in vitro* binding assays, *in vivo* binding models integrate additional features (discussed in Section 1.3) that determine the complex binding behavior of transcription factors in the cell (see Orenstein and Shamir (2014) for a comparative analysis of binding specificity models learned *in vitro* from PBM and HT-SELEX data in predicting *in vivo* binding locations determined by ChIP-seq).

Recently, ChIP-exo (Rhee and Pugh, 2011) and ChIP-nexus (He et al., 2015) increased the resolution of ChIP-derived protein-DNA binding footprints by including an exonuclease step before high-throughput sequencing the resulting digested DNA fragments. Intriguingly, Kasinathan et al. (2014) described an approach, termed ORGANIC, that measures occupied genomic regions from naturally isolated chromatin without crosslinking.

Note that there have been reports about misleading ChIP enrichments of multiple unrelated proteins at highly transcribed loci in *Saccharomyces cerevisiae* (Teytelman et al. (2013), Park et al. (2013)). Referring to Ward et al. (2014), ChIP-enrichment profiles display a combination of sequence-specific as well as hotspot targeting, supposed to originate from mechanisms other than specific DNA sequence binding (but ruling out a chromatin state that is in general more prone to ChIP).

ChIP relies on transcription factor-specific antibodies. For this reason, genome-wide ChIP experiments determine the binding locations of only one transcription factor at a time. Instead, experimental assays that measure regions of open chromatin genome-wide, using DNase I hypersensitive site sequencing such as DNase-seq (Hesselberth et al. (2009), Thurman et al. (2012), Neph et al. (2012)) and DNase-FLASH (Vierstra et al., 2014), or by sequencing transposase-accessible chromatin (ATAC-seq) (Buenrostro et al., 2013), offer an alternative without relying on antibodies. By determining accessible genomic regions, these methods reversely measure the genomic regions that are collectively occupied by the protein complement of the cell. Subsequently, however, footprints need to be assigned to individual proteins, which is not unambiguously feasible, in particular for footprints of paralogous transcription factors or transcription factors containing DNA-binding domains of the same family.

Although single-cell approaches are on the rise (Rotem et al. (2015), Buenrostro et al. (2015)), established experiments based on ChIP or open chromatin measure a population average, that is, the proportion of cells in which each site was bound. Furthermore, neither approach provides information on binding kinetics, thus ignoring transcription factor binding turnover, which was found to be fundamental for determining the functional consequences of transcription factor binding (Lickwar et al. (2012), see Koster et al. (2015) for a review). In the case of DNase-seq, factor dynamics can even dictate footprint signatures (Sung et al., 2014), while DNase I cleavage bias needs to be corrected for (He et al. (2014), Yardımcı et al. (2014)).

1.3. Biological insights into the complexity of DNA binding determinants

Specific protein-DNA binding is established by the formation of specific hydrogen bonds and electrostatic interactions at the interface between the protein and its core DNA binding site, which, geometrically, need to precisely fit together (von Hippel and Berg (1986), von Hippel (2007)). While this binding mechanism depends on the sequence alone (base readout), statistical-mechanical selection theory indicated that functional specificity is based on further properties, in addition to primary sequence recognition (Berg and von Hippel, 1987)).

Early reports (Man and Stormo (2001), Bulyk et al. (2002)) and the advent of experimental high-throughput technologies to comprehensively determine protein-DNA specificities (Section 1.2) revealed an influence of nucleotide interdependencies on *in vitro* binding specificity (Badis et al. (2009), Nutiu et al. (2011), Jolma et al. (2013)), which was partly attributed to a readout mechanism that depends on structural properties of the DNA molecule (shape readout) emerging from stacking interactions between adjacent nucleotides (Rohs et al., 2010).

Besides transcription factor concentration and binding site affinities, determinants of *in vivo* binding locations were found to be even more complex (Rohs et al. (2010), Levo and Segal (2014), Siggers and Gordân (2014), and Slattery et al. (2014)), depending on the cooperativity with transcription factors that bind to neighboring or overlapping sites, competition with nucleosomes, and overall chromatin accessibility, as demonstrated for the master regulator PU.1 (Pham et al., 2013). In the following, I address several binding specificity determinants in greater detail. Note, however, that binding determinants are highly interdependent, resulting in smooth transitions between seemingly distinct concepts.

Transcription factor abundance

Of the estimated 10^4 to 3×10^5 transcription factor molecules in a cell (Biggin (2011), Li et al. (2014a)), only a relatively small percentage (30% or less) are specifically bound to the DNA (Zabet and Adryan, 2015). Besides, transcription factors also bind non-specifically, e.g. by one-dimensional sliding on DNA (Hammar et al., 2012). Their number was calculated to lie in a similar range compared to specifically bound transcription factors (Mueller et al., 2013). Crucially, only specifically bound transcription factors were reported to be engaged in productive transcription (Morisaki et al., 2014).

Multiple binding modes

Transcription factors locate binding targets using not necessarily only a single binding mode. In fact, an increasing number of proteins are reported to employ multiple modes of binding, in part with very different binding specificities, as is the case for the lac repressor (Zuo and Stormo, 2014) and the basic leucine zipper (bZIP) transcription factor Hac1 (Fordyce et al., 2012). Moreover, bZIP transcription factors can tolerate variable spacing in between binding half-sites (Kim and Struhl (1995), Gordân et al. (2011)). Remarkably, the transcription factor SP1 was found to bind with comparable binding specificity to a G-quadruplex, a non-canonical DNA structure, in addition to its canonical DNA duplex binding sequence (Raiber et al., 2012).

The binding to primary and secondary DNA motifs can be facilitated by alternative structural conformations, as demonstrated for the glucocorticoid receptor (Meijsing et al., 2009). Here, nonspecific (but shape-recognizing) interactions with bases of the “spacer” affect the conformation of distinct regions of the DNA-binding domain (Watson et al., 2013). DNA can therefore be considered as a sequence-specific allosteric ligand (Watson et al., 2013). While some transcription factors accomplish binding to different DNA motifs by different arrangements of multiple DNA-binding domains, as achieved by the POU domains of Oct-1 (Klemm et al. (1994), Verrijzer et al. (1992)), Jolma et al. (2013) identified Elk1 to show different sequence specificities dependent on its multimeric state.

Cooperative binding

The concomitant binding of one or more cofactors can also affect the intrinsic binding specificity of a transcription factor. For instance, non-DNA-binding cofactors can modulate the binding specificity of a transcription factor by changing its specificity to core binding site (Slattery et al., 2011) or flanking (Siggers et al., 2011) nucleotides, a binding mode termed latent specificity (Slattery et al., 2011). Furthermore, the binding specificity can be changed by the cooperative binding of other DNA-binding proteins to closely located sites. The prevalence of this DNA-dependent interaction between transcription factors was systematically identified by Jolma et al. (2015), revealing an organization of transcription factor binding sites reminiscent of the enhanceosome model (Panne et al., 2007). The allosteric effect through DNA was also demonstrated by Kim et al. (2013).

Context-specific binding

The binding of transcription factors appears to depend on developmental and cellular context (Yáñez-Cuna et al., 2012). For instance, the estrogen receptor binds to high-affinity estrogen response elements in multiple cellular contexts, whereas its binding to cell-specific sites relies on interacting factors and was shown to depend on genomic context (Gertz et al., 2013).

Genomic sequences flanking the core binding site

Transcription factors that contain DNA-binding domains of the same family show highly similar sequence preferences *in vitro*. Their *in vivo* binding locations may, however, differ considerably. Carlson et al. (2010) found nucleotides flanking the core binding site to significantly influence transcription factor binding energetics. Subsequently, the impact of proximal and distal flanks was demonstrated for several bHLH transcription factors (Gordân et al. (2013), Mordelet et al. (2013), Rajkumar et al. (2013)). Changes outside of the core binding site were proposed to offer means to encode more gradual changes in binding affinity (Rajkumar et al., 2013), as was also suggested for Gcn4 and Gal4 binding (Levo et al., 2015).

One particular sequence feature that influences the binding of transcription factors is the poly(dA:dT) tract, a homopolymeric sequence of adenine nucleotides on one of the DNA strands (Segal and Widom, 2009). Its presence adjacent to core binding sites was documented by Jolma et al. (2013) and Levo et al. (2015), and its effect on transcription factor binding suggested to be mediated by DNA shape characteristics (see below).

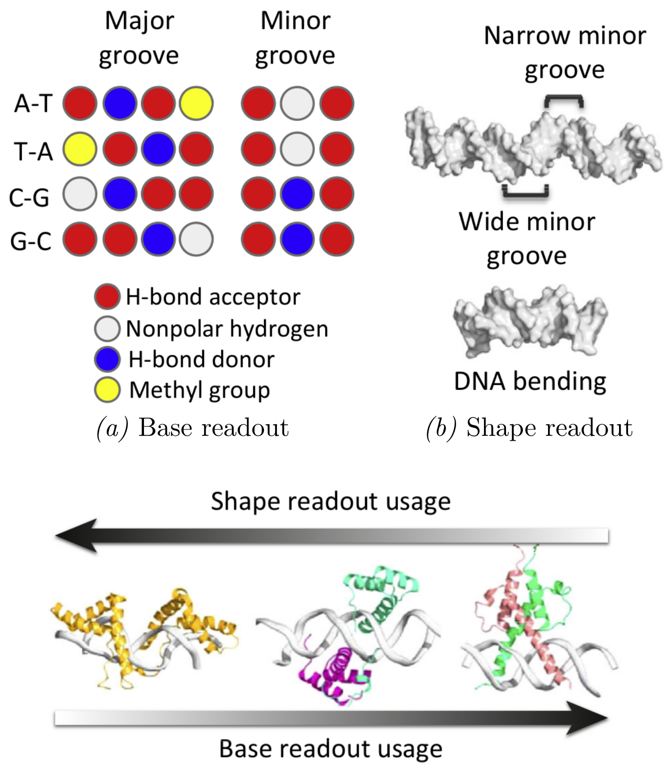
Consistent with the influence of proximal and distal flanks, White et al. (2013) found highly local sequence features (less than 50 nucleotides (nt) around the binding site) to distinguish the functional potential of Crx-bound sequences. In contrast, Dror et al. (2015) report on the widespread role of the larger binding site environment (up to 300 nt around the binding site) on determining *in vitro* and *in vivo* transcription factor binding across diverse protein families.

Afek et al. (2014) claimed that nonconsensus protein-DNA binding originating from DNA sequence symmetries contributes to *in vitro* and *in vivo* (Afek et al., 2015) protein-DNA binding affinity. In essence, this may, however, indicate that current approaches simply fail to incorporate the contributions of repetitive sequence elements, which are far from random (as discussed above), into specific (consensus) protein-DNA binding models.

Before focusing on DNA shape-based readout, I want to mention a special mode of sequence readout detected for the tumor suppressor protein p53. In this case, noncanonical Hoogsteen base pairs where observed at the central AT dimer of each half-site (Kitayner et al., 2010).

DNA shape

An increasing number of publications deals with a binding determinant attributed to DNA shape characteristics, such as minor groove width and DNA bending. Since hydrogen bonds between transcription factor amino acid residues and DNA bases in the minor groove of the DNA molecule lack specificity (Seeman et al., 1976), base-specific contacts are mainly established by the binding of transcription factors to the major groove. However, despite the degeneracy of base-specific contacts (Figure 1.1a), transcription factors bind to the minor groove by detecting local variations in DNA shape (Figure 1.1b). For instance, narrow minor grooves strongly enhance the negative electrostatic potential of the DNA, which, in turn, attracts basic side chains (Rohs et al., 2009). This potential was found to determine



(a) Base readout *(b)* Shape readout

(c) Base and shape readout usage

Figure 1.1.: **Base and shape readout contribute to transcription factor–DNA binding specificity. (a)** Base readout describes direct interactions between amino acids and the functional groups of the bases. Whereas the pattern of hydrogen bond acceptors (red) and donors (blue), heterocyclic hydrogen atoms (white) and the hydrophobic methyl group (yellow) is base pair-specific in the major groove, the pattern is degenerate in the minor groove. **(b)** Shape readout includes any form of structural readout based on global and local DNA shape features, including conformational flexibility and shape-dependent electrostatic potential. The DNA target of the IFN-β enhanceosome (PDB ID 1t2k; top) varies in minor groove shape. The human papillomavirus E2 protein binds to a DNA binding site (PDB ID 1jj4; bottom) with intrinsic curvature. **(c)** Most DNA-binding proteins use an interplay between base- and shape-readout modes to recognize their DNA binding sites. However, the contribution of each mechanism to protein-DNA binding specificity might vary across transcription factor families. Shape readout dominates for the minor groove-binding high motility group (HMG) box protein (PDB ID 2gzk; left). Base readout is a major contribution in DNA recognition by the bHLH protein Pho4 (PDB ID 1a0a; right). Both readout modes are more or less equally present in the DNA binding of a Hox–Exd heterodimer (PDB ID 2r5z; center). Figure from Slattery et al. (2014), distributed under the terms of the CC BY-NC-ND 3.0 license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

DNase I binding and cleavage rates (Lazarovici et al., 2013). The width of the minor groove was also shown to correlate with DMRT1 binding stoichiometry, indicating that DNA shape influences the formation of multimerization states (Murphy et al., 2015).

Importantly, minor-groove narrowing depends on the underlying DNA sequence and is associated with the presence of short structural elements, such as A-tracts, AT-rich sequences that exclude the flexible TpA step, originating from stacking interactions between adjacent base pairs (Rohs et al., 2009). Furthermore, the intrinsic bending of DNA regions that contain A-tract elements strongly depends on the temperature (Koo et al. (1986), Diekmann (1987)). This property is exploited by temperature-dependent binding of plant MADS-box transcription factors to the CArG-box, an A-tract with consensus CCW_6GG (Muiño et al., 2014). In contrast, DNA bending can also be induced by the binding of the transcription factor (Leonard et al., 1997).

The systematic effect of tetranucleotide sequence on B-DNA structure has been studied by microsecond molecular dynamics (Pasi et al., 2014). In the same manner, Pasi et al. (2015) found sequence-dependent potassium ion distributions around DNA. In addition, the sequence-specific formation of highly ordered and stable minor groove solvation networks was proposed to assist transcription factors in discriminating non-contacted bases (Harris et al., 2014).

Crucially, local DNA topology, as predicted from hydroxyl radical cleavage patterns, was found to be evolutionary conserved and correlated with functional non-coding regions of the human genome (Parker et al., 2009). Furthermore, distinct regions in homeodomain transcription factors were discovered to covary with either the sequence or the shape of their DNA binding sites (Dror et al., 2014). Remarkably, Abe et al. (2015) identified DNA shape-recognizing residues in Hox-DNA binding specificity, concluding that shape readout represents an independent component of binding site selection by Hox proteins, in addition to base readout (Figure 1.1c).

Note that transcription factor gene regulation strategies are markedly different between prokaryotes and eukaryotes (Wunderlich and Mirny, 2009). In eukaryotes, for instance, low-affinity binding sites were found to be important for specific and robust gene expression when existing in clusters (Crocker et al., 2015). In bilateria, however, the binding specificity of single transcription factors is highly conserved, and extends to secondary binding modes and subtle dinucleotide preferences (Nitta et al., 2015).

1.4. Computational approaches to model protein-DNA binding specificity

Historically, the binding to individual nucleotide positions within the binding site was considered to contribute independently of each other, owing to the small sample size of low-throughput experiments. This assumption limited the number of model parameters required and was implemented in the popular position weight matrix (PWM) model (Stormo et al. (1982), Staden (1984)). PWMs contain scores for each nucleotide position i and each of the four nucleotides x in the binding site. The scores are computed from a set of example binding sites by counting the fraction $p_i(x)$ of each nucleotide at each position i in the example sequences, dividing by the probability $f(x)$ of x in a background model (e.g. the fraction in the entire genome) and taking the logarithm, $\log \frac{p_i(x)}{f(x)}$. The total score of a PWM with a putative binding site sequence $x_1 \ldots x_L$ is the sum of the scores for each of the binding site nucleotides,

$$S(x_1 \dots x_L) = \sum_{i=1}^{L} \log \frac{p_i(x_i)}{f(x_i)} = \log \frac{\prod_{i=1}^{L} p_i(x_i)}{\prod_{i=1}^{L} f(x_i)}.$$

If the binding energy of the factor is additive over positions, then $k_B T$ times this score is proportional to the binding energy when the example binding sites are sampled at sufficiently low factor concentration to avoid saturating any of the binding sites. This additivity is equivalent to the assumption of independence of the probabilities of nucleotides at each position of the binding site.

Despite its rather drastic-sounding approximation, PWMs provide a good approximation to the true binding energy (Benos et al., 2002) and have been very successful as quantitative model for binding specificity, for two reasons. First, although contributions of individual base pairs are not strictly additive, the additive singleton terms in the binding energy clearly dominate over pair interaction terms. Second, models that account for correlations among nucleotides need many more parameters, and such models simply could not be trained reliably with relatively few available binding sites. To date, the PWM is the most widely used model for representing transcription factor binding sites, and there exist numerous approaches for estimating its parameters and deal with the intricacies of diverse data sources (e.g. Zhao et al. (2009), see Stormo (2013) for a review).

The advent of experimental high-throughput technologies to determine protein-DNA specificities (Section 1.2), however, strengthened the notion that the binding behavior is more complex, revealing the importance of binding determinants such as binding site context and shape readout (see Section 1.3). Characterizing the contribution of nucleotide dependencies to the overall binding energy should therefore improve the accuracy of binding models of transcription factors.

By now, numerous models incorporating nucleotide interdependencies have been developed. Some extend the PWM model to include dependencies between neighboring nucleotides (Zhang and Marr (1993), Gunewardena and Zhang (2008), Zhao et al. (2012), Kulakovskiy et al. (2013)) or among a sparse subset of pairs of positions (Zhou and Liu, 2004). Others include higher-order dependencies between neighboring binding site positions (Eggeling et al. (2014), Maaskola and Rajewsky (2014), Mathelier and Wasserman (2013)) or subsets of all positions (Ellrott et al. (2002), Barash et al. (2003), Zhao et al. (2005), Ben-Gal et al. (2005), Sharon et al. (2008), Hu et al. (2010), Keilwagen and Grau (2015)). Mixture models based on PWMs (Barash et al. (2003), Hannenhalli and Wang (2005), Georgi and Schliep (2006), Narlikar (2013)) or Markov models (Huang et al., 2006) aim to describe multiple binding modes. However, all these methods rely on heuristics to prune the dependency structure in order to avoid overfitting.

There have also been efforts to explicitly model DNA shape by deriving structural features, such as minor groove width and helical parameters, from Monte Carlo simulations (Zhou et al., 2013). Subsequently, Zhou et al. (2015) combined sequence and shape features to predict binding specificities. However, structural characteristics are a consequence of the underlying sequence and, thus, can only provide an indirect and not necessarily precise description of binding site characteristics. For this reason, regression models that rely on monomer, dimer, and trimer sequence features outperform models based on a combination of monomer and shape features, on average, in predicting universal PBM signal intensities (Zhou et al., 2015).

The question whether such more complex models can improve predictive performance over PWMs across the board has been controversially discussed (Morris et al., 2011), and several recent studies claimed simple models to be just as accurate as complex models for the large majority of transcription factors (Zhao and Stormo (2011), Weirauch et al. (2013)).

1.5. General models for the elicitation of transcription factor binding motifs

Tomovic and Oakeley (2007) investigated the statistical basis for either dependence or independence, suggesting that dependencies should be modeled in case evidence in their favor exists, whereas corrections should be omitted in case evidence for dependency is lacking. The same idea was the basis for the development of (homogeneous) interpolated Markov models (IMMs), which adjust their complexity to the amount of the available data. IMMs were introduced into bioinformatics by the famous GLIMMER software for predicting coding regions (Salzberg et al. (1998), Delcher et al. (1999)) and have also been successful in locating promoters (Ohler et al., 1999) and enhancers (Kazemian et al., 2011). IMMs learn the probability $p(x_j|x_{j-k}\dots x_{j-1})$ of a nucleotide x_j given the k preceding nucleotides $x_{j-k}\dots x_{j-1}$ by interpolating between the empirical estimate $\frac{n(x_{j-k}\dots x_j)}{n(x_{j-k}\dots x_{j-1})}$ and the lower-order probability $p(x_j|x_{j-k+1}\dots x_{j-1})$ depending on the number of counts $n(x_{j-k}\dots x_j)$. In this way, they adjust the effective order k to the amount of data available.

Here, IMMs are extended to *inhomogeneous* interpolated Markov models (iIMMs) in which the conditional probabilities $p_i(x_j|x_{j-k}\dots x_{j-1})$ depend on the position i in the model. By taking a Bayesian viewpoint, conditional probabilities are interpreted as arising from the product of a likelihood term and a Dirichlet prior whose pseudocount parameters are taken from lower-order probabilities. When enough data, that is, counts of $(k+1)$-mer $x_{j-k}\dots x_j$, are available, the data dominates over the prior. Conversely, when data gets sparse, the prior dominates and the probability will revert to a lower-order estimate. In this way, probabilities are combined from those sequence contexts for which sufficient data are available to produce good estimates, thereby adapting model complexity to the data. I found that, compared to non-interpolated inhomogeneous Markov models (iMMs), iIMMs are insensitive to parameter overfitting even when using exceedingly high model orders, without requiring prior knowledge about the prevalence of nucleotide interdependencies in the training data.

By modeling dependencies between neighboring nucleotides, iIMMs inherently learn DNA structural properties such as bendability and minor groove width, which are mainly determined by stacking interactions between neighboring bases. Furthermore, in almost all cases where single amino acids contact multiple bases simultaneously, bases appear directly adjacent to each other (Luscombe et al., 2001), with interactions between adjacent nucleotides proposed to be stronger than interactions between non-adjacent nucleotides (Jolma et al. (2013), Luscombe et al. (2001)). In agreement, O'Flanagan et al. (2005) found out that significant non-additivity, caused by DNA deformations within a protein complex, is almost exclusively limited to nearest-neighbor interactions.

iIMMs can be directly computed from aligned binding sites. However, knowledge is often lacking about the exact location of the binding site within the sequence experimentally determined to be bound by a transcription factor. For instance, binding events measured by ChIP-seq are commonly reported with a resolution of hundreds of base pairs (The ENCODE Project Consortium, 2012). One technique that was successfully applied to infer PWM models of binding motifs discovered to be enriched in training sequences, as compared to a background model, is the expectation maximization (EM) algorithm (Dempster et al., 1977). The EM algorithm has been introduced to probabilistic motif discovery by Lawrence and Reilly (1990) and has been enhanced and implemented in the popular MEME (Multiple Em for Motif Elicitation) software (Bailey and Elkan, 1994). However, no method exists to learn the more sophisticated iIMMs from unaligned binding sites. Here, I derive an EM algorithm, GIMMEmotif (General Interpolated Markov Model learning for the Elicitation of motifs), that learns iIMMs representing motifs enriched in a set of training sequences, compared to a set of background sequences, which is modeled by an IMM.

To better understand the information contained in iIMMs, the popular sequence logo (Schneider and Stephens, 1990) representation is extended to depict higher-order dependencies between neighboring sequence positions. For each model order, the higher-order sequence logo details the contribution of oligomers to the information content that is not provided by oligomers of a lower model order. In this manner, it is possible to decipher how good the approximation already is when using lower-order models, and to see what comes on top with each higher order.

I test iIMM learning on diverse DNA sequence sets, ranging from binding sites of single transcription factors, including RNA polymerase pause sites, to regulatory regions composed of complex motif architectures which are typically bound by multiple cooperatively binding proteins, such as core promoter and polyadenylation site sequences. In addition to protein-DNA interactions, I model protein-RNA interactions around PAR-CLIP crosslink sites. In order to obtain insights into the relative importance of dependencies between nucleotides with increasing inter-nucleotide distance and the prevalence of binding mechanisms, I learn PWM models and iIMMs of increasing order and compare their performance in predicting binding sites and binding affinities using cross-validated benchmark tests. While performance increases monotonously with the order of iIMMs until saturation, iIMMs consistently outperform PWM models. This argues for generally replacing PWM models by iIMMs.

2. Materials and Methods

In this Chapter, I discuss how the modeling of the binding preference of a transcription factor to a DNA sequence can be made feasible by approximating the binding free energy using the simple position weight matrix (PWM) model (Section 2.1.1) as well as more complex inhomogeneous Markov models (iMMs) (Section 2.1.2) and inhomogeneous interpolated Markov models (iIMMs) (Section 2.1.3). I also describe an approach to visualize the information contained in complex models by extending the popular sequence logo representation of PWM models (Section 2.2). Provided with the capacity to learn sequence models of varying complexity from a set of aligned sequences, I then explain how to learn models from a set of unaligned sequences using the EM algorithm (Section 2.3.1). Particularly, I show how the EM algorithm can be adapted to optimize iIMMs (Sections 2.3.2 and 2.3.3). Finally, I detail the data sets used to evaluate sequence models of varying complexity, including all processing steps and benchmark tests performed (Section 2.4).

2.1. How to model protein-DNA binding specificity?

In order to model the binding preference of a transcription factor to a DNA sequence $x_1 \dots x_L$ of length L, the DNA bases $x_1, \dots, x_L \in \{A,C,G,T\}$ are denoted by random variables $X_1, \dots, X_L$. According to Boltzmann's law, the probability of binding is related to the free energy of binding ΔE by

$$p(X_1 \dots X_L = x_1 \dots x_L) \propto e^{-\frac{\Delta E(x_1 \dots x_L)}{k_B T}}, \tag{2.1}$$

where k_B is the Boltzmann constant and T the temperature. Here, the occupation probability of a transcription factor on the sequences is assumed to be nearly zero (weak binding approximation).

In order to obtain accurate probability estimates, sufficient occurrences of all possible 4^L sequences must be present in the data. Since a transcription factor binding site can readily encompass 20 base pairs (bp), the amount of available data is however limiting. Therefore, it is necessary to find an approximation to the binding probability which is practically suitable but still able to capture the binding preference of the transcription factor in an appropriate manner. This Section covers simple and more complex models that compromise between model simplicity and accuracy.

2.1.1. PWM models

The position weight matrix (PWM) model was introduced by Stormo et al. (1982) and became the most popular model for representing transcription factor binding sites (Stormo, 2013). The model is based on the idea to approximate the binding probability

$$p(x_1 \dots x_L) = p_1(x_1)\, p_2(x_2|x_1)\, p_3(x_3|x_1x_2)\, \cdots\, p_L(x_L|x_1 \dots x_{L-1}) \tag{2.2}$$

of a transcription factor to a DNA sequence $x_1 \dots x_L$ of length L by assuming that the preferences at each position i in the binding site are independent of each other

$$p(x_1 \dots x_L) \approx \prod_{i=1}^{L} p_i(x_i). \tag{2.3}$$

Related to the free energy of binding (Equation 2.1), and by defining $\Delta E_i = -k_B T \log p_i(x_i)$, it becomes apparent that

$$\Delta E \approx \sum_{i=1}^{L} \Delta E_i, \tag{2.4}$$

in other words, the energy of binding is assumed to be a sum of independent terms over all binding bases.

For N examples of known binding sites, I may estimate the probabilities $p_i(x_i)$ of base x_i being at position i of the binding site by

$$p_i(x_i) = \frac{n_i(x_i)}{N}. \tag{2.5}$$

$n_i(x_i)$ is the number of times base x_i has been observed at site i (out of N total). However, to take account of the fact that we have insufficient knowledge to exclude bases from occurring at the binding site just because they have not yet been observed in a finite sample, I introduce pseudocounts that are proportional to the background frequency $f(x_i)$ of base x_i

$$p_i(x_i) = \frac{n_i(x_i) + \alpha_0 f(x_i)}{N + \alpha_0}, \tag{2.6}$$

where α_0 is the total number of pseudocounts applied. When sufficient data are available, the effect of the pseudocounts becomes negligible and $p_i(x_i)$ becomes essentially equal to the maximum likelihood solution. For $\alpha_0 = 0$, Equation 2.5 is recovered.

The advantage of the PWM model is the small number of parameters it requires ($3 \times L$ compared to $4^L - 1$ in the full model). However, dependencies between nucleotides contributing to the binding energy cannot be modeled. Although working well in many cases, it is highly debated as to how accurate the approximation is (see Section 1.4).

2.1.2. Inhomogeneous Markov models (iMMs)

How can the approximation in Equations 2.3 and 2.4 be improved? In Equation 2.3, I replaced all conditional probabilities by the monomer probabilities $p_i(x_i)$, thereby loosing information about correlations between positions. This corresponds to an inhomogeneous (that is, position-specific) Markov model (iMM) of order zero. In the general iMM of order k, information about correlations between $k+1$ neighboring sequence positions is retained

$$p(x_1 \dots x_L) \approx \prod_{i=1}^{L} p_i(x_i | x_{i-k} \dots x_{i-1}). \tag{2.7}$$

Here, the probability of base x_i being at position i of the binding site depends on the nature of bases at the preceding binding site positions $i-k$ to $i-1$. Contexts are restricted to the binding site. For instance, in a model of order two, $p_1(x_1|x_{-1}x_0)$ and $p_2(x_2|x_0x_1)$ are defined as $p_1(x_1)$ and $p_2(x_2|x_1)$, respectively. Similar to Equation 2.6, the conditional probability can be calculated by incorporating pseudocounts that are proportional to the monomer background frequencies

$$p_i(x_i|x_{i-k} \dots x_{i-1}) \approx \frac{n_i(x_{i-k} \dots x_i) + \alpha_k f(x_i)}{n_{i-1}(x_{i-k} \dots x_{i-1}) + \alpha_k}. \tag{2.8}$$

By taking the preceding context of binding site positions into account, iMMs, as opposed to PWM models, are able to describe stacking interactions between adjacent base pairs and short structural elements such as A-tracts (see Section 1.3). However, learning such models accurately requires sufficient data for all possible oligomers ($\gtrsim 100 \times (4^k - 1)$ counts) or prior knowledge of the dependencies between bound base pairs. Since both are lacking in many cases, iMMs are more susceptible to be affected by statistical noise and thus prone to overfitting, as compared to PWM models.

2.1.3. Inhomogeneous interpolated Markov models (iIMMs)

To cope with the increase in parameter space, I employ inhomogeneous interpolated Markov models (iIMMs) that are able to infer interdependencies of neighboring nucleotides by automatically adapting model complexity to the data (Salzberg et al., 1998). This is achieved by mixing higher-order oligomer counts with pseudocounts calculated from lower-order oligomer probabilities, instead of monomer background frequencies. By calculating conditional probabilities using

$$p_i(x_i|x_{i-k} \dots x_{i-1}) \approx \frac{n_i(x_{i-k} \dots x_i) + \alpha_k p_i(x_i|x_{i-k+1} \dots x_{i-1})}{n_{i-1}(x_{i-k} \dots x_{i-1}) + \alpha_k}, \tag{2.9}$$

iIMMs interpolate between counts and pseudocounts. Counts dominate the numerator at sequence contexts for which sufficient data are available to produce good estimates.

Conversely, higher-order probabilities fall back on lower-order probabilities when too few counts are available. For this reason, iIMMs do not require a minimum number of oligomer counts. Contrary to iIMMs, the higher the order of iMMs, the more counts are mandatory to accurately estimate the probabilities (Section 2.1.2).

The interpolation scheme is similar to the rational interpolation technique described by Ohler et al. (1999). Because dependencies between neighboring nucleotides should generally decrease with increasing distance between nucleotide positions (Jolma et al., 2013), I, however, raise the contribution of lower-order probabilities with increasing model order and suggest to set order-specific pseudocounts α_k as follows

$$\alpha_k = \begin{cases} 1, & \text{if } k = 0 \\ 20 \times \beta^{k-1}, & \text{if } k > 0, \end{cases} \tag{2.10}$$

using $\beta = 3$. The choice of pseudocounts is rather conservative but confers robustness (see Section 2.4.8). Compared to iMMs, no prior knowledge is needed about the prevalence of nucleotide interdependencies within specific binding sites.

In this work, I use iIMMs to model regulatory sequences ranging from binding sites of single transcription factors to sequence regions bound by multiple transcription factors, such as core promoters. Background sequences are modeled by homogeneous interpolated Markov models (IMMs).

2.2. Higher-order sequence logos

To visualize the content of a PWM model, Schneider and Stephens (1990) developed the sequence logo. The information content (in bits) of columns in the PWM corresponds to the height of columns in the sequence logo, and describes the importance of positions to overall protein-DNA binding specificity. The height of the four bases in each column is determined by their relative frequencies. More frequent bases are depicted on top of less frequent bases. Consequently, the consensus sequence can be assembled from the top bases, while the vertical order of bases in each column corresponds to their order of predominance.

The sequence logo was designed to reflect the characteristics of the PWM model. Therefore, it is not suited to illustrate dependencies between binding site positions. Extensions to the sequence logo have been proposed (Eden and Brunak (2004), Sharon et al. (2008), Mathelier and Wasserman (2013), Jolma et al. (2013), Keilwagen and Grau (2015)), but none of these approaches breaks down the information content into the contributions of different model orders, irrespective of the underlying model and its maximum order. To close this gap and to better understand the information contained in a higher-order iIMM (Section 2.1.3), I extend the sequence logo to depict higher-order dependencies between neighboring sequence positions. For this purpose, it is necessary to calculate the contributions of different orders to the information content.

The log-odds score is the best score to discriminate positive from negative sequences. Given position-specific conditional model probabilities $\boldsymbol{p}_i$ of order M and conditional background frequencies $\boldsymbol{f}$ of order B, the log-odds score S for the sequence $x_1 \dots x_L$ of length L can be calculated by

$$S(x_1 \dots x_L) = \sum_{i=1}^{L} \log_2 \frac{p_i(x_i|x_{i-M} \dots x_{i-1})}{f(x_i|x_{i-B} \dots x_{i-1})}. \tag{2.11}$$

This score can be decomposed into contributions from all orders zero to M, using a fixed-order background model. For instance, the following order contributions for a 1st-order

$$\begin{aligned}&\sum_{i=1}^{L} \log_2 \frac{p_i(x_i|x_{i-1})}{f(x_i|x_{i-B} \dots x_{i-1})} \\ &= \sum_{i=1}^{L} \left[\underbrace{\log_2 \frac{p_i(x_i)}{f(x_i|x_{i-B} \dots x_{i-1})}}_{0^{\text{th}}\text{-order contributions}} + \underbrace{\log_2 \frac{p_i(x_i|x_{i-1})}{p_i(x_i)}}_{1^{\text{st}}\text{-order contributions}} \right],\end{aligned}$$

and 2nd-order

$$\begin{aligned}&\sum_{i=1}^{L} \log_2 \frac{p_i(x_i|x_{i-2}x_{i-1})}{f(x_i|x_{i-B} \dots x_{i-1})} \\ &= \sum_{i=1}^{L} \left[\underbrace{\log_2 \frac{p_i(x_i)}{f(x_i|x_{i-B} \dots x_{i-1})}}_{0^{\text{th}}\text{-order contributions}} + \underbrace{\log_2 \frac{p_i(x_i|x_{i-1})}{p_i(x_i)}}_{1^{\text{st}}\text{-order contributions}} + \underbrace{\log_2 \frac{p_i(x_i|x_{i-2}x_{i-1})}{p_i(x_i|x_{i-1})}}_{2^{\text{nd}}\text{-order contributions}} \right],\end{aligned}$$

model can be obtained. For general order M, the score can be decomposed by

$$\begin{aligned}&\sum_{i=1}^{L} \log_2 \frac{p_i(x_i|x_{i-M} \dots x_{i-1})}{f(x_i|x_{i-B} \dots x_{i-1})} \\ &= \sum_{i=1}^{L} \left[\underbrace{\log_2 \frac{p_i(x_i)}{f(x_i|x_{i-B} \dots x_{i-1})}}_{0^{\text{th}}\text{-order contributions}} + \sum_{M'=1}^{M} \underbrace{\log_2 \frac{p_i(x_i|x_{i-M'} \dots x_{i-1})}{p_i(x_i|x_{i-M'+1} \dots x_{i-1})}}_{\text{higher-order contributions}} \right].\end{aligned} \tag{2.12}$$

However, how can the information content in the various orders be evaluated? Note that for sequences distributed like background (bg), we get

$$\mathbb{E}_{x_1 \dots x_L \in \text{bg}}\left[S_i(x_1 \dots x_L)\right] = 0 \tag{2.13}$$

by definition of the log-odds score.

For positive (pos) sequences, we obtain

$$
\begin{aligned}
&\mathbb{E}_{x_1 \dots x_L \in \text{pos}}\big[S_i(x_1 \dots x_L)\big] = \sum_{x_1 \dots x_L} P(x_1 \dots x_L) S(x_1 \dots x_L) \\
&= \left.\sum_{i=1}^{L} \sum_{a \in \{\text{A,C,G,T}\}} p_i(x_i = a) \log_2 \frac{p_i(x_i = a)}{f(x_i = a)} \right\} 0^{\text{th}}\text{-order} \\
&\quad \left. + \sum_{i=1}^{L} \sum_{a,b \in \{\text{A,C,G,T}\}} p_i(x_{i-1} x_i = ab) \log_2 \frac{p_i(x_i = b | x_{i-1} = a)}{p_i(x_i = b)} \right\} 1^{\text{st}}\text{-order} \\
&\quad \left. + \sum_{i=1}^{L} \sum_{a,b,c \in \{\text{A,C,G,T}\}} p_i(x_{i-2} x_{i-1} x_i = abc) \log_2 \frac{p_i(x_i = c | x_{i-2} x_{i-1} = ab)}{p_i(x_i = c | x_{i-1} = b)} \right\} 2^{\text{nd}}\text{-order} \\
&\quad \left. + \dots \right\} \text{up to } M^{\text{th}}\text{-order} \qquad (2.14)
\end{aligned}
$$

Each term specifies the contribution of its order to $\mathbb{E}[S_i(x_1 \dots x_L)]$ in bits. A contribution of 1 bit increases the probability for the positive sequences in comparison to the negative sequences by a factor 2^1 on average.

Given the contribution of each k-mer to the information content, I construct a separate sequence logo for each order $k-1$. In contrast to the sequence logo by Schneider and Stephens (1990), the height of both columns and k-mers corresponds to the contribution to the information content that is not already described in a lower order. Note that k-mers can exhibit negative contributions to the information content.

2.3. How to learn protein-DNA binding specificity models?

The models described in Section 2.1 can be easily computed from aligned binding sites. However, knowledge is often lacking about the exact location of the binding site within the sequence experimentally determined to be bound by a transcription factor. For instance, binding events measured by ChIP-seq are commonly reported with a resolution of hundreds of base pairs.

One technique that was successfully applied to infer PWM models from such unaligned binding sites is the expectation maximization (EM) algorithm (Dempster et al., 1977). The EM algorithm has been introduced to probabilistic motif discovery by Lawrence and Reilly (1990) and has been enhanced and implemented into the popular MEME software (Bailey and Elkan, 1994). However, I am not aware of any method that is capable of learning the more sophisticated iIMMs from unaligned binding sites. To achieve this objective, the EM algorithm is adapted.

To simplify the understanding of the learning procedure for iIMMs, I commence by describing the general idea of the EM algorithm and its application to learning PWM models from unaligned binding sites. Subsequently, I explain how to adapt the EM algorithm to learn iIMMs. Finally, I propose a simple heuristic that offers the opportunity to incorporate sequence weights into the learning process.

2.3.1. The EM algorithm

The EM algorithm is a general technique for finding maximum likelihood (ML) solutions for probabilistic models having some hidden or latent variables (variables that are not observed). This is achieved by alternating between the estimation of the latent variables (E step) and the unknown parameters (M step). In the E step, instead of finding a point estimate of the best hidden variables given the current estimate of the model, the EM algorithm computes a probability distribution over the hidden variables. In the M step, the unknown parameters are optimized given the posterior distribution of the hidden variables calculated in the E step. In the following, I describe the EM algorithm in more detail, based on Bishop (2006).

I denote the set of all observed data by $\boldsymbol{X}$ and the set of all latent variables by $\boldsymbol{Z}$. The set of all parameters is denoted by $\boldsymbol{\theta}$. Hence, the log likelihood function $\ln \mathcal{L}(\boldsymbol{\theta})$ is given by

$$\ln \mathcal{L}(\boldsymbol{\theta}) = \ln P(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \Big(\sum_{\boldsymbol{Z}} P(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) \Big). \tag{2.15}$$

In this formulation, the latent variables are assumed to be discrete variables. However, the EM algorithm can also be formulated for continuous latent variables by replacing the sum over $\boldsymbol{Z}$ with an integral.

Since the summation over the latent variables appears inside the logarithm, the maximization of the log likelihood is intractable. Instead, the posterior distribution of the latent variables $P(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}})$ can be calculated from a current parameter estimate $\boldsymbol{\theta}^{\text{old}}$, to find the expected value of the log likelihood evaluated for the parameter value $\boldsymbol{\theta}$, corresponding to the E step of the EM algorithm. The expectation, denoted by $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{\text{old}})$, is given by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{\text{old}}) = \sum_{\boldsymbol{Z}} P(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln P(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}). \tag{2.16}$$

Subsequently, a revised parameter estimate $\boldsymbol{\theta}^{\text{new}}$ is obtained by maximizing this function with respect to $\boldsymbol{\theta}$ in the M step

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{\text{old}}). \tag{2.17}$$

Note that the maximization is now tractable because the logarithm acts directly on the joint distribution $P(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$ in $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{\text{old}})$.

$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{\text{old}})$ and its derivative can be shown to match the log likelihood $\ln \mathcal{L}(\boldsymbol{\theta}^{\text{old}})$ and its derivative, respectively, at the current estimate $\boldsymbol{\theta}^{\text{old}}$. Furthermore, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{\text{old}})$ can be shown to be always $\leq \ln \mathcal{L}(\boldsymbol{\theta}^{\text{old}})$. Consequently, each cycle of successive E and M steps increases the log likelihood until a local maximum is reached. In order to converge to the global maximum, the EM algorithm is sensitive to the initial parameter estimate $\boldsymbol{\theta_0}$. The general EM algorithm is summarized below.

Given a joint distribution $P(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$ over observed variables $\boldsymbol{X}$ and latent variables $\boldsymbol{Z}$, governed by parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood function $P(\boldsymbol{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

1. Choose an inital setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.

2. E step: Evaluate $P(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}})$.

3. M step: Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

   $$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{\text{old}}),$$

   where

   $$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{\text{old}}) = \sum_{\boldsymbol{Z}} P(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln P(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}).$$

4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

   $$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$$

   and return to step 2.

The EM algorithm can also be used to find maximum posterior (MAP) solutions for models in which a prior $P(\boldsymbol{\theta})$ is defined over the parameters. In this case, the E step remains the same as in the ML case. In the M step, the quantity to be maximized is given by $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}) + \ln P(\boldsymbol{\theta})$.

The EM algorithm is ideally suited for probabilistic motif discovery. As stated in the beginning of this Section, the exact position of a binding site within a longer sequence is often unknown. This missing information can be formalized by discrete latent variables, resulting in a formulation of the optimization problem that is suitable to be solved by the EM algorithm.

By utilizing the EM algorithm to learn PWM models (Section 2.1.1), motifs are generally required to be enriched in a positive sequence set in comparison to a negative sequence set. In the E step, the posterior probability of the binding site to start at a certain position in the sequence is calculated given the current model estimate. In the M step, the model parameters are improved by using the posterior distribution over the start positions obtained in the E step.

2.3.2. The EM algorithm for learning inhomogeneous interpolated Markov models

In lieu of the PWM model, I want to describe motifs by the more powerful iIMMs described in Section 2.1.3. Background sequences are modeled by IMMs. The following learning scheme is based on the EM algorithm (Section 2.3.1) and was developed to detect motifs enriched in a positive sequence set as compared to a negative sequence set.

I denote the set of all observed sequences by $\boldsymbol{X} = \boldsymbol{X}^1, \ldots, \boldsymbol{X}^N$ and assume the sequences to be of equal length L. Given an iIMM of width W and maximum order M, the set of all model parameters is denoted by $\boldsymbol{\rho}$, in which parameter $\rho_j(x_i|x_{i-M} \ldots x_{i-1})$ describes the probability to observe base x_i at model position j after observing the preceding bases x_{i-M} to x_{i-1} at model positions $j-M$ to $j-1$. The indices $i \in \{1, \ldots, L\}$ and $j \in \{0, \ldots, W-1\}$ indicate sequence and model positions, respectively. For simplicity, I understand oligomers to extend only over positive indices, e.g. $\rho_j(x_2|x_{2-M} \ldots x_1) = \rho_j(x_2|x_1)$. The set of all latent variables is denoted by $\boldsymbol{z} = z_1, \ldots, z_N$, with $z_n \in \{1, \ldots, L-W+1\}$ specifying the start position of the model, or $z_n = 0$ signifying no motif occurrence, in sequence $\boldsymbol{X}^n$, $n \in \{1, \ldots, N\}$. Finally, background sequences are modeled by an IMM $\boldsymbol{f}$ with order B.

The likelihood of all observed sequences $\boldsymbol{X}$ is given by the product over the likelihoods of observed sequences $\boldsymbol{X}^n$

$$P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\rho}) = \prod_{n=1}^{N} P(\boldsymbol{X}^n|z_n,\boldsymbol{\rho}), \tag{2.18}$$

which can be calculated for known latent variables and model parameters by

$$\begin{aligned} P(\boldsymbol{X}^n|z_n = k,\boldsymbol{\rho}) = \prod_{i=1}^{k-1} f(x_i^n|x_{i-B}^n \ldots x_{i-1}^n) \times \prod_{i=k}^{k+W-1} \rho_{i-k}(x_i^n|x_{i-M}^n \ldots x_{i-1}^n) \\ \times \prod_{i=k+W}^{L} f(x_i^n|x_{i-B}^n \ldots x_{i-1}^n). \end{aligned} \tag{2.19}$$

The prior probability of latent variables

$$P(z_n = k) = \begin{cases} 1-q, & \text{if } k = 0 \\ \frac{q}{L-W+1}, & \text{if } 1 \leq k \leq L-W+1, \end{cases} \tag{2.20}$$

is chosen to conform to a zero ($k = 0$) or one ($k \geq 1$) occurrence per sequence (ZOOPS) model, in which the hyperparameter q specifies the prior probability for a sequence to contain a motif.

For a current parameter estimate $\widetilde{\boldsymbol{\rho}}$, the posterior distribution of the latent variables can now be calculated by

$$P(z_n = k|\boldsymbol{X}^n,\widetilde{\boldsymbol{\rho}}) = \frac{P(\boldsymbol{X}^n|z_n = k,\widetilde{\boldsymbol{\rho}})P(z_n = k)}{\sum_{k'=0}^{L-W+1} P(\boldsymbol{X}^n|z_n = k',\widetilde{\boldsymbol{\rho}})P(z_n = k')}, \tag{2.21}$$

using Equations 2.19 and 2.20. $\widetilde{\boldsymbol{\rho}}$ corresponds to $\boldsymbol{\theta}^{\text{old}}$ in Section 2.3.1. To accelerate the computation of the posterior distribution, I divide the likelihood $P(\boldsymbol{X}^n|z_n = k,\widetilde{\boldsymbol{\rho}})$ by the constant likelihood $P(\boldsymbol{X}^n|\boldsymbol{f})$, the likelihood of observing sequence $\boldsymbol{X}^n$ given the background model $\boldsymbol{f}$, given by

$$P(\boldsymbol{X}^n|\boldsymbol{f}) = \prod_{i=1}^{L} f(x_i^n|x_{i-B}^n \dots x_{i-1}^n). \tag{2.22}$$

Consequently, the denominator in the resulting likelihood ratio

$$\frac{P(\boldsymbol{X}^n|z_n = k,\widetilde{\boldsymbol{\rho}})}{P(\boldsymbol{X}^n|\boldsymbol{f})} = \prod_{i=k}^{k+W-1} \frac{\widetilde{\rho}_{i-k}(x_i^n|x_{i-M}^n \dots x_{i-1}^n)}{f(x_i^n|x_{i-B}^n \dots x_{i-1}^n)}, \tag{2.23}$$

can be precomputed for all sequences and sequence positions.

In the Bayesian statistical framework, one seeks to find maximum posterior (MAP) solutions. To calculate the expected value of the log likelihood in the E step of the EM algorithm, I therefore need a distribution over the model parameters to use as prior knowledge. The pseudocounts scheme described in Section 2.1.1 corresponds to assuming a Dirichlet prior distribution with parameters $\alpha f(x_i)$ over the model probabilities. Note that larger α's produce tighter distributions. Similarly, I use a Dirichlet prior with pseudocount parameters $\alpha\rho_j^*(y_{M+1}|y_2 \dots y_M)$ over iIMM parameters $\boldsymbol{\rho}$

$$P(\boldsymbol{\rho}|\boldsymbol{\rho}^*) \propto \prod_{j=0}^{W-1} \left[\prod_{y_1 \cdots y_{M+1}} \rho_j(y_{M+1}|y_1 \dots y_M)^{\alpha\rho_j^*(y_{M+1}|y_2 \cdots y_M)} \right]. \tag{2.24}$$

Hence, the expected value of the log likelihood under the posterior distribution of the latent variables, $Q(\boldsymbol{\rho}|\widetilde{\boldsymbol{\rho}})$, can be calculated using the prior distribution $P(\boldsymbol{\rho}|\boldsymbol{\rho}^*)$ over model parameters $\boldsymbol{\rho}$, by

$$
\begin{aligned}
Q(\boldsymbol{\rho}|\widetilde{\boldsymbol{\rho}}) &= \mathbb{E}_{\boldsymbol{z}|\widetilde{\boldsymbol{\rho}}}[\ln P(\boldsymbol{X},\boldsymbol{z}|\boldsymbol{\rho})] + \ln P(\boldsymbol{\rho}|\boldsymbol{\rho}^*) \\
&= \sum_{n=1}^{N} \left[\sum_{k=0}^{L-W+1} P(z_n = k|\boldsymbol{X}^n,\widetilde{\boldsymbol{\rho}}) \ln P(X^n,z_n = k,\boldsymbol{\rho}) \right] + \ln P(\boldsymbol{\rho}|\boldsymbol{\rho}^*) \\
&= \sum_{n=1}^{N} \left[\sum_{k=1}^{L-W+1} P(z_n = k|\boldsymbol{X}^n,\widetilde{\boldsymbol{\rho}}) \Big(\ln(\frac{q}{L-W+1}) \right. \\
&\qquad + \sum_{i=1}^{k-1} \ln f(x_i^n|x_{i-B}^n \dots x_{i-1}^n) \\
&\qquad + \sum_{i=k}^{k+W-1} \ln \rho_{i-k}(x_i^n|x_{i-B}^n \dots x_{i-1}^n) \\
&\qquad + \sum_{i=k+W}^{L} \ln f(x_i^n|x_{i-B}^n \dots x_{i-1}^n) \Big) \\
&\quad \left. + P(z_n = 0|\boldsymbol{X}^n,\widetilde{\boldsymbol{\rho}}) \Big(\ln(1-q) + \sum_{i=1}^{L} \ln f(x_i^n|x_{i-B}^n \dots x_{i-1}^n) \Big) \right] \\
&\quad + \alpha \sum_{j=0}^{W-1} \left[\sum_{y_1 \cdots y_{M+1}} \rho_j^*(y_{M+1}|y_2 \dots y_M) \ln \rho_j(y_{M+1}|y_1 \dots y_M) \right]. \qquad (2.25)
\end{aligned}
$$

In the M step of the EM algorithm, $Q(\boldsymbol{\rho}|\widetilde{\boldsymbol{\rho}})$ needs to be maximized with respect to $\boldsymbol{\rho}$. The partial derivative of $Q(\boldsymbol{\rho}|\widetilde{\boldsymbol{\rho}})$ is subject to the constraint

$$
\sum_{y_{M+1}} \rho_j(y_{M+1}|y_1 \dots y_M) = 1 \qquad (2.26)
$$

and can be calculated by introducing the Lagrange multipliers $\lambda_{jy_1 \cdots y_M}$

$$
\begin{aligned}
\frac{\partial Q}{\partial \rho_j(y_{M+1}|y_1 \dots y_M)} &= \overbrace{\sum_{n=1}^{N} \sum_{k=1}^{L-W+1} P(z_n = k|\boldsymbol{X}^n,\widetilde{\boldsymbol{\rho}}) I(X_{k+j-M}^n \dots X_{k+j}^n = y_1 \dots y_{M+1})}^{n_j(y_1 \cdots y_{M+1})} \\
&\quad \times \frac{1}{\rho_j(y_{M+1}|y_1 \dots y_M)} + \alpha \frac{\rho_j^*(y_{M+1}|y_2 \dots y_M)}{\rho_j(y_{M+1}|y_1 \dots y_M)} \\
&= \lambda_{jy_1 \cdots y_M} \qquad (2.27)
\end{aligned}
$$

$$\begin{aligned}
\implies n_j(y_1 \ldots y_{M+1}) + \alpha\rho_j^*(y_{M+1}|y_2 \ldots y_M) &= \lambda_{jy_1\ldots y_M}\rho_j(y_{M+1}|y_1 \ldots y_M) \\
\overset{\sum_{y_{M+1}}}{\implies} \qquad n_{j-1}(y_1 \ldots y_M) + \alpha &= \lambda_{jy_1\ldots y_M} \\
\implies \rho_j(y_{M+1}|y_1 \ldots y_M) &= \frac{n_j(y_1 \ldots y_{M+1}) + \alpha\rho_j^*(y_{M+1}|y_2 \ldots y_M)}{n_{j-1}(y_1 \ldots y_M) + \alpha} \qquad (2.28)
\end{aligned}$$

There is no closed-form solution for the maximum of $Q(\boldsymbol{\rho}|\widetilde{\boldsymbol{\rho}})$. Therefore, the question now is what pseudocounts $\rho_j^*(y_{M+1}|y_2 \ldots y_M)$ to choose. I propose a plausible recipe to iteratively update the pseudocounts parameters according to the same Equation 2.28 for a lower order, e.g.

$$\rho_j^*(y_{M+1}|y_2 \ldots y_M) = \frac{n_j(y_2 \ldots y_{M+1}) + \alpha\rho_j^*(y_{M+1}|y_3 \ldots y_M)}{n_{j-1}(y_2 \ldots y_M) + \alpha} \tag{2.29}$$

and so forth down to

$$\rho_j^*(y_{M+1}) = \frac{n_j(y_{M+1}) + \alpha f(y_{M+1})}{N + \alpha}. \tag{2.30}$$

The integration of order-dependent αs, as described in Equation 2.10, is straightforward but was omitted for clarity.

In order to initialize iIMMs in the EM algorithm, I integrated code of XXmotif, a previously developed ab initio motif finder that performed superior to related tools on various data sets (Hartmann et al., 2013). Hence, motif instances that XXmotif used to build its PWM models can be employed to initialize iIMMs. iIMMs can also be directly initialized from user-specific binding site instances or iIMMs.

2.3.3. The EM algorithm for learning inhomogeneous interpolated Markov models from weighted sequences

In addition to the sequences measured to be bound by the investigated transcription factor, several experiments provide information on binding strength or confidence in binding. For instance, PBM experiments provide fluorescence signal intensities that reflect binding affinities. ChIP-seq experiments commonly report peak statistics, such as P-values, as a measure of statistical significance. This information can be leveraged to calculate sequence weights $\boldsymbol{w}$. By using a simple heuristic, I incorporate sequence weights into Equation 2.18 and maximize the resulting weighted likelihood

$$P(\boldsymbol{X}|\boldsymbol{z},\boldsymbol{\rho},\boldsymbol{w}) = \prod_{n=1}^{N} P(\boldsymbol{X}^n|z_n,\rho)^{w_n}. \tag{2.31}$$

Consequently, the partial derivative of $Q(\boldsymbol{\rho}|\widetilde{\boldsymbol{\rho}})$ (Equation 2.27) changes to

$$\begin{aligned}\frac{\partial Q}{\partial \rho_j(y_{M+1}|y_1 \ldots y_M)} &= \overbrace{\sum_{n=1}^{N} w_n \sum_{k=1}^{L-W+1} P(z_n = k|\boldsymbol{X}^n, \widetilde{\boldsymbol{\rho}}) I(X^n_{k+j-M} \ldots X^n_{k+j} = y_1 \ldots y_{M+1})}^{n_j(y_1 \cdots y_{M+1})} \\ &\quad \times \frac{1}{\rho_j(y_{M+1}|y_1 \ldots y_M)} + \alpha \frac{\rho^*_j(y_{M+1}|y_2 \ldots y_M)}{\rho_j(y_{M+1}|y_1 \ldots y_M)} \\ &= \lambda_{j y_1 \cdots y_M}. \end{aligned} \tag{2.32}$$

For uniform weights $w_n = 1$ Equations 2.18 and 2.27 are recovered.

Sequence weights can be calculated from various sources. For instance, provided with fluorescence signal intensity measurements $\boldsymbol{I}$, a weight w_n can be assigned to sequence $\boldsymbol{X}^n$ using

$$w_n = \frac{I_n - I_{bg}}{I_{max} - I_{bg}}, \tag{2.33}$$

where I_n is the intensity of sequence $\boldsymbol{X}^n$, I_{max} is the intensity of the sequence with highest intensity, and I_{bg} is the intensity of the quantile at which no specific, but solely background signal, is expected.

Similarly, weights can be calculated from sequence ranks $\boldsymbol{R}$

$$w_n = \frac{N + 1 - R_n - (N + 1 - R_{bg})}{N - (N + 1 - R_{bg})}, \tag{2.34}$$

where R_n is the rank of sequence $\boldsymbol{X}^n$, and R_{bg} corresponds to the rank at which sequences are expected to show background signal only. Sequence ranks may e.g. be inferred for sequences with assigned P-value.

2.4. Data sets and benchmark tests

This Section describes the sources of the data sets used in the application and evaluation of GIMMEmotif, as well as details about all processing and analysis steps performed.

2.4.1. ChIP-seq data sets

I evaluated GIMMEmotif on human transcription factor ChIP-seq data sets published by The ENCODE Project Consortium (2012). The encyclopedia of DNA elements (ENCODE) March 2012 data freeze comprises 708 IDR optimal blacklist-filtered SPP (Kharchenko et al., 2008) peak sets. The irreproducible discovery rate (IDR) framework verifies the reproducibility of ChIP-seq peaks identified from replicate experiments by computing a quantitative reproducibility score (Li et al. (2011), Landt et al. (2012)). Peaks that overlap blacklisted regions were removed. These regions were empirically identified by the ENCODE Data Analysis Consortium (DAC) to show anomalous unstructured high signal in next-generation sequencing experiments independent of cell line and experiment type. The analysis was restricted to 87 RNA polymerase (RNAP) II-associated sequence-specific transcription factors characterized by Wang et al. (2012) (441 data sets) and nine additional sequence-specific transcription factors (ATF1, ATF2, Elk1, FoxM1, IRF4, SREBP2, STAT5A, TCF3, ZnF217) from subsequently conducted experiments (13 data sets).

Positive sequences were compiled from the top 5,000 peak regions (sorted best to worst according to their signal value) or all peak regions if less than 5,000 peaks are available. Sequences were extracted ±100 bp around peak summits using Biopieces (www.biopieces.org). Negative sequences were sampled from the trimer frequencies observed in positive sequences to ensure similar sequence compositions in both sequence sets. The length and number of negative sequences was the same as the length and 100 times the number of positive sequences, respectively.

In order to initialize iIMMs, I ran XXmotif (Hartmann et al., 2013) using the non-default options --revcomp, --localization, --localization-ranking, --background-model-order 2 and --merge-motif-threshold LOW and filtered the results by requiring motifs to lie localized to peak summits (--maxPvalue 0.05) and to occur in at least 5% of sequences (--minOccurrence 0.05). The motif instances that XXmotif used to calculate its top ranked PWM in each data set were employed to initialize iIMMs. Optionally, two or four uniformly initialized positions were added to both 5′- and 3′-ends of the models. The search space in training and test sequences was guaranteed to be identical and independent of the number of model positions. For instance, in order to compare the performance of models that describe the binding sites of the same transcription factor in the same data set but differ in their number of positions, I adjusted the search space in the benchmark test by extending training and test sequences of the longer model accordingly. In 446 (of all 454) data sets, corresponding to 94 transcription factors, XXmotif found at least one motif in all four cross-validation folds, independent of the length of training sequences. The remaining eight peak sets of the transcription factors c-Myc (2 data sets), E2F1 (1), ELF1 (1), PAX5 (1), PGC1A (1), and STAT1 (2) were excluded from the benchmark test.

To assess the performance of XXmotif, iMMs, and iIMMs in discriminating bound from unbound sequences, I carried out a fourfold cross-validation, that is, I trained on 75% of data, tested on holdout 25%, and pooled results over four holdout data sets. In the process, I calculated the maximum log-odds score over all possible motif positions for each positive and negative test sequence and evaluated the partial area under the receiver operating characteristic (ROC) curve (pAUC) up to a false positive rate (FPR) of 5%. Since the pAUC summarizes the part of the ROC curve which is most relevant to practical applications, it is preferable over the area under the entire ROC curve (AUC).

To evaluate the ability of iIMMs to predict *in vitro* binding affinities measured by competitive electrophoretic mobility shift assay (EMSA) for the mouse embryonic stem cell (mESC) transcription factor Klf4, I learned models using *in vivo* ChIP-seq data from Chen et al. (2008). Positive sequences were compiled from the top 5,000 peak regions (sorted best to worst according to their signal value) by extracting ±50 bp around peak region midpoints. To initialize iIMMs, I ran XXmotif with the parameter setting used in the ENCODE benchmark test (see above).

2.4.2. EMSA data sets

I used the competitive EMSA experiments for the mESC transcription factor Klf4 from Sun et al. (2013), comprising dissociation constant (K_d) measurements for 33 sequences with single mutations and 25 sequences with multiple mutations to the ten bp consensus binding site of Klf4. These dissociation constants were divided by the dissociation constant of the sequence with median K_d (single mutant sequences) or with K_d closest to the mean K_d (multiple mutant sequences), and the logarithms of the resulting ratios were calculated. Prediction scores are provided as log ratios of odds ratios, in which odds ratios are computed from the probabilities returned by the Klf4 and the background model. I compared Klf4 iIMMs of increasing order by means of the Pearson correlation between measured and predicted log ratios.

The competitive EMSA scores for 64 double-stranded oligonucleotide probes containing a potential FoxA2 binding site were taken from Levitsky et al. (2014). I calculated Spearman correlations between measured EMSA scores and log ratios predicted by FoxA2 iIMMs of increasing order. Spearman correlations to predictions from other methods and models were determined in Alipanahi et al. (2015).

2.4.3. Genomic-context PBMs

To test iIMM learning from weighted sequences (Section 2.3.3), I make use of genomic-context protein-binding microarray (gcPBM) measurements of the *S. cerevisiae* bHLH transcription factors Cbf1 and Tye7 (Gordân et al., 2013). Gordân et al. (2013) compiled 280 (Cbf1) and 312 (Tye7) 30-bp-long probes from ChIP-chip bound and ChIP-chip unbound sequences centered at the E-box CACGTG (Harbison et al., 2004). This allowed for the modeling of binding sites including up to 12-bp-long flanking regions (ensured not to contain another binding site). I calculated sequence weights from gcPBM signal intensities according to Equation 2.33, where I_{bg} was fixed to the mean signal intensity of 674 (Cbf1) and 621 (Tye7) 30-bp-long negative control probes (Gordân et al., 2013). I computed Pearson correlations between measured log signal intensities and predicted log-odds scores by applying a tenfold cross-validation, as done by Mordelet et al. (2013).

2.4.4. RNAP I/II core promoter sequences

I analyzed sequences around *Drosophila melanogaster* transcription start sites (TSSs) measured by Brown et al. (2014) using cap analysis of gene expression (CAGE) (Shiraki et al., 2003). Filtered bedGraph-formatted CAGE data sets were pooled. Before clustering TSSs, the genomic distribution of TSSs was smoothed using a 41 bp uniform kernel function. Clusters were defined by genome intervals in which the smoothed distribution of TSS counts was found to lie entirely above the genome-wide average. The mode of the distribution was used as the representative TSS in each cluster. Subsequently, the clusters were filtered by three criteria. First, clusters with less than five TSS counts were excluded. Second, clusters had to exhibit more TSS counts compared to any other cluster within 150 bp (regarding representative TSSs). Third, clusters had to lie close to FlyBase-annotated TSSs (dos Santos et al., 2015), by requiring a representative TSS to be located within 250 bp upstream of an annotated TSS or within a $5'$ untranslated region (UTR), or the cluster interval to contain an annotated TSS. The clustering resulted in 15,971 TSSs assigned to 11,536 unique genes. Genes can thus be regulated by multiple core promoters defined by distinct TSSs.

In order to assign TSS clusters to a broad and narrow transcription-initializing core promoter class, the peakedness of the TSS distributions was quantified with a TSS width score by calculating the mean absolute deviation from the median TSS location as

$$\text{TSS width} = \frac{1}{N}\sum_{i=1}^{N} \left| x_i - \text{median}(X) \right| , \tag{2.35}$$

where N is the number of TSSs within the cluster, x_i is the position of the i^{th} TSS, and $\text{median}(X)$ is the median TSS position within the cluster. The distribution of TSS widths shows a local minimum at a value of five. Therefore, clusters with a TSS width smaller than five were classified as narrow peak (NP) and the remaining as broad peak (BP) core promoters. This resulted in 7,262 NP and 8,709 BP core promoters assigned to 5,576 and 7,235 genes, respectively. Note that 1,275 genes have core promoters from both classes.

In addition to NP and BP core promoters, I modeled core promoter sequences of ribosomal protein (RP) genes, which are known to differ from NP and BP core promoters in their architecture (Lenhard et al., 2012). The RPG database (Nakao et al., 2004) maintains 87 RP genes from *D. melanogaster*. Except for RpS27A, I could assign at least one core promoter to each RP gene, six of which had two associated core promoters. I thus obtained 92 RP gene core promoters, 60 and 32 belonging to the NP and BP class, respectively. Note that RP gene core promoters were not excluded from NP and BP core promoter sequences.

The core promoter encompasses the region that lies approximately ±50 bp around the TSS (Duttke et al., 2015). Therefore, I initialized core promoter models of all three classes from 101-bp-long promoter sequences centered at their representative TSSs. The models were learned by allowing the EM algorithm to realign core promoters. The 9^{th} percentile of TSS width scores was computed for each core promoter class, resulting in 3.64 (NP), 22.71 (BP), and 10.64 (RP). Hence, I learned NP, BP, and RP gene core promoters within 4 bp, 23 bp, and 11 bp using positive sequences of length 109 bp, 147 bp, and 123 bp, respectively, centered at their representative TSSs. Negative sequences were sampled from the frequencies of trimers within 250 bp of representative TSSs. The length and number of NP, BP, and RP negative sequences was the same as the length and 100 times the number of NP, BP, and RP positive sequences, respectively.

I assessed the performance of iMMs and iIMMs in predicting TSS locations using a four-fold cross-validation procedure. To calculate the precision (fraction of true in all predictions) of the models in predicting the correct positions of TSSs, I determined the position with highest log-odds score in each test sequence, extracted ±250 bp around representative TSSs. If the position was within 4 (NP), 23 (BP), and 11 (RP) bp of the representative TSS, the prediction was judged as correct, else as false. The window size of each class corresponds to the 9$^{\text{th}}$ percentile of its TSS cluster width scores (see above). Notably, while test sequences provide an identical search space (401) for all core promoter classes, the precision of random predictions is different for NP (0.02), BP (0.12), and RP (0.06) core promoters. I picture the distributions of maximum log-odds score positions, that is, the predictions of signal locations, as enrichments compared to predictions from a random predictor.

2.4.5. RNAP II polyadenylation site sequences

I use the major transcript isoform (mTIF) annotations from Pelechano et al. (2013), obtained after clustering transcript isoforms (TIFs) from *S. cerevisiae* grown in yeast extract peptone dextrose (YPD). After selecting mTIFs covering one intact open reading frame (ORF) and summing up the sequencing reads of mTIFs with identical polyadenylation (pA) site, I selected the pA site(s) with the maximum number of sequencing reads per gene. I excluded pA sites with less than five sequencing reads. In total, I selected 4,228 pA sites from 4,173 distinct genes. 51 and 2 genes are represented by two and three pA sites, respectively.

The sequence region that surrounds pA sites shows nucleotide preferences within 70 bp upstream to 30 bp downstream of pA sites. Therefore, I modeled pA sites over the length of 101 bp covering this region. To provide a biologically relevant length of test sequences, I determined the length of 3′ UTRs from measured pA sites and *S. cerevisiae* ORF annotations from the Saccharomyces Genome Database (Cherry et al., 2012). Since more than 90% of 3′ UTRs are shorter than 300 bp, test sequences were extracted from 220 bp upstream to 180 bp downstream of pA sites. This corresponded to 301 potential pA site positions within 150 bp of measured pA sites, from which the correct pA site is to be predicted in the benchmark tests. Similarly, I calculated trimer frequencies from 220 bp upstream to 180 bp downstream of pA sites to sample 101-bp-long negative sequences, required in the EM algorithm. The number of negative sequences corresponded to 100 times the number of positive sequences.

Pelechano et al. (2013) defined mTIFs by clustering the transcripts with each of their 5′- and 3′-end sites co-occurring within five bp. On this account, I determined the precision of pA site predictions by considering predictions within five bp of measured pA site locations as correct. Hence, the precision of random pA site predictions would be 0.04. In other respects, the benchmark test is identical to the evaluation procedure performed for core promoter sequences (Section 2.4.4).

2.4.6. RNAP pause site sequences

Larson et al. (2014) measured 19,960 and 9,989 RNAP pause sites in *Escherichia coli* and *Bacillus subtilis*, respectively, using nascent elongating transcript sequencing (NET-seq), and found approximately one pause site per 100 bp across well-transcribed genes on average. I extracted test sequences that correspond to a search space of 101, centered at the pause sites. To prevent overtraining caused by overlapping training and test sequences, I excluded pause sites within 54 bp of another pause site with higher relative peak height. This reduced the number of *E. coli* and *B. subtilis* pause sites to 11,648 and 6,809, respectively. Negative sequences were randomly sampled from the *E. coli* and *B. subtilis* genomes using the NCBI Reference Sequence (RefSeq) accession numbers NC_000913.3 and NC_000964.3, respectively, totaling to 100 times the number of positive sequences.

In *E. coli*, Larson et al. (2014) identified a 16-bp-long consensus pause sequence, ten bp upstream to five bp downstream of the pause index (the 3′-end of the transcript). I additionally incorporated the two bp immediately flanking the identified 16-bp-long consensus and learned the resulting 20-bp-long models by varying the model order. Likewise, I learned 20-bp-long iIMMs of *B. subtilis* RNAP pause sites. Longer models did not further improve benchmark test results.

Except for considering pause sites predicted to lie within zero bp of measured sites to be correct, a random prediction could therefore locate pause sites with a precision of 0.01, I resort to the benchmark test described for core promoter and pA site sequences (Sections 2.4.4 and 2.4.5).

2.4.7. PAR-CLIP data sets

I modeled the binding of 25 messenger ribonucleoprotein (mRNP) biogenesis factors from *S. cerevisiae* to mRNA using published PAR-CLIP data sets (Baejen et al. (2014), Schulz et al. (2013)). After sorting PAR-CLIP crosslink sites by occupancies (number of uracil to cytosine base transitions over RNA-seq counts) and excluding crosslink sites located in tRNA transcripts, I focused on the top 2,000 protein-RNA crosslink sites, which correspond to uracil nucleosides. To learn and test crosslink site iIMMs, I extracted 25-nt-long positive sequences encompassing the central crosslink site. In order to learn to discriminate factor binding sites in the transcriptome, 20,000 uracil-centered sequences (of the same length) were randomly sampled from the *S. cerevisiae* transcriptome using mRNA annotations from Pelechano et al. (2013) and employed as negative sequences both in learning and testing the models of all RNA-binding proteins. Note that negative sequences may also contain true RNA-binding motifs.

I assessed the performance of iMMs and iIMMs in discriminating between uracil nucleosides with and without crosslink analogous to the evaluation procedure conducted in the ENCODE benchmark test (Section 2.4.1).

2.4.8. Parameter setting in benchmark tests

In all applications, I used the following parameter specifications. The pseudocounts factor of iIMMs (Equation 2.10) was set to $\alpha_0 = 1$ and $\alpha_k = 20 \times 3^{k-1}$, for $k > 0$, in order to be rather conservative. With this choice of α_k, higher-order probabilities resort to lower-order probabilities with the benefit of avoiding overtraining of higher-order models. To compare with iIMMs, the pseudocounts factor of iMMs was set to $\alpha_0 = 1$ and $\alpha_k = 5$, for $k > 0$. This choice provided the best overall performance of iMMs. As background models, I used 2nd-order IMMs with pseudocounts factor $\alpha_k = 10$, for all $k \geq 0$, the default pseudocounts parameter value in XXmotif. The hyperparameter q specifies the prior probability for a sequence to contain a motif. In order to account for false positive experimental measurements, q was fixed at 0.9 for all data sets.

3. Results and Discussion

I employ iIMMs of increasing order to model protein-DNA (Section 3.1) and protein-RNA (Section 3.2) interactions and compare their performance to iMMs in diverse benchmark tests. In the following, I refer to 0^{th}-order iMMs and iIMMs as PWM models.

3.1. Protein-DNA binding specificity models

I test iIMM learning on DNA sequence sets ranging from binding sites of single transcription factors (Section 3.1.1), including RNA polymerase pause sites (Section 3.1.4), to complex regulatory regions typically bound by multiple cooperatively binding proteins: core promoter regions (Section 3.1.2) and polyadenylation sites (Section 3.1.3).

3.1.1. Transcription factor binding specificity models

I learn and test DNA binding models of transcription factors using three approaches. First, I learn models from *in vivo* binding sites and evaluate their performance in discriminating bound from unbound sequences (Section 3.1.1.1). Second, I examine whether models learned *in vivo* are capable of predicting binding affinities measured *in vitro* (Section 3.1.1.2). Last, I learn and predict *in vitro* DNA binding (Section 3.1.1.3).

3.1.1.1. Modeling nucleotide interdependencies within binding sites

I start evaluating GIMMEmotif by learning models of human transcription factors using ChIP-seq peak sets compiled from ENCODE (The ENCODE Project Consortium, 2012), by examining up to 5,000 peaks with highest confidence. iIMMs are initialized from the binding site instances that XXmotif used to build its most significant PWM model, with two uniformly initialized positions added to both $5'$- and $3'$-ends of iIMMs. To compare the performance of iIMMs of increasing model order in discriminating between bound and unbound sequences, I evaluated the partial area under the ROC curve (pAUC) up to a false positive rate of 5% (Figure 3.1A, inset) for each model and calculated the cumulative distribution over all 446 ChIP-seq data sets.

Modeling nucleotide interdependencies within core binding sites

Figure 3.1A demonstrates that higher-order iIMMs improve the performance of PWM models (Figure 3.1A, dark blue). While the most successful models are the iIMMs with highest

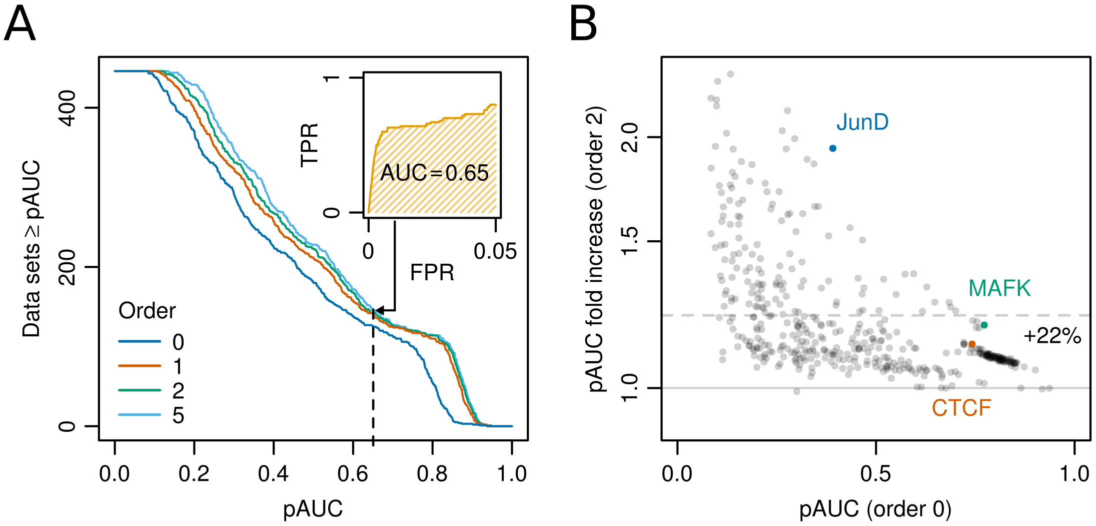


Figure 3.1.: **Core binding site iIMMs consistently outperform PWM models.** **(A)** Cumulative distributions of the partial area under the ROC curve (pAUC), up to a false positive rate of 5% (inset), over all 446 ChIP-seq data sets, using iIMMs of increasing order. **(B)** Scatter plot of the increase in performance, when increasing the model order from zero to two. The y-axis is shown in log scale. The dashed line indicates the mean fold increase.

order (Figure 3.1A, light blue), most of the improvement is already covered by 1$^{\text{st}}$-order iIMMs (Figure 3.1A, orange). Higher-order iIMMs appear to be especially beneficial in learning models from more challenging data sets, that is, data sets in which PWM models achieve low pAUC values.

The overall tendency to increase performance by increasing the model order is also reflected in the results for single transcription factor data sets, as illustrated in Figure 3.1B, comparing the pAUC values of PWM models and 2$^{\text{nd}}$-order iIMMs for each transcription factor data set. Importantly, iIMMs outperform PWM models in almost all data sets ($P = 2.5 \times 10^{-132}$, Wilcoxon one-sided signed-rank test, $n = 446$).

I exemplify the source of the performance benefit when using iIMMs by showing precision-recall curves and higher-order sequence logos for models of three transcription factors (Figure 3.2). The well-studied CCCTC-binding factor (CTCF) has been implicated in the establishment of topologically associating domains (TADs) and the formation of regulatory chromosome interactions within TADs (Ong and Corces, 2014), and its DNA binding sites have been identified to be mutational hotspots in the genomes of multiple cancer types (Katainen et al., 2015). To gain an in-depth understanding of its binding characteristics, precise binding site models are crucial.

Precision-recall curves (Figure 3.2A, left) indicate that the incorporation of dependencies between neighboring nucleotides results in more accurate CTCF models, learned in the Mcf-7 breast cancer cell line. The improved performance can be ascribed to positional interdependencies in several regions of the binding site (Figure 3.2A, right), parts of which

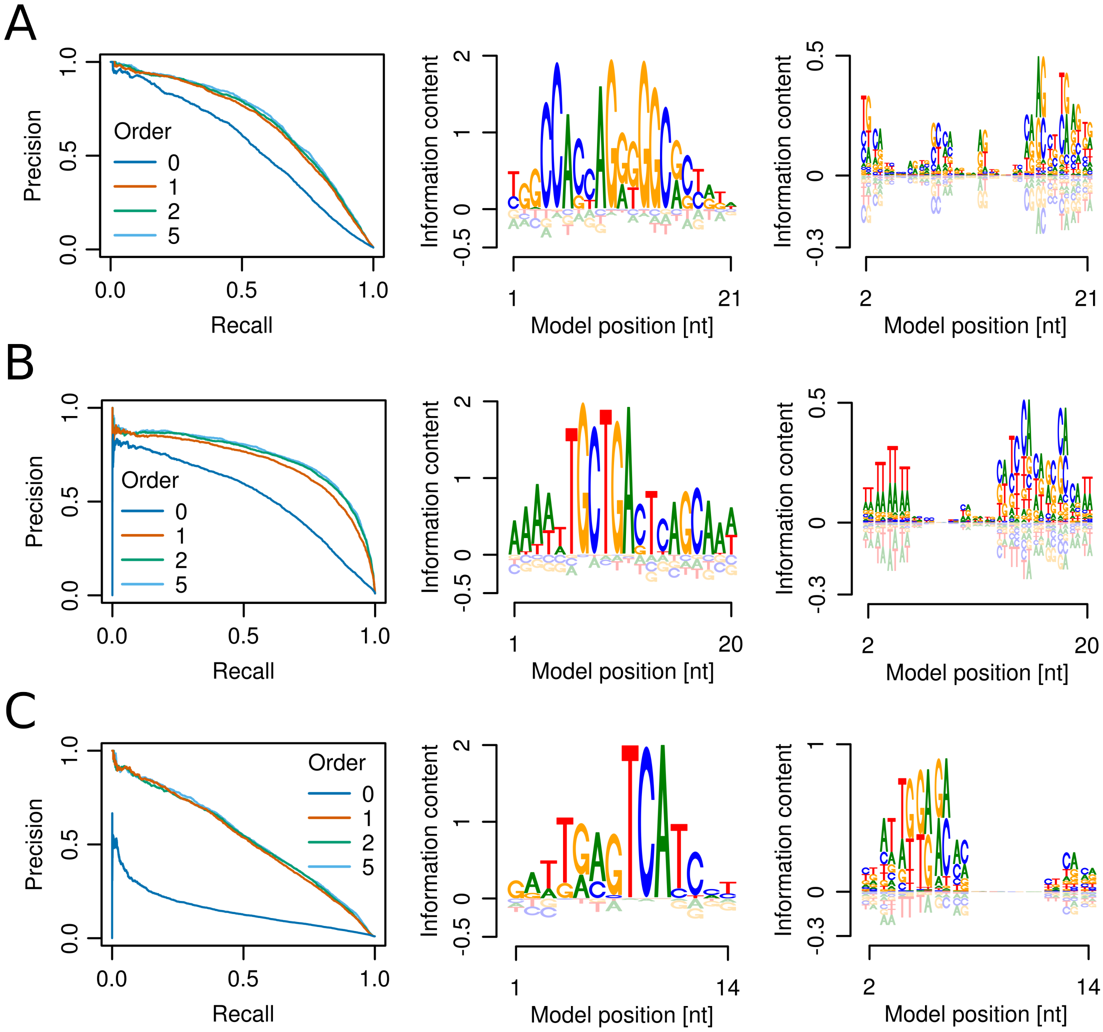


Figure 3.2.: **Core binding site iIMMs: examples. (A)** CTCF models learned in Mcf-7 cells. Precision-recall curves (left) calculated using iIMMs of increasing order. 0th-order (middle) and 1st-order (right) sequence logos depict 2nd-order iIMM. Sequence logos show CTCF model learned from all sequences. **(B)** Same as **A** but showing MAFK models learned in HepG2 cells. **(C)** Same as **A** but showing JunD models learned in HepG2 cells.

have been previously reported (Eggeling et al. (2014), Narlikar (2013)). For example, the base preferred at model position 17 is more probable to be a G in case an A is present at model position 16. In contrast, given a G at position 16, there is a higher probability of seeing a C at model position 17. This and further 1st-order dependencies, which are not evident in the standard sequence logo of order zero (Figure 3.2A, middle), may reflect the intricate interplay of a subset of its 11 zinc-finger (ZnF) domains.

The small MAF transcription factor MAFK, which belongs to the AP-1 family of basic-region leucine zipper (bZIP) proteins, targets a long palindromic sequence referred to as the MAF recognition element (MARE). MAF transcription factors are capable of accomplishing binding to two alternative MAREs: the 13-bp-long T-MARE and the 14-bp-long C-MARE, composed of a seven-bp-long TRE and a eight-bp-long CRE core sequence, respectively, which are flanked by GC elements on both sides (Kurokawa et al., 2009). Furthermore, binding to only one MARE half-site was reported, but requires a 5′-flanking AT-rich region (Yoshida et al., 2005). The alternative recognition modes depend on MAF homodimer and MAF-containing heterodimer formation.

Figure 3.2B (left) shows the performance of MAFK models learned in the Hep G2 liver carcinoma cell line. The substantial improvement of higher-order iIMMs stems from the integration of the described alternative DNA recognition modes into one model. While the T-MARE is primarily modeled in order zero of the 2^{nd}-order iIMM (Figure 3.2B, middle), the C-MARE is modeled via 1^{st}-order dependencies (Figure 3.2B, right). This indicates that the T-MARE is the predominant sequence element bound by MAFK in Hep G2 cells. Moreover, the upstream AT-rich region is also part of the model (Figure 3.2B, middle). The sequence logo of order one reveals that, in fact, this region consists of a poly(dA:dT) tract, which has also been reported for MAFG (Jolma et al., 2013). This stretch of A or T bases narrows the DNA minor groove, a structural feature that can be specifically read by DNA shape-sensitive proteins (Rohs et al., 2009). For instance, Gcn4 binding was enhanced *in vitro* by placing a poly(dA:dT) tract immediately adjacent to its core binding site (Levo et al., 2015). Besides this direct effect on transcription factor binding affinity, a nucleosome-mediated effect (Raveh-Sadka et al., 2012) seems to be relevant *in vivo*.

Similar to MAFK, the bZIP transcription factor JunD can bind to two distinct motifs, which differ in the length (one or two bp) of the spacer that separates the binding half-sites. The stoichiometry was shown to be cell-type-specific, with the primary motif selected by JunD depending on the cell-type-specific availability of oligomerization partners (Arvey et al., 2012).

In contrast to the PWM model, higher-order iIMMs can represent both binding modes prevalent in Hep G2 cells, resulting in the performance gain observed in the precision-recall curves (Figure 3.2C, left). The sequence logos of order zero (Figure 3.2C, middle) and one (Figure 3.2C, right) precisely display how the 2^{nd}-order iIMM models the variable juxtaposition of binding half-sites. Clearly, a two-component mixture model seems to be the more natural choice in this case. However, a more flexible model that is suitable for a broad range of binding motifs and regulatory sequences without prior knowledge is needed.

I provide further precision-recall curves and higher-order sequence logos for the following transcription factors in Figure A.1: BATF, c-Jun, c-Fos, HNF4α, IRF4, NF-YB, NRSF, PU.1, and ZnF143. Remarkably, I observe substantial gains in performance using iIMMs of order three or higher for some transcription factors, e.g. ZnF143 (Figure A.1I).

Incorporating nucleotides flanking core binding sites

It has been demonstrated that nucleotides flanking the core binding site add to the binding specificity of transcription factors (Gordân et al. (2013), Levo et al. (2015)). Confirming

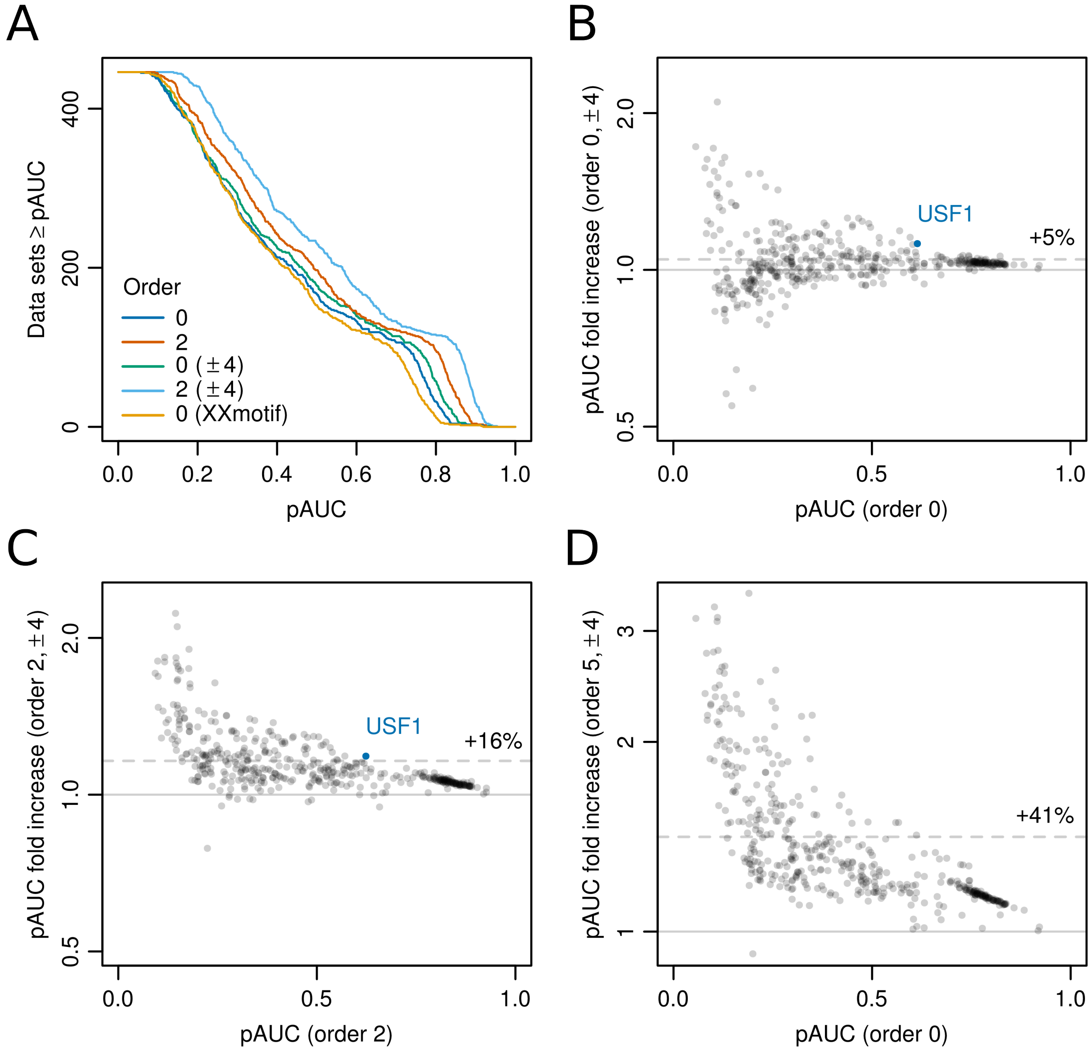


Figure 3.3.: **The impact of nucleotides flanking core binding sites. (A)** Cumulative distributions of the partial area under the ROC curve (pAUC), up to a false positive rate of 5%, over all 446 ChIP-seq data sets, using XXmotif and iIMMs of different order and size, that is, with and without adding four (±4) positions to each side of the core binding site model. **(B)** Scatter plot of the increase in performance, when extending 0^{th}-order iIMMs by adding four positions to each side (±4) of the core binding site model. The y-axis is shown in log scale. The dashed line indicates the mean fold increase. **(C)** Same as **B** but showing the increase in performance, when extending 2^{nd}-order iIMMs. **(D)** Same as **B** but showing the combined impact of increasing the model order and adding flanking nucleotides, comparing a PWM model with a 5^{th}-order iIMM extended by four additional positions on each side (±4).

these insights, core binding site surrounding sequence preferences can contribute considerably to the overall information content contained in some of the transcription factor models highlighted in Figures 3.2 and A.1.

To systematically investigate this issue, I extended the length of initial iIMMs, as determined by XXmotif, by adding four uniformly initialized positions to both sides of the models. The search space was adjusted to allow for an unbiased comparison between models of differing size.

Both PWMs and iIMMs that comprise flanking nucleotides (Figure 3.3A, green and light blue) outperform core binding site PWMs and iIMMs (Figure 3.3A, dark blue and orange), respectively, in discriminating bound from unbound sequences. Evidently, the improvement is substantially greater for iIMMs, while core binding site iIMMs already outperform elongated PWM models. Hence, the modeling of nucleotide interdependencies better brings to bear the impact of flanking nucleotides.

Importantly, the performance of PWM models determined by GIMMEmotif (Figure 3.3A, dark blue) is comparable to the performance of PWM models constructed by XXmotif (Figure 3.3A, yellow). Thus, the increase in performance using higher-order models is not the result of non-optimal PWM models learned by GIMMEmotif.

Furthermore, the length distribution of PWMs reported by XXmotif (Figure A.2) shows similar characteristics compared to that of publicly deposited PWM models. Consequently, the influence of flanking nucleotides on model performance is not caused by learning unusually short core binding site models.

While Figures 3.3B ($P = 1.3 \times 10^{-32}$, Wilcoxon one-sided signed-rank test, $n = 446$) and 3.3C ($P = 1.1 \times 10^{-119}$) validate the performance trends observed in Figure 3.3A for the vast majority of data sets, Figure 3.4 shows exemplary results for 2$^{\text{nd}}$-order iIMMs of the basic helix-loop helix (bHLH) family transcription factor USF1, learned in H1 human embryonic stem cells (H1-hESCs), which differ in the number of model positions. Despite the apparently low contribution of the eight flanking positions to the information content of the elongated model in the 0$^{\text{th}}$ order (Figure 3.4, right), the additional information accumulating over all model orders leads to a huge impact on the performance (Figure 3.4, left). Note that the core binding site turns up to be identical in both models (Figure 3.4, middle). This influence of flanking nucleotides on the binding specificity has also been demonstrated for other bHLH transcription factors (Gordân et al. (2013), Mordelet et al. (2013)), including Cbf1, the USF1 homolog in *S. cerevisiae*. Further examples of precision-recall curves and higher-order sequence logos can be found in Figure A.3 for the transcription factors GR, IRF1, and c-Fos.

To give an impression of the performance boost that can be achieved by modeling both higher-order nucleotide interdependencies and nucleotides flanking the core binding site, I compare the performance of 0$^{\text{th}}$-order non-elongated PWM models with 5$^{\text{th}}$-order, eight-bp-elongated iIMMs for single transcription factor data sets (Figure 3.3D). Strikingly, iIMMs outperform the starting point models (mean/median pAUC increase = 41/26%, $P = 5.2 \times 10^{-134}$) in all but one data set (of the estrogen receptor α in ECC-1 cells when treated with estradiol).

The data set that shows the greatest difference in the pAUC when comparing a non-elongated PWM model to a 5$^{\text{th}}$-order, eight-bp-elongated iIMM belongs to transcription

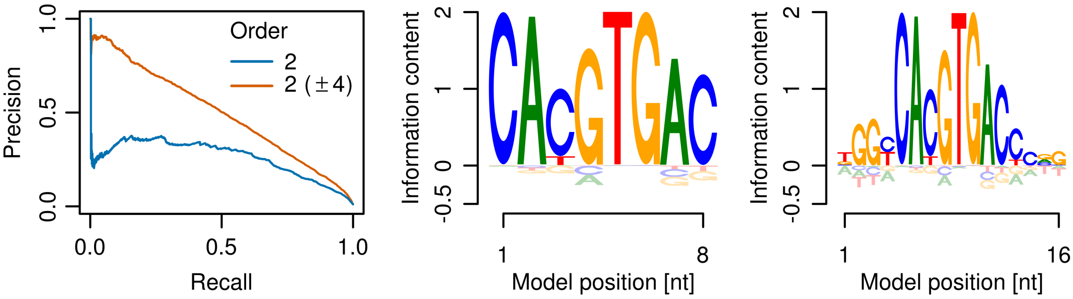


Figure 3.4.: **The impact of nucleotides flanking core binding sites: example.** USF1 models learned in H1-hESCs. Precision-recall curves (left) calculated using 2nd-order iIMMs that differ in size by four positions flanking the core binding site on each side (±4). The 0th-order sequence logo depicts 2nd-order (middle) and extended 2nd-order (right) iIMMs. Sequence logos show USF1 models learned within first cross-validation fold.

factor c-Fos measured in K562 cells. Since two motifs, the tetradecanoylphorbol acetate (TPA) response element TGACTCA as well as the CCAAT box, are similarly enriched around peak summits, iIMMs represent both motifs within one model. While higher-order iIMMs can distinguish both motifs via dependencies between neighboring nucleotides, the PWM model fails.

As expected, iIMMs appear to be less prone to overfitting, overall and in single data sets, compared to iMMs (Figure A.4). Interestingly, however, complex iMMs perform superior to PWM models (Figure A.4A), probably due to the sufficiently high number of binding site instances provided by most ChIP-seq data sets.

Finally, I revisit the CTCF models learned in Mcf-7 cells (Figures 3.1B and 3.2A). Nakahashi et al. (2013) found the 11 ZnF domains (ten of the C_2H_2 and one of the C_2HC class) of CTCF to contact DNA elements upstream and downstream of the core CTCF binding site, with elements separated by variable-length spacer sequences. To capture this complexity, I extended CTCF iIMMs of increasing order by adding 25 positions to each side of initial 17-bp-long models, thus obtaining 67-bp-long models. While the performance of the PWM model largely persists, elongated iIMMs further increase classification performance (Figures A.5A and A.5B).

Furthermore, to explore the complete binding site space of CTCF, I trained and tested the 50-bp-elongated models on all 66,592, instead of the top 5,000, ChIP-seq peaks. Remarkably, a 5th-order iIMM further improves model quality (Figures A.5A and A.5C). Importantly, in this benchmark test, the necessity to incorporate nucleotide interdependencies into binding site models becomes apparent to be even greater, since learning positional hexamers, instead of monomers, increases the pAUC by as much as 75%. As a result, dependencies between neighboring nucleotides need to be taken into account in order to comprehensively model the full complexity of CTCF binding behavior.

Part of the improvement stems from modeling varying sequence contexts from retroelements, such as Alu elements, that frequently flank CTCF binding sites due to common evolutionary history (Schmidt et al., 2012). While those flanking sequences, most likely, do not contribute to binding specificity, their inclusion into binding models (Figure A.5D) is relevant for detecting binding sites de novo.

The C_2H_2-ZnF family expanded extensively, both with the emergence of vertebrates and mammals, to become the largest class of DNA-binding domains in human (Vaquerizas et al., 2009), with C_2H_2-ZnF proteins containing roughly ten C_2H_2 ZnF domains on average (Najafabadi et al., 2015). Extended iIMMs appear to be particularly promising to determine the binding specificity of this widespread class of ZnF-containing transcription factors to presumably long and complex binding sites.

3.1.1.2. Predicting pioneer transcription factor binding affinities

In order to model gene regulatory networks, highly accurate quantitative binding specificity models are indispensable. To explore the quantitative value of iIMMs, I learned binding models of increasing complexity for the Krüppel-like factor 4 (Klf4), a pioneer transcription factor that is essential to initiate reprogramming of somatic cells to pluripotency (Takahashi and Yamanaka, 2006), using *in vivo* ChIP-seq measurements in mESCs (Chen et al., 2008).

Subsequently, I correlated affinity predictions for single and multiple mutated consensus binding sites to binding affinities measured *in vitro* by competitive EMSA experiments (Sun et al., 2013). Figure 3.5A (left) reveals that, overall, the Pearson correlation becomes greater with increasing iIMM order (solid lines). Pearson correlations calculated from affinity predictions of the PWM model determined by XXmotif serve as reference points (dashed lines).

While the PWM model of GIMMEmotif successfully predicts Klf4 affinities to single mutated binding sites, it fails at predicting affinities to multiple mutated binding sites (Figure 3.5A, middle, Pearson correlation = 0.26). In contrast, the 5^{th}-order iIMM succeeds in both tasks (Figure 3.5A, right). Thus, to accurately predict binding affinities to multiple mutated binding sites, it is crucial to model Klf4 binding sites with an iIMM of order at least three, which is also in evidence from the higher-order sequence logo of the 5^{th}-order iIMM (Figure A.6A).

Klf4 was discovered to bind partial motifs on nucleosomes, using only two of its three ZnF domains required for binding nucleosome-depleted binding sites (Soufi et al., 2015). In addition, Klf4 was found to bind distinct methylated and unmethylated motifs (Hu et al., 2013). Higher-order iIMMs appear to perform favorably in modeling this complex binding landscape.

In contrast to the ZnF domains of Klf4, the winged helix/forkhead box (Fox) DNA-binding domain of the paradigm pioneer transcription factor FoxA2 resembles that of the linker histone (Clark et al., 1993), enabling the opening of compacted chromatin (Cirillo

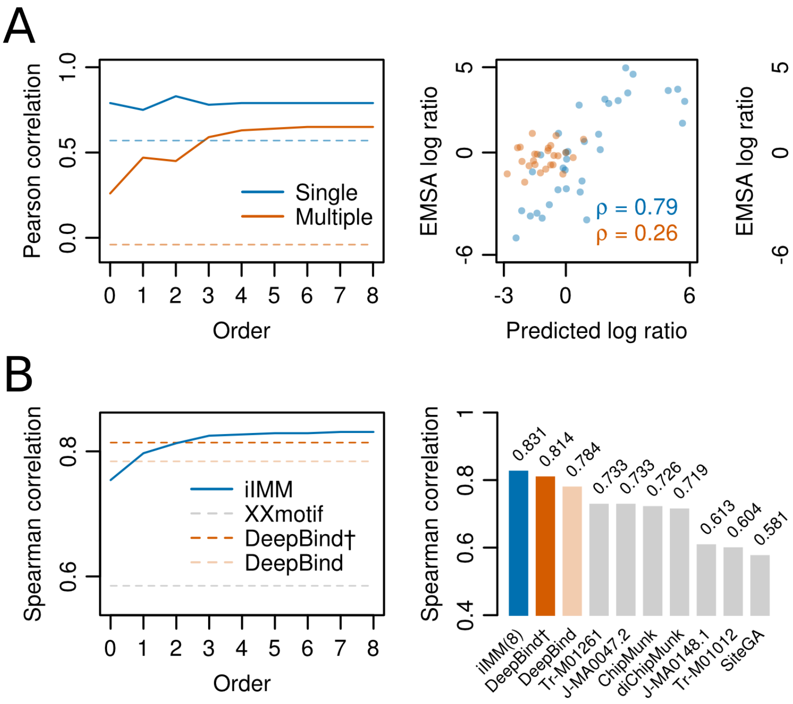


Figure 3.5.: **Higher-order iIMMs predict pioneer transcription factor binding affinities. (A)** Klf4 binding affinities to mutated consensus binding sites. Pearson correlations between predicted and measured log ratios of single and multiple mutated Klf4 binding sites (left), using iIMMs of increasing order (solid) and the most significant PWM reported by XXmotif (dashed). Scatter plot of predicted and measured log ratios of single (blue) and multiple (orange) mutated Klf4 binding sites, along with Pearson correlations, using the PWM model (middle) and the 5th-order iIMM (right) from (left). **(B)** FoxA2 binding affinities to 64 putative binding sites. Spearman correlations between EMSA scores and log ratios calculated by iIMMs of increasing order (left). The dashed lines represent Spearman correlations achieved by the most significant PWM of XXmotif (grey) and two DeepBind (orange) models. DeepBind† (dark orange) and DeepBind (light orange) models differ in length (16 and 24 bp, respectively). The bar plot (right) compares the Spearman correlations of the 8th-order iIMM and both DeepBind models to Spearman correlations attained by other methods and models (adopted from Alipanahi et al. (2015)).

et al., 2002) to initiate organogenesis (Lee et al., 2005). To test ChIP-seq FoxA2 iIMMs from HepG2 cells, I correlated predictions from iIMMs of increasing order with EMSA-measured binding affinities (Levitsky et al., 2014). Notably, Spearman correlations increase with model order, exceeding 0.83 when learning position-specific octamers (Figure 3.5B, left).

Recently, FoxA2 models learned by the deep learning technique DeepBind were shown to predict binding specificities that achieve the highest Spearman correlations among published FoxA2 models and FoxA2 models learned from the same ChIP-seq data set using published methods (Figure 3.5B, right) (Alipanahi et al., 2015). Remarkably, iIMMs of order three or higher (see higher-order sequence logos of 5th-order FoxA2 iIMM in Figure A.6B) outperform both DeepBind models (Figure 3.5B, left), without adjusting parameters. Since I use a single parameter setting for all benchmark tests there is still room to refine single models.

The results from both pioneer transcription factors implicate that although most information is captured by 2nd-order iIMMs, higher-order dependencies need to be considered to more comprehensively specify their binding behavior, e.g. to biologically relevant low-affinity binding sites (Crocker et al., 2015).

3.1.1.3. Predicting genomic-context PBM signal intensities

In Section 2.3.3, I explained how iIMMs can be learned from weighted sequences. To test this variant of iIMM learning, I determined binding specificity models for two *S. cerevisiae* bHLH transcription factors, Cbf1 and Tye7, by leveraging fluorescence signal intensities measured in gcPBM experiments (Gordân et al., 2013).

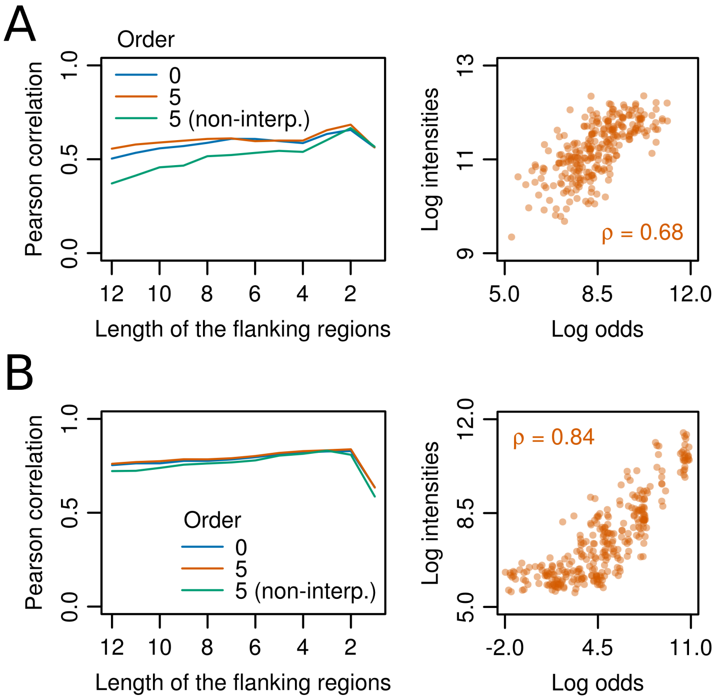


Figure 3.6.: **Higher-order iIMMs predict gcPBM signal intensities.** The influence of modeling an increasing number of flanking nucleotides on the Pearson correlation between measured gcPBM log signal intensities and predicted log-odds scores using either 5th-order iIMMs, 5th-order non-interpolating (non-interp.) iMMs, or PWM models (left). The scatter plot (right) compares measured log signal intensities and predicted log-odds scores using the 5th-order iIMM that models binding sites including two-bp-long proximal flanks from (left). **(A)** Cbf1 models. **(B)** Tye7 models.

Gordân et al. (2013) found 11 bp (Cbf1) and 5 bp (Tye7) genomic regions flanking the core E-box binding site CACGTG to influence DNA binding specificity. In contrast, iMMs and iIMMs perform best when modeling binding sites including two bp proximal flanks, independent of model complexity (Figure 3.6, left). Furthermore, the gain in performance is small when incorporating higher-order dependencies, indicating that the influence of dependencies between neighboring nucleotides on binding specificity is marginal. This argues against an influence of DNA shape, which is mainly determined by base-stacking interactions, on binding specificity, as suggested by Gordân et al. (2013).

Both transcription factors bind as homodimers to palindromic binding sites. Therefore, dependencies between non-neighboring nucleotides, corresponding to each other in palindromic half-sites, likely exist. By combining upstream and downstream flanks to calculate the sequence features of support vector regression models, Gordân et al. (2013) implicitly learn this dependency structure. Hence, the influence of flanking nucleotides on binding specificity rather stems from modeling palindromic binding site, instead of DNA shape, characteristics.

The 5$^{\text{th}}$-order iIMM of binding sites including two bp proximal flanks achieves Pearson correlations between measured log signal intensities and predicted log-odds scores of 0.68 and 0.84 for Cbf1 and Tye7, respectively (Figure 3.6, right). Without sequence weighting, the same model attains Pearson correlations of 0.48 (Cbf1) and 0.73 (Tye7). Thus, the accuracy of models learned from gcPBM measurements benefits considerably from integrating sequence weights into iIMM learning. However, iIMMs perform slightly unfavorable compared to existing approaches, as the palindromic binding behavior is modeled neither implicitly, by symmetrizing flanking sequences (Gordân et al. (2013), Mordelet et al. (2013)), nor explicitly, by learning dependencies between non-neighboring nucleotides from binding half-sites (Keilwagen and Grau, 2015).

3.1.2. Modeling RNAP I/II core promoter sequences

So far, I concentrated on the modeling of DNA sites that are bound by a single transcription factor. To investigate whether iIMMs improve PWM models of regulatory regions, which are the target of a multitude of proteins, I start with the modeling of core promoter sequences from *D. melanogaster*.

The core promoter is the DNA sequence required to initiate transcription. It encompasses the TSS in a region approximately ±50 bp relative to the TSS and comprises sequence elements, such as the TATA box, that interact with general transcription factors to recruit RNAP. Although the core promoter sequence was claimed to be the determining factor in establishing expression levels in the unicellular fungus *S. cerevisiae* (Lubliner et al., 2015), the transcriptional activity in multicellular organisms is strongly affected by distal features such as enhancers (Zabidi et al., 2015). In contrast, TSS selection was found to be largely governed by the local DNA sequence (Frith et al., 2008). Given this background, I wondered whether iIMMs are also the method of choice to model core promoter sequences and predict TSS locations.

I decided to model the two major mRNA core promoter classes pertaining to TSS precision: the narrow (NP) and broad (BP) transcription-initializing core promoters, broadly associated with developmentally regulated and housekeeping genes, respectively (Rach et al.,

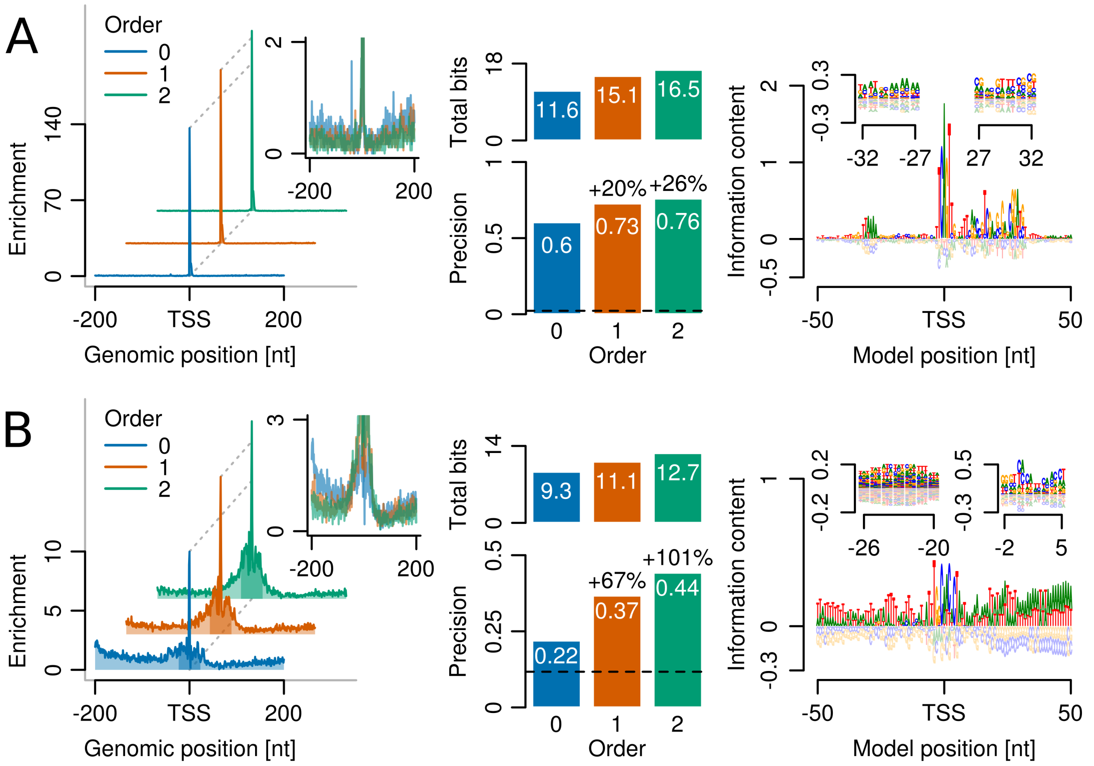


Figure 3.7.: **Higher-order iIMMs predict TSS locations. (A)** NP core promoter locations from *D. melanogaster*, predicted by iIMMs of increasing order, pictured as enrichments compared to predictions from a random predictor (left). Models were learned within four bp of measured TSSs. Inset shows a magnification of the y-axis to highlight off-site predictions. Precision of NP core promoter predictions (lower middle), measured as the fraction of predicted TSSs that are located within four bp of measured TSSs (darker shading, left). Precision improvements of higher-order iIMMs compared to the PWM model are listed in percentage atop bars. The dashed line indicates the baseline precision, as achieved by a random predictor. In addition to their performance, the total information content (in bits) of respective iIMMs is highlighted (upper middle). The 0th-order sequence logo depicts the 2nd-order iIMM learned from all sequences (right). Insets show sections of the 1st-order sequence logo in regions covering the TATA box, as well as DPE and E-box motifs, upstream and downstream of the TSS, respectively. **(B)** Same as **A** but showing models of BP core promoters from *D. melanogaster*, learned and predicted within 23 bp of measured TSSs. Logo insets show 2nd-order and 1st-order contributions in regions covering the Ohler6 and DRE, as well as Ohler1 and Ohler7 motifs, upstream and downstream of the TSS, respectively.

2009). To determine the impact of modeling higher-order dependencies on the prediction of TSS locations, I compare iIMMs of order zero, one, and two with each other.

NP core promoter models

I show the predictions of NP core promoter models as enrichments compared to predictions from a random predictor in a region ±200 bp around the TSS (Figure 3.7A, left). Clearly, the performance of the PWM model (blue line) can be improved by considering correlations between neighboring nucleotides. Consequently, iIMMs reduce the number of off-site predictions, which are widespread when using a PWM model (Figure 3.7A, left, inset). Specifically, 73% of TSSs can be located within 4 bp of the measured TSS using a 1st-order iIMM, whereas the precision drops to 0.6 when discarding 1st-order correlations (Figure 3.7A, lower middle). Prediction results improve further, albeit less markedly, using an iIMM of order two, totaling to an increase in precision of 26% compared to the PWM model.

The 0th-order sequence logo of the 2nd-order NP core promoter iIMM (Figure 3.7A, right) reveals the core promoter elements initiator (INR), TATA box, motif ten element (MTE), and downstream promoter element (DPE), which were found associated with NP core promoters (Ni et al., 2010). Details of the 1st-order sequence logo (Figure 3.7A, right, insets) show 1st-order dependencies in the region around the TATA box (left inset) and a region downstream of the TSS, in which the iIMM constitutes both the DPE and an E-box motif. In the former, nucleotide correlations partly arise from the slightly variable positioning of the TATA box with respect to the TSS. The latter case indicates how the iIMM implicitly integrates two disparate sequence elements occurring at overlapping positions relative to the TSS in distinct core promoter subsets. The complete sequence logo of the 2nd-order iIMM can be found in Figure A.7A.

The boost in precision that results from raising the model order is accompanied by a concomitant increase in the total information content of the models (Figure 3.7A, upper middle). iIMMs of higher order do not further improve the prediction of TSS locations (Figure A.9A). Importantly, however, higher-order iIMMs do not get overfit even at high orders, in contrast to iMMs. This property of iIMMs is very important in practical applications, since, in advance, it is mostly unknown up to which order the modeling is both biologically meaningful and statistically feasible.

BP core promoter models

The precision in predicting TSS locations is lower for BP compared to NP core promoters (Figure 3.7B, left and lower middle). In contrast to NP core promoters, sequence motifs were found to be less positionally fixed in BP core promoters (Rach et al., 2009), which might be the main reason for the overall lower precision. Instead, the gain in performance achieved by increasing iIMM order is much more prominent. Remarkably, the precision doubles when increasing the order from zero to two (Figure 3.7B, lower middle).

I spot several possible reasons for this finding in the sequence logo of the 2nd-order iIMM (Figure 3.7B, right). Similar to NP core promoters, the iIMM of BP core promoters represents sequence motifs that slightly vary in positioning as well as dissimilar sequence elements with identical distance to the TSS. For instance, the core promoter element Ohler6 (Ohler et al., 2002) and the DNA recognition element (DRE), which is represented in two adjacent registers, are admixed but distinguishable in the 2nd-order sequence logo (left inset). Similarly, the core promoter elements Ohler1 and Ohler7 (Ohler et al., 2002) are aligned onto each other (right inset). However, by solely modeling interdependencies between neighboring sequence positions, iIMMs are unable to model the observed co-occurrence of Ohler6 and Ohler1 as well as DRE and Ohler7 in core promoter subsets (FitzGerald et al. (2006), Ohler (2006)), which, most likely, would further improve the prediction results.

The information content at single positions in the sequence logo seems to be low but sums up across the core promoter (Figure 3.7B, upper middle, and Figure A.7B). I speculate that the nucleotide interdependencies represented in my models, at least in part, reflect DNA structural properties of core promoter sequences, which were reported to contribute to the promoter signature (Durán et al., 2013). For that purpose, an iIMM of order two seems to be sufficient, as the precision of TSS predictions cannot be further enhanced by learning iIMMs of order three or higher (Figure A.9A), similar to NP core promoters. Hence, iIMMs prove beneficial in predicting TSS locations and representing the different sequence architectures of both core promoter classes, which have been implicated in enhancer-core promoter specificity (Zabidi et al., 2015).

RP gene core promoter models

Notably, models of RP gene core promoter sequences, which are known to differ from NP and BP core promoters in their sequence architecture (Lenhard et al., 2012), do not profit from modeling higher orders (Figure A.8A). Despite the low number of RP gene core promoter sequence instances (Section 2.4.4), even a 5$^{\text{th}}$-order iIMM is not prone to overfitting, unlike iMMs of order one or higher (Figure A.9A).

3.1.3. Modeling RNAP II polyadenylation site models

Similar to the core promoter sequence, which positions RNAP II to initiate transcription, sequence elements in the 3′ UTR induce transcription termination by recruiting the cleavage and polyadenylation machinery to the pA site. In *S. cerevisiae*, 3′-end processing sequence signals were detected in the range from roughly 70 bp upstream to 30 bp downstream of the pA site (Tian and Graber, 2012). These include the UA-rich efficiency element (EE), located at least 25-40 nt upstream of the pA site, the A-rich positioning element (PE), located 10-30 nt upstream of the pA site, downstream of the EE, and U-rich elements immediately flanking the pA site. Compared to mRNA stability and translation regulation, mRNA 3′-end processing efficiency was demonstrated to exert predominant effect on expression variability, mainly mediated by the EE (Shalem et al., 2015), which seems to be the least of otherwise highly sequence variation-tolerant pA signal elements (Tian and Graber, 2012).

To predict pA site locations in analogy to core promoter TSSs, I chose to model the entire pA signal region from 70 bp upstream to 30 bp downstream of the pA site using iIMMs of increasing order. Consistent with the results on transcription factor binding sites and core promoter sequences, higher-order iIMMs outperform PWM models in predicting pA site locations by reducing the number of off-site predictions (Figure 3.8, left and lower middle). Notably, while two-thirds of the total (45%) increase in precision stem from 1$^{\text{st}}$-order nucleotide interdependencies, one-third follows from modeling 2$^{\text{nd}}$-order dependencies.

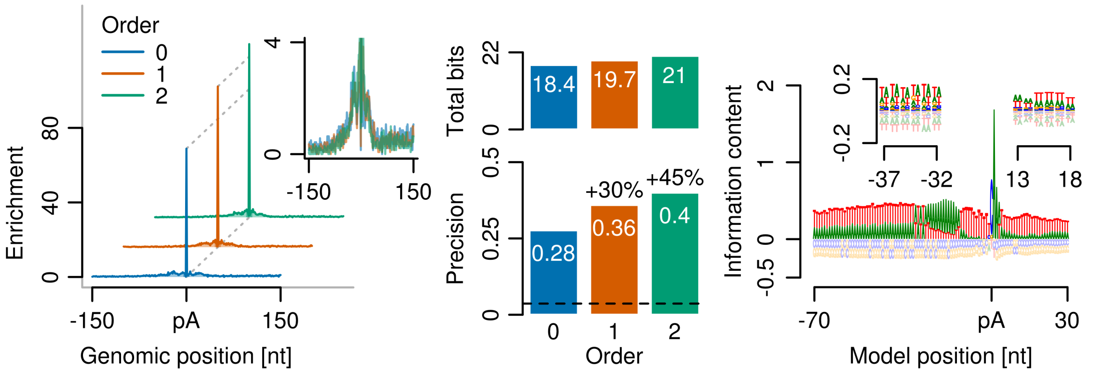


Figure 3.8.: **Higher-order iIMMs predict pA site locations.** Same as Figure 3.7A but showing models of mRNA pA sites from *S. cerevisiae*, predicted within five bp of measured pA sites. Logo insets show 1st-order contributions in regions covering the efficiency and U-rich elements, upstream and downstream of the pA site, respectively.

Thus, even a marginal increase in the total information content of higher-order relative to lower-order iIMMs (Figure 3.8, upper middle) can have a tremendous effect on prediction performance.

Besides recruiting RNA-binding proteins to emerging elements on the RNA transcript, the DNA sequence around pA sites plays a role in slowing down RNAP II transcription to allow exchange of elongation for termination factors. Therefore, I depict the sequence logo of the pA site model using the DNA alphabet.

The 2nd-order iIMM comprises all known 3′-end processing elements (Figure 3.8, right). Strikingly, the modeling of 1st-order correlations is inevitable to represent the EE (left inset) and the downstream T-rich region (right inset), which is in fact characterized by a long poly(dA:dT) tract (see Figure A.7C for the complete sequence logo). This higher-order primary structure may favor a common secondary structure to be recognized by the cleavage and polyadenylation machinery (Moqtaderi et al., 2013). Consequently, iIMMs advance the prediction of pA site locations by appropriately modeling these structures.

Similar to the results on core promoter sequences, higher-order iMMs of pA sites are prone to overfitting (Figure A.9B). In contrast, higher-order iIMMs prove to be robust, even though raising the order of iIMMs to three or higher does not further improve the prediction of pA site locations.

I successfully learned and predicted the most dominant pA site per gene (Section 2.4.5), which corresponds to the most permissive site for cleavage and polyadenylation (Moqtaderi et al., 2013). Less dominant pA sites are however distributed over a broad region, the “end zone”, downstream of the ORF terminus (Moqtaderi et al., 2013). It remains to be examined whether iIMMs are able to predict this complex cleavage and polyadenylation profile at single genes.

Since pA site cleavage and polyadenylation was reported to be highly heterogeneous between yeast species, mediated by species-specific factors (Moqtaderi et al., 2013), iIMMs show great promise for identifying universal and species-specific features by exploring pA sites in other species. In human, regulatory cleavage and polyadenylation events produce 3′ UTRs of differing lengths and were suggested to be involved in the establishment of tissue or cell identity (Ni et al., 2013). It would be exciting to dissect potential tissue-specific pA site signals using iIMMs.

3.1.4. Modeling RNAP pause sites

I complement the dependency of the transcription cycle on the DNA sequence by examining bacterial RNAP transcriptional pausing, a feature of transcription elongation, which, in prokaryotes, has been linked to the synchronization with translation (Proshkin et al., 2010). Larson et al. (2014) identified 16-bp- and 12-bp-long RNAP pause sequence signatures in *E. coli* and *B. subtilis*, respectively, suggesting that transcriptional pausing is driven by RNAP-nucleic acid interactions. By using PWM models, they ignore to what extent sequence-dependent structural properties of the DNA molecule effect pausing dynamics. I analyzed how models of higher order compare to the PWM model in predicting nucleotide-precise RNAP pausing events.

While all models predict the exact location of RNAP pausing in more than 30% of *E. coli* sequences, higher-order iIMMs increase the number of correct pause site predictions considerably (Figure 3.9, left and lower middle). For instance, the 2$^{\text{nd}}$-order iIMM correctly localizes 4,999 pause sites, 1,349 (37%) more than the PWM model, which is reflected in a 35% increase in total information content (Figure 3.9, upper middle). This suggests that DNA structural properties, implicitly modeled by iIMMs, facilitate RNAP pausing.

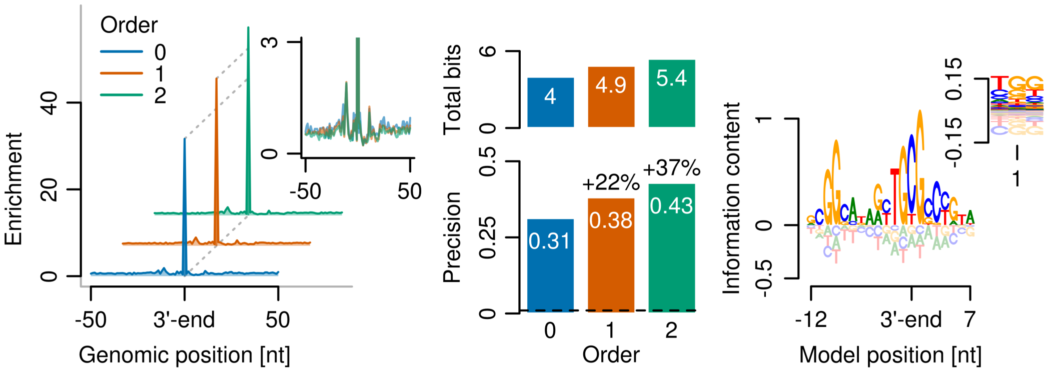


Figure 3.9.: **Higher-order iIMMs predict RNAP pause site locations.** Same as Figure 3.7A but showing models of RNAP pause sites from *E. coli*, predicted within zero bp of measured pause sites. Logo inset shows 2$^{\text{nd}}$-order contributions two bp downstream of the 3′-end of the transcript.

I observe off-site predictions upstream and downstream of the pause index, irrespective of the model order. The most prominent off-site peak is located 11 bp upstream of the pause index (Figure 3.9, inset), presumably as a result of similar sequence content in both parts of the model (Figure 3.9, right).

Beside the GpG dinucleotide at the 5′-end of the RNA-DNA hybrid, ten bp upstream of the 3′-end, another distinctive feature of the consensus sequence described by Larson et al. (2014) is the occurrence of TpG or CpG at the location of the 3′-end of the nascent transcript and incoming nucleoside triphosphate (NTP). The CpG dinucleotide of the template strand was recently shown to inhibit elongation and induce G-to-A errors when spanning the active site of RNAP (Imashimizu et al., 2015). The 2$^{\text{nd}}$-order iIMM refines this signature by unveiling that a TpG favors a G as the next nucleotide, whereas CpG is more likely to be followed by a T or C (Figure 3.9, right, inset).

In contrast to iIMMs of RNAP pause sites in *E. coli*, which show improvements up to order two, RNAP pause site models learned in *B. subtilis* enhance performance by including order three (Figures A.8B and A.9C). However, the performance in predicting pause site locations is less than half of that in *E. coli*, pointing to a weaker sequence dependence of RNAP pausing in *B. subtilis*. *E. coli* and *B. subtilis* pause site models both contain a GpG dinucleotide at the upstream and a pyrimidine at the downstream edge of the RNA-DNA hybrid, but otherwise differ substantially in their lower-order as well as higher-order content.

It will be interesting to see how the observed sequence preferences relate to those underlying RNAP II pausing in human cells (Mayer et al. (2015), Nojima et al. (2015)).

3.2. Protein-RNA binding specificity models

Protein-RNA interactions differ from protein-DNA interactions mainly due to the greater role that RNA secondary and tertiary structure plays in recruiting RNA-binding proteins. Unlike DNA-binding domains, which generally bind to related sequences, RNA-binding domains such as RNA recognition motifs (RRMs) show pronounced structural versatility to facilitate the binding to diverse sequences (Maris et al., 2005). While some RNA-binding proteins recognize structured regions (Stefl et al., 2005), in which distant stretches of nucleotides in the sequence form base-pairing stems, the vast majority of RNA-binding proteins seems to bind short single-stranded RNA sequences (Ray et al., 2013). Hence, sequence readout remains an integral part in protein-RNA interactions (Auweter et al., 2006), possibly because RNA is less structured *in vivo* as opposed to *in vitro* conditions (Rouskin et al., 2014). Therefore, I wondered whether the modeling of nucleotide interdependencies within protein-RNA binding sites increases model accuracy, without explicitly learning structural features (Li et al., 2014b).

3.2.1. Modeling PAR-CLIP crosslink sites

I examine binding sites of 25 mRNP biogenesis factors from *S. cerevisiae*, measured *in vivo* using PAR-CLIP (Baejen et al. (2014), Schulz et al. (2013)) (Section 2.4.7).

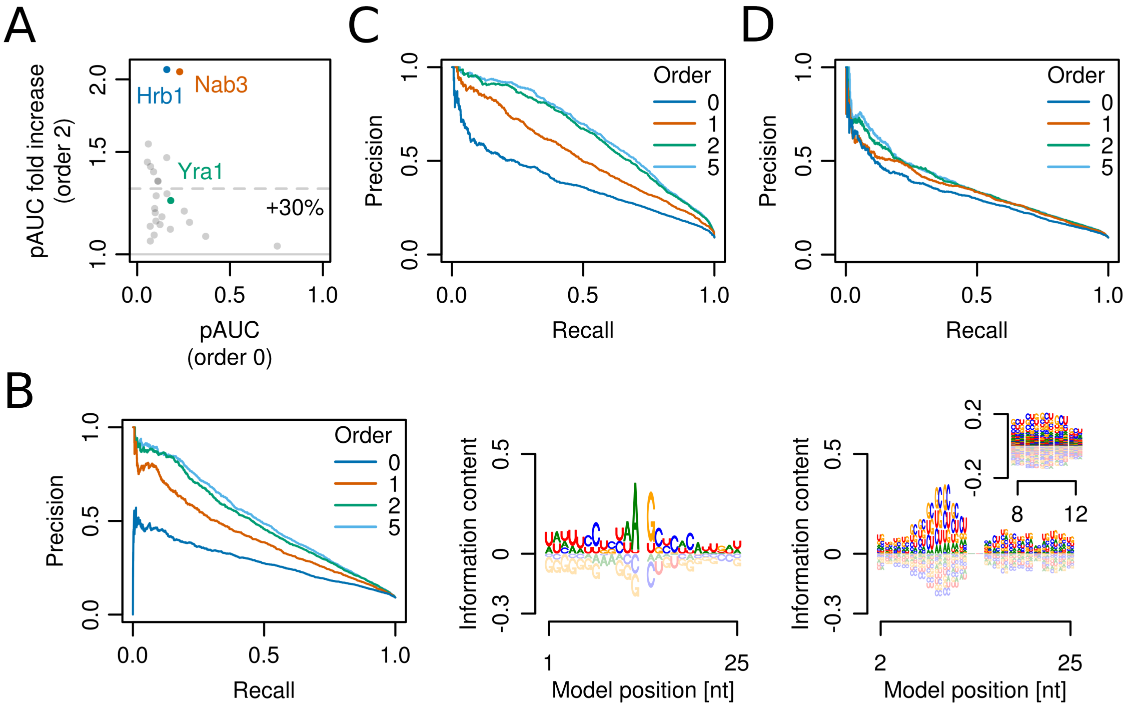


Figure 3.10.: **Higher-order iIMMs predict protein-RNA crosslink sites. (A)** Scatter plot of the increase in performance, as measured by the partial area under the ROC curve (pAUC), up to a false positive rate of 5%, when increasing the model order from zero to two. The y-axis is shown in log scale. The dashed line indicates the mean fold increase. **(B)** Precision-recall curves (left) calculated using Hrb1 iIMMs of increasing order. 0^{th}-order (middle), 1^{st}-order (right), and 2^{nd}-order (right, inset) sequence logos depict 2^{nd}-order Hrb1 iIMM. The central crosslinked U was removed from the 0^{th}-order sequence logo. Sequence logos show models learned from all sequences. **(C,D)** Precision-recall curves calculated using **(C)** Nab3 and **(D)** Yra1 iIMMs of increasing order.

Figure 3.10A shows the discriminative power of 2^{nd}-order iIMMs compared to PWM models. With the exception of Nrd1, the performance of PWM models is low (pAUC < 0.4). Despite the fact that Bacikova et al. (2014) reported broad sequence specificity in the low micromolar range for Nrd1, the PWM model seems to be almost as accurate as a 2^{nd}-order iIMM to model Nrd1 crosslink sites. For all RNA-binding proteins, iIMMs outperform PWM models ($P = 3 \times 10^{-8}$, Wilcoxon one-sided signed-rank test, $n = 25$), in two cases doubling the pAUC values that PWM models attained.

Replacing the PWM model by an iIMM had the most drastic effect for the SR-like factor Hrb1, which is implicated in the quality control of mRNAs by restricting mRNA export to spliced transcripts (Hackmann et al., 2014). While the main improvement can be attributed to 1^{st}-order dependencies, 2^{nd}-order statistics further improve model quality (Figure 3.10B, left).

The 0^{th}-order logo of the 2^{nd}-order iIMM shows that the crosslinked U (not shown) is frequently flanked by A and G upstream and downstream, respectively (Figure 3.10B, middle), which also show up around other RNA-binding protein crosslink sites (Figure

A.10). The 1st-order and 2nd-order logos highlight a CUG-rich region upstream of the crosslink site (Figure 3.10B, right), representing up to five successive CUG repeats contained in some sequences, that is not apparent in the PWM model.

CUG repeats can fold into hairpins with paired CGs and U bulges conforming to an A-form helix (Kiliszek and Rypniewski, 2014) and have been found to bind to human alternative splicing regulators of the CUG-BP, Elav-like family (CELF) (Timchenko et al., 1996) and muscleblind-like (MBNL) family (Miller et al., 2000) of RNA-binding proteins. Specifically, MBNL1 is sequestered by C(C)UG repeat expansions in myotonic dystrophy (Kanadia et al., 2003). However, different from the ZnF domains of MBNL1, which can also bind to double-stranded RNA (Konieczny et al., 2014), the three RRMs of CELF1 strongly prefer binding single-stranded RNA (Edwards et al., 2013). Hrb1, which also contains three RRMs, may bind to CUG-rich RNA by simultaneous interaction of its three RRMs to single-stranded RNA sites, akin to CELF1.

Similar to Hrb1, higher-order iIMMs of Nab3 crosslink sites markedly improve the PWM model (Figure 3.10C). In this case, higher-order iIMMs are capable of integrating the Nab3 motifs UCUU and CUUG, as well as Nrd1 motifs UGUA and GUAG (Creamer et al., 2011) into one model (Figure A.10A), consistent with Nab3 preferentially binding in close but variable spatial proximity to Nrd1 (Schulz et al., 2013).

I exemplify an iIMM with moderate improvement compared to the PWM model by the Mex67 adaptor protein Yra1 (Strässer and Hurt, 2000) (Figure 3.10D and Figure A.10B). Consistent with the applications to DNA-binding proteins, iIMMs of RNA-binding proteins are insensitive to parameter overfitting, in contrast to iMMs (Figure A.11).

Collectively, my analysis reveals that higher-order dependencies may represent secondary structure effects and interactions between multiple RNA-binding domains (Hrb1) and between multiple RNA-binding proteins (Nab3), all of which help to improve the performance of sequence models in predicting *in vivo* binding sites of RNA-binding proteins. Since RNA-binding proteins do not necessarily bind directly to the crosslinked uracil, higher-order effects may, in parts, reflect variable distances between their binding sites and the crosslinked uracils.

4. Conclusion

I modeled protein-DNA and protein-RNA binding sites using iIMMs, which only incorporate information about dependencies between adjacent nucleotides if evidence in their favor exists. Thus, model complexity is automatically adapted to the data, making iIMMs insensitive to parameter overfitting even when using exceedingly high model orders. I derived an EM algorithm, GIMMEmotif, that learns iIMMs from binding sites enriched in a positive sequence set as compared to a negative sequence set. The resulting iIMMs are made easily accessible to interpretation by visualizing the contribution of individual model orders to the overall information content.

The performance of iIMMs learned by GIMMEmotif was evaluated over a wide range of data sets and binding motifs. On 446 human ChIP-seq data sets from ENCODE, iIMMs achieve 41% mean and 26% median improvements in the partial area under the ROC curve, that is, in predicting binding instances on held-out data. The source of discovered nucleotide interdependencies appears to be manifold. For instance, detected sequence preferences seem to promote structural features that facilitate transcription factor binding either directly, within the core binding site, or indirectly, by facilitating binding site targeting, which seems to depend on a beneficial binding site environment (Dror et al., 2015). Positional dependencies in iIMMs also reflect different binding modes, depending on the cooperative binding of a cofactor or reflecting the ability to adapt to varying spacer lengths between binding half-sites by structural rearrangements within single or between multiple DNA-binding domains.

In addition to the value of iIMMs in discriminating bound from unbound binding sites, I also demonstrated their value in predicting *in vitro* binding affinities of two pioneer transcription factors. Remarkably, the accuracy of FoxA2 iIMMs was shown to be superior to models obtained from state-of-the-art methods.

Instead of binding isolated from other binding events, transcription factors act in concert to control gene expression. Therefore, I wondered whether iIMMs are able to better capture characteristics of entire regulatory regions, which are bound by a multitude of proteins. In fact, iIMMs of RNAP II core promoter sequences from fly improve the fraction of correctly predicted NP and BP TSSs over PWM models by 26% and 101%, respectively. Predictions of pA sites in yeast are enhanced by 45%. In addition to the benefits observed for single transcription factor binding sites, the modeling of nucleotide interdependencies within regulatory regions permitted the representation of variable distances between regulatory elements.

In order to examine another DNA signal that influences the dynamics of the transcription process, I chose to model bacterial RNAP pause sites. iIMMs also excel in learning transcriptional pauses, which are induced by the complex interface between the polymerase and the DNA template, improving the fraction of correctly predicted pause sites in *E. coli* and *B. subtilis* by 37% and 55%, respectively.

Lastly, I applied iIMMs to PAR-CLIP crosslink sites. Compared to PWM models, iIMMs achieve 30% mean improvements in discriminating bound from unbound binding sites, as measured by the partial area under the ROC curve. In this case, the ability of iIMMs to outperform PWM models originates from their capacity to represent a set of heterogeneous sequences in a single model.

GIMMEmotif is also capable of harnessing information on binding strength or confidence in binding by integrating this additional data as sequence weights into the EM algorithm. Leveraging fluorescence signal intensities measured in gcPBM experiments, leads to improved accuracies of iIMMs for two bHLH transcription factors.

In summary, one elaborate model can capture nucleotide interdependencies arising from different binding determinants, such as multiple binding modes, cooperative binding, and DNA shape, as well as the complex architecture of multipartite elements exhibiting variable distances between submotifs and the presence of alternative submotifs. By exhausting the information contained in the sequence, it is not necessary to make the detour via inferring and integrating DNA shape features, as proposed by Zhou et al. (2015), which are merely a consequence of the underlying sequence.

Despite the increase in parameter space, higher-order iIMM learning turns out to be highly robust and reliable, giving equivalent or better results than PWM models on all data sets, using a single default set of hyperparameters. While 2$^{\text{nd}}$-order iIMMs capture binding site characteristics in the majority of cases, there are instances in which dependencies of even higher order were apparent. Since iIMMs, as opposed to iMMs, are robust to overfitting and consistently outperform PWM models, I claim that iIMMs are always preferable to simple PWM models. Therefore, replacing MEME-based tools with GIMMEmotif is expected to lead to significant improvements over a broad range of applications in learning models of enriched binding sites and predicting novel motif instances along with their binding strengths.

4.1. Outlook

I have several ideas how to further improve the scope of my work by extending both the EM procedure to learn and the modeling of binding specificities, as well as by making my models more usable to the community.

Benchmark tests

In the ChIP-seq benchmark test, I initialized iIMMs from the motif instances that XXmotif used to construct its most significant PWM model. However, the binding of DNA-binding proteins is often influenced by secondary or cofactor binding sites. To more comprehensively describe this binding landscape, multiple iIMMs could be initialized from additional PWM models reported by XXmotif and optimized independently (or jointly using a mixture model).

Due to runtime requirements of XXmotif, I had to limit the number of ChIP-seq peak sequences to 5,000. However, the runtime of the EM algorithm is not restricting. While still initializing iIMMs from XXmotif results obtained on restricted sequence sets, I could run the EM algorithm for optimizing iIMMs on all available peak sequences. Given that the results for CTCF are promising (Figures A.5A and A.5C), we are working on applying this approach to all data sets.

EM algorithm-related extensions

In this work, I used a single default set of hyperparameters, namely q (Equation 2.20) as well as α_k and β (Equation 2.10), which performs well on all data sets, despite the diverse nature of applications. However, in single data sets, the parameter setting can potentially turn out to be too conservative to capture subtle dependencies of higher order. Ideally, position-specific pseudocounts factors are desirable. These could also be used to optimize the length of models. Extended models, e.g. ± 20 bp around seed motifs, could then be learned by default and optimized pseudocounts factors used to prune extended models, dependent on their position-specific values. Hence, the accuracy of iIMMs can be further enhanced by optimizing the hyperparameters. This can be achieved either based on cross-validating the input sequences in order to obtain training and validation sets, or by employing the generalized EM algorithm (Dempster et al., 1977), that is, by optimizing, and not maximizing, the Q function (Equation 2.25) using numerical optimization techniques.

So far, the EM algorithm implements a “zero or one occurrence per sequence” (ZOOPS) model (Equation 2.20). However, multiple nearby binding sites of the same transcription factor can appear in the same sequence. In this case, motif models may benefit from the utilization of a “multiple occurrence per sequence” (MOPS) model by increasing the effective number of instances, particularly when only few input sequences are available.

In some applications, binding sites lie localized to some reference point within the input sequences. For instance, binding events measured by ChIP-seq are expected to lie close to peak summits. This prior knowledge can be formulated using a non-uniform prior distribution over the latent variables.

The deterministic EM algorithm pursues a local search strategy. Hence, in order to converge to the global maximum of the likelihood function, it is sensitive to its initial parameter values. For this reason, stochastic variants of the EM algorithm, which allow to escape insignificant local maxima or saddle points, have been proposed (Celeux and Diebolt, 1985) and applied to motif discovery (Kilpatrick et al., 2014). It would be interesting to check the value of stochastic perturbations on improving binding models.

In Section 2.3.3, I demonstrated how sequence weights can be integrated into the EM learning scheme and describe simple approaches to calculate sequence weights from quantitative or statistical data on binding events. Evidently, there is room for refining the proposed transformations. For instance, it is conceivable to fit the quantitative data by a sigmoid function of the log-odds scores.

Finally, it will be fundamental for the user to being provided with statistical significance values such as P-values or false discovery rates. In fact, we are currently implementing an efficient computation of P-values to provide a significance evaluation of discovered iIMMs.

Modeling non-neighboring nucleotide interdependencies

As discussed in Section 1.3, dependencies between nucleotides within the transcription factor binding site become smaller with inter-nucleotide distance (Jolma et al., 2013). In agreement, the contribution of non-neighboring nucleotide interdependencies to the accuracy of binding models appeared to be marginal, except for palindromic binding sites (Keilwagen and Grau, 2015). On this account, I focused on correlations between neighboring sequence positions.

With respect to regulatory sequences harboring an ensemble of transcription factor binding sites, e.g. core promoters, it is promising to gain another look at non-neighboring nucleotide interdependencies. In this case, considering dependencies between distant nucleotides may implicitly enable the modeling of interdependencies between regulatory, e.g. core promoter, elements. Without changing the model, I can optimize the context C_i, on which the probability of the residue depends, for each position i, which is not necessarily given by the k preceding bases. This is similar to the idea of interpolated context models described by Delcher et al. (1999) but using iIMMs.

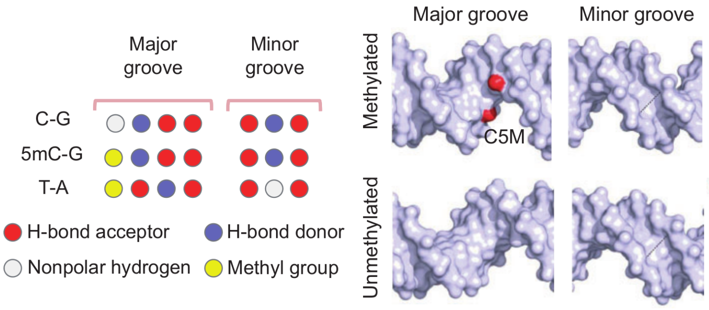
Providing a framework to circulate interpolated Markov models

Given the robust and reliable performance as well as the broad and unrestricted applicability of GIMMEmotif to nucleic acid sequences, it is important to spread the application to the community. To this end, we want to develop a server similar to the MEME suite (Bailey et al., 2009), but providing tools operating on the more sophisticated iIMMs instead of PWM models. For instance, a database maintaining and distributing iIMMs (and their sequence logos) of regulatory sequences could be used to search for matches in sequences, providing tools corresponding to FIMO (Grant et al., 2011) or MAST (Bailey and Gribskov, 1998). To provide a wide range of motif models, optimal iIMMs should be computed for all ChIP-seq and bacterial-1-hybrid data sets provided by the ENCODE consortia of human and model organisms. Certainly, the core will provide de novo searches for motifs modeled by iIMMs.

Modeling nucleic acid base modifications

So far, all protein-DNA binding specificity models comprise the standard bases adenine, cytosine, guanine, and thymine, thus using a four letter alphabet. In recent years, however, it has become evident that modifications to DNA bases, specifically modifications of cytosine residues by methylation (5mC) and hydroxymethylation (5hmC), are more important in the regulation of the binding behavior of a multitude of proteins than previously anticipated.

Similar to Figure 1.1, Figure 4.1 (from Dantas Machado et al. (2015)) illustrates how cytosine methylation affects the protein-DNA interface. The hydrophobic methyl group has a direct influence on the interface in the major groove by changing hydrophobic contacts to fine-tune the binding specificities of transcription factors (base readout). Furthermore, CpG methylation can reshape the structure and mechanical properties, such as stiffness (Pérez et al., 2012), of DNA binding sites in a sequence-dependent manner (shape readout). DNA methylation can thus have an influence on both direct and indirect readout mechanisms.



(a) Base readout *(b)* Shape readout

Figure 4.1.: **Base and shape readout of methylated DNA. (a)** Signatures of functional groups at the major groove (left) and minor groove (right) edges of C-G (top), 5mC-G (center) and T-A (bottom) base pairs. The methyl group (yellow) changes the signature of functional groups at the major groove edge of the C-G base pair, but base readout at the minor groove edge is not affected. **(b)** Presence of a 5mCpG dinucleotide (C5M carbon atom of the methyl groups shown in red) in methylated DNA (top) can affect the widths of the major (left) and minor grooves (right) compared with unmethylated DNA (bottom) as a function of its sequence context. Figure from Dantas Machado et al. (2015), distributed under the terms of the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

DNase I cleavage preferences serve as a convincing example for illuminating the interplay between DNA methylation status and protein-DNA interactions. By integrating genomic DNase I cleavage with bisulfite sequencing data for the same cell type, Lazarovici et al. (2013) showed that cleavage directly adjacent to CpG dinucleotides was enhanced at least eightfold in consequence of cytosine methylation-induced narrowing of the minor groove.

The Epstein-Barr virus transcription factor Zta was the first example of a sequence-specific transcription factor that binds methylated CpG residue-containing DNA sequence motifs to activate gene transcription (Bergbauer et al., 2010). Its DNA-binding domain is homologous to c-Fos, a member of the cellular activator protein 1 (AP-1) transcription factor family. Reminiscent of viral Zta, the c-Jun/c-Fos heterodimer was identified to reverse epigenetic silencing by inducing expression of repressed genes from methylated promoters (Gustems et al., 2014).

Transcription factors containing ZnF binding domains have also been reported to bind methylated DNA in a sequence-specific manner (Sasai et al., 2010). After elucidating structures of ZnF transcription factors in complex with methylated DNA (Buck-Koehntop et al. (2012), Liu et al. (2012)), Liu et al. (2013) proposed a common binding mode of ZnF transcription factors to methylated CpGs provided by RH-ZnF motifs, with neighboring ZnFs recognizing the sequence environment of methylated CpGs. This indicates that further sequence-specific methyl-binding ZnF transcription factors remain to be discovered. Meanwhile, the structural basis for the recognition of methylated DNA by Klf4 (Liu et al., 2014) and its intrinsic binding specificity to methylated and unmethylated motifs (Hu et al., 2013) have been elucidated. However, binding assays of higher throughput, as conducted by Hu et al. (2013) and Mann et al. (2013) using protein microarrays and PBMs, respectively, will be essential to reveal 5mC-specific DNA-binding propensities for most DNA-binding proteins. Intriguingly, cytosine methylation was found to create completely different binding sites for some transcription factors (Hu et al., 2013).

So far, I focused on methyl-DNA-binding proteins. TET proteins oxidize 5mC to 5hmC, 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC), alleged intermediates in the active demethylation pathway (Pastor et al. (2013), Kohli and Zhang (2013)). However, Hahn et al. (2013) showed that the increase of 5hmC levels in neuronal function-related genes during neurogenesis is not associated with substantial DNA demethylation, suggesting that 5hmC is a stable epigenetic mark. Particularly, methyl-CpG-binding protein 2 (MeCP2), a member of the methyl-CpG-binding domain (MBD) family of proteins (Baubec et al., 2013), was shown to bind to 5hmC (Mellén et al., 2012), although Chen et al. (2015) show evidence for specificity to non-CG methylated (mCH) DNA as neurons mature, influencing transcription and the timing of onset for Rett syndrome. The identification of dynamic readers for 5mC and its oxidized derivatives by Spruijt et al. (2013) applying mass-spectrometry-based proteomics, is another indication that further cytosine modifications have a specific biological role. Similarly, Iurlaro et al. (2013) found 5fC to be selectively bound by numerous proteins, exceeding the number of proteins with a preference for 5hmC, possibly induced by 5fC-mediated alterations of the DNA double helix (Raiber et al., 2015), suggesting functions in transcription and chromatin regulation. Moreover, while early growth response protein 1 (Egr1) and early growth response protein 1 (WT1) differentiate oxidized from unoxidized cytosine, rather than methylated from unmethylated cytosine, WT1 also recognizes 5caC within a specific DNA sequence (Hashimoto et al., 2014). Finally, in addition to cytosine modifications, adenine methylation (6mA) has recently been ascribed a potential regulatory role in gene expression (Fu et al., 2015) and transposon biology (Zhang et al., 2015) of eukaryotic organisms.

The preceding examples demonstrate that DNA modifications have a significant impact on the regulation of DNA biology and implicate the need to integrate DNA modification data into motif modeling. For instance, Luu et al. (2013) searched for motifs that discriminate between hypomethylated and hypermethylated regions. However, a logical step towards transcription factor binding specificity models that take base modifications into account would be to add DNA modifications as independent letters to the DNA base alphabet. Incorporating this extended DNA base alphabet (capturing direct readout of modified bases) into iIMMs (capturing DNA shape alterations, induced by modified bases, despite data sparsity) might result in a compelling symbiosis raising binding specificity models to the next level.

Extended alphabet models can be immediately learned from *in vitro* experiments like PBMs or HT-SELEX assays when measuring protein binding to DNA sequences with modified bases. In order to learn extended alphabet models from *in vivo* ChIP-seq experiments, genome-wide measurements of base modifications must be available in matched cell types. Favorably, experiments have been conducted for a number of modifications and organisms at single-base resolution (Booth et al. (2012), Yu et al. (2012), Song et al. (2013), Booth et al. (2014)), revealing the dynamic modification landscape across cell lines and tissues (Varley et al. (2013), Ziller et al. (2013)) as well as during development (Lister et al., 2013).

Similar to DNA, RNA nucleosides can be modified by a variety of chemical groups (Machnicka et al., 2013). To date, transcriptome-wide measurements of site-specific RNA methylation have been conducted for N^6-methyladenosine (m^6A) and 5-methylcytidine (m^5C) (Hussain et al. (2013), Linder et al. (2015)). m^6A is the predominant internal (non-cap) base modification present in eukaryotic mRNAs. Specific m^6A sites can be dynamically modulated (Schwartz et al., 2013) and selectively bound to regulate gene expression by affecting RNA stability and alternative splicing patterns (Liu et al. (2015), Wang et al. (2014), Dominissini et al. (2012)). For instance, m^6A-dependent RNA processing was shown to be crucial for circadian clock function (Fustin et al., 2013). Mechanistically, m^6A was described to control the RNA-structure-dependent accessibility of RNA binding motifs, which are otherwise buried within their local RNA structures (Liu et al., 2015).

Consequently, incorporating modified bases into (cell type-specific) DNA and RNA sequence-based binding specificity models, in particular iIMMs, by extending the four letter to a five or even six letter alphabet, may prove instrumental to further improve the accuracy of binding specificity models and, therefore, a promising new avenue to follow.

Part II.

Universality of core promoter elements?

5. Introduction

How cells locate the regions to initiate transcription is an open question, because core promoter elements (CPEs) are found in only a small fraction of core promoters (Ohler et al. (2002), Kadonaga (2012), Lenhard et al. (2012), Hartmann et al. (2013)). For example, the upstream and downstream transcription factor IIB (TFIIB) recognition elements (BRE$_\mathrm{u}$ and BRE$_\mathrm{d}$) were deduced from crystal structures of TFIIB and TATA box binding protein (TBP) bound to DNA (Nikolov et al. (1995), Lagrange et al. (1998), Deng and Roberts (2005)). To date, however, the BRE elements have not been found to be statistically enriched in human core promoter sequences by de novo motif searches.

A recent study by Venters and Pugh (2013) measured 159,117 DNA binding regions of TBP and TFIIB by ChIP-exo (chromatin immunoprecipitation with lambda exonuclease digestion followed by high-throughput sequencing) in human K562 cells. Surprisingly, four degenerate CPEs—TATA box, BRE$_\mathrm{u}$, BRE$_\mathrm{d}$, and initiator (INR)—could be located at constrained positions in nearly all of them, using IUPAC patterns (Table B.1) to describe CPEs and allowing up to three mismatches to their consensus. The authors conclude that these regions represent sites of transcription initiation marked by universal CPEs, covering RNA polymerase (RNAP) II/III TATA-containing/TATA-less coding and non-coding genes.

Short sequence motifs such as CPEs may occur frequently by chance, in particular when allowing up to three mismatched positions. To assess the biological relevance of TATA box (IUPAC pattern TATAWAWR), BRE$_\mathrm{u}$ (SSRCGCC), BRE$_\mathrm{d}$ (RTDKKKK), and INR (YYANWYY) occurrences around measured ChIP-exo peaks, it is necessary to compare their match frequencies to negative controls. For instance, the probability of observing an exact match of the INR consensus at a fixed position (search space 1) of a random sequence with 60% GC content is $P_\mathrm{Y}{}^4 P_\mathrm{A} P_\mathrm{W} = P_{(\mathrm{C\ or\ T})}{}^4 P_\mathrm{A} P_{(\mathrm{A\ or\ T})} = 0.5^4 \times 0.2 \times 0.4 = 0.005$. Hence, the probability of seeing no match within 60 possible start positions (search space 60) is approximately $(1-0.005)^{60} = 0.74$, and the probability of observing at least one match (by chance) is $1 - 0.74 = 0.26$. This and similar estimates for one to three mismatches were in strong disagreement with the negative controls shown in Figures 2c, 3e, and 6c by Venters and Pugh (2013).

Therefore, I checked the reported results using two negative control procedures. For the first negative control, I randomly permuted nucleotides in those sequence regions that were searched for matches to the CPE patterns in the original analysis, ensuring identical nucleotide composition. As a second negative control, I created nonsense patterns with information content identical to CPE patterns by alphabetically sorting their IUPAC letters to AAARTTWW (TATA box), CCCGRSS (BRE_u), DKKKKRT (BRE_d), and ANWYYYY (INR), respectively.

I show that the claimed universality of CPEs is explained by the low specificity of the patterns used and that the same match frequencies are obtained with the two negative controls. Furthermore, CPE patterns are not positionally enriched around TFIIB locations, except for TATA elements with zero or one mismatched positions around mRNA-associated TFIIB peaks. The results argue against the existence of a predictive universal core promoter sequence signature around TFIIB ChIP-exo peaks. Finally, my analyses also cast doubt on the biological significance of most of the 150,753 non-mRNA-associated ChIP-exo peaks, 72% of which lie within repetitive regions.

The main claims have been published (Siebert and Söding, 2014) and led to the retraction of the original study (Venters and Pugh, 2014).

6. Results

In my reanalysis I concentrate on TFIIB ChIP-exo peaks (Section 6.1). I start presenting the results of mRNA-associated TFIIB locations (Section 6.1.1) and continue with analyzing non-mRNA-associated TFIIB locations (Section 6.1.2). Finally, I investigate TBP ChIP-exo peaks (Section 6.2), focusing on tRNA-associated TBP locations (Section 6.2.1).

6.1. Core promoter elements around TFIIB ChIP-exo peaks

Venters and Pugh (2013) describe a total of 159,117 TFIIB ChIP-exo peak pairs in human K562 cells. 8,634 TFIIB locations could be assigned to the transcription start site (TSS) of a RNAP II-transcribed mRNA gene using the NCBI-curated RefSeq database (Pruitt et al., 2007) and requiring ChIP-exo peaks to lie within 500 bp of the corresponding TSS. This results in 6,511 non-redundant mRNA-associated TFIIB peaks. The remaining 150,753 TFIIB locations are referred to as non-mRNA-associated TFIIB locations.

In Supplemental Data 1 provided by Venters and Pugh (2013), I only identified 6,120 (instead of 6,511) unique mRNA genes assigned to at least one TFIIB ChIP-exo peak. Similarly, I worked with 150,721 (instead of 150,753) non-mRNA-associated TFIIB locations in order to avoid using peak sequences located at chromosome ends or containing non-annotated nucleotides. However, the results presented here do not dependent on the exact number of TFIIB peaks used in the analysis.

Instead of applying FIMO (Bailey et al., 2009), I used regular expressions to find occurrences of the four CPE patterns around ChIP-exo peaks. I searched for TATA elements within 168 bp (search space 161) around mRNA-associated TFIIB-bound locations on the sense strand. $\mathrm{BRE_u}$ was searched within 46 bp (40) upstream, and $\mathrm{BRE_d}$ and INR elements within 46 bp (40) and 66 bp (60) downstream of TATA elements on the sense strand, respectively. Similarly, the alphabetically sorted IUPAC patterns of BRE and INR elements were searched around AAARTTWW matches. Since no biological signal is available to orientate non-mRNA-associated TFIIB-bound locations, I searched for CPE elements within the same intervals but on both strands, thereby doubling the search space. In all analyses, I ignored sequences with matches with up to $k-1$ mismatched positions when recording matches for patterns with k mismatched positions, as done by Venters and Pugh (2013). In the following, distances between elements are defined according to motif midpoints. In the case of the TATA box, the midpoint corresponds to A or T (IUPAC letter W) at position five in the TATAWAWR consensus.

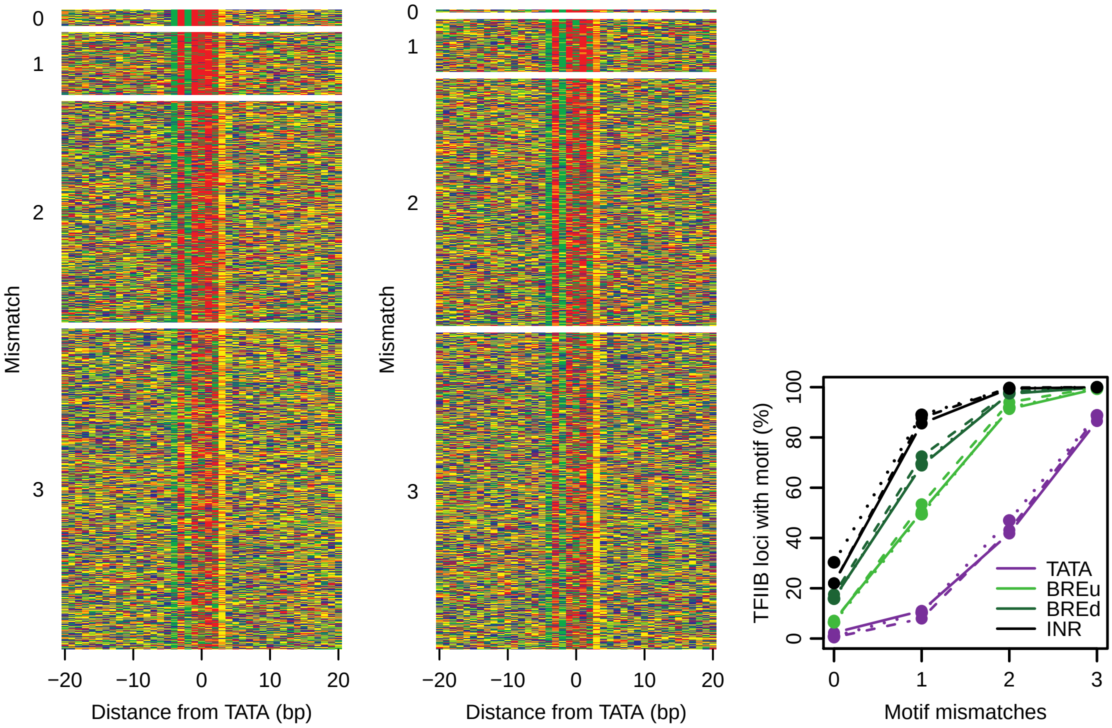
6.1.1. mRNA-associated TFIIB locations

First of all, I analyzed 6,120 TFIIB ChIP-exo peaks near annotated TSSs of mRNAs. Figure 6.1a shows the color-coded nucleotide sequences obtained by aligning 5,295 TFIIB-bound sequences found to contain one or more matches to the TATA element, allowing up to three mismatches to the consensus, thereby reproducing Figure 2b from Venters and Pugh (2013). In cases in which multiple elements were found in the same sequence, I chose the one closest to the TFIIB peak pair midpoint. 140 (535, 1,883, 2,737) TFIIB peaks had at least one match to a TATA element with zero (one, two, three) mismatched positions. The observed nucleotide patterning misleadingly implies that TATA elements are positionally constrained around TFIIB locations. However, sequences are aligned at TATA matches, instead of TFIIB peaks. I attain a very similar nucleotide distribution by aligning sequences at matches to TATA elements in randomized sequences, obtained by permuting native mRNA-associated TFIIB-bound sequences (Figure 6.1b). Exclusively, matches to the TATA consensus occur considerably less often in randomized sequences (0.4% compared to 2.3% in native sequences).

After calculating the match frequencies of TATA elements around mRNA-associated TFIIB peaks, BRE and INR elements were searched around matches to TATA elements. Although the match frequencies reported by Venters and Pugh (2013) agree with my results (Figure 6.1c, solid lines), their three negative controls match far too infrequently. In fact, I could reproduce one of the controls (60% GC random sequences) by assuming a wrong search space size of one. For this reason, I checked the reported results using the two negative control procedures detailed in Chapter 5. The negative controls closely follow the match frequencies of the four CPEs that had been observed around ChIP-exo peaks (Figure 6.1c, dashed and dotted lines). For example, while 86.5% of sequences around mRNA-associated TFIIB peaks contain a TATAWAWR instance with up to three mismatched positions, 88.3% of randomized sequences contain such a match, putting the false positive rate for TATA elements to ~100%, as opposed to ~20% determined by Venters and Pugh (2013). The results differ from insights gained from TFIIB ChIP-exo measurements in *S. cerevisiae* (Rhee and Pugh, 2012), which found 99% (instead of 41.8%) of mRNA-associated TFIIB locations to contain one or more matches to a TATA element with up to two mismatched positions, while using a 37.3% smaller search space around TFIIB peaks. In summary, the frequent occurrence of the four CPEs in the ChIP-exo peaks is fully explained by their very low specificity (owing to allowing up to three mismatches).

In spite of not being enriched in numbers, CPEs might be located at constrained positions with respect to TFIIB and/or each other, which would provide evidence for their biological relevance in mediating transcription complex assembly at the promoter. On this account, I investigated the local enrichment of CPE pattern matches around mRNA-associated ChIP-exo peaks (relating to Figures 2d and 3d in Venters and Pugh (2013)).

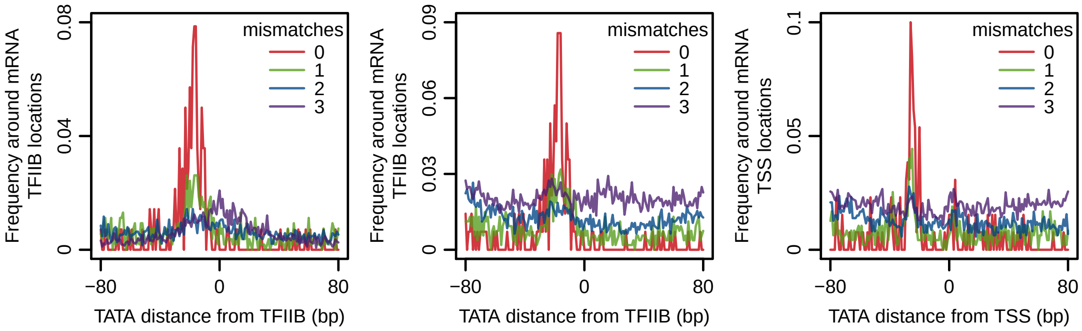
Venters and Pugh (2013) selected only the motif match closest to its assumed location,



(a) Native sequences *(b)* Randomized sequences *(c)* CPE match frequencies

Figure 6.1.: **CPE match frequencies are not enriched around mRNA-associated TFIIB peaks.** **(a)** Color-coded nucleotide sequences (A: red, C: blue, G: yellow, T: green) aligned at TATAWAWR matches (allowing zero, one, two, or three mismatches to the consensus) located around mRNA-associated TFIIB peaks. In cases in which multiple elements were found in the same sequence, I chose the one closest to the TFIIB peak. For visualization reasons, I drew a random sample of 25% of all sequences with TATA matches while preserving the relative frequency in each mismatch group. **(b)** Same as **a** but showing TATAWAWR matches in randomized sequences obtained by permuting native sequences around mRNA-associated TFIIB peaks. **(c)** Match frequencies of CPE patterns in regions around mRNA-associated TFIIB peaks (solid lines) coincide with two negative controls, using either permuted TFIIB-bound sequences (dashed) or native sequences but nonsense patterns obtained by alphabetically sorting individual CPE patterns (dotted).

claiming that this choice had no qualitative effect because multiple elements were rarely found within the same sequence. Paradoxically, I expect about 5.3 chance TATAWAWR matches per sequence on average, when allowing up to three mismatched positions in a search space of 161. By selecting TATA matches closest to TFIIB locations and recording matches dependent on the number of mismatched positions, I could reproduce an enrichment of TATAWAWR with zero or one mismatches (found in 11% of regions) between 30 and 10 nucleotides upstream of TFIIB peaks (Figure 6.2a), as expected (Lenhard et al. (2012), Hartmann et al. (2013)). However, I also distinguish peaks for TATA elements with two and three mismatched positions upstream of and centered around TFIIB peaks, respectively. These artifactual peaks disappear when all motif matches are taken into account (Figure 6.2b), and are therefore unlikely to be biologically meaningful.

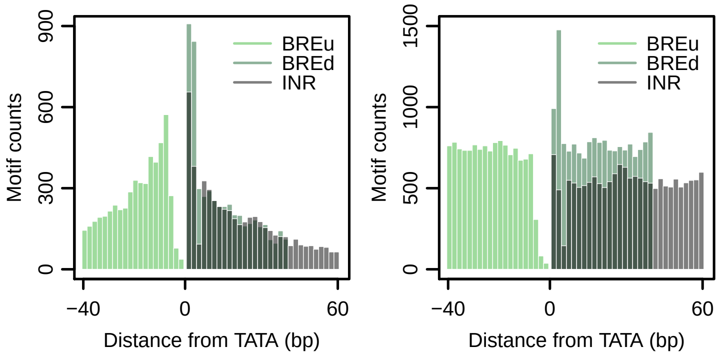


(a) Biased match positions *(b)* Unbiased match positions *(c)* Unbiased match positions

Figure 6.2.: **Distribution of TATA elements around mRNA-associated TFIIB peaks. (a)** Unsmoothed positional distributions of matches to TATAWAWR (allowing zero, one, two, or three mismatches to the consensus) around mRNA-associated TFIIB peaks, normalized by the number of sequences with corresponding motif matches. In cases in which multiple elements were found in the same sequence, I chose the one closest to the TFIIB peak. This selection bias produces artifactual peaks for TATA elements with two or three mismatched positions. TFIIB locations are orientated with respect to their corresponding RefSeq TSS. **(b)** Same as **a** but showing positional distributions of all TATAWAWR matches. TATA elements with zero and one mismatched positions are locally enriched ~20 bp upstream of TFIIB locations. Motif matches with two or three mismatched positions are not enriched. **(c)** Same as **b** but showing TATAWAWR matches around TSSs located next to mRNA-associated TFIIB peaks. The peaks of TATA elements with up to two mismatched nucleotides are sharper when sequences are aligned at their TSSs than when aligned at their TFIIB ChIP-exo peaks.

Interestingly, the peaks of the positional distributions of TATA elements with up to two mismatched nucleotides are sharper when sequences are aligned at their TSSs (Figure 6.2c) than when aligned at their TFIIB ChIP-exo peaks (Figure 6.2b). Hence, available TSS annotation appears to have a higher positional resolution than the TFIIB ChIP-exo data, apart from providing a higher coverage of core promoters.

The CPEs seemed to be positioned relative to each other (Figure 3f and Extended Data Figure 3b in Venters and Pugh (2013)). I repeated the analysis and searched for BRE and INR patterns around TATA elements. By selecting the next (instead of recording all) matches I could reproduce the artifactual peaks of BRE_u and BRE_d upstream and downstream of TATA elements, respectively (Figure 6.3a). In contrast, the INR peak at ~30 bp downstream of TATA elements was not reproducible and might be an artifact arising from the different scale used by Venters and Pugh (2013) for INR counts as compared to BRE counts. Neither INR nor BRE elements are positionally enriched with respect to TATA element midpoints when recording the positions of all matches (Figure 6.3b). The remaining peak of BRE_d immediately downstream of TATA elements can be explained by the overlap of the last three positions of the TATAWAWR motif with the first three positions of the BRE_d motif (RTDKKKK). To be more precise, A overlaps R=[AG] (A or G), W=[AT] overlaps T, and R=[AG] overlaps D=[AGT]. Hence, out of the four possible matches to AWR (without allowing mismatches), 50% also match RTD (random expectation: 9.4%).



(a) Biased match positions *(b)* Unbiased match positions

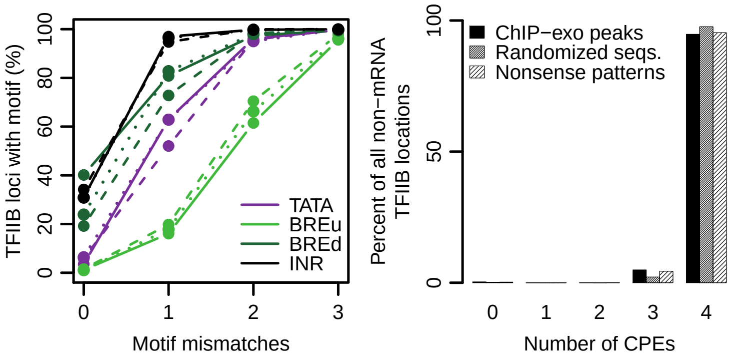
Figure 6.3.: **Distribution of $\mathbf{BRE_u}$, $\mathbf{BRE_d}$, and INR relative to TATA element midpoints.** **(a)** Distribution of matches to BRE_u, BRE_d, and INR (allowing up to three mismatches to the consensus) relative to all TATA matches found around mRNA-associated TFIIB peaks. In cases in which multiple elements were found in the same sequence, I chose the one closest to the TATA instance. This selection bias distorts the true distribution of CPE locations shown in **b**. The distances were calculated between CPE match midpoints. Each bar spans two bp. TATA elements are orientated with respect to their proximal RefSeq TSS. **(b)** Same as **a** but showing positional distributions of all CPE matches relative to TATA element midpoints.

The same reasoning applies to the bottom of BRE_u elements immediately upstream of TATA elements. In this case, the last four positions of the BRE_u consensus (SSRCGCC) and the first four positions of the TATA box consensus are incompatible with each other.

6.1.2. Non-mRNA-associated TFIIB locations

I conducted the previous analyses on the 150,753 non-mRNA-associated TFIIB peaks. Again, the negative controls closely resemble the true CPE pattern matches (Figure 6.4a). In addition, I investigated the claim that the vast majority of regions around the TFIIB peaks contain at least three of the four CPEs. I allowed up to three mismatched nucleotides per CPE as in the original analysis (Venters and Pugh, 2013). The two negative control procedures completely explain the observed CPE match frequencies around the ChIP-exo peaks (Figure 6.4b).

Since Venters and Pugh (2013) found the locations of each CPE peaked at previously defined canonical positions within their search space, they postulate the following core promoter consensus: SSRCGCC-TATAWAWR-N-RTDKKKK-$(N)^{13}$-YYANWYY. However, I found CPEs not to be positioned relative to each other around mRNA-associated TFIIB peaks. Therefore, I tested the predictive power of the core promoter consensus by searching its occurrences in the human genome, allowing up to six mismatch positions. I obtained a total of 208,814 matches, only 458 of which are contained in one of the 159,117 regions of 161 bp centered around the midpoints of TFIIB ChIP-exo peaks. I repeated the analysis with a negative control pattern in which the order of CPEs was inverted (YYANWYY-$(N)^{13}$-RTDKKKK-N-TATAWAWR-SSRCGCC). I attained similar match numbers as with



(a) CPE match frequencies *(b)* Co-localized CPE matches

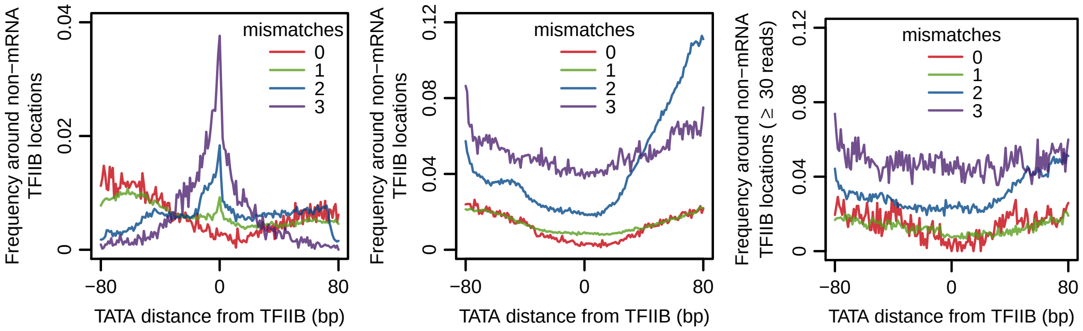
Figure 6.4.: **CPE match frequencies are not enriched around non-mRNA-associated TFIIB peaks. (a)** Match frequencies of CPE patterns in regions around non-mRNA-associated TFIIB peaks (solid lines) coincide with two negative controls, using either permuted TFIIB-bound sequences (dashed) or native sequences but nonsense patterns obtained by alphabetically sorting individual CPE patterns (dotted). **(b)** The fractions of all non-mRNA-associated TFIIB peak regions with zero to four CPE pattern matches are reproduced by two negative controls.

the consensus pattern: 602 of 185,545 matches in the genome lie in one of the TFIIB-bound regions. Hence, either the consensus pattern derived by Venters and Pugh (2013) or the TFIIB ChIP-exo peaks or both lack predictive power for core promoters.

Strikingly, TATAWAWR with up to one mismatch to the consensus is not positionally enriched around non-mRNA-associated peaks (Figure 6.5b), even though such enrichment is observed around mRNA-associated peaks (Figure 6.2b). The selection of TATA matches closest to TFIIB peaks produced artifactual peaks centered around TFIIB locations for TATA elements with one or more mismatches to the TATAWAWR consensus (Figure 6.5a). 68% of non-mRNA-associated TFIIB peaks have less than ten ChIP-exo reads. For this reason, it seemed promising to focus on the 10% strongest non-mRNA-associated ChIP-exo peaks (with a least 30 sequence reads instead of only four required by Venters and Pugh (2013)). However, I obtained similar positional distributions of TATA element matches (Figure 6.5c).

Remarkably, matches to the TATA element show a pronounced asymmetry around non-mRNA-associated TFIIB peaks, in particular when allowing exactly two mismatches to the consensus (Figure 6.5b). Similarly, the distribution of nucleotide frequencies around TFIIB locations is asymmetrical (Figure 6.6a). To further examine this characteristic, I visualized the nucleotide sequences around TATA elements found in non-mRNA-associated TFIIB peaks (Figure 6.6b). Strikingly, stripes become apparent upstream and downstream of aligned TATA elements, possibly arising from the alignment of repetitive sequences.

To investigate the influence of repetitive sequences, I first focused on Alu elements. Alu elements are a family of ~300 bp long repetitive DNA elements belonging to the class of short interspersed nuclear elements (SINEs). The human genome is composed of >10% Alu elements, which therefore are the most abundant of all mobile elements (de Koning et al.

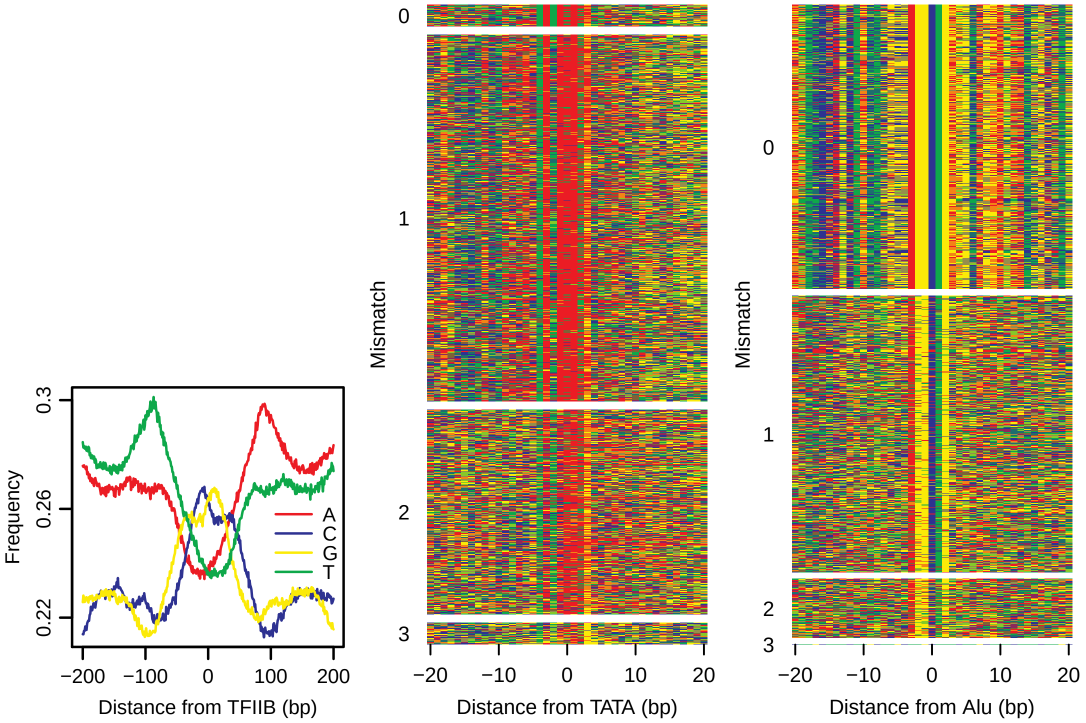


(a) Biased match positions *(b)* Unbiased match positions *(c)* Unbiased match positions

Figure 6.5.: **Distribution of TATA elements around non-mRNA-associated TFIIB peaks. (a)** Unsmoothed positional distributions of matches to TATAWAWR (allowing zero, one, two, or three mismatches to the consensus) around non-mRNA-associated TFIIB peaks, normalized by the number of sequences with corresponding motif matches. In cases in which multiple elements were found in the same sequence, I chose the one closest to the TFIIB peak. This selection bias produces artifactual peaks centered around TFIIB locations. Negative positions are upstream (5′) of the ChIP-exo peak on the Watson or Crick strand, and positive positions are downstream (3′). **(b)** Same as **a** but showing positional distributions of all TATAWAWR matches. The artefactual peaks disappeared. In contrast to Figure 6.2b, TATAWAWR matches with zero and one mismatched positions are not enriched around ~20 bp upstream of TFIIB locations. Note the pronounced asymmetry, in particular when allowing two mismatches to the consensus. **(c)** Same as **b** but showing matches to the TATAWAWR pattern around the 10% non-mRNA-associated TFIIB locations with a minimum number of 30 ChIP-exo reads.

(2011), Deininger and Batzer (2002)). To roughly estimate whether the stripes around non-mRNA-associated TFIIB peaks may originate from aligned Alu elements, I built on the work by Vansant and Reynolds (1995) who detected functional retinoic acid response elements in Alu elements. Since the origin of Alu elements can be traced back to the duplication of the 7SL RNA gene (Ullu and Tschudi, 1984), I searched for the 7SL RNA retinoic acid response element AGGCTG (allowing up to three mismatches) around non-mRNA-associated TFIIB peaks. Figure 6.6c displays the nucleotide sequences around aligned AGGCTG instances. The emerging repeat structure is most evident around perfect matches to the consensus, representing 46% of non-mRNA-associated TFIIB locations. However, Alu elements are transcribed by RNAP III (see e.g. Batzer and Deininger (2002)), and therefore are not expected to bind TFIIB.

To more thoroughly explore the repeat content around non-mRNA-associated TFIIB peaks, I used BLAST (Altschul et al., 1990) to find occurrences of human repeats from the Repbase Update database (Jurka et al., 2005) around TFIIB locations. Repbase Update (release 18.11) provides two subsets of repeat sequences: *humsub* exclusively contains Alu elements (N = 70, median length = 282 bp), while *humrep* comprises repeats other than Alu elements (N = 1043, median length = 576 bp). BLAST found matches to Alu elements in 47% of TFIIB locations (E-value < 0.01). Overall, I found 72% of non-mRNA-associated TFIIB peaks to overlap repetitive regions, by pooling Alu and the remaining repetitive



(a) Nucleotide distributions *(b)* TATAWAWR instances *(c)* AGGCTG instances

Figure 6.6.: **Non-mRNA-associated TFIIB peaks overlap sequence repeats. (a)** DNA nucleotide distributions within 200 bp of non-mRNA-associated TFIIB locations (Watson strand only) show a pronounced asymmetry. Negative positions are upstream (5′) of the ChIP-exo peak, and positive positions are downstream (3′). **(b)** Color-coded nucleotide sequences (A: red, C: blue, G: yellow, T: green) aligned at TATAWAWR matches (allowing zero, one, two, or three mismatches to the consensus) located around non-mRNA-associated TFIIB peaks. In cases in which multiple elements were found in the same sequence, I chose the one closest to the TFIIB peak. For visualization reasons, I drew a random sample of 1% of all sequences with TATA matches while preserving the relative frequency in each mismatch group. Negative positions are upstream (5′) of the ChIP-exo peak on the Watson or Crick strand, and positive positions are downstream (3′). Stripes emerge from the alignment of repetitive sequences. **(c)** Same as **b** but showing sequences aligned at matches to AGGCTG, as proxy for Alu element matches.

elements provided by Repbase Update (N = 1113, median length = 554 bp). These numbers differ greatly from the 5% ChIP-exo peaks reported to fall into repetitive regions in Extended Data Figure 1 (Venters and Pugh, 2013).

In light of the previous finding, the asymmetry around non-mRNA-associated ChIP-exo peaks probably is the result of the high proportion of sequence repeats among ChIP-exo peaks. Hence, the evidence for the biological role of most of the non-mRNA-associated TFIIB ChIP-exo peaks is inconclusive.

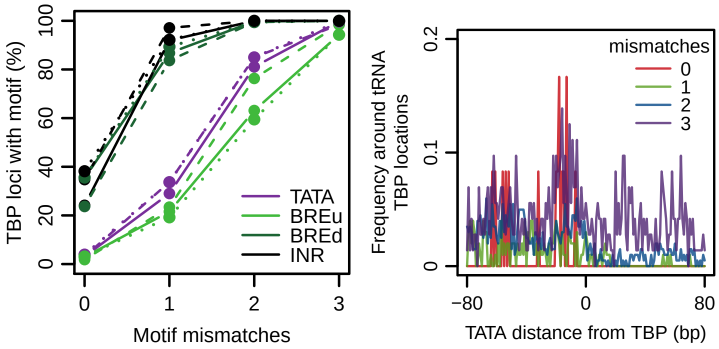
Lastly, Venters and Pugh (2013) showed averaged transcriptional activity and active histone marks around non-mRNA-associated TFIIB peaks as evidence for their biological relevance (Figure 4 in Venters and Pugh (2013)). Since only peaks near known TSSs of coding transcripts were excluded from the non-mRNA-associated TFIIB peaks, transcription initiation complexes of relatively few highly expressed non-coding transcripts and coding transcripts emanating from unannotated TSSs will be among them. Even though they will make up only a small fraction of non-mRNA-associated TFIIB peaks, I note that their contributions to averaged expression and active histone signals could be substantial.

6.2. Core promoter elements around TBP ChIP-exo peaks

In addition to TFIIB, Venters and Pugh (2013) also measured genome-wide binding of TBP using ChIP-exo. Although directly engaging the TATA box, TBP binding sites largely coincide with TFIIB-bound locations, probably due to TBP crosslinking through TFIIB (Venters and Pugh, 2013). In contrast to TFIIB, TBP also targets core promoters of RNAP III-transcribed genes, such as tRNA genes. In the following, I focus on 386 TBP ChIP-exo peaks detected around tRNA-associated TSSs.

6.2.1. tRNA-associated TBP locations

Promoters of tRNA genes have been classically described as TATA-less. More recently, however, instances of TATA elements have also been found to play a role in tRNA transcription initiation (Orioli et al., 2012). Venters and Pugh (2013) expanded this view by describing TATA and BRE elements in almost every sequence around tRNA-associated TBP peaks. Remarkably, the RNAP II-specific INR YYANWYY was not found to be enriched (see their Figure 6c).



(a) CPE match frequencies *(b)* TATA match positions

Figure 6.7.: **CPE patterns are not enriched around TBP peaks. (a)** Match frequencies of CPE patterns in regions around tRNA-associated TBP peaks (solid lines) coincide with two negative controls, using either permuted TBP-bound sequences (dashed) or native sequences but nonsense patterns obtained by alphabetically sorting individual CPE patterns (dotted). **(b)** Unsmoothed positional distributions of all matches to TATAWAWR (allowing zero, one, two, or three mismatches to the consensus) around tRNA-associated TBP peaks, normalized by the number of sequences with corresponding motif matches. TBP locations are orientated with respect to their corresponding RefSeq TSS.

Given the high false positive rates of CPE matches around TFIIB ChIP-exo peaks (Figures 6.1c and 6.4a), I wondered why the extremely degenerate INR was reported to occur only rarely around TSSs of tRNA genes. Therefore, I repeated the previous analysis and examined the 386 TBP ChIP-exo peaks located proximal to annotated tRNA TSSs using the genome coordinates provided in Supplemental Data 4 (Venters and Pugh, 2013). I obtained INR match frequencies that are in conflict with the ones published (Figure 6.7a). In addition to the INR pattern, the INR controls also match far more frequently as stated by Venters and Pugh (2013). Again, match frequencies of all four CPEs were in accordance with match frequencies of the corresponding negative controls.

Finally, I analyzed the positional enrichment of TATA elements around TBP peaks. TATAWAWR with zero mismatches (in 3% of regions) is slightly enriched between 30 to 10 nucleotides upstream of TBP peaks (Figure 6.2b). This might reflect TBP crosslinking to BRF1 (TFIIB-related factor 1), similarly to TBP crosslinking through TFIIB at RNAP II promoters. However, no positional enrichment is detectable for patterns with one or more mismatched positions.

7. Conclusion

Venters and Pugh (2013) suggest that human coding and non-coding transcription initiation complexes form at up to 500,000 regulated core promoters defined by four CPEs, providing an origin for the so-called dark matter RNA of the genome, which potentially houses a substantial portion of the missing heritability. I showed that the CPE motifs are neither enriched nor located at previously defined canonical positions around TFIIB/TBP ChIP-exo peaks, with the exception of well-established TATA box-containing promoters (Lenhard et al., 2012), comprising a subset of mRNA-associated TFIIB peaks. Therefore, the conclusions drawn by Venters and Pugh (2013) are no longer valid. In retracting their study, Venters and Pugh (2014) approve the main claims, which were published simultaneously (Siebert and Söding, 2014).

In *S. cerevisiae*, the expression of genes that are regulated by TATA-less promoters was shown to be independent of sequence-specific TBP-DNA contacts (Kamenova et al., 2014). It thus remains to be elucidated by which mechanisms other general transcription factors or coactivator subunits are directed to TATA-less promoters in order to recruit the transcription machinery.

The biological role of the majority of non-mRNA-associated TFIIB ChIP-exo peaks remains inconclusive. The fact that over two third of TFIIB locations lie within repetitive sequences does not necessarily argue against their existence. Transposable elements have been shown to contain functional transcription factor binding sites and were suggested to be a driving force for shaping gene regulatory networks during mammalian evolution (Sundaram et al., 2014). In addition, transcription from retrotransposon-driven TSSs has been documented (Frith et al., 2008). However, promoters derived from retrotransposons are transcribed by RNAP III unassisted by TFIIB, and are generally found without a TATA box or other strong spatially constrained motifs (Faulkner et al., 2009).

In the analysis, ChIP-exo peak pairs (defined by peaks 0-80 bp in the $3'$ direction from each other and on opposite strands) substantiated by more than four reads, after merging biological replicate data sets, were considered to result from TFIIB binding. This threshold seems to overestimate the signal-to-noise ratio of ChIP-exo experiments, resulting in seriously elevated false discovery rates for factor binding. Furthermore, He et al. (2015) observed signs of PCR library overamplification in the ChIP-exo data produced by Venters and Pugh (2013). TFIIB signals might also originate from incidental non-functional binding of TFIIB as a consequence of transcription-induced accessibility to open chromatin.

Part III.

Uniform transitions of the general RNAP II transcription complex

8. Introduction

Gene transcription begins with the assembly of RNA polymerase (RNAP) II and its initiation factors on promoter DNA. RNAP II then starts mRNA synthesis and exchanges initiation factors for elongation factors, which are required for chromatin passage and RNA processing (Pokholok et al. (2002), Orphanides and Reinberg (2000), Orphanides and Reinberg (2002)). Whereas RNAP II is unphosphorylated during initiation, it is phosphorylated at its C-terminal repeat domain (CTD) during elongation. The CTD is phosphorylated at Ser5 residues in the 5′ region of a gene and at Ser2 residues in the 3′ region (Komarnitsky et al. (2000), Schroeder et al. (2000)). The phosphorylated CTD recruits elongation factors to ensure cotranscriptional RNA processing and chromatin modification (Orphanides and Reinberg (2000), Orphanides and Reinberg (2002), Buratowski (2009), Perales and Bentley (2009), Meinhart et al. (2005), Hirose and Manley (2000)). A genome-wide study has shown that initiation factors are present at all active RNAP II gene promoters (Venters and Pugh, 2009), but it is unknown whether all elongation factors are recruited to all active genes, and no genome-wide studies have examined whether factor recruitment correlates with specific RNAP II phosphorylations.

To address these questions, we used high-resolution genome-wide occupancy profiling by chromatin immunoprecipitation (ChIP) coupled to tiling microarrays (ChIP-chip) of RNAP II, its phosphorylated isoforms, its elongation factors and components of the RNAP II initiation and termination machinery in proliferating yeast cells. Statistical analysis provides strong evidence for a general elongation complex—that is, one composed of all elongation factors—that mediates chromatin transcription and mRNA processing at all RNAP II genes. The general elongation complex is apparently established during a 5′ transition that is completed 150 nucleotides downstream of the transcription start site (TSS), and it is disassembled in two major steps during a 3′ transition around the polyadenylation (pA) site. Transitions are uniform and independent of gene length, type and expression. General elongation complexes are active, as their gene occupancy predicts mRNA expression levels. The results also show that CTD phosphorylation patterns previously observed at individual genes occur globally and that levels of CTD phosphorylation do not correlate with the *in vivo* occupancy of two factors that bind the phosphorylated CTD *in vitro*. This indicates CTD-independent recruitment mechanisms and CTD masking *in vivo*.

This work is the result of a collaboration between the groups of Patrick Cramer and Johannes Söding and was published in Mayer et al. (2010). To facilitate the understanding of the results, I present both experimental materials and methods (Section 9.1), as performed by Andreas Mayer, Michael Lidschreiber and Kristin Leike, as well as data processing and statistical analysis (Section 9.2), conducted by Michael Lidschreiber and me.

9. Materials and Methods

In this Chapter, I describe experimental materials and protocols (Section 9.1) used to generate and computational methods (Section 9.2) developed to process and analyze ChIP-chip data sets of the RNAP II transcription machinery. To confine the length of the thesis, I omit the Supplementary Figures and Tables referenced in this Chapter. These can be consulted in Mayer et al. (2010).

9.1. Experimental materials and methods

I begin introducing the strains (Section 9.1.1) used for ChIP of tandem affinity purification (TAP)-tagged proteins (Section 9.1.2) and RNAP II phospho-isoforms (Section 9.1.3). Subsequently, I detail quantitative real-time PCR (Section 9.1.4) and tiling microarray (Section 9.1.5) protocols. Finally, I outline the protein purifications (Section 9.1.6) discussed in Section 10.4.

9.1.1. Yeast strains and epitope tagging

S. cerevisiae BY4741 (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0) strains containing C-terminally TAP-tagged versions of target proteins were obtained from Open Biosystems. The Spt6ΔC strain lacking the 202 C-terminal residues was generated as described by Dengl et al. (2009). BY4741 untagged wild-type strain (Open Biosystems) was used for ChIP-chip analysis of the different RNAP II phospho-isoforms, for mock IP in ChIP-chip experiments, and for control in TAP purifications.

All epitope-tagged strains were validated. First, gene-specific PCR was performed to confirm that the TAP tag was at the correct genomic position. Second, Western blotting with anti-TAP (PAP, Sigma) antibodies was performed to verify whether the tagged protein of interest was properly expressed. Third, the growth of the various tagged strains compared to the non-tagged wild-type strain was monitored to rule out any influence of the epitope tag on yeast growth. This was done by serial dilutions of the various yeast strains on YPD plates at 30°C for two days (Supplementary Figure 1).

9.1.2. Chromatin immunoprecipitation with TAP-tagged yeast strains

For the yeast TAP-tagged strains, ChIP was performed as described (Aparicio et al., 2005), with modifications. Briefly, yeast strains containing TAP-tagged versions of the proteins, as well as an untagged wild-type strain (for mock IP), were grown in 600 ml YPD medium to mid-log phase (OD_{600} ~0.8). For the Spt6ΔC mutant, the cell number was doubled (1.2 L culture) since the cell lysis efficiency was reduced by twofold.

Yeast cultures were treated with formaldehyde (1%, Sigma F1635) for 20 min at room temperature. Crosslinking was quenched with 75 ml of 3 M glycine for 30 min at room temperature. All subsequent steps were performed at 4°C with pre-cooled buffers and in the presence of a fresh protease-inhibitor mix (1 mM leupetin, 2 mM pepstatin A, 100 mM phenylmethylsulfonyl fluoride, 280 mM benzamidine).

Cells were collected by centrifugation at 4000 rpm (Sorvall SLA-1500 rotor, Sorvall Evolution RC centrifuge) for 5 min, washed twice with 1 × TBS (20 mM Tris-HCL at pH 7.5, 150 mM NaCl) and twice with FA lysis buffer (50 mM HEPES-KOH at pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Na deoxycholate, 0.1% SDS, 1 × protease inhibitor mix). Cell pellets were flash-frozen in liquid nitrogen and stored at −80°C. Cell pellets were thawed on ice and resuspended in 1 ml FA lysis buffer.

Cells were disrupted by vortexing (neoLab 7-2020) in the presence of 1 ml silica-zirconia beads (Roth) for 3 min at full speed at 4°C, followed by an incubation of the sample for 2 min on ice. This was repeated 12 times. The success of the cell lysis was monitored by photometric measurements and the cell lysis efficiency was usually $> 80\%$.

The chromatin was washed twice with FA lysis buffer and sonicated by application of a Bioruptor^{TM} UCD-200 (Diagenode Inc.) to yield an average DNA fragment size of 250 bp, as determined by agarose gel electrophoresis (Supplementary Figure 1). This was achieved by sonifying the sample 35 min at the "high" intensity setting with alternating sessions of 30 sec of sonication followed by 30 sec of resting. 30 µl and 100 µl of the washed and fragmented chromatin samples were saved as input materials and for control of the average chromatin fragment size (see below), respectively.

800 µl of the remaining chromatin sample was immunoprecipitated with 20 µl IgG Sepharose^{TM} 6 Fast Flow beads (GE Healthcare) at 4°C for 4 h on a turning wheel. Immunoprecipitated chromatin was washed 3 times with FA lysis buffer, twice with high-salt FA lysis buffer (500 mM instead of 150 mM NaCl), twice with ChIP wash buffer (10 mM Tris-HCl at pH 8.0, 0.25 M LiCl, 1 mM EDTA, 0.5% NP-40, 0.5% Na deoxycholate) and one time with TE buffer (10 mM Tris-HCl at pH 7.4, 1 mM EDTA). Immunoprecipitated chromatin was eluted for 1 h at 65°C in the presence of the ChIP elution buffer (50 mM Tris-HCl at pH 7.5, 10 mM EDTA, 1% SDS). Eluted immunoprecipitated chromatin as well as input material and material for control of the average chromatin fragment size were subjected to Proteinase K (20 µl of 20 mg/ml Proteinase K from *Engyodontium album*, Sigma P4850) digestion at 37°C for 2 h and reversal of crosslinks (at 65°C overnight).

Samples used for determining the average chromatin fragment size were phenol-extracted twice and ethanol-precipitated overnight. The pellet was resuspended in 20 µl TE buffer (10 mM Tris-HCl at pH 7.4, 1 mM EDTA at pH 8.0) and incubated with 10 µl RNase A/T1

Mix (2 mg/ml RNase A, 5000 U/ml RNase T1; Fermentas) at 37°C for 1 h. The resulting DNA sample was electrophoretically separated on an 1.5% agarose gel.

DNA of the IP, mock IP and input samples was purified with the QIAquick PCR Purification Kit (Qiagen) according to the manufacturer's instructions, except that the final elution was performed with 100 µl DNase-free water. RNA was digested by adding 5 µl of RNase A (10 mg/ml, Sigma) at 37°C for 20 min. DNA was again purified with the QIAquick PCR Purification Kit (Qiagen) according to the manufacturer's instructions. In case of the IP sample, the eluate was concentrated via vacuum manifold to a final volume of 10 µl. The total volume was used for DNA amplification (Section 9.1.5).

9.1.3. Chromatin immunoprecipitation of phosphorylated RNAP II isoforms

For ChIP analysis of the RNAP II phospho-isoforms, chromatin preparation was performed as described in Section 9.1.2, except that it was conducted in the presence of a combination of phosphatase inhibitors (1 mM NaN_3, 1 mM NaF, 0.4 mM Na_3VO_4). For chromatin immunoprecipitation, a set of monoclonal antibodies with strong specificity and affinity for phosphorylated serine residues S5P (3E8), S2P (3E10) and S7P (4E12) were applied. The antibodies generated by Chapman et al. (2007) were a generous gift from Dirk Eick.

It was reported that the amount of antibody used influences the occupancy profiles obtained for RNAP II phospho-isoforms (Kim et al., 2009). Therefore, ChIP experiments were carried out with different amounts of antibody before the ChIP-chip analyses. The results of these antibody titration experiments are shown for three different regions of the two housekeeping genes ADH1 and PMA1 (Supplementary Figure 1d).

Briefly, for the 3E10 antibody, detecting S2-phosphorylated RNAP II, the occupancy behavior remained nearly identical for different amounts of antibody tested (5 to 200 µl), with peak levels towards the 3′-end of the gene. With increasing amount of the antibody, S2P levels showed only a marginal rise in the 5′ regions of ADH1 and PMA1. Acceptable fold enrichments could be reached with 25 µl 3E10 antibody.

The 3E8 antibody, directed against S5-phosphorylated RNAP II, showed the strongest reactivity at the 5′-end of genes, especially when lower amounts were used (5 to 20 µl). With increasing amount of antibody (100 to 200 µl), the signal clearly persists throughout the transcribed region with no clear peak level at 5′-end of ADH1 and PMA1. Since best fold enrichments were observed with 20 µl of 3E8 antibody, this amount was used in further ChIP experiments.

With respect to antibody 4E12, detecting S7-phosphorylated RNAP II, a change in the occupancy behavior could be observed for different amounts tested (5 to 200 µl). With lower amounts of antibody (5 µl), the signal increased towards the 3′-end of genes. When more than 25 µl were applied, the strongest signal could be detected for the 5′-ends of genes. This trend intensified with increasing amounts of antibody, resembling the occupancy behavior of S5P. Acceptable fold enrichments were obtained with 50 µl 4E12 antibody. Therefore, this amount was used in further ChIP experiments.

30 µl and 100 µl of the washed and fragmented chromatin samples were saved as input materials and as control of the average chromatin fragment size, respectively. The remaining 800 µl of sheared chromatin solution was immunoprecipitated with 20 µl, 25 µl, and 50 µl of 3E8, 3E10, and 4E12 rat monoclonal antibodies (cell culture supernatant) at 4°C overnight on a rotating wheel, respectively. 25 µl of Protein A and Protein G Sepharose were added and incubated at 4°C for 1.5 h on a rotating wheel. Immunoprecipitated chromatin was treated as described in Section 9.1.2.

9.1.4. Quantitative real-time PCR

For ChIP experiments, input and immunoprecipitated samples were assayed by quantitative real-time PCR (qPCR) in order to assess the extent of protein occupancy at different genomic regions. Primer pairs directed against promoter, coding and terminator regions of the housekeeping genes ADH1, ACT1 and PMA1, and against a heterochromatic control region of chromosome V, were designed and their PCR efficiencies determined.

All primer pairs used in this study had PCR efficiencies in the range of 95-100%. PCR reactions contained 1 µl DNA template, 2 µl of 10 µM primer pairs and 12.5 µl iTaq SYBR Green Supermix (Bio-Rad). Quantitative PCR was performed on a Bio-Rad CFX96 Real-Time System (Bio-Rad Laboratories, Inc.) using a 3 min denaturing step at 95°C, followed by 49 cycles of 30 s at 95°C, 30 s at 61°C, and 15 s at 72°C. Threshold cycle (Ct) values were determined by application of the corresponding Bio-Rad CFX Manager software (version 1.1) using the Ct determination mode "Regression". Fold enrichment of any given region over an ORF-free heterochromatic region on chromosome V was determined and calculated essentially as described by Fan et al. (2008).

9.1.5. DNA labeling and microarray handling

DNA samples were amplified and re-amplified with GenomePlex© Complete Whole Genome Amplification 2 (WGA2) Kit using the Farnham Lab WGA Protocol for ChIP-chip (O'Geen et al., 2006). DNA quantity and quality control was performed with a ND-1000 Spectrophotometer (NanoDrop Technologies) and was usually larger than 1 µg. In addition, DNA quality was monitored by agarose gel electrophoresis. The re-amplification was performed in the presence of 0.4 mM dUTP (Promega U1191) to allow later enzymatic fragmentation. The enzymatic fragmentation, labeling, hybridization and array scanning were done according to the manufacturer's instructions (Affymetrix Chromatin Immunoprecipitation Assay Protocol P/N 702238).

Enzymatic fragmentation and terminal labeling were performed by application of the GeneChip WT Double-Stranded DNA Terminal Labeling Kit (P/N 900812, Affymetrix). Briefly, re-amplified DNA was fragmented in the presence of 1.5 µl uracil-DNA-glycosylase (10 U/µl) and 2.25 µl APE1 (100 U/µl) at 30°C for 1 h 15 min. The average fragment size was in the range of 50-70 bases, as determined by automated gel electrophoresis on an Experion system (Bio-Rad Laboratories, Inc.) that allowed the analysis of small amounts of DNA. The fragmented DNA was then labeled at the 3′-end by adding 2 µl and 1 µl of terminal nucleotidyl transferase (TdT, 30 U/µl) and GeneChip DNA Labeling Reagent (5 mM), respectively.

5.5 µg of fragmented and labeled DNA were hybridized to a high-density custom-made Affymetrix tiling array (David et al., 2006) (PN 520055) at 45°C for 16 h with constant rotational mixing at 60 rpm in a GeneChip Hybridization Oven 640 (Affymetrix, SantaClara, CA). Washing and staining of the tiling arrays were performed using the FS450_0001 script of the Affymetrix GeneChip Fluidics Station 450. The arrays were scanned using an Affymetrix GeneChip Scanner 3000 7G. The resulting raw data (DAT) image files were inspected for any impairment. The CEL intensity files were used for computational analysis (Section 9.2).

9.1.6. Protein purification and identification

TAP-tagged proteins were purified by the TAP method essentially as described (Puig et al., 2001). Proteins associated with the purified TAP-tagged proteins were identified either by mass spectrometry or by western blot analysis using monoclonal antibodies directed against the HA epitope (3F10, Roche Applied Science), Rpb1 (8WG16, Santa Cruz Biotechnology, Inc.) and Rpb3 (1Y26, NeoClone Biotechnology).

9.2. Computational processing and data analysis

We developed a computational pipeline to process ChIP-chip data sets, including diverse quality control (Section 9.2.1), normalization (Section 9.2.2), and analysis (Sections 9.2.3 to 9.2.6) steps. General considerations about issues related to ChIP-chip data processing can be found in our EpiGeneSys protocol (Siebert et al., 2009).

9.2.1. Replicate measurements, reference samples, and data quality control

It is advisable to measure at least two biological replicates for each factor or condition. This allows to easily identify corrupted measurements. Furthermore, averaging over N replicates reduces the standard deviation of unsystematic noise by a factor of $\sqrt{N}$. Therefore, we analyzed at least two independent biological replicates for each factor (replicate correlations are shown in Supplementary Table 1).

Mock IP and input (genomic background) measurements were used for normalization (Section 9.2.2). Two biological replicates were used for the mock IP (Pearson correlation = 0.65). Due to the very high reproducibility/correlation between input samples of different factors (comparable to factor replicate correlations), we took input samples of three factors (Rpb3, Spt4, and Spt6) and used them as triplicate measurements to normalize all factors, except for the RNAP II phospho-isoforms and Spt6ΔC. The latter were normalized by dividing them by their matched input measurements.

The data processing was performed using R (R Core Team, 2015) and Bioconductor (Gentleman et al., 2004). We used the R package Starr (Zacher et al., 2010) for the data import of Affymetrix CEL files and the conversion into an ExpressionSet, the basic Bioconductor object class for microarray data. In order to avoid processing flawed arrays, quality assessment of each measured array was conducted by inspecting raw image files, density plots, scatter plots, and MA plots (Siebert et al., 2009).

9.2.2. Normalization using mock IP and input measurements

The normalization procedure consisted of three steps. First, we performed quantile normalization between replicate measurements (not between non-replicate measurements). Second, for each condition (including the reference measurements), we averaged the signal for each probe by calculating the geometric average over the replicate intensities. Third, data from all factors were normalized using a combined mock IP plus input reference normalization. The rationale for the last step is explained in detail below.

In ChIP-chip experiments, it is important to correct for sequence-specific and genomic region-specific biases in the efficiency of the various biochemical and biophysical steps. Fragmentation of the chromatin, PCR amplification, immunoprecipitation, labeling, and hybridization to the array can produce strong biases. Although reference-free normalization procedures that can reduce these biases have been suggested (Johnson et al., 2006), the cleanest and most efficient method consists in measuring a reference signal and dividing the true signal intensities, obtained from the immunoprecipitated protein, by the reference intensities. This is in our experience the most important step in data normalization when using arrays that contain short probes, such as the Affymetrix arrays used here, which exhibit a strong GC content bias. As reference signal, one can use the intensities obtained by hybridizing the input (genomic background) fraction to a tiling array or, alternatively, a mock IP. Both procedures are popular and it probably depends on individual experimental conditions and array platforms which of the two performs better in correcting for systematic biases without unduly increasing statistical noise.

In order to understand the effects of different normalization procedures, we developed a simple mathematical model to describe the fragmentation, amplification, labeling, and hybridization biases, as well as the bias introduced by unspecific binding in the ChIP-chip measurements. We found that, using either the mock IP or the matched input as reference signal, one can only correct for some of these effects. Based on our model, we derive a combined normalization using both mock IP and input signal that should correct for all these sources of bias. Crucial to this approach is the fact that the mock IP employs the same antibody as the factor IP, but using a wild type yeast strain with untagged protein. In this way, we measure unspecific binding of the TAP tag-directed antibody that can be used for subtracting the unspecific binding component in the factor IP.

Let x be the genomic coordinate, $B(x)$, $M(x)$, and $S(x)$ the array signals obtained from hybridizing the input (genomic background), the mock IP, and the factor IP, respectively, and $p(x)$ the occupancy profile due to the specific binding of the antibody to the factor of interest. We can model these array signals by

$$B(x) = a_B \times b(x), \tag{9.1}$$

$$M(x) = a_M \times b(x) \times u(x), \tag{9.2}$$

$$S(x) = a_S \times b(x) \times (u(x) + p(x)), \tag{9.3}$$

with a_B, a_M, and a_S unknown scaling constants for the input, mock IP, and factor IP array measurements, respectively, $b(x)$ the input intensity profile describing the fragmentation, amplification, labeling, and hybridization biases, and $u(x)$ the profile describing the effect of unspecific binding of the antibody to other DNA-bound proteins and protein complexes. We seek to obtain the occupancy profile $p(x)$.

Normalizing the factor IP (Equation 9.3) with the input (Equation 9.1), we would get

$$\frac{S(x)}{B(x)} = \frac{a_S}{a_B} \times (p(x) + u(x)), \tag{9.4}$$

showing that the unspecific binding of the antibody will lead to distortions of the ChIP enrichment signal, which will be the more serious the less specific the antibody binds to the factor and the less sequence specificity the factor has in binding to the genomic DNA.

Normalizing the factor IP (Equation 9.3) with the mock IP (Equation 9.2), we would obtain

$$\frac{S(x)}{M(x)} = \frac{a_S}{a_M} \times (1 + \frac{p(x)}{u(x)}), \tag{9.5}$$

which should work better than the previous version in cases where unspecific binding dominates the signal from the specific binding, but which introduces a bias through sequence-specific effects of the unspecific binding (e. g., through a nucleosome density-mediated GC bias of the unspecific binding signal).

Therefore, we introduce a combined normalization for the factor IP (Equation 9.3) using both input (Equation 9.1) and mock IP (Equation 9.2) signals:

$$\frac{S(x) - (a_S/a_M) \times M(x)}{B(x)} = \frac{a_S}{a_B} \times p(x). \tag{9.6}$$

If we assume that $X\%$ of the genomic regions do not bind the factor of interest, we can estimate the factor a_S/a_M as the global $\frac{X}{2}\%$ quantile of Equation 9.5,

$$\frac{a_S}{a_M} = Q_{\frac{X}{2}\%}\left[\frac{S(x)}{M(x)}\right], \tag{9.7}$$

since with this choice $\frac{X}{2}\%$ of the values of the normalized signal (Equation 9.6) will be below zero. Note that, in practice, the value of a_S/a_M depends only weakly on the value of X and can be estimated to much better than a factor two. Only when severely overestimating it (by more than a factor of two) will the correction of unspecific binding be detrimental.

To be able to give absolute occupancy values on a scale between 0% and 100%, we need to estimate the factor a_S/a_B. For this purpose, we assumed that the highest occupancies of our measured factors correspond to 100% occupancy. We estimate the highest genome-wide occupancies using the $Y\%$ quantile, instead of the genome-wide maximum probe signal, to obtain an estimation that is robust against statistical noise. In this study, we chose a quantile of $Y = 99.8\%$ for all factors, which corresponds to the highest-bound ~6,000 probes on our tiling arrays. Then, a_S/a_B is estimated as the $Y\%$ quantile of the factor occupancy (Equation 9.6):

$$\frac{a_S}{a_B} = Q_{Y\%}\left[\frac{S(x) - (a_S/a_M) \times M(x)}{B(x)}\right]. \tag{9.8}$$

Our normalization procedure should improve on the normalization procedures commonly used by correcting the signal for unspecific binding. The approach cannot correct for sequence-dependent effects of crosslinking efficiency. These effects represent, however, an inherent limitation of ChIP-chip and ChIP-seq techniques.

We used the combined mock IP plus input normalization procedure to calculate occupancy profiles for all factors except the CTD phospho-isoforms (S2P, S5P, S7P). Since for these IPs no mock IP measurements with the same antibody could be carried out, they were normalized simply by dividing through genomic input intensities.

9.2.3. Gene-wise occupancy profiles

In order to calculate occupancy profiles along genes or other genomic features the normalized occupancy signal at each nucleotide of the region was calculated as the median signal of all probes overlapping this position (6.5 probes on average). Subsequently, the probe intensities were smoothed using the sliding window smoothing procedure (window half size = 75 bp) implemented in the R package Ringo (Toedling et al., 2007).

9.2.4. Gene selection and gene-averaged profiles

We start with all nuclear *S. cerevisiae* (S288C) protein-coding genes classified as verified or uncharacterized by the Saccharomyces Genome Database (Cherry et al., 2012), corresponding to 5,769 genes. To align gene profiles across entire transcripts, only genes with available TSS and pA site assignments from RNA-seq experiments (Nagalakshmi et al., 2008) were taken into account (4,366 genes). Genes with TSS and pA site measurements downstream and upstream of the annotated ATG and Stop codon, respectively, were excluded. To remove potentially wrongly annotated TSSs and pA sites, we only included genes with TSS and pA site annotations showing a distance less than 200 bp to the corresponding downstream ATG and upstream Stop codon, respectively (3,448 genes). As a result of the limited ChIP-chip resolution and the compactness of the yeast genome exhibiting short intergenic regions (median inter-ORF length = 368 bp, median inter-transcript length = 259 bp), a gene's factor occupancy profile can have spurious contributions from flanking genes. To minimize these spillover effects, we focused on genes with a minimum ORF and transcript distance to flanking genes of 250 bp and 200 bp, respectively (1,786 genes). Furthermore, we restricted the analysis to the 50% highest expressed nuclear protein-coding genes, according to measurements by Dengl et al. (2009) (1,140 genes, ALL set).

We grouped genes into four ORF length classes: Xtremely Short (XS) ranging from 256 to 511 bp, Short (S) 512 to 937 bp, Medium (M) 938 to 1,537 bp, and Long (L) 1,538 to 2,895 bp, comprising 93, 266, 339, and 299 genes, respectively (Supplementary Figure 9a). Gene-averaged profiles within these groups were scaled to median gene length.

We calculated gene-averaged profiles by taking the median over gene-wise profiles (Section 9.2.3). For facilitating the comparison of elongation factor occupancies, in particular their slope in the region around the TSS, we shifted the traces in Figures 10.1e and 10.1f by up to 0.1 on the occupancy scale such that they overlapped in the region [TSS − 250 bp,TSS].

9.2.5. Pairwise profile correlations and correlation network

The following analyses were done using the 4,366 genes with available TSS and pA site annotations (Section 9.2.4). Pairwise Pearson correlations over factor occupancy profiles were calculated between concatenated gene profiles ranging each from TSS − 250 bp to pA site + 250 bp. The correlation-based network was calculated using the Neato algorithm implemented in the Graphviz software (Gansner and North, 2000), which employs an edge-weighted, spring-embedded layout procedure attempting to minimize a global energy function, equivalent to statistical multi-dimensional scaling.

9.2.6. Singular value decomposition

For each of the nine elongation factors f and for each of 4,366 genes (Section 9.2.4), we calculated the 90% quantile of the occupancies within the region [TSS − 250 bp,TSS + 250 bp] as a robust proxy for peak occupancies. This resulted in a $9 \times 4{,}366$ matrix. From each matrix element, we subtracted the average over its row, that is, over its factor. The resulting matrix X_{fg} was subjected to singular value decomposition (SVD), yielding singular values $\sigma_1 \geq \ldots \geq \sigma_9 \geq 0$ and unit-length, orthogonal, singular vectors $u_1,\ldots,u_9,v_1,\ldots,v_9$, such that $X_{fg} = \sum_{i=1}^{9} \sigma_i \times u_{if} \times v_{ig}$. The k^{th} term in this sum, $\sigma_k \times u_{kf} \times v_{kg}$, can explain a fraction $\sigma_k^2 / \sum_{i=1}^{9} \sigma_i^2$ of the data variance (Figure 10.2c, left). We repeated the SVD analysis using all 15 factors (three initiation factors, Cet1, Rpb3, nine elongation factors, and Pcf11) for Figure 10.2c, right. For Figure 10.2d, we included Cet1, Rpb3, three Rpb1 phospho-isoforms, nine elongation factors, and Pcf11.

To reveal correlations contained in the 14.4% of the total variance that was not contributed by the first term in the SVD (Figure 10.2d), we subtracted from the data matrix the values from the first term of the SVD, that is, $X_{fg} - \sigma_1 \times u_{1f} \times v_{1g}$. This resulted in the matrix of residual correlations shown in Figure 10.2d. To ensure that the residual correlations were not caused by spillover effects among neighboring genes, we used a very stringently filtered set of 97 spatially well-separated genes. First, we demanded that the distances to the nearest verified or uncharacterized nuclear ORF, snoRNA, snRNA, ncRNA, CUT or SUT (according to Cherry et al. (2012) and Xu et al. (2009)) be at least 500 bp. Second, we used only genes whose neighboring genomic transcripts were both annotated to be transcribed from the same strand. In this way, we made sure that TSSs and pA sites of neighboring transcripts were well separated. The resulting matrix of residual correlations is very similar to the one shown in Figure 10.2d (Supplementary Figure 8b, bottom), confirming the validity of the analysis on the full set of 4,366 genes. We determined the standard errors of each of the correlation coefficients by taking bootstrap samples from the columns of matrix $X_{fg} - \sigma_1 \times u_{1f} \times v_{1g}$ and obtained errors of ±0.047 and below. Hence, the residual correlations are not caused by statistical noise but rather mirror actual physical and functional associations.

10. Results

This Chapter presents the results of the ChIP-chip data analysis. To confine the length of the thesis, I did not include the Supplementary Figures and Tables referenced in this Chapter. These can be consulted in Mayer et al. (2010).

10.1. Genome-wide profiling reveals RNAP II on the majority of genes

We determined genome-wide occupancy profiles by ChIP in exponentially growing *S. cerevisiae* strains expressing TAP–tagged proteins (Section 9.1.1). ChIP was performed as described in Section 9.1.2. Enriched DNA fragments of an average size of 250 nt (Supplementary Figure 1) were analyzed with tiling microarrays that cover the yeast genome at 4-nt resolution (Section 9.1.5). For data normalization, we developed a procedure that corrects for nonspecific antibody binding by using input measurements as well as mock immunoprecipitations (Section 9.2.2). Data from two or three highly reproducible replicates were averaged (Section 9.2.1). The profile for the RNAP II subunit Rpb3 (Figure 10.1) matched previous profiles (Jasiak et al., 2008) obtained with different strains, experimental protocols and array platforms, but the new profile showed more details (Supplementary Figure 2).

RNAP II was observed at genes encoding proteins, small nuclear RNA and small nucleolar RNA, and at regions producing cryptic unstable and unannotated transcripts (Xu et al., 2009), but was lacking at genes transcribed by RNAP I and RNAP III (Figure 10.1a and Supplementary Figure 3). Of 4,366 yeast genes with annotated TSS and pA sites (Nagalakshmi et al., 2008), 2,465 (56%) showed RNAP II peak occupancies above 20%, consistent with transcription of most of the genome (David et al., 2006).

To average RNAP II profiles over genes, we examined the 50% most highly expressed genes (Dengl et al., 2009) that were at least 200 nt away from neighboring genes. These were sorted into four main length classes, scaled to adjust for length differences and aligned by their TSS and pA sites (Section 9.2.4). The pA site marks the point of RNA $3'$ cleavage and polyadenylation, but transcription continues beyond this site until termination. Consistent with this, the gene-averaged Rpb3 profile revealed RNAP II occupancy through the transcribed region and into the region flanking the pA site on the $3'$ side (Figures 10.1b and 10.1c, and Supplementary Figure 4).

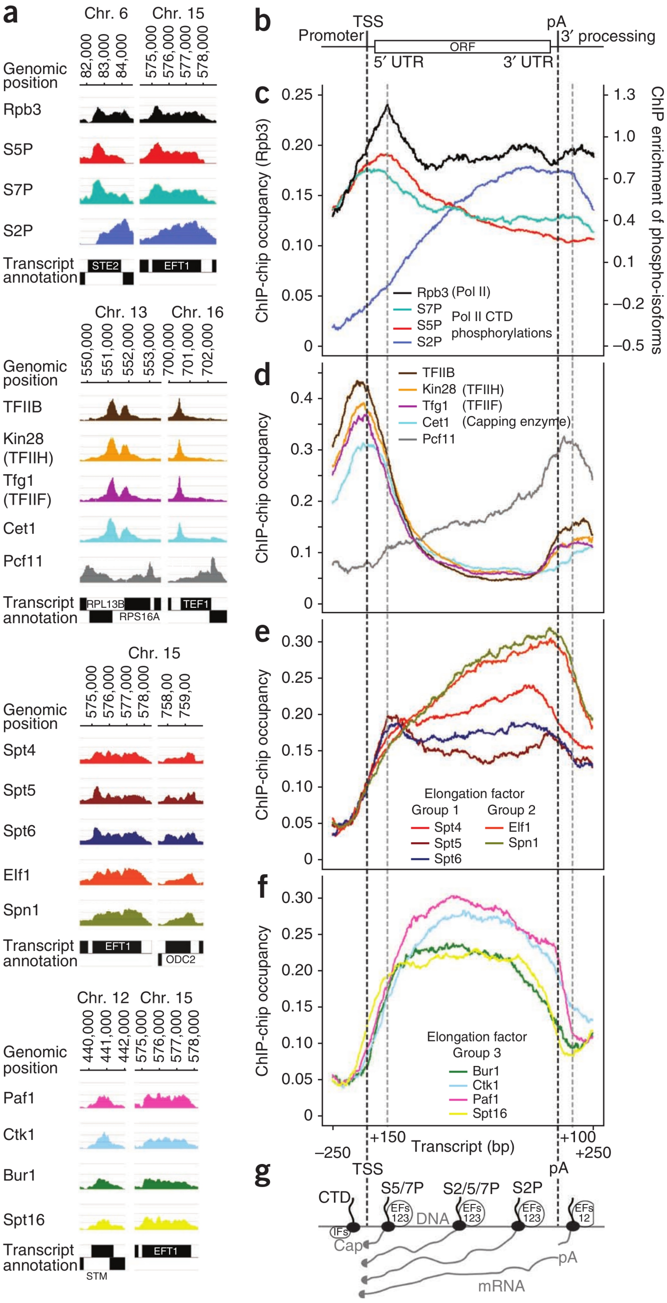


Figure 10.1.: **Genome-wide occupancy profiling of the RNAP II machinery. (a)** Factor occupancy on selected genes. Colored profiles represent normalized factor occupancies smoothed by a 150-nt window running median. The color code is used throughout figures. Black boxes indicate transcripts (David et al., 2006) on the Watson (top) and Crick strands (bottom). **(b)** DNA frame with promoter, 5′ UTR, ORF and 3′ UTR. Dashed black lines indicate the TSS and pA site. Dashed gray lines mark the positions 150 nt downstream of the TSS and 100 nt downstream of the pA site. **(c)** Gene-averaged profiles for the median gene length class (1,238 ± 300 nt, 339 genes) of RNAP II and its phosphorylated forms. Profiles of other length classes are generally similar (Supplementary Figure 9). Occupancies and signal intensities are given for Rpb3 and phosphorylated RNAP II on the left and right y axes, respectively. For details, see Section 9.2. **(d)** Gene-averaged profiles as in **c** for initiation (TFIIB, TFIIF, TFIIH), 5′ capping (Cet1) and termination (Pcf11) factors. **(e,f)** Gene-averaged profiles as in **c** for elongation factors of groups 1 (Spt4, Spt5, Spt6), 2 (Elf1, Spn1) and 3 (Spt16, Paf1, Ctk1, Bur1). **(g)** Cartoon representation of RNAP II (black filled circles) and its CTD (black lines) transcribing along DNA (horizontal gray line) from left to right, to produce mRNA (gray lines). IFs, initiation factors; EFs 123, elongation factors of groups 1, 2 and 3; S2/5/7P, phosphorylation of CTD serines 2, 5 and 7.

10.2. Initiation and termination factors flank the transcribed region

Gene-averaged profiles for the initiation factors TFIIB, TFIIF and TFIIH showed a single strong peak 50–30 nt upstream of the TSS (Figure 10.1d, Supplementary Figure 4, and Supplementary Table 2). This indicates the presence of initiation complexes at promoters and is consistent with a scanning mechanism for TSS location in yeast (Kuehner and Brow (2006), Kostrewa et al. (2009)). TFIIF was found only at promoters and not within transcribed regions, indicating that its reported elongation-stimulatory activity *in vitro* (Renner et al., 2001) is restricted *in vivo* to early RNA elongation and to downstream sites of transient association.

The weaker peaks for initiation factors observed downstream of the pA site are mostly due to residual spillover effects from closely spaced genes on the same strand. When we averaged only over convergently transcribed genes, the peaks were reduced two- to threefold (Supplementary Figure 5). The remaining peaks may indicate gene looping at selected genes.

Occupancy of the capping enzyme subunit Cet1 peaked just downstream of the TSS, consistent with capping when the nascent RNA appears on the RNAP II surface. The symmetric peaks of averaged occupancy of initiation factor and capping enzyme indicate that these factors are restricted to defined locations just upstream and downstream, respectively, of the TSS.

Occupancy of the 3′ processing and termination factor Pcf11 peaked downstream of the pA site, consistent with transcription and completion of mRNA 3′-end formation downstream of the pA site. Thus, representative initiation and termination factors show peak occupancies outside the transcribed region and are apparently not present during mRNA chain elongation.

10.3. Elongation factors enter during a single 5′ transition

Elongation factor profiles did not correlate with profiles of initiation or termination factors (Figure 10.2). Elongation factors were absent at the promoter, but their occupancies sharply increased downstream of the TSS within a narrow window of ~50 nt, indicating coordinated elongation complex assembly during a single 5′ transition (Figures 10.1e and 10.1f, and Supplementary Table 2). Spt16 was an exception, entering ~30 nt further upstream.

Elongation factors showed characteristic distributions over the transcribed region. We observed three distinct profile shapes and used them to group the factors (Figure 10.2b). Group 1 includes Spt4, Spt5 and Spt6, group 2 includes Spn1 and Elf1, and group 3 includes Spt16, Paf1 and the CTD kinases Bur1 and Ctk1.

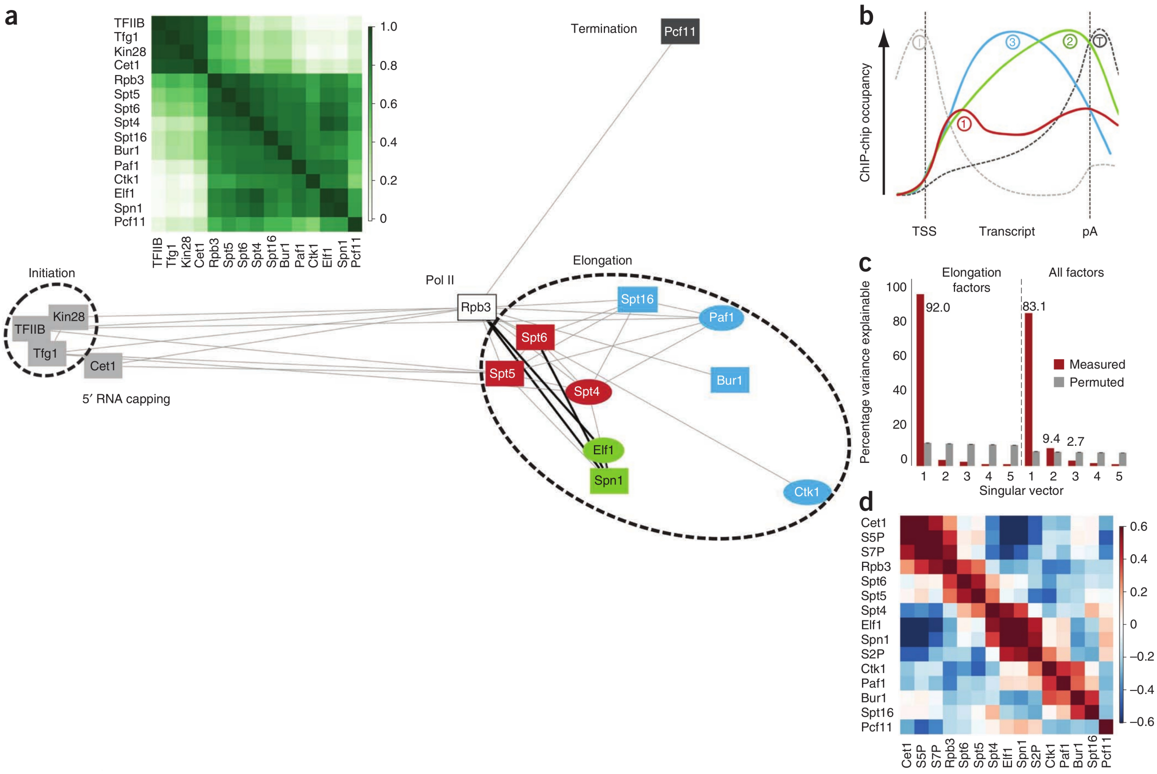


Figure 10.2.: **Statistical analysis indicates a general elongation complex.** **(a)** Correlation analysis of genome-wide occupancy profiles. Initiation factors TFIIB, Tfg1 (TFIIF), Kin28 (TFIIH) and the capping enzyme Cet1 have similar profiles that are distinct from those of the nine elongation factors and the termination factor Pcf11. Elongation factors of groups 1 (dark red), 2 (light green) and 3 (blue) cluster in two dimensions when Pearson correlation coefficients between occupancy profiles are provided as similarity metric. Lines represent direct and functional interactions previously known (gray) or described here (black). Factors represented by ovals are not essential in yeast. **(b)** RNAP II factors can be grouped by their gene-averaged profiles. **(c)** SVD analysis. The contributions of the first five singular vectors to the variance (red) are shown in comparison to a control with randomly permuted matrix elements (gray). SVD reveals that 92% of the variance of peak occupancies of elongation factors at each gene can be explained by strictly covarying factor occupancies as a contribution from the first singular vector (left). When all factors are included, 83.1% of the variance is explained by covariation (right). **(d)** Residual correlations described by all but the first singular vector reveal a modular substructure among factors and phosphorylated RNAP II forms, suggesting physical and functional interactions.

10.4. Spn1 and Elf1 interact within a RNAP II complex

The similar gene-averaged profiles of the poorly characterized factors Spn1 and Elf1 suggested that these factors interact. To test this, we purified Spn1 from yeast using a TAP tag (Section 9.1.6). Spn1 copurified with Elf1 and RNAP II (Supplementary Figure 6), consistent with an interaction between Spn1 and Elf1.

To probe for a direct Spn1-Elf1 interaction, we coexpressed the two factors in bacteria. Spn1 and Elf1 did not copurify after coexpression (data not shown). These results suggest that Spn1 and Elf1 interact indirectly within a RNAP II complex, and the profiling data suggest that their recruitment and functions during the transcription cycle are distinct from those of other elongation factors.

10.5. Elongation factors exit during a two-step 3′ transition

Around the pA site, two steps of a 3′ transition could be distinguished. The two-step transition was most easily seen at genes with high factor occupancies, such as ribosomal protein genes (Figure 10.3 and Supplementary Table 2).

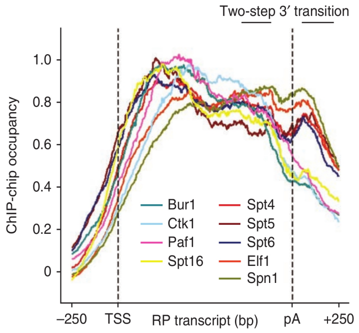


Figure 10.3.: **Two-step 3′ transition observed at ribosomal protein genes.** Shown are averaged elongation factor profiles on selected ribosomal protein (RP) genes. Dashed lines indicate the TSS and pA site. The regions of the two-step 3′ transition are indicated. In contrast to the averaged profiles in Figure 10.1, the very high occupancies did not allow us to align profiles with their promoter minima along the y axis.

Whereas group 3 factor occupancies sharply decreased upstream of the pA site, factors from groups 1 and 2 apparently exited further downstream, suggesting they are present during RNA 3′-end formation and possibly during transcription termination. Spn1 and Elf1 peaked just upstream of the pA site, and 100 nt downstream of this site they were still present at about 80% of their peak occupancies (Figure 10.1e and Supplementary Table 2).

10.6. A general elongation complex for chromatin transcription

High correlations between elongation factor profiles (Figure 10.2a and Supplementary Figure 7) suggested that all elongation factors co-occupy active genes. To investigate this, we measured covariation in the data sets by singular value decomposition (SVD). We calculated peak occupancies for nine elongation factors within 4,366 genes (Section 9.2.6). After subtracting the row mean of the 9 × 4,366 matrix from each element, we subjected the resulting matrix to SVD. The first singular value, which describes strict covariation, explained 92% of the total variance (Figure 10.2c, left, and Supplementary Figure 8a). Thus, elongation factor occupancies covaried over all genes, consistent with a general composition of the elongation complex.

The apparent elongation complex composition and coordinated assembly during a 5′ transition were independent of gene length, expression, function, transcript type, size of

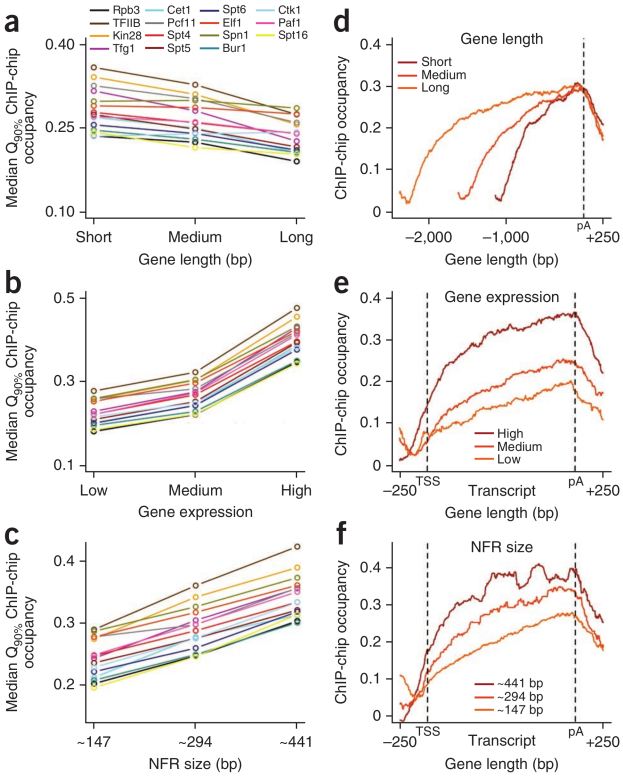


Figure 10.4.: **Transcription complex composition and transitions are independent of gene length, expression and NFR size. (a-c)** Medians of the peak factor occupancies covary between different length classes **(a)**, expression level classes **(b)** and nucleosome-free promoter region (NFR) size classes **(c)** (Hesselberth et al. (2009), Yuan et al. (2005)). $Q_{90\%}$, 90% quantiles of gene occupancies, used as a proxy for peak occupancies. **(d–f)** Gene-averaged profiles of the representative elongation factor Elf1 have shapes and transition points that are independent of gene length **(d)**, expression level **(e)** and NFR size **(f)**. The same holds for all other profiled factors and also for genes grouped by transcript type and functional class (Supplementary Figure 9).

the nucleosome-depleted promoter region and the presence of introns (Figure 10.4 and Supplementary Figure 9). Although differences in the composition of elongation complexes in individual cells cannot be ruled out, these results indicate a general initiation-elongation transition and a general elongation complex composition on RNAP II genes.

10.7. CTD phosphorylation profiles depend on TSS location

To investigate how the observed profiles and transitions correlate with CTD phosphorylation, we determined ChIP-chip profiles for RNAP II phosphorylated at CTD residues Ser7, Ser5 and Ser2 using site-specific antibodies (Chapman et al., 2007) (Section 9.1.3). The averaged profiles revealed broad peaks of Ser7 and Ser5 phosphorylation at around 20 and 120 nt, respectively, downstream of the TSS (Figure 10.1c, Supplementary Figure 4, and Supplementary Table 2). Ser7 and Ser5 phosphorylation decreased over the transcribed region, whereas Ser2 phosphorylation increased, saturated at 600–1,000 nt downstream of the TSS

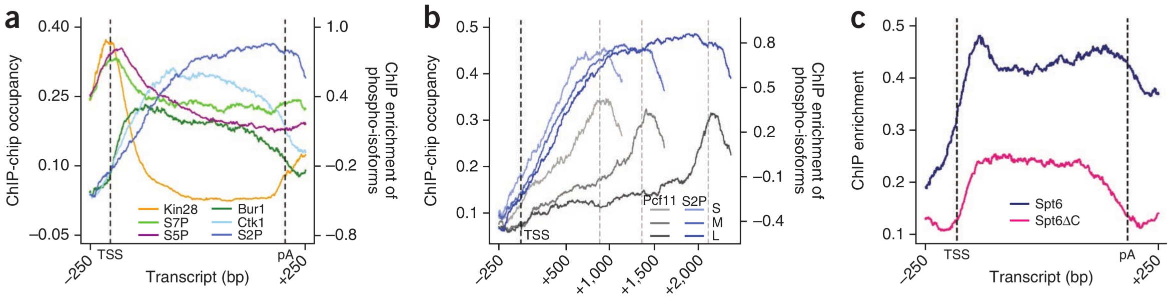


Figure 10.5.: **RNAP II phosphorylation patterns and factor occupancy.** **(a)** Gene-averaged profiles (for length class L only) of CTD phosphorylations, CTD Ser5 kinase Kin28 and Ser2 kinases Bur1 and Ctk1. **(b)** Profiles for Ser2-phosphorylated RNAP II and Pcf11 aligned at the TSS and pA site, respectively, and averaged for length classes short (S), medium (M) and long (L). Occupancy and signal intensity for Pcf11 and S2P are plotted on the left and right y axes, respectively. **(c)** Gene-averaged profiles of Spt6 and the C-terminal deletion variant Spt6ΔC (lacking the 202 C-terminal residues) (Dengl et al., 2009). As 100% occupancy levels are not expected for Spt6ΔC, the y axis shows ChIP enrichments obtained by normalization with input measurements as well as mock IPs (Sections 9.2.1 and 9.2.2) without scaling to 100% occupancy.

and sharply decreased 100–200 nt downstream of the pA site (Figure 10.5, Supplementary Figures 4 and 9, and Supplementary Table 2). The point where full Ser2 phosphorylation was reached did not depend on pA site location, but rather on TSS location (Figure 10.5b).

Although we cannot rule out changes in ChIP efficiency due to the accessibility of the phosphorylated CTD to antibodies, our results overall indicate that CTD phosphorylation patterns previously observed on individual genes (Komarnitsky et al. (2000), Schroeder et al. (2000)) occur globally and depend on TSS location.

10.8. Recruitment of CTD kinases explains CTD phosphorylations

The Ser7 and Ser5 peaks just downstream of the TSS are consistent with the presence of the Ser7 and Ser5 kinase Kin28 (Kim et al. (2009), Akhtar et al. (2009)) just upstream of the TSS (Figure 10.5a). Furthermore, the early peak of Ser7 phosphorylation is consistent with dependence of Ser7 phosphorylation on the co-regulatory Mediator complex (Boeing et al., 2010), which binds at promoters.

The subsequent increase in Ser2 phosphorylation is consistent with the Ser2 kinases Bur1 and Ctk1 (Murray et al. (2001), Liu et al. (2009)) entering during the 5′ transition and remaining present in transcribed regions (Figure 10.5a). Thus, specific CTD phosphorylations can be explained by the recruitment of corresponding specific CTD kinases at defined distances from the TSS.

Unfortunately, we were not able to obtain satisfactory profiles of CTD phosphatases to compare their recruitment with the observed decreases in CTD phosphorylation.

10.9. CTD phosphorylation and factor recruitment

To clarify the relationships between CTD phosphorylations and factor occupancies, we subjected all profiles to SVD and correlated residual profiles lacking the contributions of the first SVD term; the eliminated first term described 85.6% of the covariation of factor occupancies (Supplementary Figure 8b).

Ser7 and Ser5 phosphorylation correlated with the occupancy of the capping enzyme Cet1 (Figure 10.2d), as expected from binding of capping enzyme to Ser5-phosphorylated CTD *in vitro* (Rodriguez et al., 2000). As Ser5 phosphorylation levels peaked more than 100 nt downstream from the Cet1 peak, the capping enzyme may already be bound when the first Ser5 residues are phosphorylated. Cet1 occupancy dropped very sharply further downstream, whereas Ser5-phosphorylation levels remained high, suggesting an active mechanism to release the capping enzyme from the CTD.

Ser2-phosphorylation correlated strongly with Spn1 and Elf1 occupancy (Figure 10.2d and Supplementary Figure 8), suggesting these factors are stabilized within the elongation complex by Ser2 phosphorylation.

10.10. Possible CTD masking and CTD-independent recruitment

CTD Ser2 phosphorylation did not correlate with occupancy of Pcf11 and Spt6, although these factors bind to the Ser2-phosphorylated CTD *in vitro* (Yoh et al. (2007), Barillà et al. (2001)). Pcf11 was recruited mainly at the pA site (Figure 10.5b and Supplementary Figure 9), consistent with the known role of Pcf11 in RNA 3′ processing. This may be explained if the Ser2-phosphorylated CTD becomes accessible to Pcf11 only after pA site passage and 3′ RNA cleavage. Alternatively, Pcf11 crosslinking may be increased by cooperative interactions of factors and RNA around the pA site or by conformational changes in the elongation complex.

Spt6 entered early, during the 5′ transition, suggesting a recruitment mechanism independent of CTD Ser2 phosphorylation. To investigate this, we determined the ChIP-chip profile of a variant of Spt6 lacking its CTD-binding C-terminal domain (Spt6ΔC), using a yeast strain that expresses only a truncated Spt6 lacking the last 202 residues (Dengl et al., 2009). Deletion of the Spt6 CTD-binding domain led to much less recruitment of Spt6 but did not abolish its entry during the 5′ transition (Figure 10.5c and Supplementary Figure 10). Thus, Spt6 is apparently recruited in a CTD-independent manner during the 5′ transition, but full recruitment requires the CTD-binding domain. The CTD-binding domain was required for retaining Spt6 until the pA site was reached (Figure 10.5c), consistent with Spt6's preference for binding the Ser2-phosphorylated CTD.

These results indicate that binding of a factor to the phosphorylated CTD *in vitro* cannot predict factor recruitment *in vivo*. This suggests that the CTD may be transiently masked and its accessibility regulated, and that factor recruitment includes CTD-independent and

CTD phosphorylation type–independent recruitment.

10.11. No evidence for promoter-proximally stalled RNAP II

In higher eukaryotes, RNAP II is often stalled early during elongation near the promoter (Nechaev et al. (2010), Zeitlinger et al. (2007), Core et al. (2008)) and can be released by activators (Rahl et al., 2010). Our data do not provide evidence for the presence of such promoter-proximally stalled RNAP II in growing yeast. Genes with stalled RNAP II would show Ser5 but not Ser2 phosphorylation, or at least more Ser5 than Ser2 phosphorylation. However, we did not find genes with a peak for Ser5 phosphorylation and no peak for Ser2 phosphorylation (data not shown). SVD analysis of initiation and elongation factor profiles revealed a high covariance of 83.1% (Figure 10.2c, right, and Supplementary Figure 8c), suggesting that initiation complexes are generally efficiently converted to elongation complexes.

Although Rpb3 occupancy peaks around 150 nt downstream of the TSS (Figure 10.1c), this does not indicate polymerase stalling, as stalling generally occurs closer to the TSS. Instead, this RNAP II peak may be explained by the $5'$ transition being slow because of capping, phosphorylation and assembly events, leading to an apparent accumulation of RNAP II. Alternatively or additionally, the peak may reflect transient pausing of RNAP II between the +1 and +2 nucleosomes, which are positioned around 40 and 210 nt downstream of the TSS, respectively (Jiang and Pugh, 2009).

Our data from exponentially growing yeast also do not show evidence for polymerase peaks upstream of the TSS, as observed in yeast during stationary growth (Radonjic et al., 2005).

10.12. General elongation complexes are productive

To investigate whether general elongation complexes are active on most genes, we correlated averaged Rpb3 and elongation factor occupancies with mRNA levels (Dengl et al., 2009). The mRNA level should be proportional to the mRNA synthesis rate of a single elongation complex times its occupancy, divided by the mRNA decay rate (see legend of Figure 10.6). For constant mRNA synthesis and decay rates, we would therefore expect a linear dependence of occupancy on mRNA level, corresponding to a slope of one in a log-log plot.

We found a robust correlation of 0.65 between the log occupancy and the log mRNA levels (Figure 10.6a); however, the slope was only ~0.5. A correlation of 0.71 was obtained when we used the distance-filtered gene set. This shows that increased mRNA levels are due not only to greater elongation complex occupancy, but also to a higher ratio of mRNA synthesis rates over decay rates.

The same dependence leads to a high correlation of 0.79 between the observed average occupancy and the expected occupancy calculated from the mRNA level (Figure 10.6b). These correlations indicate that most general elongation complexes are active in producing mRNA and that occupancy of genes by the general elongation complex is a good predictor for gene expression level.

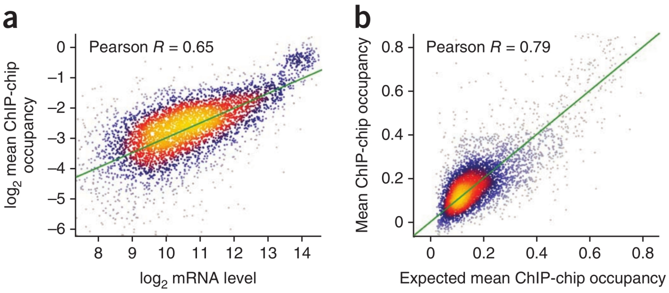


Figure 10.6.: **Elongation complex occupancy predicts mRNA expression.** **(a)** The logarithm of the average elongation factor and Rpb3 transcript occupancy is highly correlated with the logarithm of mRNA levels. For constant mRNA synthesis rates (elongation complex speeds) v and decay rates r, we expect a linear relationship between elongation factor occupancy o and mRNA levels c, because at equilibrium the rates of mRNA synthesis and decay are equal, and thus $ov = cr$. This would result in a linear dependence between $\log(o)$ and $\log(c)$ with slope one: $\log(o) = \log(c) + \log(r/v)$. The actual slope of 0.49 (green) implies that the ratio of RNAP II speed to decay rate increases slightly with increasing mRNA level ($v/r \propto c^{0.51}$). For ribosomal protein genes, v/r is about threefold higher than average. **(b)** The averaged transcript occupancy of Rpb3 and the nine elongation factors is highly correlated with the occupancy expected on the basis of the relationship between occupancies and observed mRNA levels in **a**.

11. Discussion

Here, we have established an improved protocol for obtaining high-resolution genome-wide occupancy profiles for components of the RNAP II transcription machinery, to investigate the chromatin-transcription cycle *in vivo* (Figure 10.1g). We demonstrate that RNAP II elongation factors, which are required for chromatin passage and RNA processing on individual genes, associate with all transcribed RNAP II genes in proliferating yeast cells. Elongation factors show three distinct, nonrandom patterns of distribution over genes, and these distribution patterns are independent of gene length, type, function or nucleosome structure. The underlying general elongation complex is established and disassembled during uniform transitions at the beginnings and the ends of genes. Elongation factors enter during a sharp $5'$ transition just downstream of the TSS and exit in a two-step $3'$ transition around the pA site. Our genome-wide RNAP II phosphorylation profiles match patterns observed at individual genes and are explained by the recruitment of specific CTD kinases at defined distances from the TSS. Analysis of CTD phosphorylation profiles does not support the existence of promoter-proximally stalled RNAP II in growing yeast. CTD phosphorylation is not predictive of the recruitment of factors that bind the phosphorylated CTD *in vitro*. Instead, we obtained evidence that CTD accessibility is regulated by transient CTD masking and that recruitment mechanisms are CTD independent. Finally, occupancy of genes by the general elongation complex predicts the resulting mRNA levels, suggesting that most or all elongation complexes are active.

Published biochemical and genetic data suggest that the $5'$ transition corresponds to a coordinated conversion of a general initiation complex to a general elongation complex. The conversion includes initiation factor dissociation, which liberates the RNAP II clamp domain (Chen et al., 2007) for binding Spt5 (Hirtreiter et al., 2010). Spt5 could coordinate entry of group 1 factors because it binds Spt4 *in vitro* (Guo et al., 2008) and associates with Spt6 *in vivo* (Lindstrom et al., 2003). Group 1 factors could recruit group 2 factors, as Spt6 binds Spn1 (Lindstrom et al. (2003), Krogan et al. (2002)). Consistent with this, group 1 factors interact genetically with Elf1 (Prather et al., 2005) and, as we show here, Spn1 and Elf1 interact within a RNAP II complex. Recruitment of group 3 factors may commence with CTD Ser5 phosphorylation; this recruits Bur1 (Qiu et al., 2009), which in turn phosphorylates Spt5, thereby recruiting Paf1 (Liu et al. (2009), Zhou et al. (2009), Laribee et al. (2005)). Spt16 enters around 30 nt upstream from other elongation factors (Figure 10.1f and Supplementary Table 2), perhaps because it binds to the +1 nucleosome (Stuwe et al., 2008); this is consistent with its role as a histone chaperone (Belotserkovskaya et al., 2003). Initiation factors are not present when the 5′ transition is completed around 150 nt downstream of the TSS, consistent with Ctk1 promoting dissociation of initiation factors (Ahn et al., 2009).

The general two-step 3′ transition we observed is consistent with ChIP data obtained at individual genes. The early exit of group 3 factors Paf1, Ctk1 and Bur1, and the continued presence of Spt4, Spt5 and Spt6, have previously been observed (Kim et al. (2004), Keogh et al. (2003)). Our results, however, show that the reported Spt16 occupancy downstream of the pA site (Kim et al., 2004) does not occur globally. Also, our observation of peak levels of the bona fide 3′-processing factor Pcf11 downstream of the pA site challenges the idea of an early loading of 3′-processing factors in the 5′ region of genes (Glover-Cutter et al., 2008). We observed continued presence of group 1 and group 2 factors downstream of the pA site, consistent with their reported roles in mRNA 3′ processing (Kaplan et al., 2005), mRNA export (Yoh et al., 2007) and re-establishment of chromatin structure after RNAP II passage (Kaplan et al., 2003).

Our results also provide insights into the role of CTD phosphorylation during transcription complex transitions and in the coordination of transcription-coupled events. First, peak levels of Ser7 and Ser5 phosphorylation were generally observed in the 5′ regions of genes, and peak Ser2 phosphorylation in the 3′ regions of genes. Second, the 5′ transition occurs before any substantial Ser2 phosphorylation, suggesting that the assembly of the general elongation complex is independent of Ser2 phosphorylation, consistent with the observation that the Ser2 kinase Ctk1 is not required for association of elongation factors with transcribing RNAP II (Ahn et al., 2004). Third, peak levels of Ser2 phosphorylation are always reached 600–1,000 nt downstream of the TSS, regardless of the position of the pA site. This argues against a role of Ser2 phosphorylation in triggering the 3′ transition, although Ser2 phosphorylation is required for cotranscriptional 3′ RNA processing (Ahn et al., 2004). Fourth, the recruitment of Pcf11 and Spt6, which both bind the Ser2-phosphorylated CTD *in vitro*, cannot be explained solely by factor binding to the Ser2-phosphorylated CTD *in vivo*. Instead, late Pcf11 entry suggests CTD masking within the transcribed region and an increase in CTD accessibility upon RNA cleavage at the pA site, allowing for Pcf11 binding. Furthermore, Spt6 enters during the 5′ transition even when it lacks its CTD-binding domain, indicating that a CTD-independent recruitment mechanism exists. The CTD-binding domain seems to be more important for retaining Spt6 until the pA site is reached than for recruiting it during the 5′ transition.

In conclusion, our data support the following view of a productive chromatin transcription cycle. The initiation complex forms ~30–50 nt upstream of the TSS and contains unphosphorylated RNAP II and initiation factors. The complex then scans for the TSS downstream, begins RNA synthesis and triggers RNA 5′ capping. Next, the complex is converted into a general elongation complex during a sharp and efficient but presumably slow 5′ transition that is completed around 150 nt downstream of the TSS, where Ser5 phosphorylation levels peak. During subsequent elongation, Ser2 phosphorylation increases until it reaches peak levels 600–1,000 nt downstream of the TSS. During a two-step $3'$ transition, a group of elongation factors exits upstream of the pA site, whereas another group persists downstream, where it is joined by additional factors such as Pcf11, resulting in mRNA $3'$ processing and transcription termination.

11.1. Follow-up studies and computational framework applications

Our processing, quality control, normalization and analysis pipeline was essential to a number of further ChIP-chip studies with the purpose of uncovering the intricate interplay between various factors and the core RNAP II transcription machinery.

The RNAP II CTD is a structural feature that provides a “landing pad” for numerous regulatory factors. However, as discussed in Section 10.10, Figure 10.5b indicated that the CTD may also be transiently masked and its accessibility be regulated. In fact, Mayer et al. (2012) could demonstrate that CTD heptad repeats phosphorylated at the tyrosine residue (Tyr1) can impair termination factor recruitment to RNAP II. Subsequently, the cleavage and polyadenylation factor subunit Glc7 was discovered to be the Tyr1 phosphatase required for triggering transcription termination (Schreieck et al., 2014). The general initiation-elongation transition of RNAP II (Sections 10.3 and 10.6) was further elucidated by Lidschreiber et al. (2013), suggesting that cap completion stimulates productive Pol II elongation.

In a collaboration with the group of Katja Sträßer, our computational framework could finally be applied to analyze ChIP-chip binding profiles of factors involved in the coupling of mRNA synthesis to mRNA export via direct binding to the RNAP II CTD (Meinel et al., 2013).

Part IV.

Appendix

A. Supplementary material (Part I)

A.1. Modeling nucleotide interdependencies within core binding sites

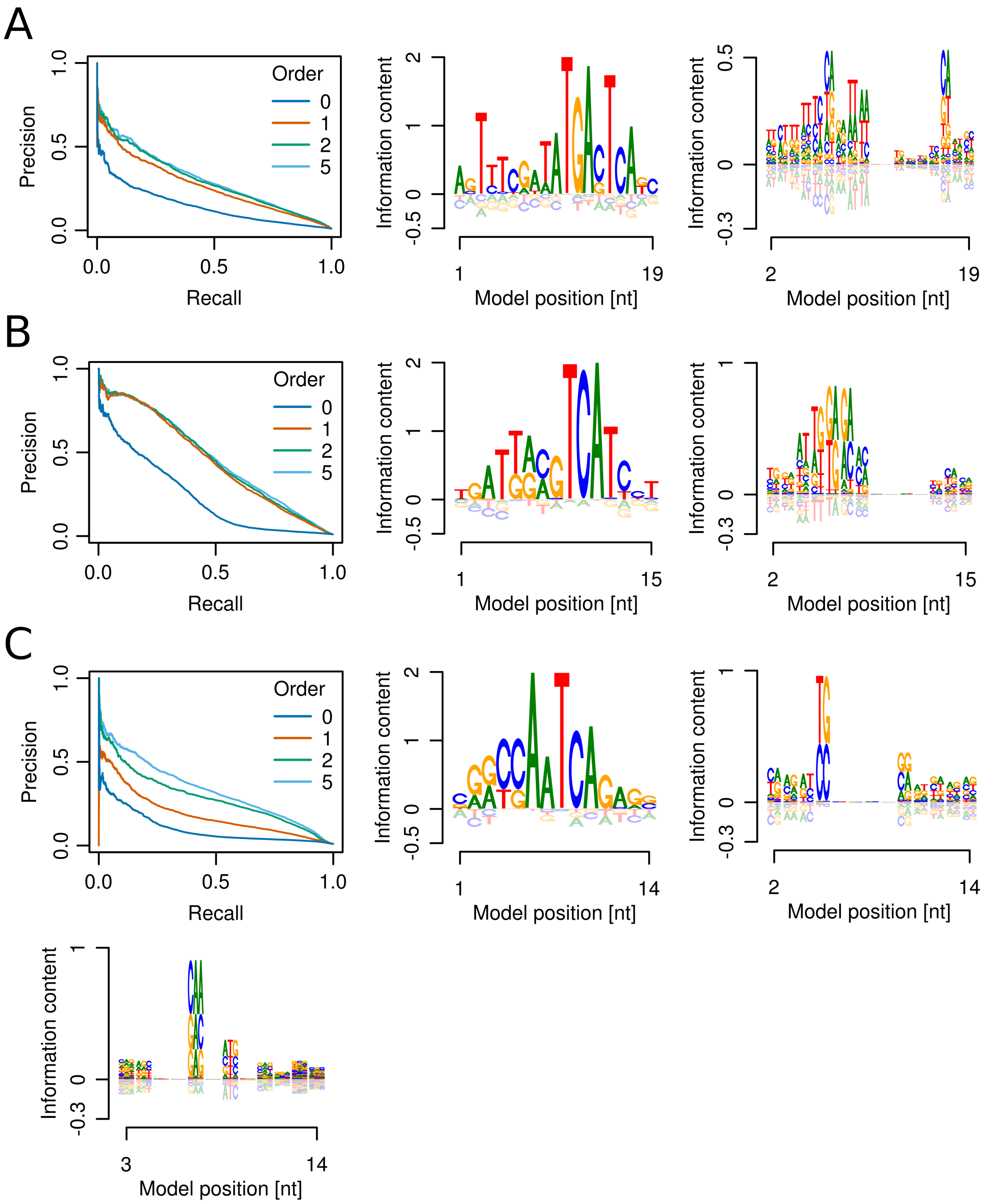


Figure A.1.: (Continued on the following page.)

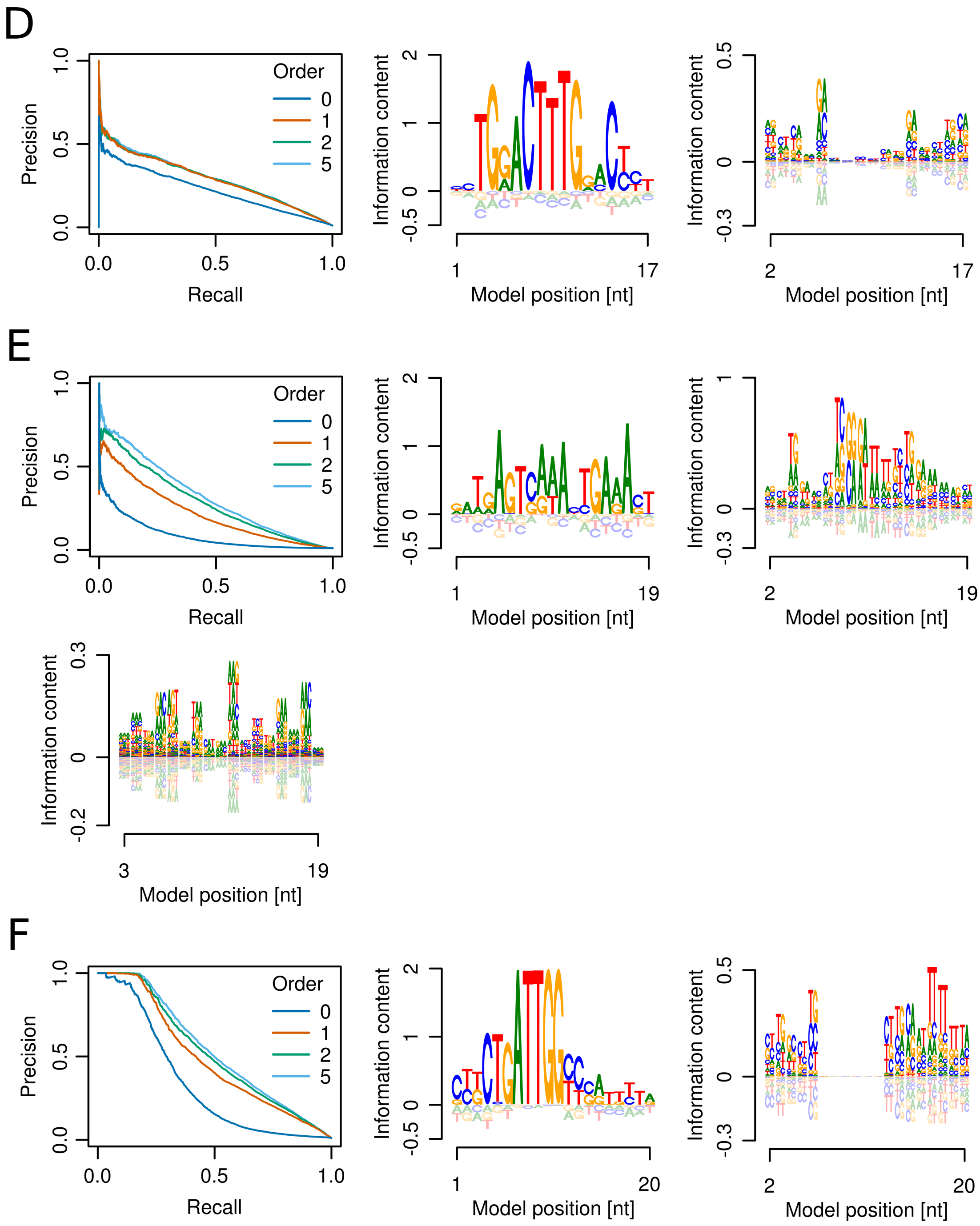


Figure A.1.: (Continued on the following page.)

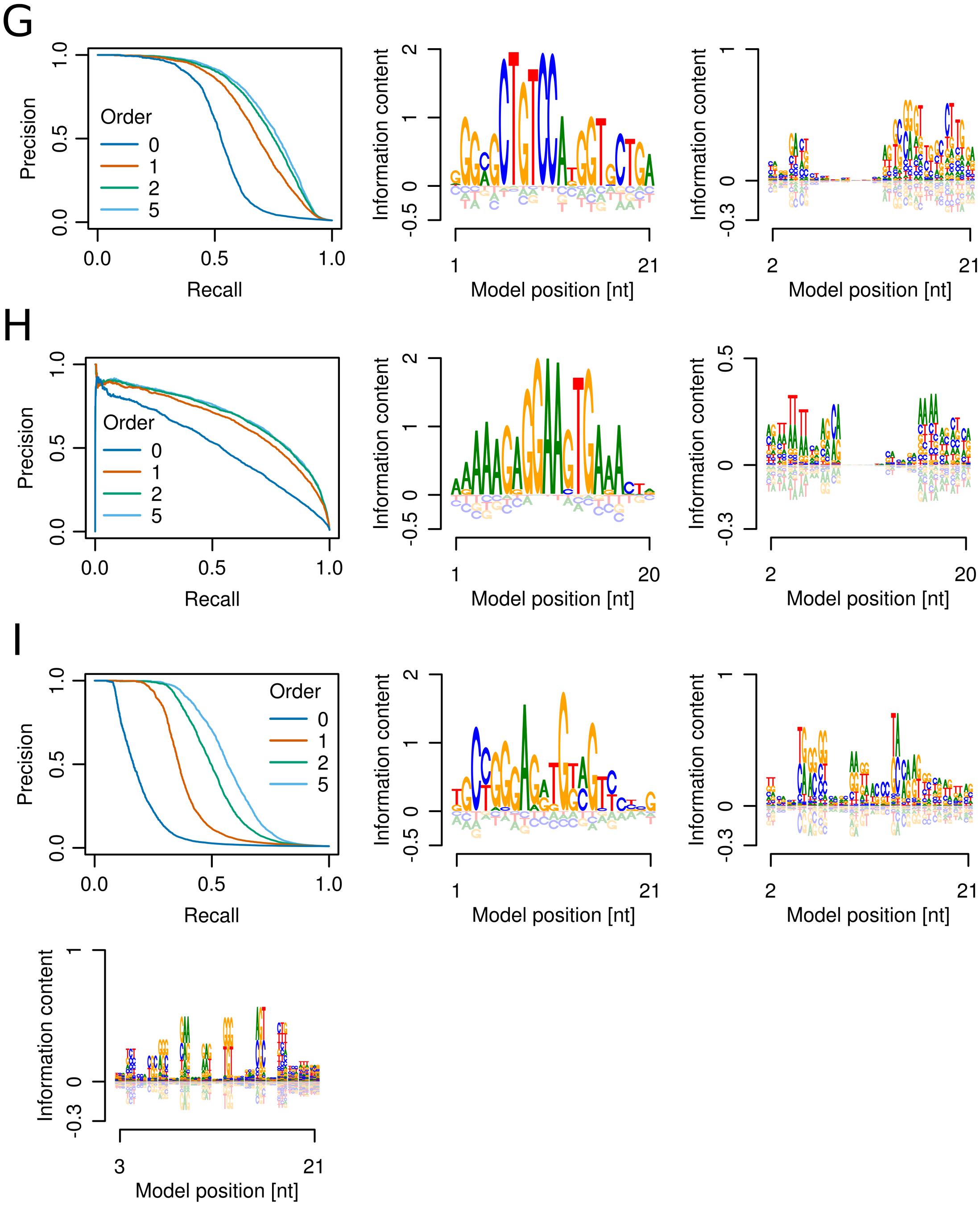


Figure A.1.: **Core binding site iIMMs: further examples. (A)** BATF models learned in GM12878 cells. Precision-recall curves (left) calculated using iIMMs of increasing order. 0th-order (middle) and 1st-order (right) sequence logos depict 2nd-order iIMM. Sequence logos show BATF models learned from all sequences. **(B-I)** Same as **A** but showing **(B)** c-Jun models learned in HepG2 cells, **(C)** c-Fos models learned in K562 cells, **(D)** HNF4α models learned in HepG2 cells, **(E)** IRF4 models learned in GM12878 cells, **(F)** NF-YB models learned in K562 cells, **(G)** NRSF models learned in PFSK-1 cells, **(H)** PU.1 models learned in GM12891 cells, and **(I)** ZnF143 models learned in H1-hESC cells. The sequence logos of the 2nd-order iIMMs in **C**, **E**, and **I** are shown up to the 2nd order.

A.2. Incorporating nucleotides flanking core binding sites

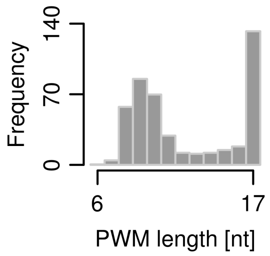


Figure A.2.: **The length distribution of XXmotif's PWM models.** Due to runtime requirements of XXmotif, the maximum number of PWM positions is limited to 17. Note that 2^{nd}-order, eight-bp-elongated iIMMs improve 2^{nd}-order, non-elongated iIMMs independent of the number of initial iIMM positions (Figure 3.3C).

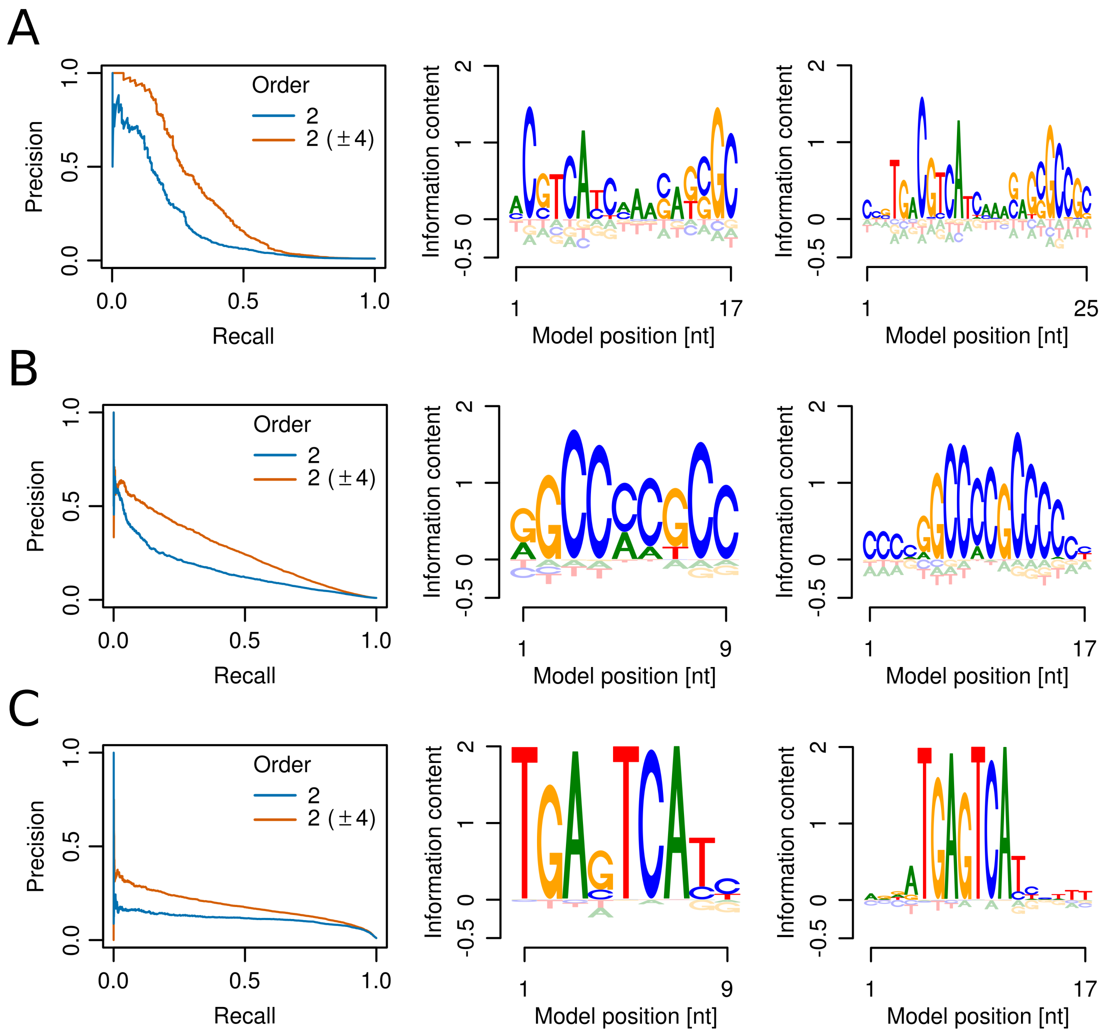


Figure A.3.: **The impact of nucleotides flanking core binding sites: further examples. (A)** GR models learned in HepG2 cells. Precision-recall curves (left) calculated using 2nd-order iIMMs that differ in size by four positions flanking the core binding site on each side (±4). The 0th-order sequence logo depicts 2nd-order (middle) and extended 2nd-order (right) iIMMs. Sequence logos show GR models learned from all sequences. **(B,C)** Same as **A** but showing **(B)** IRF1 models learned in K562 cells, and **(C)** c-Fos models learned in Mcf-10a cells.

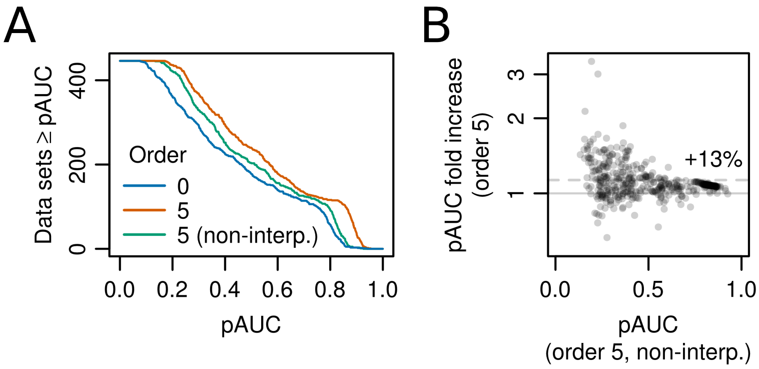


Figure A.4.: **iIMMs of transcription factor binding sites outperform iMMs.** The performance of 5th-order, eight-bp-elongated iIMMs and non-interpolating (non-interp.) iMMs is compared by showing **(A)** cumulative distributions of the partial area under the ROC curve (pAUC), up to a false positive rate of 5%, over all 446 ChIP-seq data sets, and **(B)** the increase in pAUC for single ChIP-seq data sets. The y-axis in **B** is shown in log scale. The dashed line in **B** indicates the mean fold increase.

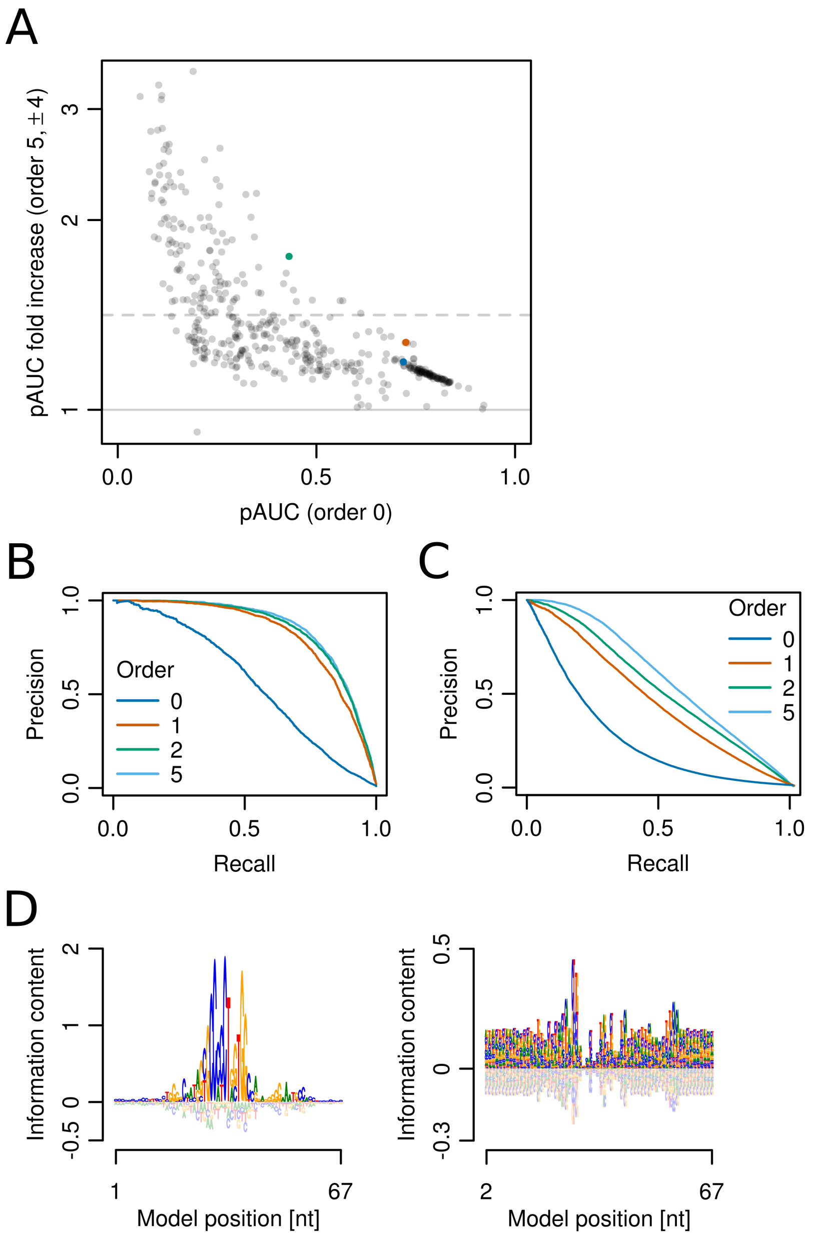


Figure A.5.: **Extended CTCF models. (A)** Same as Figure 3.3D but highlighting the comparison of pAUCs obtained by CTCF models learned in Mcf-7 cells (blue). In addition, I compare pAUCs calculated using PWM models and 5th-order iIMMs that were both extended by 25 positions to each side of initial models and either learned and tested on the top 5,000 (orange) or all 66,592 (green) ChIP-seq peaks available in Mcf-7 cells. **(B)** Same as Figure 3.2A but showing 50-bp-elongated CTCF models. **(C)** Same as **B** but showing CTCF models learned and tested on all 66,592 ChIP-seq peaks. Note that the number of negative sequences in **C** exceeds 13 times the number of negative sequences in **B**. **(D)** 0th-order (left) and 1st-order (right) sequence logos of 2nd-order iIMM from **C**. Sequence logos show CTCF model learned within first cross-validation fold. Note that the CTCF logo is reverse complementary to that in Figure 3.2A.

A.3. Predicting pioneer transcription factor binding affinities

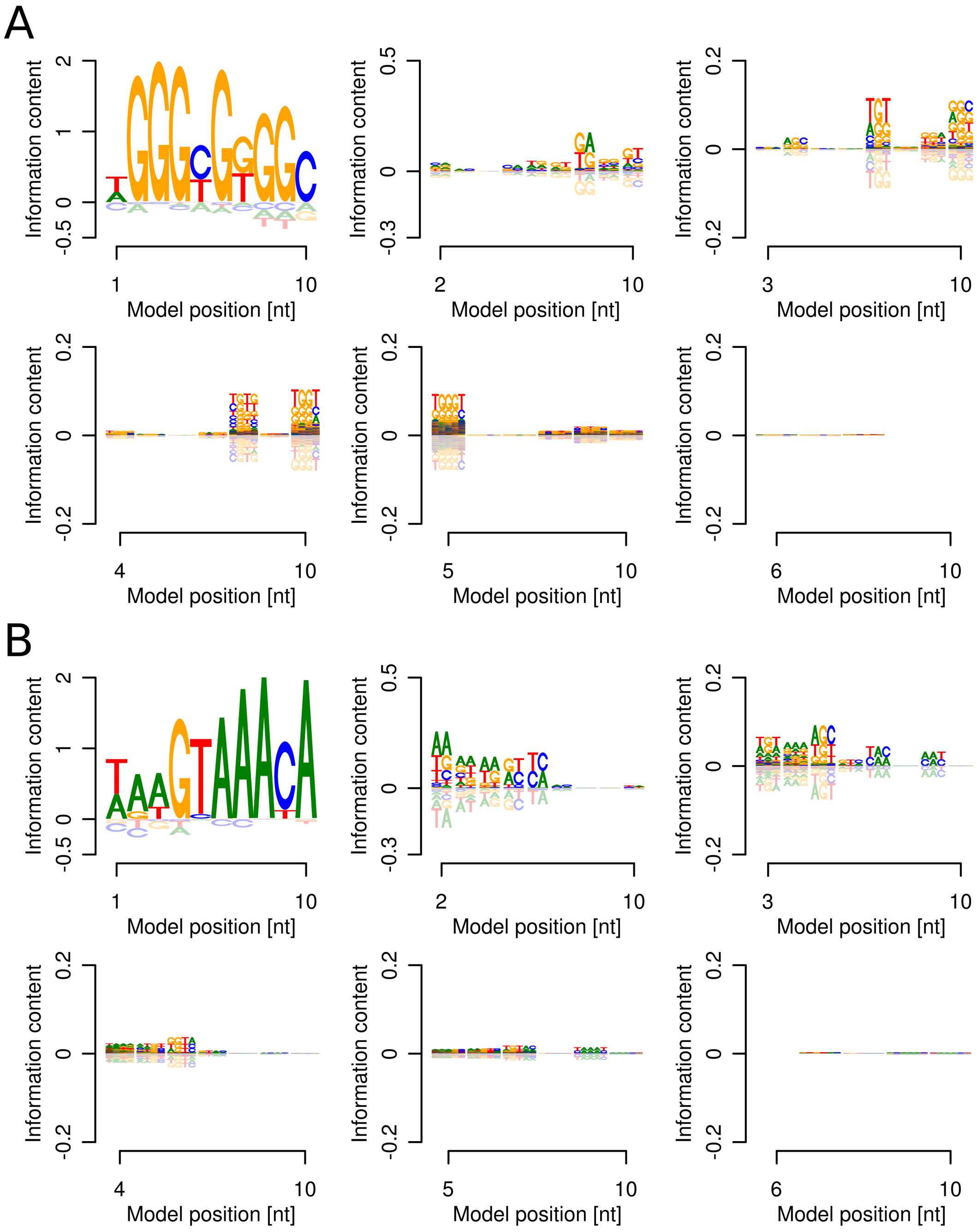


Figure A.6.: **Higher-order sequence logos of pioneer transcription factor iIMMs.** 5th-order iIMM of **(A)** Klf4 and **(B)** FoxA2, shown from 0th up to 5th order (top left to bottom right).

A.4. Modeling nucleotide interdependencies within complex regulatory regions

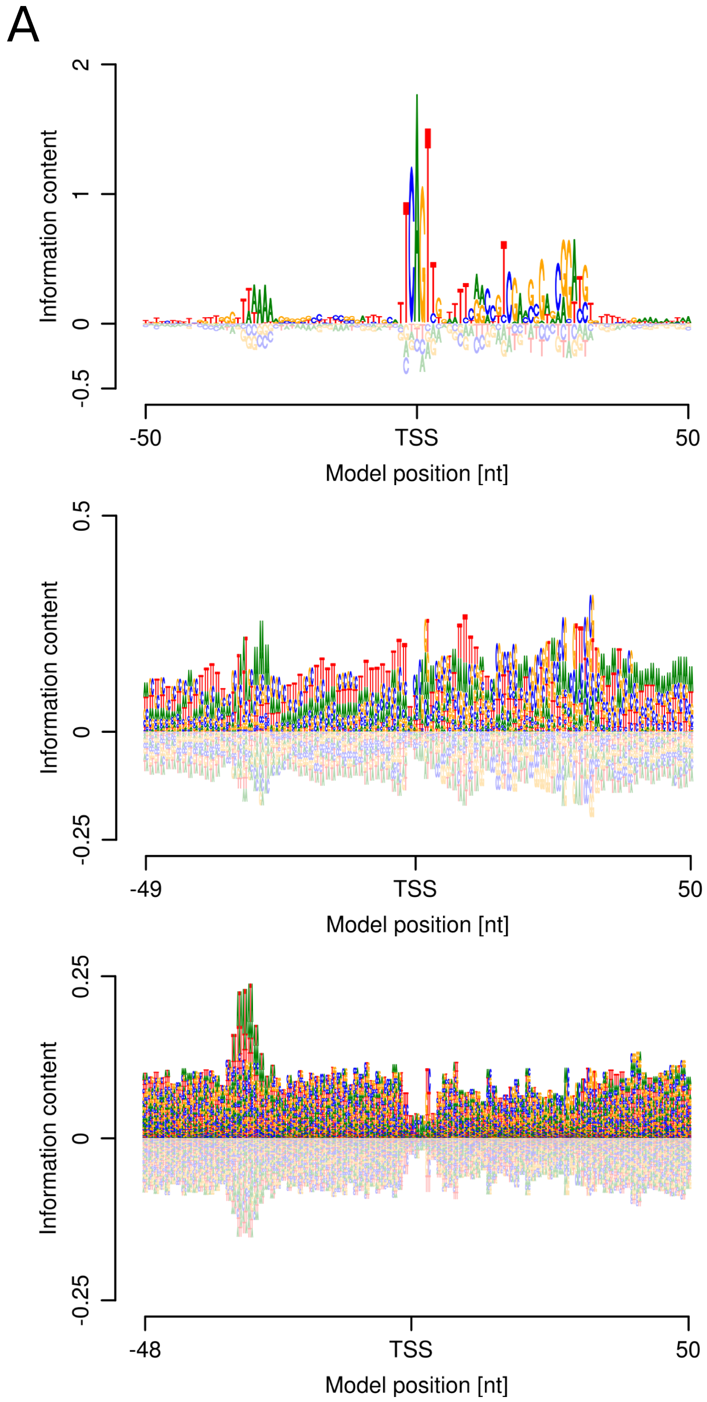


Figure A.7.: (Continued on the following page.)

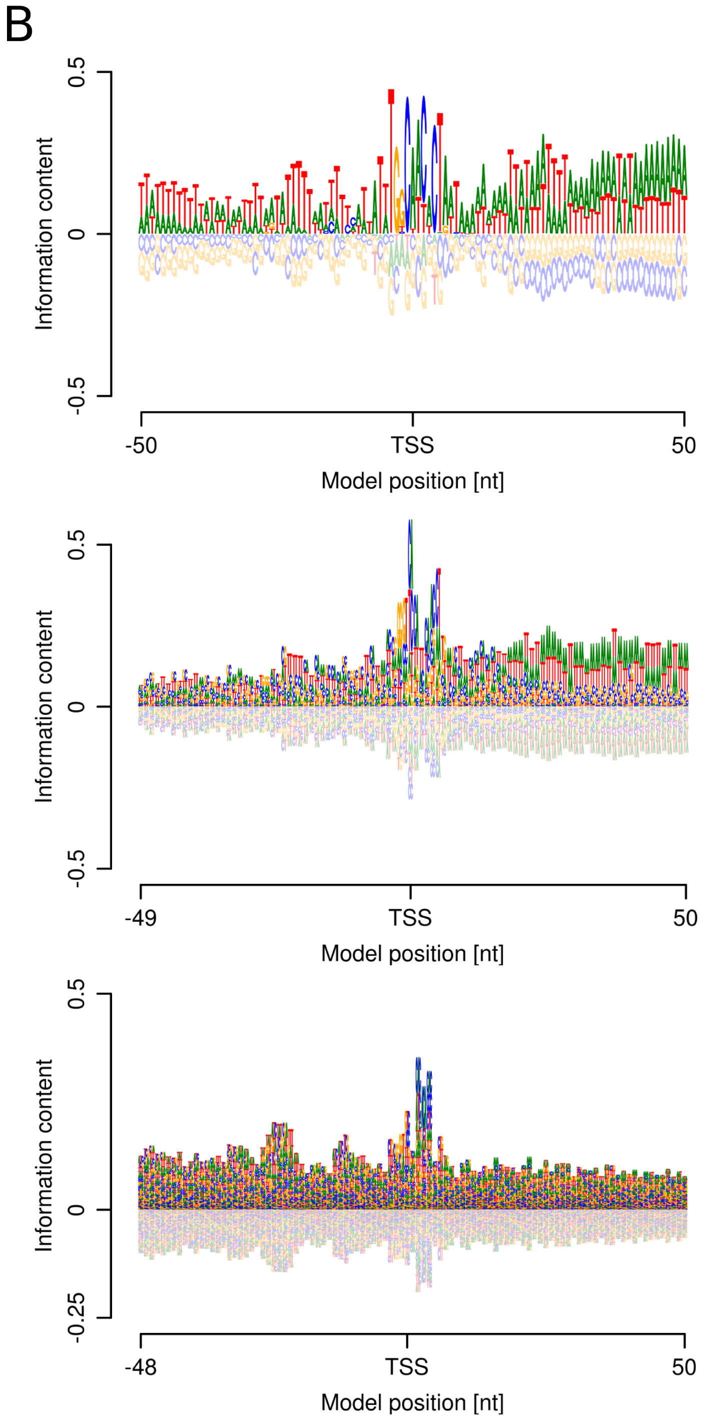


Figure A.7.: (Continued on the following page.)

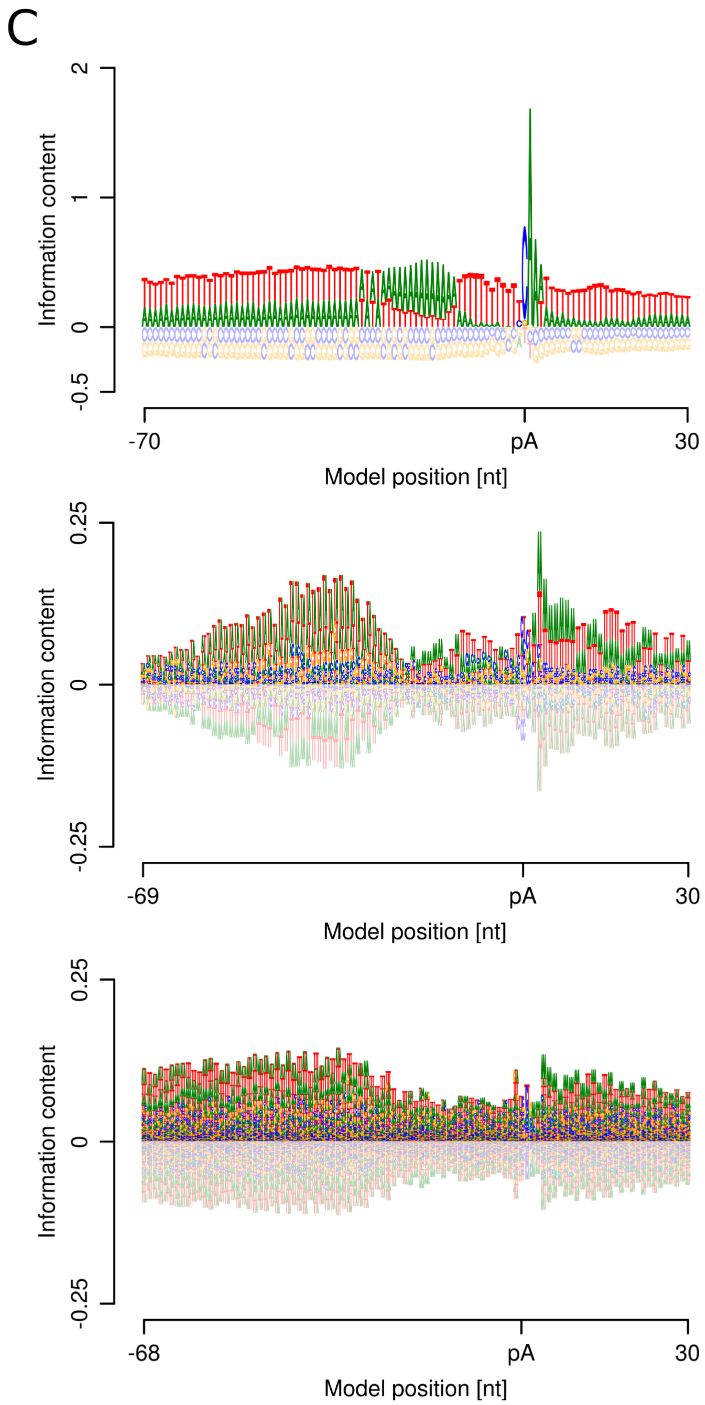


Figure A.7.: **Higher-order sequence logos of complex regulatory regions.** Complete higher-order sequence logos showing 2$^{\mathrm{nd}}$-order iIMMs learned from all **(A)** NP and **(B)** BP core promoters in *D. melanogaster*, and **(C)** pA sites in *S. cerevisiae*. Panels depict (top) 0$^{\mathrm{th}}$-order, (middle) 1$^{\mathrm{st}}$-order, and (bottom) 2$^{\mathrm{nd}}$-order sequence logos.

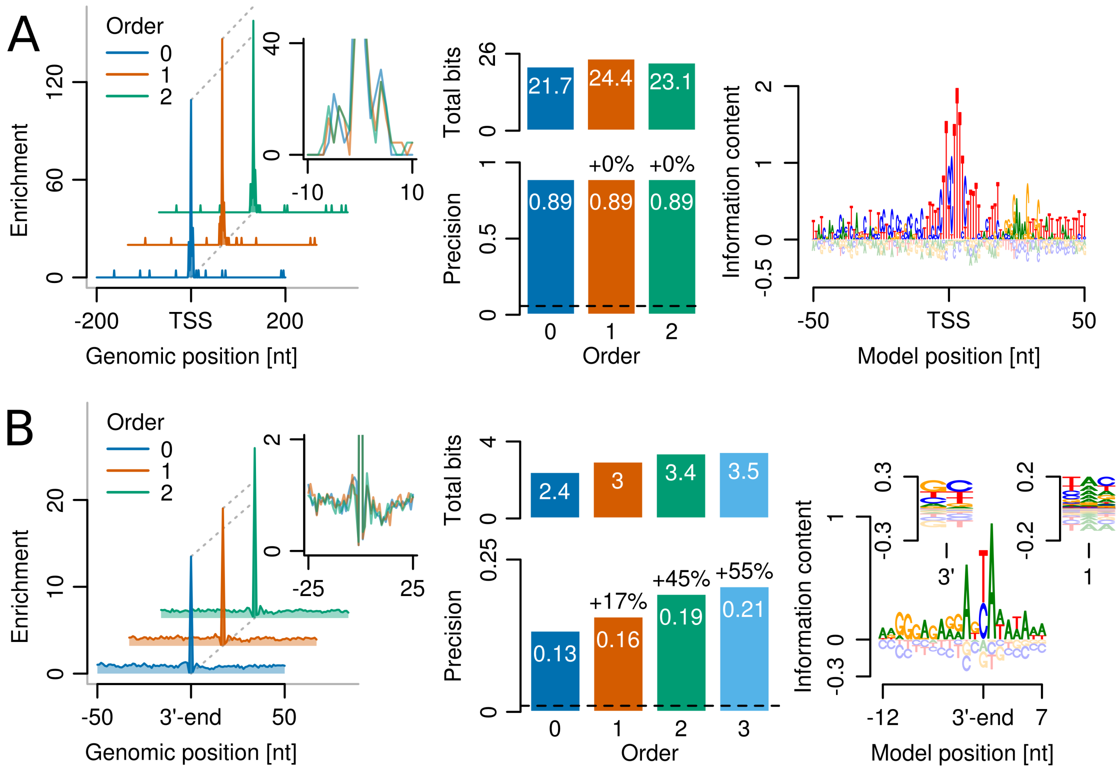


Figure A.8.: **Higher-order iIMMs predict locations of complex regulatory regions. (A)** Same as Figure 3.7A but showing models of RP core promoters from *D. melanogaster*, learned and predicted within 11 bp of measured TSSs. The 0$^{\text{th}}$-order sequence logo depicts the PWM model (right). **(B)** Same as Figure 3.9 but showing models of RNAP pause sites from *B. subtilis*, predicted within zero bp of measured pause sites. Logo insets show 1$^{\text{st}}$-order and 2$^{\text{nd}}$-order contributions at and two bp downstream of the 3′-end of the transcript, respectively.

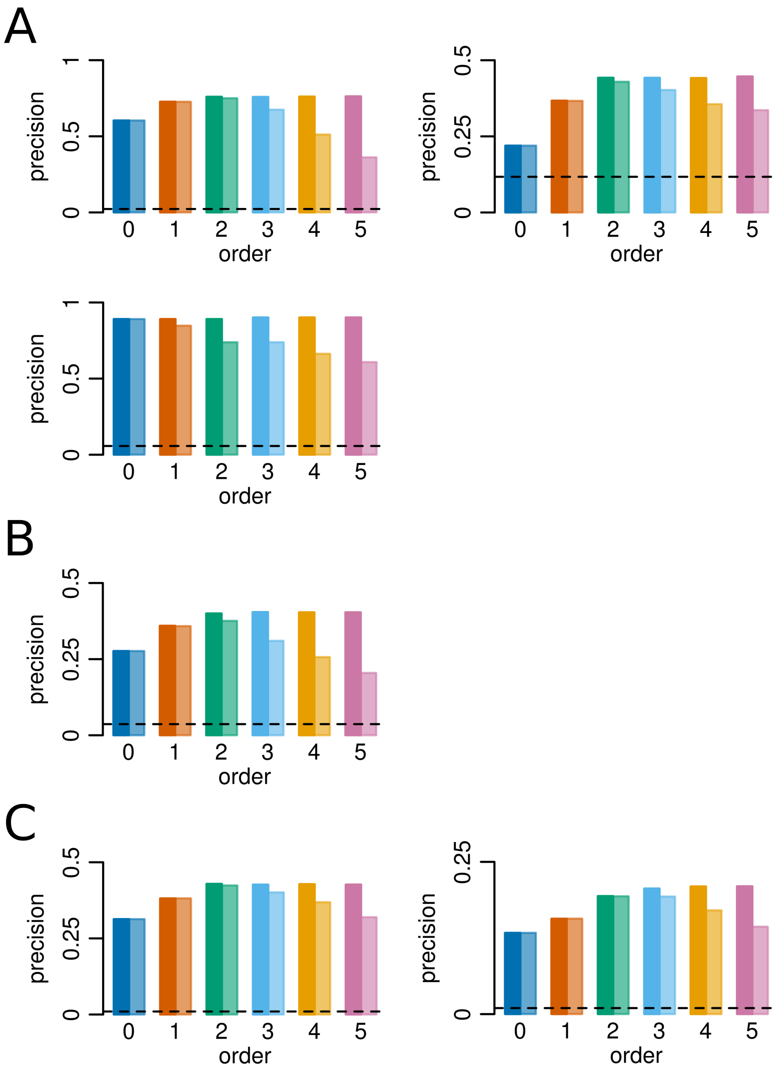


Figure A.9.: **iIMMs of complex regulatory regions outperform iMMs.** The dark and light bars show the precision of iIMMs and iMMs of increasing order, respectively. The dashed line indicates the baseline precision, as achieved by a random predictor. Precision of models learned from **(A)** NP (left), BP (right), and RP (bottom) core promoters in *D. melanogaster*, **(B)** pA sites in *S. cerevisiae*, and **(C)** RNAP pause sites in *E. coli* (left) and *B. subtilis* (right). In contrast to iMMs, iIMMs are not prone to overfitting when learning models of higher order.

A.5. Modeling PAR-CLIP crosslink sites

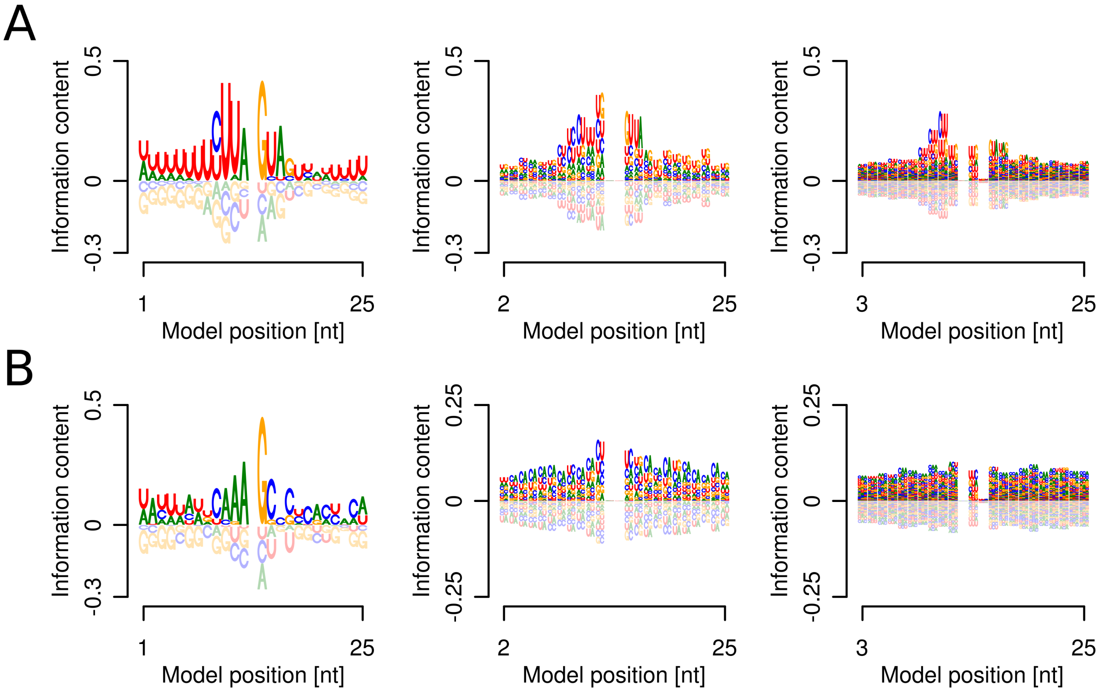


Figure A.10.: **Higher-order sequence logos of protein-RNA crosslink sites. (A)** 0th-order (left), 1st-order (middle), and 2nd-order (right) sequence logos for 2nd-order iIMM of Nab3. The central crosslinked U was removed from the 0th-order sequence logo. Sequence logos show the model learned from all sequences. **(B)** Same as **A** but depicting the Yra1 model.

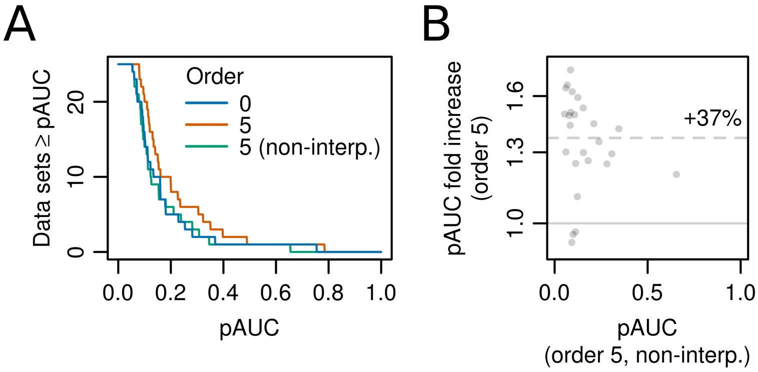


Figure A.11.: **iIMMs of protein-RNA crosslink sites outperform iMMs. (A)** Cumulative distributions of the partial area under the ROC curve (pAUC), up to a false positive rate of 5%, over all 25 PAR-CLIP data sets, comparing the performance of iIMMs to non-interpolating (non-interp.) iMMs of 5th order. **(B)** Scatter plot of the increase in pAUC for single PAR-CLIP data sets, comparing the performance of iIMMs to iMMs of 5th order. The y-axis is shown in log scale. The dashed line indicates the mean fold increase. In the majority of data sets, iIMMs outperform iMMs.

B. Supplementary material (Part II)

B.1. The IUPAC letter nomenclature

Symbol	Meaning	Origin of designation
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
W	A or T	Weak interaction (2 hydrogen bonds)
S	C or G	Strong interaction (3 hydrogen bonds)
R	A or G	puRin
Y	C or T	pYrimidine
K	G or T	Ketone
M	A or C	aMino
B	C, G, or T	not A, B follows A in the alphabet
D	A, G, or T	not C, D follows C
H	A, C, or T	not G, H follows G
V	A, C, or G	not T (not U), V follows U
N	A, C, G, or T	aNy

Table B.1.: The IUPAC letter nomenclature represents an extended alphabet to deal with incompletely specified bases in nucleic acid sequences. The nomenclature permits the allocation of a single-letter symbol in cases where two or more bases are allowed at a particular sequence position (adapted from Cornish-Bowden (1985)).

Bibliography

N. Abe, I. Dror, L. Yang, M. Slattery, T. Zhou, H. J. Bussemaker, R. Rohs, and R. S. Mann. Deconvolving the recognition of DNA shape from sequence. *Cell* **2015**;*161*(2):307–318.

A. Afek, H. Cohen, S. Barber-Zucker, R. Gordân, and D. B. Lukatsky. Nonconsensus protein binding to repetitive DNA sequence elements significantly affects eukaryotic genomes. *PLoS Comput Biol* **2015**;*11*(8):e1004429.

A. Afek, J. L. Schipper, J. Horton, R. Gordân, and D. B. Lukatsky. Protein-DNA binding in the absence of specific base-pair recognition. *Proc Natl Acad Sci USA* **2014**;*111*(48):17140–17145.

S. H. Ahn, M.-C. Keogh, and S. Buratowski. Ctk1 promotes dissociation of basal transcription factors from elongating RNA polymerase II. *EMBO J* **2009**;*28*(3):205–212.

S. H. Ahn, M. Kim, and S. Buratowski. Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and $3'$ end processing. *Mol Cell* **2004**;*13*(1):67–76.

M. S. Akhtar, M. Heidemann, J. R. Tietjen, D. W. Zhang, R. D. Chapman, D. Eick, and A. Z. Ansari. TFIIH kinase places bivalent marks on the carboxy-terminal domain of RNA polymerase II. *Mol Cell* **2009**;*34*(3):387–393.

B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **2015**;*33*(8):831–838.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol* **1990**;*215*(3):403–410.

O. Aparicio, J. V. Geisberg, E. Sekinger, A. Yang, Z. Moqtaderi, and K. Struhl. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Curr Protoc Mol Biol* **2005**;*Chapter 21*:Unit 21.3.

A. Arvey, P. Agius, W. S. Noble, and C. Leslie. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res* **2012**;*22*(9):1723–1734.

S. D. Auweter, F. C. Oberstrass, and F. H.-T. Allain. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res* **2006**;*34*(17):4943–4959.

V. Bacikova, J. Pasulka, K. Kubicek, and R. Stefl. Structure and semi-sequence-specific RNA binding of Nrd1. *Nucleic Acids Res* **2014**;*42*(12):8024–8038.

G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C.-F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk. Diversity and complexity in DNA recognition by transcription factors. *Science* **2009**;*324*(5935):1720–1723.

C. Baejen, P. Torkler, S. Gressel, K. Essig, J. Söding, and P. Cramer. Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition. *Mol Cell* **2014**;*55*(5):745–757.

T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **2009**; *37*(Web Server issue):W202–W208.

T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA, USA, **1994**; pages 28–36.

T. L. Bailey and M. Gribskov. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **1998**;*14*(1):48–54.

Y. Barash, G. Elidan, N. Friedman, and T. Kaplan. Modeling dependencies in protein-DNA binding sites. In Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology, RECOMB '03. ACM Press, New York, NY, USA, **2003**; pages 28–37.

D. Barillà, B. A. Lee, and N. J. Proudfoot. Cleavage/polyadenylation factor IA associates with the carboxyl-terminal domain of RNA polymerase II in Saccharomyces cerevisiae. *Proc Natl Acad Sci USA* **2001**;*98*(2):445–450.

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* **2007**; *129*(4):823–837.

A. Battle, Z. Khan, S. H. Wang, A. Mitrano, M. J. Ford, J. K. Pritchard, and Y. Gilad. Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **2015**;*347*(6222):664–667.

M. A. Batzer and P. L. Deininger. Alu repeats and human genomic diversity. *Nat Rev Genet* **2002**; *3*(5):370–379.

T. Baubec, R. Ivánek, F. Lienert, and D. Schübeler. Methylation-dependent and -independent genomic targeting principles of the MBD protein family. *Cell* **2013**;*153*(2):480–492.

R. Belotserkovskaya, S. Oh, V. A. Bondarenko, G. Orphanides, V. M. Studitsky, and D. Reinberg. FACT facilitates transcription-dependent nucleosome alteration. *Science* **2003**;*301*(5636):1090–1093.

I. Ben-Gal, A. Shani, A. Gohr, J. Grau, S. Arviv, A. Shmilovici, S. Posch, and I. Grosse. Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics* **2005**; *21*(11):2657–2666.

P. V. Benos, M. L. Bulyk, and G. D. Stormo. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res* **2002**;*30*(20):4442–4451.

D. Benveniste, H.-J. Sonntag, G. Sanguinetti, and D. Sproul. Transcription factor binding predicts histone modifications in human cell lines. *Proc Natl Acad Sci USA* **2014**;*111*(37):13367–13372.

O. G. Berg and P. H. von Hippel. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol* **1987**;*193*(4):723–750.

M. Bergbauer, M. Kalla, A. Schmeinck, C. Göbel, U. Rothbauer, S. Eck, A. Benet-Pagès, T. M. Strom, and W. Hammerschmidt. CpG-methylation regulates a class of Epstein-Barr virus promoters. *PLoS Pathog* **2010**;*6*(9):e1001114.

M. F. Berger, A. A. Philippakis, A. M. Qureshi, F. S. He, P. W. Estep, III, and M. L. Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* **2006**;*24*(11):1429–1435.

M. D. Biggin. Animal transcription networks as highly connected, quantitative continua. *Dev Cell* **2011**;*21*(4):611–626.

C. M. Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Secaucus, NJ, USA, **2006**.

S. Boeing, C. Rigault, M. Heidemann, D. Eick, and M. Meisterernst. RNA polymerase II C-terminal heptarepeat domain Ser-7 phosphorylation is established in a mediator-dependent fashion. *J Biol Chem* **2010**;*285*(1):188–196.

M. J. Booth, M. R. Branco, G. Ficz, D. Oxley, F. Krueger, W. Reik, and S. Balasubramanian. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **2012**;*336*(6083):934–937.

M. J. Booth, G. Marsico, M. Bachman, D. Beraldi, and S. Balasubramanian. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat Chem* **2014**;*6*(5):435–440.

J. B. Brown, N. Boley, R. Eisman, G. E. May, M. H. Stoiber, M. O. Duff, B. W. Booth, J. Wen, S. Park, A. M. Suzuki, K. H. Wan, C. Yu, D. Zhang, J. W. Carlson, L. Cherbas, B. D. Eads, D. Miller, K. Mockaitis, J. Roberts, C. A. Davis, E. Frise, A. S. Hammonds, S. Olson, S. Shenker, D. Sturgill, A. A. Samsonova, R. Weiszmann, G. Robinson, J. Hernandez, J. Andrews, P. J. Bickel, P. Carninci, P. Cherbas, T. R. Gingeras, R. A. Hoskins, T. C. Kaufman, E. C. Lai, B. Oliver, N. Perrimon, B. R. Graveley, and S. E. Celniker. Diversity and dynamics of the Drosophila transcriptome. *Nature* **2014**;*512*(7515):393–399.

B. A. Buck-Koehntop, R. L. Stanfield, D. C. Ekiert, M. A. Martinez-Yamout, H. J. Dyson, I. A. Wilson, and P. E. Wright. Molecular basis for recognition of methylated and specific DNA sequences by the zinc finger protein Kaiso. *Proc Natl Acad Sci USA* **2012**;*109*(38):15229–15234.

J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **2013**;*10*(12):1213–1218.

J. D. Buenrostro, B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **2015**;*523*(7561):486–490.

M. L. Bulyk, P. L. F. Johnson, and G. M. Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res* **2002**;*30*(5):1255–1261.

S. Buratowski. Progression through the RNA polymerase II CTD cycle. *Mol Cell* **2009**;*36*(4):541–546.

C. D. Carlson, C. L. Warren, K. E. Hauschild, M. S. Ozers, N. Qadir, D. Bhimsaria, Y. Lee, F. Cerrina, and A. Z. Ansari. Specificity landscapes of DNA binding molecules elucidate biological function. *Proc Natl Acad Sci USA* **2010**;*107*(10):4544–4549.

G. Celeux and J. Diebolt. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly* **1985**;*2*(1):73–82.

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* **2007**;*318*(5857):1780–1782.

H.-T. Chen, L. Warfield, and S. Hahn. The positions of TFIIF and TFIIE in the RNA polymerase II transcription preinitiation complex. *Nat Struct Mol Biol* **2007**;*14*(8):696–703.

L. Chen, K. Chen, L. A. Lavery, S. A. Baker, C. A. Shaw, W. Li, and H. Y. Zoghbi. MeCP2 binds to non-CG methylated DNA as neurons mature, influencing transcription and the timing of onset for Rett syndrome. *Proc Natl Acad Sci USA* **2015**;*112*(17):5509–5514.

X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y.-H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W.-K. Sung, N. D. Clarke, C.-L. Wei, and H.-H. Ng. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **2008**;*133*(6):1106–1117.

J. M. Cherry, E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E. T. Chan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, B. C. Hitz, K. Karra, C. J. Krieger, S. R. Miyasato, R. S. Nash, J. Park, M. S. Skrzypek, M. Simison, S. Weng, and E. D. Wong. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res* **2012**;*40*(Database issue):D700–D705.

L. A. Cirillo, F. R. Lin, I. Cuesta, D. Friedman, M. Jarnik, and K. S. Zaret. Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol Cell* **2002**;*9*(2):279–289.

K. L. Clark, E. D. Halay, E. Lai, and S. K. Burley. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **1993**;*364*(6436):412–420.

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

A. Cornish-Bowden. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res* **1985**;*13*(9):3021–3030.

T. J. Creamer, M. M. Darby, N. Jamonnak, P. Schaughency, H. Hao, S. J. Wheelan, and J. L. Corden. Transcriptome-wide binding sites for components of the Saccharomyces cerevisiae non-poly(A) termination pathway: Nrd1, Nab3, and Sen1. *PLoS Genet* **2011**;*7*(10):e1002329.

J. Crocker, N. Abe, L. Rinaldi, A. P. McGregor, N. Frankel, S. Wang, A. Alsawadi, P. Valenti, S. Plaza, F. Payre, R. S. Mann, and D. L. Stern. Low affinity binding site clusters confer Hox specificity and regulatory robustness. *Cell* **2015**;*160*(1-2):191–203.

A. C. Dantas Machado, T. Zhou, S. Rao, P. Goel, C. Rastogi, A. Lazarovici, H. J. Bussemaker, and R. Rohs. Evolving insights on how cytosine methylation affects protein-DNA binding. *Brief Funct Genomics* **2015**;*14*(1):61–73.

L. David, W. Huber, M. Granovskaia, J. Toedling, C. J. Palm, L. Bofkin, T. Jones, R. W. Davis, and L. M. Steinmetz. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci USA* **2006**;*103*(14):5320–5325.

A. P. J. de Koning, W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* **2011**;*7*(12):e1002384.

P. L. Deininger and M. A. Batzer. Mammalian retroelements. *Genome Res* **2002**;*12*(10):1455–1465.

A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **1999**;*27*(23):4636–4641.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **1977**;*39*(1):1–38.

W. Deng and S. G. E. Roberts. A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes Dev* **2005**;*19*(20):2418–2423.

S. Dengl, A. Mayer, M. Sun, and P. Cramer. Structure and in vivo requirement of the yeast Spt6 SH2 domain. *J Mol Biol* **2009**;*389*(1):211–225.

S. Diekmann. Temperature and salt dependence of the gel migration anomaly of curved DNA fragments. *Nucleic Acids Res* **1987**;*15*(1):247–265.

D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, K. Cesarkas, J. Jacob-Hirsch, N. Amariglio, M. Kupiec, R. Sorek, and G. Rechavi. Topology of the human and mouse m^{6}A RNA methylomes revealed by m^{6}A-seq. *Nature* **2012**;*485*(7397):201–206.

X. Dong, M. C. Greven, A. Kundaje, S. Djebali, J. B. Brown, C. Cheng, T. R. Gingeras, M. Gerstein, R. Guigó, E. Birney, and Z. Weng. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol* **2012**;*13*(9):R53.

G. dos Santos, A. J. Schroeder, J. L. Goodman, V. B. Strelets, M. A. Crosby, J. Thurmond, D. B. Emmert, W. M. Gelbart, and the FlyBase Consortium. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res* **2015**;*43*(Database issue):D690–D697.

I. Dror, T. Golan, C. Levy, R. Rohs, and Y. Mandel-Gutfreund. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res* **2015**; *25*(9):1268–1280.

I. Dror, T. Zhou, Y. Mandel-Gutfreund, and R. Rohs. Covariation between homeodomain transcription factors and the shape of their DNA binding sites. *Nucleic Acids Res* **2014**;*42*(1):430–441.

E. Durán, S. Djebali, S. González, O. Flores, J. M. Mercader, R. Guigó, D. Torrents, M. Soler-López, and M. Orozco. Unravelling the hidden DNA structural/physical code provides novel insights on promoter location. *Nucleic Acids Res* **2013**;*41*(15):7220–7230.

S. H. C. Duttke, S. A. Lacadie, M. M. Ibrahim, C. K. Glass, D. L. Corcoran, C. Benner, S. Heinz, J. T. Kadonaga, and U. Ohler. Human promoters are intrinsically directional. *Mol Cell* **2015**; *57*(4):674–684.

E. Eden and S. Brunak. Analysis and recognition of 5′ UTR intron splice sites in human pre-mRNA. *Nucleic Acids Res* **2004**;*32*(3):1131–1142.

J. M. Edwards, J. Long, C. H. de Moor, J. Emsley, and M. S. Searle. Structural insights into the targeting of mRNA GU-rich elements by the three RRMs of CELF1. *Nucleic Acids Res* **2013**; *41*(14):7153–7166.

R. Eggeling, A. Gohr, J. Keilwagen, M. Mohr, S. Posch, A. D. Smith, and I. Grosse. On the value of intra-motif dependencies of human insulator protein CTCF. *PLoS One* **2014**;*9*(1):e85629.

K. Ellrott, C. Yang, F. M. Sladek, and T. Jiang. Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics* **2002**;*18 Suppl 2*:S100–S109.

X. Fan, N. Lamarre-Vincent, Q. Wang, and K. Struhl. Extensive chromatin fragmentation improves enrichment of protein binding sites in chromatin immunoprecipitation experiments. *Nucleic Acids Res* **2008**;*36*(19):e125.

G. J. Faulkner, Y. Kimura, C. O. Daub, S. Wani, C. Plessy, K. M. Irvine, K. Schroder, N. Cloonan, A. L. Steptoe, T. Lassmann, K. Waki, N. Hornig, T. Arakawa, H. Takahashi, J. Kawai, A. R. R. Forrest, H. Suzuki, Y. Hayashizaki, D. A. Hume, V. Orlando, S. M. Grimmond, and P. Carninci. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **2009**;*41*(5):563–571.

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

P. M. Fordyce, D. Pincus, P. Kimmig, C. S. Nelson, H. El-Samad, P. Walter, and J. L. DeRisi. Basic leucine zipper transcription factor Hac1 binds DNA in two distinct modes as revealed by microfluidic analyses. *Proc Natl Acad Sci USA* **2012**;*109*(45):E3084–E3093.

M. C. Frith, E. Valen, A. Krogh, Y. Hayashizaki, P. Carninci, and A. Sandelin. A code for transcription initiation in mammalian genomes. *Genome Res* **2008**;*18*(1):1–12.

Y. Fu, G.-Z. Luo, K. Chen, X. Deng, M. Yu, D. Han, Z. Hao, J. Liu, X. Lu, L. C. Doré, X. Weng, Q. Ji, L. Mets, and C. He. N^6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. *Cell* **2015**;*161*(4):879–892.

J.-M. Fustin, M. Doi, Y. Yamaguchi, H. Hida, S. Nishimura, M. Yoshida, T. Isagawa, M. S. Morioka, H. Kakeya, I. Manabe, and H. Okamura. RNA-methylation-dependent RNA processing controls the speed of the circadian clock. *Cell* **2013**;*155*(4):793–806.

E. R. Gansner and S. C. North. An open graph visualization system and its applications to software engineering. *Software Practice & Experience* **2000**;*30*(11):1203–1233.

M. Geertz, D. Shore, and S. J. Maerkl. Massively parallel measurements of molecular interaction kinetics on a microfluidic platform. *Proc Natl Acad Sci USA* **2012**;*109*(41):16540–16545.

R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **2004**;*5*(10):R80.

B. Georgi and A. Schliep. Context-specific independence mixture modeling for positional weight matrices. *Bioinformatics* **2006**;*22*(14):e166–e173.

J. Gertz, D. Savic, K. E. Varley, E. C. Partridge, A. Safi, P. Jain, G. M. Cooper, T. E. Reddy, G. E. Crawford, and R. M. Myers. Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol Cell* **2013**;*52*(1):25–36.

J. Gertz, K. E. Varley, T. E. Reddy, K. M. Bowling, F. Pauli, S. L. Parker, K. S. Kucera, H. F. Willard, and R. M. Myers. Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. *PLoS Genet* **2011**;*7*(8):e1002228.

D. S. Gilmour and J. T. Lis. Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proc Natl Acad Sci USA* **1984**;*81*(14):4275–4279.

K. Glover-Cutter, S. Kim, J. Espinosa, and D. L. Bentley. RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. *Nat Struct Mol Biol* **2008**;*15*(1):71–78.

R. Gordân, K. F. Murphy, R. P. McCord, C. Zhu, A. Vedenko, and M. L. Bulyk. Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. *Genome Biol* **2011**;*12*(12):R125.

R. Gordân, N. Shen, I. Dror, T. Zhou, J. Horton, R. Rohs, and M. L. Bulyk. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep* **2013**;*3*(4):1093–1104.

C. E. Grant, T. L. Bailey, and W. S. Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **2011**;*27*(7):1017–1018.

S. Gunewardena and Z. Zhang. A hybrid model for robust detection of transcription factor binding sites. *Bioinformatics* **2008**;*24*(4):484–491.

M. Guo, F. Xu, J. Yamada, T. Egelhofer, Y. Gao, G. A. Hartzog, M. Teng, and L. Niu. Core structure of the yeast Spt4-Spt5 complex: a conserved module for regulation of transcription elongation. *Structure* **2008**;*16*(11):1649–1658.

M. Gustems, A. Woellmer, U. Rothbauer, S. H. Eck, T. Wieland, D. Lutter, and W. Hammerschmidt. c-Jun/c-Fos heterodimers regulate cellular genes via a newly identified class of methylated DNA sequence motifs. *Nucleic Acids Res* **2014**;*42*(5):3059–3072.

A. Hackmann, H. Wu, U.-M. Schneider, K. Meyer, K. Jung, and H. Krebber. Quality control of spliced mRNAs requires the shuttling SR proteins Gbp2 and Hrb1. *Nat Commun* **2014**;*5*:3123.

M. A. Hahn, R. Qiu, X. Wu, A. X. Li, H. Zhang, J. Wang, J. Jui, S.-G. Jin, Y. Jiang, G. P. Pfeifer, and Q. Lu. Dynamics of 5-hydroxymethylcytosine and chromatin marks in mammalian neurogenesis. *Cell Rep* **2013**;*3*(2):291–300.

P. Hammar, P. Leroy, A. Mahmutovic, E. G. Marklund, O. G. Berg, and J. Elf. The lac repressor displays facilitated diffusion in living cells. *Science* **2012**;*336*(6088):1595–1598.

S. Hannenhalli and L.-S. Wang. Enhanced position weight matrices using mixture models. *Bioinformatics* **2005**;*21 Suppl 1*:i204–i212.

C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature* **2004**;*431*(7004):99–104.

L.-A. Harris, L. D. Williams, and G. B. Koudelka. Specific minor groove solvation is a crucial determinant of DNA binding site recognition. *Nucleic Acids Res* **2014**;*42*(22):14053–14059.

H. Hartmann, E. W. Guthöhrlein, M. Siebert, S. Luehr, and J. Söding. P-value-based regulatory motif discovery using positional weight matrices. *Genome Res* **2013**;*23*(1):181–194.

H. Hashimoto, Y. O. Olanrewaju, Y. Zheng, G. G. Wilson, X. Zhang, and X. Cheng. Wilms tumor protein recognizes 5-carboxylcytosine within a specific DNA sequence. *Genes Dev* **2014**; *28*(20):2304–2313.

H. H. He, C. A. Meyer, S. S. Hu, M.-W. Chen, C. Zang, Y. Liu, P. K. Rao, T. Fei, H. Xu, H. Long, X. S. Liu, and M. Brown. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat Methods* **2014**;*11*(1):73–78.

Q. He, J. Johnston, and J. Zeitlinger. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat Biotechnol* **2015**;*33*(4):395–401.

J. R. Hesselberth, X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, W. S. Noble, S. Fields, and J. A. Stamatoyannopoulos. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* **2009**;*6*(4):283–289.

Y. Hirose and J. L. Manley. RNA polymerase II and the integration of nuclear events. *Genes Dev* **2000**;*14*(12):1415–1429.

A. Hirtreiter, G. E. Damsma, A. C. M. Cheung, D. Klose, D. Grohmann, E. Vojnic, A. C. R. Martin, P. Cramer, and F. Werner. Spt4/5 stimulates transcription elongation through the RNA polymerase clamp coiled-coil motif. *Nucleic Acids Res* **2010**;*38*(12):4040–4051.

M. Hu, J. Yu, J. M. G. Taylor, A. M. Chinnaiyan, and Z. S. Qin. On the detection and refinement of transcription factor binding sites using ChIP-seq data. *Nucleic Acids Res* **2010**;*38*(7):2154–2167.

S. Hu, J. Wan, Y. Su, Q. Song, Y. Zeng, H. N. Nguyen, J. Shin, E. Cox, H. S. Rho, C. Woodard, S. Xia, S. Liu, H. Lyu, G.-L. Ming, H. Wade, H. Song, J. Qian, and H. Zhu. DNA methylation presents distinct binding sites for human transcription factors. *eLife* **2013**;*2*:e00726.

W. Huang, D. M. Umbach, U. Ohler, and L. Li. Optimized mixed Markov models for motif identification. *BMC Bioinformatics* **2006**;*7*:279.

S. Hussain, J. Aleksic, S. Blanco, S. Dietmann, and M. Frye. Characterizing 5-methylcytosine in the mammalian epitranscriptome. *Genome Biol* **2013**;*14*(11):215.

M. Imashimizu, H. Takahashi, T. Oshima, C. McIntosh, M. Bubunenko, D. L. Court, and M. Kashlev. Visualizing translocation dynamics and nascent transcript errors in paused RNA polymerases in vivo. *Genome Biol* **2015**;*16*:98.

M. Iurlaro, G. Ficz, D. Oxley, E.-A. Raiber, M. Bachman, M. J. Booth, S. Andrews, S. Balasubramanian, and W. Reik. A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol* **2013**;*14*(10):R119.

F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **1961**;*3*:318–356.

A. J. Jasiak, H. Hartmann, E. Karakasili, M. Kalocsay, A. Flatley, E. Kremmer, K. Strässer, D. E. Martin, J. Söding, and P. Cramer. Genome-associated RNA polymerase II includes the dissociable Rpb4/7 subcomplex. *J Biol Chem* **2008**;*283*(39):26423–26427.

C. Jiang and B. F. Pugh. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* **2009**;*10*(3):161–172.

D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **2007**;*316*(5830):1497–1502.

W. E. Johnson, W. Li, C. A. Meyer, R. Gottardo, J. S. Carroll, M. Brown, and X. S. Liu. Model-based analysis of tiling-arrays for ChIP-chip. *Proc Natl Acad Sci USA* **2006**;*103*(33):12457–12462.

A. Jolma, T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J. M. Vaquerizas, J. Yan, M. J. Sillanpää, M. Bonke, K. Palin, S. Talukder, T. R. Hughes, N. M. Luscombe, E. Ukkonen, and J. Taipale. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res* **2010**;*20*(6):861–873.

A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N. M. Luscombe, T. R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale. DNA-binding specificities of human transcription factors. *Cell* **2013**;*152*(1-2):327–339.

A. Jolma, Y. Yin, K. R. Nitta, K. Dave, A. Popov, M. Taipale, M. Enge, T. Kivioja, E. Morgunova, and J. Taipale. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **2015**;*527*(7578):384–388.

M. Jovanovic, M. S. Rooney, P. Mertins, D. Przybylski, N. Chevrier, R. Satija, E. H. Rodriguez, A. P. Fields, S. Schwartz, R. Raychowdhury, M. R. Mumbach, T. Eisenhaure, M. Rabani, D. Gennert, D. Lu, T. Delorey, J. S. Weissman, S. A. Carr, N. Hacohen, and A. Regev. Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science* **2015**;*347*(6226):1259038.

J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **2005**;*110*(1-4):462–467.

J. T. Kadonaga. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip Rev Dev Biol* **2012**;*1*(1):40–51.

I. Kamenova, L. Warfield, and S. Hahn. Mutations on the DNA binding surface of TBP discriminate between yeast TATA and TATA-less gene transcription. *Mol Cell Biol* **2014**;*34*(15):2929–2943.

R. N. Kanadia, K. A. Johnstone, A. Mankodi, C. Lungu, C. A. Thornton, D. Esson, A. M. Timmers, W. W. Hauswirth, and M. S. Swanson. A muscleblind knockout model for myotonic dystrophy. *Science* **2003**;*302*(5652):1978–1980.

C. D. Kaplan, M. J. Holland, and F. Winston. Interaction between transcription elongation factors and mRNA 3′-end formation at the Saccharomyces cerevisiae GAL10-GAL7 locus. *J Biol Chem* **2005**;*280*(2):913–922.

C. D. Kaplan, L. Laprade, and F. Winston. Transcription elongation factors repress transcription initiation from cryptic sites. *Science* **2003**;*301*(5636):1096–1099.

R. Karlić, H.-R. Chung, J. Lasserre, K. Vlahovicek, and M. Vingron. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci USA* **2010**;*107*(7):2926–2931.

S. Kasinathan, G. A. Orsi, G. E. Zentner, K. Ahmad, and S. Henikoff. High-resolution mapping of transcription factor binding sites on native chromatin. *Nat Methods* **2014**;*11*(2):203–209.

R. Katainen, K. Dave, E. Pitkänen, K. Palin, T. Kivioja, N. Välimäki, A. E. Gylfe, H. Ristolainen, U. A. Hänninen, T. Cajuso, J. Kondelin, T. Tanskanen, J.-P. Mecklin, H. Järvinen, L. Renkonen-Sinisalo, A. Lepistö, E. Kaasinen, O. Kilpivaara, S. Tuupanen, M. Enge, J. Taipale, and L. A. Aaltonen. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet* **2015**;*47*(7):818–821.

M. Kazemian, Q. Zhu, M. S. Halfon, and S. Sinha. Improved accuracy of supervised CRM discovery with interpolated Markov models and cross-species comparison. *Nucleic Acids Res* **2011**; *39*(22):9463–9472.

J. Keilwagen and J. Grau. Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Res* **2015**;*43*(18):e119.

M.-C. Keogh, V. Podolny, and S. Buratowski. Bur1 kinase is required for efficient transcription elongation by RNA polymerase II. *Mol Cell Biol* **2003**;*23*(19):7005–7018.

P. V. Kharchenko, M. Y. Tolstorukov, and P. J. Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **2008**;*26*(12):1351–1359.

A. Kiliszek and W. Rypniewski. Structural studies of CNG repeats. *Nucleic Acids Res* **2014**; *42*(13):8189–8199.

A. M. Kilpatrick, B. Ward, and S. Aitken. Stochastic EM-based TFBS motif discovery with MITSU. *Bioinformatics* **2014**;*30*(12):i310–i318.

H. Kilpinen, S. M. Waszak, A. R. Gschwind, S. K. Raghav, R. M. Witwicki, A. Orioli, E. Migliavacca, M. Wiederkehr, M. Gutierrez-Arcelus, N. I. Panousis, A. Yurovsky, T. Lappalainen, L. Romano-Palumbo, A. Planchon, D. Bielser, J. Bryois, I. Padioleau, G. Udin, S. Thurnheer, D. Hacker, L. J. Core, J. T. Lis, N. Hernandez, A. Reymond, B. Deplancke, and E. T. Dermitzakis. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **2013**;*342*(6159):744–747.

J. Kim and K. Struhl. Determinants of half-site spacing preferences that distinguish AP-1 and ATF/CREB bZIP domains. *Nucleic Acids Res* **1995**;*23*(13):2531–2537.

M. Kim, S.-H. Ahn, N. J. Krogan, J. F. Greenblatt, and S. Buratowski. Transitions in RNA polymerase II elongation complexes at the 3′ ends of genes. *EMBO J* **2004**;*23*(2):354–364.

M. Kim, H. Suh, E.-J. Cho, and S. Buratowski. Phosphorylation of the yeast Rpb1 C-terminal domain at serines 2, 5, and 7. *J Biol Chem* **2009**;*284*(39):26421–26426.

S. Kim, E. Broströmer, D. Xing, J. Jin, S. Chong, H. Ge, S. Wang, C. Gu, L. Yang, Y. Q. Gao, X. Su, Y. Sun, and X. S. Xie. Probing allostery through DNA. *Science* **2013**;*339*(6121):816–819.

M. Kitayner, H. Rozenberg, R. Rohs, O. Suad, D. Rabinovich, B. Honig, and Z. Shakked. Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat Struct Mol Biol* **2010**;*17*(4):423–429.

J. D. Klemm, M. A. Rould, R. Aurora, W. Herr, and C. O. Pabo. Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. *Cell* **1994**;*77*(1):21–32.

R. M. Kohli and Y. Zhang. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **2013**;*502*(7472):472–479.

P. Komarnitsky, E. J. Cho, and S. Buratowski. Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes Dev* **2000**;*14*(19):2452–2460.

P. Konieczny, E. Stepniak-Konieczna, and K. Sobczak. MBNL proteins and their target RNAs, interaction and splicing regulation. *Nucleic Acids Res* **2014**;*42*(17):10873–10887.

H. S. Koo, H. M. Wu, and D. M. Crothers. DNA bending at adenine-thymine tracts. *Nature* **1986**; *320*(6062):501–506.

M. J. E. Koster, B. Snel, and H. T. M. Timmers. Genesis of chromatin and transcription dynamics in the origin of species. *Cell* **2015**;*161*(4):724–736.

D. Kostrewa, M. E. Zeller, K.-J. Armache, M. Seizl, K. Leike, M. Thomm, and P. Cramer. RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature* **2009**;*462*(7271):323–330.

N. J. Krogan, M. Kim, S. H. Ahn, G. Zhong, M. S. Kobor, G. Cagney, A. Emili, A. Shilatifard, S. Buratowski, and J. F. Greenblatt. RNA polymerase II elongation factors of Saccharomyces cerevisiae: a targeted proteomics approach. *Mol Cell Biol* **2002**;*22*(20):6979–6992.

J. N. Kuehner and D. A. Brow. Quantitative analysis of in vivo initiator selection by yeast RNA polymerase II supports a scanning model. *J Biol Chem* **2006**;*281*(20):14119–14128.

I. Kulakovskiy, V. Levitsky, D. Oshchepkov, L. Bryzgalov, I. Vorontsov, and V. Makeev. From binding motifs in ChIP-seq data to improved models of transcription factor binding sites. *J Bioinform Comput Biol* **2013**;*11*(1):1340004.

H. Kurokawa, H. Motohashi, S. Sueno, M. Kimura, H. Takagawa, Y. Kanno, M. Yamamoto, and T. Tanaka. Structural basis of alternative DNA recognition by Maf transcription factors. *Mol Cell Biol* **2009**;*29*(23):6232–6244.

J. C. Kwasnieski, C. Fiore, H. G. Chaudhari, and B. A. Cohen. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res* **2014**;*24*(10):1595–1602.

T. Lagrange, A. N. Kapanidis, H. Tang, D. Reinberg, and R. H. Ebright. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev* **1998**;*12*(1):34–44.

S. G. Landt, G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein, K. I. Fisher-Aylor, G. Euskirchen, M. Gerstein, J. Gertz, A. J. Hartemink, M. M. Hoffman, V. R. Iyer, Y. L. Jung, S. Karmakar, M. Kellis, P. V. Kharchenko, Q. Li, T. Liu, X. S. Liu, L. Ma, A. Milosavljevic, R. M. Myers, P. J. Park, M. J. Pazin, M. D. Perry, D. Raha, T. E. Reddy, J. Rozowsky, N. Shoresh, A. Sidow, M. Slattery, J. A. Stamatoyannopoulos, M. Y. Tolstorukov, K. P. White, S. Xi, P. J. Farnham, J. D. Lieb, B. J. Wold, and M. Snyder. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **2012**;*22*(9):1813–1831.

R. N. Laribee, N. J. Krogan, T. Xiao, Y. Shibata, T. R. Hughes, J. F. Greenblatt, and B. D. Strahl. Bur kinase selectively regulates H3 K4 trimethylation and H2B ubiquitylation through recruitment of the PAF elongation complex. *Curr Biol* **2005**;*15*(16):1487–1493.

M. H. Larson, R. A. Mooney, J. M. Peters, T. Windgassen, D. Nayak, C. A. Gross, S. M. Block, W. J. Greenleaf, R. Landick, and J. S. Weissman. A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science* **2014**;*344*(6187):1042–1047.

C. E. Lawrence and A. A. Reilly. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* **1990**;*7*(1):41–51.

A. Lazarovici, T. Zhou, A. Shafer, A. C. D. Machado, T. R. Riley, R. Sandstrom, P. J. Sabo, Y. Lu, R. Rohs, J. A. Stamatoyannopoulos, and H. J. Bussemaker. Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc Natl Acad Sci USA* **2013**;*110*(16):6376–6381.

C. S. Lee, J. R. Friedman, J. T. Fulmer, and K. H. Kaestner. The initiation of liver development is dependent on FoxA transcription factors. *Nature* **2005**;*435*(7044):944–947.

B. Lenhard, A. Sandelin, and P. Carninci. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* **2012**;*13*(4):233–245.

D. A. Leonard, N. Rajaram, and T. K. Kerppola. Structural basis of DNA bending and oriented heterodimer binding by the basic leucine zipper domains of Fos and Jun. *Proc Natl Acad Sci USA* **1997**;*94*(10):4913–4918.

V. G. Levitsky, I. V. Kulakovskiy, N. I. Ershov, D. Y. Oshchepkov, V. J. Makeev, T. C. Hodgman, and T. I. Merkulova. Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-seq data. *BMC Genomics* **2014**;*15*:80.

M. Levo and E. Segal. In pursuit of design principles of regulatory sequences. *Nat Rev Genet* **2014**; *15*(7):453–468.

M. Levo, E. Zalckvar, E. Sharon, A. C. Dantas Machado, Y. Kalma, M. Lotan-Pompan, A. Weinberger, Z. Yakhini, R. Rohs, and E. Segal. Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res* **2015**;*25*(7):1018–1029.

J. J. Li, P. J. Bickel, and M. D. Biggin. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2014**a;*2*:e270.

J. J. Li and M. D. Biggin. Gene expression. Statistics requantitates the central dogma. *Science* **2015**;*347*(6226):1066–1067.

Q. Li, J. B. Brown, H. Huang, and P. J. Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics* **2011**;*5*(3):1752–1779.

X. Li, H. Kazan, H. D. Lipshitz, and Q. D. Morris. Finding the target sites of RNA-binding proteins. *Wiley Interdiscip Rev RNA* **2014**b;*5*(1):111–130.

C. R. Lickwar, F. Mueller, S. E. Hanlon, J. G. McNally, and J. D. Lieb. Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature* **2012**; *484*(7393):251–255.

M. Lidschreiber, K. Leike, and P. Cramer. Cap completion and C-terminal repeat domain kinase recruitment underlie the initiation-elongation transition of RNA polymerase II. *Mol Cell Biol* **2013**;*33*(19):3805–3816.

B. Linder, A. V. Grozhik, A. O. Olarerin-George, C. Meydan, C. E. Mason, and S. R. Jaffrey. Single-nucleotide-resolution mapping of m^{6}A and m^{6}Am throughout the transcriptome. *Nat Methods* **2015**;*12*(8):767–772.

D. L. Lindstrom, S. L. Squazzo, N. Muster, T. A. Burckin, K. C. Wachter, C. A. Emigh, J. A. McCleery, J. R. Yates, and G. A. Hartzog. Dual roles for Spt5 in pre-mRNA processing and transcription elongation revealed by identification of Spt5-associated proteins. *Mol Cell Biol* **2003**;*23*(4):1368–1378.

R. Lister, E. A. Mukamel, J. R. Nery, M. Urich, C. A. Puddifoot, N. D. Johnson, J. Lucero, Y. Huang, A. J. Dwork, M. D. Schultz, M. Yu, J. Tonti-Filippini, H. Heyn, S. Hu, J. C. Wu, A. Rao, M. Esteller, C. He, F. G. Haghighi, T. J. Sejnowski, M. M. Behrens, and J. R. Ecker. Global epigenomic reconfiguration during mammalian brain development. *Science* **2013**;*341*(6146):1237905.

N. Liu, Q. Dai, G. Zheng, C. He, M. Parisien, and T. Pan. N^6-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature* **2015**;*518*(7540):560–564.

Y. Liu, Y. O. Olanrewaju, Y. Zheng, H. Hashimoto, R. M. Blumenthal, X. Zhang, and X. Cheng. Structural basis for Klf4 recognition of methylated DNA. *Nucleic Acids Res* **2014**;*42*(8):4859–4867.

Y. Liu, H. Toh, H. Sasaki, X. Zhang, and X. Cheng. An atomic model of Zfp57 recognition of CpG methylation within a specific DNA sequence. *Genes Dev* **2012**;*26*(21):2374–2379.

Y. Liu, L. Warfield, C. Zhang, J. Luo, J. Allen, W. H. Lang, J. Ranish, K. M. Shokat, and S. Hahn. Phosphorylation of the transcription elongation factor Spt5 by yeast Bur1 kinase stimulates recruitment of the PAF complex. *Mol Cell Biol* **2009**;*29*(17):4852–4863.

Y. Liu, X. Zhang, R. M. Blumenthal, and X. Cheng. A common mode of recognition for methylated CpG. *Trends Biochem Sci* **2013**;*38*(4):177–183.

S. Lubliner, I. Regev, M. Lotan-Pompan, S. Edelheit, A. Weinberger, and E. Segal. Core promoter sequence in yeast is a major determinant of expression level. *Genome Res* **2015**;*25*(7):1008–1017.

N. M. Luscombe, R. A. Laskowski, and J. M. Thornton. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res* **2001**; *29*(13):2860–2874.

P.-L. Luu, H. R. Schöler, and M. J. Araúzo-Bravo. Disclosing the crosstalk among DNA methylation, transcription factors, and histone marks in human pluripotent cells through discovery of DNA methylation motifs. *Genome Res* **2013**;*23*(12):2013–2029.

J. Maaskola and N. Rajewsky. Binding site discovery from nucleic acid sequences by discriminative learning of hidden Markov models. *Nucleic Acids Res* **2014**;*42*(21):12995–13011.

M. A. Machnicka, K. Milanowska, O. Osman Oglou, E. Purta, M. Kurkowska, A. Olchowik, W. Januszewski, S. Kalinowski, S. Dunin-Horkawicz, K. M. Rother, M. Helm, J. M. Bujnicki, and H. Grosjean. MODOMICS: a database of RNA modification pathways–2013 update. *Nucleic Acids Res* **2013**;*41*(Database issue):D262–D267.

S. J. Maerkl and S. R. Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **2007**;*315*(5809):233–237.

T. K. Man and G. D. Stormo. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res* **2001**;*29*(12):2471–2478.

I. K. Mann, R. Chatterjee, J. Zhao, X. He, M. T. Weirauch, T. R. Hughes, and C. Vinson. CG methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4 heterodimer that is active in vivo. *Genome Res* **2013**;*23*(6):988–997.

C. Maris, C. Dominguez, and F. H.-T. Allain. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J* **2005**;*272*(9):2118–2131.

A. Mathelier and W. W. Wasserman. The next generation of transcription factor binding site prediction. *PLoS Comput Biol* **2013**;*9*(9):e1003214.

A. Mayer, J. di Iulio, S. Maleri, U. Eser, J. Vierstra, A. Reynolds, R. Sandstrom, J. A. Stamatoyannopoulos, and L. S. Churchman. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* **2015**;*161*(3):541–554.

A. Mayer, M. Heidemann, M. Lidschreiber, A. Schreieck, M. Sun, C. Hintermair, E. Kremmer, D. Eick, and P. Cramer. CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. *Science* **2012**;*336*(6089):1723–1725.

A. Mayer, M. Lidschreiber, M. Siebert, K. Leike, J. Söding, and P. Cramer. Uniform transitions of the general RNA polymerase II transcription complex. *Nat Struct Mol Biol* **2010**;*17*(10):1272–1278.

S. H. Meijsing, M. A. Pufall, A. Y. So, D. L. Bates, L. Chen, and K. R. Yamamoto. DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science* **2009**;*324*(5925):407–410.

D. M. Meinel, C. Burkert-Kautzsch, A. Kieser, E. O'Duibhir, M. Siebert, A. Mayer, P. Cramer, J. Söding, F. C. P. Holstege, and K. Sträßer. Recruitment of TREX to the transcription machinery by its direct binding to the phospho-CTD of RNA polymerase II. *PLoS Genet* **2013**; *9*(11):e1003914.

A. Meinhart, T. Kamenski, S. Hoeppner, S. Baumli, and P. Cramer. A structural perspective of CTD function. *Genes Dev* **2005**;*19*(12):1401–1415.

M. Mellén, P. Ayata, S. Dewell, S. Kriaucionis, and N. Heintz. MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* **2012**;*151*(7):1417–1430.

J. W. Miller, C. R. Urbinati, P. Teng-Umnuay, M. G. Stenberg, B. J. Byrne, C. A. Thornton, and M. S. Swanson. Recruitment of human muscleblind proteins to $(\mathrm{CUG})_\mathrm{n}$ expansions associated with myotonic dystrophy. *EMBO J* **2000**;*19*(17):4439–4448.

Z. Moqtaderi, J. V. Geisberg, Y. Jin, X. Fan, and K. Struhl. Species-specific factors mediate extensive heterogeneity of mRNA 3′ ends in yeasts. *Proc Natl Acad Sci USA* **2013**;*110*(27):11073–11078.

F. Mordelet, J. Horton, A. J. Hartemink, B. E. Engelhardt, and R. Gordân. Stability selection for regression-based models of transcription factor-DNA binding specificity. *Bioinformatics* **2013**; *29*(13):i117–i125.

T. Morisaki, W. G. Müller, N. Golob, D. Mazza, and J. G. McNally. Single-molecule analysis of transcription factor binding at transcription sites in live cells. *Nat Commun* **2014**;*5*:4456.

Q. Morris, M. L. Bulyk, and T. R. Hughes. Jury remains out on simple models of transcription factor specificity. *Nat Biotechnol* **2011**;*29*(6):483–484.

F. Mueller, T. J. Stasevich, D. Mazza, and J. G. McNally. Quantifying transcription factor kinetics: at work or at play? *Crit Rev Biochem Mol Biol* **2013**;*48*(5):492–514.

J. M. Muiño, C. Smaczniak, G. C. Angenent, K. Kaufmann, and A. D. J. van Dijk. Structural determinants of DNA recognition by plant MADS-domain transcription factors. *Nucleic Acids Res* **2014**;*42*(4):2138–2146.

M. W. Murphy, J. K. Lee, S. Rojo, M. D. Gearhart, K. Kurahashi, S. Banerjee, G.-A. Loeuille, A. Bashamboo, K. McElreavey, D. Zarkower, H. Aihara, and V. J. Bardwell. An ancient protein-DNA interaction underlying metazoan sex determination. *Nat Struct Mol Biol* **2015**;*22*(6):442–451.

S. Murray, R. Udupa, S. Yao, G. Hartzog, and G. Prelich. Phosphorylation of the RNA polymerase II carboxy-terminal domain by the Bur1 cyclin-dependent kinase. *Mol Cell Biol* **2001**;*21*(13):4089–4096.

U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **2008**;*320*(5881):1344–1349.

H. S. Najafabadi, S. Mnaimneh, F. W. Schmitges, M. Garton, K. N. Lam, A. Yang, M. Albu, M. T. Weirauch, E. Radovani, P. M. Kim, J. Greenblatt, B. J. Frey, and T. R. Hughes. C_2H_2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol* **2015**;*33*(5):555–562.

H. Nakahashi, K.-R. K. Kwon, W. Resch, L. Vian, M. Dose, D. Stavreva, O. Hakim, N. Pruett, S. Nelson, A. Yamane, J. Qian, W. Dubois, S. Welsh, R. D. Phair, B. F. Pugh, V. Lobanenkov, G. L. Hager, and R. Casellas. A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep* **2013**;*3*(5):1678–1689.

A. Nakao, M. Yoshihama, and N. Kenmochi. RPG: the Ribosomal Protein Gene database. *Nucleic Acids Res* **2004**;*32*(Database issue):D168–D170.

L. Narlikar. MuMoD: a Bayesian approach to detect multiple modes of protein-DNA binding from genome-wide ChIP data. *Nucleic Acids Res* **2013**;*41*(1):21–32.

S. Nechaev, D. C. Fargo, G. dos Santos, L. Liu, Y. Gao, and K. Adelman. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science* **2010**;*327*(5963):335–338.

S. Neph, J. Vierstra, A. B. Stergachis, A. P. Reynolds, E. Haugen, B. Vernot, R. E. Thurman, S. John, R. Sandstrom, A. K. Johnson, M. T. Maurano, R. Humbert, E. Rynes, H. Wang, S. Vong, K. Lee, D. Bates, M. Diegel, V. Roach, D. Dunn, J. Neri, A. Schafer, R. S. Hansen, T. Kutyavin, E. Giste, M. Weaver, T. Canfield, P. Sabo, M. Zhang, G. Balasundaram, R. Byron, M. J. MacCoss, J. M. Akey, M. A. Bender, M. Groudine, R. Kaul, and J. A. Stamatoyannopoulos. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **2012**;*489*(7414):83–90.

T. Ni, D. L. Corcoran, E. A. Rach, S. Song, E. P. Spana, Y. Gao, U. Ohler, and J. Zhu. A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat Methods* **2010**;*7*(7):521–527.

T. Ni, Y. Yang, D. Hafez, W. Yang, K. Kiesewetter, Y. Wakabayashi, U. Ohler, W. Peng, and J. Zhu. Distinct polyadenylation landscapes of diverse human tissues revealed by a modified PA-seq strategy. *BMC Genomics* **2013**;*14*:615.

D. B. Nikolov, H. Chen, E. D. Halay, A. A. Usheva, K. Hisatake, D. K. Lee, R. G. Roeder, and S. K. Burley. Crystal structure of a TFIIB-TBP-TATA-element ternary complex. *Nature* **1995**;*377*(6545):119–128.

K. R. Nitta, A. Jolma, Y. Yin, E. Morgunova, T. Kivioja, J. Akhtar, K. Hens, J. Toivonen, B. Deplancke, E. E. M. Furlong, and J. Taipale. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife* **2015**;*4*:e04837.

T. Nojima, T. Gomes, A. R. F. Grosso, H. Kimura, M. J. Dye, S. Dhir, M. Carmo-Fonseca, and N. J. Proudfoot. Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **2015**;*161*(3):526–540.

R. Nutiu, R. C. Friedman, S. Luo, I. Khrebtukova, D. Silva, R. Li, L. Zhang, G. P. Schroth, and C. B. Burge. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat Biotechnol* **2011**;*29*(7):659–664.

R. A. O'Flanagan, G. Paillard, R. Lavery, and A. M. Sengupta. Non-additivity in protein-DNA binding. *Bioinformatics* **2005**;*21*(10):2254–2263.

H. O'Geen, C. M. Nicolet, K. Blahnik, R. Green, and P. J. Farnham. Comparison of sample preparation methods for ChIP-chip assays. *Biotechniques* **2006**;*41*(5):577–580.

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

U. Ohler, S. Harbeck, H. Niemann, E. Nöth, and M. G. Reese. Interpolated Markov chains for eukaryotic promoter recognition. *Bioinformatics* **1999**;*15*(5):362–369.

U. Ohler, G. Liao, H. Niemann, and G. M. Rubin. Computational analysis of core promoters in the Drosophila genome. *Genome Biol* **2002**;*3*(12):RESEARCH0087.

C.-T. Ong and V. G. Corces. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* **2014**;*15*(4):234–246.

Y. Orenstein and R. Shamir. A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Res* **2014**;*42*(8):e63.

A. Orioli, C. Pascali, A. Pagano, M. Teichmann, and G. Dieci. RNA polymerase III transcription control elements: themes and variations. *Gene* **2012**;*493*(2):185–194.

G. Orphanides and D. Reinberg. RNA polymerase II elongation through chromatin. *Nature* **2000**; *407*(6803):471–475.

G. Orphanides and D. Reinberg. A unified theory of gene expression. *Cell* **2002**;*108*(4):439–451.

D. Panne, T. Maniatis, and S. C. Harrison. An atomic model of the interferon-β enhanceosome. *Cell* **2007**;*129*(6):1111–1123.

D. Park, Y. Lee, G. Bhupindersingh, and V. R. Iyer. Widespread misinterpretable ChIP-seq bias in yeast. *PLoS One* **2013**;*8*(12):e83506.

S. C. J. Parker, L. Hansen, H. O. Abaan, T. D. Tullius, and E. H. Margulies. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* **2009**;*324*(5925):389–392.

M. Pasi, J. H. Maddocks, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham, III, P. D. Dans, B. Jayaram, F. Lankas, C. Laughton, J. Mitchell, R. Osman, M. Orozco, A. Pérez, D. Petkevičiūtė, N. Spackova, J. Sponer, K. Zakrzewska, and R. Lavery. μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res* **2014**; *42*(19):12272–12283.

M. Pasi, J. H. Maddocks, and R. Lavery. Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res* **2015**;*43*(4):2412–2423.

W. A. Pastor, L. Aravind, and A. Rao. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nat Rev Mol Cell Biol* **2013**;*14*(6):341–356.

V. Pelechano, W. Wei, and L. M. Steinmetz. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **2013**;*497*(7447):127–131.

R. Perales and D. Bentley. "Cotranscriptionality": the transcription elongation complex as a nexus for nuclear transactions. *Mol Cell* **2009**;*36*(2):178–191.

A. Pérez, C. L. Castellazzi, F. Battistini, K. Collinet, O. Flores, O. Deniz, M. L. Ruiz, D. Torrents, R. Eritja, M. Soler-López, and M. Orozco. Impact of methylation on the physical properties of DNA. *Biophys J* **2012**;*102*(9):2140–2148.

T.-H. Pham, J. Minderjahn, C. Schmidl, H. Hoffmeister, S. Schmidhofer, W. Chen, G. Längst, C. Benner, and M. Rehli. Mechanisms of in vivo binding site selection of the hematopoietic master transcription factor PU.1. *Nucleic Acids Res* **2013**;*41*(13):6391–6402.

D. K. Pokholok, N. M. Hannett, and R. A. Young. Exchange of RNA polymerase II initiation and elongation factors during gene expression in vivo. *Mol Cell* **2002**;*9*(4):799–809.

K. Poorey, R. Viswanathan, M. N. Carver, T. S. Karpova, S. M. Cirimotich, J. G. McNally, S. Bekiranov, and D. T. Auble. Measuring chromatin interaction dynamics on the second time scale at single-copy genes. *Science* **2013**;*342*(6156):369–372.

D. Prather, N. J. Krogan, A. Emili, J. F. Greenblatt, and F. Winston. Identification and characterization of Elf1, a conserved transcription elongation factor in Saccharomyces cerevisiae. *Mol Cell Biol* **2005**;*25*(22):10122–10135.

S. Proshkin, A. R. Rahmouni, A. Mironov, and E. Nudler. Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science* **2010**;*328*(5977):504–508.

K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **2007**; *35*(Database issue):D61–D65.

M. Ptashne. The chemistry of regulation of genes and other things. *J Biol Chem* **2014**;*289*(9):5417–5435.

O. Puig, F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, and B. Séraphin. The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **2001**;*24*(3):218–229.

H. Qiu, C. Hu, and A. G. Hinnebusch. Phosphorylation of the Pol II CTD by Kin28 enhances Bur1/Bur2 recruitment and Ser2 CTD phosphorylation near promoters. *Mol Cell* **2009**;*33*(6):752–762.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, **2015**.

E. A. Rach, H.-Y. Yuan, W. H. Majoros, P. Tomancak, and U. Ohler. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. *Genome Biol* **2009**;*10*(7):R73.

M. Radonjic, J.-C. Andrau, P. Lijnzaad, P. Kemmeren, T. T. J. P. Kockelkorn, D. van Leenen, N. L. van Berkum, and F. C. P. Holstege. Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon S. cerevisiae stationary phase exit. *Mol Cell* **2005**;*18*(2):171–183.

P. B. Rahl, C. Y. Lin, A. C. Seila, R. A. Flynn, S. McCuine, C. B. Burge, P. A. Sharp, and R. A. Young. c-Myc regulates transcriptional pause release. *Cell* **2010**;*141*(3):432–445.

E.-A. Raiber, R. Kranaster, E. Lam, M. Nikan, and S. Balasubramanian. A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro. *Nucleic Acids Res* **2012**; *40*(4):1499–1508.

E.-A. Raiber, P. Murat, D. Y. Chirgadze, D. Beraldi, B. F. Luisi, and S. Balasubramanian. 5-Formylcytosine alters the structure of the DNA double helix. *Nat Struct Mol Biol* **2015**;*22*(1):44–49.

A. S. Rajkumar, N. Dénervaud, and S. J. Maerkl. Mapping the fine structure of a eukaryotic promoter input-output function. *Nat Genet* **2013**;*45*(10):1207–1215.

V. K. Rakyan, T. A. Down, D. J. Balding, and S. Beck. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* **2011**;*12*(8):529–541.

T. Raveh-Sadka, M. Levo, U. Shabi, B. Shany, L. Keren, M. Lotan-Pompan, D. Zeevi, E. Sharon, A. Weinberger, and E. Segal. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat Genet* **2012**;*44*(7):743–750.

D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A. Smith, C. A. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. F. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, and T. R. Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **2013**;*499*(7457):172–177.

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* **2000**;*290*(5500):2306–2309.

D. B. Renner, Y. Yamaguchi, T. Wada, H. Handa, and D. H. Price. A highly purified RNA polymerase II elongation control system. *J Biol Chem* **2001**;*276*(45):42601–42609.

H. S. Rhee and B. F. Pugh. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **2011**;*147*(6):1408–1419.

H. S. Rhee and B. F. Pugh. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **2012**;*483*(7389):295–301.

C. R. Rodriguez, E. J. Cho, M. C. Keogh, C. L. Moore, A. L. Greenleaf, and S. Buratowski. Kin28, the TFIIH-associated carboxy-terminal domain kinase, facilitates the recruitment of mRNA processing machinery to RNA polymerase II. *Mol Cell Biol* **2000**;*20*(1):104–112.

R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann. Origins of specificity in protein-DNA recognition. *Annu Rev Biochem* **2010**;*79*:233–269.

R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig. The role of DNA shape in protein-DNA recognition. *Nature* **2009**;*461*(7268):1248–1253.

A. Rotem, O. Ram, N. Shoresh, R. A. Sperling, A. Goren, D. A. Weitz, and B. E. Bernstein. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol* **2015**; *33*(11):1165–1172.

S. Rouskin, M. Zubradt, S. Washietl, M. Kellis, and J. S. Weissman. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **2014**;*505*(7485):701–705.

S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* **1998**;*26*(2):544–548.

N. Sasai, M. Nakao, and P.-A. Defossez. Sequence-specific recognition of methylated DNA by human zinc-finger proteins. *Nucleic Acids Res* **2010**;*38*(15):5015–5022.

D. Schmidt, P. C. Schwalie, M. D. Wilson, B. Ballester, A. Gonçalves, C. Kutter, G. D. Brown, A. Marshall, P. Flicek, and D. T. Odom. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **2012**;*148*(1-2):335–348.

T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **1990**;*18*(20):6097–6100.

A. Schreieck, A. D. Easter, S. Etzold, K. Wiederhold, M. Lidschreiber, P. Cramer, and L. A. Passmore. RNA polymerase II termination involves C-terminal-domain tyrosine dephosphorylation by CPF subunit Glc7. *Nat Struct Mol Biol* **2014**;*21*(2):175–179.

S. C. Schroeder, B. Schwer, S. Shuman, and D. Bentley. Dynamic association of capping enzymes with transcribing RNA polymerase II. *Genes Dev* **2000**;*14*(19):2435–2440.

D. Schulz, B. Schwalb, A. Kiesel, C. Baejen, P. Torkler, J. Gagneur, J. Söding, and P. Cramer. Transcriptome surveillance by selective termination of noncoding RNA synthesis. *Cell* **2013**; *155*(5):1075–1087.

B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. Global quantification of mammalian gene expression control. *Nature* **2011**;*473*(7347):337–342.

S. Schwartz, S. D. Agarwala, M. R. Mumbach, M. Jovanovic, P. Mertins, A. Shishkin, Y. Tabach, T. S. Mikkelsen, R. Satija, G. Ruvkun, S. A. Carr, E. S. Lander, G. R. Fink, and A. Regev. High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell* **2013**;*155*(6):1409–1421.

N. C. Seeman, J. M. Rosenberg, and A. Rich. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci USA* **1976**;*73*(3):804–808.

E. Segal and J. Widom. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr Opin Struct Biol* **2009**;*19*(1):65–71.

O. Shalem, E. Sharon, S. Lubliner, I. Regev, M. Lotan-Pompan, Z. Yakhini, and E. Segal. Systematic dissection of the sequence determinants of gene $3'$ end mediated expression control. *PLoS Genet* **2015**;*11*(4):e1005147.

E. Sharon, S. Lubliner, and E. Segal. A feature-based approach to modeling protein-DNA interactions. *PLoS Comput Biol* **2008**;*4*(8):e1000154.

T. Shiraki, S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa, S. Fukuda, D. Sasaki, A. Podhajska, M. Harbers, J. Kawai, P. Carninci, and Y. Hayashizaki. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci USA* **2003**; *100*(26):15776–15781.

M. Siebert, M. Lidschreiber, H. Hartmann, and J. Söding. A guideline for ChIP-chip data quality control and normalization. *EpiGeneSys* **2009**;*protocol 47*.

M. Siebert and J. Söding. Universality of core promoter elements? *Nature* **2014**;*511*(7510):E11–E12.

T. Siggers, M. H. Duyzend, J. Reddy, S. Khan, and M. L. Bulyk. Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol Syst Biol* **2011**; *7*(1):555.

T. Siggers and R. Gordân. Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Res* **2014**;*42*(4):2099–2111.

M. Slattery, T. Riley, P. Liu, N. Abe, P. Gomez-Alcala, I. Dror, T. Zhou, R. Rohs, B. Honig, H. J. Bussemaker, and R. S. Mann. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* **2011**;*147*(6):1270–1282.

M. Slattery, T. Zhou, L. Yang, A. C. Dantas Machado, R. Gordân, and R. Rohs. Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci* **2014**;*39*(9):381–399.

C.-X. Song, K. E. Szulwach, Q. Dai, Y. Fu, S.-Q. Mao, L. Lin, C. Street, Y. Li, M. Poidevin, H. Wu, J. Gao, P. Liu, L. Li, G.-L. Xu, P. Jin, and C. He. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **2013**;*153*(3):678–691.

A. Soufi, M. F. Garcia, A. Jaroszewicz, N. Osman, M. Pellegrini, and K. S. Zaret. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **2015**; *161*(3):555–568.

C. G. Spruijt, F. Gnerlich, A. H. Smits, T. Pfaffeneder, P. W. T. C. Jansen, C. Bauer, M. Münzel, M. Wagner, M. Müller, F. Khan, H. C. Eberl, A. Mensinga, A. B. Brinkman, K. Lephikov, U. Müller, J. Walter, R. Boelens, H. van Ingen, H. Leonhardt, T. Carell, and M. Vermeulen. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **2013**;*152*(5):1146–1159.

R. Staden. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res* **1984**; *12*(1 Pt 2):505–519.

R. Stefl, L. Skrisovska, and F. H.-T. Allain. RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. *EMBO Rep* **2005**;*6*(1):33–38.

G. D. Stormo. Modeling the specificity of protein-DNA interactions. *Quant Biol* **2013**;*1*(2):115–130.

G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht. Use of the ‘Perceptron’ algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res* **1982**;*10*(9):2997–3011.

G. D. Stormo and Y. Zhao. Determining the specificity of protein-DNA interactions. *Nat Rev Genet* **2010**;*11*(11):751–760.

K. Strässer and E. Hurt. Yra1p, a conserved nuclear RNA-binding protein, interacts directly with Mex67p and is required for mRNA export. *EMBO J* **2000**;*19*(3):410–420.

T. Stuwe, M. Hothorn, E. Lejeune, V. Rybin, M. Bortfeld, K. Scheffzek, and A. G. Ladurner. The FACT Spt16 “peptidase” domain is a histone H3-H4 binding module. *Proc Natl Acad Sci USA* **2008**;*105*(26):8884–8889.

W. Sun, X. Hu, M. H. K. Lim, C. K. L. Ng, S. H. Choo, D. S. Castro, D. Drechsel, F. Guillemot, P. R. Kolatkar, R. Jauch, and S. Prabhakar. TherMos: estimating protein-DNA binding energies from in vivo binding profiles. *Nucleic Acids Res* **2013**;*41*(11):5555–5568.

V. Sundaram, Y. Cheng, Z. Ma, D. Li, X. Xing, P. Edge, M. P. Snyder, and T. Wang. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res* **2014**;*24*(12):1963–1976.

M.-H. Sung, M. J. Guertin, S. Baek, and G. L. Hager. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol Cell* **2014**;*56*(2):275–285.

K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **2006**;*126*(4):663–676.

L. Teytelman, D. M. Thurtle, J. Rine, and A. van Oudenaarden. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc Natl Acad Sci USA* **2013**; *110*(46):18602–18607.

The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**;*489*(7414):57–74.

R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B.-K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature* **2012**; *489*(7414):75–82.

B. Tian and J. H. Graber. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip Rev RNA* **2012**;*3*(3):385–396.

L. T. Timchenko, J. W. Miller, N. A. Timchenko, D. R. DeVore, K. V. Datar, L. Lin, R. Roberts, C. T. Caskey, and M. S. Swanson. Identification of a $(\mathrm{CUG})_\mathrm{n}$ triplet repeat RNA-binding protein and its expression in myotonic dystrophy. *Nucleic Acids Res* **1996**;*24*(22):4407–4414.

J. Toedling, O. Skylar, O. Sklyar, T. Krueger, J. J. Fischer, S. Sperling, and W. Huber. Ringo–an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics* **2007**;*8*:221.

A. Tomovic and E. J. Oakeley. Position dependencies in transcription factor binding sites. *Bioinformatics* **2007**;*23*(8):933–941.

J. Toth and M. D. Biggin. The specificity of protein-DNA crosslinking by formaldehyde: in vitro and in Drosophila embryos. *Nucleic Acids Res* **2000**;*28*(2):e4.

B. M. Turner. Defining an epigenetic code. *Nat Cell Biol* **2007**;*9*(1):2–6.

E. Ullu and C. Tschudi. Alu sequences are processed 7SL RNA genes. *Nature* **1984**;*312*(5990):171–172.

G. Vansant and W. F. Reynolds. The consensus sequence of a major Alu subfamily contains a functional retinoic acid response element. *Proc Natl Acad Sci USA* **1995**;*92*(18):8229–8233.

J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **2009**;*10*(4):252–263.

K. E. Varley, J. Gertz, K. M. Bowling, S. L. Parker, T. E. Reddy, F. Pauli-Behn, M. K. Cross, B. A. Williams, J. A. Stamatoyannopoulos, G. E. Crawford, D. M. Absher, B. J. Wold, and R. M. Myers. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res* **2013**; *23*(3):555–567.

B. J. Venters and B. F. Pugh. A canonical promoter organization of the transcription machinery and its regulators in the Saccharomyces genome. *Genome Res* **2009**;*19*(3):360–371.

B. J. Venters and B. F. Pugh. Genomic organization of human transcription initiation complexes. *Nature* **2013**;*502*(7469):53–58.

B. J. Venters and B. F. Pugh. Retraction: Genomic organization of human transcription initiation complexes. *Nature* **2014**;*513*(7518):444.

C. P. Verrijzer, M. J. Alkema, W. W. van Weperen, H. C. Van Leeuwen, M. J. Strating, and P. C. van der Vliet. The DNA binding specificity of the bipartite POU domain and its subdomains. *EMBO J* **1992**;*11*(13):4993–5003.

J. Vierstra, H. Wang, S. John, R. Sandstrom, and J. A. Stamatoyannopoulos. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nat Methods* **2014**;*11*(1):66–72.

P. H. von Hippel. From "simple" DNA-protein interactions to the macromolecular machines of gene expression. *Annu Rev Biophys Biomol Struct* **2007**;*36*:79–105.

P. H. von Hippel and O. G. Berg. On the specificity of DNA-protein interactions. *Proc Natl Acad Sci USA* **1986**;*83*(6):1608–1612.

J. Wang, J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder, and Z. Weng. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* **2012**;*22*(9):1798–1812.

X. Wang, Z. Lu, A. Gomez, G. C. Hon, Y. Yue, D. Han, Y. Fu, M. Parisien, Q. Dai, G. Jia, B. Ren, T. Pan, and C. He. N^6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **2014**;*505*(7481):117–120.

L. D. Ward, J. Wang, and H. J. Bussemaker. Characterizing a collective and dynamic component of chromatin immunoprecipitation enrichment profiles in yeast. *BMC Genomics* **2014**;*15*:494.

C. L. Warren, N. C. S. Kratochvil, K. E. Hauschild, S. Foister, M. L. Brezinski, P. B. Dervan, G. N. Phillips, Jr, and A. Z. Ansari. Defining the sequence-recognition profile of DNA-binding molecules. *Proc Natl Acad Sci USA* **2006**;*103*(4):867–872.

L. C. Watson, K. M. Kuchenbecker, B. J. Schiller, J. D. Gross, M. A. Pufall, and K. R. Yamamoto. The glucocorticoid receptor dimer interface allosterically transmits sequence-specific DNA signals. *Nat Struct Mol Biol* **2013**;*20*(7):876–883.

M. T. Weirauch, A. Cote, R. Norel, M. Annala, Y. Zhao, T. R. Riley, J. Saez-Rodriguez, T. Cokelaer, A. Vedenko, S. Talukder, DREAM5 Consortium, H. J. Bussemaker, Q. D. Morris, M. L. Bulyk, G. Stolovitzky, and T. R. Hughes. Evaluation of methods for modeling transcription factor sequence specificity. *Nat Biotechnol* **2013**;*31*(2):126–134.

J. W. Whitaker, Z. Chen, and W. Wang. Predicting the human epigenome from DNA motifs. *Nat Methods* **2015**;*12*(3):265–272.

M. A. White, C. A. Myers, J. C. Corbo, and B. A. Cohen. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc Natl Acad Sci USA* **2013**;*110*(29):11952–11957.

Z. Wunderlich and L. A. Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet* **2009**;*25*(10):434–440.

Z. Xu, W. Wei, J. Gagneur, F. Perocchi, S. Clauder-Münster, J. Camblong, E. Guffanti, F. Stutz, W. Huber, and L. M. Steinmetz. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **2009**;*457*(7232):1033–1037.

J. O. Yáñez-Cuna, H. Q. Dinh, E. Z. Kvon, D. Shlyueva, and A. Stark. Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding. *Genome Res* **2012**;*22*(10):2018–2030.

G. G. Yardımcı, C. L. Frank, G. E. Crawford, and U. Ohler. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res* **2014**; *42*(19):11865–11878.

S. M. Yoh, H. Cho, L. Pickle, R. M. Evans, and K. A. Jones. The Spt6 SH2 domain binds Ser2-P RNAPII to direct Iws1-dependent mRNA splicing and export. *Genes Dev* **2007**;*21*(2):160–174.

T. Yoshida, T. Ohkumo, S. Ishibashi, and K. Yasuda. The 5′-AT-rich half-site of Maf recognition element: a functional target for bZIP transcription factor Maf. *Nucleic Acids Res* **2005**; *33*(11):3465–3478.

M. Yu, G. C. Hon, K. E. Szulwach, C.-X. Song, L. Zhang, A. Kim, X. Li, Q. Dai, Y. Shen, B. Park, J.-H. Min, P. Jin, B. Ren, and C. He. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **2012**;*149*(6):1368–1380.

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* **2005**;*309*(5734):626–630.

N. R. Zabet and B. Adryan. Estimating binding properties of transcription factors from genome-wide binding profiles. *Nucleic Acids Res* **2015**;*43*(1):84–94.

M. A. Zabidi, C. D. Arnold, K. Schernhuber, M. Pagani, M. Rath, O. Frank, and A. Stark. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **2015**; *518*(7540):556–559.

B. Zacher, P. F. Kuan, and A. Tresch. Starr: Simple Tiling ARRay analysis of Affymetrix ChIP-chip data. *BMC Bioinformatics* **2010**;*11*:194.

J. Zeitlinger, A. Stark, M. Kellis, J.-W. Hong, S. Nechaev, K. Adelman, M. Levine, and R. A. Young. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet* **2007**;*39*(12):1512–1516.

G. Zhang, H. Huang, D. Liu, Y. Cheng, X. Liu, W. Zhang, R. Yin, D. Zhang, P. Zhang, J. Liu, C. Li, B. Liu, Y. Luo, Y. Zhu, N. Zhang, S. He, C. He, H. Wang, and D. Chen. N^6-methyladenine DNA modification in Drosophila. *Cell* **2015**;*161*(4):893–906.

M. Q. Zhang and T. G. Marr. A weight array method for splicing signal analysis. *Comput Appl Biosci* **1993**;*9*(5):499–509.

X. Zhao, H. Huang, and T. P. Speed. Finding short DNA motifs using permuted Markov models. *J Comput Biol* **2005**;*12*(6):894–906.

Y. Zhao, D. Granas, and G. D. Stormo. Inferring binding energies from selected binding sites. *PLoS Comput Biol* **2009**;*5*(12):e1000590.

Y. Zhao, S. Ruan, M. Pandey, and G. D. Stormo. Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics* **2012**;*191*(3):781–790.

Y. Zhao and G. D. Stormo. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat Biotechnol* **2011**;*29*(6):480–483.

K. Zhou, W. H. W. Kuo, J. Fillingham, and J. F. Greenblatt. Control of transcriptional elongation and cotranscriptional histone modification by the yeast Bur kinase substrate Spt5. *Proc Natl Acad Sci USA* **2009**;*106*(17):6956–6961.

Q. Zhou and J. S. Liu. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics* **2004**;*20*(6):909–916.

T. Zhou, N. Shen, L. Yang, N. Abe, J. Horton, R. S. Mann, H. J. Bussemaker, R. Gordân, and R. Rohs. Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc Natl Acad Sci USA* **2015**;*112*(15):4654–4659.

T. Zhou, L. Yang, Y. Lu, I. Dror, A. C. Dantas Machado, T. Ghane, R. Di Felice, and R. Rohs. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res* **2013**;*41*(Web Server issue):W56–W62.

M. J. Ziller, H. Gu, F. Müller, J. Donaghey, L. T.-Y. Tsai, O. Kohlbacher, P. L. De Jager, E. D. Rosen, D. A. Bennett, B. E. Bernstein, A. Gnirke, and A. Meissner. Charting a dynamic DNA methylation landscape of the human genome. *Nature* **2013**;*500*(7463):477–481.

Z. Zuo and G. D. Stormo. High-resolution specificity from DNA sequencing highlights alternative modes of lac repressor binding. *Genetics* **2014**;*198*(3):1329–1343.