Understanding Protein-DNA Binding Events

Anja Sophie Kiesel

München 2017

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

Understanding Protein-DNA Binding Events

Anja Sophie Kiesel
aus München, Deutschland

2017

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, den ....

Anja Sophie Kiesel

Dissertation eingereicht am 10.10.2017

1. Gutachter: Dr. Johannes Söding
2. Gutachter: Prof. Dr. Julien Gagneur

Mündliche Prüfung am 30.11.2017

Acknowledgments

First and foremost I would like to thank my Ph.D. supervisor Dr. Johannes Soeding for giving me the opportunity to work in his group and become a member of the Soeding team. With several collaborative projects, I had the chance to dive into a broad variety of biological problems.

Thanks to all the members of my dissertation committee, Prof. Dr. Julien Gagneur, Prof. Dr. Förstermann, Dr. Dietmar Martin, Prof. Dr. Karl-Klaus Conzelmann and Prof. Dr. Karl Peter Hopfner. I appreciate your willingness to read and revise my work.

During the past years, I had the opportunity to work in several collaborations with bioinformaticians and biochemists which I really enjoyed. Therefore I would like to thank Schulle, Dr. Poony and the Tork-man for the amazing NUTs project, Alexandra for the exciting Histone collaboration, and the members of the Gaullab for a long-term project to dissect the Drosophila core promoter secrets. In this extent, I also want to thank Alessio, George, Martha, and Andrea for a great time during my lab internship at Alessio's bench. Thanks to the Tresh lab and the Gagneur lab, working with you and spending time with you really made my days. Thanks to the Göttinger Lab for being so much fun at the group retreats. Special Thanks go to Wanwan, you do a fantastic job, and I hope that your work will find its deserved appreciation. Thank you Schlauchi and Maria for always having something to laugh about and distributing this amazingly happy mood in the office. Thanks to Markus for lively discussions about the web server and the road trip from Fischbachau to Ingolstadt. Thank you, Vincent and Juri for being great office neighbors and many discussions about the most abstract things I can imagine. Thank you, Carina for not only being a great colleague but also for being my friend, conference room buddy and best B&B in Göttingen. Thank you so much, Holger and Matthias for being my mentors during the past year, not only concerning work-related topics. My biggest Thanks go to my long lasting warriors Susi and Mark, without you I may not have been able to finish. Mark, thank you for so much advice and never being annoyed by my never-ending questions and concerns, thanks for introducing the world of board games to me. Susi, you inspire me everyday with your never-ending enthusiasm and motivation and thank you for making me finally jump out of an airplane, indescribable!

I want to thank all my friends and family for the long lasting support and believe. Even if many of you do not understand what I am working on you still tried to listen to my explanations without being annoyed by another story about "these proteins in the cell". Thank you, Mom, and thank you, Dad, for your love and support especially the web server part and the encouragement to stick to the "physics" til the end. My beloved Bro's, I thank you for being there when I needed your help and showing interest in my "wicked" work.

Lastly, and most importantly, I want to express my greatest appreciation for the unconditional support from the best man on earth, Niklas you are my hero! Without you, I would not even have started a Ph.D. and probably dropped out several times. You always encouraged me to go on and made me perform at my best. I love you for being who you are, for supporting me in any part of our life, by always being there when I need you and by helping to make the most out of everything. You are incredible, and I am thankful for having you in my life. 5 years ago, you promised me a Porsche when I finish my Ph.D.. Now that the end is near, I would like to mention that desires change and that I would like to replace the Porsche for a Donkey.

Summary

DNA binding proteins regulate essential biological processes such as DNA replication, transcription, repair, and splicing. Transcription factors (TFs) are in the focus of this work because they have the largest effect of activating and repressing gene expression by influencing transcription rates. It is important to model TF binding affinity to DNA and to predict protein-DNA binding events to understand how they regulate cell mechanisms.

Higher order Markov models bring *de-novo* motif discovery to the next level. BaMM!motif has been shown to provide robust predictions of these more sophisticated binding models. Here I introduce the BaMM!motif web application, a web-based platform which combines *de-novo* motif discovery with motif enrichment and motif-motif comparison tools and a database of known motifs. This web application enables the usage of the BaMM!motif algorithm in a straightforward and robust environment.

Post-translational histone modifications and linker histone incorporation regulate chromatin structure and genome activity. How these systems interface on a molecular level is unclear. Using biochemistry, one observes that the modification behavior of N-terminal histone H3 tails depends on the nucleosomal contexts. I found that linker histones inhibit modifications of different H3 sites on a genome-wide level.This proposes that alterations of H3 tail-linker DNA interactions by linker histones execute basal control mechanisms of chromatin function.

Pervasive transcription of eukaryotic genomes stems to a large extent from bidirectional promoters that synthesize mRNA and divergent noncoding RNA (ncRNA). Here, I show that early termination that relies on the essential RNA-binding factor Nrd1 attenuates transcription of 32 genes in yeast. Further, depletion of Nrd1 from the nucleus results in 1,526 Nrd1-unterminated transcripts (NUTs) that originate from nucleosome-depleted regions (NDRs) and can deregulate mRNA synthesis by antisense repression and transcription interference.

Publications

The BaMM!motif webserver for *de-novo* motif discovery and regulatory sequence analysis

Kiesel A, Meier M, Ge W, Roth C, Wess M, Soeding J
(est 2018) Nucleic Acids Res(Web Server issue), *Manuscript in preparation*

Modulations of DNA Contacts by Linker Histones and Post-translational Modifications Determine the Mobility and Modifiability of Nucleosomal H3 Tails

Stützer A, Liokatis S, **Kiesel A**, Schwarzer D, Sprangers R, Soeding J, Selenko P, Fischle W
(2016) Mol Cell, DOI:10.1016/j.molcel.2015.12.2015

Patient-specific driver gene prediction and risk assesment through integrated network analysis of cancer omics profiles

Bertrand D, Chng KR, Sherbaf FG, **Kiesel A**, Chia BK, Sia YY, Huang SK, Hoon DS, Liu ET, Hillmer A, Nagarajan N
(2015) Nucleic Acids Res, DOI:10.1093/nar/gku1393

Transcriptome surveillance by selective termination of noncoding RNA synthesis

Schulz D, Schwalb B, **Kiesel A**, Baejen C, Torkler P, Gagneur J, Soeding J, Cramer P
(2013) Cell, DOI:10.1016/cell.2013.10.024

Contents

List of Figures

Chapter 1

Introduction to protein-DNA interactions

Protein-DNA interactions are commonly found in all living organisms. They play a major role in many essential biological processes such as DNA replication, transcription, repair, and splicing. Histones are proteins in the eukaryotic cell nuclei. They are involved in chromosome packaging, a mechanism to regulate DNA accessibility. The enzyme DNA Polymerase is needed to transcribe DNA into mRNA. Nucleases cleave DNA and glycosylases repair DNA breaks. Transcription factors (TFs) are in focus of this work on protein-DNA interactions because they have the largest effect of activating and repressing gene expression by influencing transcription rates. TFs are proteins that have at least one DNA-binding domain (DBD), which interacts with a DNA sequence pattern. It is important to model TF binding affinity to DNA and to predict protein-DNA binding events to understand how they regulate cell mechanisms.

1.1 Protein-DNA binding controls transcription

The rate of gene transcription determines protein abundance to a larger extent than post-transcriptional processes such as translation or degradation of mRNAs and proteins. Independent studies [14, 70, 85] show that transcription rates explain the majority of protein level variation. Schwanhäusser et al. [129] challenged this view by claiming that differences in translation rates dominate, with transcription rates explaining only 34% of the variance in protein abundances. However, Li et al. [86] ascribed 73% of protein level changes to transcription rates based on error-corrected estimates. Thus, transcription control mechanisms play a fundamental role in regulating protein abundance.

In 1961, Jacob and Monod [63] found that TF binding to regulatory DNA elements controls protein synthesis. This highlights TFs as the primary regulators for transcription rate [116]. Experimental measures of the correlation between genome-wide chromatin modification patterns and gene expression lead to the discovery of the epigenetic code, which questioned this claim [13, 150]. Soon computational approaches were published which use histone modification states to predict gene expression [40, 73]. These correlations turned out to be indirect effects instead of causal implications [46, 79]. TF binding can accurately predict histone modification patterns [16].

The binding of TFs to specific sequence pattern of DNA elements mediates transcriptional control and therefore also protein abundance. It is essential to model binding specificity and genome-wide binding patterns of TFs and other proteins to uncover regulatory networks.

1.2 Measuring protein-DNA binding

Many *in vitro* and *in vivo* experiments have been developed to detect protein-DNA binding specificities using high-throughput measurements [142].

Protein-binding microarrays (PBMs) expose short DNA fragments as a binding platform for a target TF. Due to the number of fragments and combinatorics, the PBM represents all possible ten-base-long DNA sequences with every eight-base-long fragment occurring 32 times each. A fluorescent antibody then binds to the protein and is used to measure binding specificity [6, 18, 48]. Cognate site identifier (CSI) arrays [158] also use DNA fragments to measure relative binding. In comparison to PBMs, single-stranded DNAs fold back to form double-stranded DNA binding sites to eliminate the need for primer-directed DNA synthesis.

Other high-throughput approaches include the systematic evolution of ligands by exponential enrichment (SELEX) [68, 69] or microfluidic devices that mechanically induce trapping of molecular interactions (MITOMI) [45, 92].

The *in vivo* approaches, such as chromatin immunoprecipitation followed by microarray (ChIP-chip) [120] or sequencing (ChIP-seq) [67], determine the locations within the genome that the TFs bind. This provides candidate genes that are likely to be regulated. Their resolution is not high enough to define exact binding sites. Motif discovery algorithms need to analyze the sequence information afterward. Gilmour and Lis first introduced ChIP of proteins cross-linked to DNA in 1984 [47]. In a ChIP experiment, all DNA-binding proteins are crosslinked to DNA, i.e., via formaldehyde. Next, sonication shears the chromatin into short fragments. The fragments have a length of 200-600 bps. Specific antibodies against the protein of interest precipitate the target protein-DNA complexes. Reverse crosslinking releases the DNA fragments of interest from the complexes. The obtained solution with fragments is purified, amplified via PCR, sequenced and mapped to a reference genome to get the binding sites of the protein of interest. ChIP-exo [121] and ChIP-nexus [56] increased the resolution of ChIP-derived protein-DNA binding footprints by introducing an extra exonuclease step, which digests the ends of the DNA fragments before high-throughput sequencing.

A target-specific antibody is necessary to perform a ChIP-seq experiment. Alternatives for measuring binding events without antibodies *in vivo* include DNase-seq [60, 111, 149] and ATAC-seq [21]. In both cases, regions of open chromatin are measured in a genome-wide fashion by cutting accessible DNA and sequencing the fragment ends. DNase-seq uses DNase I, while ATAC-seq uses transposase to digest chromatin. These experimental assays offer an antibody-free alternative to measure more than one TF at a time, by determining collectively occupied genomic regions. These footprints need to be assigned to proteins computationally, which is difficult, especially for paralogous TFs.

1.3 Protein-DNA binding patterns

A protein binds to DNA by forming specific hydrogen bonds and electrostatic interactions. These interactions take place at the geometrically precisely fitting intersection of the DNA core binding site and the protein's interface [154, 155]. While this binding mechanism solely depends on the sequence (base read-out), statistical-mechanical selection theory indicated further properties that influence functional specificity [17]. The dependency of neighboring nucleotides within a binding site was reported for several proteins [22, 93] and could partly be attributed to a readout mechanism that depends on structural properties of the DNA molecule (shape readout), such as stacking interactions [123].

Determinants of binding locations were found to be even more complex, since secondary effects, such as cooperativity with other TFs, chromatin accessibility or competition with nucleosomes, also affect protein-DNA binding events [84, 123, 133, 136]. Binding determinants are highly interdependent, resulting in smooth transitions between base-readout and shape-readout concepts.

Only a small amount ($< 30\%$) [166] of the estimated 10^4 to 3×10^5 TF molecules present in a cell [19, 87] are specifically bound to DNA at any given time. A similar amount of factors bind in an unspecific manner [107] by i.e., sliding along the DNA [53]. Only specifically binding TFs have been found to interact in productive transcription [106].

1.4 Modeling transcription factor binding sites

In order to model the binding preference of a TF to a DNA sequence x of length L, the DNA bases $x_1, \ldots, x_L \in \{A, C, G, T\}$ are denoted by random variables $X_1, \ldots, X_L$.

According to Boltzmann's law, the probability of binding is related to the Gibbs free energy of binding ΔE by

$$p\left(X_1 \ldots X_L = x_1 \ldots x_L\right) \propto e^{-\frac{\Delta E(x_1 \ldots x_L)}{k_B^T}} \tag{1.1}$$

Where k_B is the Boltzmann constant, and T is the temperature. Applying Bolzmann's law assumes that the TF occupancy on the sequence is close to zero (weak binding approximation).

This Section covers models that compromise between simplicity and accuracy to estimate TF binding affinities to DNA sequence.

Position weight matrix

The most common model for protein binding sites is the position weight matrix (PWM), introduced by Stormo [141]. It defines the importance of the four bases adenine (A), cytosine (C), guanine (G) and thymine (T) to the binding affinity between a TF and a DNA sequence $x_1 \dots x_L$.

The model is based on the idea to approximate the binding probability of a TF to a DNA sequence by assuming that each position i in the binding site contributes to the binding energy independently (see Equation 1.2).

$$p(x_1 \dots x_L) = p_1(x_1)\, p_2(x_2|x_1)\, p_3(x_3|x_1x_2) \dots p_L(x_L|x_1 \dots x_{L-1}) \tag{1.2}$$

With this assumption, the binding probability can be approximated by the product of the individual probabilities $p_i(x_i)$ of base $x_i \in \{A, C, G, T\}$ being present at position i. (see Equation 1.3).

$$p(x_1 \dots x_L) \approx \prod_{i=1}^{L} p_i(x_i) \tag{1.3}$$

For N sequences of known binding sites, one can estimate the probability $p_i(x_i)$ of base x_i being at position i of the binding site by:

$$p_i(x_i) = \frac{n_i(x_i)}{N} \tag{1.4}$$

Where $n_i(x_i)$ is the number of times base x_i has been observed at position i in all N sequences.

For accurate probability estimates, a sufficient amount of occurrences of all possible 4^L sequences would need to be reflected in the data. Since the length of a TF binding site can easily reach up to 20 base pairs (bp), the amount of data may be a limiting factor for the accuracy of the model. It is necessary to compensate missing data depth by introducing pseudocounts. In general pseudocounts are designed to be proportional to the background frequency $f(x_i)$ of base x_i:

$$p_i(x_i) = \frac{n_i(x_i) + \alpha f(x_i)}{N + \alpha} \tag{1.5}$$

Where α is the total amount of pseudo counts to add. The influence of pseudo counts decreases and becomes negligible with increasing data depth, i.e., high counts $n_i(x_i)$.

The significant advantage of the PWM model is the small number of parameters and robustness for little data depth. It requires $3 \times L$ compared to $4^L - 1$ parameters in the full model. Due to the assumption of independence between positions, the PWM model can only capture base readout but lacks any contributions to the binding energy based on a shape readout. Although PWMs work well in many cases and are still widely used for evaluating protein-DNA binding specificity [36, 81, 140], their accuracy is debated [81].

inhomogeneous Markov Models (iMMs)

Inhomogeneous means position-specific while the order defines the positional range of the dependency. A PWM corresponds to an inhomogeneous Markov model (iMM) of order zero. Since in a PWM model, the conditional probabilities are approximated with its monomer probabilities $p_i(x_i)$, information about correlations between positions is lost. iMMs of higher order k can retain information about the correlations between $k+1$ neighboring positions:

$$p\left(x_1 \ldots x_L\right) \propto \prod_{i=1}^{L} p_i\left(x_i | x_{i-k} \ldots x_{i-1}\right) \tag{1.6}$$

The probability of base x_i at position i is now depending on its preceding binding site positions $i-k$ to $i-1$. The conditional probabilities can be calculated from the sequence counts. They are corrected with pseudo counts which are proportional to the monomer background frequencies (see Equation 1.7).

$$p_i\left(x_i | x_{i-k} \ldots x_{i-1}\right) \propto \frac{n_i\left(x_{i-k} \ldots x_i\right) + \alpha f\left(x_i\right)}{n_{i-1}\left(x_{i-k} \ldots x_{i-1}\right) + \alpha} \tag{1.7}$$

By taking the context into account, iMMs can capture stacking interactions between neighboring bases and local structural properties. However, the model complexity rises with the model order k, leading to more parameters. More data is needed to reflect all possible oligomers to obtain an accurate model. This makes iMMs more likely to be affected by statistical noise. When using an optimization process to learn a TF binding site model, this means that iMMs are more prone to overfitting than PWMs.

inhomogeneous Interpolated Markov Models (iIMMs)

To reduce the risk of overfitting, Salzberg et al. introduced interpolated Markov models (IMMs) [124]. inhomogeneous Interpolated Markov Models (iIMMs) combine the ideas of interpolation and inhomogeneous adaptation. Here, adjusting the higher-order oligomer counts with pseudo counts based on the lower-order oligomer probabilities instead of using fixed monomeric background frequencies infers dependencies of neighboring bases (see Equation 1.8).

$$p_i\left(x_i|x_{i-k}\ldots x_{i-1}\right) \propto \frac{n_i\left(x_{i-k}\ldots x_i\right)+\alpha_k p_i\left(x_i|x_{i-k+1}\ldots x_{i-1}\right)}{n_{i-1}\left(x_{i-k}\ldots x_{i-1}\right)+\alpha_k} \qquad (1.8)$$

iIMMs interpolate between counts and pseudo counts of the order below. For oligomers that are frequent in the data, the counts dominate over the pseudo counts, while the lower-order pseudo counts drive probabilities of underrepresented oligomers. This means, iIMMs do not require a minimum number of oligomer counts, thus are more robust to statistical noise than iMMs.

The BaMM!motif algorithm

The BaMM!motif algorithm was published by Siebert et al. in 2016 [132]. BaMM!motif is a Bayesian approach for motif discovery using iIMMs in which conditional probabilities of order k-1 act as priors for those of order k. This Bayesian Markov model (BaMM) training automatically adapts model complexity to the amount of available data.

The pseudo counts of the iIMMs are weighted in an order specific fashion (see Equation 1.9) because the interaction between neighboring nucleotides decreases with distance ([69]).

$$\alpha_k = 1,\ \text{if } k = 0; 20\times 3^{k-1},\ \text{if } k > 0 \qquad (1.9)$$

BaMM!motif derives an EM algorithm for *de-novo* discovery of enriched motifs.

The goal is to estimate the model parameter $p_{motif}(K)$, which is a vector containing the $W \times 4K+1$ conditional probabilities $p_j(K)(x_{K+1}|x_{1:K})$ for any $K+1$-mer $x_{1:K+1}$. The EM algorithm cycles between the E- and M-steps. The E-step estimates the probabilities for a motif to be present at position i of sequence n.

$$r_{ni} := p(z_n = i|x_n, p_{motif}(K)) \tag{1.10}$$

They use the zero-or-one-occurrence-per-sequence (ZOOPS) model [7]. The hidden variable z_n indicates where the motif occurs in sequence n. In the M-step they use the new r_{ni} to update the model parameter $p_{motif}(k)$ for all orders $k = 0, \ldots, K$. This update equation is equal to Equation 1.8 except that in this model the counts $n_j(x_{1:k})$ are interpreted as fractional counts (see Equation 1.11).

$$n_j(x_{1:k}) := \sum_n r_{ni} \mathbb{1}\left(x^n_{i+j-k:i-j-1} = x_{i:k}\right) \tag{1.11}$$

The indicator function returns 1 if the logical expression is true and 0 otherwise. The update of model parameters in the M-step runs through all orders from $k = 0 \ldots K$, using the just updated model parameters from the order below each time.

For transcription factor binding, BaMM!motif achieves significantly better motif models. In 97% of 446 ChIP-seq ENCODE datasets, the cross-validated partial AUC outperforms PWMs by an average improvement of 36% [132]. BaMM!motif also learns complex multipartite motifs, improving predictions of polyadenylation sites, transcription start sites, bacterial pause sites, and RNA binding sites. BaMM!motif never performed worse than PWMs. These robust improvements argue in favor of generally replacing PWMs by BaMM!motif derived models, called BaMMs.

1.5 Visualization of binding motifs

A sequence logo can visualize the content of a PWM [127]. The information content (in bits) for each position in a PWM is represented by the height of the

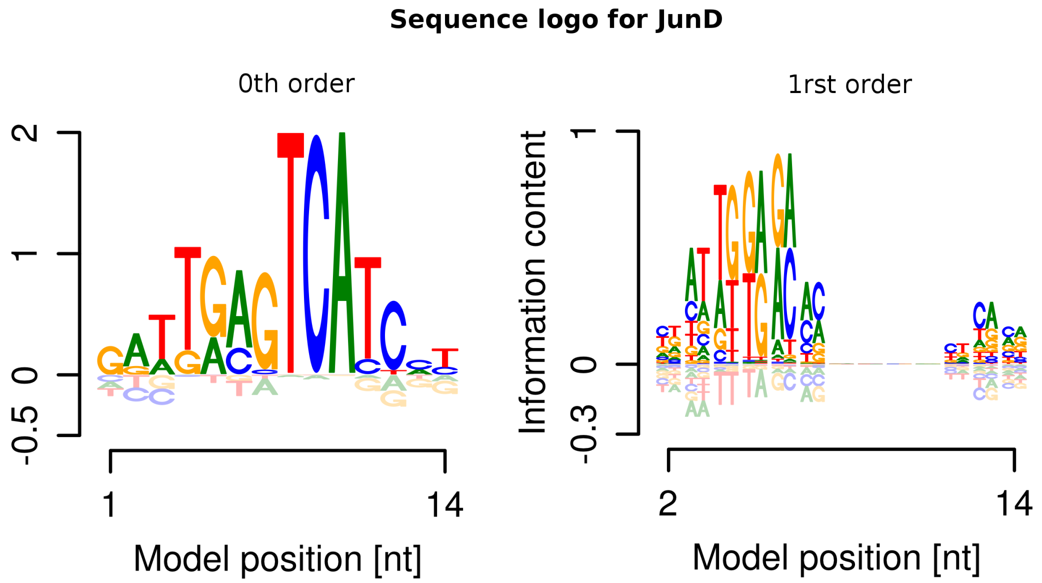


Figure 1.1: Sequence Logo for the transcription factor JunD. From the sequence logo for zeroth-order (left) one can infer the consensus motif. The sequence logo of the first-order informational gain (right) shows a variable spacer (one or zero bps) in between the two half sides of the homodimer.

columns in the logo. Their relative frequency determines the height of the four bases. The most frequent base is shown on top, this way it is possible to assemble the consensus motif. The overall height of each column describes the importance of each position in the motif for the protein-binding specificity. Since the sequence logo is designed as a position-specific visualization for PWMs, it does not reflect positional interdependencies. Extensions to the sequence logo have been proposed [41, 69, 75, 98, 131] by using a matrix based solution for showing higher-order dependencies. Siebert et al. [132] are the first ones that break the information content down to each order. This way, the informational gain for each order is depicted in a separate plot. Thus, the total motif information is realized as a collection of $K+1$ sequence logos (see Figure 1.1).

Chapter 2

BaMM!motif: A web application for *de-novo* motif discovery

To make higher order motif discovery available to a broad public of experimentalists and scientific computationalists, I have implemented the BaMM!motif web application.

The platform provides tools for *de-novo* motif discovery, motif occurrence search, motif-motif comparisons, as well as a database of previously predicted higher order TF binding models by the BaMM!motif algorithm.

First, I give an overview of available online platforms for motif discovery, and model databases (see Section 2.1). Next, I will explain and discuss which techniques I used to implement the BaMM!motif web application (see Sections 2.3-2.7), and introduce all available features (see Section 2.8).

2.1 Web applications for motif discovery

Several motif discovery and search tools have been described in literature [9, 42, 59, 66, 89, 115, 135, 147, 148]. The MEME suite [9] is the most prominent web platform, offering tools for motif discovery (DREME [8], MEME [7]), motif occurrence analysis (FIMO [49]) and motif comparison (TOMTOM [50]).

Common motif analysis tools work with the PWM representation of protein-

DNA bindings, hence neglecting interdependencies of adjacent nucleotides. Thus, the BaMM!motif web application aims to provide motif analysis tools similar to the MEME suite but operating on the more sophisticated higher-order motif models instead of PWMs.

Many databases such as JASPAR, HOCOMOCO, SwissRegulon or TRANSFAC [78, 97, 100, 114] provide thousands of TF binding site PWM models. First-order iMMs, sometimes called dinucleotide PWMs, are added to the JASPAR and HOCOMOCO databases. HOCOMOCO's first-order iMMs yielded better results than simple PWMs on average [78], and JASPAR's dinucleotide PWMs perform significantly better than PWMs for 21% of 96 tested datasets [99].

Hence, the BaMM!motif web application includes a database maintaining and distributing higher order models and is connected with motif search and comparison tools, while the BaMM!motif *de-novo* discovery algorithm provides the core analysis of the web application.

2.2 Technical aspects of the web application

Various tools and technologies are available for web development, to design not only a modern looking web interface but also to create a stable, fast and secure backend server. The following sections describe which technologies I used to implement the application logic (see Section 2.2.1). The web application backend setup focuses on flexibility to allow deployment (see Section2.2.2), scheduling asynchronous tasks (see Section 2.2.3) as well as ensuring robust and secure job calculations (see Section 2.2.4).

2.2.1 Web application framework

A web application is composed of a web server and web client. Both interact with each other. The server stores data and processes them while the client contains the user interface and delivers tasks to the server [108]. In former days the connectivity and the entire construct of the client-server architecture would have been implemented by hand solely aided by a common gateway interface (CGI) [61]. Since the connectivity and the core attributes are the same for each web application regardless of the content, web application frameworks are used to provide a common infrastructure for the client-server communication. Many frameworks are open source and use common scripting languages suitable for web applications. I have decided to use the web application framework Django for this project. It is mainly written in Python and integrates HTML templates for the client interface.

Django combines multiple tasks of a web application into a framework, hence, gives organization to the client-server architecture by providing a programming infrastructure and taking care of rudimentary communications between the front and back end. Django supports Python 3 and is the most popular Python-based web framework [37]. Since the BaMM!motif web application is meant to run long-term, maintainability, security as well as a commonly used programming language which facilities extendability, are essential. Django combines many of these features and is therefore selected as the web application framework for BaMM!motif.

Django's Model-View-Controller structure

Django uses a Model-View-Controller (MVC) like design pattern to structure the web application [113]. The code for defining and accessing data (the model) is separate from the user interface (the view), which again is separate from the request routing logic (the controller) [83].

A Model is defined in a python class and describes the design of a database table [113]. These model classes can be used to create, retrieve, update or delete records in the database using python commands rather than repetitive SQL statements. Django migrations propagate changes to the database schema. They automatically recognize differences in the database structure, store them in a separate migration directory and apply those to the complete database. Thus, migrations not only circumvent problems due to invalid database entries or logic but also work as a version control system for the database schema, since a new migration only stores changes to the previous version.

Figure 2.1 depicts a schema of all Django models from the BaMM!motif web application and their connectivity with each other. The database consists of six

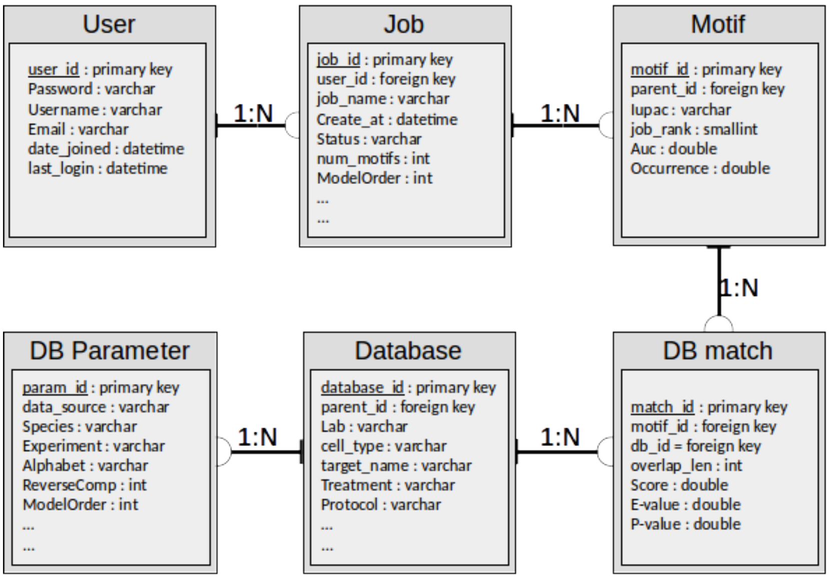


Figure 2.1: Organization of the BaMM!motif web server database which is used within Django to store models and results from user jobs as well as the database of predicted models for motif matching.

tables. The table "Job" contains information about a submitted job to the web server, like the input data location, BaMM!motif parameters, and the current status of the job. The table "Motif" holds the predicted binding site models from finished "Jobs". The table "Database" holds information about the publicly available input datasets and motifs predicted from them. Currently, this contains 446 entries based on ChIP-seq data sets from ENCODE [32]. Since several data sets from the same type (i.e., from ChIP-seq) are treated equally in the database, the BaMM!motif specific parameters which were used for motif discovery are saved separately in the table "DB-Parameter". The table "DB match" holds information about motif-motif comparisons between newly predicted motifs and motifs stored in the database. While a direct link between "Motif" and "Database" would lead to an N:N relation, the separate table enables to conclude 1:N relationships and offers room to store pair specific information, like the motif-motif comparison score and *P*-value.

A View is an endpoint that can be accessed by a client to retrieve information [164]. A Python function defines each view. Based on the input (i.e., request or further parameter), a specific template is loaded and rendered. Loading requires finding the correct template for a given identifier and preprocessing it by compiling it to an in-memory representation. This enables the developer to perform database searches and requesting specific information which is dynamically integrated into a static HTML output template and displayed in the client browser.

In the BaMM!motif web application, each web page is connected to a specific view function. If a client requests dynamic information, e.g., a result page from a finished job, this page will be built dynamically from the job entry in the model database according to the requested job id and the static HTML representation frame for result pages. Thus, the HTML template for result pages can be reused for any job type requested (i.e., *de-novo* motif discovery, motif occurrence search, or database entry). Python logic infers the differences of these pages and reduces templates to a minimum.

Views define which HTML templates will be filled and presented to the client side. Together with the URL pattern definition that connects a URL to each view, it reflects the business logic of the controller layer. BaMM!motif organizes URLs in (a) job submission, (b) job result, (c) database, and (d) general requests. Where applicable, the URL pattern contains the primary key of a submitted job or database entry which will be used within the view function to retrieve the corresponding information from the related model entry.

Figure 2.2 sketches how the Django framework reacts to a user request. When a user sends a URL request (1), the controller reads the URL and retrieves information which is needed for the corresponding view from the model (2). The model accesses the database (3) and sends the retrieved data (4) back to the controller (5). The controller communicates with the requested view (6), which integrates the database information and responds to the URL request (7). Django comes with default installed applications that offer functionality concerning administration, user authorization, messaging and static file management [37].

2.2.2 Deployment and portability

During the setup phase of a new web application, its workload and traffic are hard to estimate. The infrastructure and hardware which hosts the web application is shared with other projects and may change or become updated over time, while the

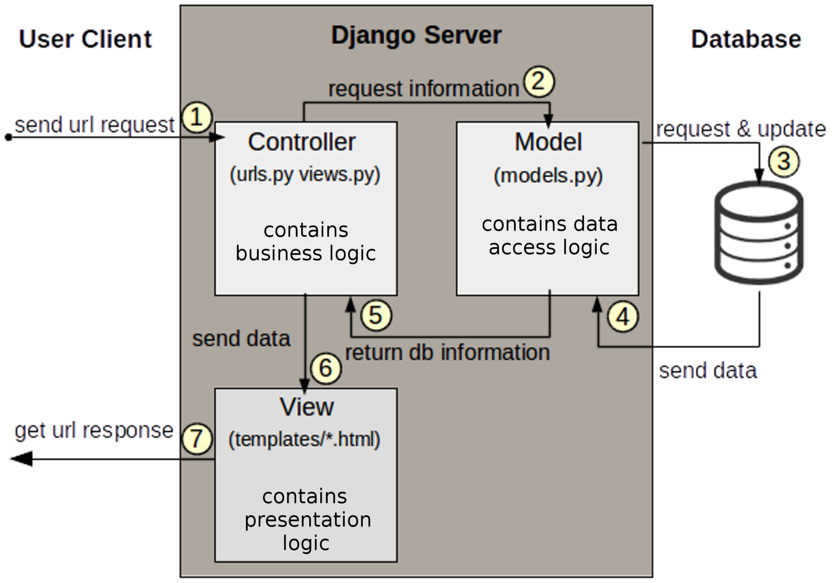


Figure 2.2: BaMM!motif webserver: Django's Model-View-Controller structure. When a client sends a URL request to the Django web server, the controller communicates with the model to obtain the requested data base information and sends it to the view which gives the requested URL response.

application itself might expand to multiple modules. A web application needs a flexible and easy to adapt deployment strategy to prepare for such circumstances.

The most common approaches are virtual machines and container-based systems. They enable to use the same environment during development and production phase. Software portability encapsulated in a running environment enormously speeds up deployment because it eliminates typical problems such as different system versions or missing plugins. Among all container solutions, Docker currently leads the market [38].

Docker containers versus virtual machines

Docker is a container virtualization technology, comparable to a lightweight virtual machine (VM). Each Docker container provides a separate encapsulated software environment. However, Docker containers differ in several points from VMs. A VM is a complete copy of an operating system running on a hypervisor on top of the physical hardware, which then executes a custom application. In contrast, a

Docker container interacts directly with the capabilities of the host's kernel level (see Figure 2.3).

Consequently, containers are much more lightweight than VMs, thus require less memory and are faster. Further, multiple Docker containers share supporting systems from the host, while VMs each need their own instance of a kernel. This is the reason why VMs need several minutes to launch while Docker containers can start up in less than a second [64, 71]. While Docker containers can all access one instance of shared resources [102], VMs need an instance for each client, making them less scaleable and harder to maintain.Note, that although Docker requires a Linux host system, it is possible to deploy Docker containers on MacOS and Windows, on which a stripped down Linux-based virtual machine is automatically configured as a host system.

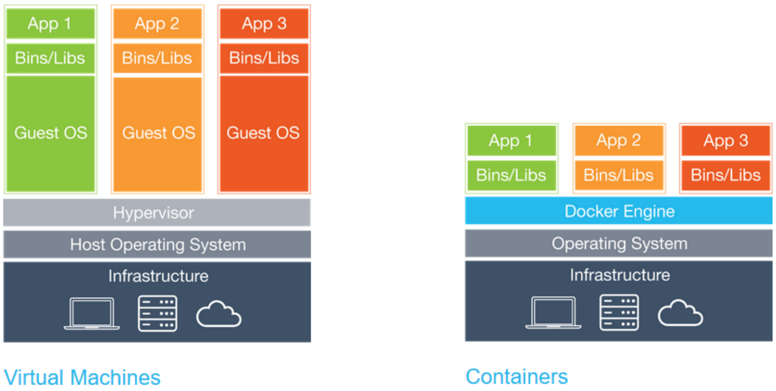


Figure 2.3: Docker Containers are more lightweight than VMs because they share the same supporting systems from the host, making them quicker in startup and less memory consuming. (Illustration is taken from https://mondedie.fr/d/7164-Tuto-Utilisation-de-Docker/2)

However, the full isolation of VMs has one significant advantage over Docker containers. Due to its complete separation from the host system, VMs offer an excellent protection by isolating for unwanted input/output (I/O) traffic. If a user or process manages to disrupt the kernel, it only impacts the current VM. Since Docker containers share a common kernel, its disruption would affect all Docker containers running on the same host system. This makes VMs the safer and more robust virtualization technology compared to conventional Docker containers. Nevertheless, several options and isolation levels can be applied to Docker containers to elevate safety and stability (Section 2.2.4 discusses details about Docker container security). Regarding the advantages of Docker containers over VMs and the ability to downsize their drawbacks, I have decided to package BaMM!motif into a Docker container virtualization.

Docker compose for multiple Docker containers

The idea of Docker containers suggests encapsulating one particular tool within a container. Most software systems, such as web applications need several tools to work efficiently, thus consist of multiple containers. To create and appropriately integrate several containers, Docker has introduced its extension Docker-compose [39]. It can be used to unify a setup routine for a group of containers and define interactions and access between containers. While pooling administration of multiple containers, Docker compose allows individual container settings to maintain high-level scalability.

The BaMM!motif Docker container infrastructure

The BaMM!motif web application consists of four containers; one web container and three service containers (see Figure 2.4).

The web container encapsulates the Django framework and the core of the web application. A MySQL service container holds the MySQL database where job results, as well as the precalculated binding site models, are stored. The Redis service container works in collaboration with the Celery service container to manage asynchronous tasks (described in detail in Section 2.2.3).

All containers run in the same Docker compose instance, and a separate Docker network connects them with each other. A Dockerfile defines the settings for the Docker instance. The Docker-compose client uses the information from the Dockerfile and the settings from the docker-compose.yml to interact with the Docker daemon, i.e., the Docker engine to launch the Docker compose instance.

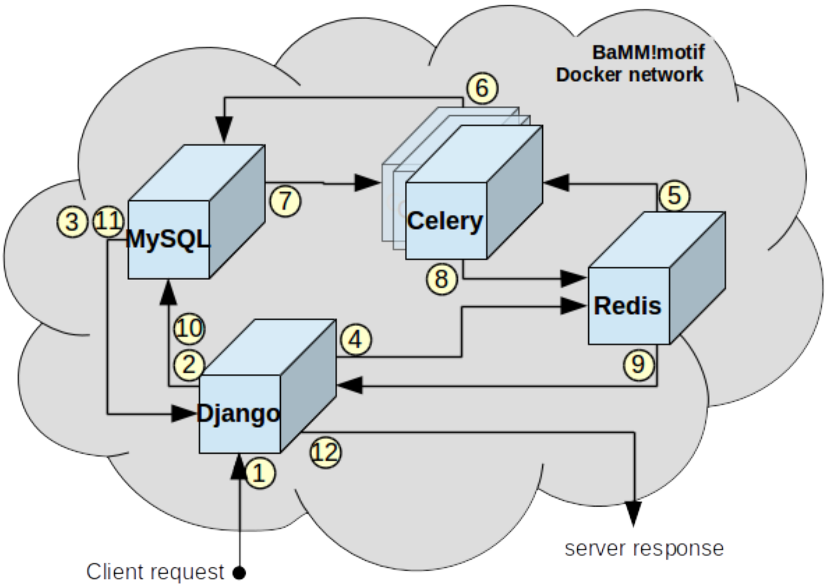


Figure 2.4: Docker Containers are launched in a separate Docker network. A user request enters into the Django container (1), which initiates a new job submission within the MySQL database (2) and retrieves its job ID (3). This is submitted to the broker container Redis(4), who schedules the job and assign a Celery worker to it (5). Depending on the workload, several Celery worker containers may be launched. The Celery container runs the job by accessing the corresponding MySQL entry and storing the results in it (6,7). When finished, the Celery worker signals to the broker (8) who forwards the finished job ID to the Django web container (9). Django then retrieves the results from the MySQL database (10,11) and responds to the user request (12).

2.2.3 Asynchronous job schedueling

When a user pushes the "BaMM!" button, he/she triggers the execution of the selected program with the provided input(s), i.e., submits a job. The main task of the web application is responding to requests from various users. A job scheduling layer takes care to organize and prioritize tasks asynchronously by sending them to the background. This way, the web application is not blocked by a job running in the foreground.

Celery is an asynchronous task queue manager that allows running asynchronous code by building up tasks queues [27]. It requires a system to send and receive messages, a message broker, which is realized as a separate service. Redis is a distributed messaging platform that handles the income of tasks and works as a backend cache system [119]. Due to the backend cache, Redis speeds up queries that run on the backend.

Celery helps the task managing in Django, Redis helps the database and all other sources in executing, managing, caching as well as messaging what is required from Celery to work efficiently. This way multiple tasks can be submitted and processed without disrupting the web page's response to user requests.

I have decided to use Celery as an asynchronous task queue manager with Redis as message broker because they are both available as Docker images, hence comfortable to set up as containers.

This also gives the opportunity to scale Celery according to the load balance by increasing the number of workers without unnecessarily upscaling other containers. Further, Django is capable of integrating Celery as an application into the web framework with a lightweight configuration file written in Python.

Once Django is setup to work with Celery, one can define Celery tasks written as Python functions. These tasks can be called from within Django views synchronously or asynchronously. This provides a comfortable maintaining environment having a uniform programming language.

2.2.4 Security aspects

Security systems are an important aspect to consider when operating a web application. The BaMM!motif web application handles sensitive private data uploads from clients, which need to be guaranteed to not be available to third parties. Further, malicious user actions may disrupt or even destroy the entire server.

Therefore, web applications, their environment, as well as the underlying system should be protected against attacks. The following section describes which actions are taken to decrease the BaMM!motif web application's vulnerability to a minimum.

Django's builtin security options

Django offers a way to protect against Cross-Site-Request-Forgery (CSRF). This is, an anonymous user tries to execute tasks under the credentials of another user without that user knowing. It currently ranks as the most vulnerable technique

[28]. By using the CSRF module appropriately, Django asks for a secret in each POST request based on a user-specific cookie.

If a browser connects with an HTTP connection instead of an HTTPS connection, it is possible for existing cookies to be leaked. Since CSRF protection uses cookies, BaMM!motif uses the Django options to work with secure sessions and CSRF cookies only.

Therefore, all HTTP requests are redirected over HTTPS connections using HTTP Strict Transport Security (HSTS) [11].

In certain cases, host headers provided by the client construct URLs. Even though sanitation prevents cross site scripting attacks, a false host value could lead to CSRF. Thus, Django validates host headers against a list of allowed hosts.

Docker's approach to increase security

Docker containers interact directly with the host's kernel system and have been criticized for being less secure than fully encapsulated virtual machines. Recent Docker releases follow best practices and valuable configurations were added to reduce security risks while maintaining a lightweight and fast structure [31].

Docker containers are built from Docker images which are downloaded from the public community DockerHub. BaMM!motif only uses ContentTrust certified images provided from Docker. Custom images may include invasive malware and are not recommended, thus also not used for the BaMM!motif web application.

Some file systems from the Linux kernel need to be mounted in a container to achieve proper process running. This is critical since mounting would also enable writing to folders, which is highly depreciated for system folders.

Therefore, Docker mounts these system files in a "read-only" state, elevating the isolation level for processors and memory access [2].

On default, a Docker container runs in root user mode. This enables that each process can be run regardless of its properties. However, this certainly increases the impact of malicious commands when having root privileges. Thus, all containers from the BaMMM!motif web applications are launched in user mode without root rights. This namespaced restricted mode eliminates the risk from container breakouts [64].

A separate network for the BaMM!motif Docker container further reduces the namespace. The network is defined within the Docker compose setup file and only contains BaMM!motif relevant resources. This eliminates potential crosstalks from third-party Docker containers running on the same machine.

Nginx as reverse proxy

Exposed I/O ports from Docker are only accessible via a Nginx gateway. The Nginx works as a reverse proxy that retrieves user data and communicates with one or more servers [112]. This way, it is possible to launch more than one web application on the same machine using the same ports (80 for HTTP; 443 for HTTPS).

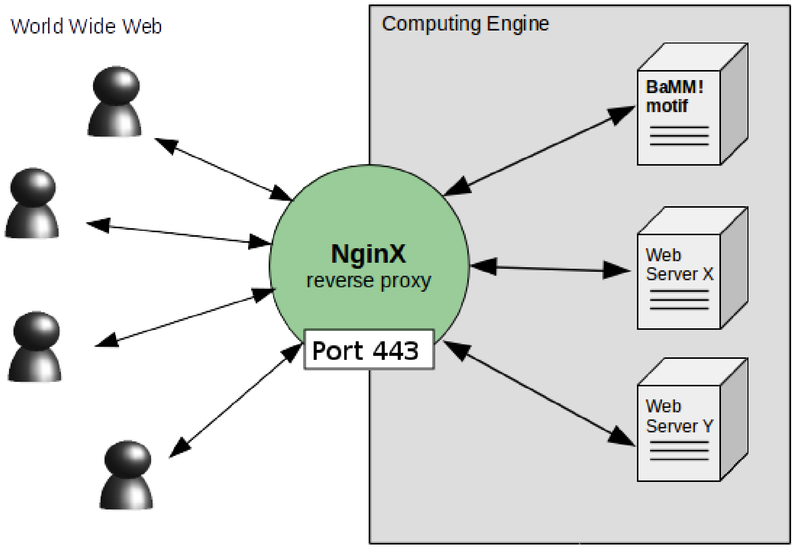


Figure 2.5: The Nginx works as a reverse proxy to operate several web applications via the same I/O ports while performing SSL encryption.

Additionally, Nginx performs SSL encryption which prevents unauthorized listening and enormously speeds up server response, thus only allowing HTTPS connections via port 443. While Django and Docker settings only protect the encapsulated web application, the Nginx functions as a gateway for the complete host server system.

Even though the Nginx makes the server better maintainable, it is crucial that each web application launched on the server follows a standard level of security by checking any user data, such as cookies, HTML header or parameter.

2.3 Operating the BaMM!motif web application

The BaMM!motif web application makes BaMM!motif an easy-to-use web-based tool. It provides *de-novo* motif discovery (see Section 2.4), evaluates model performance, compares new predictions with a database of known higher order BaMM!motif models (see Section 2.5), offers a search tool for motif occurrences (see Section 2.6), and contains a database of 446 BaMM!motif models from ChIP-seq ENCODE datasets (see Section 2.7). These functionalities and additional features (see Section 2.8) of the web application are explained in detail in the following sections.

Operating a task on the BaMM!motif web server follows a simple three-step process:

1. Task Selection: The main page provides three possible tasks from which the user can choose. (a) *De-novo* Motif Discovery, (b) Motif Occurrence Search, (c) Browse Motif Database or (d) Search with a Motif though a Motif Database.

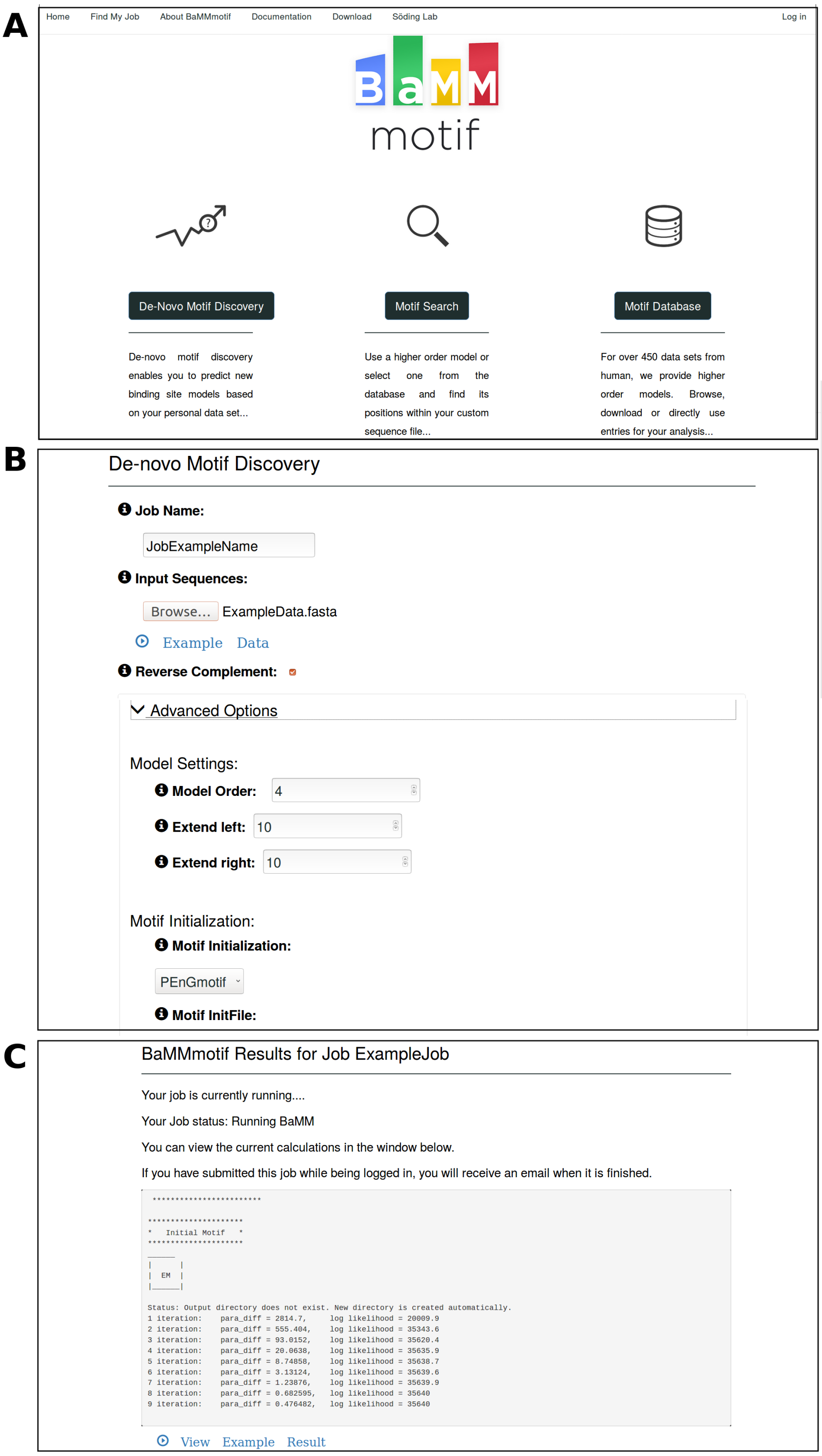
2. Data Upload: Based on the selected task the user is demanded to provide (a) a sequence set, (b) a motif model or (c) the name of a TF of interest.

3. Submission: Pressing the "BaMM!" button submits the desired query.

This straightforward and short procedure helps users to operate the BaMM!motif tool quickly and comfortable. More experienced users have the opportunity to adjust the process by entering the advanced sections for data input and parameter setups.

Figure 2.6 shows an exemplary workflow for operating *de-novo* motif discovery. The BaMM!motif main algorithm for *de-novo* motif discovery was developed by Matthias Siebert et al.[132] and introduced in Chapter 1. By selecting

"De-novo Motif Discovery" from the main home page, the user can start a new job (see Figure 2.6.A).



De-novo motif discovery enables you to predict new binding site models based on your personal data set...

Use a higher order model or select one from the database and find its positions within your custom sequence file...

For over 450 data sets from human, we provide higher order models. Browse, download or directly use entries for your analysis...

De-novo Motif Discovery

Job Name: JobExampleName

Input Sequences: Browse... ExampleData.fasta

Example Data

Reverse Complement: ☑

Advanced Options

Model Settings:

Model Order: 4

Extend left: 10

Extend right: 10

Motif Initialization:

Motif Initialization: PEnGmotif

Motif InitFile:

BaMMmotif Results for Job ExampleJob

Your job is currently running....

Your Job status: Running BaMM

You can view the current calculations in the window below.

If you have submitted this job while being logged in, you will receive an email when it is finished.

```
 ************************

*********************
*   Initial Motif   *
*********************
______
|     |
|  EM |
|_____|

Status: Output directory does not exist. New directory is created automatically.
1 iteration:    para_diff = 2814.7,     log likelihood = 20009.9
2 iteration:    para_diff = 555.404,    log likelihood = 35343.6
3 iteration:    para_diff = 93.0152,    log likelihood = 35620.4
4 iteration:    para_diff = 20.0638,    log likelihood = 35635.9
5 iteration:    para_diff = 8.74858,    log likelihood = 35638.7
6 iteration:    para_diff = 3.13124,    log likelihood = 35639.6
7 iteration:    para_diff = 1.23876,    log likelihood = 35639.9
8 iteration:    para_diff = 0.682595,   log likelihood = 35640
9 iteration:    para_diff = 0.476482,   log likelihood = 35640
```

View Example Result

Figure 2.6: (**A**) Task selection: The main page provides three possible tasks from which the user can choose. (a) *De-novo* Motif Discovery, (b) Motif Occurrence Search or (c) Browse Motif Database. (**B**) Data upload for *de-novo* motif discovery. The user provides input sequences and can optionally set parameters in the advanced option drop-down menu before submitting the job. (**C**) The status of a running job can be monitored in the automatically updating report page.

The minimal input required by the user is an input sequence set in fasta format and the decision whether to consider reverse complement sequences, based on the strand specificity of the protein of interest and the given input sequence information.

The 'Advanced Options' drop-down menu contains suggested default options for BaMM!motif, which can be modified (see Figure 2.6.B).

The user can specify the order of the optimized Bayesian Markov model. A fourth order model learns the frequencies of 5-mer nucleotides to model the correlations between nearby positions; i.e., a first-order model consists of dinucleotide dependencies, while a zeroth order model represents a standard PWM without interdependencies. The algorithm is robust against overfitting so that the model order can be set high even for low depth input data. However, a higher order model contains more parameters, which takes longer to optimize. The default model order is set to four as a compromise for speed and precision.

For initialization, BaMM!motif requires a start motif. This motif can either be uploaded by the user or generated automatically using PEnGmotif (*https://github.com/soedinglab/PEnG-motif*). PEnGmotif, which is developed by Markus Meier, is a fast seeding and merging algorithm to obtain overrepresented k-mers from the input sequence set.

BaMM!motif initializes with a PWM representation of fixed length. The user can extend the model length to each side of the initial PWM by 0-100 bases. This gives the possibility to investigate flanking regions of PWMs that contain higher order information which cannot be captured in a PWM. By default, the initial PWM extends by ten bases to each side.

The user can upload a background sequence set from which the statistical background model is learned. The background model describes how nucleotides and k-mers are distributed on 'normal' sequences. BaMM!motif will learn motifs which are enriched in the input sequence set in comparison to the expectation derived from the background model. When no background sequence set is provided, a background model is learned based on a second order homogeneous Markov model from the input sequences. The higher the background model order, the higher is the biological content it describes. Short motifs may already be described by the background model if the background model order is set too high, while long, complex motifs are better modeled using a higher-order homogeneous Markov model as the background model.

BaMM!motif models are optimized using an expectation maximization (EM) approach. The user can deselect this optimization for obtaining PEnGmotif's initialization PWM. EM specific parameters such as the convergence criteria for parameter differences and the maximum number of optimization iterations can be adjusted. The EM q-value defines the percentage of input sequences expected to obtain a true TF binding site. In ChIP-seq experiments, we expect that roughly 90% of the obtained sequences contain a real binding site. Therefore the default value is set to 0.9.

The user can further adjust the options of the false discovery rate estimation (FDR), a motif performance measure implemented by Wanwan Ge. Here a 'mFold' larger background sequence set as the input sequence set is sampled based on a background model of the selected FDR sampling order. BaMM!motif executes FDR estimation in a cross-validated fashion, where the user can select the number of cross-validations with 'cvFold'. On default, BaMM!motif performs false discovery rate estimation on a 100-times larger sampled background sequence set with a 5-fold cross-validation.

Lastly, the optimized model will be compared against the BaMM!motif database (see Section 2.7). The result page reports all database matches that score better than the selected P-value cut off (more about motif-motif comparison in Section 2.5).

After uploading all necessary files and setting parameters, the job is ready for submission to the BaMM!motif server. The user obtains the job id of a successfully submitted job, which can be used to obtain its status and results.

A submitted job and its progress can be viewed in an embedded command line window which refreshes automatically (see Figure 2.6.C). Once the job has finished, the user will be redirected to the resulting output page.

2.4 *De-novo* motif discovery results

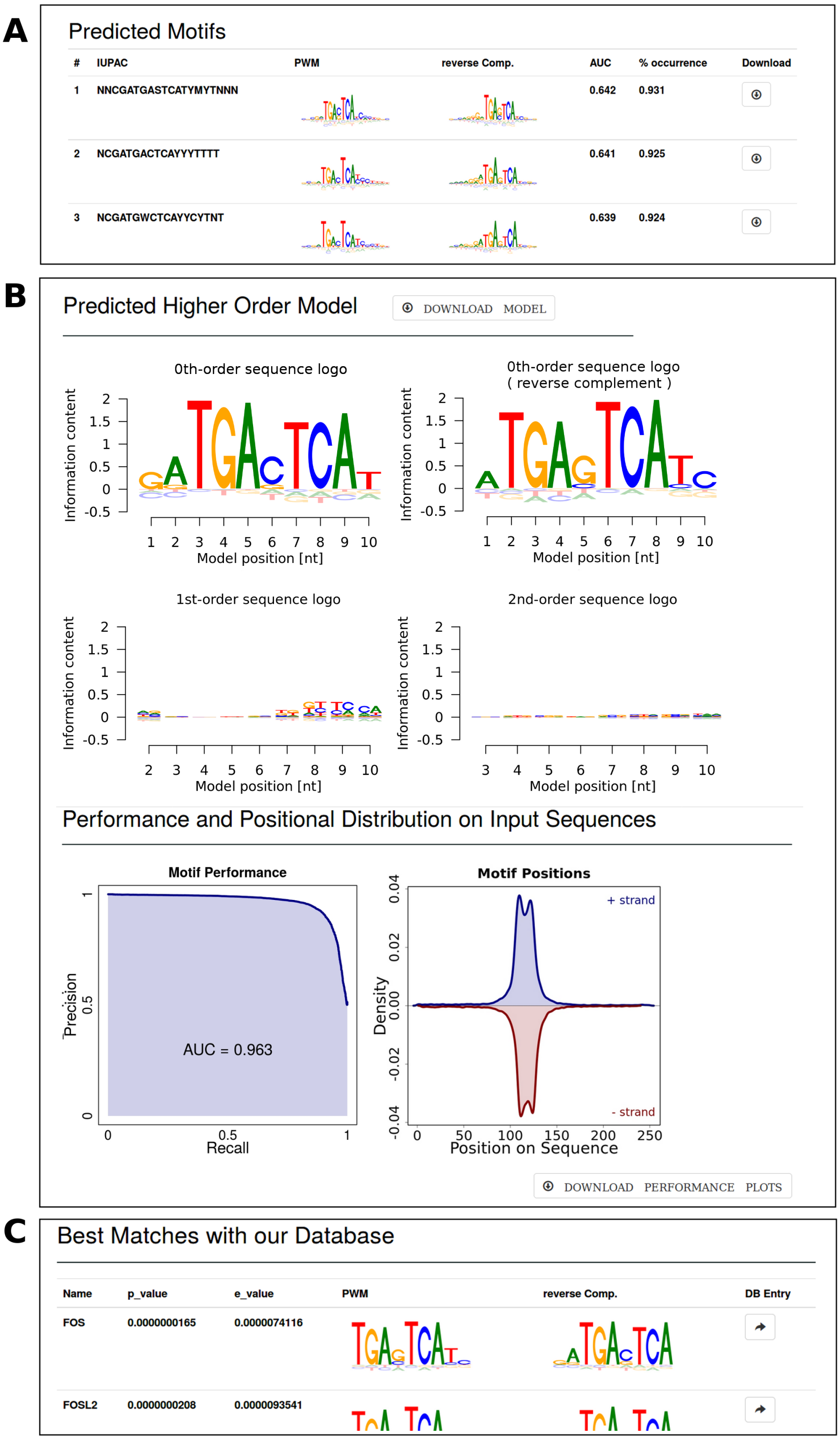
The result page consists of three blocks. First, the settings and input files for the particular BaMM!motif run are summarized. Next, a compact list of all found motifs gives an overview of the results. Third, each predicted motif is explained in a detailed result box.

The summary list shows a compact overview over all found motifs ordered by their E-value (see Figure 2.7.A). Each motif is listed with its IUPAC string and web logos of its zeroth order contribution and its reverse complement. The estimated area under the precision recall curve (AUC) value and the percentile of occurrence in the input sequence set give insights into the motif performance. A download button provides access to all motif-related data.

By clicking on one motif from the list, the page will jump to the appropriate motif detail box (see Figure 2.7.B). Sequence logos showing the zeroth order contribution as PWM and its reverse complement visualize the model. Higher order sequence logos show the informational gain upon the lower order to the model for first and second order.

The precision recall curve is based on the cross-validated false discovery rate estimation performed on sampled negative sequences. The AUC value gives an estimate of the motif strength. The motif enrichment in the input sequence set is shown as a positional distribution plot. By default, BaMM!motif runs in reverse complement mode which means that all sequences in the input dataset and their reverse complements are scanned for motif occurrences of the predicted model.

A motif-motif comparison links known TF binding sites from the BaMM!motif database to the *de-novo* predicted motif, based on a similarity score (for details see Section 2.5). Motifs, which have a higher similarity to the new motif than the selected P-value cutoff, are listed at the bottom of each motif result box (see Figure 2.7.C). The P-value and E-value indicate the significance of the comparison. A target name, a web logo (also of the reverse complement) and a link for further information represent the corresponding database entry.



Predicted Motifs

#	IUPAC	PWM	reverse Comp.	AUC	% occurrence	Download
1	NNCGATGASTCATYMYTNNN			0.642	0.931	
2	NCGATGACTCAYYYTTTT			0.641	0.925	
3	NCGATGWCTCAYYCYTNT			0.639	0.924	

Predicted Higher Order Model

Performance and Positional Distribution on Input Sequences

Best Matches with our Database

Name	p_value	e_value	PWM	reverse Comp.	DB Entry
FOS	0.0000000165	0.0000074116			
FOSL2	0.0000000208	0.0000093541			

Figure 2.7: (**A**) The BaMM!motif result page contains a summary list of all predicted motifs with IUPAC string, zeroth order web logo and its reverse complement, the AUC value as well as the percentile of occurrence in the input sequence set, and a download button. Each motif is represented in detail by its web logos for zeroth - second order informational gain and performance plots (**B**) and its motif-motif comparison results based on the BaMM!motif database (**C**).

2.5 Motif-Motif comparison

A newly identified motif often raises the question whether it resembles a previously identified TF binding site. To address this problem, I have developed a motif-motif comparison tool which matches a query motif with the BaMM!motif database (for details on the database see Section 2.7). The tool ranks all motifs from the database by the similarity of their zeroth order model components to the query motif and computes P-values and E-values.

2.5.1 Variable motif length and shift adjustment

The similarity between zeroth order motifs is the maximum similarity score $s(\cdot,\cdot)$ evaluated in the overlapping regions when the two motifs of length l and l' are shifted by d with respect to each other.

The indices defining the overlap region in the two PWMs are:

$$
\begin{aligned}
&j_1 = \max\{0, d\}, j_2 = \min\{l-1, l'-1+d\}\\
&\text{and}\\
&j'_1 = \max\{0, -d\}, j'_2 = \min\{l'-1, l-1-d\}
\end{aligned}
$$

The underhangs, which come from the offset, are padded with the background model of the query motif. The padding is particularly important for cases where a core motif has been extended by low information positions to either side in different ways between the query motif and the database motif. If the extensions have low information, their comparison to the background will not add to the similarity score. If the extensions contain more information, their comparison to the background model will negatively contribute to the overall similarity score.

Without padding, the scoring tends to favor partly overlapping motifs that only match in their flanking regions. This would lead to high similarity scores due to similar low informative contexts. The minimum number of overlapping nucleotides is set to four. Figure 2.8 shows an example where two motifs contain the same core motif $MOTIF$ but are extended to different sites by low information positions X and Y. The padding will fill up the underhangs with the background model B.

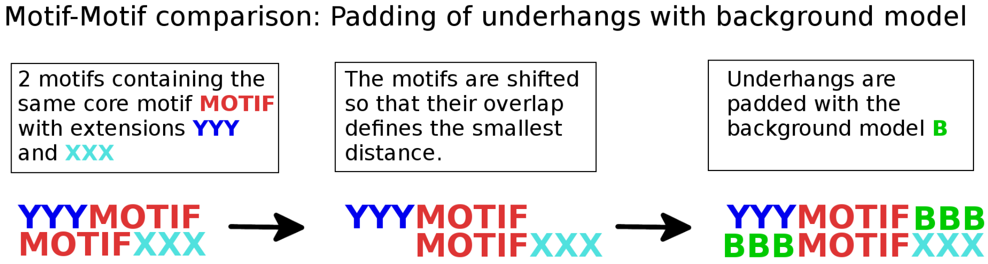

Motif-Motif comparison: Padding of underhangs with background model
2 motifs containing the same core motif MOTIF with extensions YYY and XXX
The motifs are shifted so that their overlap defines the smallest distance.
Underhangs are padded with the background model B
YYYMOTIF
MOTIFXXX
YYYMOTIF
MOTIFXXX
YYYMOTIFBBB
BBBMOTIFXXX

Figure 2.8: Two motifs contain the same core motif $MOTIF$ but are extended to different sites by lower informative positions X and Y. Padding will fill up the underhangs with the background model B in order to provide a complete length comparison without neglecting overhanging motif parts.

2.5.2 Local randomization of motifs as control cases

Control cases are generated as randomized versions of the original query motif to validate motif-motif comparison scores. Here a control case is a randomly permuted original motif where the neighboring positions are kept in proximity, by a localized shuffling of the motif positions. Each position j of the original motif obtains a randomly assigned number $R_j = -j + 2Z$, with a normally distributed variable $Z \sim \mathrm{N}(0,1)$. The positions j are then sorted in ascending order of R_j.

Each column will be replacing its probabilities for A by T and vice versa with $p_{switch} = 0.5$, to assure that randomized motifs differ enough from the original motif. The same applies to the substitution of C by G. Thus, the AT and CG content is preserved, but a higher shuffling of the original motif is achieved. This approach is especially important for motifs with many similar columns, (i.e., A-rich motifs) where shuffling the positions will not have a high impact. Figure 2.9 depicts a motif for the TF JunD and a randomized version of the original motif. A randomized motif may not have a higher similarity score than $\frac{1}{2}S_{max}(M_{orig.}, M_{orig.})$ to the original motif. Otherwise, the randomization is invalid and will be repeated until the similarity score criterion is met.

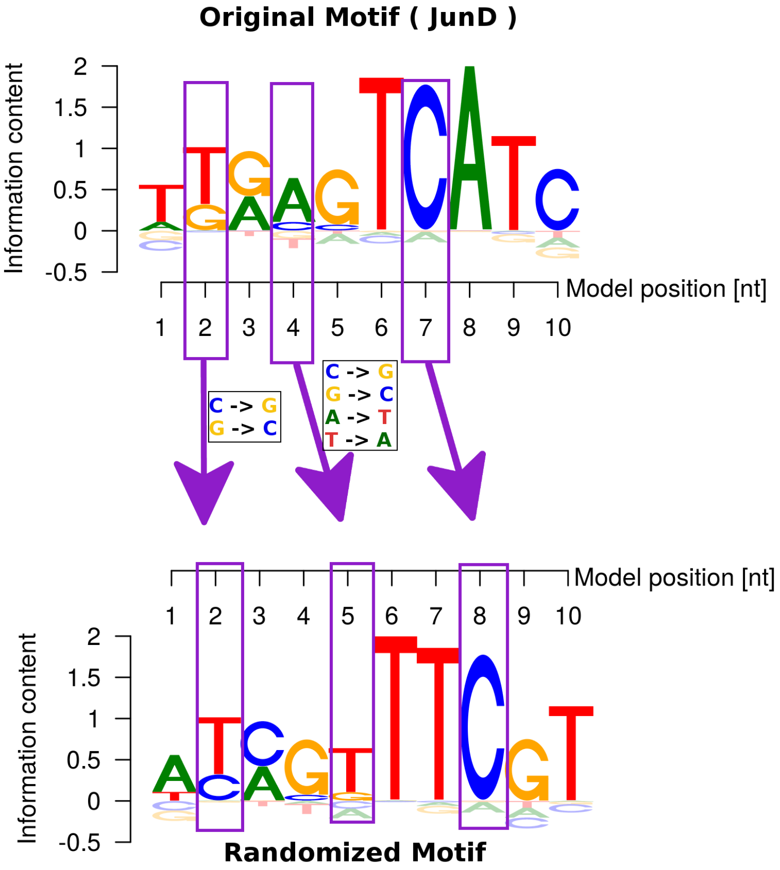


Figure 2.9: Control cases for similarity score calculations are obtained by locally randomizing the positions j in the original motif. For this example, motif column #2 keeps its position, but probabilities for C and G are switched. Motif column #4 moves to position #5 and changes both, A with T and C with G probabilities. The motif column #7 moves to position #8 without altering nucleotide probabilities.

2.5.3 Scoring motif-motif similarity

Let's assuming a score $s(\cdot,\cdot)$ for motifs of same length W exists. The score between models M_1 and M_2 of different lengths is defined as the maximum over all possible alignments of M_1 and M_2 where overhangs have been padded with a zeroth order background distribution, as described above. Two alternative scores S_1 and S_2 have been tested for motif-motif comparisons for motifs of equal length.

The similarity score S_1 between two PWMs is computed as described in Equation 2.1:

$$S_1(M_1,M_2) = \sum_{j=j_1}^{j_2} \left(\alpha \left[d(M_{1j},M_1^{\mathrm{bg}}) + d(M_{2j-d},M_1^{\mathrm{bg}})\right] - d(M_{1j},M_{2j-d})\right) \quad (2.1)$$

Here, M_1^{bg} is a PWM that contains the zeroth order background model for query motif M_1 in each column. As distance $d(M_{1j},M_{1j'})$ between two PWM columns M_{1j} and $M_{2j'}$ the sum of Jenssen Shannon divergence is used. The Jenssen Shannon divergence is defined as the arithmetic average over the relative entropies of each column with their average distribution $\bar{M}_{1ja}$ where $a \in [A,C,G,T]$. (see Equation 2.2).

$$\bar{M}_{1ja} := \frac{M_{1ja} + M_{2j'a}}{2} \quad (2.2)$$

Hence the distance between PWMs $M1_j$ and $M2_{j'}$ is:

$$\begin{aligned} d(M_{1j},M_{2j'}) &= \frac{1}{2}\left(H(M_1||\bar{M}_{1j}) + H(M_{2j'}||\bar{M}_{1j})\right) \\ &= \frac{1}{2}\sum_{a}^{\mathrm{A,C,G,T}} \left(M_{1ja}\log_2 M_{1ja} + M_{2j'a}\log_2 M_{2j'a} - 2\bar{M}_{1ja}\log_2 \bar{M}_{1ja}\right) \end{aligned} \quad (2.3)$$

The score S_1 depends on the weight of the motif strength to the overall similarity score, α (see in Equation 2.1). The higher the α, the higher the impact of the query motif strength. With a too low α, the score tends to report similarity based on weak information columns, leading to matches of flanking regions. Figure 2.10 shows the similarity score distributions for a motif-motif comparison of self-matches (red) and control cases (blue). 446 ENCODE ChIP-seq datasets from 96 distinct TFs were used as search space, and similarity scores of comparisons for each motif to itself (= self-match; left) and ten randomized versions to the original motif (= control cases; right) are shown.

For α values in the range $0.1-1$ the variance in similarity scores increases for original motifs while decreasing for randomized motif scores. The original motifs represent 96 distinct TFs which have no connection to each other. The randomized motifs all represent nonsense motifs. Thus, the variance of similarity

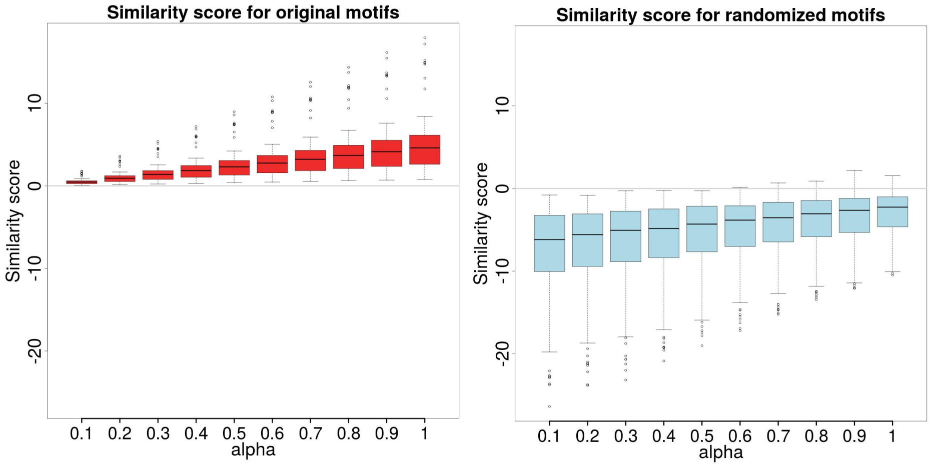


Figure 2.10: For α values ranging from 0.1 to 1, the similarity score distributions for original motifs and randomized versions of the same motifs are depicted on the left and right, respectively. The score distribution within original motifs increases with higher α while the score variance decreases for randomized motifs with higher α.

scores from the original motifs should be maximized while the difference in the similarity scores from the randomized motifs should be minimized. Both criteria are met when selecting $\alpha = 1$. A high value for α increases the importance of motif strengths in the similarity score calculation.

The second score S_2 for equal-length motifs M_1 and M_2 is defined as

$$S_2(M_1, M_2) = \log_2 \sum_{x_1=1}^{4} \dots \sum_{x_W=1}^{4} \frac{p(x_{1:W}|M_1)\, p(x_{1:W}|M_2)}{p(x_{1:W}|\mathrm{bg})} \tag{2.4}$$

where $p(x_{1:i}|M)$ denotes the probability of generating the nucleotide sequence $x_{1:i} = (x_1, \dots, x_i)$ using the first $i \leq W$ positions of a model M and bg is the zeroth order background model. The normalization of the coemission probability with the background model probability ensures that the expectation value of the score of a motif with any other unrelated motif is 0. This is important as it eliminates the length bias. If the expectation value was positive, for example, alignments with higher overlaps would get better scores on average than alignments with shorter overlaps.

The terms in the sum factorize over the positions, and the sum of products can be written as a product of sums:

$$\begin{aligned} S_2(M_1,M_2) &= \log_2 \sum_{x_1=1}^{4} \cdots \sum_{x_W=1}^{4} \prod_{i=1}^{W} \frac{p(x_i|M_1)\,p(x_i|M_2)}{p(x_i|\mathrm{bg})} \\ &= \log_2 \prod_{i=1}^{W} \sum_{a=1}^{4} \frac{p(x_i=a|M_1)\,p(x_i=a|M_2)}{p(x_i=a|\mathrm{bg})} \\ &= \sum_{i=1}^{W} \log_2 \left(\sum_{a=1}^{4} \frac{p(x_i=a|M_1)\,p(x_i=a|M_2)}{p(x_i=a|\mathrm{bg})} \right) \end{aligned} \qquad (2.5)$$

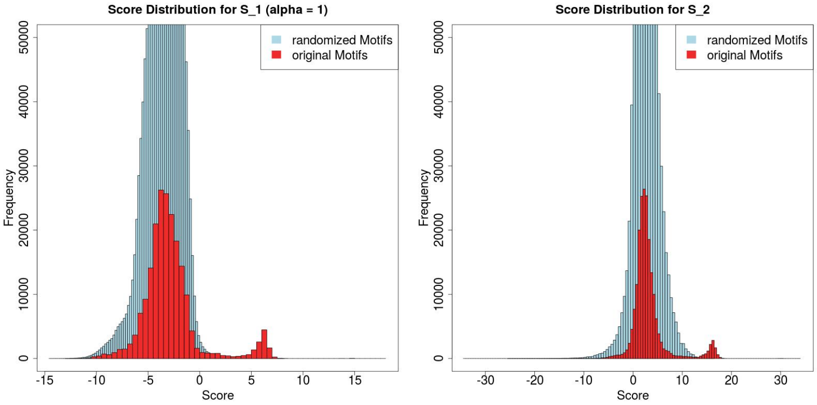


Figure 2.11: Similarity scores between original motifs(red) and randomized motifs as control cases (blue) compared to the BaMM!motif database. The control cases define the shape of original motifs which are not similar to the query motif, while true motif-motif matches describe the small peak on the right.

Figure 2.11 shows the similarity score distributions from a comparison of 446 motifs with each other (red; predicted on ChIP-seq data) and 10 times more randomized motifs as control cases (blue). Most original motifs are from a different TF than the current query motif and describe nonmatching similarity scores. Therefore the score distribution based on randomized query motifs largely overlaps with the score distribution based on original query motifs. Similarity scores derived from comparisons between original motifs describing the same TF score higher (peak on the right). Both scores S_1 and S_2 can distinguish between true motif similarity and false control cases. The score S_2 has the significant advantage that no further parameter optimization is necessary and that the approach can easily be expanded for higher order motif comparisons (see Section 2.10).

2.5.4 Evaluating scores via E-value

P-values and E-values are computed for the scores S_1 and S_2 to evaluate the similarity strength between a query motif and the database match. Background scores are generated by performing $M = 50$ searches with query motifs in which the original positions have been permuted randomly, as described in Section 2.9. Scores from the original query motif and its permutations are denoted as $\{S_1^+,\dots,S_{N^+}^+\}$ and $\{S_1^-,\dots,S_{N^-}^-\}$, respectively. The list of $N^+ + N^-$ positive- and negative-set scores are sorted jointly in descending order. The cumulated number of scores from the negative set up to rank l in this list is denoted by FP_l. For computing the P-value of entry l with score S_l in that list, $S_l^{\text{lower}} = \max\{S_n^- : S_n^- \leq S_l\}$ and $S_l^{\text{higher}} = \min\{S_n^- : S_n^- \geq S_l\}$ are defined as the nearest lower or equally ranked and nearest higher or equally ranked negative scores, respectively. The P-value is interpolated between these two scores with FP_l and $\mathrm{FP}_l + 1$ false positive counts:

$$P\text{-value}(S_l) = \frac{1}{N^-}\left(\mathrm{FP}_l + \frac{S_l^{\text{higher}} - S_l}{S_l^{\text{higher}} - S_l^{\text{lower}} + \varepsilon}\right). \tag{2.6}$$

The ε is very small (10^{-5}) and serves to avoid the fraction being 0 when $S_l = S_l^{\text{higher}} = S_l^{\text{lower}}$.

The estimate becomes inaccurate when only a few or no negatives are higher than S_l, i.e., for low P-values. In that regime, it is better to rely on a parametric fit of the exponentially falling part of the cumulative distribution. The limit is set, beyond which will be used the exponential extrapolation, at $l = n_{\text{top}} = \min\{50, 0.1 \times N^-\}$. This ensures that the extrapolation is applied at most to the top 10% of points, and, if more data are available, only to the top 50. For $l \geq n_{\text{top}}$, a maximum-likelihood estimate based on the top n_{top} data points is used (see Equation 2.7).

$$P\text{-value}(S_l) = \frac{n_{\text{top}}}{N^-}\exp(-(S_l - S_{n_{\text{top}}}^-)/\lambda) \tag{2.7}$$

with

$$\lambda := \frac{1}{n_{\text{top}}}\sum_{l=1}^{n_{\text{top}}}\left(S_l - S_{n_{\text{top}}}^-\right) \tag{2.8}$$

For $S_{n_{\text{top}}}$ this equation yields $P\text{-value}(S_{n_{\text{top}}}) = \frac{n_{\text{top}}}{N^-}$, which is the empirical P-value. The E-values are obtained from the P-value as described in Equation 2.9.

$$E\text{-value} = N^{+} \times P\text{-value}. \tag{2.9}$$

Figure 2.12 shows the performance for the exponential fit of the n_{top} data points for both scores S_1 and S_2 for two example query motifs Fosl and Gata3. The average distance between the exponential extrapolation and the cumulative distribution for all 446 query motifs is shown as distance distribution histogram.

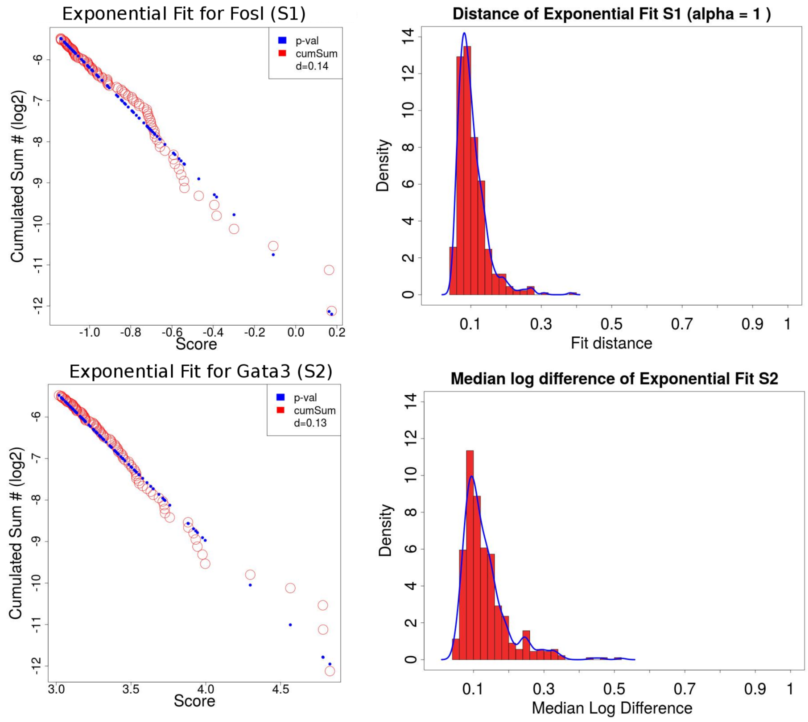


Figure 2.12: n_{top} data points are fitted by an exponential fit (blue: p-value) to the sampled negative distrstibution (red: cumulated sum of false positives). Examples for the TFs Fosl and Gata3 for the similarity scores S_1 and S_2 are shown. The histograms visualize the overall mean log2 deviation between the cumulated sum of false positives to the computed p-values. Extrapolation ensures that P-values stay accurate even if the number of negatives that score higher than S_l are sparse.

2.5.5 Validation of the BaMM-match method

For validating the motif-motif comparison, I computed similarity scores for 446 ENCODE ChIP-seq datasets. Ranked by their P-value, the number of true positives (TPs) found before the first false positive (FP) defines the performance. For a given query motif, I defined all datasets with the same Ensembl target ID as TPs. FPs are scores that come from randomized motifs as explained in Section 2.5.2. Since the amount of TPs varies for TFs, the counts are scaled.

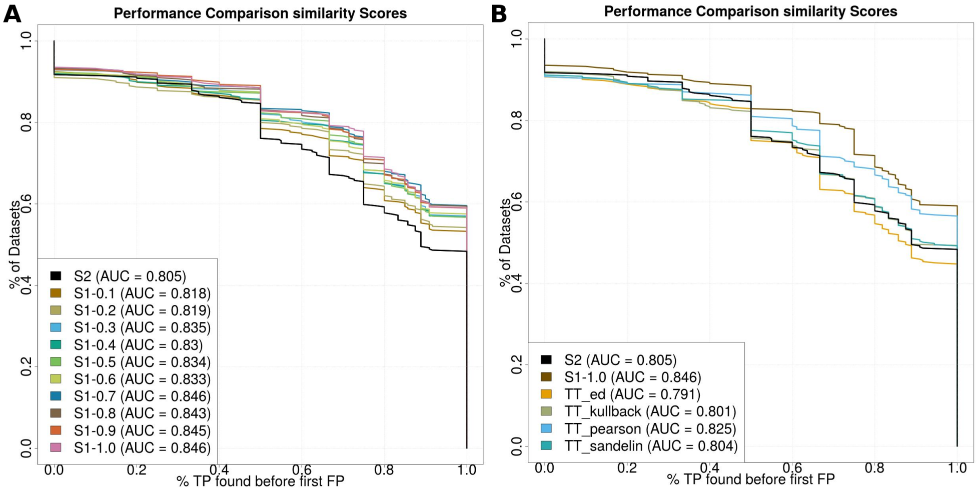


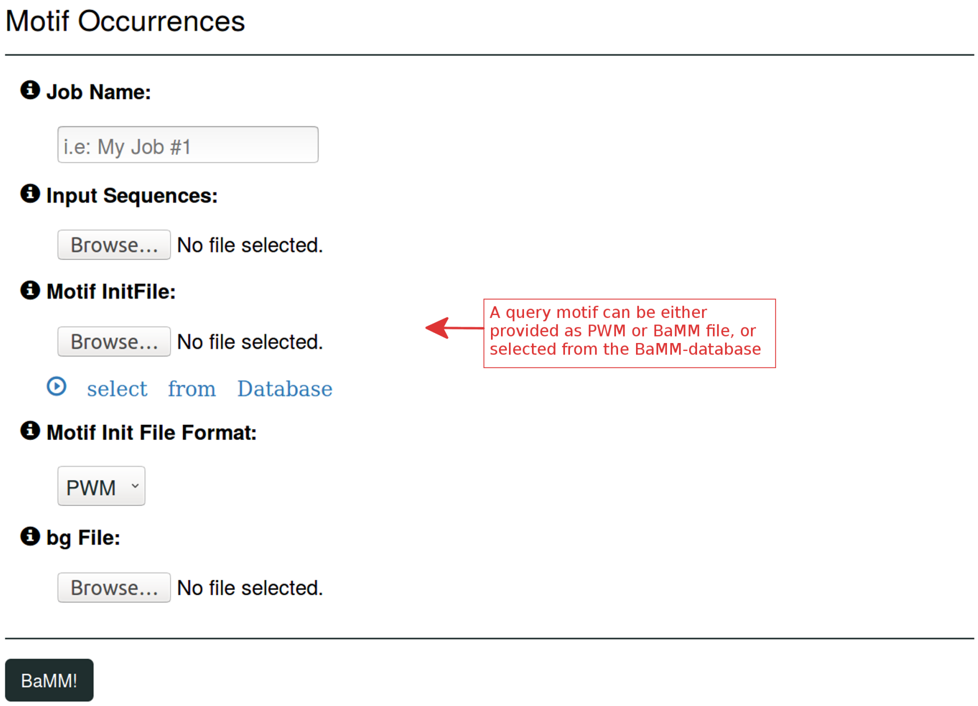
Figure 2.13: The amount of motif-motif comparisons between datasets from the same transcription factor (TP) that rank higher than the best comparison to randomized motifs (first FP) is shown as the fraction of the total dataset of size $N = 446$. S_2 performs better than TOMTOMs P-values based on euclidian distance and equally good as TOMTOM based on kullback or sandelin distance measures.S_1 with $\alpha = 1$ outperforms TOMTOM for any available distance measure.

S_1 reaches a higher AUC than S_2 (see Figure 2.13, left). In general, a higher α value in the S_1 score leads to better performance, which is conform with the analysis of the score distributions earlier.

Comparing the similarity scores with the most common motif-motif comparison tool TOMTOM [50] shows that the S_2 score performs better than TOMTOMs P-values based on euclidian distance and equally good as TOMTOM based on kullback or sandelin distance measures (see Figure 2.13, right). S_1 score outperforms TOMTOM for any available distance measure. Thus, S_2 and S_1 offer a valuable addition to the BaMM!motif web application by providing meaningful motif-motif comparison information. S_2 can be extended to use the full information of higher order motifs. I expect that this extension of the S_2 score will increase its performance.

2.6 Motif enrichment in DNA sequences

I extended the BaMM!motif algorithm with a search tool to scan sequences for motif scores to find occurrences of a known motif in a nucleotide sequence set. The user can select a model from the BaMM!motif database or provide a motif file in PWM or BaMM format (see Figure 2.14). Input sequences that will be scanned for motif occurrences are uploaded by the user in fasta format.



Motif Occurrences

Job Name:

i.e: My Job #1

Input Sequences:

Browse... No file selected.

Motif InitFile:

Browse... No file selected.

○ select from Database

Motif Init File Format:

PWM

bg File:

Browse... No file selected.

A query motif can be either provided as PWM or BaMM file, or selected from the BaMM-database

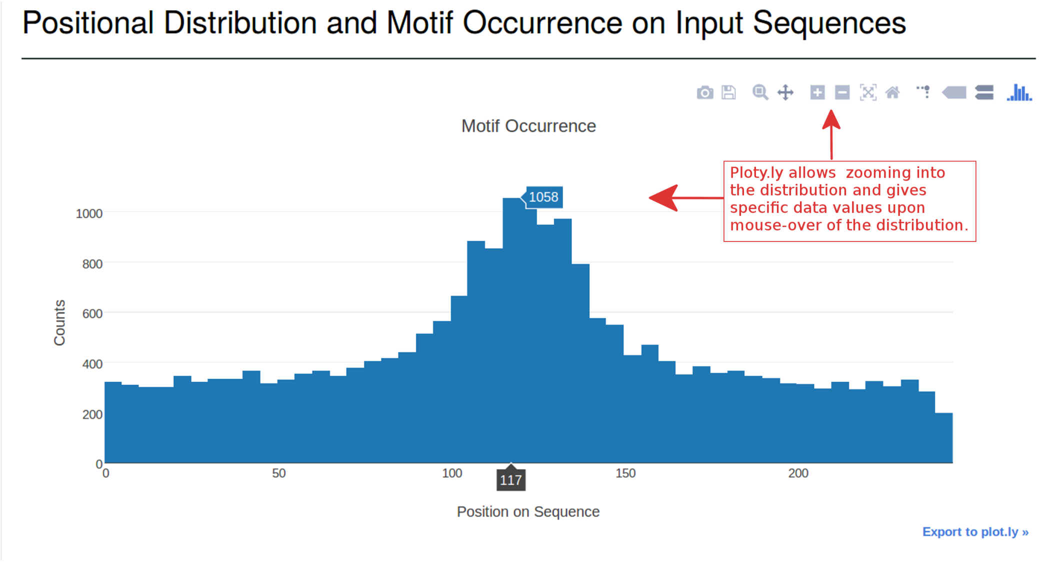
Figure 2.14: The input panel for a motif enrichment analysis is similar to a *de-novo* motif discovery input, which makes it easy to operate. It is possible to scan for a custom motif by uploading a PWM or BaMM file, or selecting a motif from the BaMM!motif database as query motif.

For a given sequence set of size N, MN negative sequences of the same size and length as the positive sequence set are generated using a second order background model and a choice of $M \approx \min\{10^6/N, 1\}$. The background model is obtained from the 3-mer distributions in the positive sequence set. A large M ensures that E-values can be estimated down to order one empirically. The log-odds score of a motif p_{motif} of order K for the sequence $x_1 \ldots x_L$ can be calculated by:

$$S = \sum_{i=1}^{L} log_2 \frac{p_i(x_i|x_{i-K} \ldots x_{i-1})}{p_{bg}(x_{i-B} \ldots x_{i-1})} \tag{2.10}$$

Here, B is the order of the backrgound model p_{bg}. $N^+ = \sum_{n=1}^{N}(L_n - W + 1)$ scores are computed from the positive set and $N^- = M\sum_{n=1}^{N}(L_n - W + 1)$ scores from the negative set. Here, W denotes the motif length and L_n the length of sequence n. Scores from the positive and negative sets are denoted as $\{S_1^+, \ldots, S_{N^+}^+\}$ and $\{S_1^-, \ldots, S_{N^-}^-\}$, respectively and ranked together in descending order. P-values and E-values are computed in the same way as described for motif-motif comparisons (see 2.5.4).

The result page for a motif enrichment analysis is similar to a *de-novo* motif discovery result. It shows the web logos of the provided motif, its reverse complement and an interactive motif distribution plot of all occurrences that have a P-value smaller than a defined cutoff (see Figure 2.15).

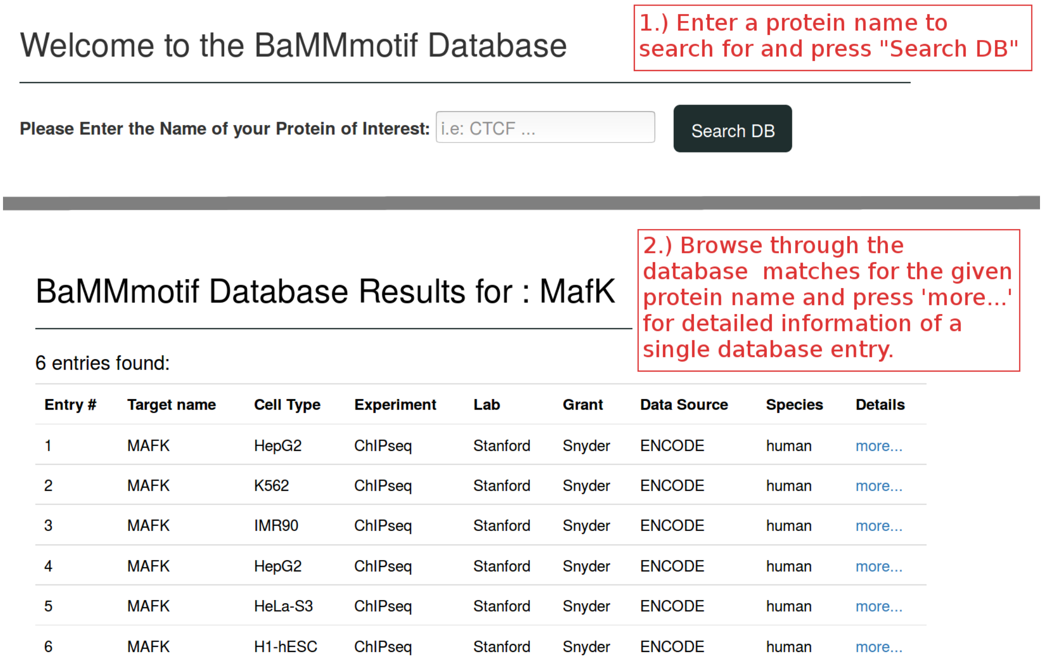


Ploty.ly allows zooming into the distribution and gives specific data values upon mouse-over of the distribution.

Figure 2.15: The motif occurrences found by the motif enrichment tool are visualized with an interactive distribution plot. It shows the positional preferences of a motif according to the sequence space.

2.7 Higher order motif database

The BaMM!motif database contains higher order models of TF binding sites. ChIP-seq datasets published by The ENCODE Project Consortium [32] were restricted to 96 sequence-specific transcription factors characterized by Wang et al. [156] based on sequencing quality (446 data sets).



Welcome to the BaMMmotif Database

1.) Enter a protein name to search for and press "Search DB"

Please Enter the Name of your Protein of Interest: i.e: CTCF ...

BaMMmotif Database Results for : MafK

2.) Browse through the database matches for the given protein name and press 'more...' for detailed information of a single database entry.

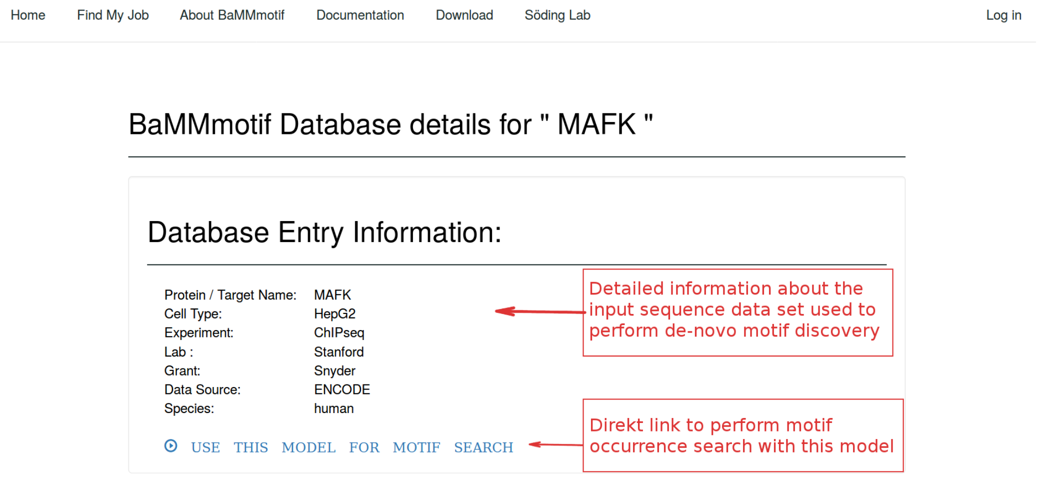
6 entries found:

Entry #	Target name	Cell Type	Experiment	Lab	Grant	Data Source	Species	Details
1	MAFK	HepG2	ChIPseq	Stanford	Snyder	ENCODE	human	more...
2	MAFK	K562	ChIPseq	Stanford	Snyder	ENCODE	human	more...
3	MAFK	IMR90	ChIPseq	Stanford	Snyder	ENCODE	human	more...
4	MAFK	HepG2	ChIPseq	Stanford	Snyder	ENCODE	human	more...
5	MAFK	HeLa-S3	ChIPseq	Stanford	Snyder	ENCODE	human	more...
6	MAFK	H1-hESC	ChIPseq	Stanford	Snyder	ENCODE	human	more...

Figure 2.16: The BaMM!motif database is operated by inserting the name of a TF of interest. A filtered summary of all available database entries for a given TF name is listed. The "more..." button in each row redirects to the detailed database entry for the selected dataset.

The ChIP-seq datasets were prepared as follows to obtain the presented predictions. From each dataset the sequence region surrounding the predicted peak summit was cut out as a 251 bp window (125 bp + summit + 125 bp). Each sequence input set was processed with BaMM!motif to predict the best motif of order $k = 4$. Background models are obtained by the background distribution of 3-mers from the positive sequence sets. Other BaMM!motif parameters were set to the suggested default values.

The user can search for database entries by typing the TF name into the search panel (see Figure 2.16, top). All entries that contain the selected TF name are listed with web logo and additional information on the ENCODE input data, as well as a link to the detailed database entry page (see Figure 2.16, bottom).



BaMMmotif Database details for " MAFK "

Database Entry Information:

Protein / Target Name:	MAFK
Cell Type:	HepG2
Experiment:	ChIPseq
Lab :	Stanford
Grant:	Snyder
Data Source:	ENCODE
Species:	human

⦿ USE THIS MODEL FOR MOTIF SEARCH

Detailed information about the input sequence data set used to perform de-novo motif discovery

Direkt link to perform motif occurrence search with this model

Figure 2.17: An exemplary database entry for the TF MafK predicted from an ENCODE ChIP-seq dataset. The top part of the database entry page shows a summary of the input sequence information and provides a shortcut to use the described model in a further motif enrichment analysis.

The database entry page has a similar outline as the result page from *de-novo* motif discovery (see Section 2.4) and motif enrichment analysis (see Section 2.6), since they also show predicted higher order motif models.

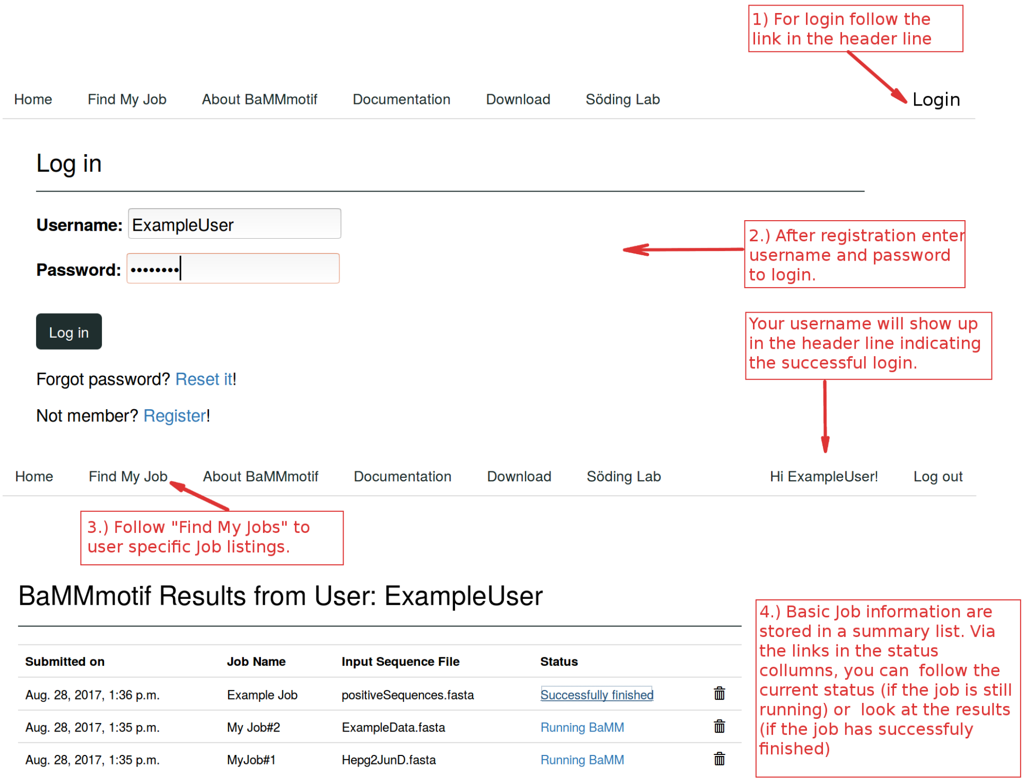
In addition to the presentation in web logos, motif performance and positional distribution on the input sequence set, the database entry page also contain summary information about the input sequence set from which the model was predicted (see Figure 2.17).

The BaMM!motif settings that were used for the prediction are listed at the bottom. All database entries can be utilized as input for a motif enrichment analysis to find specific binding sites on a custom input sequence set.

2.8 Additional features

Login

Users can operate all tools provided on the BaMM!motif web application without registration. However, it is possible to create a user account which gives access to a private user page. This page lists all submitted jobs for easy monitoring of multiple jobs and results (see Figure 2.18). Since the storage capacity of the BaMM!motif server is limited, data upload sizes are restricted, and older jobs will be deleted on a regular basis. As a registered user, the size of uploaded files increases (from 2.5*Mb* to 250*Mb*) and the lifetime of submitted jobs is extended (from 4 weeks to 2 years). Further, the user will be informed about a soon "expiring" job via email, as a reminder to download the data if not done so yet.



1) For login follow the link in the header line

Log in

Username: ExampleUser

Password: ••••••••

2.) After registration enter username and password to login.

Your username will show up in the header line indicating the successful login.

Forgot password? Reset it!

Not member? Register!

3.) Follow "Find My Jobs" to user specific Job listings.

BaMMmotif Results from User: ExampleUser

Submitted on	Job Name	Input Sequence File	Status	
Aug. 28, 2017, 1:36 p.m.	Example Job	positiveSequences.fasta	Successfully finished	🗑
Aug. 28, 2017, 1:35 p.m.	My Job#2	ExampleData.fasta	Running BaMM	🗑
Aug. 28, 2017, 1:35 p.m.	MyJob#1	Hepg2JunD.fasta	Running BaMM	🗑

4.) Basic Job information are stored in a summary list. Via the links in the status collumns, you can follow the current status (if the job is still running) or look at the results (if the job has successfuly finished)

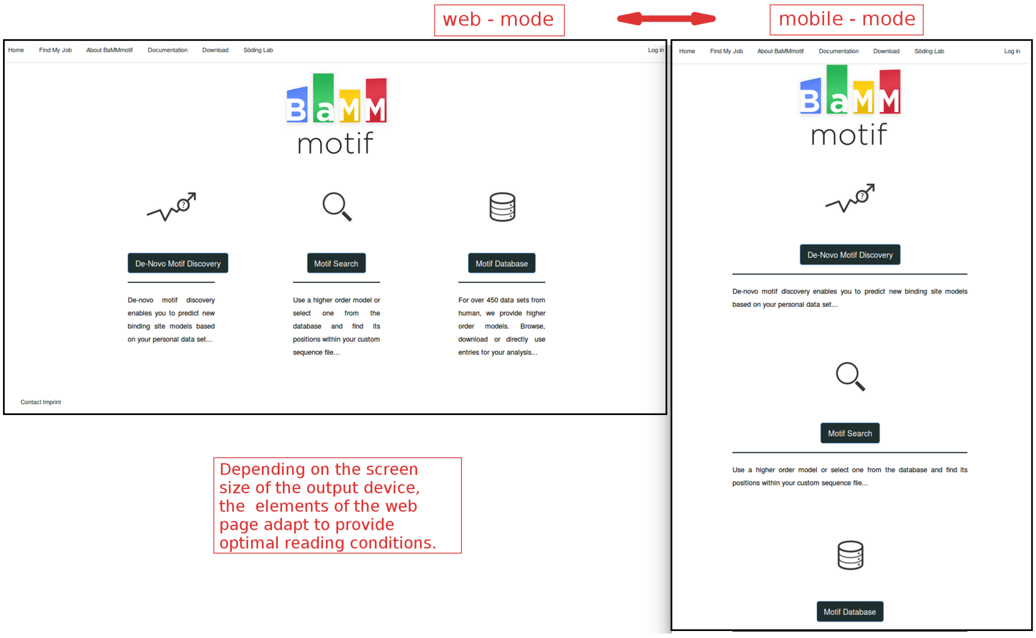
Figure 2.18: Creating a user account on BaMM!motif is possible. When logged in, one can track all user jobs in a job list, delete them and access their status and result page without storing the job ID.

Documentation

An example input sequence set and pre-computed results allow the user to get a quick overview of the server's usage and output. Mouse-over explanations for all input parameters as well as a general documentation page with detailed information help the user to understand parameters and interpret results.

Responsive web design

Responsive web design (RWD) allows the user to view web pages designed for desktop platforms on other output sources such as tablets and smartphones in a screen-size appropriate fashion. An RWD designed web page adapts element sizes, and layouts to the user device while maintaining content, design, and performance [125]. The BaMM!motif web application works with a fluid grid concept from the bootstrap library to support a variety of output devices (see Figure 2.19).



web - mode

mobile - mode

De-novo motif discovery enables you to predict new binding site models based on your personal data set...

Use a higher order model or select one from the database and find its positions within your custom sequence file...

For over 450 data sets from human, we provide higher order models. Browse, download or directly use entries for your analysis...

Depending on the screen size of the output device, the elements of the web page adapt to provide optimal reading conditions.

Figure 2.19: The BaMM!motif web pages are constructed with responsive elements which adapt their spacing and orientation based on the output device's screen size. This leads to optimal readability even for small displays like mobile devices.

2.9 Conclusion

The BaMM!motif web application combines sophisticated higher-order modeling of TF binding sites with valuable downstream analyses such as motif enrichments and motif-motif comparisons. The BaMM!motif database contains predicted motif models which have been shown to outperform PWMs in informational content and precision [132]. The server has a clean interface with well-structured input and result pages. It is unique in offering higher order *de-novo* motif discovery and analysis tools while providing a database of higher order motif models.

2.10 Outlook

The BaMM!motif web application will be extended in several ways. BaMM!motif is currently specialized for using fasta formatted sequences as obtained from ChIP-seq experiments. However, HT-SELEX and PBM data add another level of information by providing intensities for each single sequence. The BaMM!motif team aims to extend the algorithm for processing such data and including the given sequence weights. Further, the EM algorithm has an order specific weight for the interpolated pseudo counts, which is currently set to a fixed value. Experience has shown that this factor is not optimal for all datasets, penalizing higher order contributions too strongly in some cases. Thus, a new approach using collapsed Gibbs sampling for optimizing the pseudo count weights for each dataset individually is currently investigated and developed. The motif-motif comparison currently scores motif similarity between zeroth order models. The $S2$ score can be extended to use higher order information for scoring motif-motif similarity. We expect this extended score to perform better than scores based on zeroth order information only. The motif-motif comparison tool is only available as an add-on to BaMM!motif's de-novo motif discovery at the moment. The tool will be extracted and made available as a standalone tool on the web page. The BaMM!motif database will be expanded to other publicly available data to provide a broad range to higher order motif models.

Chapter 3

Influence of linker Histones on histone modifications

3.1 Introduction

The packaging of DNA into chromatin establishes an essential control mechanism of gene expression in eukaryotes. Nucleosome core particles are the basic building blocks of chromatin, with 147 base pairs (bps) of double-stranded DNA wrapped around an octameric protein complex made up of two copies of the four core histones H2A, H2B, H3, and H4. Linker DNA of variable length connects individual nucleosome core particles and accommodates binding of linker histones. Nucleosome core particles with additional linker DNA establish the repeating unit of chromatin, termed nucleosomes [91].

Linker histone protein family H1 has the most variable histones in sequence across all species. H1 contains a central globular domain, a long C- and a short N-terminal tail [118]. Nuclease digestion and DNA footprinting experiments showed that the H1 core globular domain localizes close to the nucleosome dyad, where its binding protects approximately 15-30 bps of DNA [10].

All histones are post-translationally modified (PTM), dynamically and reversibly. Histone modifications act as transcriptional activators or repressors, influence chromosome packaging, DNA damage, and repair. The N-terminal side chains of histones can be acetylated, methylated, phosphorylated or ubiquitinated. Acetylation is done by the enzymatic addition of an acetyl group (COCH_3), methylation by transferring one to three methyl groups to arginine or lysine residues.

Different PTMs have been found to mark chromatin states as transcriptionally active (i.e. H3K4me1, H3K4me3, H3K9ac, H3K27ac, H3K36me3) or repressive (i.e. H3K9me3, H3K27me3). This constitutes a major regulatory layer of chromatin function. Indeed, different histone PTMs correlate with transcriptionally active or repressive chromatin states [12].

The selective incorporation of linker histones represents another mechanism to alter chromatin states on a structural and functional level. In comparison to PTMs, the binding of linker histones to nucleosomes mostly correlates with repressive chromatin states and is thought to impede transcription activity [24, 54, 80].

From a structural point of view, core histones are divided into unstructured tail regions that flank folded core domains. Since the tails stick out from the nucleosomes and are accessible to modifying enzymes [90], most PTMs reside in these regions. Out of the four core histones, the N terminus of H3 features the most extended tail sequence (35 residues) with more than 30 known modification events [12, 29, 151]. Both H3 tails originate from nucleosomes at structurally equivalent positions close to the entry and exit sites of DNA [90].

Compared to the minimal 147 bps nucleosome core particle, efficient linker histone binding requires additional 2 × 20 bp of DNA [162]. This places linker histones and H3 tail residues in the same spatial context. Experimental evidence suggests that a functional relationship between linker histone binding, nucleosomal H3 tail architecture [72], and H3 tail PTMs [58] exists.

The sequences of linker histones are more diverse than those of core histones, but they all feature a central, conserved globular domain that is essential for nucleosome binding [54]. In humans, 11 genes encode linker histones, with H1.1 to H1.5 being the most abundant subtypes in somatic tissues [54]. Linker histone variants may be positioned differently in nucleosomal contexts, with their globular domains being either at the nucleosomal dyad axis and in contact with both stretches of linker DNA or in asymmetric configurations with contacts to only one of the linker DNA elements [103, 145, 167, 168].

Irrespective of core domain binding, linker histones contain extended N- and

C-terminal segments of variable lengths that are rich in basic amino acids and structurally disordered [54]. In particular, the C-terminal linker histone domain (CTD) contributes to high-affinity chromatin binding *in-vivo* [57] and induces altered linker DNA conformations *in-vitro* [15, 43, 52, 103, 145].

Given the spatial proximity of histone H3 tails and linker histones within nucleosomes, I set out to obtain insights into the PTM behavior of H3 tail residues in different nucleosomal contexts.

I show that linker histone incorporation greatly lowers the overall modification efficiencies of individual H3 sites within nucleosomes on a genome-wide level. These results establish that the presence of linker histones constitutes a basal control layer of H3 modifiability and thereby the functional states of chromatin.

3.2 Material and methods

For investigating the genome-wide influence of linker histone H1 on histone H3 PTMs, I processed 15 individual ChIP-seq experiments from four laboratories containing eight different H3 modifications, two H1 linker histone variants, and one H4 control modification.

After preprocessing and data normalization, I applied a batch-effect correction to eliminate lab-related correlations. A correlation analysis of the log2 IP signal over input ratios was performed on the data sets.

General H3 tail modifications should be reduced in chromosomal regions with high linker histone content, based on previous H3 and H1 dynamics analysis [143]. To address this question, I made use of ChIP-seq data from mouse embryonic stem cells, for which linker histone occupancy and histone PTM occurrence have been measured [25, 34, 74, 104].

To minimize fluctuations in the different data sets caused by experimental variations, and to compare individual data quality, I selected GEO database entries that contained at least two annotated histone PTMs analyzed in the same study and laboratory.

I partitioned the mouse genome into 500 bp-sized bins and determined the enrichment of linker histones and PTMs in each bin.

3.2.1 Data sets

Cao et al. [25] established a knock-in system and showed that the N-terminally tagged H1 proteins H1d and H1c in mouse embryonic stem cells (m-ESCs) are functionally interchangeable to their endogenous counterparts *in-vivo*. They provide a high-resolution mapping of these two H1 variants along with H3 PTMs H3K4me3, H3K9me3, and H3K27me3 (Gene Expression Omnibus (GEO) Accession ID GSE46134).

Over four billion bases of sequence from chromatin immunoprecipitated DNA were used to generate genome-wide chromatin-state maps of m-ESCs for H3K4me3, H3K9me3, H3K27me3, H3K36me3, and H4K20me3 by Mikkelsen et al. [104]. The study (GSE12241) provides not only duplicates of Cao's H3 modifications but adds one additional H3 methylations and a control H4 methylation to compare effects to different nucleosome core histones.

Acetylation of Histone H3 data is provided by Karmodiya et al. [74], adding two transcriptional repressive marks H3K9ac and H3K14ac to the analysis (GSE31284).

Creygthon et al. [34] provide the monomethylation of H3 lysine (H3K4me1), a not modified H3 measurement and H3K4me3 as well as H3K27ac (GSE24164). They investigated enhancer activity upon genome-wide binding targets of enhancer-associated H3 modifications.

3.2.2 Data preprocessing

Previously mentioned datasets were downloaded from the Sequence Read Archive (SRA) at NCBI in raw SRA file formats and converted into the fastq format using the sra-toolkit [109].

Fastq is a text-based format that contains not only a nucleotide sequence but also its corresponding quality scores. Fastq format is the standard for storing high-throughput sequencing outputs [30].

Depending on the fastqc quality report in each fastq file, adapter trimming, quality filtering (minimal sequencing quality of 20) and removal of leading bases with abnormal nucleotide distribution was performed using the fastx-toolkit $(http : //hannonlab.cshl.edu/fastx_toolkit/index.html)$. Sequences shorter than 20 bases were removed.

Filtered reads in the processed fastq files were mapped to mouse genome version mm9 with bowtie2 using standard mapping parameters.

Mapped reads for each ChIP-seq and Input-seq library were assigned to non-overlapping bins of size 500 bp covering the entire genome. Reads that overlap two bins were counted as 0.5 reads in each bin.

3.2.3 Normalization

Read counts in each bin were slightly smoothed by mixing them with the two upstream and two downstream neighboring bins using weights 0.1, 0.15, 0.5, 0.15, 0.1. This step was done to account for typical ChIP-seq resolution limitations. Upon inspection in a genome browser, around 98% of the genome in each of the ChIP-seq profiles showed uniformly low background counts. ChIP-counts that fell below the 98% quantile were set to the value of the 98% quantile to avoid computing correlations which are based on background noise.

Due to the lack of input information, the unmodified H3 ChIP-seq signal was used for normalization of the GSE24164 data sets [44]. The log2 of this ratio was used for further analysis.

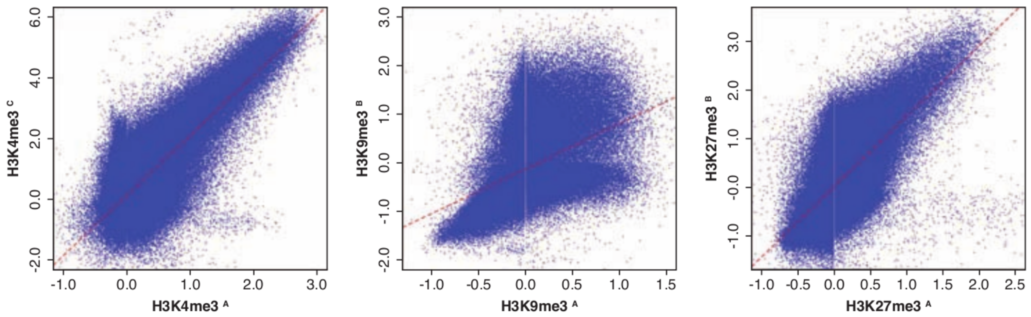


Figure 3.1: 2-Dimensional scatter plots of the indicated PTMs from distinct studies based on log2 ratios of IP over background signals. The superscript denotes the datasets: A = [25]; B = [104]; C =[34]; The red, dashed regression line marks the main trend.

3.2.4 Batch-effect correction

For data sets containing more than four different ChIP-seq datasets, lab related batch effects were corrected. Batch effects are technical artifacts that are not associated with the underlying signal. They correspond to unrelated factors, such as laboratory conditions, reagent lots and laboratory personnel differences. [126].

The mean of all ChIP-seq analysis from one lab was subtracted from each analyzed histone PTM [82]. This method increases the correlation between measurements of the same histone modification from different laboratories while decreasing the correlation between measurements coming from the same lab (see Figure 3.1).

3.2.5 Genome-wide correlation analysis

To analyze in how far H1 linker histone occupancy influences genome-wide histone H3 modifications, the log2 IP over input ratio was calculated and used to define a Pearson correlation coefficient between binned profiles of all histone modifications and linker histones. The Pearson correlation coefficient explains a linear correlation between two variables x and y in the range $[-1, 1]$.

A negative Pearson correlation coefficient corresponds to anti-correlation while a positive value close to 1 illustrates strong correlation. A Pearson correlation coefficient of 0 defines no correlation between variables x and y.

The Pearson correlation coefficient is calculated by dividing the covariance of two variables by the product of their standard deviations.

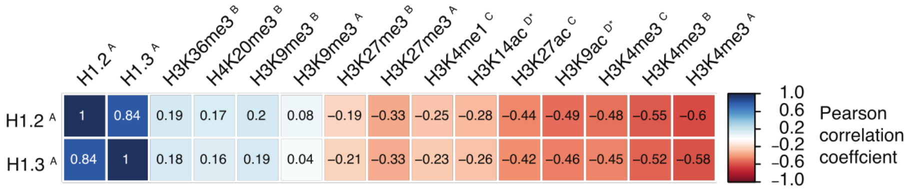
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.1}$$

Here n is the number of binned log2 IP over input ratio values, x_i defines the i-th value for a specific H3 histone modification and y_i the i-th value for a H1 histone. $\bar{x}$ and $\bar{y}$ are the sample means for each dataset.

3.3 Results

H3 tail modifications and linker Histone occupancy segregate on a genome-wide level

I found that the linker histone isoforms H1c and H1d correlate well over the entire genome (see Figure 3.2). Similarly, data on individual histone modifications from different laboratories showed good global overlap.



	H1.2^{A}	H1.3^{A}	H3K36me3^{B}	H4K20me3^{B}	H3K9me3^{B}	H3K9me3^{A}	H3K27me3^{B}	H3K27me3^{A}	H3K4me1^{C}	H3K14ac^{D*}	H3K27ac^{C}	H3K9ac^{D*}	H3K4me3^{C}	H3K4me3^{B}	H3K4me3^{A}
H1.2^{A}	1	0.84	0.19	0.17	0.2	0.08	−0.19	−0.33	−0.25	−0.28	−0.44	−0.49	−0.48	−0.55	−0.6
H1.3^{A}	0.84	1	0.18	0.16	0.19	0.04	−0.21	−0.33	−0.23	−0.26	−0.42	−0.46	−0.45	−0.52	−0.58

Figure 3.2: Correlation coefficients between mouse linker histones H1.2 or H1.3 (mouse H1c and H1d, respectively), eight different H3 modifications, and one H4 modification in mESCs. The superscript denotes the origin of the data set: A, Cao et al.(2013); B, Mikkelsen et al. (2007); C, Creyghton et al.(2010); D*, Karmodiya et al. (2012). The Pearson correlations were computed between log2 ratios of IP over background signals for the entire genome. The data set marked by an asterisk could not be corrected for lab-related batch effects.

The overall distribution of H3 Lys4 tri-methylation reported by three independent studies [25, 34, 104] and the profiles of Lys27 and Lys9 tri-methylation reported by two independent studies [25, 104] were well correlated (see Figure 3.1). From these results, I conclude that data normalization and processing procedures were working satisfactorily.

Both linker histone isoforms display a strong anti-correlation with H3 Lys4 tri-methylation over the entire genome (see Figure 3.3,top), which is evidently preserved on the level of individual gene loci (see Figure 3.3, bottom).

Mono-methylated H3 Lys4, tri-methylated H3 Lys27, and also acetylated H3 Lys9, Lys14, and Lys27 display a similar negative correlation with linker histone occupancy. This demonstrates that activating and repressing H3 modification marks follow the expected trend.

For H3 Lys9 and Lys36 tri-methylation, I detect a weakly positive correlation with linker histone occupancy. This might be explained by the linker histone-mediated, direct recruitment of the Lys9-specific methyltransferase SU(VAR)3-9 [88]. Similar recruitment scenarios might account for the positive correlation of H3 Lys36 trimethylation.

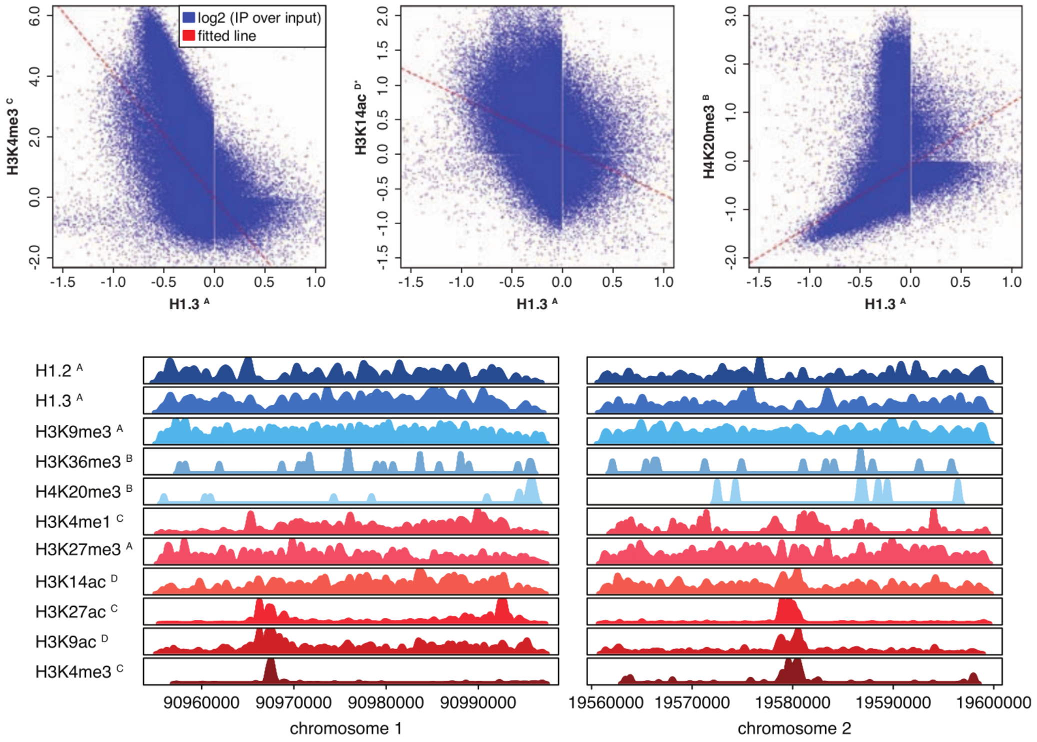


Figure 3.3: Top: 2-Dimensional scatter plots of the indicated PTMs over H1.3 based on log2 ratios of IP over background signals. The superscript denotes the datasets: A = [25]; B = [104]; C =[34]; D* = [74]; The red, dashed regression line marks the main trend. Bottom: Genome browser coverage tracks of library-normalized reads per 500 bp of H1.2, H1.3, and different core histone PTMs along the indicated mouse chromosomes. For histone modifications, which were available in more than one dataset, one representative experiment was selected for display.

Lastly, tri-methylated H4 Lys20 displays a weak correlation with linker histone occupancy, which is in agreement with the H4 modification being less affected by linker histone binding to nucleosomes.

Overall, results from this genome-wide analysis confirm the repressive effects of linker histone occupancy on the general modification status of nucleosomal H3 tail residues.

3.4 Conclusion

This project needed a major data processing pipeline to provide a robust data set for a rather simple correlation analysis. The challenge was not in calculating a genome-wide correlation between H1 linker histone occupancy and H3 PTMs in comparison to H4 PTMs, but more in processing the data correctly. Without quality control from the raw data, normalization for library size and experimental resolution, and batch effect correction to eliminate lab related correlations, a proper result about the connection between linker histones and core histone modifications would have been imprecise. Due to many replicative data, the corrections and filtering steps could be validated. The Pearson correlation analysis then showed the genome-wide correlation and anti-correlation of eight different histone H3 and H4 modifications and linker histone H1 occupancy.

Further details and findings can be found in our MolCell publication *Modulations of DNA Contacts by Linker Histones and Post-translational Modifications Determine the Mobility and Modifiability of Nucleosomal H3 Tails* [143].

Chapter 4

Promoter proximal termination

4.1 Introduction

The Nrd1-Nab3-Sen1 complex directs transcriptional termination of non-coding RNAs. Nrd1 and Nab3 bind to specific RNA sites within the first 1000 base pairs of nascent RNA and initiate promoter proximal termination. Studies have shown that this pathway also terminates individual coding genes, such as Nrd1 itself [3, 4]. In these cases, the Nrd1 pathway induces transcription attenuation.

Here I investigate if Nrd1 induces transcription attenuation genome-wide or only on a distinct group of protein-coding RNAs. I integrate ChIP-seq and RNA-seq data from a Nrd1 anchor-away and a control experiment into a genomic analysis of the yeast species *Saccharomyces Cerevisiae*. Additional PAR-CLIP data is used to define Nrd1 and Nab3 binding sites on the nascent mRNAs. I found that transcriptional attenuation is rare and only applies to 32 genes. Further, the project lead to the discovery of a new type of non-coding RNA, Nrd1-dependent un-terminated transcripts (NUTs), which occur upon Nrd1-depletion from the nucleus.

The yeast genome and non-coding RNAs

The yeast genome is annotated extensively, which enables to map experimental data to the corresponding genes [5, 144]. The yeast species used in this project is the baker's yeast *Saccharomyces Cerevisiae*. It was the first eukaryotic genome to be sequenced in 1996 [35]. Its genome is about 12 million base pairs long and contains roughly 6000 genes. The genome is massively packed with transcribed genes. It also contains non-coding RNAs, which do not encode for a protein. New non-coding RNAs were discovered in yeast. The relevant ones to this project are described below.

Small nuclear and small nucleolar RNAs, or snRNAs and snoRNAs, are short non-coding transcripts [96]. They are known to be terminated by the Nrd1-Nab3-Sen1 pathway and are non-polyadenylated transcripts. Unlike that of other non-coding RNAs, the function of snRNA/snoRNAs is known. snoRNAs are necessary for RNA silencing and telomerase maintenance [96]. In higher eukaryotes, snoRNAs regulate alternative splicing.

Cryptic unstable transcripts (CUTs) were discovered in *Saccharomyces Cerevisiae* in 2005 [163]. CUTs are non-coding transcripts that are unstable and rapidly degraded by the nuclear exosome [165]. In contrast to the degradation of coding transcripts, the CUT degradation is mediated by the Nrd1-Nab3-Sen1-dependent pathway. CUTs derive from unannotated regions widely distributed in the genome [110]. They are often found on the antisense strand of a coding gene emerging from its promoter region. Thus, CUTs may represent by-products of divergent transcription [3]. It has been shown that non-termination of CUTs can reduce gene expression attenuation, transcriptional interference, and alternative start site selection [165].

Xu et al. discovered another type of non-coding RNA, stable unannotated or uncharacterized transcripts (SUTs) [165]. SUTs are longer and more stable than CUTs. SUTs are primarily degraded by cytoplasmic 5'-3' degradation and nonsense-mediated decay (NMD). SUTs may be processed similar to mRNAs [134].

Xrn1-sensitive unstable transcripts (XUTs) were introduced by van Dijk EL et al. in 2011 [152]. These short non-coding transcripts are polyadenylated. XUTs occur mainly antisense to open reading frames and are destabilized by the Xrn1 5'-3' RNA exonuclease. Genes with antisense XUTs are silenced when XUT termination is inhibited.

Termination of Polymerase II transcription

In yeast, two different pathways can terminate Polymerase II (Pol II) transcription [62, 105]. Termination of mRNA transcription requires cleavage and polyadenylation factors. These factors bind via the polyadenylation signal of the nascent mRNA and introduce termination with the exosome pathway [94]. Termination of ncRNA transcription depends on the Nrd1-Nab3-Sen1 pathway [137].

Nrd1 and its role in transcription termination

Nrd1 is an essential protein that interacts with Pol II by binding to the serine-5 phosphorylated form of the Pol II C-terminal domain (CTD). Nrd1 binds to the nascent RNA near the 5'-end and promotes premature transcription termination [26]. Nrd1 binds RNA in a sequence-specific manner via the RNA recognition motif UGUA [33, 137, 153] and interacts with Nab3 and Sen1 [138]. CUTs [4, 146] and SUTs are terminated by the Nrd1 pathway [95].

Nrd1-dependent termination restricts antisense transcription from bidirectional promoters by terminating divergent transcription and initiating their rapid degradation by the exosome [23, 65, 130, 159]. This hypothesis confirmed *in-vivo* cross-linking of Nrd1 and Nab3 to CUTs [161].

Nrd1 binds to many mRNAs [33, 161], and is recruited to mRNA according to chromatin immunoprecipitation experiments [101]. Nrd1-dependent termination to attenuate mRNA transcription has been observed for Nrd1, Hrp1, and Imd2 [3, 139], Ura2, Ura8, and Ade12 [77, 146], and Fks2 [76].

Nrd1 is responsible for recruiting termination factors to the promoter proximal region of some coding-genes and most noncoding regions. So far only the effect of Nrd1 on single genes has been examined. Recently, it was shown that Nrd1 is recruited to the 5'-region of all protein-coding mRNAs. This raises the question whether Nrd1 has a more general function than terminating short non-coding RNAs. Therefore, this project investigated transcriptional changes upon depletion of Nrd1 from the nucleus. By combining ChIP-seq, RNA-seq and PAR-CLIP data, the general function of Nrd1 was investigated.

Anchor-away for essential proteins

Nrd1 is an essential protein and cannot be knocked out without killing the yeast cells. Anchor-away is a technique for investigating the influence of essential proteins by pulling them out from the nucleus into the cytoplasm [55]. The protein, here Nrd1, is still in the cell while being depleted in the nucleus. Because transcription takes place in the nucleus, anchor-away enables investigation of transcription in a Nrd1-depleted condition.

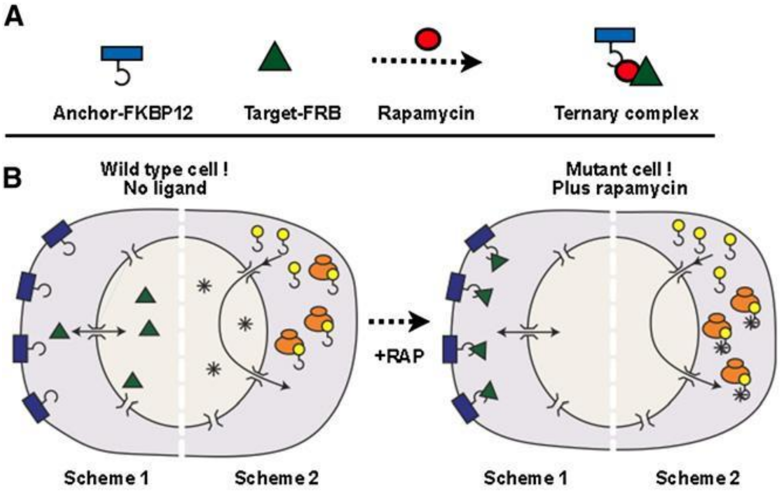


Figure 4.1: The experimental procedure of anchor-away for essential proteins as described in detail by Haruki et al. [55]. A protein of interest, the target, will be pulled out of the nucleus by attaching it to an abundant cytoplasmic protein, the anchor, via gene tagging by rapamycin-dependent heterodimerization.

A protein of interest (target) will be pulled out from the nucleus by attaching it to an abundant cytoplasmic protein (anchor) via gene tagging due to rapamycin-dependent heterodimerization (see Figure 4.1). A protein of the large subunit of the ribosome was selected as the anchor (RPL13A), due to the massive flow of ribosomal proteins during maturation. The target protein Nrd1 is tagged with the human FKBP12 rapamycin-binding domain (FRB). FKB12 is attached to the anchor protein. Adding rapamycin will form the ternary complex with target-FRB and FKBP12-Anchor fusion proteins. This complex formation results in fast depletion of Nrd1 from the nucleus [55].

4.2 Material and methods

4.2.1 Data sets

The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition. RNA-seq is used to quantify expression levels of transcripts, including non-coding and small RNAs [157]. Comparison between wild-type expression and expression after Nrd1-anchor-away reflects the influence of the Nrd1 complex on expression levels.

To find out where Nrd1 and Nab3 bind specifically in the RNA transcripts, Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) [51, 117, 160] experiments were performed. The presence of a binding site also verifies active transcription, since a protein cannot bind to a non-existing RNA.

RNA Pol II profiles are obtained from ChIP-seq measurements. They indicate transcriptional activity. This is important because short transcripts are fast degraded and may therefore not be visible in RNA-seq profiles. From these Pol II profiles, the regions where early termination of transcription is inhibited due to Nrd1-depletion are extracted. This is done by investigating differential profiles from the two conditions, untreated and Nrd1-depleted.

4.2.2 Normalization

To compare the two conditions, I sum up the read counts from the ChIP-seq inferred profiles for each position from both replicates of one condition. One pseudo count is added to prevent singularities.

I calculate a size factor [1] for each condition, based on the open reading frame transcript (ORF-T) regions as described by Xu et al. [165] and divide the read counts by it. This is important to correct for library size and sequencing depth variations. I choose only to use ORF-T regions for the size factor calculation since a significant variation in ncRNA expression is observed, which does not represent overall expression variability.

Because I am interested in the change of Pol II behavior between the two conditions, I calculated differential profiles, as the log2 ratio (see red profile in

Figure 4.2) between Nrd1 depleted (blue profile) and wild-type (green profile) read count pileups, to investigate their variation. The following analysis is based on these differential profiles.

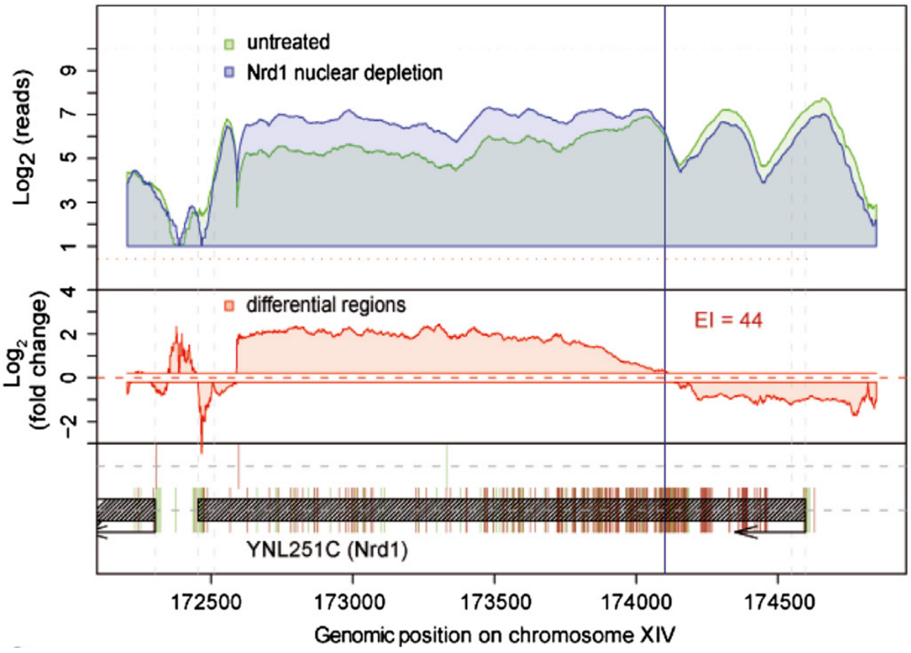


Figure 4.2: Log2 Pol II reads from ChIP-seq around the Nrd1 gene locus before (green) and after nuclear depletion of Nrd1 (blue) and calculated log2 differences in ChIP signal (red). The vertical black line indicates the derived early termination/attenuation site. RNA-binding sites of Nrd1 and Nab3 as determined by PAR-CLIP are shown as green and brown vertical lines at the bottom.

4.2.3 Escape Index

Brannan et al. introduced the Escape Index (EI) in 2012 [20]. It describes how much Pol II escapes the promoter proximal region for productive transcription. Hence, it can be used as a scale for the quantity of premature termination of Pol II transcription. The higher the EI, the lower the influence of premature termination effects. The gene separates into the transcript body (+301 bps to the polyA site) and the promoter-proximal region (-100 bps to +300 bps with respect to the transcription start site(TSS)) to calculate a transcript-specific EI (see Figure 4.3). The

EI is defined as the log2-ratio between the transcript body fold change (TBF) and the promoter proximal fold change (PPF).

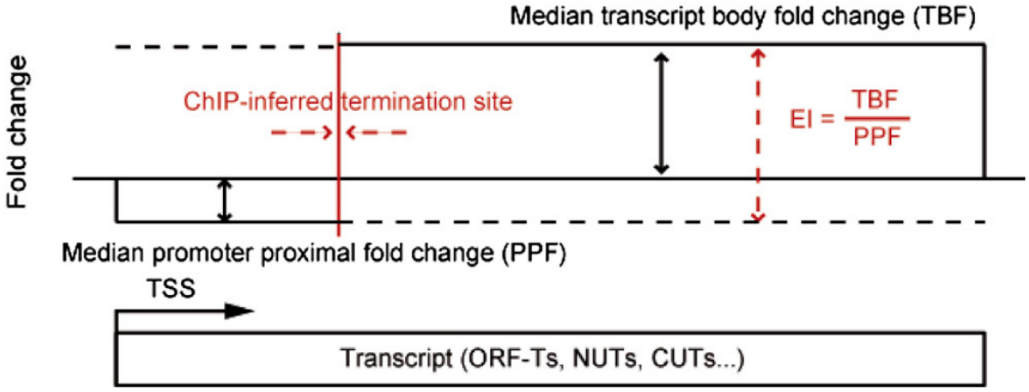

Median transcript body fold change (TBF)
Fold change
ChIP-inferred termination site
EI = TBF / PPF
Median promoter proximal fold change (PPF)
TSS
Transcript (ORF-Ts, NUTs, CUTs...)

Figure 4.3: Scheme illustrating the determination of termination site and EI from ChIP-seq data. EIs were calculated as the median fold-change in the transcrit body (TBF) divided by the median fold-change in the promoter proximal region (PPF) upon nuclear depletion of Nrd1.

The flexible Escape Index

Since it is known that Nrd1 binds in the first 1000 bps of an ORF, the border between the promoter-proximal region and transcript body region was adapted for each transcript according to the Nrd1 binding preference.

The border between the two gene segments (PPF and TBF) is determined by fitting a piecewise constant curve to the differential profile within the first 1000 bps downstream of the transcript TSS or, for shorter transcripts within the first half of the transcript. The fitting was done by using the R-function "tillingArray" from the Bioconductor package "GRanges".

This border is defined as the predicted termination site, where Nrd1-dependent early termination is expected to take place. Escape Indices were subsequently calculated as the ratio of median transcript body fold-change (second segment) and median promoter proximal region fold-change (first segment).

4.2.4 Error-correction and thresholding

EIs are weighted to yield coverage-dependent quantities by the following factor:

$$EI_{weight} = \sqrt{\frac{26}{l_{PPF}}\left(\frac{1}{\sum r_{ND_{PPF}}} + \frac{1}{\sum r_{WT_{PPF}}}\right) + \frac{26}{l_{TBF}}\left(\frac{1}{\sum r_{ND_{TBF}}} + \frac{1}{\sum r_{WT_{TBF}}}\right)} \quad (4.1)$$

Here l is the length of the segment in bp, r is the number of readcounts from the normalized ChIP-seq pileups, ND is the Nrd1 depleted sample and WT the untreated sample (wild-type). The sums over readcounts for a defined region (i.e. $\sum r_{ND_{PPF}}$ sums over readcounts r from an Nrd1 depleted sample ND in the promoter proximal segment PPF) is normalized by the read length of 26.

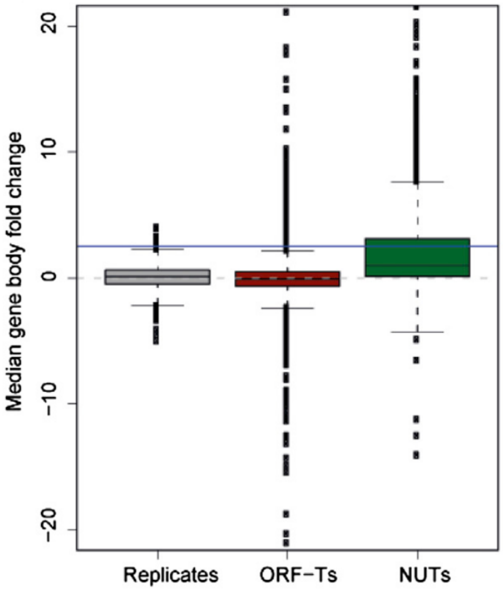


Figure 4.4: Distributions of median changes in Pol II occupancy in ORF-Ts, NUTs, and for a null distribution obtained by using two replicate measurements of Pol II ChIP-seq. The threshold to define attenuated genes is shown as a blue horizontal line.

I calculated EIs on differential profiles of conditional replicates to define a threshold for significantly high EIs. These EIs represent differential expression due to biological fluctuations and are used to estimate a cutoff for the signal. EIs which exceed the 95% quantile from this noise level distribution are taken as identifiers for Nrd1 dependent promoter-proximal termination (see Figure 4.4).

Further, the distribution of the transcript gene body fold change has been used to further restrict Nrd1 dependent transcripts to those who show a 2-fold or more change in this region (see Figure 4.5).

Since the introduced EI is calculated based on fold-changes in the form of differential profiles, it has to be interpreted differently than the original EI. Here a high value of the flexible EI infers strong influence of Nrd1-dependent early termination.

4.3 Results

Based on the EI and further input from the RNA-seq and PAR-CLIP data analysis I was able to define which genes are terminated by the Nrd1-Nab3-Sen1 pathway. This chapter summarizes the criteria used to define the set of Nrd1-dependent promoter proximal terminated genes and visualizes the results.

Transcriptional attenuation of 32 genes

A gene needs to fulfill three criteria to be defined as terminated by Nrd1. First, the EI based on the ChIP-seq data has to be above the defined threshold (see Section 4.2.4). This means that the transcript has more Pol II binding at the beginning of its ORF-T than at the transcript body. The EI is the indicator that transcription is terminated and Pol II is released early from the transcription process.

Second, the RNA-seq profile has to show differential expression for the two conditions untreated and Nrd1-depleted. With this criteria, I check for normal expression of genes in the untreated and a disrupted transcriptional behavior upon Nrd1-depletion.

Third, a gene has to have Nrd1 and Nab3 binding-sites in proximity to the predicted termination site, based on PAR-CLIP data. This asures that differential expression and transcriptional early termination are linked to the Nrd1-Nab3-Sen1 pathway.

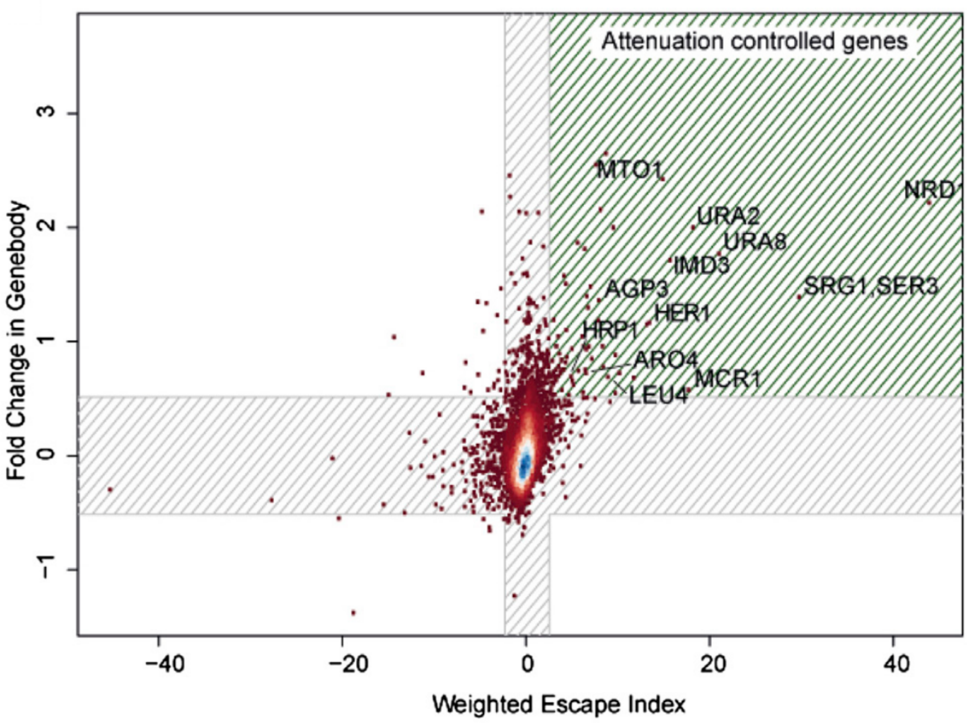


Figure 4.5: Attenuation of mRNA genes upon nuclear depletion of Nrd1 is rare under optimum growth conditions. Only 32 genes show de-attenuation upon nuclear depletion of Nrd1, as indicated by a weighted EI > 2.5 and a > 1.4-fold change in ChIP-seq genebody signal (green shaded region).

Nrd1 frequently attenuates transcription

Based on this definition of promoter proximal termination caused by the Nrd1-Nab3-Sen1 complex, I assigned early termination to 32 ORF-Ts. Single gene analysis from previous findings match with these results [3, 76, 77, 139, 146]. I could confirm Nrd1-dependent promoter proximal termination for Imd2, Hrp1, Ura2, Ura8, and Nrd1 (see Figure 4.5).

The shaded area in the scatter plot marks our predicted set of genes. This depicts that transcription attenuation is rare, based on the experimental conditions.

When aligning the differential profiles of the 32 selected ORF-Ts according to their TSS position, one can see that the differential Pol II occupancy is decreased in the promoter-proximal region, meaning there is no significant change in transcription in this region between the two conditions (see Figure 4.6 upper

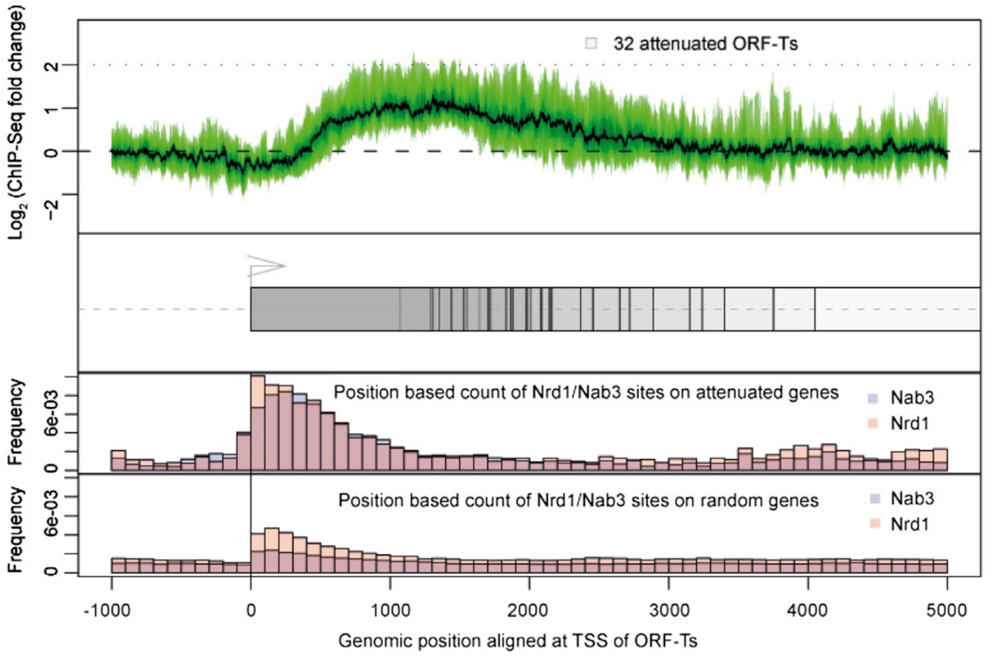


Figure 4.6: Differential profile of 32 attenuated ORF-Ts (green distribution plot) aligned and their TSSs (gray arrow). Nab3 (blue) and Nrd1 (red) binding sites according to PAR-CLIP data are peaked around +400 bps after the TSS of the 32 attenuated genes (upper panel) compared to 100 x 32 randomly drawn genes (lower panel).

panel).

In the transcript body (starting around 400 bps after the TSS), the differential Pol II occupancy is increased, which reflects higher read-through within the Nrd1-depleted condition. This shows Nrd1-dependent early termination of these 32 ORF-Ts.

Further, PAR-CLIP provides a 3.3-fold higher density of Nrd1- and Nab3-binding sites for these 32 ORF-Ts compared to randomly drawn ORF-Ts (see Figure 4.6 lower panels).

Nrd1 terminates ncRNA

Since the early termination of ORF-Ts is unlikely to be the main function of the Nrd1 pathway, analyzing Pol II binding before and after nuclear depletion of Nrd1 determined termination sites of ncRNAs. Upon Nrd1-depletion, most ncR-

NAs result in elongated transcripts, which run into downstream located ORF-T. These elongated transcripts are named Nrd1-dependent unterminated transcripts, or NUTs.

Therefore I investigated the distances between the predicted termination site of sense upstream located ncRNAs and downstream ORF-T preinitiation complexes (PICs) (see Figure 4.7). The positions of the downstream ORF-T PICs is based on ChIP-exo data of TFIIB [122], a subunit of the Pol II PIC. I could show that Nrd1 terminates transcription of ncRNAs before they would interfere with the PIC of a downstream ORF-T.

This reveals that Nrd1 is responsible for terminating transcription of ncRNAs that would otherwise disrupt transcription of ORF-Ts.

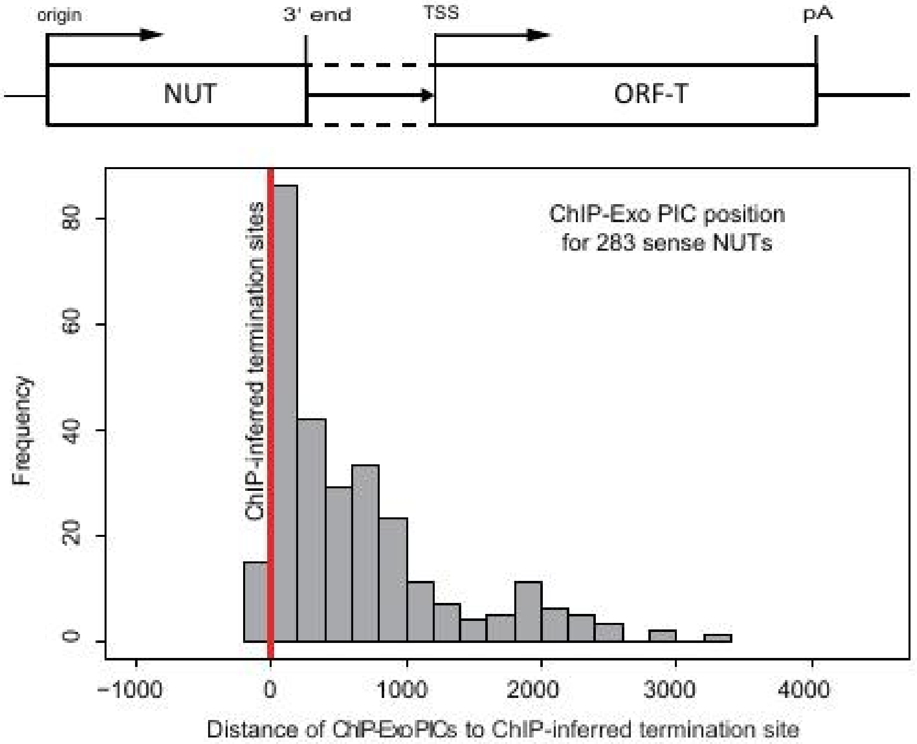


Figure 4.7: Schematic drawing of a sense NUT upstream of an ORF-T (upper panel). Predicted Termination Sites of NUTs are upstream of downstream ORF-T PICs (lower panel).

4.4 Conclusion

Previous findings document that the Nrd1-Nab3-Sen1 complex is responsible for termination of snoRNAs, snRNAs, and some ORF-Ts. With this genome-wide study, a set of 32 genes was identified which is early terminated by the Nrd1-Nab3-Sen1 complex.

Nrd1 terminates non-coding transcripts shortly after transcription initiation in proximity to the promoter to control unwanted transcription. This general role of Nrd1 is essential for proper transcription of genes because these non-coding RNAs and their transcription machinery would otherwise interrupt clean transcription of downstream ORF-T regions. Nrd1's main task as a genome-wide necessary factor is the termination of unwanted transcripts.

The analysis of ChIP-seq, RNA-seq and PAR-CLIP experimental data identified the general role of Nrd1. The diversity of the measured data gave an overall picture of the processes and helped to validate the hypothesis about the Nrd1-Nab3-Sen1 complex independently. Since all three data sets represent a different information level of the transcription process, the combined message of them is robust and can be trusted.

Further details and findings, especially those based on the RNA-seq and PAR-CLIP data sets, can be found in our CELL publication *Transcriptome surveillance by selective termination of non-coding RNA synthesis* [128].

Bibliography

[1] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010.

[2] Charles Anderson. Docker [Software engineering]. *IEEE Software*, 32(3):102–c3, 2015.

[3] John T Arigo, Kristina L Carroll, Jessica M Ames, and Jeffry L Corden. Regulation of Yeast< i> NRD1 Expression by Premature Transcription Termination. *Molecular cell*, 21(5):641–651, 2006.

[4] John T. Arigo, Daniel E. Eyler, Kristina L. Carroll, and Jeffry L. Corden. Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. *Molecular cell*, 23(6):841–851, 2006.

[5] Christopher D Armour and Pek Yee Lum. From drug to protein: using yeast genetics for high-throughput target discovery. *Curr Opin Chem Biol*, 9(1):20–24, feb 2005.

[6] 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*, 324(5935):1720–1723, 2009.

[7] T. L Bailey and C. Elkan. Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36, 1994.

[8] Timothy L. Bailey. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27(12):1653–1659, 2011.

[9] Timothy L. Bailey, Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Research*, 37(SUPPL. 2), 2009.

[10] V. V. Bakayev and G. P. Georgiev. Heterogeneity of Chromatin Subunits in Vitro and Location of Histone HI. *Nucleic Acids Research*, 3(2):477–492, 1976.

[11] Bastian Ballmann. Inside django security. *Informatik-Spektrum*, 35(3):182–189, 2012.

[12] Andrew J Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395, 2011.

[13] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4):823–837, 2007.

[14] A. Battle, Z. Khan, S. H. Wang, A. Mitrano, M. J. Ford, J. K. Pritchard, and Y. Gilad. Impact of regulatory variation from RNA to protein. *Science*, 347(6222):664–667, 2014.

[15] J. Bednar, R. A. Horowitz, S. A. Grigoryev, L. M. Carruthers, J. C. Hansen, A. J. Koster, and C. L. Woodcock. Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. *Proceedings of the National Academy of Sciences*, 95(24):14173–14178, 1998.

[16] Dan Benveniste, Hans-Joachim Sonntag, Guido Sanguinetti, and Duncan Sproul. Transcription factor binding predicts histone modifications in human cell lines. *Proceedings of the National Academy of Sciences of the United States of America*, 111(37):13367–72, 2014.

[17] Otto G. Berg and Peter H. von Hippel. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *Journal of Molecular Biology*, 193(4):723–743, 1987.

[18] Michael F Berger, Anthony A Philippakis, Aaron M Qureshi, He S Fangxue, Preston W Estep, and Martha L Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology*, 24(11):1429–1435, 2006.

[19] Mark D. Biggin. Animal Transcription Networks as Highly Connected, Quantitative Continua, 2011.

[20] Kris Brannan, Hyunmin Kim, Benjamin Erickson, Kira Glover-Cutter, Soojin Kim, Nova Fong, Lauren Kiemele, Kirk Hansen, Richard Davis, Jens Lykke-Andersen, and David L Bentley. mRNA decapping factors and the exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription. *Mol Cell*, 46(3):311–324, 2012.

[21] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–8, 2013.

[22] Martha L Bulyk, Philip L F Johnson, and George M Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Research*, 30(5):1255–1261, 2002.

[23] Stephen Buratowski. Progression through the RNA polymerase II CTD cycle. *Molecular cell*, 36(4):541–546, 2009.

[24] Michael Bustin, Frédéric Catez, and Jae Hwan Lim. The dynamics of histone H1 function in chromatin, 2005.

[25] Kaixiang Cao, Nathalie Lailler, Yunzhe Zhang, Ashwath Kumar, Karan Uppal, Zheng Liu, Eva K. Lee, Hongwei Wu, Magdalena Medrzycki, Chenyi Pan, Po Yi Ho, Guy P. Cooper, Xiao Dong, Christoph Bock, Eric E.

Bouhassira, and Yuhong Fan. High-resolution mapping of h1 linker histone variants in embryonic stem cells. *PLoS genetics*, 9(4):e1003417, 2013.

[26] Kristina L Carroll, Dennis A Pradhan, Josh A Granek, Neil D Clarke, and Jeffry L Corden. Identification of cis elements directing termination of yeast nonpolyadenylated snoRNA transcripts. *Molecular and cellular biology*, 24(14):6241–6252, 2004.

[27] Celeryproject. First steps with celery.

[28] Cenzic. Application Vulnerability Trends Report: 2014, 2015.

[29] Ping Chi, C. David Allis, and Gang Greg Wang. Covalent histone modifications — miswritten, misinterpreted and mis-erased in human cancers. *Nature Reviews Cancer*, 10(7):457–469, 2010.

[30] Peter J A Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, 2009.

[31] Theo Combe, Antony Martin, and Roberto Di Pietro. To Docker or Not to Docker: A Security Perspective. *IEEE Cloud Computing*, 3(5):54–62, 2016.

[32] Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2013.

[33] Tyler J. Creamer, Miranda M. Darby, Nuttara Jamonnak, Paul Schaughency, Haiping Hao, Sarah J. Wheelan, and Jeffry L. Corden. Transcriptome-wide binding sites for components of the Saccharomyces cerevisiae non-poly (A) termination pathway: Nrd1, Nab3, and Sen1. *PLoS genetics*, 7(10):e1002329, 2011.

[34] Menno P Creyghton, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J Steine, Jacob Hanna, Michael a Lodato, Garrett M Frampton, Phillip a Sharp, Laurie a Boyer, Richard a Young, and

Rudolf Jaenisch. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50):21931–6, 2010.

[35] Lior David, Wolfgang Huber, Marina Granovskaia, Joern Toedling, Curtis J Palm, Lee Bofkin, Ted Jones, Ronald W Davis, and Lars M Steinmetz. A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences*, 103(14):5320–5325, 2006.

[36] Abdollah Dehzangi, Yosvany López, Sunil Pranit Lal, Ghazaleh Taherzadeh, Jacob Michaelson, Abdul Sattar, Tatsuhiko Tsunoda, and Alok Sharma. PSSM-Suc: Accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction. *Journal of Theoretical Biology*, 425:97–102, 2017.

[37] Django Software Foundation. Django User Guide, 2013.

[38] Docker. Docker Releases.

[39] Docker. Overview of Docker Compose.

[40] Xianjun Dong, Melissa C Greven, Anshul Kundaje, Sarah Djebali, James B Brown, Chao Cheng, Tom R Gingeras, Mark Gerstein, Roderic Guigó, Ewan Birney, and Zhiping Weng. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology*, 13(9):R53, 2012.

[41] E. Eden and S. Brunak. Analysis and recognition of 5??? UTR intron splice sites in human pre-mRNA. *Nucleic Acids Research*, 32(3):1131–1142, 2004.

[42] Laurence Ettwiller, Benedict Paten, Mirana Ramialison, Ewan Birney, and Joachim Wittbrodt. Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nature methods*, 4(7):563–565, 2007.

[43] He Fang, David J. Clark, and Jeffrey J. Hayes. DNA and nucleosomes direct distinct folding of a linker histone H1 C-terminal domain. *Nucleic Acids Research*, 40(4):1475–1484, 2012.

[44] Christoffer Flensburg, Sarah A. Kinkel, Andrew Keniry, Marnie E. Blewitt, Alicia Oshlack, Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. A comparison of control samples for ChIP-seq of histone modifications. *Frontiers in Genetics*, 11(10):733–739, 2014.

[45] Marcel Geertz, David Shore, and Sebastian J Maerkl. Massively parallel measurements of molecular interaction kinetics on a microfluidic platform. *Proceedings of the National Academy of Sciences*, 109(41):16540–16545, 2012.

[46] Jason Gertz, Katherine E. Varley, Timothy E. Reddy, Kevin M. Bowling, Florencia Pauli, Stephanie L. Parker, Katerina S. Kucera, Huntington F. Willard, and Richard M. Myers. Analysis of dna methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. *PLoS Genetics*, 7(8), 2011.

[47] David S Gilmour and John T Lis. Detecting protein-DNA interactions in vivo: Distribution of RNA polymerase on specific bacterial genes. *Biochemistry*, 81:4275–4279, 1984.

[48] Raluca Gordân, Ning Shen, Iris Dror, Tianyin Zhou, John Horton, Remo Rohs, and Martha L. Bulyk. Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. *Cell Reports*, 3(4):1093–1104, 2013.

[49] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[50] Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Noble. Quantifying similarity between motifs. *Genome Biology*, 8(2):R24, 2007.

[51] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano Jr, Anna-Carina Jungkamp, Mathias Munschauer, and Others. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, 2010.

[52] A Hamiche, P Schultz, V Ramakrishnan, P Oudet, and A Prunell. Linker histone-dependent DNA structure in linear mononucleosomes. *Journal of molecular biology*, 257(1):30–42, 1996.

[53] Petter Hammar, Prune Leroy, Anel Mahmutovic, Erik G Marklund, Otto G Berg, and Johan Elf. The lac repressor displays facilitated diffusion in living cells. *Science (New York, N.Y.)*, 336(6088):1595–8, 2012.

[54] Nicole Happel and Detlef Doenecke. Histone H1 and its isoforms: Contribution to chromatin structure and function, 2009.

[55] Hirohito Haruki, Junichi Nishikawa, and Ulrich K Laemmli. The anchor-away technique: rapid, conditional establishment of yeast mutant phenotypes. *Molecular cell*, 31(6):925–932, 2008.

[56] Qiye He, Jeff Johnston, and Julia Zeitlinger. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat Biotechnol*, 33(4):395–401, 2015.

[57] Michael J. Hendzel, Melody A. Lever, Ellen Crawford, and John P H Th'Ng. The C-terminal Domain Is the Primary Determinant of Histone H1 Binding to Chromatin in Vivo. *Journal of Biological Chemistry*, 279(19):20028–20034, 2004.

[58] J. E. Herrera, K. L. West, R. L. Schiltz, Y. Nakatani, and M. Bustin. Histone H1 Is a Specific Repressor of Core Histone Acetylation in Chromatin. *Molecular and Cellular Biology*, 20(2):523–529, 2000.

[59] Gerald Z Hertz and Gary D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics (Oxford, England)*, 15(7-8):563–77, 1999.

[60] Jay R Hesselberth, Xiaoyu Chen, Zhihong Zhang, Peter J Sabo, Richard Sandstrom, Alex P Reynolds, Robert E Thurman, Shane Neph, Michael S Kuehn, William S Noble, Stanley Fields, and John A Stamatoyannopoulos. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods*, 6(4):283–289, 2009.

[61] Adrian Holovaty and Jacob Kaplan-Moss. *The Definitive Guide to Django: Web Development Done Right*. Development, 2009.

[62] Jing-Ping Hsin and James L Manley. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev*, 26(19):2119–2137, 2012.

[63] François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318–356, 1961.

[64] Douglas M Jacobsen and Richard Shane Canon. Contain This, Unleashing Docker for HPC. *Cray User Group 2015*, page 14, 2015.

[65] Alain Jacquier. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nature Reviews Genetics*, 10(12):833–844, 2009.

[66] Victor X. Jin, Jeff Apostolos, Naga Satya Venkateswara Ra Nagisetty, and Peggy J. Farnham. W-ChIPMotifs: A web application tool for de novo motif discovery from ChIP-based high-throughput data. *Bioinformatics*, 25(23):3191–3193, 2009.

[67] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)*, 316(5830):1497–502, 2007.

[68] Arttu Jolma, Teemu Kivioja, Jarkko Toivonen, Lu Cheng, Gonghong Wei, Martin Enge, Mikko Taipale, Juan M. Vaquerizas, Jian Yan, Mikko J. Sillanpää, Martin Bonke, Kimmo Palin, Shaheynoor Talukder, Timothy R.

Hughes, Nicholas M. Luscombe, Esko Ukkonen, and Jussi Taipale. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research*, 20(6):861–873, 2010.

[69] Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M. Vaquerizas, Renaud Vincentelli, Nicholas M. Luscombe, Timothy R. Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, 2013.

[70] Marko Jovanovic, Michael S Rooney, Philipp Mertins, Dariusz Przybylski, Nicolas Chevrier, Rahul Satija, Edwin H Rodriguez, Alexander P Fields, Schraga Schwartz, Raktima Raychowdhury, Maxwell R Mumbach, Thomas Eisenhaure, Michal Rabani, Dave Gennert, Diana Lu, Toni Delorey, Jonathan S Weissman, Steven A Carr, Nir Hacohen, and Aviv Regev. Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science (New York, N.Y.)*, 347(6226):1259038, 2015.

[71] Ann Mary Joy. Performance comparison between Linux containers and virtual machines. In *Conference Proceeding - 2015 International Conference on Advances in Computer Engineering and Applications, ICACEA 2015*, pages 342–346, 2015.

[72] P.-Y. Kan, X. Lu, J. C. Hansen, and J. J. Hayes. The H3 Tail Domain Participates in Multiple Interactions during Folding and Self-Association of Nucleosome Arrays. *Molecular and Cellular Biology*, 27(6):2084–2091, 2007.

[73] Rosa Karlić, Ho-Ryun Chung, Julia Lasserre, Kristian Vlahovicek, and Martin Vingron. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 107(7):2926–2931, 2010.

[74] Krishanpal Karmodiya, Arnaud R Krebs, Mustapha Oulad-Abdelghani, Hiroshi Kimura, and Laszlo Tora. H3K9 and H3K14 acetylation co-occur at

many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics*, 13(1):424, 2012.

[75] Jens Keilwagen and Jan Grau. Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Research*, 43(18), 2015.

[76] Ki-Young Kim and David E Levin. Mpk1 MAPK association with the Paf1 complex blocks Sen1-mediated premature transcription termination. *Cell*, 144(5):745–756, 2011.

[77] Jason N Kuehner and David A Brow. Regulation of a eukaryotic gene by GTP-dependent start site selection and transcription attenuation. *Mol Cell*, 31(2):201–211, jul 2008.

[78] Ivan V. Kulakovskiy, Ilya E. Vorontsov, Ivan S. Yevshin, Anastasiia V. Soboleva, Artem S. Kasianov, Haitham Ashoor, Wail Ba-Alawi, Vladimir B. Bajic, Yulia A. Medvedeva, Fedor A. Kolpakov, and Vsevolod J. Makeev. HOCOMOCO: Expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Research*, 44(D1):D116–D125, 2016.

[79] Jamie C. Kwasnieski, Christopher Fiore, Hemangi G. Chaudhari, and Barak A. Cohen. High-throughput functional testing of ENCODE segmentation predictions. *Genome Research*, 24(10):1595–1602, 2014.

[80] P J Laybourn and J T Kadonaga. Role of nucleosomal cores and histone H1 in regulation of transcription by RNA polymerase II. *Science*, 254(5029):238–45, 1991.

[81] Nguyen Quoc Khanh Le, Quang Thai Ho, and Yu Yen Ou. Incorporating deep learning with convolutional neural networks and position specific scoring matrices for identifying electron transport proteins, 2017.

[82] Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tackling the widespread and critical impact of batch

effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

[83] Avraham Leff and James T. Rayfield. Web-application development using the Model/View/Controller design pattern. In *Proceedings - 5th IEEE International Enterprise Distributed Object Computing Conference*, 2001.

[84] Michal Levo and Eran Segal. In pursuit of design principles of regulatory sequences. *Nature reviews. Genetics*, 15(7):453–68, 2014.

[85] J J Li and M D Biggin. Statistics requantitates the central dogma: Transcription, not translation, chiefly determines protein abundance in mammals. *Science*, 347(6226):1066–1067, 2015.

[86] Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics*, 5(3):1752–1779, 2011.

[87] Xiao Li, Hilal Kazan, Howard D. Lipshitz, and Quaid D. Morris. Finding the target sites of RNA-binding proteins, 2014.

[88] X. Lu, S. N. Wontakal, H. Kavi, B. J. Kim, P. M. Guzzardo, A. V. Emelyanov, N. Xu, G. J. Hannon, J. Zavadil, D. V. Fyodorov, and A. I. Skoultchi. Drosophila H1 Regulates the Genetic Activity of Heterochromatin by Recruitment of Su(var)3-9. *Science*, 340(6128):78–81, 2013.

[89] Sebastian Luehr, Holger Hartmann, and Johannes Söding. The XXmotif web server for eXhaustive, weight matriX-based motif discovery in nucleotide sequences. *Nucleic Acids Research*, 40(W1), 2012.

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

[91] Karolin Luger, Mekonnen L. Dechassa, and David J. Tremethick. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nature Reviews Molecular Cell Biology*, 13(7):436–447, 2012.

[92] Sebastian J Maerkl and Stephen R Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science (New York, N.Y.)*, 315(5809):233–237, 2007.

[93] 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 research*, 29(12):2471–2478, 2001.

[94] Corey R Mandel, Yun Bai, and Liang Tong. Protein factors in pre-mRNA 3 -end processing. *Cellular and Molecular Life Sciences*, 65(7-8):1099–1122, 2008.

[95] Sebastian Marquardt, Dane Z Hazelbaker, and Stephen Buratowski. Distinct RNA degradation pathways and 3'extensions of yeast non-coding RNA species. *Transcription*, 2(3):145–154, 2011.

[96] A Gregory Matera, Rebecca M Terns, and Michael P Terns. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol*, 8(3):209–220, 2007.

[97] Anthony Mathelier, Oriol Fornes, David J. Arenillas, Chih Yu Chen, Grégoire Denay, Jessica Lee, Wenqiang Shi, Casper Shyr, Ge Tan, Rebecca Worsley-Hunt, Allen W. Zhang, François Parcy, Boris Lenhard, Albin Sandelin, and Wyeth W. Wasserman. JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 44(D1):D110–D115, 2016.

[98] Anthony Mathelier and Wyeth W. Wasserman. The Next Generation of Transcription Factor Binding Site Prediction. *PLoS Computational Biology*, 9(9), 2013.

[99] Anthony Mathelier, Xiaobei Zhao, Allen W. Zhang, François Parcy, Rebecca Worsley-Hunt, David J. Arenillas, Sorana Buchman, Chih Yu Chen, Alice Chou, Hans Ienasescu, Jonathan Lim, Casper Shyr, Ge Tan, Michelle Zhou, Boris Lenhard, Albin Sandelin, and Wyeth W. Wasserman. JASPAR

2014: An extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 42(D1), 2014.

[100] V Matys, O V Kel-Margoulis, E Fricke, I Liebich, S Land, A Barre-Dirrie, I Reuter, D Chekmenev, M Krull, K Hornischer, N Voss, P Stegmaier, B Lewicki-Potapov, H Saxel, A E Kel, and E Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(Database issue):D108–10, 2006.

[101] Andreas Mayer, Martin Heidemann, Michael Lidschreiber, Amelie Schreieck, Mai Sun, Corinna Hintermair, Elisabeth Kremmer, Dirk Eick, and Patrick Cramer. CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. *Science*, 336(6089):1723–1725, 2012.

[102] Dirk Merkel. Docker: lightweight Linux containers for consistent development and deployment, 2014.

[103] Sam Meyer, Nils B. Becker, Sajad Hussain Syed, Damien Goutte-Gattat, Manu Shubhdarshan Shukla, Jeffrey J. Hayes, Dimitar Angelov, Jan Bednar, Stefan Dimitrov, and Ralf Everaers. From crystal and NMR structures, footprints and cryo-electron-micrographs to large and soft structures: Nanoscale modeling of the nucleosomal stem. *Nucleic Acids Research*, 39(21):9139–9154, 2011.

[104] Tarjei S. Mikkelsen, Manching Ku, David B. Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, William Brockman, Tae-Kyung Kim, Richard P. Koche, William Lee, Eric Mendenhall, Aisling O'Donovan, Aviva Presser, Carsten Russ, Xiaohui Xie, Alexander Meissner, Marius Wernig, Rudolf Jaenisch, Chad Nusbaum, Eric S. Lander, and Bradley E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, 2007.

[105] Hannah E Mischo and Nick J Proudfoot. Disengaging polymerase: Terminating RNA polymerase II transcription in budding yeast. *Biochimica et*

Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 1829(1):174–185, jan 2013.

[106] Tatsuya Morisaki, Waltraud G Müller, Nicole Golob, Davide Mazza, and James G McNally. Single-molecule analysis of transcription factor binding at transcription sites in live cells. *Nature communications*, 5:4456, 2014.

[107] Florian Mueller, Timothy J Stasevich, Davide Mazza, and James G McNally. Quantifying transcription factor kinetics: at work or at play? *Crit Rev Biochem Mol Biol*, 48(5):492–514, 2013.

[108] Daniel Nations. Improve Your Understanding of Web Applications, 2016.

[109] Ncbi. SRA-toolkit.

[110] Helen Neil, Christophe Malabat, Yves D.Aubenton-Carafa, Zhenyu Xu, Lars M Steinmetz, and Alain Jacquier. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature*, 457(7232):1038–1042, 2009.

[111] Shane Neph, Jeff Vierstra, Andrew B Stergachis, Alex P Reynolds, Eric Haugen, Benjamin Vernot, Robert E Thurman, Sam John, Richard Sandstrom, Audra K Johnson, Matthew T Maurano, Richard Humbert, Eric Rynes, Hao Wang, Shinny Vong, Kristen Lee, Daniel Bates, Morgan Diegel, Vaughn Roach, Douglas Dunn, Jun Neri, Anthony Schafer, R Scott Hansen, Tanya Kutyavin, Erika Giste, Molly Weaver, Theresa Canfield, Peter Sabo, Miaohua Zhang, Gayathri Balasundaram, Rachel Byron, Michael J MacCoss, Joshua M Akey, M A Bender, Mark Groudine, Rajinder Kaul, and John A Stamatoyannopoulos. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, 2012.

[112] Nginx. Nginx documentation, 2015.

[113] Ariel Ortiz. Web development with python and django. *Proceedings of the 43rd ACM technical symposium on Computer Science Education - SIGCSE '12*, page 686, 2012.

[114] Mikhail Pachkov, Piotr J. Balwierz, Phil Arnold, Evgeniy Ozonov, and Erik Van Nimwegen. SwissRegulon, a database of genome-wide annotations of regulatory sites: Recent updates. *Nucleic Acids Research*, 41(D1), 2013.

[115] Giulio Pavesi, Paolo Mereghetti, Federico Zambelli, Marco Stefani, Giancarlo Mauri, and Graziano Pesole. MoD Tools: Regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. *Nucleic Acids Research*, 34(WEB. SERV. ISS.), 2006.

[116] Mark Ptashne. The chemistry of regulation of genes and other things, 2014.

[117] Michal Rabani, Joshua Z Levin, Lin Fan, Xian Adiconis, Raktima Raychowdhury, Manuel Garber, Andreas Gnirke, Chad Nusbaum, Nir Hacohen, Nir Friedman, and Others. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature biotechnology*, 29(5):436–442, 2011.

[118] V. Ramakrishnan, J. T. Finch, V. Graziano, P. L. Lee, and R. M. Sweet. Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature*, 362(6417):219–223, 1993.

[119] Redis. Introduction to Redis.

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

[121] Ho Sung Rhee and B. Franklin Pugh. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419, 2011.

[122] Ho Sung Rhee and B Franklin Pugh. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, 483(7389):295–301, 2012.

[123] Remo Rohs, Xiangshu Jin, Sean M West, Rohit Joshi, Barry Honig, and Richard S Mann. Origins of specificity in protein-DNA recognition. *Annual review of biochemistry*, 79:233–69, 2010.

[124] Steven L. Salzberg, Arthur L. Deicher, Simon Kasif, and Owen White. Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26(2):544–548, 1998.

[125] Amy Schade. Responsive Web Design (RWD) and User Experience, 2014.

[126] Andreas Scherer. *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. Wiley Blackwell, 2009.

[127] Thomas D. Schneider and R. Michael Stephens. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100, 1990.

[128] Daniel Schulz, Bjoern Schwalb, Anja Kiesel, Carlo Baejen, Philipp Phillipp Torkler, Julien Gagneur, Johannes Soeding, Patrick Cramer, Baejen Carlo, Philipp Phillipp Torkler, Julien Gagneur, Johannes Soeding, and Patrick Cramer. Transcriptome surveillance by selective termination of non-coding RNA synthesis. *Cell*, 155(5):1075–1087, nov 2013.

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

[130] Amy C Seila, Leighton J Core, John T Lis, and Phillip A Sharp. Divergent transcription: a new feature of active promoters. *Cell Cycle*, 8(16):2557–2564, 2009.

[131] Eilon Sharon, Shai Lubliner, and Eran Segal. A feature-based approach to modeling protein-DNA interactions. *PLoS Computational Biology*, 4(8), 2008.

[132] Matthias Siebert and Johannes Soeding. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Research*, 44(13):6055–6069, 2016.

[133] Trevor Siggers and Raluca Gordân. Protein-DNA binding: Complexities and multi-protein codes. *Nucleic Acids Research*, 42(4):2099–2111, 2014.

[134] Navjot Singh, Zhuo Ma, Trent Gemmill, Xiaoyun Wu, Holland Defiglio, Anne Rossettini, Christina Rabeler, Olivia Beane, Randall H Morse, Michael J Palumbo, and Steven D Hanes. The Ess1 prolyl isomerase is required for transcription termination of small noncoding RNAs via the Nrd1 pathway. *Mol Cell*, 36(2):255–266, 2009.

[135] S. Sinha and M. Tompa. YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research*, 31(13):3586–3588, 2003.

[136] Matthew Slattery, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordân, and Remo Rohs. Absence of a simple code: How transcription factors read the genome, 2014.

[137] E J Steinmetz and D A Brow. Repression of gene expression by an exogenous sequence element acting in concert with a heterogeneous nuclear ribonucleoprotein-like protein, Nrd1, and the putative helicase Sen1. *Mol Cell Biol*, 16(12):6993–7003, 1996.

[138] Eric J Steinmetz, Nicholas K Conrad, David A Brow, and Jeffry L Corden. RNA-binding protein Nrd1 directs poly (A)-independent 3 -end formation of RNA polymerase II transcripts. *Nature*, 413(6853):327–331, 2001.

[139] Eric J Steinmetz, Christopher L Warren, Jason N Kuehner, Bahman Panbehi, Aseem Z Ansari, and David A Brow. Genome-wide distribution of yeast RNA polymerase II and its control by Sen1 helicase. *Molecular cell*, 24(5):735–746, 2006.

[140] Gary D. Stormo. Modeling the specificity of protein-DNA interactions. *Quantitative biology*, 1(2):115–130, 2013.

[141] Gary D. Stormo, Thomas D. Schneider, Larry Gold, and Andrzej Ehrenfeucht. Use of the ’perceptron’ algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research*, 10(9):2997–3011, 1982.

[142] Gary D Stormo and Yue Zhao. Determining the specificity of protein-DNA interactions. *Nature reviews. Genetics*, 11(11):751–60, 2010.

[143] Alexandra Stützer, Stamatios Liokatis, Anja Kiesel, Dirk Schwarzer, Remco Sprangers, Johannes Söding, Philipp Selenko, and Wolfgang Fischle. Modulations of DNA Contacts by Linker Histones and Post-translational Modifications Determine the Mobility and Modifiability of Nucleosomal H3 Tails. *Molecular Cell*, 61(2):247–259, 2016.

[144] Erin Styles, Ji-Young Youn, Mojca Mattiazzi Usaj, and Brenda Andrews. Functional genomics in the study of yeast cell polarity: moving in the right direction. *Philos Trans R Soc Lond B Biol Sci*, 368(1629):20130118, 2013.

[145] Sajad Hussian Syed, Damien Goutte-gattat, Nils Becker, Sam Meyer, and Manu Shubhdarshan. Single-base resolution mapping of H1–nucleosome interactions and 3D organization of the nucleosome. *Proceedings of the National Academy of Sciences*, 107(21):1–6, 2010.

[146] Marilyne Thiebaut, Elena Kisseleva-Romanova, Mathieu Rougemaille, Jocelyne Boulay, and Domenico Libri. Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. *Molecular cell*, 23(6):853–864, 2006.

[147] Morgane Thomas-Chollier, Olivier Sand, J. V. Turatsinze, R. Janky, Matthieu Defrance, Eric Vervisch, Sylvain Brohée, and Jacques van Helden. RSAT: regulatory sequence analysis tools. *Nucleic acids research*, 36(Web Server issue), 2008.

[148] William A. Thompson, Lee A. Newberg, Sean Conlan, Lee Ann McCue, and Charles E. Lawrence. The gibbs centroid sampler. *Nucleic Acids Research*, 35(SUPPL.2), 2007.

[149] Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, Kavita Garg, Sam John, Richard Sandstrom, Daniel Bates, Lisa Boatman, Theresa K Canfield, Morgan Diegel, Douglas Dunn, Abigail K Ebersol, Tristan Frum, Erika Giste, Audra K Johnson, Ericka M Johnson, Tanya Kutyavin, Bryan Lajoie, Bum-Kyu Lee, Kristen Lee, Darin London, Dimitra Lotakis, Shane Neph, Fidencio Neri, Eric D Nguyen, Hongzhu Qu, Alex P Reynolds, Vaughn Roach, Alexias Safi, Minerva E Sanchez, Amartya Sanyal, Anthony Shafer, Jeremy M Simon, Lingyun Song, Shinny Vong, Molly Weaver, Yongqi Yan, Zhancheng Zhang, Zhuzhu Zhang, Boris Lenhard, Muneesh Tewari, Michael O Dorschner, R Scott Hansen, Patrick A Navas, George Stamatoyannopoulos, Vishwanath R Iyer, Jason D Lieb, Shamil R Sunyaev, Joshua M Akey, Peter J Sabo, Rajinder Kaul, Terrence S Furey, Job Dekker, Gregory E Crawford, and John A Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012.

[150] Bryan M Turner. Defining an epigenetic code. *Nature Cell Biology*, 9(1):2–6, 2007.

[151] Bryan M. Turner. Nucleosome signalling; An evolving concept, 2014.

[152] E L Van Dijk, C L Chen, Y D Aubenton-Carafa, S Gourvennec, M Kwapisz, V Roche, C Bertrand, M Silvain, P Legoix-Ne, S Loeillet, and Others. XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature*, 475(7354):114–117, 2011.

[153] Lidia Vasiljeva, Minkyu Kim, Hannes Mutschler, Stephen Buratowski, and Anton Meinhart. The Nrd1–Nab3–Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. *Nature structural & molecular biology*, 15(8):795–804, 2008.

[154] P H von Hippel and O G Berg. On the specificity of DNA-protein interactions. *Proc. Natl. Acad. Sci. U.S.A.*, 83(6):1608–1612, 1986.

[155] Peter H von Hippel. From "simple" DNA-protein interactions to the macromolecular machines of gene expression. *Annual review of biophysics and biomolecular structure*, 36:79–105, 2007.

[156] Jie Wang, Jiali Zhuang, Sowmya Iyer, XinYing Lin, Troy W. Whitfield, Melissa C. Greven, Brian G. Pierce, Xianjun Dong, Anshul Kundaje, Yong Cheng, Oliver J. Rando, Ewan Birney, Richard M. Myers, William S. Noble, Michael Snyder, and Zhiping Weng. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*, 22(9):1798–1812, 2012.

[157] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, jan 2009.

[158] Christopher L Warren, Natasha C S Kratochvil, Karl E Hauschild, Shane Foister, Mary L Brezinski, Peter B Dervan, George N Phillips, and Aseem Z Ansari. Defining the sequence-recognition profile of DNA-binding molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 103(4):867–72, 2006.

[159] Wu Wei, Vicent Pelechano, Aino I Järvelin, and Lars M Steinmetz. Functional consequences of bidirectional promoters. *Trends in Genetics*, 27(7):267–276, 2011.

[160] Lukas Windhager, Thomas Bonfert, Kaspar Burger, Zsolt Ruzsics, Stefan Krebs, Stefanie Kaufmann, Georg Malterer, Anne L'Hernault, Markus Schilhabel, Stefan Schreiber, and Others. Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Research*, 22(10):2031–2042, 2012.

[161] Wiebke Wlotzka, Grzegorz Kudla, Sander Granneman, and David Tollervey. The nuclear RNA polymerase II surveillance system targets polymerase III transcripts. *EMBO J*, 30(9):1790–1803, 2011.

[162] Christopher L. Woodcock, Arthur I. Skoultchi, and Yuhong Fan. Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. *Chromosome Research*, 14(1):17–25, 2006.

[163] Françoise Wyers, Mathieu Rougemaille, Gwenaël Badis, Jean-Claude Rousselle, Marie-Elisabeth Dufour, Jocelyne Boulay, Béatrice Régnault, Frédéric Devaux, Abdelkader Namane, Bertrand Séraphin, and Others. Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly (A) polymerase. *Cell*, 121(5):725–737, 2005.

[164] Xiaonuo Gantan. Writing simple views for your first python django application, 2013.

[165] Zhenyu Xu, Wu Wei, Julien Gagneur, Fabiana Perocchi, Sandra Clauder-Münster, Jurgi Camblong, Elisa Guffanti, Françoise Stutz, Wolfgang Huber, and Lars M Steinmetz. Bidirectional promoters generate pervasive transcription in yeast. *Nature*, 457(7232):1033–1037, 2009.

[166] Nicolae Radu Zabet and Boris Adryan. Estimating binding properties of transcription factors from genome-wide binding profiles. *Nucleic Acids Research*, 43(1):84–94, 2015.

[167] B.-R. Zhou, H. Feng, H. Kato, L. Dai, Y. Yang, Y. Zhou, and Y. Bai. Structural insights into the histone H1-nucleosome complex. *Proceedings of the National Academy of Sciences*, 110(48):19390–19395, 2013.

[168] Bing Rui Zhou, Jiansheng Jiang, Hanqiao Feng, Rodolfo Ghirlando, T. Sam Xiao, and Yawen Bai. Structural Mechanisms of Nucleosome Recognition by Linker Histones. *Molecular Cell*, 59(4):628–638, 2015.