# Nucleosome Occupancy and Dynamics in Yeast

## Genome-wide and Promoter-level Analyses and Modeling

Michael Roland Wolff

Munich, 2020

# Nucleosome Occupancy and Dynamics in Yeast

## Genome-wide and Promoter-level Analyses and Modeling

Michael Roland Wolff

Doctoral thesis
presented to the Faculty of Physics
of the Ludwig–Maximilians–Universität
München

by

Michael Roland Wolff
from Frankfurt (Oder)

Munich, 20.08.2020

Erstgutachter: Prof. Dr. Ulrich Gerland
Zweitgutachter: Prof. Dr. Erwin Frey
Tag der mündlichen Prüfung: 07.10.2020

# Contents

# Zusammenfassung

Der zentralen Baustein des Chromatins ist ein Protein-DNA Komplex, das Nukleosom, bestehend aus 147 bp DNA, die sich um einen Proteinkern, ein Oktamer aus Histonen, wickelt. Erste Aufnahmen mittels Elektronenmikroskopie zeigten Nukleosome wie Perlen auf einer Kette. Die Positionen dieser “Perlen” sind nicht zufällig, sondern formen reguläre Arrays um nukleosomarme Regionen (nucleosome depleted region, NDR), wie genomweites Nukleosom-mapping gezeigt hat. Diese sogenannten stereotypischen NDR-Array Muster kann man oft in Promotorregionen finden und sie sind von zentraler Bedeutung für fundamentale Regulierungsprozesse wie Transkription, Replikation oder DNA Reparatur. Deshalb spielt in der Chromatinforschung heute die Untersuchung der Bestimmungsfaktoren, Auswirkungen und Mechanismen der Nukleosompositionierung und -besetzung eine große Rolle. Während die Positionierung beschreibt, wo auf dem Genom Nukleosome typischerweise sitzen, misst man mit der Nukleosombesetzung, wie häufig, also in wie vielen Zellen, sie an einer bestimmten Position sind.

Die Abwesenheit oder Gegenwart von Nukleosomen auf einem Promotor entscheidet, ob ein bestimmtes Gen zu RNA transkribiert wird, um später das entsprechende Protein herzustellen, oder nicht. In Hefe werden die meisten Gene kontinuierlich transkribiert und haben einen nukleosomarmen Promotor. Die Promotorregion von inaktiven Genen ist jedoch oft mit Nukleosomen besetzt. Ein Beispiel für ein induzierbares Gen in *Saccharomyces cerevisiae*, der Bäckerhefe, ist das *PHO5* Gen. Die Nukleosome des *PHO5* Promotors werden unter Phosphatmangel entfernt, als Voraussetzung für die Aktivierung der Transkription. Damit spielen nicht nur Nukleosompositionierung und Besetzung an Promotoren eine große Rolle in der Genregulation, sondern auch die Dynamik von Nukleosomaufbau und -abbau. Diese entscheidet zum Beispiel, wie schnell sich eine Zelle an eine sich ändernde Umgebung anpassen kann. In dieser Dissertation werden interdisziplinäre Projekte zwischen molekularer Biochemie, Bioinformatik und theoretischer Biophysik präsentiert, die die Nukleosombesetzung und Nukleosomdynamik auf unterschiedlichen Beschreibungsebenen untersuchen.

Das einleitende erste Kapitel gibt eine kurze Zusammenfassung neuerer biologische Erkenntnisse der Chromatinforschung, beginnend mit den Eigenschaften von Nukleosomen und unterschiedlichen Messtechniken, die Besetzung und Position von Nukleosomen entlang der DNA zu bestimmen. Wir wiederholen kurz physikalische Modelle der Nukleosompositionierung, die auf das eindimensionale Gasmodell von Tonks für harte Teilchen aufbauen und geben eine Einführung in den Forschungsbereich der Chromatinremodellierung, die die Rolle von ATP-abhängigen Enzymen untersucht, welche die Nukleosomstruktur und -position manipulieren.

Im zweiten Kapitel wird die Entwicklung und die Anwendung einer neuen genomweiten Methode, um die absolute Besetzung des Hefegenoms zu messen, geschildert. Diese ist definiert über die Anzahl von Zellen mit einem Nukleosom (oder einem anderen DNA-bindenden Faktor) an einer bestimmten genomischen Position, dividiert durch die Gesamtanzahl untersuchter

Zellen. Etablierte Methoden messen diese Besetzung nur relativ und manchmal zusätzlich verzerrt durch systematische Messabweichungen. Nach der Verifikation über mehrere orthogonale experimentelle Ansätze, jeder mit seiner eigenen neuentwickelten Analyse, präsentieren wir eine hochauflösende absolute Besetzungskarte des Hefegenoms. Mit Hilfe dieser bestimmen wir zum Beispiel die durchschnittliche Nukleosomenzahl pro Zelle zu 57 000 bis 60 000 und die durchschnittliche genomweite absolute Besetzung zu 78%. Weiterhin finden wir im Allgemeinen keine Korrelationen zwischen der absoluten Besetzung eines Gens und dessen Transkriptionsrate. Wir analysieren auch Daten neuer Sequenziertechniken für lange DNA Fragmente und geben hier einen Ausblick auf die Vorteile dieser Technik im Gegensatz zu anderen Methoden, die nur Durchschnittsergebnisse über viele Zellen liefern.

Das dritte Kapitel handelt von Computersimulationen unterschiedlicher idealisierter Mechanismen zur Chromatinremodellierung, die auf ein eindimensionales Gas von Nukleosomen mit weichem Kern wirken. Aus der Perspektive der Physik führen die Aktionen dieser Remodellierer zu Nichtgleichgewichtssystemen. Mit dem Hinzufügen eines aktiven direktionalen Nukleosomverschiebemechanismus ändert sich die stationäre Dichte des Systems, kann aber mit dem ursprünglichen Nukleosomgasmodell im Gleichgewicht ohne Remodellierer, dafür mit nun weicheren Nukleosomen, gefittet werden. Also hängt die scheinbare Nukleosomweichheit in der stationären Dichte auch von anderen Faktoren, nicht nur reinen Nukleosomeigenschaften, ab. Wir vergleichen auch Mechanismen zur Nukleosombeseitigung und erforschen qualitative Effekte von DNA-bindenden Remodellierungsenzymen. Um jedoch quantitative Untersuchungen durchzuführen, sind experimentelle Daten über Nukleosomkonfigurationen, also simultane Informationen über benachbarte Nukleosome, unter dem Einfluss von Remodellierungsenzymen notwendig.

Im vierten Kapitel benutzen wir Daten zur Häufigkeit solcher Nukleosomkonfigurationen und erforschen einen neuen Modellierungsansatz für die Dynamik von Nukleosomkonfigurationen auf dem *PHO5* Promotor in Hefe. Mit einem einzigartigen, mittels Elektronenmikroskopie gewonnen, Datensatz für den *PHO5* Promotor als Ausgangpunkt, entwickeln und untersuchen wir eine neue Klasse von Modellen, "regulated on-off-slide models", welche den Nukleosomaufbau, -abbau und Nukleosomverschiebungen innerhalb von Konfigurationen beschreiben. Nach dem systematischen Testen aller Modelle bis zu einer fixierten Parameterzahl, präsentieren wir die kleine Zahl von Modellen, die in der Lage sind die unterschiedlichen gemessenen Datensätze zu reproduzieren: Nukleosomkonfigurationen unterschiedlicher Promotoraktivierungsgrade, Besetzungsdaten von *PHO5* Promotormutanten und Daten zur Dynamik des Nukleosomeinbaus. Diese Daten zur Dynamik ermöglichen auch die Berechnung der Zeitskalen der Modelle. In all diesen, sehr ähnlichen, Modellen wird nur die Rate von Prozessen zum Nukleosomaufbau variiert, um zwischen den unterschiedlichen Promotoraktivierungsgraden zu wechseln, was das vorherrschende Bild zur Chromatinregulierung auf dem *PHO5* Promotor herausfordert.

# Abstract

The central building block of chromatin is a protein-DNA complex, the nucleosome, consisting of 147 bp of DNA wrapped around a protein core, a histone octamer. The nucleosomal structure is highly conserved between eukaryotes, from yeast to human. First visualization by electron microscopy revealed that nucleosomes look like beads on a string. The "bead" positions are not random, but form regular arrays around nucleosome depleted regions (NDR), as shown by genome-wide nucleosome mapping. These so-called stereotypical NDR-array patterns are often found at promoter regions, and they are essential for fundamental regulatory processes like transcription, replication or DNA repair. Thus, today a major part of chromatin research is to investigate the determinants, mechanisms and impacts of nucleosome positioning and nucleosome occupancy. While nucleosome positioning denotes where on the genome nucleosomes are in the cell average, nucleosome occupancy monitors how often, i.e. in how many cells, a certain position is covered by a nucleosome.

The absence or presence of nucleosomes on the promoter determines whether a certain gene is transcribed to RNA to later yield the corresponding protein or not. In yeast, most genes are constantly transcribed and exhibit a nucleosome depleted promoter. The promoter region of inactive genes, however, is often occupied by nucleosomes. An example for an inducible gene in *Saccharomyces cerevisiae*, the budding yeast, is the *PHO5* gene. Some nucleosomes in the *PHO5* promoter are removed as a prerequisite for transcription activation under phosphate starvation. Therefore, not only nucleosome positioning and occupancy at promoters, but also the dynamics of nucleosome assembly and disassembly play an important role in gene regulation, determining, for instance, how fast a cell can adapt to a changing environment. This thesis presents interdisciplinary projects between molecular biochemistry, bioinformatics and theoretical biophysics investigating nucleosome occupancy and dynamics at different descriptive levels.

The introductory Chapter 1 gives a short summary of recent biological results in chromatin research, starting with known nucleosome properties and the different established measurement techniques to determine nucleosome occupancy and positioning along the DNA. We also shortly review physical models of nucleosome positioning based on the one-dimensional Tonks gas model for hard particles and give an introduction to chromatin remodeling research, to provide insight into the role of ATP-dependent enzymes manipulating nucleosome structure and positions.

In Chapter 2 we describe the development and applications of a new genome-wide method to measure the absolute occupancy in yeast, defined by the number of cells with a nucleosome (or other DNA-binding factor) at a certain position divided by the total number of investigated cells. Established methods measure occupancy only in a relative manner, sometimes further distorted by measurements biases. After ensuring that different orthogonal experimental approaches, each with its own newly developed analysis pipeline, yield the same results, we present a high resolution (9 bp) absolute occupancy map of the yeast genome. We use this map to find for example the average number of nucleosomes per yeast cell at around 57 000 to 60 000, and a global average absolute occupancy of 78%. Furthermore we find no correlation between absolute occupancy of gene bodies and their transcription rate. We also analyze data from new long read sequencing techniques and give here an outlook on its benefits in contrast to short read methods which only yield cell-averaged results.

Chapter 3 deals with computational simulations of different idealized chromatin remodeling mechanisms acting on a one-dimensional soft-core nucleosome gas. From a physics point of view, these remodeler actions lead to non-equilibrium systems. When adding an active directional nucleosome sliding mechanism the steady state density of the system changes, but can be fitted by the original equilibrium nucleosome gas model without remodelers, albeit with now softer nucleosomes. Thus, the apparent softness of nucleosomes in the steady state density also depends on other factors, not only pure nucleosome properties. We also compare different nucleosome eviction mechanisms and explore qualitative effects of DNA-binding of remodeler enzymes. However, to perform quantitative investigations experimental data on nucleosome configurations, i.e. information on the combined state of neighboring nucleosomes, under the influence of remodelers are necessary.

In Chapter 4, we use measured occurrences of such nucleosome configurations and explore a new modeling approach for the dynamics of nucleosome configurations on the *PHO5* promoter in yeast. Using a unique electron microscopy data set for this promoter, we develop and investigate a new class of models, “regulated on-off-slide models”, describing nucleosome assembly, disassembly and sliding within nucleosome configurations. After systematically testing all models up to a fixed number of parameters, we present the very few models able to fit three different types of data: the nucleosome configuration data set for different activation states, occupancy data of *PHO5* promoter mutants and data describing the dynamics of nucleosome incorporation, which enable the calculation of model time scales. All these models are quite similar and only vary the rate of assembly processes to regulate between the different activation states, challenging the current view of chromatin regulation on the *PHO5* promoter.

# List of Figures

# List of Tables

# 1 Introduction to Chromatin, Nucleosomes and Remodelers

This introductory chapter first gives an overview of chromatin and nucleosome properties and the importance of nucleosomes for gene regulation. The second section provides a brief review of experimental methods to measure nucleosome positioning and occupancy along the genome, which is especially needed for Chapter 2, followed by a section on physical models of nucleosome positioning, as a prerequisite for Chapter 3. Lastly we sum up recent results regarding chromatin remodeling, with general results being helpful in Chapter 3 and Chapter 4.

## 1.1 Nucleosome properties

The term "chromatin" was first coined by Walther Flemming, one of the pioneers of mitosis research, over 140 years ago in 1879, to describe the easily stainable fibrous scaffold inside the nucleus of eukaryotes [1]. Later "chromosomes" were introduced, to label the nuclear threads Flemming already observed inside the nucleus in preparation to cell division (Figure 1.1). He rightfully suspected they consists of the same nuclear material as chromatin, just at different phases in the cell cycle [1].

Since 1974, we know that the central building block of chromatin is a protein-DNA complex, the nucleosome, consisting of 147 bp of DNA wrapped around a protein core, a histone octamer, as first suggested by Kornberg [3] (Figure 1.2). Nucleosomes are the first layer of DNA packaging inside the nucleus and part of the gene expression control mechanism. If viewed as a disk, the nucleosome has a diameter of 11 nm and a height of 5.7 nm [4]. The bound DNA forms a distorted left-handed superhelix of 147 bp length corresponding to 1.7 turns around the core. The center base pair is called "nucleosome dyad". The core is structured in a H3-H4 tetramer that binds to two H2A-H2B dimers and has a relative molecular mass of 206 000. The H3-H4 tetramer always binds to the first and last part of the 147 bp as well as the central part. The H2A-H2B dimers bind to the remainder of the 147 bp long histone bound DNA. In higher eukaryotes, there is an additional H1 histone, which binds to the DNA next to the core. The nucleosome cores are highly conserved and occur every $(200 \pm 40)$ bp throughout all eukaryotic genomes [5].

Because of the interactions between histones and DNA, the binding of other factors to the DNA that is part of a nucleosome is inhibited. Thus, the positioning and occurrence of nucleosomes can regulate DNA-dependent processes, like the binding of transcription factors [6, 7], replication machinery [8] and DNA repair [9, 10]. When the DNA is transcribed or replicated, the nucleosomes need to be disassembled at least partially, and reassembled again after. Nucleosomes regulate DNA-dependent processes not only by merely occupying DNA, but also by

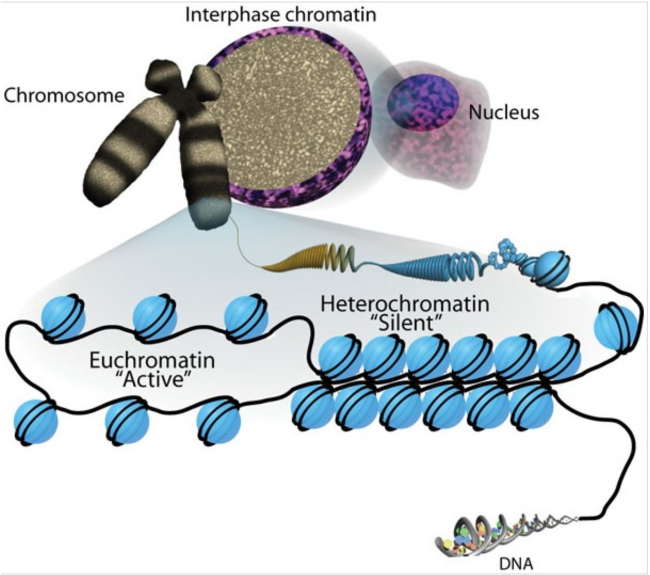


**Figure 1.1:** Illustration of chromatin inside the cell nucleus during interphase as well as before cell devision (chromosomes). Light blue spheres denote nucleosomes in the different chromatin states of higher eukaryotes: euchromatin is more accessible to other factors and permissible for transcription. Heterochromatin is more densely packed and harder to access. From: [2] under CC BY 3.0 license.

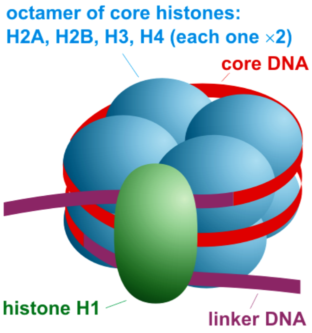


**Figure 1.2:** Schematic nucleosome organization with histone octamer (blue), 147 bp of core DNA (red), linker DNA (purple) and H1 histone with occurs more often in higher eukaryotes (green). Nucleosome and DNA diameters are not to scale. For a realistic cryo-EM reconstruction with additional bound enzyme see Figure 1.5. From: Wikimedia Commons: nucleosome organization by Darekk2 under CC BY-SA 3.0 license.

carrying different histone variants and posttranslational modification of the histones allowing short-term as well as long-term regulation [11]. In higher eukaryotes, the regions that are densely packed and thus harder to access by other factors are called heterochromatin, whereas euchromatin is more accessible (Figure 1.1).

Nucleosomes are constantly disassembled, reassembled, moved or modified by ATP-dependent remodeler enzymes (see Section 1.4). The resulting nucleosome dynamics can be quantified throughout the genome to some degree by measuring nucleosome turnover rates [12, 13], which we will use in Chapter 4. Once assembled, nucleosomes are not static and the DNA of a nucleosome can partially unwrap beginning from the ends of the core region and make binding to the unwrapped part possible [14, 15, 16, 17]. This “nucleosome breathing” also allows nucleosomes to invade each other, leading to a partial overlap of the 147 bp core region [18]. The unwrapping rate decreases drastically with distance inside the nucleosome [19]. Additionally, nucleosome breathing motivates biophysical soft-core interacting particle models (see Section 1.3).

## 1.2 Measurement and features of relative nucleosome density and occupancy in yeast

Nucleosomes can be visualized using an electron microscope, where they appear to look like beads on a string, but other techniques are needed to investigate where on the genome nucleosomes form (for a recent reviews see [20, 21]). We begin with reviewing MNase-based nucleosome detection methods, where only the nucleosomal DNA is sequenced. Then we will explain chemical cleavage methods, where nucleosomes are modified to cut the DNA and investigate general features of genome-wide high resolution maps obtained by both methods.

In micrococcal nuclease (MNase) based methods the linker DNA between nucleosomes is (partially) digested, as MNase cuts the accessible linker DNA. The histone bound DNA, however, remains intact and can be sequenced (MNase-seq) [22]. Fragments with typical nucleosome footprint length indicate nucleosome positions on the genome. The MNase concentration as well as the digestion duration can be tuned to obtain low or high digestion degrees. Care must be taken to find the “correct” digestion degree, as also nucleosomal DNA is slowly digested. Since non-nucleosomal binding factors can have a similar footprint as nucleosomes, one can additionally include an immunoprecipitation step with anti-histone antibodies, to make sure only nucleosomal DNA fragments are sequenced (MNase-ChIP-seq) [23, 24]. The dyad position can be estimated by the genomic position of the fragment center. This dyad signal can be convoluted with a certain footprint length, typically 147 bp, to obtain an estimated “nucleosome occupancy” or a shorter footprint length, like 50 bp, to obtain a smoothed dyad signal for easier visual separation of peaks. Signal peaks then correspond to typical nucleosome positions, while peak heights scale with the likelihood of finding a nucleosomes at a given position. However this simple interpretation has several caveats: (i) MNase-seq / MNase-ChIP-seq measures cell averages, since the signal is a sum of all sequenced fragments of many cells. Thus, signal peaks are well-positioned nucleosomes, i.e. a large portion of cells has a nucleosome at the peak location, while randomly positioned nucleosomes averaged over many cells do not lead to peaks. (ii) Since only the nucleosomal fragments are sequenced, it is only possible to make relative statements of nucleosome occupancy, e.g. between different positions of the genome, while the absolute occupancy, i.e. the fraction of cells that has a nucleosome at a given position, remains unknown. Trying to find the right normalization, e.g. by taking the maximum peak height and setting it to 1, usually is not possible because (iii) even relative statements suffer from several biases of MNase based methods e.g. the preferred cutting of AT-rich DNA, even within nucleosomes. But new theoretical frameworks and spike-in controls can be useful to control theses biases [25].

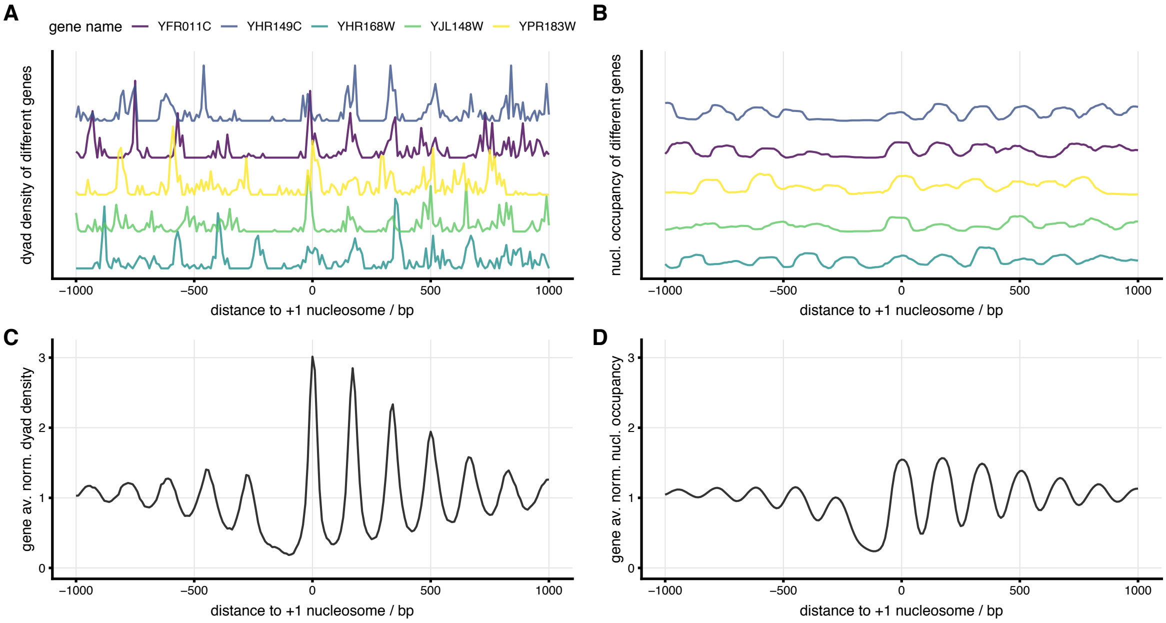


**Figure 1.3:** N+1 aligned dyad density and nucleosome occupancy in *S. cerevisiae* measured by chemical cleavage [26] showing the nucleosome depleted region (NDR) and nucleosome positions (.., N-1, N+1, N+2, ...). (**A**) Exemplary dyad densities (normalized to a mean of 1) of five random genes shifted successively by 5 in y-direction for better visibility. Each gene is aligned with respect to the N+1 position (0 on the x-axis), calculated with the peak calling algorithm DANPOS [27] as in [28]. (**B**) To obtain a nucleosome occupancy, the dyads were extended to a footprint of 101 bp and the density normalized again to 1. This reduced footprint was used to increase the visibility of the linker regions [26]. Same scale, genes and N+1 alignment as in panel A. (**C**) Gene averaged dyad density of 4785 N+1 aligned genes normalized to a mean density of 1. (**D**) As in panel C but for the 101 bp nucleosome footprint occupancy.

Another experimental approach are modified histones that contain chemical cleavage sites [29, 26]. The DNA is then cut at symmetrically located sites very close to the dyad of the nucleosome and sequenced. Ideally, this provides the locations of the two nucleosomes where the DNA has been cut and the resulting data can even be interpreted as next-neighbor pair density [29], allowing the analysis of distances between two neighboring nucleosomes. However, DNA might also be cut without a nucleosome present, leading to a false signal. A more precise chemical cleavage method to locate nucleosomes uses two cut sites exactly 51 bp apart centered around the nucleosome dyad [26], with the resulting 51 bp fragments still being long enough to be accurately mapped to the genome. This allows a very accurate measurement of nucleosome positions, but as MNase-seq, it scores nucleosomal DNA only, not allowing absolute occupancy statements. Like in MNase-seq or MNase-ChIP-seq the dyad densities can be convoluted with a certain footprint length to obtain a nucleosome occupancy (Figure 1.3A and B). This technique works very well in yeast like *Saccharomyces cerevisiae*, but higher organisms can have many more copies of histone genes, and all of them would need to be properly mutated to introduce the chemical cleavage sites at each nucleosome.

As a side note, it is also possible to measure the position of linker DNA between nucleosome cores using hyperactive transposase. Here, sequencing adapters are inserted by the Tn5 transposase into nucleosome-free DNA and the resulting fragments sequenced (assay for transposase-accessible chromatin, ATAC-seq) [30].

Genome-wide high resolution maps, as generated for example by MNase-seq or chemical cleavage methods, revealed a stereotypical distribution of the nucleosomes on individual genes in yeast, called "stereotypical NDR-array pattern", with a nucleosome depleted region (NDR), at the promoter, followed by an array of peaks downstream of the transcription start site (TSS) (Figure 1.3A and B). These peaks correspond to well-positioned nucleosomes and historically these positions, even though they are not always strictly fixed, were named "N+1", then "N+2", "N+3" and so on. Conversely, the positions upstream of the TSS are called "N-1", "N-2", ... By convention, there is no "N±0" position.

This stereotypical NDR-array pattern at TSSs observed in yeast occurs mainly at constitutively active genes (housekeeping or growth genes), however, not at inducible or silent genes. This leads to a division into stereotypical (or canonical) and non-stereotypical (non-canonical) genes, either exhibiting the stereotypical NDR-array pattern, or not, respectively. This distinction becomes more important for higher eukaryotic genomes, like *Drosophila* or human genomes, which have more silent genes than yeast [31, 32]. The promoters of non-canonical genes exhibit for instance a narrowed NDR, a NDR that is shifted upstream, or no nucleosome depleted region at all. An example of non-canonical genes in yeast are the *PHO* genes that are induced by phosphate starvation, especially *PHO5*. Here, upon induction, up to three promoter nucleosomes are removed, providing access to transcription factor binding sites and finally activating transcription [33]. In Chapter 4, we investigate the nucleosome configurations of the *PHO5* promoter in different activation states in detail.

When averaging over the N+1 aligned gene dyad densities or occupancies, one obtains a very generic average array pattern (Figure 1.3C and D) with the highest peak at 0, the N+1 position. The gene averaged N+1 peak is then followed by a periodic array of slowly declining peaks in downstream direction. Before the N+1 peak at 0, i.e. in upstream direction, is a region of low average density/occupancy which is larger than the linker regions between the downstream peaks, followed by much lower average N-1, N-2,... peaks. The lower average peak levels downstream of the NDR can be explained by the different NDR lengths in individual genes (Figure 1.3A) which can lead to a misalignment of the individual N-1, N-2,... positions. These gene averaged patterns have been an interesting starting point for biophysical modeling (see Section 1.3) and are often used to investigate the effects of mutations for instance of chromatin remodeling enzymes (see Section 1.4) on nucleosome positioning.

In Chapter 2 we present newly developed methods to directly measure the absolute occupancy along the genome, with occupancy values between 0 and 1 corresponding to the fraction of occupied DNA-molecules over all investigated DNA-molecules at each measurement site. Additionally, all up to now mentioned methods measure cell averages of nucleosome density, occupancy or nucleosome-free DNA. Single-cell measurement techniques, however, are becoming more and more important. One example is the measurement of nucleosome configurations, i.e. simultaneous observations of several nucleosomes on the same DNA molecule, at the *PHO5* promoter [34]. We will use this data set together with other data to systematically find effective models for the dynamics of promoter nucleosome configurations in Chapter 4. Another example of single-cell measurements are long read sequencing techniques based on nanopore sequencing, which allow the sequencing of reads much longer than the usual couple hundred base pairs. Since the occupancy signal on each long read is a snapshot of the occupancy of a single cell, this allows inference of the positions of several nucleosomes on the same DNA molecule. We will investigate this method in Section 2.5.

## 1.3 Physical models of the gene-averaged nucleosome density

In the following we will review physical models that are able to effectively describe the measured gene-averaged nucleosome density pattern. All of these models are based on one-dimensional grand canonical systems, with hard or soft interaction potential and homogeneous or heterogeneous external potential. The mathematical basics of these systems and calculations from potentials to densities as well as inverse methods are briefly reviewed in Section B.1.

In 1988, Kornberg and Stryer analyzed data from MNase digestion and gel electrophoresis and concluded that random nucleosome positions near a boundary result in an array of regularly spaced nucleosomes at seemingly non-random locations [35]. Later experimental studies used MNase-ChIP-seq data in yeast to claim that “the organization of nucleosomes throughout genes is largely a consequence of statistical packing principles” [36]. This idea of “statistical positioning” corresponds to the one-dimensional Tonks gas model with hard core interaction near a fixed boundary [37] and leads to an effective description of the gene-averaged nucleosome density or occupancy by equilibrium gas models of extended particles.

It has been firmly established, that the DNA-binding affinity of histones is dependent on the DNA sequence, mostly because the biophysical properties of DNA, as it bendedness and bendability, depend on the sequence [38, 39]. Furthermore active remodeling changes the nucleosome landscape [24]. However, modeling the (relative) nucleosome density in *S. cerevisiae* aligned at N+1 positions and averaged over many genes (like shown in Figure 1.3C) is quite successful without knowing the underlying sequence affinities or remodeler mechanics. It was possible to achieve quantitative agreement downstream and upstream of the nucleosome free region just using the Tonks gas model and different boundary conditions [40], i.e. a directly positioned N+1 nucleosome and a N-1 nucleosome that is statistically positioned by a nucleosome-repelling DNA region. The different boundary conditions of N+1 and N-1 correspond to the observation that the N-1 aligned pattern in upstream direction is not just a mirrored version of the N+1 aligned pattern in downstream direction. The amplitude of the N-1 aligned pattern is indeed much lower [40].

Trying to find a unified physical model for 12 yeast species, Tonks gas model was extended with soft instead of hard particles [41]. Such a soft nucleosome gas takes into account nucleosome breathing, i.e. transient unwrapping of DNA segments that are usually part of the nucleosome, which is parameterized by an effective energy cost for DNA unwrapping and an effective

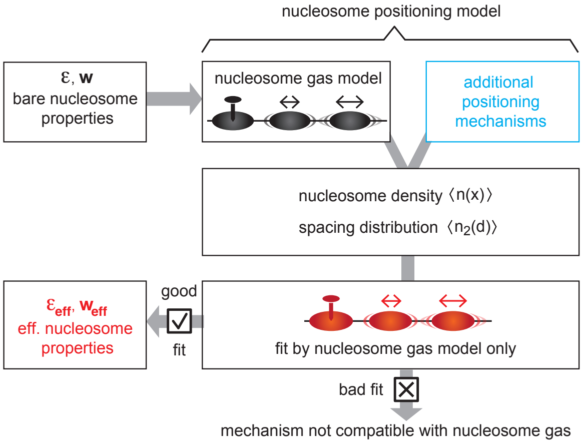

nucleosome positioning model
ε, w
bare nucleosome properties
nucleosome gas model
additional positioning mechanisms
nucleosome density ⟨n(x)⟩
spacing distribution ⟨n2(d)⟩
εeff, weff
eff. nucleosome properties
good
☑
fit
fit by nucleosome gas model only
bad fit ☒
mechanism not compatible with nucleosome gas

**Figure 1.4:** General framework to analyze the effect of specific nucleosome positioning mechanisms on the soft nucleosome gas. The nucleosome density $\langle n(x) \rangle$ and spacing distribution $\langle n_2(d) \rangle$ are the results from the interplay of the "nucleosome gas model" with bare nucleosome parameters $\varepsilon$ and $w$ (stiffness and size) and additional positioning mechanisms. If $\langle n(x) \rangle$ and $\langle n_2(d) \rangle$ can be described by the nucleosome gas model alone (fit good), the effective nucleosome properties $\varepsilon_{\text{eff}}$ and $w_{\text{eff}}$ may differ from the bare ones. Adapted from [44] under CC BY-NC 4.0 license.

nucleosome size. It was possible to fit 11 of the 12 species with the same soft nucleosome parameters, and to accommodate the pattern differences between species just by their different average nucleosome density [41]. Furthermore, soft nucleosomes significantly speed up the nucleosome filling process compared to hard particles in theoretical models, for example during reassembly of densely packed nucleosomes after DNA replication [42, 43].

The apparent contradiction that "simple" nucleosome gas models are able to fit measured nucleosome density data with supposedly several underlying positioning mechanisms, like DNA-dependence or remodeling activity was investigated by Nübler et al. [44]. They illustrate that the declining periodic density pattern near a boundary and the corresponding spacing (next neighbor distance) distribution are very robust with respect to different additional positioning mechanisms (general framework in Figure 1.4). "Statistical positioning emerges as an effective description from the complex interplay of different positioning mechanisms, which ultimately only renormalize the model parameter quantifying the effective softness of nucleosomes" [44]. Thus, even including a DNA sequence-dependent nucleosome positioning landscape, active directional sliding remodeling (see Chapter 3) or nucleosome spacing by an effective moderate attractive interaction at a preferred spacing distance still results in a qualitatively equal gene-averaged declining periodic density pattern that can be very well described by a soft nucleosome gas without these positioning mechanisms. Note that while some mechanisms lead to apparent nucleosome softening, like seemingly random, uncorrelated positioning landscapes and active directional sliding (Section 3.2), others, like effective attractive interactions and periodic trapping positioning landscapes have the opposite effect and make nucleosomes effectively stiffer. The robustness of the declining periodic density pattern and the spacing distribution also has limits, however, as in the case of strong effective nucleosome interaction or correlations in the positioning landscapes that are aligned with the boundary [44].

As a side note, the declining periodic density pattern at a boundary so central to many nucleosomal positioning studies, can be eliminated completely in theoretical models by a suitable external potential near the boundary leading to a completely constant nucleosome density throughout a bounded system (see Section B.2).

## 1.4 Chromatin remodeling

### 1.4.1 General role of chromatin remodelers

The aim of this section is to present some existing experimental approaches to investigate chromatin remodeling enzymes. This will be useful to explore the qualitative behavior of theoretical remodeler systems in Chapter 3 and to model the effective dynamics of promoter nucleosome configurations in Chapter 4.

A study by Zhang et al. [24] shows the effects of chromatin remodeling enzymes on in vitro assembled nucleosomes, using salt gradient dialysis. In this method purified core histones are incubated with DNA strands, e.g. a yeast plasmid library, in a buffer with high starting concentration of NaCl. As the salt concentration is reduced, the screening of electrostatic interactions decreases and nucleosomes spontaneously assemble on the DNA. Depending on the amount of available histones very closely packed nucleosomes can be achieved [45]. The nucleosome positions on this reconstituted chromatin are usually much less regular than in vivo. When aligned with respect to the in vivo N+1 or TSS position, one usually finds only a very shallow NDR and almost no gene averaged array pattern with N+1, N+2,... peaks [24]. Zhang et al. then showed that the in vivo pattern can be partly restored in vitro by incubating the assembled chromatin with whole cell extract and adenosine triphosphate (ATP) providing evidence that "biochemical reconstitution of proper nucleosome positioning, spacing, and occupancy levels was achieved across the 5' ends of most yeast genes by adenosine triphosphate-dependent trans-acting factors" [24].

These ATP-dependent trans-acting factors that act on nucleosomes in vivo are chromatin remodeling complexes (remodelers). They convert the energy of ATP hydrolysis to move, assemble, disassemble or restructure nucleosomes. These enzymes belong to the Snf2 family of ATP-dependent DNA and RNA helicases (Figure 1.5), thus are able to translocate on DNA like helicases. They track along one strand of the DNA double helix in a 3' to 5' direction, however, they do not separate the DNA strands [46]. Five subfamilies of ATP-dependent chromatin remodeling factors (SWI/SNF, ISWI, INO80/SWR1, CHD/Mi2 and ATRX) have been discovered so far, with each remodeler subfamily named after the first complex discovered in that subfamily [47]. To provide an idea of the "world" of remodelers, the following sections will present some properties of the SWI/SNF and the ISWI subfamily. Note that there have been many studies in investigating the mechanics of remodelers and their effects on nucleosomes (reviewed in [47, 48, 46, 49, 50]) but fully understanding the detailed dynamics of nucleosome remodeling, especially in a multi nucleosome setting remains a challenge.

Regarding the effect of remodelers on the nucleosome pattern, in a follow-up study of [24], Krietenstein et al. achieved a similar in vitro reconstitution of the in vivo pattern without whole cell extract but using only pure proteins: yeast genomic DNA, histones, sequence-specific general regulatory factors Abf1/Reb1, and remodelers RSC, ISW2, INO80, and ISW1a [52].

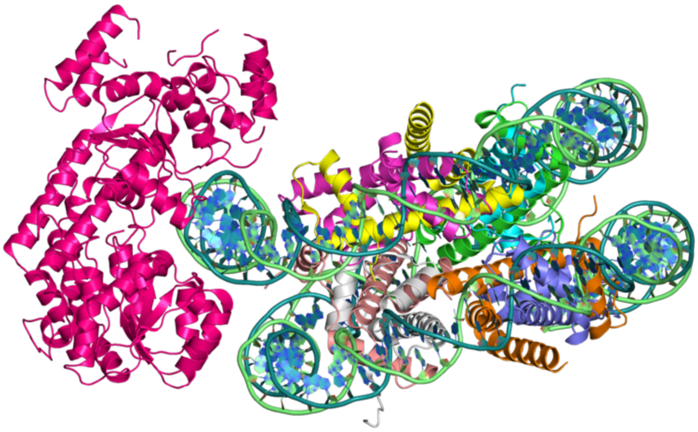


**Figure 1.5:** Cryo-EM reconstruction of *S. cerevisiae* Snf2 ATPase domain (left in pink) in complex with a nucleosome (right). Remodeler complexes consists of several other additional domains and can be significantly larger than a nucleosome [47]. From: Wikimedia Commons under public domain, original source: EMBL-EBI PDBe and [51].

To study the remodeler effects in vivo, a typical experimental approach is to knock out one or more remodelers in a mutated strain and monitor the possibly changed nucleosome positioning and occupancy, given that the new mutated strain is still viable. In vivo knock-out experiments for Isw1 and Chd1 remodelers (in yeast) show that regular positioning of the majority of nucleosomes is lost, especially in coding regions. Exceptions include the region upstream of the promoter, the +1 nucleosome, and a subset of locations distributed throughout coding regions where other factors are likely to be involved [53]. The ISW1 and CHD1 double knock-out shows a very strong reduction of the regularity of the declining periodicity pattern for the gene averaged occupancy [53]. Note that remodelers are redundant to some degree, as just knocking out the ISW1, ISW2 or CHD1 remodelers does not have such a strong effect (CHD1 seems to be the one with the least redundancy here though). Gkikopoulos et al. end their study with the remark: “Given the substantial defects to chromatin after deletion of ISW1 and CHD1, it is perhaps surprising that this strain survives reasonably well. Our data indicate that substantial transcription is possible in the absence of correct nucleosome spacing within coding regions. Chromatin organization within open reading frames may have a more important role in tuning the sensitivity and kinetics of transcriptional responses rather than as an obligate requirement.” A similar later study by Ocampo et al. [54] is based on the fact that the three known yeast nucleosome spacing enzymes (CHD1, ISW1 and ISW2) form arrays with different spacing in vitro. After analyzing the in vivo knock out strains they suggest that “CHD1 directs short spacing, resulting in eviction of H1 and chromatin unfolding, whereas ISW1 directs longer spacing, allowing H1 to bind and condense the chromatin. Thus, competition between the two remodelers to set the spacing on each gene may result in a highly dynamic chromatin structure.”

Another question is whether the mentioned spacing activity of some remodelers depends on the nucleosome density. In one-dimensional gas models with extended particles that do not attract each other for example, the average distance between neighbors increases with reduced particle density. However, different in vitro and in vivo studies presented evidence for constant nucleosome spacing even for reduced histone density, named “clamping activity” [55]

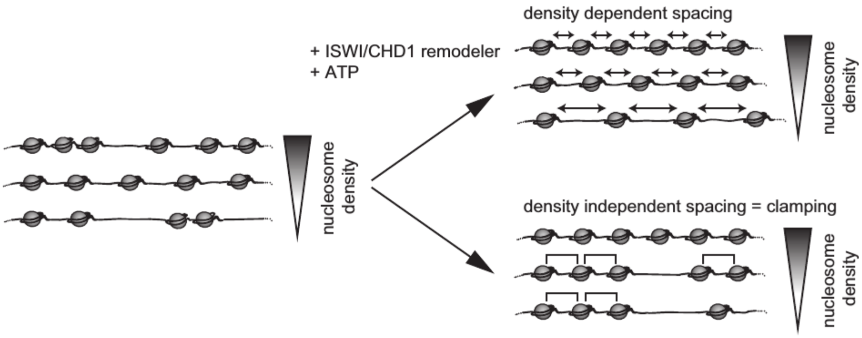


**Figure 1.6:** Two theoretical scenarios emerge in remodeling assays with varied nucleosome density. In density dependent spacing, the nucleosomal repeat length scales reciprocally with nucleosome density (double arrows), while density independent spacing ("clampling") leads to similar distances between neighbors (square brackets). In the case of ISWI/CHD1 remodeling, the study [55] by Lieleg et al. provides evidence for the latter scenario. From: [55], reprinted with permission from ASM.

(Figure 1.6). Lieleg et al. [55] showed in a purified system, that "ISWI- and CHD1-type nucleosome remodelers have a clamping activity such that they not only generate regularly spaced nucleosome arrays but also generate constant spacing regardless of nucleosome density. This points to a functionally attractive nucleosome interaction that could be mediated either directly by nucleosome-nucleosome contacts or indirectly through the remodelers."

### 1.4.2 SWI/SNF remodeler subfamily

The SWI/SNF subfamily is named after the SWI/SNF remodeler (SWItch/Sucrose Non-Fermentable). In yeast there is only one other family member, the RSC remodeler (Remodels the Structure of Chromatin). The subfamily contains homologs of the yeast SWI2/SNF2 ATPase with typically 8 to 15 subunits, which is quite large. Both remodelers function as monomers and play roles in activation and repression of genes [56].

Different remodeler families have different abilities. In contrast to the ISWI subfamily, the SWI/SNF subfamily remodelers "do not require a minimal length of linker DNA to move nucleosomes and can displace adjacent nucleosomes" [47]. Additionally, "SWI/SNF and related remodelers (...) can push the histone octamer beyond one end of a short piece of DNA, transfer histone octamers or create unusual dinucleosomal species in vitro, functions that ISWI remodelers, for example, normally lack" [46].

Remodelers do not always only act on one nucleosome at a time, as nucleosome disassembly by SWI/SNF requires a minimum of two nucleosomes on the same DNA template: "As the SWI/SNF complex moves one nucleosome on DNA toward the second nucleosome, it pushes against the second nucleosome until the DNA is displaced and spooled into the mobilized nucleosome. The second nucleosome is eventually displaced as more DNA is actively displaced from its surface and spooled into the other" [47]. Similar results have been found for RSC [48].

RSC is involved in keeping promoters nucleosome free [46] and Lorch et al. [57] found that RSC activity can be influenced by specific DNA sequences: "the AT-rich sequences present in many NFRs [i.e. nucleosome free regions] have little effect on the stability of nucleosomes. Rather, these sequences facilitate the removal of nucleosomes by the RSC chromatin remodeling complex.”

Epigenetic markers also play a role in chromatin remodeling, for instance acetylation of histone H3 tails increases binding as well as remodeling rate for yeast SWI/SNF and RSC remodelers [47]. While many remodeler studies are based on yeast, remodelers also play a very important role in human biology, for example “in human SWI/SNF, a frequently used method for regulating the complex in a tissue-specific manner is to switch particular subunits” and different diseases (mainly cancer) are related to different mutated subunits of the human SWI/SNF [47].

Regarding the mechanical properties of the RSC remodeler action, Sirinakis et al. [58] characterized the real-time activity of a minimal RSC translocase motor on bare DNA using high-resolution optical tweezers. On dsDNA, they observed “a processivity of about 35 bp, a speed of about 25 bp/s, and a step size of 2.0 (±0.4, s.e.m.) bp. Surprisingly, the motor is capable of moving against high force, up to 30 pN, making it one of the most force-resistant motors known.” Additionally, RSC and SWI/SNF have dramatically increased pulling force when remodeling nucleosomes compared to free DNA [47].

### 1.4.3 ISWI remodeler subfamily

The ISWI subfamily contains homologs of the *Drosophila* ISWI ATPase (ISWI: Imitation SWI/SNF), which are typically 2 to 4 subunits large [56]. ISWI subfamily remodelers require a minimal length of linker DNA for mobilizing nucleosomes [59] and the ISWI remodeler shows a linker DNA length-dependent regulation of nucleosome movement to evenly space nucleosomes. There are several possible explanations for this observation [47]: “One is that the affinity of ISWI is reduced as the linker DNA length is shortened, causing the complex to fall off and search for new substrates with more appropriate lengths of linker DNA. (...) Another explanation is that the rate at which the helicase domain translocates along nucleosomal DNA is regulated by the length of the linker DNA.”

Yeast has three homologs, the ISW1a, ISW1b and ISW2 remodelers. ISW1a and ISW1b are recruited to the gene body, and ISW1b is an example of remodelers recruited by histone tail modifications, in this case trimethylation of H3 or H3K36me3 [59]. The Isw2 complex has been shown to shift nucleosomes on promoter sequences to inhibit transcription initiation [46]. The ISW1a remodeler in yeast preferably binds to nucleosomes with two linkers of about 33 bp without subsequent action. If bound when only one linker exists, the nucleosome is moved in direction of the linker until the other linker is long enough (33 bp) and can be bound to as well [60]. It has also been suggested that ISW1a acts as a ‘protein ruler’ by binding to two neighboring nucleosomes simultaneously (dinucleosome binding) [61]. ISW1b shows no such spacing activity [47]. Like ISW1a, the ISW2 complex slides in direction of bound linker DNA [62]. In experiments using mononucleosomes, i.e. single nucleosomes on short single DNA fragments, ISW2 complexes slide preferentially to the center of DNA [62].

The following two single-molecule FRET (fluorescence resonance energy transfer) studies investigated the effect of remodelers on mononucleosomes by leveraging the dependency of the

FRET signal on the distance between two light-sensitive molecular markers bound to DNA and histones, respectively. Deindl et al. [63] probed mononucleosome translocation by ISWI-family remodelers (ISW1b, ISW2 and Isw2p, the main subunit of ISW2, for yeast) and suggest the following remodeling mechanism: "DNA is first translocated toward the nucleosomal exit side by the ATPase domain, 1 bp at a time, generating strain on the entry-side DNA; after 7 bp of translocation, the strain becomes sufficiently strong to trigger an enzyme action at the nucleosomal entry side that draws DNA into the nucleosome; this action partially releases the strain and allows three additional base pairs of DNA to be translocated to the exit side; this 3 bp step then repeats to generate processive DNA translocation across the nucleosome.". Blosser et al. studied the human ACF (ATP-dependent Chromatin-assembly Factor) remodeler in single-molecule FRET assays, "revealing previously unknown remodelling intermediates and dynamics. In the presence of ACF and ATP, the nucleosomes exhibit gradual translocation along DNA interrupted by well-defined kinetic pauses that occurred after approximately seven or three to four base pairs of translocation. The binding of ACF, translocation of DNA and exiting of translocation pauses are all ATP-dependent, revealing three distinct functional roles of ATP during remodelling. At equilibrium, a continuously bound ACF complex can move the nucleosome back-and-forth many times before dissociation, indicating that ACF is a highly processive and bidirectional nucleosome translocase." [64].

# 2 Measurement and Analysis of Absolute Nucleosome Occupancy [1]

## 2.1 Introduction

As described in Section 1.2, even though the precision of mapping nucleosomes increased, current methods to measure nucleosome occupancy on a genome-wide scale give only relative but not absolute values. Naturally, absolute occupancy contains more information than relative occupancy, as it can be translated into absolute particle numbers and allows a direct probabilistic interpretation.

There are absolute occupancy measurements at single loci either based on the differential accessibility of nucleosomal versus nonnucleosomal DNA for restriction enzymes (REs) or DNA methyltransferases (DNMTs). REs have a specific binding motif, usually 4 bp or 6 bp long, and are able cut the DNA double strand, if the binding site is accessible. They have been used for RE accessibility measurements for example at the *PHO5* and *PHO8* promoter [65, 66]. DNMTs methylate bases of the double strand, usually cytosine or adenine, using S-adenosyl methionine (SAM) as the methyl donor. DNA methylation footprinting was developed using prokaryotic DNMTs methylating either CpG, i.e. CG, sites or GpC, i.e. GC, sites [67, 68, 69]. If cut and uncut, or methylated and unmethylated DNA is measured, the fraction of not cut or not methylated fragments of all fragments corresponds to the absolute occupancy. For correct measurement, it is important that the enzymes cut/methylate all accessible sites which is usually tested by comparing samples with different digestion duration and enzyme amounts. If the sample with longer digestion or more enzyme yields the same absolute occupancy, the RE- or DNMT-catalyzed reaction is saturated. Therefore it is crucial that the nucleosome dynamics is frozen, like for ex vivo-prepared or in vitro-assembled chromatin under physiological buffer and temperature conditions [70, 24]. In vivo, however, nucleosomes are assembled, disassembled and restructured by ATP-dependent nucleosome remodeling enzymes, possibly changing the nucleosome occupancy over time as well as generating transient DNA accessibility also within nucleosomes [47, 71].

There are already genome-wide methods using REs or DNMTs, but they have not yet provided reliable absolute occupancies. RE-based methods scored only cut fragments [72, 73] and DNMT-based methods, for example NOMe-seq [74], had the disadvantage of too low sequencing coverage and DNA methylation being either not saturated or chosen to match MNase-seq results [28]. In the following we present an RE-based as well as an DNMT-based method to measure absolute occupancy genome-wide and apply them to the *Saccharomyces cerevisiae* genome.

[1] Large parts of this chapter are adapted from our publication [28] under CC BY-NC 4.0 license.

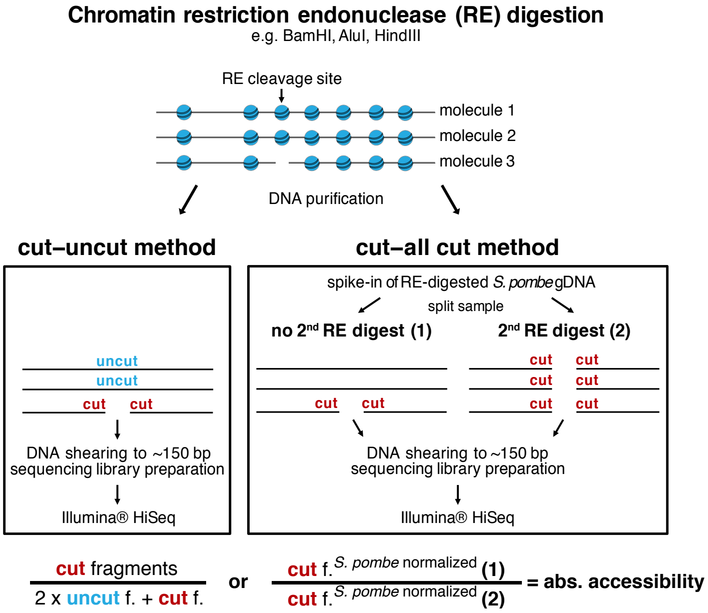


**Figure 2.1:** Method overview for genome-wide absolute occupancy measurement by restriction enzymes and high throughput sequencing (ORE-seq). Absolute accessibility is defined as 1 - absolute occupancy. The presented formulas for the cut-uncut and cut-all cut method are simplified and the full formulas and derivations can be found in Appendix A.

This interdisciplinary project was a close collaboration with the group of PD Dr. Philipp Korber[2] and lead to a publication in 2019 [28] which this chapter is closely based on. Elisa Oberbeckmann performed most of the presented experiments and Mark Heron was responsible for data analysis and analysis development during the first years of the project, which was continued by the author of this thesis starting 2017.

## 2.2 Absolute occupancy by restriction enzymes (ORE-seq)

### 2.2.1 Method development

We developed and tested the genome-wide measurement of absolute nucleosome occupancy using the wild type (WT) strain BY4741 of the budding yeast *Saccharomyces cerevisiae* during logarithmic growth in full media. After the ex vivo isolation of chromatin, it was digested at different RE concentrations and for different incubation times to ensure the RE digestion was saturated. We investigated two methods (Figure 2.1) to quantify the DNA molecules cut at the respective RE sites out of all molecules genome-wide. Both methods differ in the way the amount of all molecules at the RE sites is measured. In the first method, we aimed to quantify the ratio of cut out of all fragments, cut as well as uncut, at each RE site. Therefore, after the RE digest the DNA was purified and sheared to ca. 150 bp long fragment by sonication to allow high throughput Illumina sequencing. After mapping to the reference genome fragments

[2] Molecular Biology Division, Biomedical Center, Faculty of Medicine, Ludwig-Maximilians-Universität München

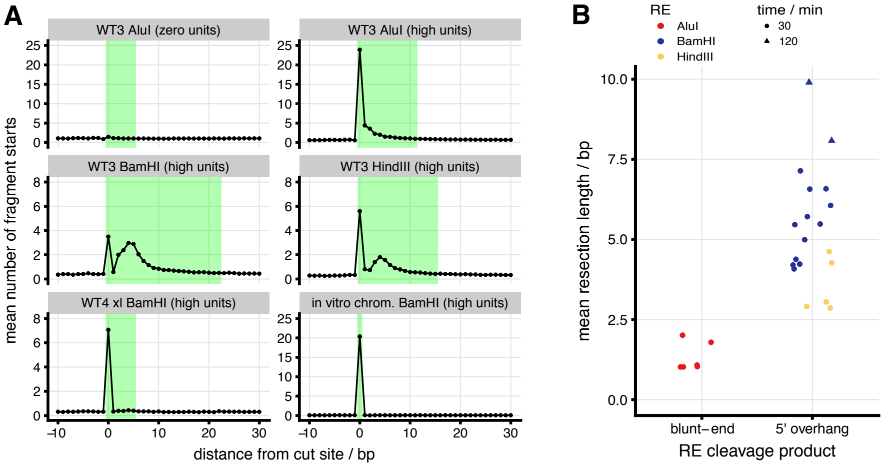


**Figure 2.2:** Analysis of RE cut site resection, i.e. unwanted shortening of fragments that were cut at RE sites. (**A**) Histograms of DNA fragments starting at the indicated distance downstream of the RE cut sites for the different ORE-seq samples (all 30 min incubations). Counts at a given distance were averaged over all RE sites of the sample. Read starts within green areas are counted as “cut by the RE” in the analysis (count windows, Section A.2). “xl” denotes crosslinked samples. Analyzing fragment ends near cut site shows the same behavior. (**B**) Mean RE cut site resection lengths, i.e. the mean distance from the cut site of all fragments starting/ending within the count windows, dependent on RE type and incubation time.

cut at RE sites and fragments covering RE sites without having been cut were counted. To correct for sonication breaks at RE sites, we also scored fragment ends in regions without RE sites (Section A.2).

Different REs have different motif sequences and also cut the DNA in different ways. Some, like AluI, cut both strands between the same base pairs at the center at the motif (blunt-end). Others, like BamHI and HindIII cut the + strand upstream of the motif center and the − strand downstream of the motif center (5’ overhang). See the following example for the 6 bp motif of HindIII, with ’||’ indicating the cut positions:

```
+ strand:  5’-...A||A G C T T...-3’
- strand:  3’-...T T C G A||A...-5’
```

3’ overhangs are possible as well. Depending on the RE, we measured different distributions of fragment end counts near RE sites (Figure 2.2A, high unit samples in row one and two). A likely reason were endogenous yeast exonucleases copurified with the ex vivo chromatin that resected RE cut ends during the treatment with REs. To score resected RE cuts as proper RE cuts as well, we used a window-based counting algorithm (Section A.2). Since resection was stronger for BamHI and HindIII (Figure 2.2B), we assumed the resecting exonucleases were likely 5’-3’ exonucleases, reducing predominantly 5’ overhangs. In a later test, we also crosslinked the chromatin with formaldehyde, a chemical introducing covalent bonds between DNA and proteins. We then saw no resection at all (Figure 2.2A, bottom left), likely because the endogenous nucleases present were inactivated by the crosslinking step.

Since each cut leads to two ends that can be detected independently, while each non-cut leaves one fragment, we calculated the ratio of fragments ending at or near this site over the sum of these fragments plus twice the fragments spanning over this site to obtain an estimate of the absolute accessibility at each RE site. The absolute occupancy is given by one minus the absolute accessibility. We tested this “cut-uncut method” with artificial mixtures of completely cut and uncut purified yeast genomic DNA (gDNA). As there are no nucleosomes or other proteins present anymore, gDNA is cut completely when digested with REs. For AluI, BamHI (short for BamHI-HF) and HindIII we mixed such predigested gDNA with undigested gDNA at 10%, 30%, 50%, 70% and 90% uncut fraction within pipetting accuracy. Additionally we tested undigested and fully digested gDNA. While the measurements of the completely uncut sample and the completely cut sample yielded values very close to 100% and 0% absolute occupancy (averaged over all sites), respectively, the other samples showed lower mean absolute occupancy values as prepared with the largest deviation for the 30%, 50% and 70% samples (Figure 2.3A, left graph). The average deviation over all samples was 7.6%. We suspected some technical bias towards DNA fragments with RE cut ends, effectively causing fragments with ends cut by sonication to be less likely sequenced than fragments cut by REs.

With a direct comparisons of cut and uncut fragments seeming problematic, we implemented the second method (“cut-all cut method”) that scores only the fragments with RE cuts (Figure 2.1, right). As a proxy to the number of all molecules we decided to use the number of cuts in a parallel sample, where the DNA was completely cut at all cut sites (“all cut sample”). After the digestion with RE and DNA purification, we split the samples into two halves. One of them was treated again with the RE, yielding the all cut sample. We counted the number of RE cut ends after sequencing as in the method before and absolute accessibility was measured as the ratio of RE cut ends in the half without over the half with second RE digest (details in Section A.3). As this is now the ratio of fragment counts of the same type, the bias from the previous method before should be circumvented. We also corrected for possible loss of material during parallel treatment of both halves by normalization to RE digested *S. pombe* gDNA spiked-in before splitting into the two halves. We tested this method for the same calibration samples as before and found that the agreement between measured and prepared mean absolute occupancy was now much better (2.5% average deviation, Figure 2.3A, center graph). While the measured mean values were much closer to the prepared value, the standard deviation between sites (error bars in Figure 2.3A) was increased compared to the cut-uncut method at lower prepared occupancies, indicating a much larger spread of individual values at the sites. This was caused by the estimation of the accessibility by the ratio of two on average similarly large (for low occupancy) cut counts from independent samples (the two halves), while the cut-uncut method only uses one sample, having the mathematical advantage of a standard deviation decrease towards 0% as well as towards 100% occupancy.

Having seen that we could measure the correct prepared mean absolute occupancies with the more involved cut-all cut method, we wanted to improve the previous cut-uncut method by introducing a correction factor for each RE to eliminate the bias towards RE cut fragments (“corrected cut-uncut method”). This correction factor was fitted for each RE to minimize the deviation from the prepared occupancy (Figure 2.3B, see Section A.4) leading to an average deviation of 1% (Figure 2.3A, right graph). Since the best fit values of the correction factor did not deviate much between REs, we additionally did a combined fit using AluI, BamHI and HindIII samples together to obtain a correction factor to be used for other REs later on. Unless specified otherwise, we used the corrected cut-uncut method for occupancy measurement via restriction enzymes and high throughput sequencing (ORE-seq).

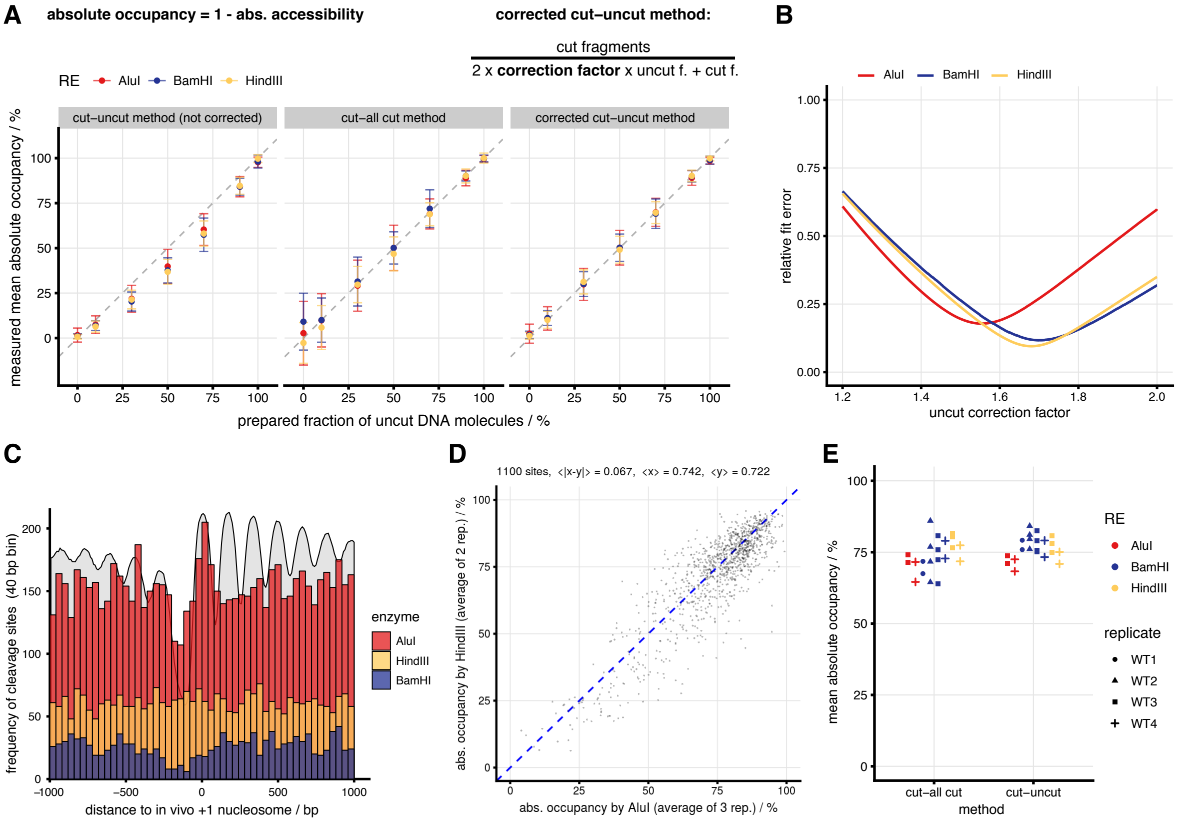


**Figure 2.3:** Calibration and testing of ORE-seq measurements. (**A**) Calibration curves for three variants of genome-wide detection of RE accessibility and absolute occupancy. Completely cut and non-cut yeast gDNA preparations were mixed to produce the indicated fraction of uncut DNA (x-axis), which was then measured as mean absolute occupancy by the indicated methods. Error bars correspond to the standard deviation between sites. (**B**) Dependency of the relative fit error for the corrected cut-uncut method (A, right plot) on the correction factor for each RE. A relative fit error of 1 corresponds to the uncorrected cut-uncut method in A (left plot) (**C**) +1 nucleosome-aligned site distributions with 40 bp bins with the final ODM-seq absolute occupancy map to indicate the average nucleosome positions (grey background, Figure 2.11B). (**D**) Correlation plot of ORE-seq data at HindIII sites obtained from AluI versus HindIII (HindIII cut sites are a subset of AluI cut sites). <x> and <y> denote the mean absolute occupancy measured by AluI or HindIII, respectively, over 1100 compared sites. The HindIII map was obtained by averaging 2 replicates at each site, the AluI map by averaging 3 replicates. (**E**) Comparison of mean absolute occupancy values always after 30 min incubation time and for low and high RE concentrations but analyzed either by the cut-all cut or the corrected cut-uncut method.

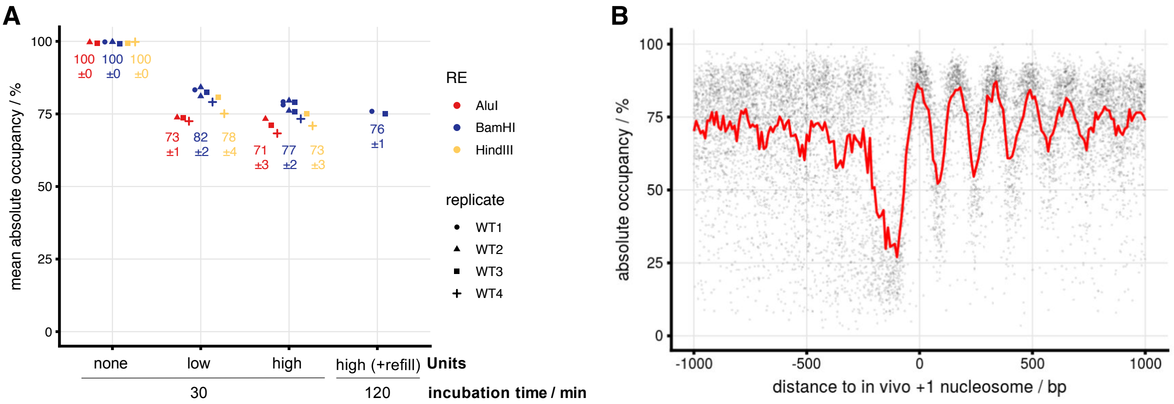


**Figure 2.4:** ORE-seq absolute occupancy results. (**A**) Absolute occupancy averaged over all sites (mean absolute occupancy) obtained by ORE-seq for different biological replicates (WT1 to WT4) under the indicated conditions of RE concentration (Units) and incubation time. For longer incubation, fresh enzyme was refilled after 60 min (indicated with refill). Numbers give average values ± standard deviation of replicates. (**B**) Absolute occupancy at genomic cut sites aligned to in vivo +1 nucleosome positions. Each dot represents the value of one genomic site and the red line shows the 10 bp bin mean occupancy of aligned sites. To calculate the ORE-seq map we averaged the occupancy values at the same sites in all valid high unit RE samples (AluI, BamHI and HindIII) with equal weights.

### 2.2.2 ORE-seq with ex vivo yeast chromatin

When applying ORE-seq to biological replicates of ex vivo yeast chromatin with AluI, BamHI and HindIII, we found good reproducibility. Additionally, each replicate clearly saturated during RE digestion, as the mean absolute occupancy values for each RE were within five percent points for samples with increasing concentrations (low versus high units) or incubation times (Figure 2.4A). We measured mean absolute occupancy plateau values (averaged mean absolute occupancies using the high units samples) from 71% to 77% for different REs with the different values possibly being caused by the different distribution of RE sites with different motifs along the genome (Figure 2.3C, Table A.3). We also exploited that each 6 bp HindIII site contains a 4 bp AluI site as a control by comparing the absolute occupancy measured by HindIII and AluI at HindIII sites. Averaging over several replicates, we obtained 7% absolute difference for the average site and 2% difference in mean absolute occupancy averaged over all sites (Figure 2.3D). We checked that the mean absolute occupancy values measured by the corrected cut-uncut versus the cut-all cut method centered around the same plateau (Figure 2.3E) leading to the conclusion that the correction factor derived from the gDNA calibration samples could also be used for ex vivo chromatin. The mean absolute occupancy values varied less between replicates for the corrected cut-uncut than for the cut-all cut method (Figure 2.3E). This was another reason so stick with the corrected cut-uncut method. Since our newly developed genome-wide method is a successor of single site methods based on Southern blotting, we also wanted to compare the results of ORE-seq at these singe sites. Therefore we analyzed DNA purified from some ex vivo samples after the same RE digestion as for ORE-seq also by classical secondary cleavage and Southern blotting for occupancy at a BamHI site in the *PHO5* promoter and a HindIII site in the *PHO8* promoter. The occupancy values measured by ORE-seq versus Southern blotting agreed well within 7% for the BamHI site, but differed by 20% for the HindIII site (data not shown here, see [28]). The difference at the HindIII site may have been due a an unusually large estimation error that can occur at individual sites with larger standard deviation between replicates. For ORE-seq the standard deviation between replicates averaged over sites was 5-6% (Table A.3). We then used quality criteria like saturation of digestion and sequencing coverage at each RE site (Section A.4) to select and combine absolute occupancy data from different samples into a genome-wide ORE-seq map (Figure 2.4B) with low average resolution of ca. 870 bp.

In the course of this project we also used additional REs, like HhaI, another HindIII variant named HindIII-HF and a combination of BamHI-HF with KpnI, but did not repeat the calibration described above and instead used the combined correction factor fit value. For this reason we did not include them into the final ORE-seq map. Their mean absolute occupancy values for WT replicates, however, were within the range of AluI, BamHI-HF and HindIII (Figure 2.5A, B). The crosslinked samples showed increased mean absolute occupancy and we will discuss this when comparing ORE-seq results with the methylation based method described in the next section. As mentioned before, crosslinked samples showed no resection, independent of the RE (Figure 2.5C, now also for the additional single REs). Furthermore we analyzed a mutant strain that undergoes a histone H3/H4 depletion after a medium switch [75] and were able to report a strong absolute occupancy decrease compared to its non-depleted state at the level of our WT strains (Figure 2.5B, D). Another mutant strain previously reported to show histone depletion and reduced nucleosome occupancy, a yeast *nhp6a/b* mutant [76], however, did not give lower mean absolute occupancy than its wild-type variant or our WT replicates (Figure 2.5B, E).

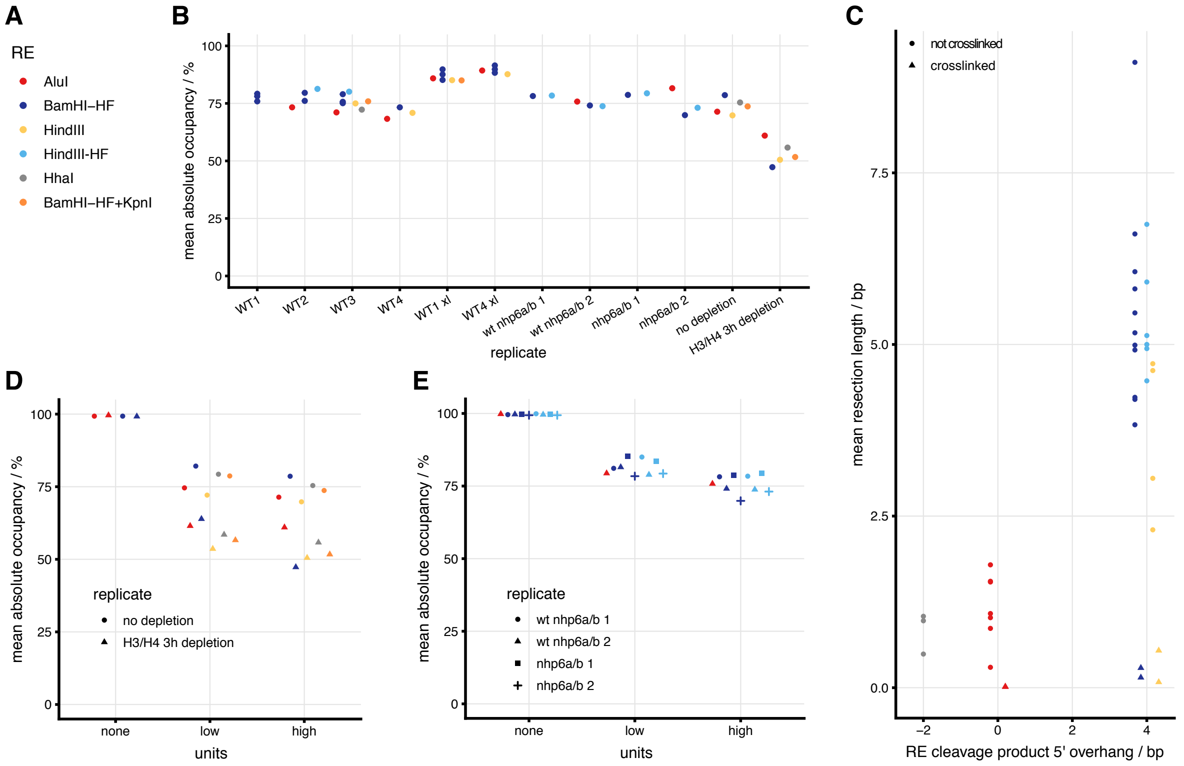


**Figure 2.5:** Additional ORE-seq results including more REs and mutant strains. (**A**) Colors indicate the used restriction enzymes throughout this figure. Here samples with restriction enzymes HindIII-HF, HhaI and BamHI combined with KpnI are included. For these additional RE, we did not have calibration samples and we used the correction factor that optimized the combined fit error of the Alu, BamHI-HF and HindIII calibration samples for the cut-uncut method. In the combined BamHI and KpnI samples, the sites of both REs were analyzed independently (e.g. independent count window length) except all sites where considered when determining the distance to next neighbor sites for filtering and the background region away from all sites. (**B**) Mean absolute occupancy of all replicates including crosslinked WT replicates, *nhp6a/b* wild type and mutant replicates, as well as a histone depletion strain with and without histone depletion. (**C**) As Figure 2.2B, but only 30 min samples and including additional REs and crosslinked samples. (**D**) As Figure 2.4A but for the histone depletion strain. (**E**) As Figure 2.4A but for the *nhp6a/b* wild type and mutant replicates. For strain information see Section A.1.

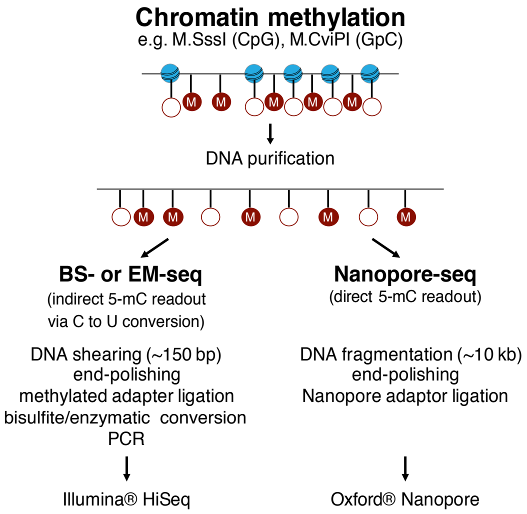


**Figure 2.6:** Method overview for genome-wide absolute occupancy measurement by DNA methylation and high throughput sequencing (ODM-seq) with its variants using bisulfite conversion after methylation (BS-seq), enzymatic conversion after methylation (EM-seq) and direct methylation calling (Nanopore-seq). In BS-seq and EM-seq only unmethylated cytosines are converted to uracil and can be detected as “a wrong base” after the read is mapped to the reference genome.

## 2.3 Absolute occupancy by DNA methylation (ODM-seq)

### 2.3.1 Method development

To improve the low ORE-seq resolution and to complement it with an orthogonal method, we sought to measure genome-wide absolute occupancy by differential cytosine methylation at position C-5 in CpG or GpC motifs. We detected methylated cytosines mostly by treatment with bisulfite, converting only unmethylated cytosines to uracil followed by Illumina sequencing of approx. 150 bp sonicated fragments (Figure 2.6). Thus, only the methylated cytosines in CpG or GpC motifs remain unchanged. In Illumina sequencing after a PCR (polymerase chain reaction) step, uracil bases are detected as thymine (or adenine on the complementary PCR strand instead of a guanine), since uracil pairs with adenine during the PCR. Special mapping tools have been developed to align these manipulated fragment sequences to the reference genome [77, 78, 79], allowing the analysis of the methylation states of the CpG or GpC sites (also see Section A.5). As an enzymatic alternative to bisulfite, we also used EM-seq (enzymatic methyl sequencing) for the conversion step. Additionally, we used a direct methylation readout of long DNA fragments in Oxford Nanopore sequencing without PCR with Nanopolish [80], which has additional benefits that we will discuss in Section 2.5.

In a first step we tested if DNA methylation saturates at a plateau or if it eventually invades nucleosomes (“overmethylation”). In case of saturation we also compared the plateau value to ORE-seq results. For a direct comparison, chromatin reconstituted in vitro by salt gradient dialysis (SGD) with fly histones for a yeast whole genome plasmid library [81] was used. In contrast to ex vivo chromatin, this was nuclease free (Figure 2.2A, lower right) and consisted only of canonical nucleosomes, which allowed us to investigate REs and DNMTs specifically regarding their action towards nucleosomes. Furthermore this enabled us to vary the DNA to

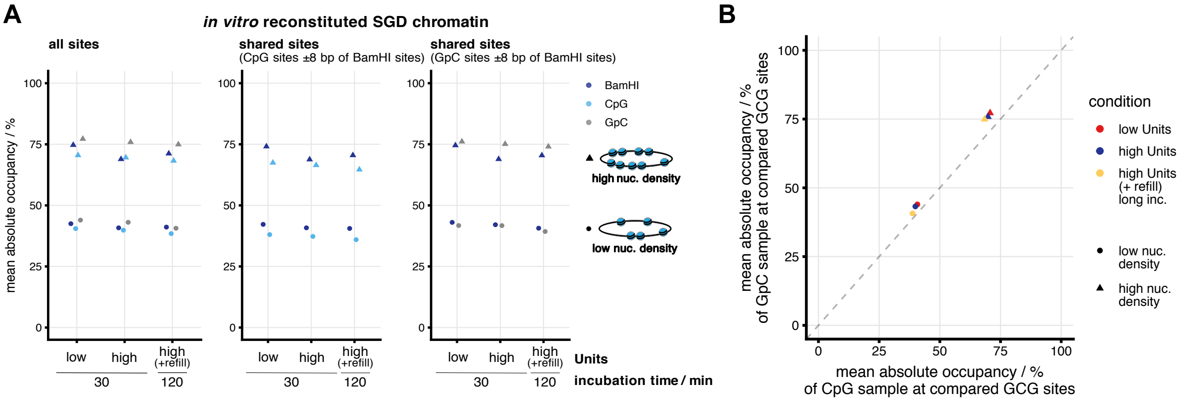


**Figure 2.7:** In vitro reconstituted chromatin measured with ORE- and ODM-seq. (**A**) Comparison of mean absolute occupancy averaged over the indicated site subsets between ODM-seq (BS-seq) using M.SssI (CpG) or M.CviPI (GpC) and ORE-seq using BamHI for genome-wide in vitro reconstituted salt gradient dialysis (SGD) chromatin of high or low nucleosome density and for different enzyme concentrations (Units) and incubation times. For longer incubation, fresh enzyme and for DNMTs fresh SAM (methyldonor S-adenosyl methionine) was refilled after 60 minutes (+refill). (**B**) Mean absolute occupancy data of GpC vs CpG methylated SGD chromatin as in panel A, but only at GCG sites.

histones ratio during reconstitution and to test the RE and DNMT results at two different, yet previously unknown, absolute nucleosome occupancies. To allow $Mg^{2+}$-dependent RE cleavage, we added 1.5 mM $MgCl_2$ to the buffer pioneered for DNMTs in the Kladde group [82]. Bisulfite conversion rate and coverage were used as quality criteria to discard low quality reads and sites, respectively (Section A.5).

Testing high as well as as low nucleosome density SGD chromatin, we measured the mean absolute nucleosome occupancy by the RE BamHI, the CpG DNMT M.SssI and the GpC DNMT M.CviPI. All showed saturation at very similar plateaus with mean difference between lowest and highest value of 5% and maximal difference of 7%. The results did not change if all cut/modified sites were compared (Figure 2.7A left graph) or only BamHI sites close to CpG (Figure 2.7A, center graph) or GpC (Figure 2.7A, right graph) methylation sites. The GpC DNMT systematically gave on average 5% higher mean absolute occupancy values than the CpG DNMT at the same GCG sites (Figure 2.7B). Still, we showed that both RE and DNMT approaches crossvalidate each other within acceptable error range. Thus, both could be equally used to measure genome-wide absolute nucleosome occupancy. We named the new DNMT-based method “ODM-seq” (occupancy measurement via DNA methylation and high throughput sequencing).

However, when applying this method to ex vivo prepared yeast chromatin, we were not able to achieve saturation or reached plateau values much lower than with ORE-seq unless magnesium was added or the chromatin was formaldehyde crosslinked, where proteins on the DNA are covalently bound to the DNA (Figure 2.8A). All RE samples included magnesium since it is needed by REs to properly function, so this was the first time we investigated the absolute occupancy of an ex vivo prepared sample without extra magnesium, with the intend to inhibit $Mg^{2+}$-dependent endogenous yeast exo- and endonucleases, which copurify with ex vivo chromatin and may interfere with the analysis, especially during long incubations.

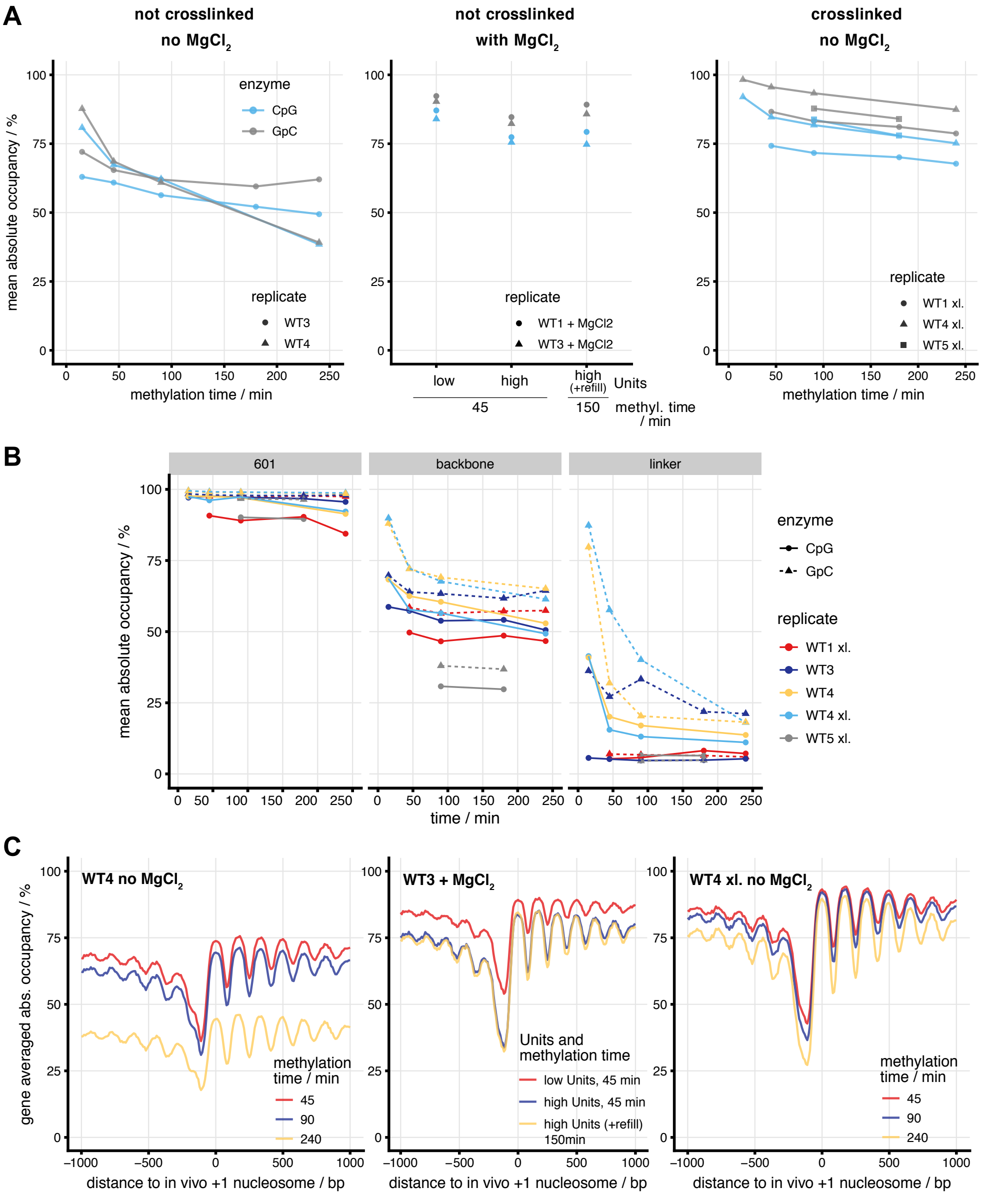


**Figure 2.8:** Ex vivo chromatin measured with ODM-seq with or without $MgCl_2$ and crosslinking. (**A**) Mean absolute occupancy of different replicates (WT1, WT3-WT5) under the indicated conditions. "xl." denotes in vivo formaldehyde crosslinked samples. (**B**) As panel A but for indicated regions on the 25mer 601 array plasmid SGD chromatin spiked into the indicated ex vivo chromatin replicates. 601 regions are expected be fully occupied, while linker regions easily accessible. Spike-in was not from the same preparation, i.e. occupancy can vary, except the same spike-in chromatin was used for CpG and GpC methylation of each replicate and the same for the comparison of ± crosslinking for WT4. (**C**) +1 nucleosome aligned gene averaged absolute occupancy measured for the indicated ex vivo chromatin replicates, conditions and time points corresponding to panel A.

All samples included a spike-in of SGD assembled plasmids with a 25mer array of the Widom 601 nucleosome positioning sequence [83] which generates predictable nucleosome positions during SGD [55]. The highly repetitive array forced us to map the spike-in reads to a single linker with ending and starting 601 reference sequence instead of the real 25mer plasmid sequence, but we still could calculate the occupancy at the CpG or GpC sites (Figure A.5). Sites well within the linker and well within the 601 sequence were each averaged to single occupancy values for linker and 601 sequence. Additionally the sites on backbone of the plasmid were averaged to a single backbone occupancy value. Occupancies of the spike-in chromatin mostly saturated at occupancy values of the expected relative magnitude: high for sites in the 601 sequence, where nucleosomes preferentially assemble, intermediate for backbone sites, which belong to nucleosomal and non-nucleosomal regions, and low for linker sites, where nucleosomes are expected to be scarce (Figure 2.8B). The respective spike-in plateaus could differ between experiments since different SGD chromatin preparations were used. We saw that the GpC DNMT was sometimes less efficient in reaching saturation, especially in linker methylation and decided to use complete methylation of 25mer 601 array linker DNA (less than 20% occupancy) as an internal control for saturating methylation.

We concluded that diffusible nucleases or proteases stemming from the ex vivo chromatin were not the reason for the low or not saturating absolute occupancy in the ex vivo chromatin without magnesium (Figure 2.8A, left), since these would also have affected the spike-in. The spike-in results together with those obtained for genomic plasmid library SGD (Figure 2.7A) also argued against overmethylation by DNMTs. As overmethylation and enzymatic degradation of nucleosomes was ruled out, we conjectured that yeast endogenous nucleosomes may be intrinsically unstable during prolonged incubation due to their known structural properties [84] and differ in that aspect from nucleosomes generated with fly histone octamers for the SGD assemblies.

The presence of $Mg^{2+}$ during methylation, however, led to saturation and mean absolute occupancies similar as via ORE-seq, at least for the CpG DNMT, while the GpC DNMT yielded higher occupancy values (Figure 2.8A, center) but also smaller differences between linker and nucleosome occupancy in N+1 aligned composite plots (Figure 2.8C, center) than the ORE-seq map (Figure 2.4). Further investigation showed that, as suspected, the added magnesium activates endogenous exo- and endonucleases that depending on the replicate lead to more or less strong bias in measured occupancy in low occupancy regions of long incubated samples. For a full discussion of this problem please the supplemental material of our paper [28].

To stabilize the yeast nucleosomes as well as inhibit endogenous nucleases we turned to in vivo formaldehyde crosslinking and verified saturation of methylation as well as mean absolute occupancy plateaus very similar to the ORE-seq values for most samples. The GpC methylation of the crosslinked WT4 replicate was not saturating according to the occupancy values for the linker region of the 25mer 601 SGD spike-in chromatin (Figure 2.8B) and was excluded from further analysis. We also did not find any sign of nuclease activity in crosslinked samples [28] and variation of crosslinking time between one and 20 minutes did not change the results (Figure 2.9A). When measuring samples with combined $Mg^{2+}$ and formaldehyde crosslinking with ORE-seq or ODM-seq, we found absolute occupancies with higher values than just with $Mg^{2+}$ or formaldehyde treatment alone (Figure 2.9B). We suspect this could be due to overcompaction, decreasing enzyme accessibility also via higher order chromatin compaction.

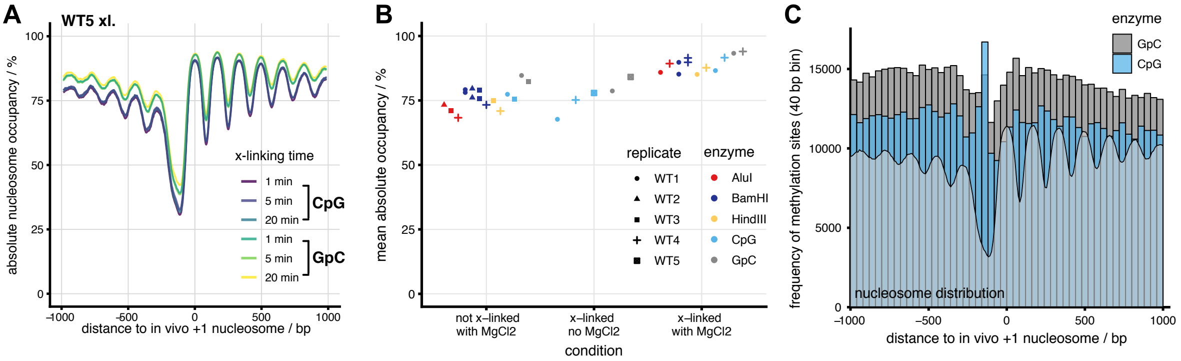


**Figure 2.9:** Additional results for ODM-seq. (**A**) +1 nucleosome aligned gene averaged absolute occupancy for different crosslinking times. (**B**) As Figure 2.4A but including ODM-seq samples, crosslinked (x-linked) chromatin preparations and only for 30 min incubation time and high RE concentration. (**C**) +1 nucleosome-aligned site distributions of DNA methylation sites with 40 bp bins and ODM-seq absolute occupancy map (grey background, Figure 2.11B)

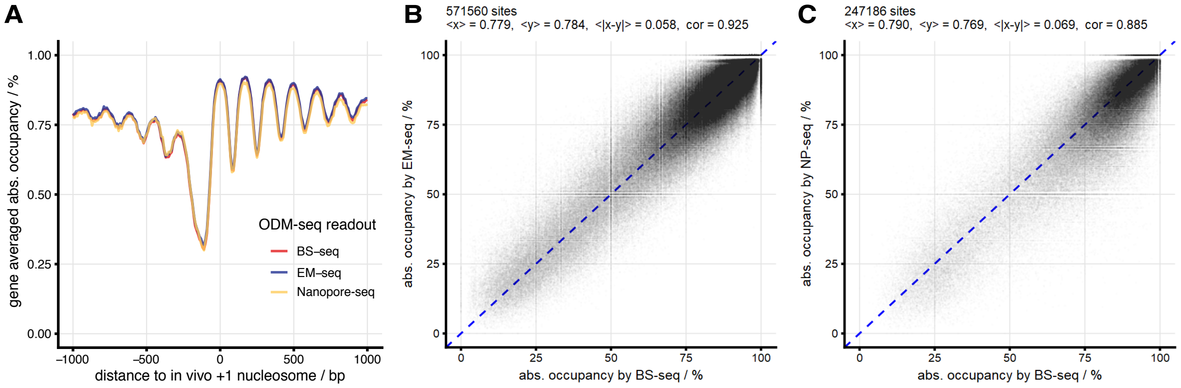


**Figure 2.10:** Comparison of different CpG ODM-seq readouts for WT5 (crosslinked). (**A**) N+1 aligned gene averaged absolute occupancy with different readouts methods. (**B**) Correlation plot of absolute occupancy values of all CpG sites measured by EM-seq vs BS-seq (**C**) As in B, but Nanopore-seq (short NP-seq) vs BS-seq. Since Nanopolish groups sites closer than 11 bp (Section A.5), close sites where also grouped and averaged in BS-seq for this plot.

Parallel to searching and finding a robust biochemical protocol for methylation incubation without nucleosome loss or overcompaction we investigated alternative readout methods to bisulfite followed by Illumina sequencing (BS-seq). EM-seq, the enzymatic alternative to bisulfite as well as direct readout of 5mC in long DNA fragments by Nanopore sequencing (Nanopore-seq) gave almost identical N+1 aligned composite plots for the same chromatin sample (Figure 2.10A). Comparing the absolute occupancies measured at each site lead to very good correlations of 92.5% between BS-seq and EM-seq (Figure 2.10B) and 88.5% between BS-seq and Nanopore-seq (Figure 2.10C) which are slightly higher than the 87.9% correlation between plus and minus strand of the same sample in BS-seq (Figure A.6). This high agreement controls against any systematic errors in the bisulfite sequencing and bioinformatics pipeline. We could also exclude a contribution of DNA from unlysed cells as we would have seen resulting unmethylated long DNA fragments in Nanopore-seq. REs or DNMTs are not able to access DNA from unlysed cells and in short read methods like ORE-seq, BS-seq or EM-seq, the DNA fragments of unlysed cells would be indistinguishable from cells with high occupancy on the fragment, possibly leading to systematically distorted measurements.

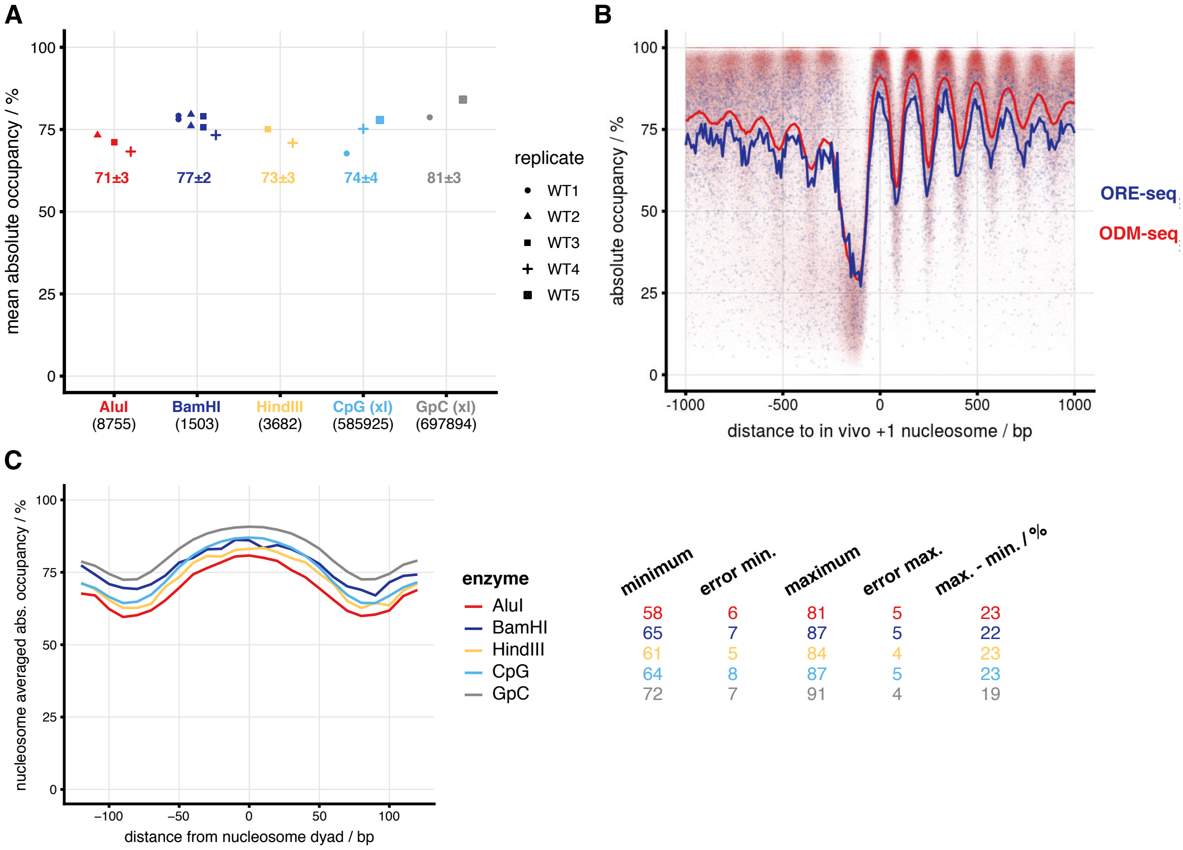
### 2.3.2 ODM-seq with crosslinked ex vivo chromatin and comparison with ORE-seq

To sum up, we had absolute occupancy measurements across the yeast genome for two independent methods (ORE-seq and ODM-seq) involving two different conditions (non-cross-linked chromatin in RE buffer with $Mg^{2+}$ vs. cross-linked chromatin in DNMT buffer without $Mg^{2+}$). We used five independent enzymes (AluI, BamHI, HindIII, M.SssI (CpG DNMT), M.CviPI (GpC DNMT)), five independent biological replicates (WT1 to WT5), and obtained a comparison between BamHI and the two DNMTs for purified in vitro chromatin (Figure 2.7A). All these independent and partially orthogonal measurements showed mean absolute occupancy in the range of 71% to 81% (Figure 2.11A). It was crucial that this set of approaches yielded values in a similar range to cross-validated each other, as there was no genome-wide precedent for such absolute occupancy values.

We combined the results of the CpG and GpC samples presented in (Figure 2.11A) to an ODM-seq map and compared it with the ORE-seq map (Figure 2.11B). The ORE-seq map tended to yield lower mean occupancy than the ODM-seq map (Figure 2.11B), which seemed to be mainly caused by the higher occupancy measured by GpC methylation (Figure 2.11A). Comparing the mean occupancy measured by the different enzymes in a pairwise fashion, we found that the errorbars (standard deviation over samples) overlapped in most cases, except for example GpC compared with AluI and HindIII. Of course, mean absolute occupancy values across the genome are dependent on the different positions of sites for each enzyme in nucleosomes and nonnucleosomal regions (Figure 2.3C and Figure 2.9C).

In order to correct for such different site positions, we plotted absolute occupancy values for each enzyme as averaged over 10 bp bins around the dyads of all called nucleosomes[3] mapped

[3] The chemical cleavage data allows a very accurate measurement of nucleosome dyad positions but does not give absolute occupancy values. The average nucleosome positions ("called/typical nucleosomes") can be obtained by peak calling algorithms applied to the cell-averaged chemical cleavage signal. Also see Section 2.4.1.



| enzyme | minimum | error min. | maximum | error max. | max. - min. / % |
|---|---|---|---|---|---|
| AluI | 58 | 6 | 81 | 5 | 23 |
| BamHI | 65 | 7 | 87 | 5 | 22 |
| HindIII | 61 | 5 | 84 | 4 | 23 |
| CpG | 64 | 8 | 87 | 5 | 23 |
| GpC | 72 | 7 | 91 | 4 | 19 |

**Figure 2.11:** Genome-wide absolute occupancy measurement by restriction enzymes and DNA methyltransferases. (**A**) Mean absolute occupancy obtained by ORE-seq (Supplemental Table S3 of [28]) or ODM-seq (Supplemental Table S4 of [28]) for the indicated enzymes and biological replicates at saturation conditions. The number of sites implemented for each enzyme is indicated in parentheses. "xl." indicates in vivo formaldehyde crosslinked samples. As Figure 2.9B, but only non-crosslinked ORE-seq samples with $MgCl_2$ as well as crosslinked ODM-seq samples. (**B**) Absolute occupancy from ORE-seq and ODM-seq maps (averaged occupancy values over all included samples) and aligned at in vivo +1 nucleosome positions. Each dot corresponds to the value of one genomic site, and the lines show the 10 bp bin mean occupancy of aligned sites. (**C**) Absolute occupancy values averaged in 10 bp bins around nucleosome dyads called from chemical cleavage-seq data [26] and averaged over all replicates for the indicated enzymes. On the right, absolute occupancy values and errors (mean over sites in the bin of the standard deviation among samples) are shown for the maxima and minima of each plot as well as the difference between maximum and minimum values for each enzyme. See Figure 2.13C for the distribution of occupancy values of the ODM-seq map at the same called dyad positions (averaged values within ±20 bp of the dyads).

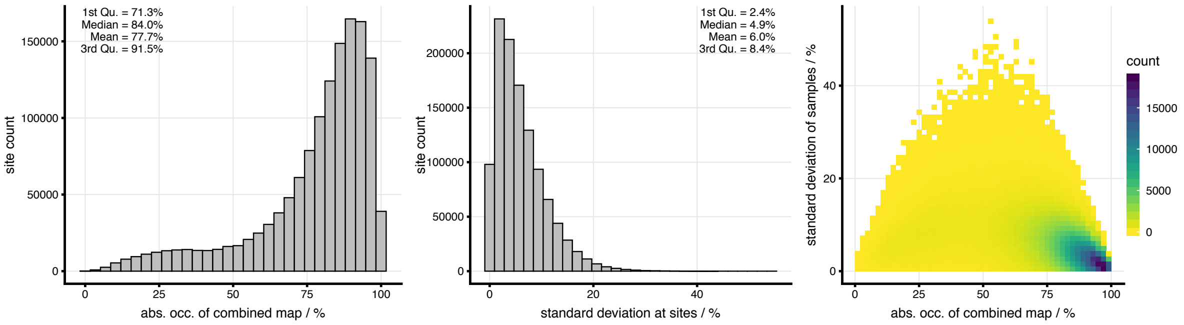


**Figure 2.12:** ODM-seq map site statistics. Left: histogram of absolute occupancies of ODM-seq map at DNA methylation sites. Center: histogram of standard deviation between replicates at DNA methylation sites. Right: correlation plot (color indicates number of occurrences) for map absolute occupancy versus standard deviation of samples at DNA methylation sites.

by chemical cleavage [26] (Figure 2.11C). For this we used individual enzyme maps, where the occupancy values of all valid samples of a given enzyme are averaged at each site. This illustrates how each enzyme measured the absolute occupancy near the nucleosome dyads, the nucleosome flanks and the linker region. Even though the maximum occupancy values varied from 81% to 91%, the differences of maximum and minimum were very similar (ca. 23%) except for GpC. The errorbars at maximum and minimum (mean over sites in the bin of the standard deviation among samples) overlapped for most pairwise comparisons, leaving us confident that we obtained an accurate measure for absolute occupancy at low resolution using REs and at high resolution using DNMTs.

## 2.4 Applications of the ODM-seq absolute occupancy map

### 2.4.1 Absolute nucleosome occupancy

The restriction enzymes enabled us to cross-validate the absolute occupancy values of the ODM-seq results, but they did not contribute many sites, less than 1% compared to all DNMT sites (Figure 2.11A, Figure 2.13A and Table A.3). So we decided to just use the ODM-seq absolute occupancy map with a mean resolution of 9 bp and an average error of 6% (mean over sites of standard deviation between samples) (Figure 2.12, center and right; Table A.3) in the following analysis. The +1 nucleosome-aligned composite plot of the ODM-seq map (Figure 2.11B) is very similar to those using MNase-seq, MNase-anti-histone-ChIP-seq [24, 23] or chemical cleavage-seq data [26] (Figure 2.13B). However, our absolute occupancy map provides meaningful peak heights in absolute terms, whereas the other methods show more fluctuations in nucleosome peak heights (Figure 2.13A) and do not agree, for instance whether the average relative occupancies for the +1 and +2 nucleosomes are the same, or one of the two is higher than the other. The ODM-seq map now shows that, the first three nucleosomal peaks in the N+1 aligned plot have almost the same absolute occupancy.

Our measurements revealed generally high occupancy, with a median site occupancy of 84% (Figure 2.12, left) but also a substantial fraction of sites with lower occupancies. Neither ORE-seq nor ODM-seq are able to distinguish which bound factors contributes to the measured

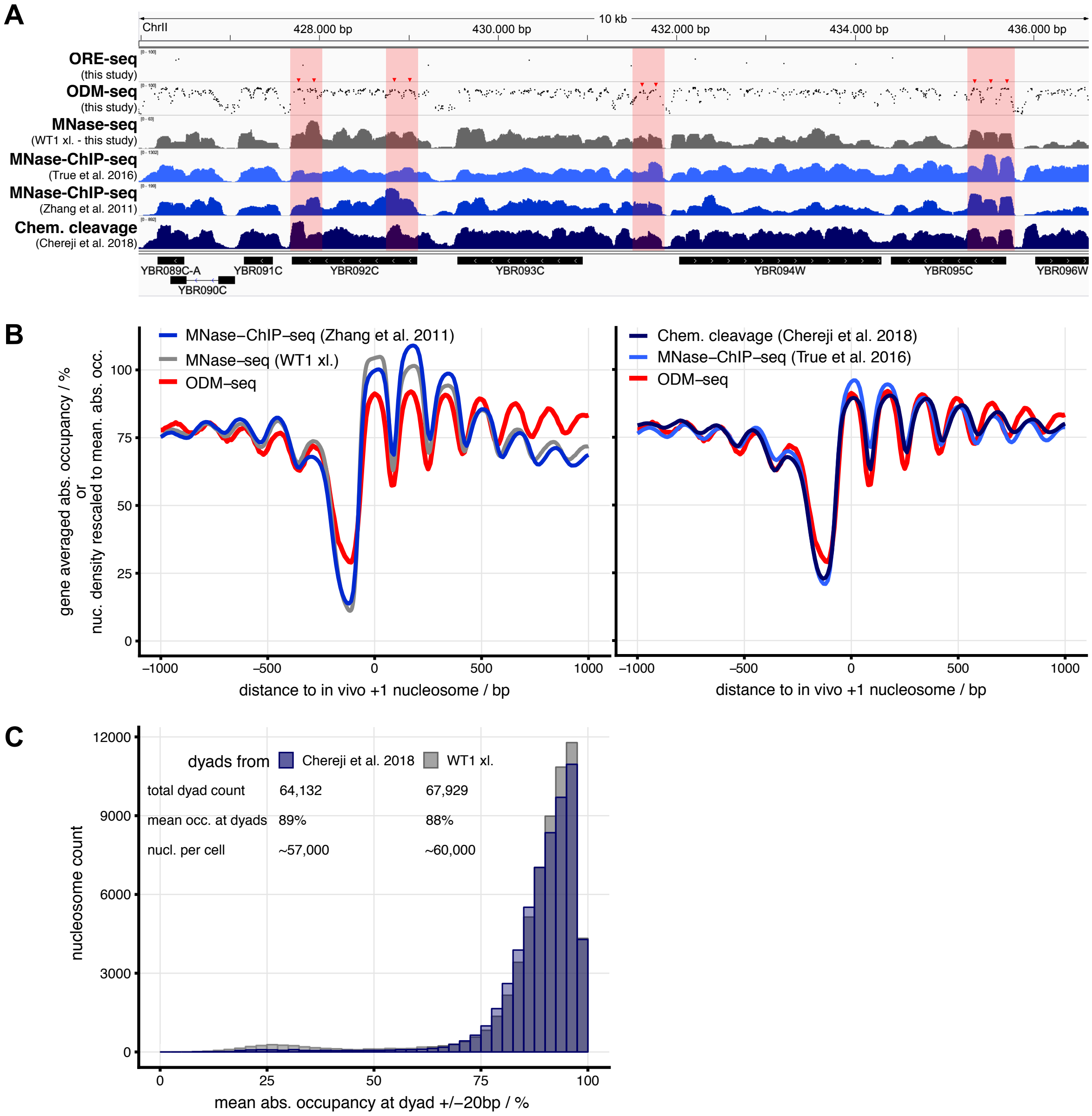


**Figure 2.13:** Comparison of the ODM-seq absolute occupancy map with other occupancy measurements. (**A**) Integrated Genome Viewer (IGV) browser shots comparing different data sets with our ORE-seq and ODM-seq absolute occupancy data. (Details in Supplemental Table S2 of [28], external data: True et al. 2016 [23], Zhang et al. 2011 [24], Chereji et al. 2018 [26].) Regions in light red highlight pronounced differences in occupancy/signal between methods. (**B**) N+1 aligned gene averaged absolute occupancy ODM-seq map and other indicated data sets. The lines show the 10 bp bin mean occupancy of aligned sites. Because the external data do not provide absolute occupancy, we globally rescaled their signal to have the same genomic mean as the absolute occupancy map. Nucleosome dyads of external data sets were extended to 147 bp. (**C**) Histogram of absolute occupancy at nucleosome positions called from chemical cleavage [26] and our own MNase data set.

occupancy. To remedy this, we combined our data with nucleosome-specific mapping data to obtain an absolute nucleosome occupancy map, similar as to in Figure 2.11C. We now used two nucleosome dyad data sets: we called dyads in our own MNase-seq data set for the crosslinked WT1 sample (using fragments with typical nucleosome footprint length only) or used nucleosome dyad cluster medians (typical nucleosomes) called from chemical cleavage-seq (kindly provided by Razvan Chereji) [26]. For each of the called dyads position, we averaged the absolute occupancy of sites within ±20 bp and the corresponding histograms (Figure 2.13C) for MNase-seq or chemical cleavage-seq show rather narrow distributions with means of 88% and 89%, respectively. Using these values and the total dyad count of each dyad data set we calculated the absolute number of nucleosomes in a yeast cell as approx. 60 000 for MNase-seq calling and approx. 57 000 for chemical cleavage-seq called dyads. The different values likely stemmed from method-specific limitations for scoring nucleosomes. MNase-seq can score nonnucleosomal complexes in nucleosome free regions (NFRs) as “nucleosomes” [85], which might cause an overestimation of the total nucleosome number and a slight bump in the histogram around 25% occupancy which is not present for chemical cleavage-seq called dyads (Figure 2.13C). Conversely, some nucleosomes in gene bodies are not called properly in the chemical cleavage based data set and the number of nucleosomes is underestimated [28].

### 2.4.2 Absolute occupancy of nonnucleosomal DNA-bound factors

When comparing the histograms of absolute occupancy at CpG/GpC sites (Figure 2.11C, left) and absolute nucleosome occupancy (Figure 2.13C) we find a significant population of CpG/GpC sites with occupancy below 50%, but almost no nucleosomes with such low occupancy. These sites are mostly positioned in NFRs and linkers (Figure 2.11B) and especially NFRs are probably occupied by nonnucleosomal factors (Figure 2.16A) thus preventing the absolute occupancy to reach zero. Similar to using called nucleosome dyads from external data sets and investigating the absolute occupancy map at the dyad positions, we also used external data sets with the positions of DNA binding factors like the general regulatory factors (GRFs) Rap1, Abf1 and Rep1 measured with SLIM-ChIP (short-fragment-enriched, low-input, indexed MNase ChIP; a high resolution transcription factor mapping protocol) [86]. We were able to detect absolute occupancy peaks for Abf1 and Rap1, but not Reb1 (Figure 2.14A-D) which might be due the binding site motifs and different factor footprints not covering CpG/GpC sites. These GRF peaks were not present in MNase-seq or chemical cleavage signals (Figure 2.14B, right), as expected, but the surrounding regularly spaced nucleosome arrays were similar compared with ODM-seq. These arrays were much less regular and at lower occupancy around Rap1 sites, than around Abf1 and Reb1. Rap1 sites also showed a broader distribution relative to the TSSs (Figure 2.16A) and their sites often come in neighboring pairs, impeding the array visualization in site aligned composite plots. We also investigated absolute GRF occupancy for Mcm1, Cbf1 and Orc1, using PWM hit sites in [28].

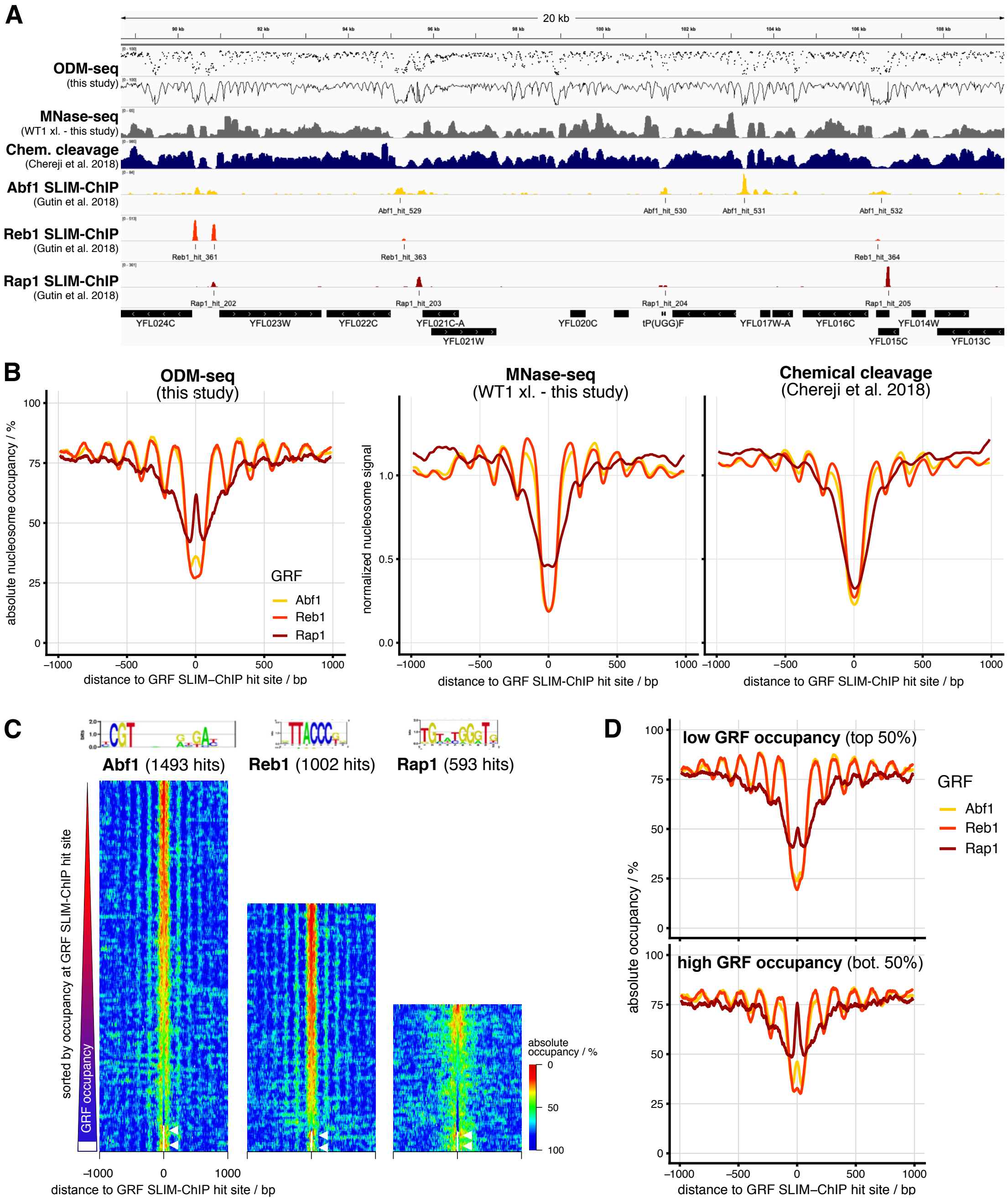


**Figure 2.14:** ODM-seq measures not only absolute nucleosome but also absolute GRF occupancy. (**A**) IGV browser shot comparison of different data sets (details in Supplemental Table S2 of [28], external data: Chereji et al. 2018 [26], Gutin et al. 2018 [86]). ODM-seq data are given both as individual (top) and as connected data points (second from top). (**B**) GRF site aligned and averaged plots of absolute occupancy (left) or normalized signal (center and right) for the indicated data sets, where signals are normalized to a mean of one. (**C**) GRF site aligned heat maps of absolute occupancy sorted from top to bottom according to increasing absolute occupancy at GRF sites. The position weight matrix [87] and the number of binding sites detected by SLIM-ChIP for the indicated GRFs is given above the heat maps. White color denotes absence of signal (highlighted by white triangles). (**D**) As in B, left graph, but for genes of low and high GRF occupancy according to the SLIM-ChIP sorting in C.

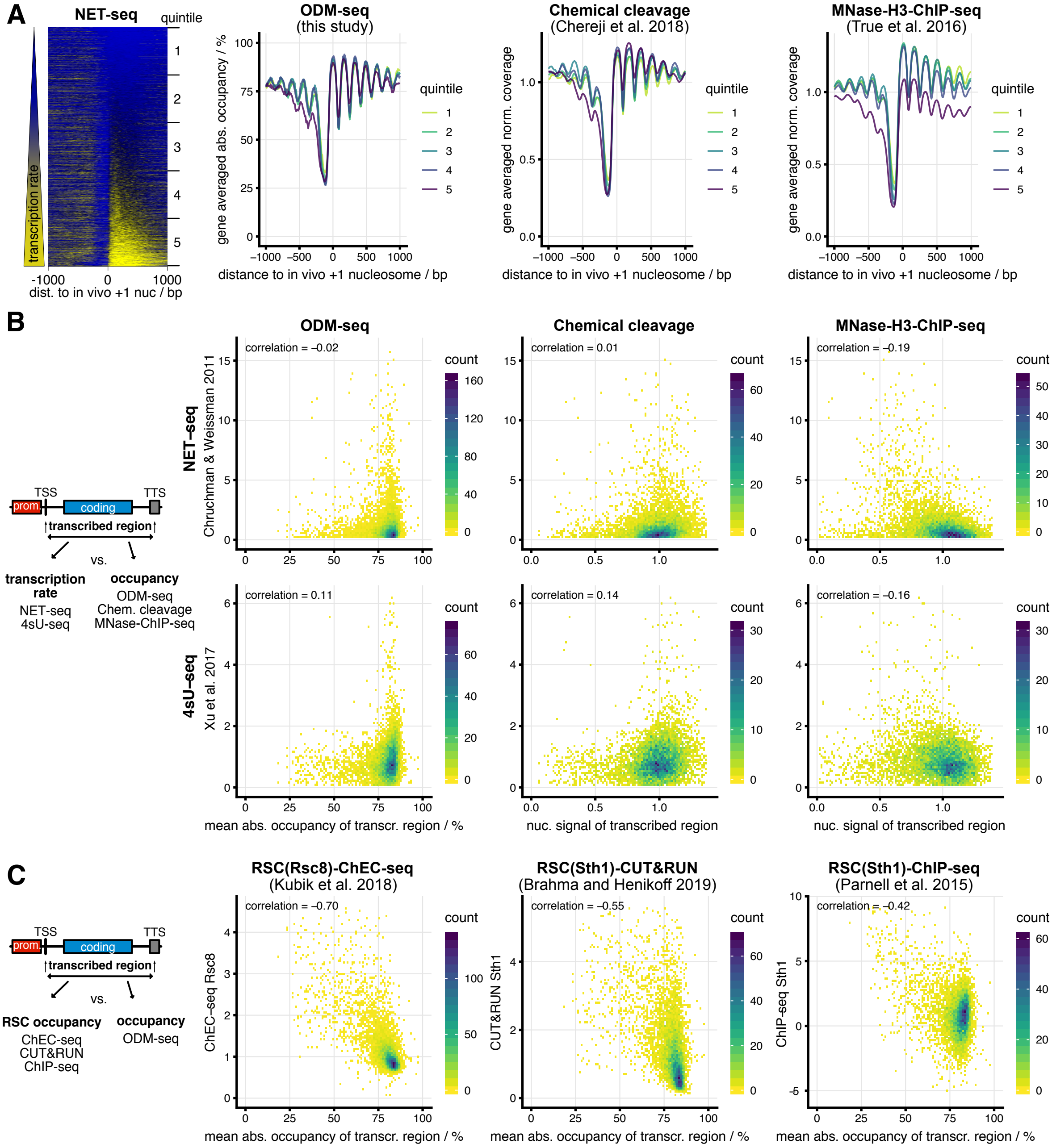


**Figure 2.15:** Correlation of absolute occupancy with biological features. (**A**, left) In vivo +1 nucleosome aligned heat map of NET-seq data monitoring nascent RNA bound to RNA polymerase [88] sorted from top to bottom by increasing signal over the gene body. (Right) As in Figure 2.13B but for the indicated data sets and genes subdivided according to quintiles of sorting in heat map on the left. (**B**) Correlation plots (color indicates number of occurrences) of transcription rate (NET-seq as in A or 4sU-seq [Xu et al. 2017]) against the absolute occupancy or coverage averaged over transcribed regions for the indicated data sets as in A. (**C**) As in B but correlation of absolute occupancy averaged over transcribed regions with RSC binding measured by the indicated methods. External data: Chereji et al. 2018 [26], True et al. 2016 [23], Churchman and Weissman [88], Xu et al. 2017 [89], Kubik et al. 2018 [90], Brahma and Henikoff 2019 [91], Parnell et al. 2015 [92], details in Supplemental Table S2 of [28].

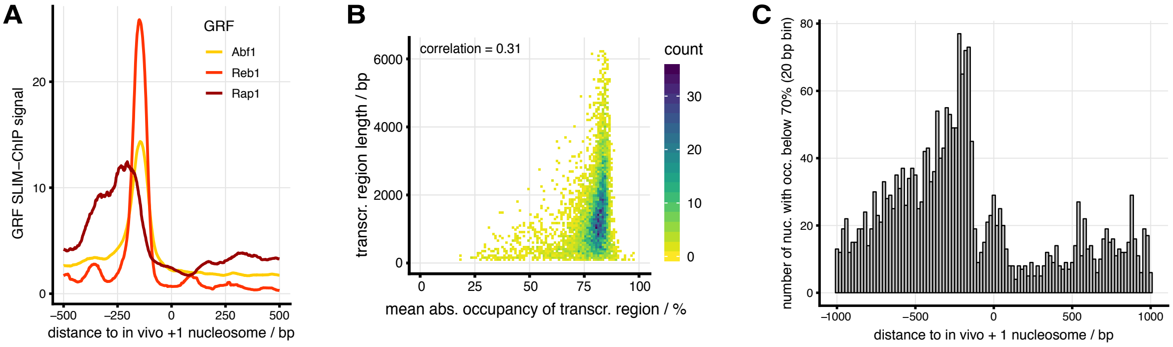


**Figure 2.16:** Correlation of absolute occupancy with biological features (part 2). (**A**) +1 nucleosome-aligned composite plots of SLIM-ChIP data for the indicated GRFs. (**B**) Correlation of average ODM-seq absolute occupancy on transcribed regions with region lengths. (**C**) +1 Nucleosome-aligned histogram (accumulated in 20 bp bins) of nucleosomes dyads called by chemical cleavage-seq with $< 70\%$ absolute occupancy.

### 2.4.3 No correlation of absolute occupancy with transcription

97% of all nucleosomes called by chemical cleavage mapping have an absolute occupancy greater or equal than 70%, with a mean ± standard deviation of $(90 \pm 6)\,\%$ (Figure 2.13C). With such a flat occupancy landscape we did not expect strong correlations between absolute occupancies and biological features. Nevertheless, we investigated which biological feature correlated with absolute occupancy.

Previously, very high expression levels were reported to correlate with nucleosome loss over gene bodies, for example for heat shock-induced genes [93]. As measure for transcription activity, we used NET-seq data (native elongating transcript sequencing, detecting nascent RNA bound to RNA polymerase) [88] to calculate gene quintiles and compared +1 nucleosome aligned composite plots of absolute occupancy, chemical cleavage signal [26] and MNase-H3-ChIP-seq signal [23] and other mapping data (see [28]) for each quintile (Figure 2.15A). The MNase-H3-ChIP-seq signal was reduced for the most highly transcribed genes, but this was much less the case for chemical cleavage-seq and ODM-seq. Furthermore, for each gene, we correlated the occupancy signal averaged over the transcripted region with the NET-seq signal and the gene signal of another method, which labels new RNA with the uridine analog 4-thiouridine (4sU-seq) [89]. We found no anti-correlation when occupancy was measured by ODM-seq or chemical cleavage-seq and only very poor anti-correlation for MNase-H3-ChIP-seq (Figure 2.15B).

We also checked for a correlation between the transcribed region length and absolute occupancy. Long genes showed mostly high absolute occupancy, while shorter genes assumed a larger range of absolute occupancies (Figure 2.16B).

### 2.4.4 Correlation of absolute occupancy decrease with RSC remodeling complex binding

The chromatin remodeling complex RSC is a major nucleosome displacing factor in yeast [94, 95] (Section 1.4.2) and mainly depletes nucleosomes upstream of the TSSs [96, 52]. We

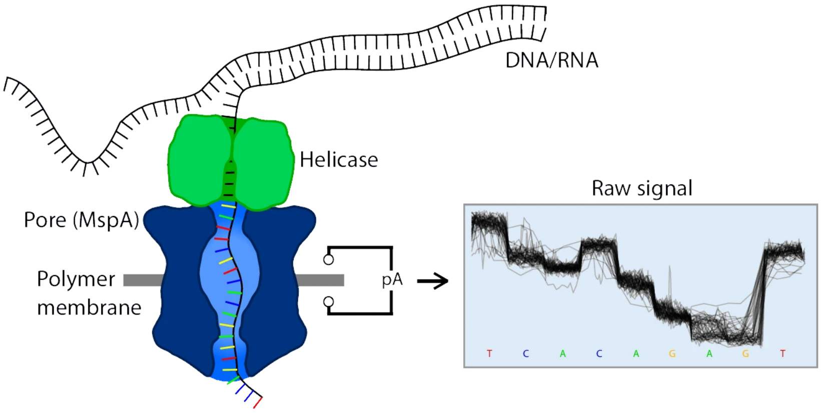


**Figure 2.17:** Nanopore sequencing principle. The DNA/RNA double strand is separated by a helicase and one strand guided through the nanopore, e.g. *Mycobacterium smegmatis* porin A (MspA). The ion flow through the channel is changed by the nucleotides in the sensing region. Using a basecalling algorithm, the raw current trace is converted to a nucleotide sequence. Adapted from: [97] under CC BY 4.0 license.

tested three different RSC mapping data sets and found inverse correlation between the RSC signal and the absolute occupancy, both averaged over individual gene regions (Figure 2.15C). Additionally, nucleosome dyads called by chemical cleavage with lower than 70% absolute occupancy were mainly upstream of the TSSs (Figure 2.16C). To sum up, absolute nucleosome occupancy is constant for nucleosomes across the yeast genome unless RSC depletes nucleosomes, which happens mostly in regulatory regions upstream of TSSs.

## 2.5 Nucleosome detection on long reads

This section focuses on nanopore sequencing, which we already used in the previous sections of this chapter to, in addition to normal base sequencing, directly call methylations (Nanopore-seq). Here, we want to start leveraging the fact that this method allows the sequencing of much longer reads, than for example the standard Illumina sequencing. First, we begin with a short introduction to nanopore sequencing in general.

### 2.5.1 Nanopore sequencing

Nanopore sequencing has been investigated since the 1980s in several laboratories. [98] provides a good historic overview which will be quickly outlined in this paragraph. The main idea of nanopore sequencing is that while a polymer, e.g. DNA or RNA single strand, is translocated through a nanopore from one side of a membrane to the other side, the ionic current between the two membrane sides is influenced by the nucleotides within the nanopore sensing region, i.e. the region of the pore with the smallest diameter (Figure 2.17). Early setups used pore proteins found in biology like the $\alpha$-hemolysin from *Staphylococcus aureus* and later engineered mutants of *Mycobacterium smegmatis* porin A (MspA) with a pore diameter of approximately 1 nm each and a sensing region length of 5 nm and 0.6 nm, respectively. A

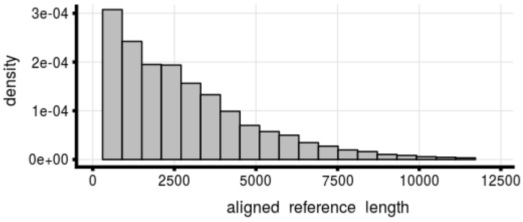


**Figure 2.18:** Mapped reference lengths (bp) in the nanopore ODM-seq WT4 xl CpG 180min sample show the mapped read length distribution.

smaller sensing region reduces the number of nucleotide affecting the current level. From 2005 on all four DNA nucleobases could be distinguished from each other in immobilized strands, but only by 2012 the use of controlled translocation mechanisms made continuous sequencing possible. In 2014 one of the first available nanopore sequencing devices, the very compact (90 g weight) MinION device from Oxford Nanopore Technologies (ONT), used 512 pores in parallel to achieve high throughput sequencing of DNA strands with lengths of several thousands base pairs. Since there is no DNA synthesis step needed as in other high-throughput sequencing technologies, the nanopore sequencing itself does not set a maximal read length and thus later even read lengths up to 1 000 000 bp were reported, with a sequencing speed of 450 bp/ sec using R9.4 flow cells [99]. Basecalling accuracies ranged from initially 66% to up to later 95% depending on the exact pore chemistry, basecalling algorithm and application. Later ONT launched also larger devices with even more pores sequencing in parallel for laboratory use. More and more refined basecalling algorithms have been developed, including deep learning methods [100, 101]. As outlined in [102], basecalling accuracy still remains one of the disadvantages of nanopore sequencing when compared with short read sequencing methods with more than 99% accuracy. Additionally detecting homopolymers with the correct length is challenging, since in these cases the raw signal often does not change enough to properly "count" the number of nucleotides passing through the pore.

However the possibility to sequence very long reads has far reaching benefits. For instance, it allows de novo genome assembly with up to 99% accuracy when polishing the draft assembly with Nanopolish [103] using the raw data. With short read methods, obtaining the genomic sequence of highly repetitive regions, like the rDNA loci of *S. cerevisiae*, is problematic. Another advantage of the lack of a DNA synthesis step is the direct observation of base modifications such as methylations [104, 80] (and more recently [105, 106]) as well as the direct sequencing of RNA molecules containing uracil bases [99]. Direct long read RNA sequencing is now also used to uncover the diversity of alternative splicing isoforms and their expression levels, but with higher error rates than for DNA sequencing [99]. Very recent studies report methods to distinguish N6-methyladenosine from unmodified adenosine on long reads [107, 106, 108] which leads to a drastic resolution increase compared to CpG and GpC methylation sites.

### 2.5.2 Methylation calling with Nanopolish

In the following we will focus on the ability to detect the 5-methylcytosine (5-mC) modification using Nanopolish (updated for methylation calling in [80]), in order to detect accessible and

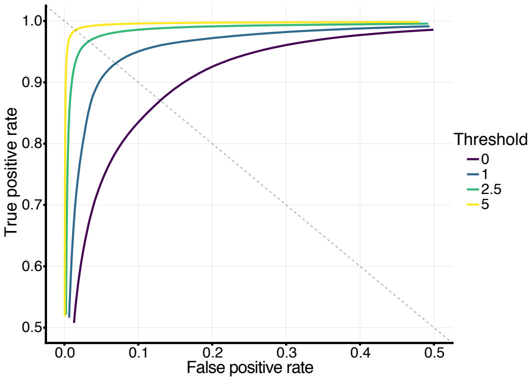


**Figure 2.19:** ROC curves for methylation calling with Nanopolish using two test data sets: a completely unmethylated and a completely methylated sample. Lower calling thresholds for the logarithmic likelihood ratios result in lower ROC curves.

occupied regions on single long reads. With Nanopore-seq, the analyzed read lengths easily extend 2000 bp (Figure 2.18), much more than the typical 50 bp of paired-end BS-seq, allowing to observe the occupancy for regions corresponding to several neighboring nucleosomes.

Nanopolish uses a hidden Markov model to detect methylated CpG sites with model parameters fitted to training sets. Its output is a log likelihood ratio for each CpG site on each read and then a threshold is used to make a call at a given site: absolute log likelihood values below the threshold are called “ambiguous” and these site are ignored to improve the overall accuracy [80], while the remaining positive (negative) log likelihood ratios are called “methylated” (“unmethylated”). We used our own test data sets of untreated as well as completely methylated DNA to test the accuracy (Figure 2.19). According to a preprint [109] from the lab that developed Nanopolish, it is also able to call CpG as well as GpC methylation sites simultaneously, using the right training data, but we did not succeed in finding good quantitative agreement for GpC sites when trying this with our own test sets (data not shown).

The long read lengths allow the analysis of highly repetitive sequences, where short reads could not have been mapped accordingly. As an example consider the 25mer 601 plasmid spike-in. It consists of 12 identical sections with the same 601 and linker sequence and 13 sections of slightly changed 601 and linker sequences (the changes were done to introduce restriction enzyme motifs that were not used in this project). To map the 50 bp long reads from bisulfite ODM-seq, we used one generic 601-linker-601 sequence to map all reads to one linker section with neighboring 601 sequences (Figure A.5). The much longer nanopore reads can be mapped to the exact plasmid sequence allowing to analyze all 25 nucleosome positions independently as well as simultaneously on the same reads, given the read is long enough to cover all positions (Figure 2.20).

As seen earlier the ODM-seq variant using direct methylation calling (Nanopore-seq) gives the same gene-averaged accessibility as bisulfite (BS-seq) or enzymatic methylation (EM-seq) measurements using the same methylated chromatin (Figure 2.10A). Additionally we obtained a high correlation between BS-seq and Nanopore-seq when comparing all CpG sites individually (Figure 2.10C). Given the high agreement between the different methods we could then investigate long reads from genic regions and observe the typical nucleosomal array pattern on individual reads (Figure 2.22).

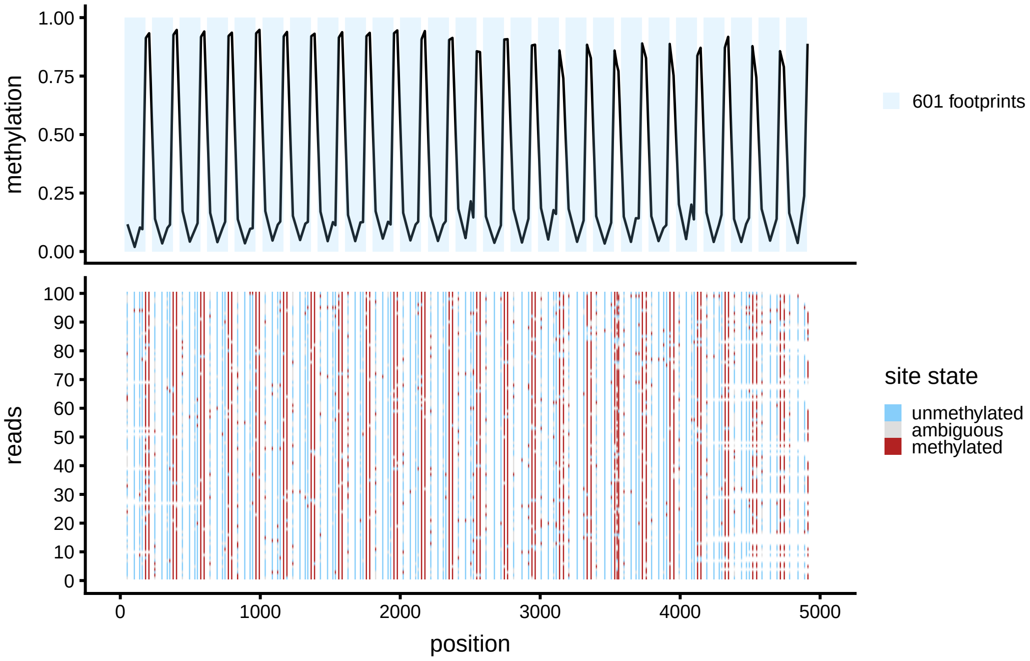


**Figure 2.20:** Measured methylation of the 601 25mer plasmid spike-in. Top: methylation at each site, averaged over reads. The Widom 601 sequences (marked by light blue rectangles) position nucleosome during salt gradient dialysis with high precision and only the linker regions between achieve high methylation ratios. Bottom: 100 individual reads covering the 601 25mer sequence. In contrast to the BS-seq method (spike-in data in Figure A.5), nanopore sequencing allows to investigate highly repetitive sequences as well as much longer individual reads.

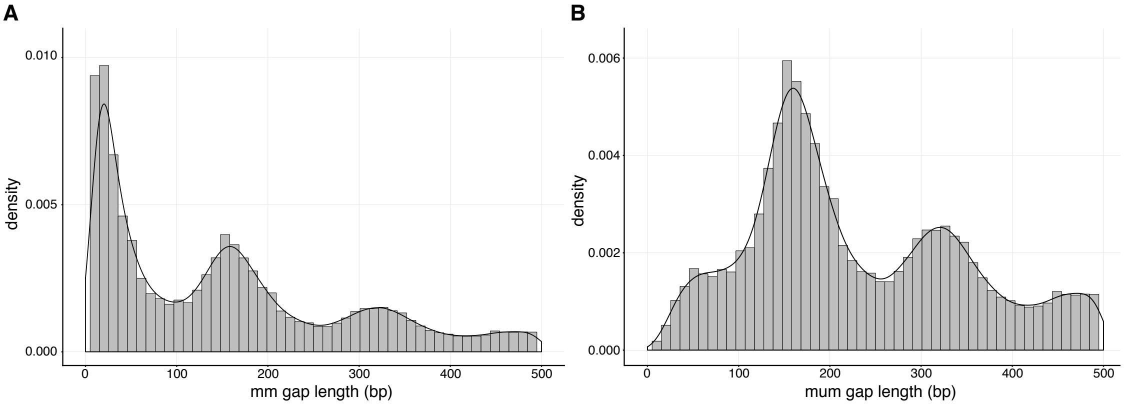


**Figure 2.21:** Histograms of ODM-seq nanopore mm- and mum-gap lengths for the WT4 xl CpG 180min sample. (**A**) mm-gaps are the distances between two methylated sites, with no other methylated sites in between. (**B**) mum-gaps are mm-gaps with at least one unmethylated site in between thus indicating a nucleosome or other bound factor between the methylated gap end points.

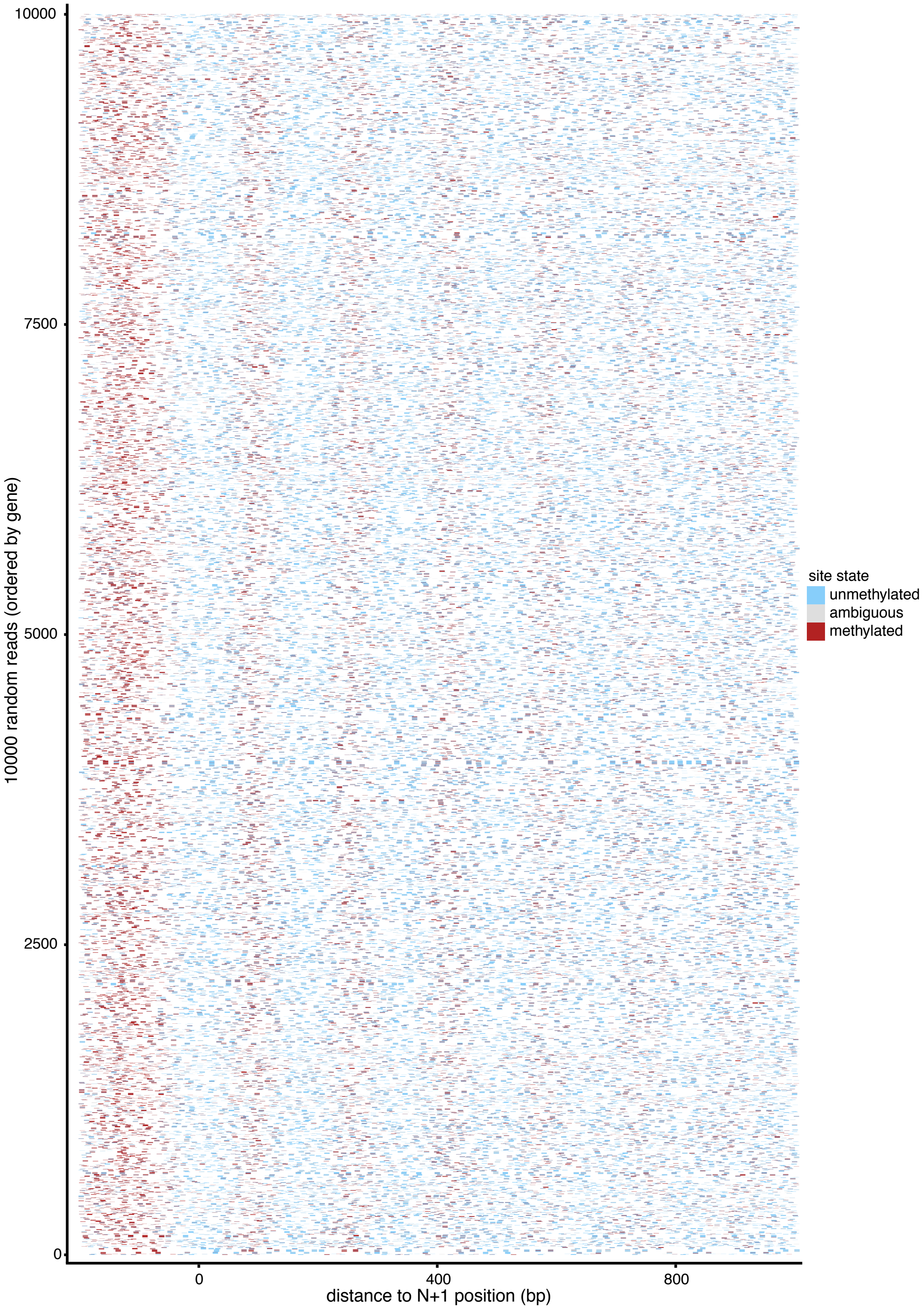


**Figure 2.22:** 10000 random genic reads of the WT4 xl CpG 180min ODM-seq nanopore sample covering positions -200 to 1000 with respect to the corresponding N+1 position and ordered by gene.

To investigate the Nanopore-seq data further, we defined the distance between two methylated methylation sites with no other methylated site in between as “mm-gap” and a more refined version, the “mum-gap” as an mm-gap with at least one unmethylated site between the methylated gap end points.[4] Short mum-gaps approximate the maximal size of nucleosomes or other bound factors while longer mum-gaps can be caused by dinucleosomes or higher “clusters” of nucleosomes (Figure 2.21). However, the quality of the approximation depends on the underlying methylation site resolution and the exact site positions. If there is not methylation site in the linker between two nucleosomes, the resulting mum-gaps will span both particles and if there is no site close to the end of a nucleosome the mum-gaps will overestimate the nucleosome footprint. Even though the distribution of mum gaps shows the highest peak near the typical footprint size of a nucleosome, many gaps are shorter or much longer. This is due to the limited resolution of CpG sites on the genome as well as ambiguous sites and calling errors. For high enough resolution and perfect methylation calling we would expect mum-gaps of mostly of the size of nucleosomes and smaller bound factors only. For mum gaps with a typical nucleosome footprint size a simple heuristic placing the nucleosome dyad into the center of the gap might work reasonable well, but determining nucleosome positions in larger mum-gaps, especially placing several nucleosomes inside the same mum-gap, remains an unsolved challenge. However, this investigation is still ongoing and hopefully leads to a more detailed analysis of nucleosome configurations, i.e. information on many neighboring nucleosomes on the same read, in the future.

## 2.6 Discussion

In this chapter we presented the measurement methods and analysis of the first genome-wide high-resolution absolute occupancy map in yeast. We used orthogonal approaches with restriction enzymes (ORE-seq) and DNA-methyltransferases (ODM-seq) as well as different methylation detection methods that cross-validated each other. We developed the necessary bioinformatics pipeline and correction methods for ORE-seq as well as ODM-seq (Appendix A). To investigate the absolute nucleosome occupancy, we combined our measurements with data from nucleosome-specific mapping approaches, like MNase-seq and chemical cleavage sequencing and we found that nucleosome occupancy is surprisingly uniform. We were able to calculate the total number of nucleosomes per yeast cell to 57 000 to 60 000, depending on the nucleosome specific method, with 97% of all nucleosomes called by chemical cleavage sequencing having an absolute occupancy greater or equal to 70%. Our absolute occupancy ODM-seq map has an overall average of 78%. According to our measurements, highly transcribed genes rarely exhibit low, but mostly high absolute occupancy, which argues that RNA polymerase passage fosters high nucleosome occupancy. MNase-based methods suggested an inverse relationship between transcription activity and nucleosome occupancy in gene bodies, but MNase is not reliable in this regard [25]. We were also able to measure the absolute occupancy of nonnucleosomal DNA-bound factors like the general regulatory factors Rap1 and Abf1 at their binding sites and found a negative correlation between RSC activity and absolute occupancy.

[4] The analysis of mm- and mum-gaps was implemented by Maryam Khatami and later further developed by Matthias Hanke.

For our ODM-seq approach, we used bisulfite or enzymatic conversion followed by short read (Illumina) sequencing or direct nanopore sequencing with methylation calling by Nanopolish. An alternative would be to combine one of the conversion steps with regular nanopore sequencing to detect the converted sites. Preliminary tests using a typical long read mapping tool (minimap 2) for the converted reads showed that this is possible in principle (data not shown), but a specific mapping tool for long converted reads would be needed for a proper quantitative analysis.

Our methodology can be applied to any chromatin preparation with frozen nucleosome dynamics and provides absolute occupancy of nucleosomes and other factors along the genome. However, one has to be careful to titrate DNA methylation into saturating conditions. In the future it can be helpful to investigate occupancy changes in processes like DNA replication, DNA repair or aging.

# 3 Chromatin Remodeling Simulations

## 3.1 Introduction

In the previous chapter we have developed several independent methods to measure the absolute occupancy of nucleosomes and other factors along the genome, which are steady state quantities averaged over many cells. Now we will focus on the underlying mechanisms that possibly establish these steady states provided by chromatin remodeling complexes (“remodelers”). Remodelers move, disassembly or modify nucleosomes in an ATP-dependent manner, but it has not yet been possible to observe their actions on single DNA molecules involving several nucleosomes (see Section 1.4). In this project we sought to explore the qualitative behavior of theoretical remodeler systems that are motivated by experimental observations. These theoretical, usually active, model systems are based on one-dimensional equilibrium soft-core nucleosome gas models used to fit the typical N+1 aligned average occupancy or dyad density patterns (see Section 1.3).

As base model without remodelers, we use a grand canonical ensemble of nucleosomes adsorbing to and desorbing from a one-dimensional substrate, the DNA. In this equilibrium setting, the steady state one- and two-particle densities can be calculated directly from the given external potential and the interaction potential between nearest neighbors. Solutions to the inverse problems, where potentials are calculated from given densities are also available, but in some cases not robust with respect to perturbations/errors in the densities or only implicitly given (details in Section B.1).

We extend this grand canonical model with remodeler enzymes, which are able to bind to nucleosomes on the DNA and then perform actions like sliding or rapid disassembly of the nucleosome (Figure 3.1). We also include the option of nucleosome remodeler complexes binding directly to empty DNA. Sliding remodelers can be given a preferred sliding direction, either upstream or downstream, which is decided upon binding to the nucleosome and can be tuned from an unbiased random walk ($\gamma = 0$) keeping detailed balance and thus the system in equilibrium to an almost completely deterministic step in a fixed direction ($\gamma \to \infty$). During sliding, the interaction between neighboring nucleosomes or remodeler complexes is also considered.

The nucleosome interaction potential we use here is based on the observation of nucleosome breathing [14, 15, 16, 17], i.e. “the spontaneous transient unwrapping of nucleosomal DNA from either end of the fully wrapped nucleosome ‘ground state’ represented by the crystal structure” [41]. Möbius et al. [41] used a nucleosome breathing model where base pairs can unwrap from either side, with unwrapping energy $\varepsilon$, theoretically until the nucleosome center, the dyad, is reached. The nucleosome footprint length $a = 2w + 1$ is another fit parameter, with $w$ being the maximum number of unwrapped base pairs on each side. This effective

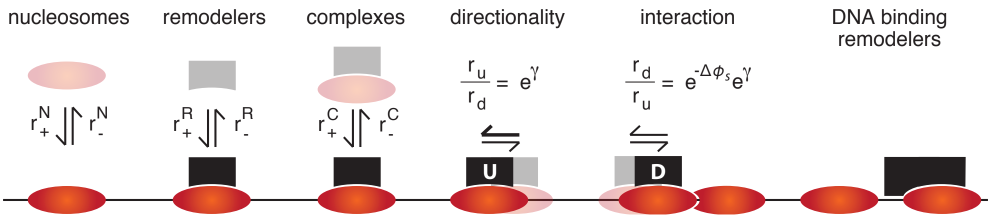


**Figure 3.1:** Possible "ingredients" of nucleosome remodeling simulations from left to right: Assembly and disassembly of nucleosomes along the DNA, (un-)binding of remodelers to nucleosomes, (un-)binding of remodeling complexes, sliding of remodeling complexes along the DNA possibly with preferred direction upstream ("U") or downstream ("D") governed by $\gamma$ and affected by the interaction with neighboring particles. Finally we also investigate sliding remodelers that additionally bind directly to the DNA. The interaction potential $\phi$ affects sliding as well as nucleosome and complex (un-)binding. $\mathrm{k_B T}$ is set to 1 throughout this chapter. Rate calculations are described in Section B.3.

nucleosome footprint $a$ is typically larger than the footprint of the crystal structure as this effective description also includes steric constraints of chromatin. The interaction potential can then be calculated by counting all possible configurations of unwrapped base pairs of two nucleosomes at a fixed distance (details in Section B.3.1). Specifically, we use this interaction potential with the parameters fitted in [44] (Figure 3.2C and Figure B.1A)

Due to the lack of experimental data we first assumed the interaction with neighbors to remain unchanged when remodelers bind to nucleosomes, which is a strong assumption, given that some remodelers exceed the size of nucleosomes by several factors [47]. Since experimental data suggests that remodelers also bind to DNA directly [60], we wanted to investigate the theoretical implications within our models as well, and thus gave some remodeler complexes a larger footprint, with an extended interaction potential.

So far, there is no experimental method to track the dynamics of individual nucleosomes in a multi nucleosome configuration under the influence of remodelers. Consequently it very hard to tune the parameters of the models to give realistic results. Thus, we do not aim for a quantitative description, but a qualitative investigation of interesting phenomena in this setup, providing some insight into possible behaviors of such systems and hopefully guiding future experiments. The next section investigates the effects of active directional sliding mediated by remodelers followed by sections on different nucleosome eviction mechanisms and DNA-binding sliding remodelers.

## 3.2 Directional sliding remodelers near a boundary [1]

In this section we discuss the effects of remodelers that enable nucleosome sliding. In this setup, after a remodeler binds to a nucleosome, the complex moves upstream or downstream with rates $r_u$ or $r_d$, respectively (Figure 3.2A). There are two types of complexes, upstream movers, where $r_u \geq r_d$ and downstream movers where $r_u \leq r_d$. The parameter $\gamma$ increases or decreases $r_u$ and $r_d$ such that $r_u/r_d = e^{\pm\gamma}$ with the plus sign for upstream movers and

[1] Parts of this section are adapted from our manuscript [44] under CC BY-NC 4.0 license.

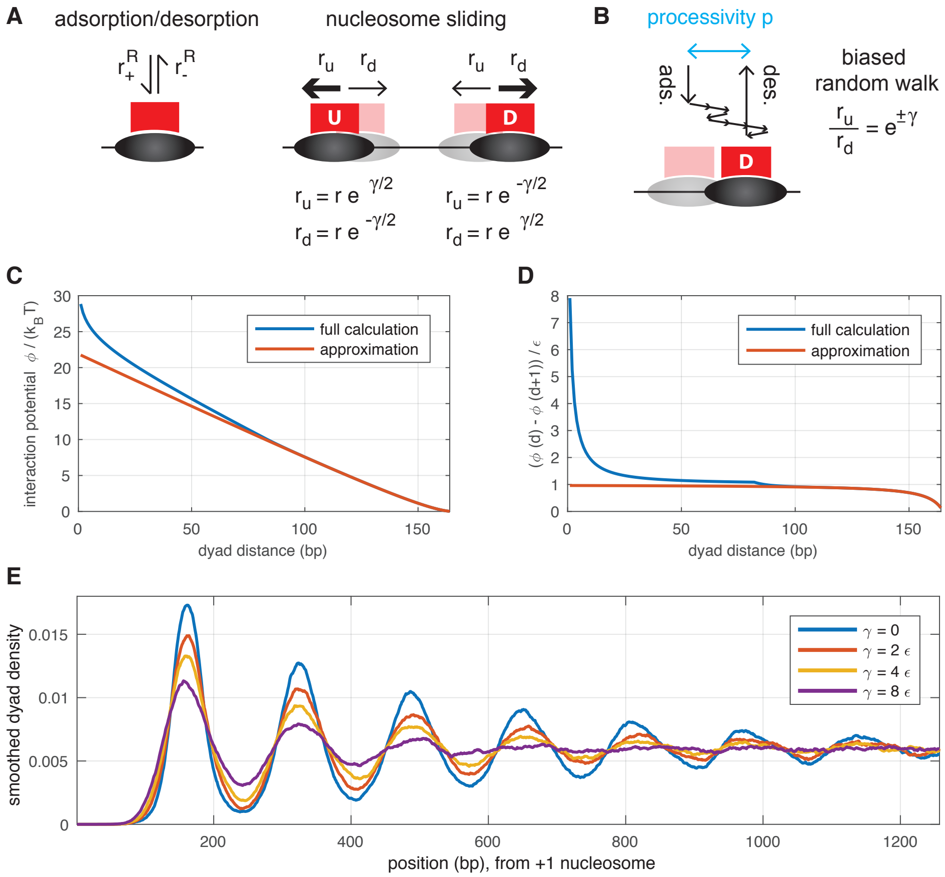


**Figure 3.2:** Directional sliding affects the steady state density. (**A**) Model setup with nucleosome adsorption/desorption and remodeler mediated directional sliding. Remodeler complexes perform a biased random walk with $r_u/r_d = e^{\pm\gamma}$ (plus sign for upstream mover and the minus sign for downstream mover) (**B**) The processivity $p$ measures the average displacement in the preferred direction, i.e. the drift distance, in a free environment until unbinding of the remodeler or the complex. (**C**) Unwrapping potential (full calculation in Section B.3) with approximation derived in [41] for $\varepsilon = 0.152$ and $w = 82$ (where we set 1 bp to length 1). (**D**) Interaction penalty in units of $\varepsilon$ for sliding one bp further into a neighboring nucleosome. (**E**) Smoothed dyad density of nucleosomes next to the fixed N+1 nucleosome at 0 ("soft boundary") for different values of the movement bias parameter $\gamma$. Given the complex moves without neighbor interaction, $\gamma = 2\varepsilon$ ($\gamma = 8\varepsilon$) corresponds to a probability of 0.5754 (0.7714) to move in the preferred direction. Simulation parameters can be found in Table B.1.

minus sign for downstream movers. The preferred direction is decided at random with equal probability upon binding to the nucleosome.

If $\gamma = 0$ the complexes perform a symmetric random walk and detailed balance holds throughout the system. In this case the addition of the remodelers to the model changes the dynamics of the nucleosomes compared to the basic grand canonical model without remodelers, as nucleosomes can now slide along the DNA, but all steady state quantities regarding the nucleosomes, like dyad density or next-neighbor distance distribution, remain unchanged.

If $\gamma > 0$, the complexes perform a biased random walk. We define the processivity $p$ of each type, upstream or downstream mover, as the average displacement in upstream or downstream direction, respectively, between remodeler binding and unbinding in case of free movement, i.e. without neighbor interaction (Figure 3.2B). It can be calculated from the average drift speed (exemplary for a downstream mover below), with $s = 1\,\mathrm{bp}$ being the jump size and $1/r_- := 1/(r_-^R + r_-^C)$ the average life time of the remodeler complex on the DNA:

$$p = s(r_d - r_u)/r_- = 2sr\sinh(\gamma/2)/r_- \tag{3.1}$$

$$\approx sr\gamma/r_- \quad (\text{for } \gamma << 1) \tag{3.2}$$

When a remodeler complex moves towards a neighboring nucleosome, the movement rates are also affected by the interaction potential (Figure 3.2C). The movement bias towards the neighbor is neutralized, if the interaction penalty for a 1 bp jump $\Delta\phi$ is equal to the bias parameter $\gamma$. Assuming the remodeler and the whole remodeler complex do not unbind, this happens quite early for $\gamma < \varepsilon/2$, but already $\gamma = \varepsilon$ allows a drift far into the neighboring nucleosome and $\gamma = 8\varepsilon$ up until the neighboring dyad (Figure 3.2D). Note that remodeler complex as well as nucleosome unbinding become increasingly likely with dyads moving closer to each other, as the interaction potential also applies to adsorption and desorption rates when neighbors are nearby.

Due to the directionality of the remodelers, detailed balance is broken and the model a non-equilibrium system. We found that as $\gamma$ increases, the amplitude of the nucleosome array pattern decreases (Figure 3.2E), presumably because the directional sliding enables movement into neighboring nucleosomes until the energy penalty due to the interaction counteracts the directional movement. Additionally, if $\gamma > 0$, increasing the base sliding rate $r$ further decreases the amplitude of the nucleosome array pattern [44].

Note that the new array patterns are very similar to array patterns of the basic grand canonical nucleosome gas, but with different parameters for the nucleosome interaction. Indeed, when fitting the remodeled densities with the basic model, we find very good agreement (Figure 3.3A). Repeating this fit for densities from remodeler systems with different parameters $\gamma$ and $r$ yields flow diagrams of the fit parameters $w_{\text{eff}}$ and $\varepsilon_{\text{eff}}$ (Figure 3.3B) showing a minor decrease in $w_{\text{eff}}$ with increasing remodeler processivity $p$ and a strong decrease in $\varepsilon_{\text{eff}}$. This corresponds to effective nucleosome softening, as the nucleosome interaction strength decreases with $\varepsilon_{\text{eff}}$. Note that remodeler system can reach the same processivity with different combinations of $\gamma$ and $r$ but have different steady state quantities and thus different effective fit parameters $w_{\text{eff}}$ and $\varepsilon_{\text{eff}}$.

As shown in [44], "effective nucleosome softening is [also] a generic phenomenon of averaging over nucleosome positioning landscapes" similar to remodeler mediated directional sliding

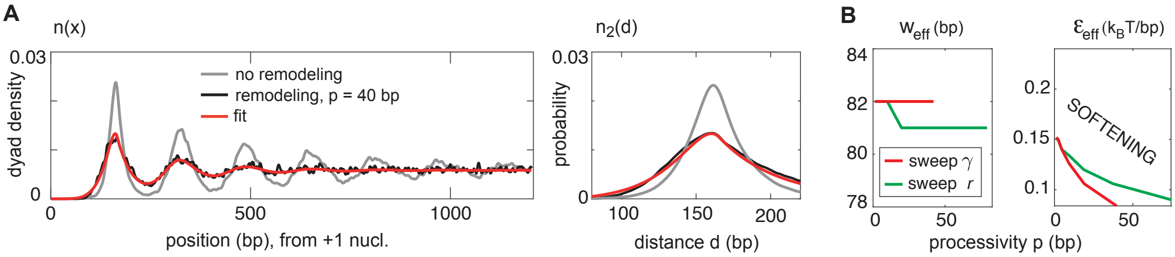


**Figure 3.3:** Apparent nucleosome softening by remodeler mediated directional nucleosome sliding. (**A**) One-particle densities and next-neighbor distance densities of the system without remodeling, with directional remodeling and a fit to the remodeled densities without remodeling. (**B**) Parameters of the interaction potentials fitted to remodeled systems with $w = 82\,(\mathrm{bp})$ and $\varepsilon = 0.152\,(\mathrm{k_BT/bp})$. Simulation parameters can be found in Table B.1.

and the opposite, effective nucleosome stiffening can be generated by remodeler mediated nucleosome attraction or periodic energy landscapes for example by trapping on AT rich sequences.

To sum up, adding directional, and thus active, sliding mediated by remodelers to the basic nucleosome gas affects steady state observables. The steady state density and next-neighbor distance distribution, however, can be fitted by the basic nucleosome gas model with parameters that describe softer nucleosomes as compared to the initial active model.

## 3.3 Nucleosome eviction by chromatin remodeling

Since most genes in yeast exhibit a nucleosome depleted upstream of the transcription start site, we wondered which remodeler activities can lead to a reduced dyad density and thus reduced absolute occupancy within a certain region. Note that the reduced dyad density could also directly be forced by an external potential simulating a strong DNA sequence dependence that leads to less nucleosome binding. The nucleosome dynamics at promoters is likely determined by a complicated interplay of mechanisms, and here we only want to illustrate that different eviction remodeling mechanisms can in principle give similar results.

To this end we simulate two different eviction mechanisms in a periodic system. We first focus on the effects of remodelers that are locally recruited and destabilize the nucleosome by strongly increasing its disassembly rate (Figure 3.4A). This mechanism breaks detailed balance, as there is no corresponding increase in nucleosome assembly rate. As expected the life time of nucleosomes on the DNA is strongly reduced within the remodeler recruitment region (Figure 3.4B) compared to the rest of the system.

We found that using the same remodeler recruitment peak for directional sliding remodelers (upstream movers as well as downstream movers) has a similar depletion effect, as remodeler complexes slide out of the recruitment region or move deeply into neighboring nucleosomes until the neighbor disassembles due to the interaction (Figure 3.4C). A similar process has already been observed in experiments using a dinucleosomal model system [18] and has been investigated in the SWI/SNF remodeler subfamily (Section 1.4.2). In our simulations, the

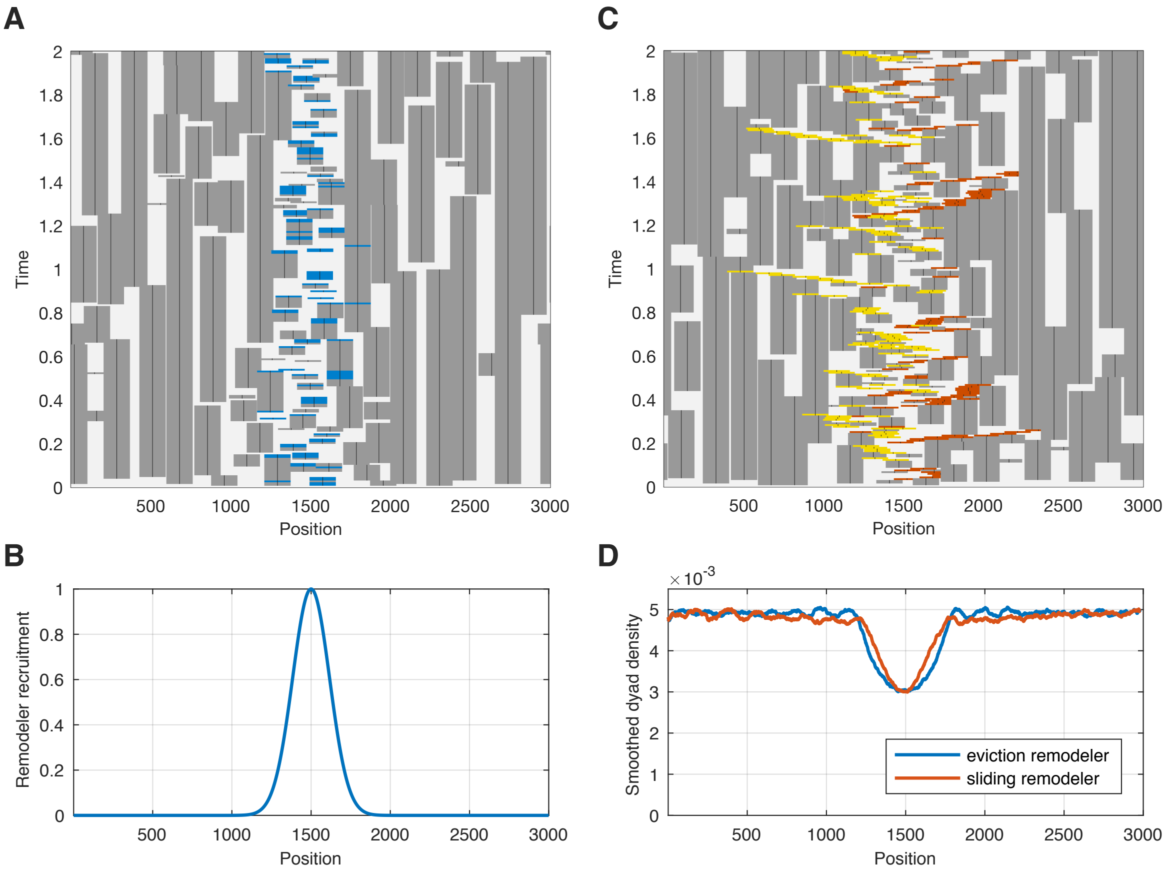


**Figure 3.4:** Different nucleosome eviction remodeling mechanisms can achieve similar density profiles. (**A**) Kymograph of assembling and disassembling nucleosomes, where the disassembly rate is strongly increased by a remodeler (blue) that predominantly binds to nucleosomes around position 1500. (**B**) All remodeler binding to nucleosomes is proportional to the same recruitment landscape (centered at 1500 with standard deviation 165). (**C**) As panel A, but here remodelers slide nucleosomes in a preferred direction, either downstream (red) or upstream (yellow). With $\gamma = 10\varepsilon$ and a jump size of $s = 10$, remodeler complexes are able to move into a neighboring nucleosome (Figure 3.2D), which becomes increasingly likely to disassemble because of the soft nucleosome interaction potential. Here remodeler nucleosome complexes themselves do not bind to or unbind from the DNA directly, but are only assembled and disassembled step by step. (**D**) Averaged nucleosome dyad densities of both mechanisms at simulation time 10 for 20000 realizations. Simulation parameters can be found in Table B.2.

nucleosome remodeler complexes can only disassemble step by step by unbinding of the remodeler followed by disassembly of the nucleosome (complex assembly is also step by step), but the results are similar for enabled direct complex binding and unbinding. We tuned our two different remodeler mechanisms such that, using the same local recruitment, base remodeler and nucleosome binding/unbinding rates, their resulting dyad densities show a drop to the same value at the center of the recruitment peak (Figure 3.4D), illustrating that different eviction mechanisms can lead to very similar nucleosome densities.

In both cases the nucleosome depletion causes only slight density oscillations outside the depleted region, indicating that this recruitment peak alone is not suitable to restrict the positioning of nucleosomes next to the depletion region, as it is the case for in vivo N+1 nucleosomes when aligning different genes with respect to their transcription start site. This underlines the importance of more informative observables than nucleosome density alone.

Here, the two theoretical scenarios could be distinguished by the steady state distribution of the remodelers, as the sliding remodeler complexes move further away from the recruitment peak, where no eviction remodelers would bind. An experimentally accessible observable that could distinguish both mechanisms without investigating the remodelers themselves is the dynamics of Flag/Myc-tagged H3 histones [110, 12, 13]. In this method, at the beginning only Myc-tagged H3 histones are available for assembly while the concentration of Flag-tagged H3 histones increases slowly over time. During the Flag-tagged H3 histone concentration increase, positions with very stable nucleosomes are less likely to incorporate a Flag-tagged histone, as for example around position 1000 and 2000 just outside the depleted region in Figure 3.4A. Nucleosomes in the same area that are more likely to disassemble and reassemble because of contact with sliding complexes from the depleted region (Figure 3.4C), however, might show a faster uptake in Flag-tagged H3 histones.

## 3.4 DNA-binding sliding remodelers

It has been experimentally observed that remodelers that slide nucleosomes can also bind extranucleosomal DNA, for example the ISW1a remodeler in yeast, which binds DNA on both sides of the nucleosome [60] or the drosophila ACF remodeler that can bind as a dimer, with each monomer also binding the DNA to one side of the nucleosome [111] (also see Section 1.4.3). Here, we investigate sliding remodelers that in addition to nucleosomes also bind to neighboring stretches of DNA. We assume that the extranucleosomal DNA is only bound on one side of the nucleosome and that we can combine DNA binding and nucleosome binding into one event. Additionally we assume that it has no other effects on the remodeler actions despite changing the interaction with neighbors. We explore the effect of such extranucleosomal DNA binding on sliding remodelers, with or without directional movement. The directional movement creates two configurations, either a preferred movement into the direction of the DNA binding ("pull remodeler") or against ("push remodeler") (Figure 3.5).

Here we study DNA binding remodelers in a simple periodic system with a fixed number of nucleosomes with initially equal distance to each other, that do not assemble or disassemble to keep the average nucleosome density fixed when switching from non-DNA-binding remodelers to DNA-binding remodelers. Since the DNA-binding remodelers compete with the nucleosome for DNA sites, the chemical potential of nucleosomes would have to be corrected to

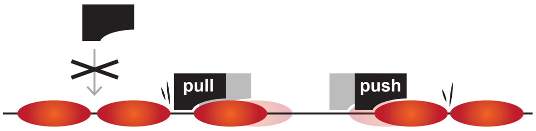


**Figure 3.5:** DNA-binding remodelers with push and pull configurations. In our model, “push” remodeler complexes preferably move away from the side where the remodeler binds the DNA. “Pull” remodeler complexes preferably move towards the DNA binding side of the remodeler. The interaction range with other nucleosomes is extended by the amount of remodeler bound DNA, which can lead to blocked remodeler binding, blocked sliding or even nucleosome displacement, if nucleosomes are allowed to disassemble in the specific simulation.

keep an equal nucleosome density. For such a simplified system, we found that the DNA binding configuration strongly affects the system dynamics. We use a system with unidirectional sliding remodelers without DNA-binding (Figure 3.6A) as comparison. Just adding the DNA-binding for unidirectional sliding did not change the dynamics qualitatively (Figure 3.6B). This changes, however, if DNA-binding is combined with directionality, in this case by setting $\gamma = 0.5\varepsilon$. In a pull configuration, where the remodelers bind DNA in direction of the preferred movement, nucleosomes appear to be much more regular spaced (Figure 3.6C). Systems with push remodelers show again different dynamics, with quickly appearing clusters, i.e. two or more nucleosomes that are within interaction range most of the time (Figure 3.6D). These clusters appeared to act as effective particles moving much slower than individual remodeled nucleosomes with only very rare cluster dissolution events (dissolution not shown in Figure 3.6D). This “cluster diffusion” seems to depend on cluster size with more than two nucleosomes being apparently completely frozen (Figure 3.6D).

Since the systems here are translation invariant due to the periodic boundary condition, the average one-particle densities are all constant. The normalized void distributions, i.e. next neighbor distance densities, fit qualitative descriptions above (Figure 3.6E). Initially the nucleosomes start with a distance equal to the footprint length of one DNA-binding remodeler nucleosome complex (165 bp + 82 bp = 247 bp). Without DNA-binding the most likely next neighbour distance was 169 bp. DNA-binding with unidirectional remodelers ($\gamma = 0$) has similar void distribution with a global maximum at 171 bp and a second lower maximum at 245 bp. The position of the lower maximum is close to the footprint length of a remodeler nucleosome complex. Setting $\gamma = \varepsilon/2$ changes the void distribution drastically. In the case of pull remodelers we find a strong peak with a maximum at 240 bp, which is a bit shorter than the the complex footprint length, most likely due to the directionality which drives the complex further into the neighbor before the interaction counteracts the directional bias. The much sharper void distribution corresponds to the more regular nucleosome distances in Figure 3.6C. In the case of push remodelers, we find a similarly sharp distribution, but shifted to a maximum at 155 bp, with a very shallow second maximum at 253 bp. Since the remodeler DNA-binding site is not in the direction of the preferred movement anymore, the remodelers can push the nucleosome much closer to each other and even around 10 bp into the interaction range due to the directionality. At this distance the nucleosome start to form clusters, since it is almost impossible for a push remodeler to bind inbetween nucleosomes with outward directed preferred sliding while binding with inward directed preferred movement is much more likely since the DNA at the cluster border is more accessible to the remodeler. Note that due to the much slower emerging cluster dynamics in the case of push remodelers, it is not clear

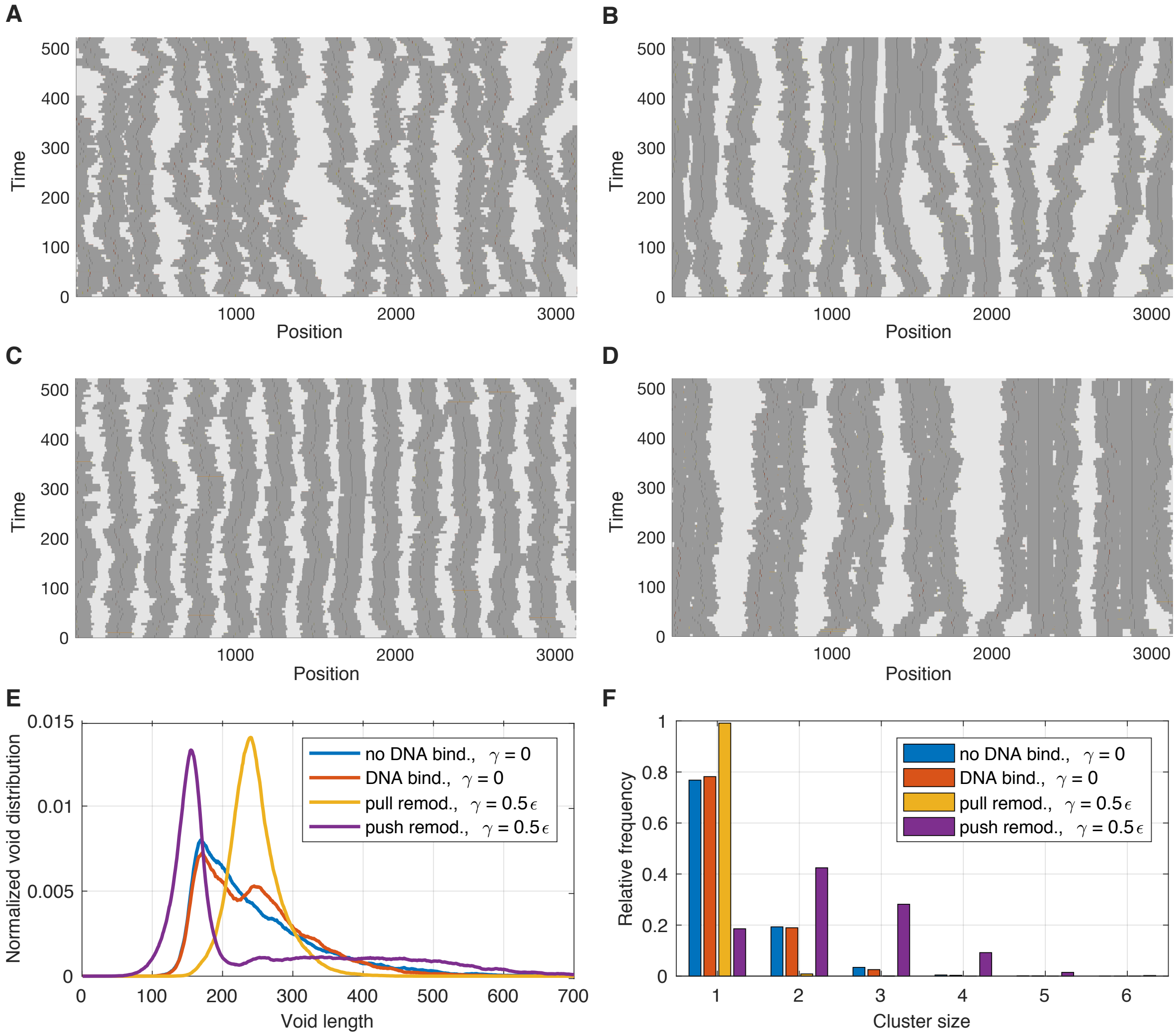


**Figure 3.6:** DNA-binding sliding remodeler effects on systems with fixed nucleosome number, i.e. without nucleosome assembly/disassembly. (**A-D**) Exemplary kymographs of nucleosomes (interaction footprint of 165 bp in dark gray) which are moved by remodelers (not visualized here, as the remodeler binding/unbinding dynamics is too fast) which (**A**) only bind the nucleosome and have no preferred sliding direction, (**B**) bind 82 bp of DNA on one random side next to the nucleosome (not visualized) and have no preferred sliding direction, (**C**) bind the nucleosome in pull configuration, (**D**) bind the nucleosome in push configuration. (**E**) Smoothed and normalized void (distance between two neighboring nucleosome dyads) distributions for all four cases. (**F**) Relative frequencies with which nucleosomes are part of clusters of different sizes for all four cases. Nucleosomes are part of the same cluster if their interaction footprints overlap. Results of E and F were calculated at time 1000 from 1000 realizations each with system size 7410 and 30 nucleosomes. Further simulation parameters can be found in Table B.3.

whether the simulation time is long enough to reach a steady state. With stronger movement bias ($\gamma = \varepsilon$) of the pull remodeler, the peak of the void distribution becomes even sharper with a slightly reduced maximum position, whereas the void distribution of the push remodeler is shifted far towards lower distances with a maximum at 84 bp, pushing nucleosomes deep into each other (data not shown). This large shift is also a consequence of this setup where nucleosome do not disassemble even if a neighbor is very close.

To investigate the cluster formation in each case, we calculated the relative frequencies with which nucleosomes are part of clusters of different sizes. Two neighboring nucleosomes belong to the same cluster if they are within interaction range (165 bp). In the unidirectional settings we find similarly decaying relative frequencies with 77% and 78% of nucleosomes being without a close neighbor, without and with DNA-binding, respectively (Figure 3.6F). In the case of pull remodelers, 99% of all nucleosomes are without close neighbor, since pull remodelers would preferably bind in outward direction of any given cluster and dissolve it quickly. Using push remodelers, however, only 19% of nucleosome have no close neighbor and most nucleosomes are moved to form clusters of at least two nucleosomes, with approximately 42%, 28%, 9.2%, 1.4% of nucleosomes being part of clusters of size 2, 3, 4, 5, respectively. Again with the caveat, that the snapshot of cluster sizes of push remodeler might not be in steady state and the steady state cluster frequencies might shift even more towards larger clusters.

## 3.5 Discussion

The aim of this project was to explore the space of one-dimensional nucleosome gas models including an active remodeler action that can take place after remodeler binding to the nucleosome. We analyzed several mechanisms independently to understand their implications. Starting from a basic grand canonical nucleosome gas, we investigated the effects of remodeler mediated sliding without and with preferred direction. We found that a movement bias in a preferred direction changes the steady state one-particle density in very similar way to just making the nucleosomes softer in the basic grand-canonical model. This should be taken into account in the future when trying to determine the underlying nucleosome properties by fitting to experimental data. We then moved towards eviction mechanisms and illustrated that a disassembly remodeler, that increases the nucleosome disassembly rate, and directional sliding remodelers both recruited with the same strength, can reduce the dyad density at the recruitment peak in a very similar way. In the third part we explored the effects of one-sided DNA-binding of sliding remodelers. While DNA-binding without directional bias did not have large effects on the void distributions, including a directional movement bias gave strongly different void distributions and cluster sizes depending on the binding configuration (pull or push remodelers). While push remodelers showed interesting clustering behavior we decided against further investigation due to the lack of experimental evidence for this mechanism as well as due to the fact that typical measured nucleosome patterns often appear to be rather regularly spaced than clustered.

Note that the here presented models are not intended to make quantitative predictions about biological systems. Future studies might be able to build on more detailed experimental data, especially of the dynamics of nucleosome configurations under the influence of remodelers, to tune the models accordingly. In a next step one could take on the challenge of comparing simulations of several remodeling mechanisms with in vivo measurements, where many types of remodelers with different mechanisms are acting at the same time.

# 4 Effective Dynamics of Nucleosome Configurations at the *PHO5* Promoter [1]

In this chapter, we show that the availability of single cell nucleosome configurations data instead or in addition to cell averaged occupancy data allows a much more detailed modeling approach of nucleosome dynamics. As an example, we use the *PHO5* promoter, where the interplay of nucleosomes and transcription factors has been studied intensively in its activated as well as inactivated state and which also serves as a paradigm for human promoters [33].

The here presented interdisciplinary project was a close collaboration with Philipp Korber and his group[2] and lead to a manuscript that is currently under review [112]. The described promoter mutant experiments ("sticky N-3 mutants") were performed by Andrea Schmidt and Philipp Korber.

## 4.1 Introduction

The *PHO5* gene is regulated via the availability of intracellular inorganic phosphate. If enough phosphate is available inside the cell, the *PHO5* gene is repressed as its promoter region is occupied by four well-positioned nucleosomes numbered N-1 to N-4 relative to the gene start. Especially nucleosomes N-1 and N-2 block transcription factor binding sites that are critical for gene induction. N-1 prevents access of the TATA-box binding protein (TBP) to the TATA-box, and N-2 hinders the transactivator Pho4 from binding the UASp2 element (Upstream Activating Sequence phosphate regulated 2). Upon phosphate removal or depletion, a cascade of signals leads to activation of Pho4 by inhibition of its phosphorylation and by increasing its nuclear concentration [113, 114]. Pho4 triggers a complicated nucleosome remodeling process involving up to five different nucleosome remodeling enzymes, some of which have redundant and some crucial roles [115]. It also triggers histone acetylation, histone chaperones and probably more cofactors that finally leads to more or less complete removal of nucleosomes N-1 to N-5 and transcription of the *PHO5* gene (reviewed in [33]). Note that this chromatin transition was among the first shown to be a prerequisite and not consequence of transcription initiation [65, 116, 6]. Thus, it strongly argued for the now widely accepted view that chromatin structure, positioned nucleosomes in particular, are not just packaging DNA, but represent an important level of regulation.

While the involved cofactors are known for this model system exceptionally well, the resulting nucleosome dynamics are still not yet understood. To investigate these, the full promoter

[1] Large parts of this chapter are adapted from our manuscript [112] under CC BY 4.0 license.

[2] Molecular Biology Division, Biomedical Center, Faculty of Medicine, Ludwig-Maximilians-Universität München

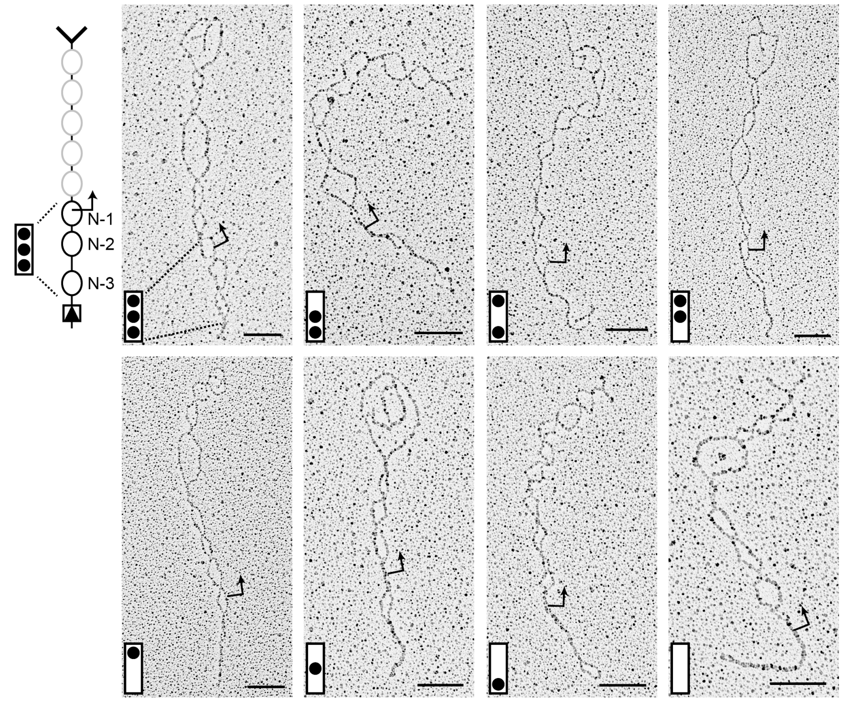


**Figure 4.1:** *PHO5* gene molecules with promoter nucleosome configurations measured in [34]. *PHO5* gene chromatin rings were formed in vivo, then isolated. Crosslinking with psoralen only covalently bound DNA double strands in linker regions between nucleosomes with each other. After denaturation and specific cutting, the former nucleosomes appeared as bubbles when analyzed with an electron microscope. To distinguish the two molecule ends, a "fork" was introduced at the end without the *PHO5* promoter. The configurations of several hundred molecules were analyzed for different promoter states (repressed wild type as well as mutants with different degrees of activation). From: [34] under CC BY 4.0 license.

nucleosome configurations in different states are needed. For the *PHO5* promoter this information is available, at least for the subsystem of the N-1, N-2 and N-3 sites, in activated, half-activated and repressed states [34]. The N-1 to N-3 subsystem of *PHO5* promoter chromatin has been validated before to recapitulate the regulation behavior of the full-length promoter [116], allowing us to restrict the following analysis to these three nucleosome positions/sites. In [34], Brown et al. measured the occurrences of promoter nucleosome configurations of separate *PHO5* gene molecules by crosslinking linker DNA and imaging the molecules with electron microscopy (Figure 4.1). Additionally, simple biologically motivated Markov models were used to describe the dynamics of promoter nucleosome configurations and we will discuss this approach in more detail in the following Section 4.2. Importantly, in [117], Brown et al. also showed that the variation in nucleosome configurations is intrinsically stochastic, i.e., it is not the result of other stochastic events in the nucleus. Thus the usage of Markov models, which evolve only depending on the current system state, not on its history or other external factors, for the dynamics of promoter nucleosome configurations is valid. In [34], however, such models were not systematically investigated and only a few models that individually fit the data of different promoter states presented (more details and a systematic extension are discussed in Section 4.2). There are also other more detailed computational models of nucleosome remodeling with base-pair resolution, for example the study of Kharerin et al. [118], which needed many assumptions to fit the data and therefore is not consistent with a fully unbiased approach.

With current methods it is impossible to observe changes in individual nucleosome configurations at the same locus/promoter over time in vivo or in vitro, thus these dynamics need to be inferred by systematic and unbiased theoretical modeling. In theory, many equilibrium and non-equilibrium models can reproduce the same measured steady state distributions of promoter configurations. The challenge in modeling the promoter nucleosome dynamics is to find well-motivated restrictions and assumptions to decrease the number of fit parameters to a reasonable level, while still staying unbiased and modeling on a similar level as the available experimental data. In our case we also combined as many different experimental data sets as possible within the same model.

It is important to note that the *PHO5* promoter nucleosome dynamics can be viewed in two different “levels” of detail, depending on the experimental data available. The site-centric point of view concentrates on the three positions/sites N-1, N-2 and N-3 and remodeling at these sites individually, i.e. without considering the neighboring sites (Figure 4.2A). Experimental examples are typically based on cell-averaged measurements like accessibility measurements by restriction enzymes at these positions, where the measurements are taken at each position separately. More detailed single cell experiments are able to keep track of all the eight promoter nucleosome configurations (“promoter configurations”), i.e. simultaneously observe which of the three sites are occupied in each cell (Figure 4.2B). As a starting point for our simultaneous fit, we decided to use the configurational data for three mutants (Table 4.1) corresponding to different activation states of the *PHO5* promoter (“promoter states”): repressed (wild-type), weakly activated (*pho4[85-99] pho80*Δ *TATA* mutant) and activated (*pho80*Δ mutant) [34]. These three promoter state have different steady state distributions for the occurrences of promoter configurations (Figure 4.2C). We did not include the additional mutant strains provided by the Brown et al. study, since their steady state distributions are very similar to the ones just mentioned. Note that it is possible to obtain the three absolute site accessibilities

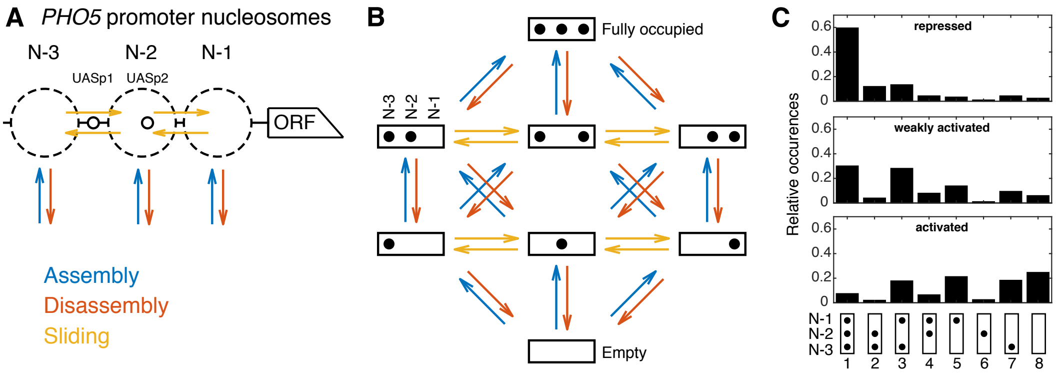


**Figure 4.2:** (**A**) Simplified nucleosome dynamics at the *PHO5* promoter including assembly, disassembly and sliding from a site-centric point of view. Possible nucleosome positions are indicated by dashed circles. Small circles represent Pho4 binding sites (UASp elements). (**B**) Modeling approach with 8 promoter configurations and 32 reactions. Arrow color code as in panel A. (**C**) Measured relative occurrences of the 8 promoter configurations indicated at the bottom as in panel B but rotated by 90° and for three different "promoter states": the repressed wild-type, a weakly activated mutant (*pho4[85-99] pho80Δ TATA*) and the activated mutant (*pho80Δ*), using data from [34], shown in Table 4.1.

| State | Mutant | Conf. 1 | Conf. 2 | Conf. 3 | Conf. 4 | Conf. 5 | Conf. 6 | Conf. 7 | Conf. 8 |
|---|---|---|---|---|---|---|---|---|---|
| Repressed | wild type | 125 | 25 | 28 | 9 | 7 | 2 | 9 | 5 |
| Weakly act. | *pho4[85-99] pho80Δ* | 61 | 8 | 57 | 16 | 28 | 2 | 19 | 12 |
| Activated | *pho80Δ* | 15 | 4 | 36 | 13 | 43 | 5 | 37 | 50 |

**Table 4.1:** Occurrences of each of the eight promoter nucleosome configurations obtained from electron microscopy of single *PHO5* gene molecules in [34]. All cells were grown in high-phosphate conditions which leads to a repressed *PHO5* promoter in wild-type. *pho80* knockout mutants simulate phosphate starvation and take on the activated *PHO5* promoter state. Configuration numbering as in Figure 4.2C.

from the promoter configuration occurrences, but this calculation cannot be inverted.

In the following sections, our goal is to design an unbiased class of effective minimal models featuring assembly, disassembly and sliding processes, which describe promoter configurations with the same detail as the available data sets and then analyze all models within this class to look for the least complex models to fit the experimental data.

## 4.2 Previous modeling approaches

At first, we describe the fitting procedure used in [34] in more detail as well as our systematic extension of this approach, which ultimately lead to the design of a more powerful and easier to motivate model class ("regulated on-off-slide models") presented in the next section.

After counting the occurrences of the eight promoter configurations, Brown et al. fitted simple Markov models to the different promoter states individually [34]. The simplest model with just assembly reactions and disassembly reactions, each with the same rate, respectively, did not fit the data well (Figure 4.3A, B). After extending this base model with sliding reactions, that move the nucleosome from the center position (N-2) away to the outside positions (N-1 and

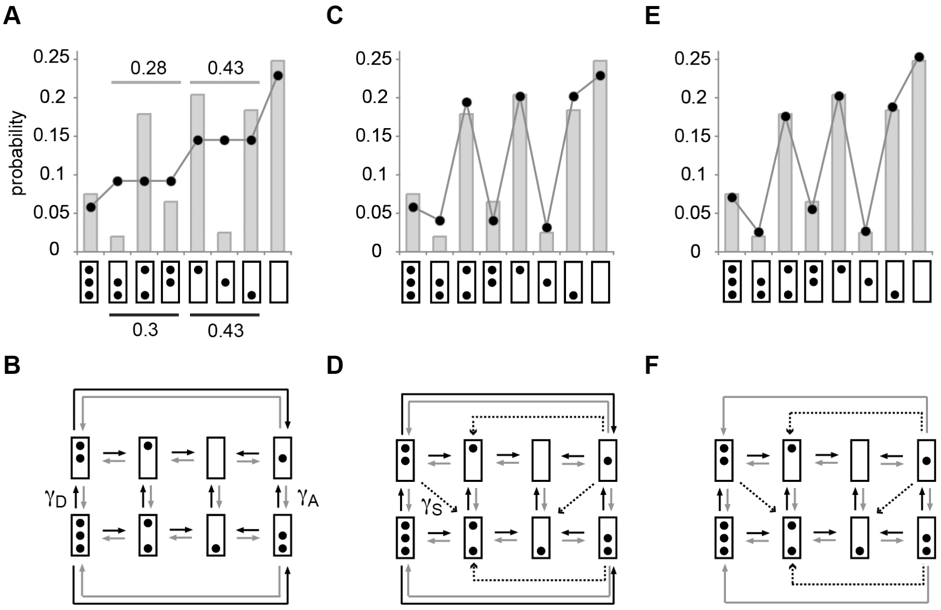


**Figure 4.3:** Model fits to the activated promoter state in the modeling approach from [34]. Top: measured distribution of promoter configurations in the activated promoter state (bars) with best fit steady state distributions (lines with dots) of the different models shown at the bottom: all assembly reactions (black arrows) have the rate $\gamma_A = 1$, disassembly reactions (gray arrows) the rate $\gamma_D$ and sliding reactions (dotted arrows) the rate $\gamma_S$. In this approach the network topology decides which assembly, disassembly and sliding reactions are allowed and the parameter values $\gamma_D$ and $\gamma_S$ were obtained by maximizing the likelihood of the data of a given network topology. To display our models, we use a more symmetric, but equivalent layout (Figure 4.2B). (**A**, **B**) A very simple network with all assembly reactions, all disassembly reactions and no sliding reactions fails to reproduce the measured data. (**C**, **D**) Adding sliding reactions from N-2 to N-1 and N-3 drastically improves the fit. (**E**, **F**) Removing certain disassembly reactions from the network further improves the fit. Adapted from [34] under CC BY 4.0 license.

N-3) the fit improved drastically (Figure 4.3C, D). In a last step, Brown et al. removed two of the four disassembly reactions of the N-1 nucleosome, further improving the fit (Figure 4.3E, F). It is unclear, however, whether they tried to systematically set single reaction rates to zero, to find the best network topology.

Brown et al. then fitted the configuration occurrences of five other promoter states using the model topologies of Figure 4.3D and F, obtaining reasonable good fit results. In each fit, two parameter values were optimized, the disassembly rate $\gamma_D$ and the sliding rate $\gamma_S$, while the assembly rate $\gamma_A$ is set to 1. The network topology is a more abstract “parameter” that has drastic effects on the quality of the fit as well as the biological interpretation of the model. A change in the network topology means that some reactions that were previously allowed are now forbidden, and vice versa. A priori it is unclear, why the activated state and the repressed state should be governed by different network topologies leading to such a drastic and discontinuous change in reaction rates. Thus we set our first goal to systematically search network topologies that simultaneously fit multiple promoter states well.

To investigate different network topologies methodically we first need to formalize the fitting procedure in [34]. To uniquely describe a given network topology we introduce three sets of allowed reactions, i.e. reactions that have a non-zero rate in this topology, for assembly ($R_A$), disassembly ($R_D$) and sliding ($R_S$). Each of these three reaction sets contains pairs of nucleosome promoter configurations $(i, j)$, with each pair representing a reaction from $i$ to $j$ allowed in the given network topology. We call this class of models "reaction set models". We assume that the reaction sets contain enough reactions to form a network connecting all eight configurations, in the sense that each configuration can be reached by every other configuration, leading to a unique steady state. We only allow $R_S$ to be empty, to model systems without sliding reactions. To sum up, any

$$R_A \subset \{(i,j)|i \to j \text{ is an assembly reaction}\} \tag{4.1}$$

$$R_D \subset \{(i,j)|i \to j \text{ is a disassembly reaction}\} \tag{4.2}$$

$$R_S \subset \{(i,j)|i \to j \text{ is a sliding reaction}\} \tag{4.3}$$

$$\gamma_A = 1, \;\; \gamma_D, \gamma_S > 0 \tag{4.4}$$

For example, the most basic network topology shown in Figure 4.3B, is described by the reactions sets $R_A$ containing all assembly reactions, $R_D$ containing all disassembly reactions and, since there is no sliding, $R_S$ containing no reactions. $\gamma_A$, $\gamma_D$ and $\gamma_S$ denote the rate values of the reactions in the corresponding sets. One rate is set to 1 to fix the time scale (here the assembly rate).

To fit a given reaction set model to the available data of nucleosome configurations, we use a maximum likelihood approach as in [34]. First, for a given model $M$, defined by $R_A, R_D, R_S, \gamma_A, \gamma_D, \gamma_S$, the rate transition matrix $Q$ of the time-continuous Markov process is obtained from the reaction sets and rate parameters,

$$Q_{ij} := \begin{cases} \gamma_A & \text{if } (i,j) \in R_A \\ \gamma_D & \text{if } (i,j) \in R_D \\ \gamma_S & \text{if } (i,j) \in R_S \\ 0 & \text{else for } i \neq j \end{cases} \tag{4.5}$$

$$\text{and } Q_{ii} := -\sum_{j \neq i} Q_{ij}. \tag{4.6}$$

The dynamics of the probability distribution $\boldsymbol{p}(t)$ of a time-continuous Markov process is governed by $\frac{\partial}{\partial t} p_j = \sum_i p_i Q_{ij}$, starting from a fixed initial condition. Thus, the steady state distribution of the model, $\boldsymbol{p}$ is the solution to $\sum_i p_i Q_{ij} = 0$. Since each measured promoter nucleosome configuration was independent of the others, the likelihood of the data set $\boldsymbol{n}$, containing the number of occurrences of the eight different configurations for a chosen promoter state, is given by a multinomial distribution,

$$P(\boldsymbol{n}|R_A, R_D, R_S, \gamma_A, \gamma_D, \gamma_S) = \binom{\sum_i n_i}{n_1, n_2, ..., n_8} \prod_i p_i^{n_i} \tag{4.7}$$

with the multinomial coefficient $\binom{\sum_i n_i}{n_1, n_2, ..., n_8} = \frac{(\sum_i n_i)!}{\prod_i n_i!}$.

We have several data sets $\boldsymbol{n}^{(\sigma)}$ corresponding to different promoter states $\sigma$ due to varying cell conditions or mutations [34] that we now aim to fit simultaneously. First, let $M^{(\sigma)*}$ be the model(s) with parameter values that lead to the highest likelihood for each of the data set $\boldsymbol{n}^{(\sigma)}$. This corresponds to individual and completely independent fits of each promoter state $\sigma$,

$$M^{(\sigma)*} := \underset{R_A,R_D,R_S,\gamma_D,\gamma_S}{\arg\max} \quad P(\boldsymbol{n}^{(\sigma)}|R_A,R_D,R_S,1,\gamma_D,\gamma_S). \tag{4.8}$$

If we assume that the different data sets originate from the same network topology, i.e. the same reaction sets, the differences between steady state distributions are the result of different rate parameter values only. However, we need to choose which parameters we allow to vary, i.e. "regulate". If we only regulate one rate parameter, for example $\gamma_D$, to approximate each data set, we get the best model by optimizing as in the definition of $M_D^*$,

$$M_D^* := \underset{R_A,R_D,R_S,\gamma_S,\gamma_D^{(1)},\gamma_D^{(2)},\ldots}{\arg\max} \left(\prod_\sigma P(\boldsymbol{n}^{(\sigma)}|R_A,R_D,R_S,1,\gamma_D^{(\sigma)},\gamma_S)\right). \tag{4.9}$$

Now the values of $\gamma_D^{(1)}, \gamma_D^{(2)}, ...$, are supposed to regulate the steady state distribution in the different promoter states, while $\gamma_S$ and $\gamma_A = 1$ do not change. Correspondingly, we can also regulate $\gamma_S$ (with best model $M_S^*$) or $\gamma_A$ (with best model $M_A^*$, setting one of the other two parameters to 1). In general, a different regulated parameter can lead to different best likelihood values for fixed reaction sets, i.e. the same reaction sets perform differently well, depending on which parameter is regulated. This leaves us with three methods regulating one rate parameter to model regulation between the different promoter states.

If we assume two regulated rate parameters, we get the best model(s) as in the definition of $M_{D,S}^*$. Then the choice of the parameter which is set to one and not regulated is arbitrary and does not change the fit results.

$$M_{D,S}^* := \underset{R_A,R_D,R_S,\gamma_S^{(1)},\gamma_S^{(2)},\ldots,\gamma_D^{(1)},\gamma_D^{(2)},\ldots}{\arg\max} \left(\prod_\sigma P(\boldsymbol{n}^{(\sigma)}|R_A,R_D,R_S,1,\gamma_D^{(\sigma)},\gamma_S^{(\sigma)})\right) \tag{4.10}$$

Note that in this case, the combined likelihood to fit all promoter states is just the product of the likelihoods to fit the promoter states individually, since different promoter states have no common parameter anymore (except the time scale, which only matters when investigating dynamics).

The best parameter values or the best reaction sets might not be unique, which needs to be taken under consideration when investigating the best fit results. The possibilities for choosing $R_A$, $R_D$, $R_S$ are limited by the following considerations: Our eight configurations with 3 sites that can be occupied or not lead to twelve assembly, twelve disassembly and eight sliding reactions (Figure 4.2B). This gives $2^{12} = 4096$ different reactions sets for assembly and disassembly. The theoretical number of sliding reaction sets of $2^8 = 256$. Of course many combinations for only few allowed reaction might not results in a unique steady state and we decided to ignore these.

We tested this systematic modeling approach by analyzing all network topologies with at least 10 assembly reactions in $R_A$, at least 10 disassembly reaction in $R_D$ and $R_S$ representing sliding that is independent of the third site (i.e. the site not involved in the sliding step).

We found several network topologies with good simultaneous fit results regulating two rate parameters between the promoter states (data not shown).

This was already an improvement over the modeling approach in [34], however, we also realized several disadvantages with this approach. (i) The reaction sets needed to be restricted a priori to reduce the total number of models. It seems possible to overcome this issue by additional optimization and computing time, however the approach does not scale well, if the number of nucleosomes to be modeled increases. Each additional nucleosome increases the number of possible reactions and reaction sets drastically, and with it, the total model count. (ii) The four different simultaneous fit methods need to be taken into account. Using two parameters to regulate between different promoter state instead of one always improves the fit, but in the end a reasonable criterion is needed to decide whether the additional parameter values are worth the improvement. (iii) It is unclear how to rigorously compare the complexity of different models. Each has the same number of fit parameters (within the same simultaneous fit method) and only the topology, i.e. the reactions that are set to zero, change. The number of these forbidden reactions could be used as a proxy for model complexity. But deleting several related reactions, for example assembly at the same position, might be achieved by the same biological process, which should then be taken into account. Finally the most important disadvantage is that (iv), the models can only forbid reactions and a less restrictive inhibition or even enhancement is impossible, yet these are very common in biological processes and should be included. A solution to (iv) are additional reaction sets with independent rate parameters that govern a smaller subset of assembly, disassembly or sliding reactions. This idea leads to a new model class ("regulated on-off-slide models") presented in the next section. This class is also designed such that it solves (ii) and (iii) and allows a more straight-forward way to interpret the models biologically.

## 4.3 Effective configuration dynamics with regulated on-off-slide models

In this section we compile a large and unbiased collection of possible models within our class of "regulated on-off-slide models" and then select the models that are consistent with experimental data. Like the already presented reaction set models, regulated on-off-slide models include assembly and disassembly of N-1, N-2 and N-3 nucleosomes as well as nucleosome sliding from one occupied position to an unoccupied neighboring position and can mimic regulated transitions from repressed over weakly activated to fully activated promoter state dynamics without changing the network topology. Unlike reaction set models, we allow several independent parameters to govern reactions of the same type (like assembly, disassembly or sliding) taking into account the total number of fit parameters and the biological interpretation of the involved processes. Instead of the network topology, these involved processes will define each model.

To further restrict the models, in addition to the *PHO5* promoter nucleosome configuration data from [34] of repressed, weakly activated and activated cells, we also used data from our own restriction enzyme accessibility experiments of two different "sticky (= lower accessibility) N-3" mutant promoters, to address the coupling between remodeling of the N-3 and N-2 nucleosome, and two data sets of Flag-/Myc-tagged histone exchange dynamics experiments [12, 13] to obtain a time scale.

### 4.3.1 Regulated on-off-slide models

An on-off-slide model consists of a set of processes for nucleosome assembly, disassembly and possibly sliding that allow reactions from one configuration to another (all possible reactions in Figure 4.2B). In the following we compile a list of increasingly specific processes. Global (i.e. *PHO5* promoter-wide) processes govern all assembly ("A"), disassembly ("D") and sliding ("S") reactions, respectively (Figure 4.4A). Invoking the principle of Occam's razor, we start with global processes and then overwrite some reactions with more specific processes, making the simplest models more and more complex until we find agreement with the considered data. A similar method was used to model combinatorial acetylation patterns on histones, but with fixed global disassembly and without the option of sliding and regulation [119].

A regulated on-off-sliding model is defined by its set of processes combined with the information which of the processes are regulated. "Regulated" processes can adopt a different rate value for each promoter state whereas "constitutive" processes keep the same rate value. Fitting the data of three promoter states, regulated processes have three fit parameters, whereas constitutive processes have one. One degree of freedom of the fit represents the overall time scale of each model.

Brown et al. showed that just only global assembly and disassembly processes are not sufficient to fit the measured occurrences of all eight configurations, even to describe only one promoter state (e.g. the activated promoter in Figure 4.3A). The addition of global sliding is not enough to obtain a good fit (data not shown). There have to be local modifications of at least one global processes (assembly, disassembly or sliding). In [34], and our extension with reaction set models, modifications were introduced by setting certain reaction rates in the reaction network topology to zero, searching for a good network topology in the discrete space of all possible topologies. In contrast, we give regulated on-off-slide models the ability to continuously deviate from the global process rate values for a given set of reactions. The simplest modification are site-specific processes. They are motivated biologically by sequence-dependent effects and recruitment or inhibition of remodeling factors in a site-specific way. For each of the three nucleosome positions N-1, N-2 and N-3, we add the three site-specific assembly processes "A1", "A2" and "A3", and the three disassembly processes "D1", "D2" and "D3", respectively, to the pool of optional processes (see for example Figure 4.4B). While each model has to have a global assembly and a global disassembly process, all additional processes are optional.

If two different processes govern the same reactions within a certain model, as for example the global and a site-specific assembly process, the more specific process, i.e. the one governing the rate values of less reactions, overwrites the more general process, but only for these reactions. This rule allows an increased as well as a decreased rate value for the reactions of the more specific process.

For sliding, five additional processes are included. One process represents sliding reactions leaving from the N-2 site to model start site-specific non-directional sliding ("S2*"). Two processes allow directional sliding between N-1 and N-2 sites and two directional sliding between

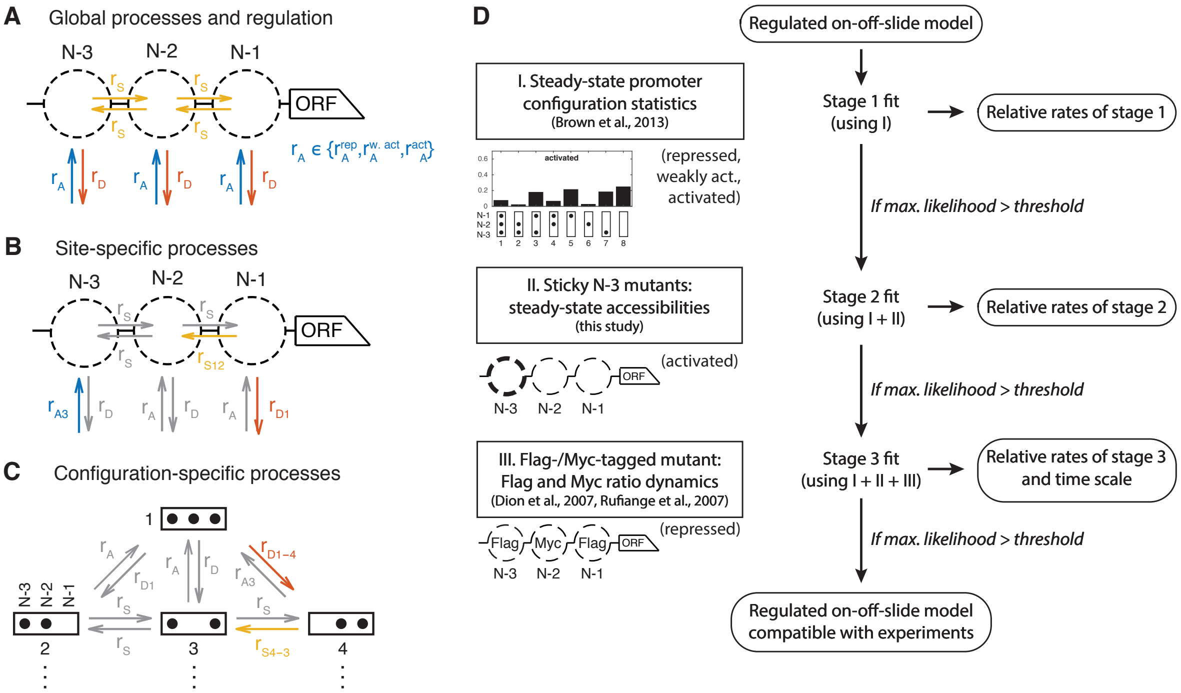

A Global processes and regulation
N-3
N-2
N-1
ORF
r_S
r_A
r_D
r_A ∈ {r_A^rep, r_A^w. act, r_A^act}
B Site-specific processes
N-3
N-2
N-1
ORF
r_S
r_S12
r_A3
r_D
r_A
r_D1
C Configuration-specific processes
1
2
3
4
N-3
N-2
N-1
r_A
r_D1
r_D
r_D1−4
r_A3
r_S
r_S4−3
D
Regulated on-off-slide model
I. Steady-state promoter configuration statistics
(Brown et al., 2013)
activated
(repressed, weakly act., activated)
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8
II. Sticky N-3 mutants: steady-state accessibilities
(this study)
(activated)
III. Flag-/Myc-tagged mutant: Flag and Myc ratio dynamics
(Dion et al., 2007, Rufiange et al., 2007)
Flag
Myc
(repressed)
Stage 1 fit (using I)
Relative rates of stage 1
If max. likelihood > threshold
Stage 2 fit (using I + II)
Relative rates of stage 2
Stage 3 fit (using I + II + III)
Relative rates of stage 3 and time scale
Regulated on-off-slide model compatible with experiments

**Figure 4.4:** Definition and evaluation workflow of regulated on-off-slide models. On-off-slide models are defined by processes from different hierarchies: (**A**) Global processes for assembly, disassembly and (optional) sliding. Global processes govern all reactions of the corresponding type with the same rate $r_A$, $r_D$ or $r_S$. To fit multiple promoter states simultaneously, at least one process has to be regulated, employing different rate values depending on the promoter state. Here, the global assembly process is regulated. (**B**) Optional site-specific processes for assembly and disassembly at each position (for example here with rates $r_{A3}$ and $r_{D1}$) and for sliding between each neighboring pair of positions (here $r_{S12}$). Reactions in gray have not been overruled by more specific processes (here: site-specific processes) and thus are still governed by the rate parameters of processes on the less specific hierarchy level (here: global processes). (**C**) The last hierarchy level consists of optional configuration-specific processes governing only one reaction (here with rates $r_{D1-4}$ and $r_{S4-3}$). Only the promoter configurations 1 to 4 are shown. (**D**) Each regulated on-off-slide model up to a certain number of fit parameters is evaluated successively using the experimental data on the left-hand side (promoter states during the experiment given in parentheses). Models are rejected if they do not match the maximum likelihood threshold after each stage. With each additional experimental data set, the fit can result in different new optimal relative rate values for each model. The dynamic Flag-/Myc-tagged histone measurements enable us to also fit the model time scales in the last stage.

N-2 and N-3 sites, named “Sxy” for sliding from site x to site y. These sliding processes are not only dependent on the state of one site, but on the origin and the destination. The origin has to be occupied and the destination empty. This introduces a correlation between neighboring sites. However, since these processes do not take into account the full configuration, we still call them “site-specific”.

We also include configuration-specific processes which only govern a single reaction rate and overrule any other process for this reaction (see for example Figure 4.4C), to allow even more specific modulations. This gives another 32 optional processes, one for each reaction. They are named by the reaction type and the involved configurations, for instance, the disassembly process from configuration 1 to configuration 4, is denoted with “D1-4”, sliding from configuration 4 to configuration 3 with “S4-3”.

To assess the complexity of a given model, we use the number of fit parameters. Starting with the simplest regulated on-off-slide models having 4 parameters, we increased the maximal parameter number up to 7 to obtain the first models that simultaneously fit all data sets well. 7 parameter values allow models with 1 regulated process and 1 to 4 constitutive processes, as well as models with 2 regulated processes and 1 constitutive processes. Models with zero constitutive processes, i.e. only regulated processes, are ignored since they decouple the time scales of the different promoter states and in this case the fit to steady state observables is not harmed by setting any of the regulated processes to a constitutive process. After going through all combinations of processes with up to 7 parameters and ignoring effectively identical models constructed with different processes (e.g. models where one process is completely overwritten by others, or partially overwritten such that it could replaced by another more specific process) we ended up with 68 145 regulated on-off-slide models in total. The relative occurrence of individual processes in this initial model set is equal for all optional assembly and disassembly processes and slightly lower for non-global sliding processes, since the number of effectively identical, and thus ignored, sliding process combinations is higher (Figure 4.5A).

Comparing this model class with the reaction set models discussed in Section 4.2, regulated on-off-slide models can approximate network topologies of reaction set models by incorporating configuration-specific parameters and setting their rate close to zero. However, regulated on-off-slide models are also able to only slightly decrease or even increase these additional process rates compared to the global rates. This, together with the possibility to overwrite global processes also by site-specific processes yields a much more suitable model class.

Using this new class of models we determined which are capable to reproduce the four experimental data sets in a step by step (“staged”) fitting procedure (Figure 4.4D). This allowed us to dissect the contributions of the different data sets to the model selection and reduced the model count in the later, computationally more involved, stages.

### 4.3.2 Promoter configuration statistics

#### Maximum likelihood fit

In the first stage the parameter values, i.e. the rate values of the involved processes, of each model were fitted by maximizing the likelihood to observe the measured *PHO5* promoter

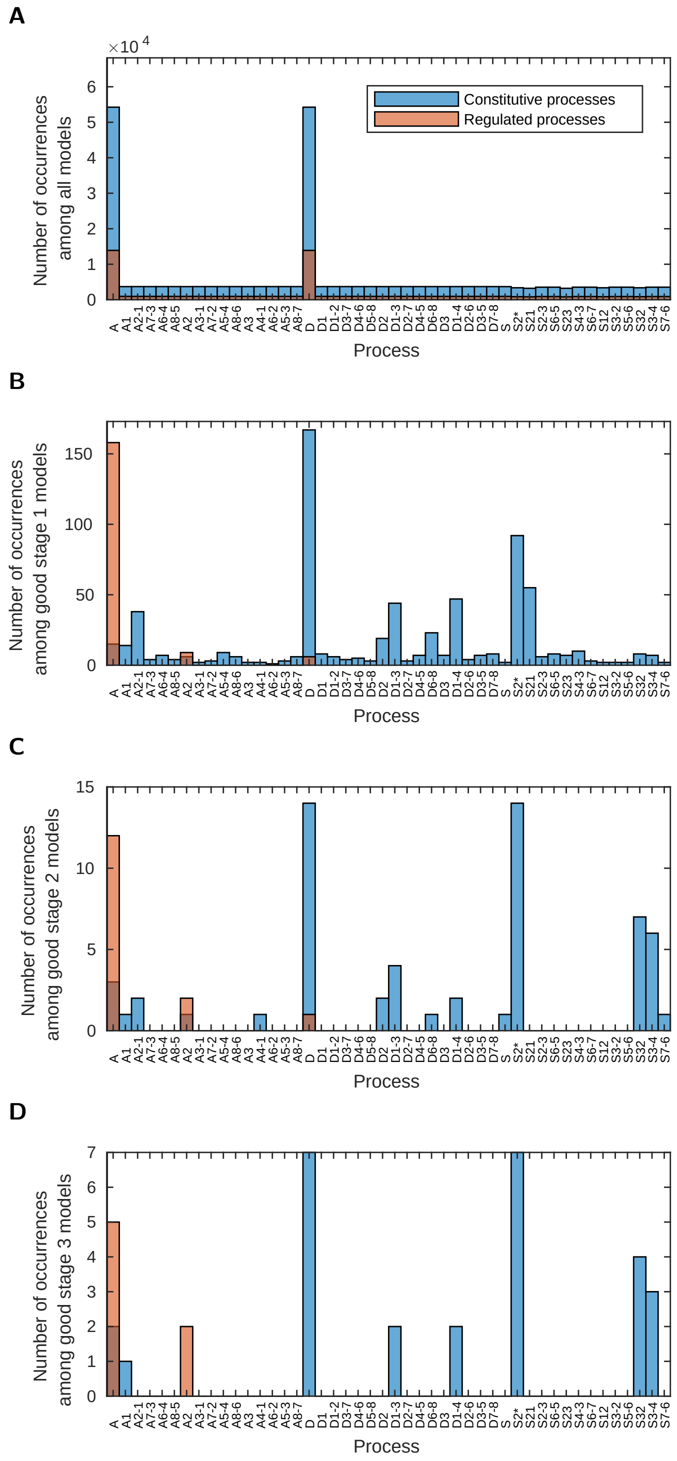


**Figure 4.5:** Occurrences of the different processes in the models with satisfactory likelihood at the different stages. (**A**) in all 68 145 analyzed models, (**B**) in all 173 not rejected models of stage 1, (**C**) in all 15 not rejected models of stage 2, (**D**) in all 7 satisfactory models after stage 3. In each plot the y-axis limit is the number of the considered models allowing the comparison of the relative occurrences as more and more experimental data sets are fitted.

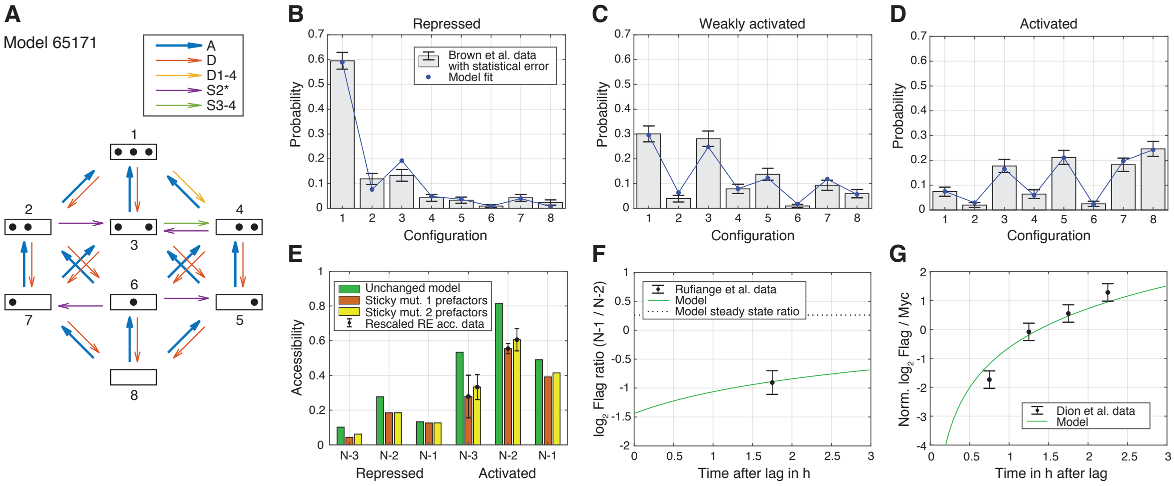


**Figure 4.6:** Best regulated on-off-slide model compatible with all the experimental data sets in stage 3. All fits were done simultaneously. (**A**) Regulated on-off-slide model with the regulated process A (global assembly, thick arrows) and the constitutive processes D (global disassembly), D1-4 (disassembly from configuration 1 to 4, overwrites D), S2* (sliding away from N-2) and S3-4 (sliding from configuration 3 to 4). (**B, C, D**) Combined fits to the steady-state promoter nucleosome configuration occurrences in repressed, weakly activated and activated state. Only the changing rate of the regulated global assembly process accounts for differences in the three distributions. The other processes (D, D1-4, S2* and S3-4) are constitutive and their rates do not change. The model fits in stage 1 (ignoring all other data) and stage 2 (ignoring Flag-/Myc-tagged histone exchange data) are only slightly better for this model (data not shown). (**E**) Fit to the sticky N-3 RE accessibility data of two mutants in the activated state (error bars with standard deviation of rescaled RE accessibility). Only reaction rates involving the N-3 were allowed to divert from the previously fitted parameter values. (**F**) Fit to the Rufiange et al. data [13] of Flag amounts at N-1 over N-2 after 2h (minus lag time) of Flag expression. The error bars are the standard deviations for two measurements. (**G**) Fit to the Dion et al. data [12] of Flag over Myc amounts at N-1 at four time points after Flag expression. y-axis points are normalized by their mean to account for a sloppy fit parameter in the treatment of the data in [12]. Error bars are estimated experimental standard deviations used in the fit.

configurations (Figure 4.2C and Table 4.1). The global assembly parameter (for the activated state) was set to 1 to fix the time scale at this point of the analysis, resulting in relative rates for the remaining 6 parameter values. The model with the processes and reaction network shown in Figure 4.6A serves as an example throughout the fit procedure. After the first maximum likelihood fit, one can compare the different steady state distributions for the three promoter states with the measured configuration statistics. The inclusion of the data sets discussed below into the fit did not visibly worsen the fit of the configuration statistics data for this specific model (data not shown), thus we already show the final fit results of the last stage (Figure 4.6B, C and D).

Specifically, we optimized the parameter values $\boldsymbol{r}$ by maximizing the log10 likelihood $L_I(\boldsymbol{r})$. With the rate parameter notation introduced in Figure 4.4, we have for the example of Figure 4.6A:

$$\boldsymbol{r} = (r_A^{\text{rep}}, r_A^{\text{w. act}}, r_A^{\text{act}}, r_D, r_{D1-4}, r_{S2^*}, r_{S3-4})^\top . \tag{4.11}$$

Let $Q(\boldsymbol{r}^\sigma)$ be the transition rate matrix of the Markov process defined by the model for promoter state $\sigma$, where $\boldsymbol{r}^\sigma$ denotes the vector containing the regulated parameter value(s) of promoter state $\sigma$ and all constitutive parameter values. A non-diagonal entry $Q_{ij}(\boldsymbol{r}^\sigma)$ is the rate to go from configuration $i$ to configuration $j$ and is non-zero only for valid assembly, disassembly and sliding reactions and then given by the entry of $\boldsymbol{r}^\sigma$ which holds the parameter value of the process that governs this reaction in the given model. If sliding reactions are not governed by any sliding process within the model, their rate is set to zero. Diagonal entries are given by $Q_{ii}(\boldsymbol{r}^\sigma) = -\sum_{j\neq i} Q_{ij}(\boldsymbol{r}^\sigma)$. In the example of Figure 4.6A, the transition rate matrix in the activated state is given by (with "..." representing the diagonal entries)

$$Q(\boldsymbol{r}^{\text{act}}) = \begin{pmatrix} \dots & r_D & r_D & r_{D1-4} & 0 & 0 & 0 & 0 \\ r_A^{\text{act}} & \dots & r_{S2*} & 0 & 0 & r_D & r_D & 0 \\ r_A^{\text{act}} & 0 & \dots & r_{S3-4} & r_D & 0 & r_D & 0 \\ r_A^{\text{act}} & 0 & r_{S2*} & \dots & r_D & r_D & 0 & 0 \\ 0 & 0 & r_A^{\text{act}} & r_A^{\text{act}} & \dots & 0 & 0 & r_D \\ 0 & r_A^{\text{act}} & 0 & r_A^{\text{act}} & r_{S2*} & \dots & r_{S2*} & r_D \\ 0 & r_A^{\text{act}} & r_A^{\text{act}} & 0 & 0 & 0 & \dots & r_D \\ 0 & 0 & 0 & 0 & r_A^{\text{act}} & r_A^{\text{act}} & r_A^{\text{act}} & \dots \end{pmatrix} . \tag{4.12}$$

The steady state distribution $p_{i\sigma}$ of $Q(\boldsymbol{r}^\sigma)$ is the solution of $p_{j\sigma} = \sum_i p_{i\sigma} Q_{ij}(\boldsymbol{r}^\sigma) = 0$. Then $L_I(\boldsymbol{r})$ can be calculated using the multinomial distribution, similar to (4.7),

$$L_I(\boldsymbol{r}) = \sum_\sigma \log_{10} \left[ \frac{(\sum_i n_{i\sigma})!}{\prod_i n_{i\sigma}!} \prod_i p_{i\sigma}^{n_{i\sigma}} \right] , \tag{4.13}$$

with $n_{i\sigma}$ being the number of observations of the corresponding promoter configurations (Table 4.1).

We maximized $L_I(\boldsymbol{r})$ numerically for each tested model, using the MATLAB function fmincon. To obtain the steady state distribution to high numeric accuracy we used the state reduction algorithm [120, 121, 122]. The range of parameter values was limited to $[10^{-2}; 10^2]$, with 1 being the rate value of the global assembly process for the activated state, thus all assembly and disassembly reactions have a rate greater than zero, ensuring a unique steady state for each model. A wider range of $[10^{-3}; 10^3]$ did not affect the fit results. We used 100 random sets of initial parameter values for each model to ensure a robust maximum. In 3.6% of all models the 100 tries found at least two different maximal likelihood values, which were always extremely low. In 2.4% of all models, the found maximum likelihood parameter values were not unique. However, none of these problematic models were among models with relatively high maximal likelihoods.

### Stage 1 fit results

To select the models that advance to the next stage, we used the logarithmic likelihood ratio with respect to the perfect fit, $R_1 = -L_I + L_0$, where $L_0$ is the highest possible log10 likelihood to obtain the configuration data, corresponding to a perfect fit. Thus, $L_0$ is calculated by taking the relative occurrences of the configurations as parameters in a multinomial distribution and then calculating the likelihood for these occurrences. The best model achieved $R_1 = 4.02$ (distribution of $R_1$ for all models in Figure C.1A) and our example model in Figure 4.6 reached $R_1 = 4.38$. To define models that are in good agreement with the measured configuration statistics, we set an upper threshold to $R_1$, $R_{max} = 6$. We used the same $R_{max}$ in the following stages and this threshold, corresponding to a likelihood $\approx 100$ times lower than the current best model, gave enough room for fitting models to additional data. We found 173 models with $R_1 < R_{max}$, with the top 30 models presented in Figure C.2. Two of these 173 models used only 5 instead of the 6 free fit parameters (with $R_1 = 5.32$ and 5.88, Figure C.1D and Figure C.3).

For a first comparison with the performance of the models used in [34] (Figure 4.3D and F), we calculated the corresponding logarithmic likelihood ratio, using the best network topologies (network D or network F) to fit the different promoter states, as described in [34]. This combined model has effectively 6 fit parameters (disassembly and sliding for each state, with assembly set to 1) and results in $R_1 = 4.93$. Thus, we found models with higher likelihood, as well as models with one parameter less and not much worse likelihood, proving the power of our systematic approach.

The set of models with $R_1$ below the threshold was still quite heterogeneous, making it difficult to state biological interpretations. For example there were models where all sliding reactions have positive rates, and some models with no sliding at all (Figure C.1C), like the second best model (Figure C.2). Some models did not use any configuration-specific processes (Figure C.1B). The regulated processes were mostly global assembly, while some models used regulated global disassembly or site-specific assembly at N-2 (Figure 4.5B). In the following stages, we aimed to further sort out models by including additional data sets to obtain a more homogeneous picture.

#### Assembly-disassembly symmetry in equilibrium models

Note that most regulated on-off-slide models are non-equilibrium models. Equilibrium regulated on-off-slide models are models that fulfill detailed balance or, equivalently, where the net fluxes between all promoter configurations in steady state are zero in all activation states. These are models that only use global processes, where the symmetry is trivial, as well as models without any sliding nor any configuration-specific processes and they have an assembly-disassembly symmetry. That means, swapping all assembly processes to the equivalent disassembly process and vice versa (e.g. A to D, A1 to D1, D2 to A2), yields a model with equal maximum likelihood (at possibly different parameter values). Consequently, for each equilibrium model with a regulated assembly parameter there is a symmetric equilibrium model with the corresponding regulated disassembly parameter and equal maximum likelihood, and vice versa. Non-equilibrium models, however, do not share this symmetry. There are 196 equilibrium regulated on-off-slide models among the 68 145 investigated models, and none of them fit the data well, with the best equilibrium model having $R_1 \approx 11$.

### 4.3.3 Integration of *PHO5* promoter mutant data

#### Accessibility experiments for *PHO5* promoter mutants

In the following stages, we fitted the models that passed stage 1 to additional experimental data and sorted out models that are not able to fit all data simultaneously. In the second

| | Wild type | Sticky N-3 mutant 1 | | Sticky N-3 mutant 2 | |
|---|---|---|---|---|---|
| | accessibility | accessibility | wt fold-change | accessibility | wt fold-change |
| N-2 RE (ClaI) | $(64.0 \pm 1.4)\,\%$ | $(43.5 \pm 2.1)\,\%$ | $0.68 \pm 0.04$ | $(47.5 \pm 4.9)\,\%$ | $0.74 \pm 0.08$ |
| N-2 Brown et al. | $82\,\%$ | | | | |
| N-3 RE (HhaI) | $(58.5 \pm 2.1)\,\%$ | $(30.5 \pm 13.4)\,\%$ | $0.52 \pm 0.23$ | $(36.5 \pm 7.8)\,\%$ | $0.62 \pm 0.13$ |
| N-3 Brown et al. | $55\,\%$ | | | | |

**Table 4.2:** Restriction enzyme (RE) accessibility of N-2 and N-3 sites in phosphate starved, i.e. activated, cells measured in this study and corresponding accessibility values of Brown et al. [34] (RE accessibility with mean ± standard deviation of two independent biological replicates and the fold-change standard deviation calculated using standard error propagation). The sticky N-3 mutants feature manipulated DNA sequences at the N-3 site, which decrease the RE accessibility at the N-3 site compared to the wild-type. In the RE measurements, this sticky N-3 also decreases the accessibility of the N-2 site. From stage 2 on, we tested which regulated on-off-slide models with compatible configuration distribution in stage 1 can at the same time yield the accessibility fold-changes at sites N-2 and N-3 for both sticky N-3 mutants.

stage we used data from two *PHO5* promoter mutants published by Small et al. in [69] where the DNA sequence underlying N-3 was mutated such as to increase certain dinucleotide periodicities that favor nucleosome formation and may increase intrinsic nucleosome stability [123]. With an at that time new DNA methylation assay for probing nucleosome occupancies and configurations, these authors claimed that N-3 at these mutated *PHO5* promoters was hardly removed after *PHO5* induction. Thus these mutated *PHO5* promoters offered an interesting parameter modulation and we could have tested our models using these nucleosome occupancy data. However, we questioned the published data (see Section 4.4) and checked them by classical and well-documented restriction enzyme (RE) accessibility assay [124] (see Section C.1.1). Using the exact same strains from Small et al., we measured N-2 and N-3 occupancies at the wild type and mutant *PHO5* promoters (Table 4.2). Our experiments confirmed that the sticky N-3 mutant promoters show reduced removal of both N-2 and N-3 upon *PHO5* induction, but not that these nucleosomes were hardly removed at all. We used our data as additional constraints for our models in stage 2 and designed a method to investigate which models show the same interdependence of N-2 and N-3 accessibility.

### Accessibility fold-changes

The Small et al. strains included corresponding isogenic wild type strains. Our measured RE accessibilities for this wild type in the activated state were lower than the accessibilities in the activated state of the wild type calculated from the data from Brown et al. [34] (Table 4.2), a discrepancy that likely stems from different experimental conditions. Brown et al. used a different strain background, YS18, and the *pho80* allele in high phosphate conditions for induction, while we used the S288C background and over night phosphate starvation to achieve direct comparison with the Small et al. data. We found repeatedly that S288C strains do not result in as high ClaI accessibility values in the induced state as the YS18 background, which was formerly used in classical *PHO5* studies reporting such high degree of *PHO5* promoter nucleosome remodeling ([125] and data not shown). We bypassed this discrepancy by using the accessibility fold-changes of mutants compared to the wild-type (Table 4.2) to investigate our models, effectively normalizing the accessibility values coming from different experiments.

### Modeling approach

We opted to fit the experimental fold-changes together with the stage 1 data, minimizing $R_2 = -(L_I + L_{II}) + L_0$ with $L_{II}$ being the log10 likelihood to obtain the accessibility fold changes. We ignored additive constants to the log10 likelihood, so that after including the next data set into the fit, a perfect agreement between model and the additional data set already with the values $\boldsymbol{r}$ from the previous stage, would lead to the same log likelihood value as before. Since the mutants only differ in the DNA sequence at the N-3 position from the wild type, we needed to test for each model whether reasonably large changes in reaction rates involving the N-3 nucleosome could reproduce the experimental behavior. Since the exact consequences of the sticky N-3 mutation on the reaction rates are unknown, we systematically considered many different possibilities of changes in reactions where the N-3 nucleosome is involved. This was achieved by including prefactors $\boldsymbol{\kappa}^m$ for these 12 reaction rates (Figure 4.7), which were fitted together with the model parameters $\boldsymbol{r}$ to the configuration statistics data and the accessibility fold changes of both sticky N-3 mutants ($m = 1$ or $m = 2$) and allowed each prefactor to vary between 1/5 and 5. Assuming the experimental fold changes are normally distributed the log10 likelihood of a model to reproduce the new data up to an additive constant is given by

$$L_{II}(\boldsymbol{r}) \ln(10) = \sum_{m=1}^{2} \max_{\boldsymbol{\kappa}^m} \sum_{s=2}^{3} -\frac{(f_s(\boldsymbol{r}, \boldsymbol{\kappa}^m) - f_{sm}^{\text{mean}})^2}{2 f_{sm}^{\text{var}}}, \tag{4.14}$$

with $f_{sm}^{\text{mean}}$ and $f_{sm}^{\text{var}}$ being mean and variance, respectively, of the measured accessibility fold changes in active state of sticky N-3 mutant $m$ at nucleosome site $s$ (2 for N-2 and 3 for N-3), $f_s(\boldsymbol{r}, \boldsymbol{\kappa}^m)$ being the corresponding model fold change, and $\boldsymbol{\kappa}^m$ being the values of the rate prefactors of sticky mutant $m$.

To obtain $f_s(\boldsymbol{r}, \boldsymbol{\kappa}^m)$ for each model, we calculated a modified transition rate matrix for each mutant using the non-diagonal part of the transition rate matrix $Q(\boldsymbol{r}^{\text{act}})$ and multiplied it component-wise with the matrix $W(\boldsymbol{\kappa}^m)$ containing the prefactor values $\boldsymbol{\kappa}^m$ for the affected reactions for mutant $m$ (and 1 otherwise). Using the modified transition rate matrices we calculated the mutant steady state distributions and finally the corresponding fold ratios of accessibilities at N-2 and N-3.

We used 4 prefactors per sticky N-3 mutant, one for each group of reactions, assembly at N-3, disassembly at N-3, sliding from N-3 to N-2 and from N-2 to N-3 (Figure 4.7), respectively, leading to $\boldsymbol{\kappa}^m = (\kappa_{a3}^m, \kappa_{d3}^m, \kappa_{s23}^m, \kappa_{s32}^m)^\top$. The off-diagonal part of $W(\boldsymbol{\kappa}^m)$ is then given by

$$W(\boldsymbol{\kappa}^m) = \begin{pmatrix} ... & 1 & 1 & \kappa_{d3}^m & 1 & 1 & 1 & 1 \\ 1 & ... & 1 & 1 & 1 & \kappa_{d3}^m & 1 & 1 \\ 1 & 1 & ... & \kappa_{s32}^m & \kappa_{d3}^m & 1 & 1 & 1 \\ \kappa_{a3}^m & 1 & \kappa_{s23}^m & ... & 1 & 1 & 1 & 1 \\ 1 & 1 & \kappa_{a3}^m & 1 & ... & 1 & 1 & 1 \\ 1 & \kappa_{a3}^m & 1 & 1 & 1 & ... & \kappa_{s23}^m & 1 \\ 1 & 1 & 1 & 1 & 1 & \kappa_{s32}^m & ... & \kappa_{d3}^m \\ 1 & 1 & 1 & 1 & 1 & 1 & \kappa_{a3}^m & ... \end{pmatrix}, \tag{4.15}$$

where the diagonal does not matter due the component-wise multiplication with the non-diagonal part of $Q(\boldsymbol{r}^{\text{act}})$ in order to obtain the non-diagonal part of the mutant rate transition matrices.

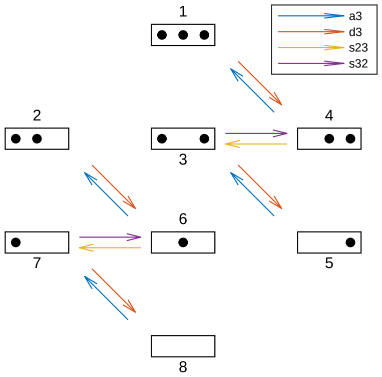


**Figure 4.7:** Reactions involving the N-3 position. For each model the reaction rates are governed by the model's processes as in stage 1. "a3", "d3", "s23" and "s32" define sets of reactions. Depending on the model, reactions in any of the four sets can be governed by different processes. To investigate models for the compatibility with the sticky N-3 mutation experiments, the model's reaction rates of assembly at N-3 (a3), disassembly at N-3 (d3), sliding from N-2 to N-3 (s23) and sliding from N-3 to N-2 (s32), each obtain a prefactor whose values are found by maximizing the combined likelihood of the configurational data and the sticky N-3 mutant accessibility fold-changes. The prefactors have no direct effect on the fit to the configurational data, but the combined likelihood fit allows a trade-off when a change in the process rates might benefit the accessibility fold change fit more than the configurational data fit.

We also tested 12 prefactors per mutant, one for each affected reaction, giving the mutations even more freedom to change the reaction networks. This test yielded almost the same maximum likelihood values for most models and two additional good models at this stage which then dropped out in the final stage. Note that the exact values of prefactors found during the optimization depended on their initial condition, as their best values were often sloppy or not unique, but still resulted in the same maximum likelihood.

For equilibrium models, the ratio of a pair of reverting reaction rates can be interpreted by $e^{\Delta E/\mathrm{k_B T}}$, with $\Delta E$ denoting the energy difference between the two configurations. This leads to a maximal possible absolute change in N-3 nucleosome binding energy modeled by the prefactors of $|\Delta E_{\text{wild type}} - \Delta E_{\text{N-3 mutant}}| \approx 3.2\,\mathrm{k_B T}$. An increased prefactor range between 1/10 and 10 did only improve the results of some models, but not yield more satisfactory models in the last stage.

Using the same likelihood threshold as in stage 1, stage 2 resulted in 15 models that fit both sticky N-3 mutants well (Figure C.4A), with the best logarithmic likelihood ratio of $R_2 = 4.38$ for the model in Figure 4.6. The example in Figure 4.6E shows a perfect fit to the sticky N-3 fold changes (including the data in stage 3), indicating that for this model, the prefactors had suitable influence to exactly reproduce the experimental fold changes. Since many models did not pass the threshold at this stage anymore, this influence strongly depends on the individual regulated on-off-slide model.

Out of these 15 models, only three models without configuration-specific processes remained (Figure C.4B). Furthermore, we also excluded models without sliding processes (Figure C.4C) as well as models with less than seven fit parameters (Figure C.4D) during this stage.

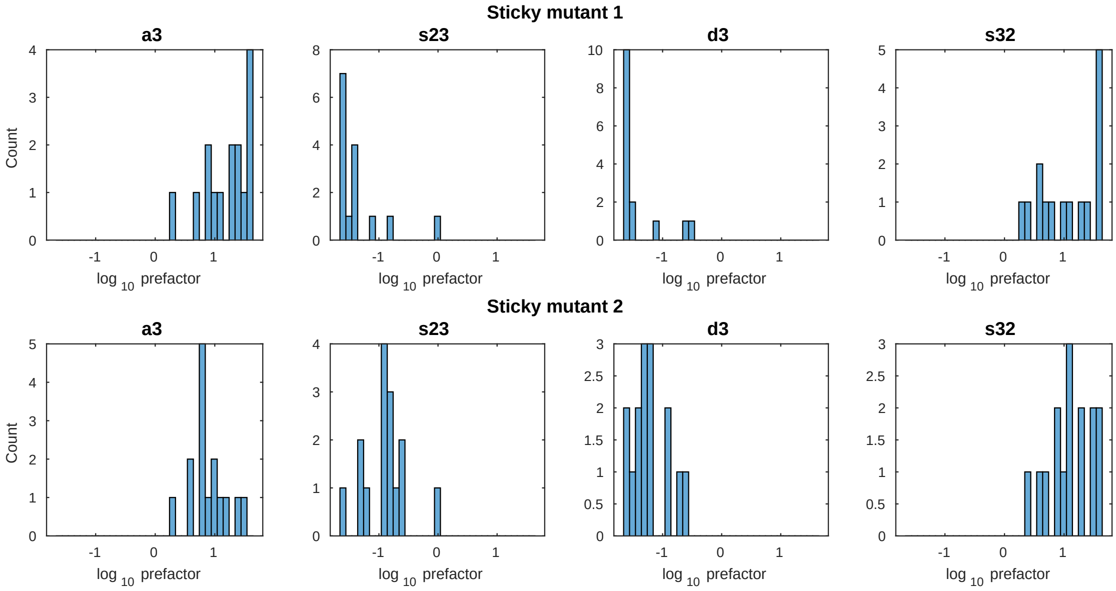


**Figure 4.8:** Histograms of the four prefactor values for the reaction sets a3, s23, d3 and s32, respectively, for both sticky mutants and all 15 models with maximum likelihood above the threshold in stage 2. To achieve the experimental accessibility fold changes, assembly at N-3 and sliding from N-3 to N-2 was increased for most models, while disassembly at N-3 and sliding from N-2 to N-3 was decreased by the prefactor values.

The fitted prefactor values showed the expected qualitative behavior for the models with $R_2 < R_{max}$ in both mutants (Figure 4.8): consistent with the reduced experimental accessibility at N-3, assembly at N-3 was increased by a prefactor greater than 1, while disassembly at N-3 was decreased by a prefactor smaller than 1. Furthermore, the prefactors of sliding reactions decreased sliding from N-2 to N-3 and increased sliding from N-3 to N-2. While this favored higher accessibility at N-3 again, it lead to the concomitant accessibility decrease at N-2.

Using the sticky N-3 data in our fit had a strong impact on the relative occurrences of sliding processes, as it strongly favors models with either S32 or S3-4 processes, both of which allow the sliding from N-3 to N-2 (Figure 4.5C). Now, in stage 2, all 15 good models include at least one sliding process (Figure C.5).

### 4.3.4 Flag-/Myc-tagged nucleosome exchange modeling

In stage 3, we further constrained our models using histone dynamics measurements and determined, for the first time, the optimal time scale of each model. Histone dynamics measurements reflect the appearance/disappearance of nucleosomes regardless if via assembly/disassembly or sliding. These dynamics are measured in cells that constitutively express Myc-tagged H3 histones and are then induced to express also Flag-tagged H3 histones, which, over time, are incorporated into nucleosomes [110]. We used the histone pool model as well as the measured average Flag over Myc amount ratio at the N-1 position in MNase-ChIP-chip assays of Dion et al. [12] and the measured Flag at N-1 over Flag at N-2 amount ratio in Flag-tagged H3 MNase-ChIP-qPCR assays of Rufiange et al. [13]. To investigate the nucleo-

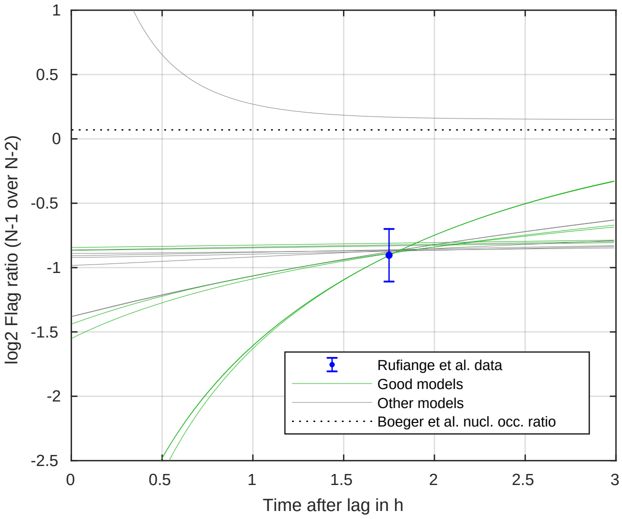


**Figure 4.9:** Flag amount ratio of N-1 over N-2, calculated for all models with satisfactory likelihood after stage 2. Green lines represent the satisfactory models of stage 3. The steady state ratio, that eventually all models should reach closely, is indicated by the dashed line. The three groups of green lines match the three model groups in the text (Section 4.3.6) and Figure 4.11: almost constant lines (group 3), slightly rising lines (group 1) and quickly rising lines (group 2).

some configuration dynamics of each model we kept track which nucleosome contains a Flag- or Myc-tagged histone, given the theoretical Flag- and Myc-tagged histone distribution in the histone pool. This enabled us to calculate the dynamics of the ratio of Flag at N-1 over Flag at N-2 amount (Figure 4.6F and Figure 4.9) and the ratio of Flag over Myc amount at the N-1 (Figure 4.6G and Figure 4.10) for each model in order to determine the agreement with the two histone dynamics data sets. We obtained 7 models in good agreement with all data sets and a best logarithmic likelihood ratio of $R_3 = -(L_I + L_{II} + L_{III}) + L_0 = 4.61$, with $L_{III}$ being the summed log10 likelihoods of the third and fourth data set. Thus, the threshold value of $R_{max} = 6$ denotes a likelihood $\approx 25$ times lower as the best achieved value of $R_3$.

The following technical part describes the optimization procedure in stage 3 in more detail, starting with the histone pool and nucleosome turnover model in the Dion et al. study. To obtain the Flag and Myc amounts in a given model with given parameter values, we assumed that the Myc H3 and Flag H3 amounts in the histone pool are given by

$$\begin{aligned} M(t) &= \frac{\alpha_M}{\beta_M} \\ F(t) &= \begin{cases} 0, \text{ for } t < t_0 \\ \frac{\alpha_F}{\beta_F}(1 - e^{\beta_F(t-t_0)}), \text{ for } t \geq t_0, \end{cases} \end{aligned} \tag{4.16}$$

where we used the production rates $\alpha_F = 50\,/\,\text{min}$, $\alpha_M = 10\,/\,\text{min}$, the degradation rates $\beta_F = 0.01\,/\,\text{min}$, $\beta_M = 0.03\,/\,\text{min}$ and the lag time $t_0 = 15\,\text{min}$ which were fitted in [12]. For

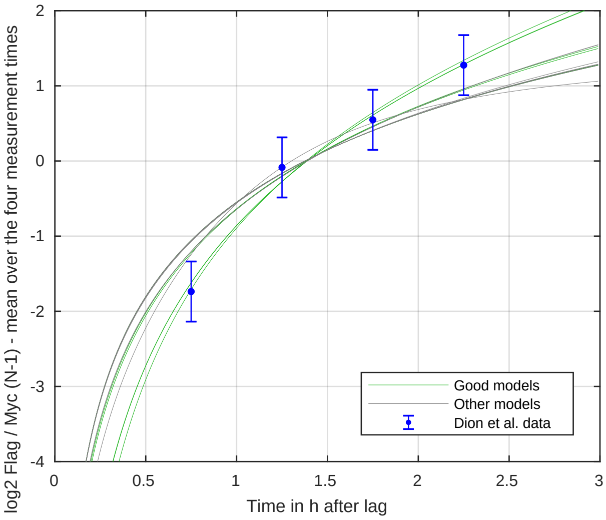


**Figure 4.10:** Shifted Flag over Myc amount ratio at N-1 position, calculated for all models with satisfactory likelihood after stage 2. Green lines represent the satisfactory models of stage 3. Errorbars indicate the estimated standard deviations before the linear transformation needed to remove a sloppy component in the data as described in the text.

$t > t_0$, the probability that a newly assembled nucleosome contains a Flag H3 is given by

$$\begin{aligned} P^+(t|N) &= \frac{F(t)}{F(t)+M(t)} \\ &= 1\Big/\left(1+\frac{\alpha_M\beta_F}{\alpha_F\beta_M}\Big/\left(1-e^{-\beta_F(t-t_0)}\right)\right) \\ &\to 1\Big/\left(1+\frac{\alpha_M\beta_F}{\alpha_F\beta_M}\right) = 0.9375, \text{ for } t\to\infty. \end{aligned} \tag{4.17}$$

Following Dion et al., the conditional probability that a given nucleosome at site $l$ at time $t$ contains a Flag H3 then fulfills the ordinary differential equation

$$\frac{d}{dt}P_l(t|N) = \lambda_l\left(P^+(t|N) - P_l(t|N)\right), \tag{4.18}$$

with $\lambda_l$ being an effective turnover rate at probe position $l$. In our case, the dynamics of the three promoter nucleosomes are coupled, determined by the transition rate matrix $Q(\boldsymbol{r}^\sigma)$ of a given regulated on-off-slide model. At this stage, we included different nucleosome types (i.e. Flag and Myc) into the model, replacing the eight promoter configurations by all 27 possibilities to arrange no, a Flag or a Myc nucleosome at each of the three sites. Based on $Q(\boldsymbol{r}^\sigma)$ and $P^+(t|N)$, we define an extended Flag/Myc transition rate matrix $E(\boldsymbol{r}^\sigma, P^+(t|N))$. Each "new" assembly reaction rate in $E(\boldsymbol{r}^\sigma, P^+(t|N))$ is given by the corresponding "old" assembly rate in $Q(\boldsymbol{r}^\sigma)$ times either $P^+(t|N)$ or $1-P^+(t|N)$, for a new Flag or Myc nucleosome, respectively. To find the corresponding "old" reaction any extended Flag/Myc configuration is projected to one of the eight normal nucleosome configurations simply by ignoring the Flag/Myc tag information. For example, denoting Flag- and Myc-tagged nucleosomes with "F" and "M", respectively, an assembly reaction from the state (F, M, 0) to the state (F, M,

M) in the extended model corresponds to an assembly reaction from state (1, 1, 0) to the state (1, 1, 1) in the normal model, and its reaction rate is multiplied by $1 - P^+(t|N)$ in the extended model, since the new nucleosome is Myc-tagged. The rates of sliding and disassembly of Flag or Myc nucleosomes are assumed to be equal to the corresponding normal sliding and disassembly rates. The probability of extended configuration $i$ at time $t$ is the $i$-th entry of $\boldsymbol{q}^*(t)$, the solution of

$$\frac{\partial}{\partial t}\boldsymbol{q} = E^\top \left(\boldsymbol{r}^\sigma, P^+(t|N)\right) \boldsymbol{q}, \tag{4.19}$$

where $\sigma$ is fixed in the repressed state, in which all histone exchange experiments took place. The log2 ratios of Flag at N-1 over Flag at N-2 amount, $g(\boldsymbol{r}, t)$, and Flag over Myc amounts at N-1, $h(\boldsymbol{r}, t)$, of each model then correspond to log2 ratios of sums of $q_i^*(t)$ over suitable configurations $i$ with Flag or Myc nucleosomes at the wanted sites.

Regarding the actual maximum likelihood fit, $L_{III}$ has two contributions, one for each histone H3 exchange experiment:

$$L_{III}(\boldsymbol{r}) = L^1_{III}(\boldsymbol{r}) + L^2_{III}(\boldsymbol{r}) \tag{4.20}$$

For the first contribution, to fit the data from Rufiange et al., we used

$$L^1_{III}(\boldsymbol{r}) \ln(10) = -\frac{(g(\boldsymbol{r}, t') - g^{\text{mean}})^2}{2g^{\text{var}}} \tag{4.21}$$

with $g^{\text{mean}}$ and $g^{\text{var}}$ being the mean and variance, respectively, of the measured log2 ratios of Flag amounts at N-1 over N-2 (Flag-H3 MNase-ChIP in [13], ratio values 0.591 and 0.483 for replicate 1 and 2, respectively) and $g(\boldsymbol{r}, t')$ the corresponding log2 ratio of the model as described above for measurement time $t' = 2\,\text{h}$ (not corrected for the lag time).

For the second contribution, let $h_j^{\text{mean}}$ denote the measured normalized mean log2 ratios of Flag amount over Myc amount at N-1, with $j = 1, 2, 3, 4$ indicating the four different time points. We obtained $\boldsymbol{h}^{\text{mean}} = (-0.417, 1.24, 1.87, 2.60)^\top$ from Dion et al. as follows: we recalculated the normalization constant of each time point using the measured mean log2 Myc/Flag signal ratios as described (supplementary material of [12] using the whole-genome commercial microarrays (Agilent) data with the nucleosome pool parameters as in the section above) and then took the normalized results of the probe at the N-1 position of the *PHO5* promoter (chr2:431049-431108). Unfortunately, neighboring probes were only in linker regions between promoter nucleosome positions. As mentioned in Dion et al. and corroborated by our own calculations, the values $h_j^{\text{mean}}$ have large uncertainties, mostly due to an additive sloppy global normalization constant leading to systematic errors, while the differences between time points were determined with reasonable accuracy. Thus, we decided to fit the measured values only after a transformation that eliminates the sloppy global normalization constant by choosing the average over the four time points as a reference: $\tilde{h}_j^{\text{mean}} = h_j^{\text{mean}} - 1/4 \sum_{k=1}^4 h_k^{\text{mean}}$. Let $C$ be the resulting covariance matrix after this linear transformation, assuming an independent estimated experimental standard deviation of 0.4 before the transformation. This estimate was informed by the standard deviation of the Rufiange et al. data as well as from perturbations of the nucleosome pool parameters when recalculating the normalization constants. The corresponding normalized values of the model are denoted by $\tilde{h}_j(\boldsymbol{r})$, calculated from the log2 ratios of Flag amount over Myc amount at N-1, $h(\boldsymbol{r}, t_j)$, using the same transformation, with $t_j$ denoting the measurement time points. Since the four measurements at different time points were linearly mapped to always have an average of zero, the estimated density follows a degenerate multivariate normal distribution with the covariance matrix $C$. Thus the log10 likelihood can be calculated with

$$L^2_{III}(\boldsymbol{r})\ln(10) = -\frac{1}{2}\sum_{j=1}^{4}\sum_{k=1}^{4}(\tilde{h}_j(\boldsymbol{r}) - \tilde{h}_j^{\mathrm{mean}})C^+_{jk}(\tilde{h}_k(\boldsymbol{r}) - \tilde{h}_k^{\mathrm{mean}}), \tag{4.22}$$

where $C^+$ is the Moore-Penrose inverse (pseudoinverse) of $C$.

### 4.3.5 Sensitivity analysis

To determine how sloppy the found best parameter values for a given model are, we performed a simple sensitivity analysis: we calculated the log10 likelihood $L_I + L_{II} + L_{III}$ along certain directions from the best fit point in logarithmic parameter space. An approximation of the likelihood function by a second order Taylor expansion at the best fit point worked only in a small area, as expected in a highly non-linear setting, and too small to properly estimate parameter sloppiness.

As a compromise between properly scanning the parameter space and computational feasibility, we decided to use the following test directions: each fitted parameter value individually, the eigenvectors of the numerical Hessian of the likelihood function at the best fit, as well as the numerical gradient, which can be non-zero if the best fit point lies on the boundary of the parameter space. We ignored the parameter boundaries during the sensitivity analysis, to also take into account sloppiness that "reaches over" the boundary. Along these directions we tested in exponentially increasing steps from the best fit position which positions in parameter space lead to a decrease of the likelihood by $\approx 50\%$, i.e. a log10 likelihood ratio change of $\approx 0.30$. This change is of similar order as the log10 likelihood differences within our group of satisfactory models. "Error bars" were calculated for each parameter by taking the largest deviation of the log10 parameter value from the best value at the 50% likelihood level found in all tested directions (Table 4.3 and Table 4.4).

### 4.3.6 Properties of satisfactory models

#### Processes and rate values

Of the 68 145 tested models with at most 7 fitted parameter values, 7 models satisfied our threshold for the likelihood in the combined fit in the last stage to the nucleosome configuration data, the sticky N-3 mutant accessibility data and the H3 exchange data. These “satisfactory” models, their processes and rate values are presented in Figure 4.11. Figure 4.12 gives an overview of the same results from a different perspective, showing all 32 reaction rates in each model.

All satisfactory models had one regulated process and 4 constitutive processes (Figure 4.11). Models with regulation by two processes in combination with 1 constitutive process (also summing up to 7 fit parameters) did not fit the data well enough, likely because regulation with 2 processes leaves only 1 possible constitutive process slot and only three processes in

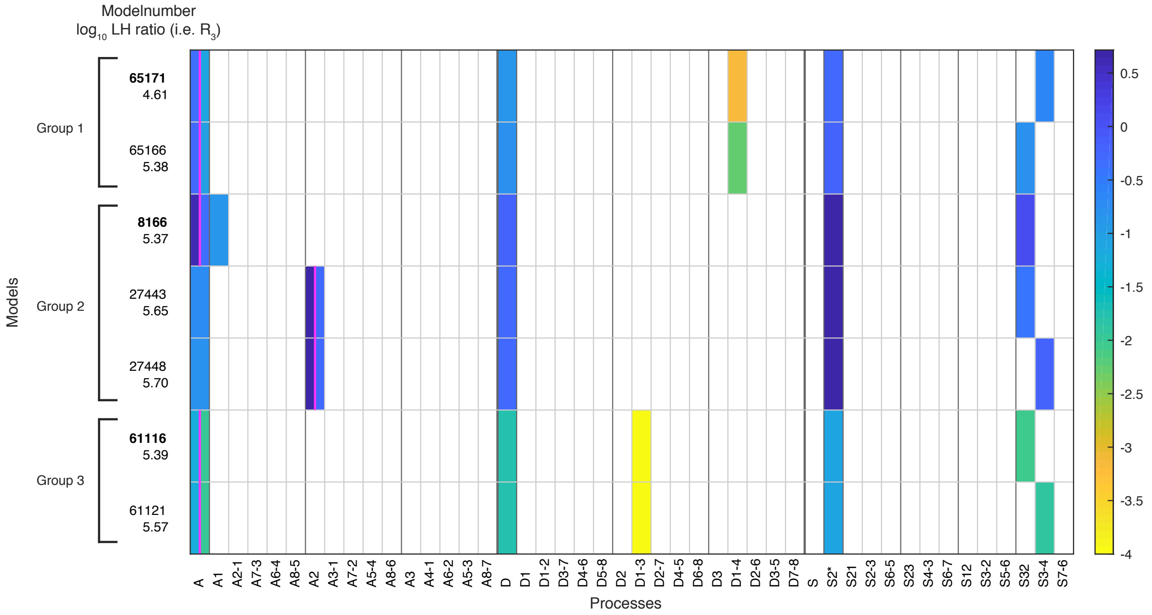


**Figure 4.11:** The 7 models that agree with all data sets after stage 3. The x-axis lists all possible processes and the colored boxes in each row show each model's processes with their rate values. White boxes indicate the absence of a process in a model. Regulated processes are separated into two differently colored boxes for repressed (left half) and activated (right half) promoter state. Weakly activated rate values are not presented here. The exact rate values are given in Table 4.3 and Table 4.4. On the left side next to each row are the model number and the log10 likelihood ratio $R_3$. The models are grouped with respect to similarities in the site-centric net fluxes (Figure C.10). The model numbers of group representatives are printed in bold with net fluxes compared in Figure 4.13.

| model | time scale | name | value | name | value | name | value | name | value |
|---|---|---|---|---|---|---|---|---|---|
| 65171 | $-1.1^{+0.29}_{-0.39}$ | D | $0.23^{+0.023}_{-0.023}$ | D1-4 | $-2.0^{+1.4}_{-\infty}$ | S2* | $0.87^{+0.099}_{-0.096}$ | S3-4 | $0.45^{+0.12}_{-0.12}$ |
| 65166 | $-1.0^{+0.27}_{-0.33}$ | D | $0.23^{+0.025}_{-0.025}$ | D1-4 | $-1.2^{+0.85}_{-\infty}$ | S2* | $0.85^{+0.094}_{-0.094}$ | S32 | $0.24^{+0.10}_{-0.10}$ |
| 8166 | $-0.34^{+0.096}_{-0.090}$ | D | $0.11^{+0.029}_{-0.048}$ | A1 | $-0.56^{+0.37}_{-0.64}$ | S2* | $1.0^{+0.070}_{-0.070}$ | S32 | $0.41^{+0.089}_{-0.089}$ |
| 27443 | $-0.70^{+0.14}_{-0.14}$ | A | 0 | D | $0.44^{+0.13}_{-0.13}$ | S2* | $1.4^{+0.16}_{-0.16}$ | S32 | $0.31^{+0.16}_{-0.16}$ |
| 27448 | $-0.77^{+0.14}_{-0.14}$ | A | 0 | D | $0.50^{+0.13}_{-0.13}$ | S2* | $1.5^{+0.16}_{-0.16}$ | S3-4 | $0.57^{+0.21}_{-0.20}$ |
| 61116 | $-2.0^{+1.0}_{-\infty}$ | D | $0.23^{+0.024}_{-0.024}$ | D1-3 | $-2.0^{+1.3}_{-\infty}$ | S2* | $0.93^{+0.065}_{-0.067}$ | S32 | $-0.042^{+0.16}_{-0.16}$ |
| 61121 | $-2.0^{+0.97}_{-\infty}$ | D | $0.23^{+0.027}_{-0.023}$ | D1-3 | $-2.0^{+1.4}_{-\infty}$ | S2* | $0.93^{+0.069}_{-0.070}$ | S3-4 | $0.10^{+0.19}_{-0.20}$ |

**Table 4.3:** Rate values of constitutive processes (name value pairs with relative value on log10 scale) for each of the satisfactory models after the combined fit to all data sets in stage 3 (as shown in Figure 4.11). The time scale is given in 1/h and also on log10 scale and needs to be added to each individual relative log10 rate value. We also investigated the sensitivity of our models with respect to rate changes in certain directions in parameter space (see Section 4.3.5). ± values for each process correspond to the highest found change from the best process rate value in all tested parameter directions which lead to an a $\approx 50\%$ decrease in likelihood and indicate the sloppiness of a rate value ($-\infty$ indicates that the log10 rate value could be made arbitrarily small).

| model | time scale | name | rep. value | weakly act. value | act. value |
|---|---|---|---|---|---|
| 65171 | $-1.1^{+0.29}_{-0.39}$ | A | $0.80^{+0.051}_{-0.050}$ | $0.41^{+0.045}_{-0.040}$ | 0 |
| 65166 | $-1.0^{+0.27}_{-0.33}$ | A | $0.80^{+0.051}_{-0.049}$ | $0.41^{+0.041}_{-0.040}$ | 0 |
| 8166 | $-0.34^{+0.096}_{-0.090}$ | A | $0.94^{+0.060}_{-0.058}$ | $0.46^{+0.045}_{-0.045}$ | 0 |
| 27443 | $-0.70^{+0.14}_{-0.14}$ | A2 | $1.4^{+0.14}_{-0.14}$ | $0.94^{+0.15}_{-0.15}$ | $0.43^{+0.19}_{-0.19}$ |
| 27448 | $-0.77^{+0.14}_{-0.14}$ | A2 | $1.5^{+0.14}_{-0.14}$ | $1.0^{+0.15}_{-0.15}$ | $0.51^{+0.18}_{-0.18}$ |
| 61116 | $-2.0^{+1.0}_{-\infty}$ | A | $0.79^{+0.048}_{-0.048}$ | $0.41^{+0.041}_{-0.040}$ | 0 |
| 61121 | $-2.0^{+0.97}_{-\infty}$ | A | $0.79^{+0.048}_{-0.048}$ | $0.41^{+0.040}_{-0.040}$ | 0 |

**Table 4.4:** Rate values of regulated processes (values on log10 scale) for each of the satisfactory models after the combined fit to all data sets in stage 3 shown in Figure 4.11. Same as Table 4.3, but for regulated instead of constitutive processes.

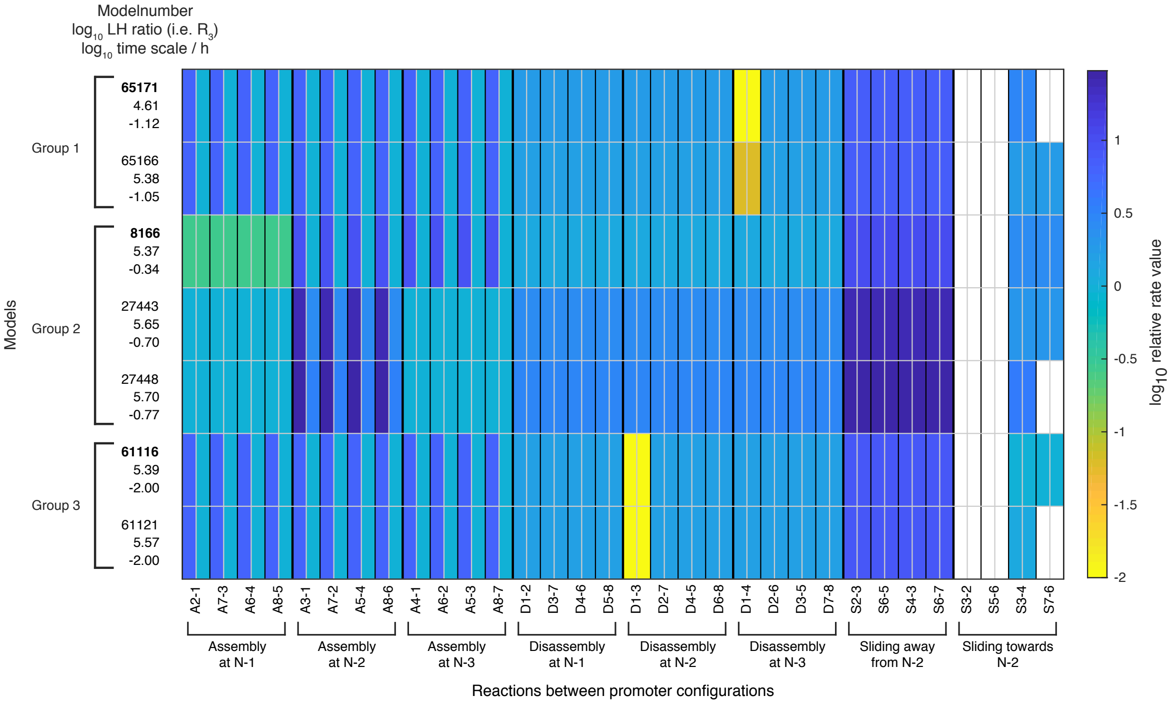


**Figure 4.12:** Relative rate values for each of the 32 reactions (x-axis) when using the processes and their rate values of each model (y-axis) as presented in Figure 4.11. Colored boxes indicate the relative rate values with respect to the global assembly process rate in activated state. White fields correspond to a reaction rate of zero. For a given reaction, the left half represents the repressed value, the right half the activated value, which only differ if the reaction is governed by a regulated process. Weakly activated rate values are not presented here. The left column shows for each model: the model number, the log10 likelihood ratio $R_3$ and the log10 time scale parameter value.

total are not enough to achieve good agreement with the data. Thus, when keeping the total number of fit parameter fixed at 7, the rather simple regulation of only one process is preferred over regulation of several processes.

Within the 7 satisfactory models, the regulated processes are global assembly, A (5x) and assembly at N-2, A2 (2x) (Figure 4.11 and Figure 4.5D). The preference for regulation by assembly rather than disassembly can already be observed after the maximum likelihood fit to the configuration data in stage 1 (Figure 4.5B). The rates of all regulated assembly processes decrease from the repressed over the weakly activated to the activated state. This is in agreement with the promoter-wide nucleosome occupancy decrease during induction.

We automatically investigated whether sliding is needed to fit all data sets, since all sliding processes are optional. Sliding occurs in all 7 models with at least two different sliding processes, consequently only global sliding with the same rate for each sliding reaction is not sufficient. Every satisfactory model employs the sliding process away from N-2 to N-1 and N-3 (S2*) and a process that allows sliding from N-3 to N-2 (S32 or the configuration specific process S7-6, see Figure 4.11 and Figure 4.5D).

On-off-slide models combine processes from different hierarchies: global, site-specific and configuration-specific. Each model has a global assembly and global disassembly process. Global sliding is optional and not used in any of the satisfactory models. Each of these models has one to three site-specific processes out of A1, A2, S2* and S32 (with A1 and A2 overruling the global assembly process). Two models have no configuration-specific processes, models 8166 and 27443 (Figure 4.11). All other satisfactory models have one or two configuration-specific processes. Thus, configuration-specific processes can be beneficial, but are not needed to achieve agreement with the data. The lack of configuration-specific processes does not mean that the three nucleosome sites are independent of each other, since the sliding reactions always introduce coupling between neighboring sites.

### Fluxes

We calculated the directional fluxes (Figure C.6 and Figure C.7) as well as the net fluxes (Figure C.8, Figure C.9 and for three representative models Figure 4.13) between all configurations for each satisfactory model and for different promoter states. Since there are no experimental methods to measure these fluxes yet, modeling approaches provide the only view into the internal promoter configuration dynamics. As mentioned before, the vast majority of all regulated on-off-slide models are non-equilibrium models, i.e. have non-zero net fluxes. In the repressed promoter state, the seven satisfactory models showed the highest fluxes as well as net fluxes occurring between the configurations with three or two nucleosomes (Figure C.6 and Figure C.8). In all seven model the net fluxes are predominantly cyclic from configuration 1 to 2 to 3 and back to 1. Despite qualitative similarities, the fluxes and net fluxes also highlight different behavior, for example only five models showed cyclic net fluxes from configuration 1 to 4 to 3 and back to 1. In the activated promoter state, the higher fluxes and net fluxes between the first four configurations are lost and the differences between the seven models become more pronounced.

Going back to the view point of nucleosome sites (Figure 4.2A), we define effective "site-centric net fluxes" by summing all assembly/disassembly net fluxes at each site and sliding net fluxes between N-1 and N-2 as well as N-2 and N-3 (Figure C.10 and for three representative models Figure 4.13). The site-centric net fluxes provide a simplified picture of the net paths of the nucleosomes on the promoter, ignoring the events on neighboring sites, which are only correctly depicted in the directional or net fluxes between the promoter configurations. In all satisfactory models the N-2 site had a central role with the strongest net nucleosome influx in all promoter states. But we found differences regarding the other site-centric net fluxes and divided the satisfactory models into three groups. 2 models (group 1) exhibit site-centric net influx only at N-2 and N-3 for the repressed state and only at N-2 for the activated state. The remaining models have site-centric net influx only at N-2 for all promoter states, but show very different flux amounts. 3 models (group 2) exhibit a repressed N-2 site-centric net influx of $\approx 0.59/\mathrm{h}$, while the last 2 models (group 3) have $\approx 0.013/\mathrm{h}$. The time scale parameters of group 3 are not properly determined by the given data: Here, the time scales are 10-fold lower than for the other two groups and have larger error bars (Table 4.3), allowing the time scale to speed up by up to a factor of 10 as well as becoming arbitrarily small, while still meeting the fit threshold.

For each group, we picked a representative model with the highest likelihood within the group and present the net fluxes and the site-centric net fluxes in Figure 4.13.

### Maximal reaction rates

After fixing the time scale for each model in the last stage, we could investigate the rate values for each reaction of all satisfactory models (Table 4.3 and Table 4.4). The highest rate of any assembly and disassembly processes were 5/h and 0.6/h, respectively. Sliding rate values had a maximum of 5/h, corresponding to 0.2 bp/s assuming an unidirectional travel of 160 bp with constant speed between two sites. These sliding rates are within the capabilities of yeast ySWI/SNF or RSC remodeling complexes with translocation speeds of up to 13 bp/s [126].

### Bounds for chromatin opening and closing times

*PHO5* induction is the result of consecutive signal transduction, promoter chromatin opening, transcription initiation and downstream processes that finally lead to functional Pho5 acid phosphatase gene product. On the level of *PHO5* mRNA or acid phosphatase activity, induction begins about two hours after phosphate starvation of the cells [127, 128], while after the same time the increase of chromatin accessibility is usually complete, i.e. the kinetics of *PHO5* promoter chromatin opening are faster [129, 128, 130, 131]. Since we modeled here the effective dynamics of nucleosomes at the *PHO5* promoter, we can give approximate upper bounds for the chromatin opening rates for each regulated on-off-slide model.

The effective trajectory of the regulated process rate in time, from the value of the repressed state to the value of the activated state, is dependent on how quickly the cell senses the phosphate starvation and on subsequent signal processes. To calculate a reasonable upper bound for the chromatin opening rate, we assumed the regulation happens instantaneously, i.e. the activated rate value of the regulated processes applies instantaneously at the change of the medium for a population in repressed state. Then the promoter configuration distribution

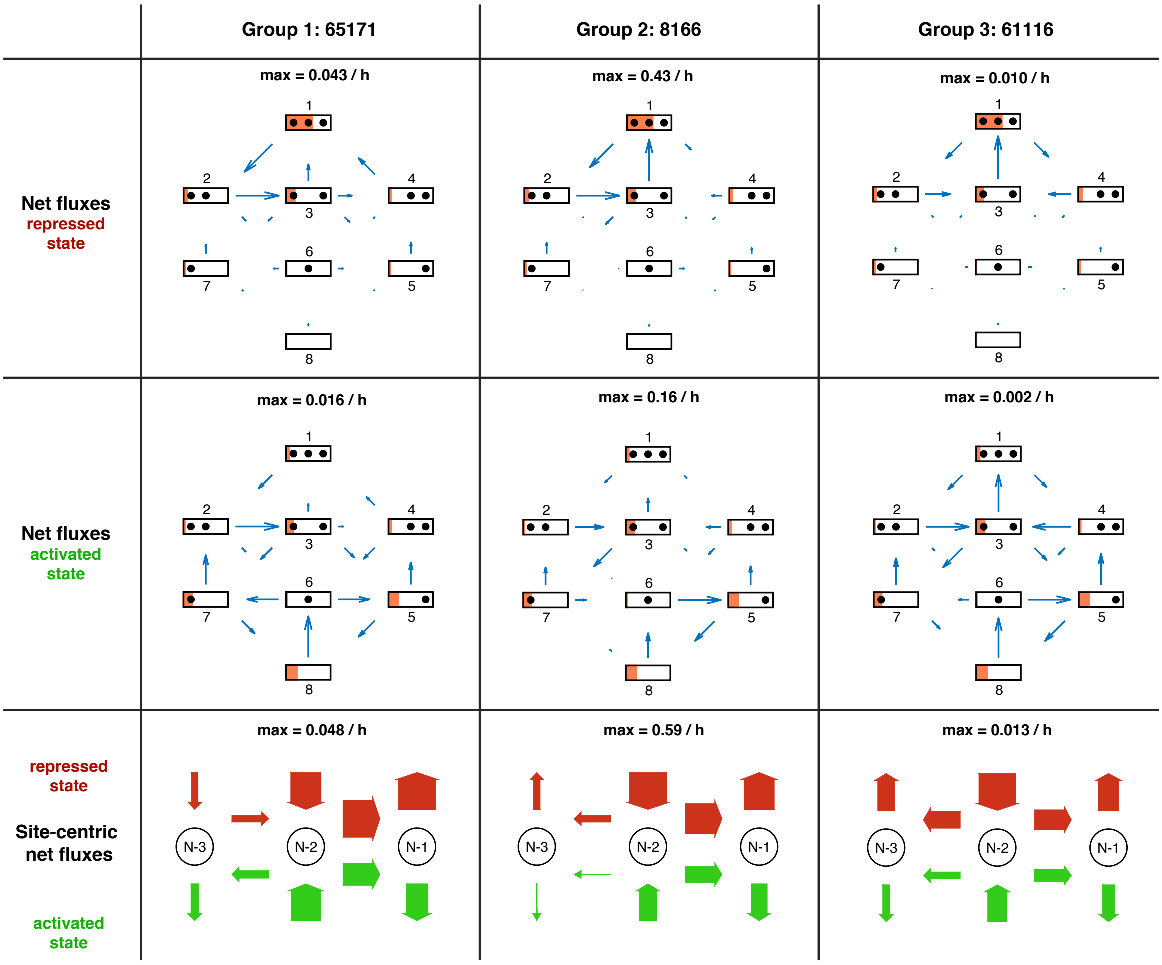


**Figure 4.13:** Overview of net fluxes and site-centric net fluxes for the three group representatives. First two rows: net fluxes in repressed and activated promoter state (arrows) with configuration probabilities as orange horizontal bars (a filled promoter rectangle represents probability 1). Arrow lengths show the relative flux amounts within a flux network, with the maximum stated above each plot. Third row: site-centric net fluxes in repressed (red) and activated (green) state, calculated by summing all assembly/disassembly net fluxes at each site and sliding net fluxes between N-1 and N-2 as well as N-2 and N-3. Here the flux amount is given by the arrow thickness, with the maximum stated above.

decays exponentially towards the activated steady state with a rate that can be approximated by the negative eigenvalue of the transition rate matrix closest (but not equal) to zero, taking into account the fitted time scale. This “effective chromatin opening rate” represents an upper bound of how quickly a given model is able to switch to the activated state. Conversely, we did the same calculations for the “effective chromatin closing rate”, an upper bound of how fast a given model can switch to the repressed state.

Group 1 models had an effective chromatin opening rate close to 0.2/h. Group 2 exhibited a faster effective chromatin opening rate of $\approx 0.8$/h, while group 3 yielded only $\approx 0.02$/h. Since these values are proportional to the fitted time scale, they inherit its error bars. For group 2, the effective chromatin opening rate was high enough to reflect the experimentally measured kinetics. The remaining two groups could also match these kinetics, although just barely, if the maximum time scale uncertainty was considered (approximately an increase by a factor of 2 for group 1 and a factor of 10 for group 3, see Table 4.3).

The ratio of the effective chromatin closing rate over the opening rate was between 2.9 and 3.5 for all satisfactory models. Already after the first fit in stage 1, all investigated assembly-regulated models in agreement with the configurational data had a ratio from 1.5 to 4.5. The few disassembly-regulated models after stage 1 had a ratio smaller than 1, ranging from 0.25 to 0.65. This further supports the assembly-regulated models since promoter chromatin closing and repression of *PHO5* transcription were experimentally shown to be faster than chromatin opening and transcription activation: after the shift from phosphate-free to phosphate-containing medium, repression of *PHO5* transcription was almost complete within 20 min [110] and 65% of chromatin closing was achieved within 45 min [129].

### Extending the likelihood threshold

When we increased the logarithmic likelihood ratio threshold $R_{max}$ from 6 to 7, we obtained 28 models with $R_3 < R_{max}$ (data not shown). The above described properties of satisfactory models, however remained stable. A notable exception was the appearance of the first model with only one sliding process (S2*) and models with both, global sliding and sliding away from N-2 (S2*).

## 4.4 Discussion

We introduced a new method for modeling nucleosome dynamics at promoters: regulated on-off-slide models. They feature nucleosome assembly, disassembly and sliding and enable a simultaneous fit of data for different promoter activation states. The hierarchical approach of global, site-specific and configuration-specific processes enabled us to represent different network topologies, i.e., turning “off” reactions by using a strongly decreased rate of a non-global process, and continuous variations among network topologies. We investigated all 68 145 regulated on-off-slide models with up to 7 fit parameters and realized a completely unbiased modeling approach that provides insight into previously hidden nucleosome dynamics.

We used the *PHO5* promoter as an example, where regulated on-off-slide models provided a unique integration of four different data sets in nucleosome resolution: multi-nucleosome configuration measurements, our own nucleosome accessibility experiments in sticky N-3 mutants and two Flag/Myc-tagged histone H3 exchange experiments.

We obtained and used the same two sticky N-3 mutant strains as published [69], but we decided not to use their published chromatin data, which were generated by an at that time novel single molecule DNA methylase footprinting method and seemed less trustworthy for the following reasons. First, it was unclear if the DNA methylation reactions were saturated possibly leading to too low accessibility values (for a detailed discussion see Section 2.3 as well as [28]). Second, accessibility at the N-3 position in wild type cells was reduced (7%) in the activated compared to the repressed (28%) state, contrary to chromatin opening upon activation and the data by [34] with 55% accessibility at N-3 in the activated promoter state. Third, the authors did not detect any case where all three N-1 to N-3 nucleosomes were removed upon activation, even though this configuration was among the most frequent in the study by Brown et al. [34]. Both the second and third point may be caused by incomplete methylation and/or by erroneous interpretation of methylation footprint patterns as not only nucleosomes but also other factors like the transcription initiation machinery could obstruct methylation.

To investigate these doubts, we used the classical and well-documented restriction enzyme accessibility assay [132, 124] to measure *PHO5* chromatin opening in the sticky N-3 mutants. In this assay, both sticky N-3 mutants still displayed considerable chromatin opening but corroborated the general conclusion by Small et al., that opening of both the N-2 and N-3 nucleosomes was less extensive than for the wild type promoter. Even though the effect was weaker than claimed by Small et al., it still substantiates earlier measurement at the yeast *PHO8* and *PHO84* promoters [133] (which are co-regulated with the *PHO5* promoter) about the role of underlying DNA sequence in stabilizing nucleosomes against remodeling. Recent in vitro experiments started to reveal how chromatin remodeling enzymes are affected by nucleosomal DNA sequences [57, 134, 52]. The sticky N-3 *PHO5* promoter mutants pioneered by Small et al. may provide an impressive example how this is relevant in vivo.

With these four data sets in mind, we designed our models to consider full nucleosomes, not individual histones, during assembly and disassembly and used an effective description over a more detailed method with base-pair resolution and transcription factor dynamics [118]. In this way, steady state data combined with dynamical data yielded new insights into the nucleosome configuration dynamics. Our regulated on-off-slide models were arranged such that at least one process was regulated, i.e. its rate value differed depending on the promoter state. This resulted in the possibility to examine how regulation between different promoter states was most likely achieved and we found that regulation from repressed over weakly activated to activated promoter states can be surprisingly simple. Note that only using the configurational data of [34] was insufficient to restrict our model set. For example, the best and second best models after stage 1 were almost equally likely to reproduce the data. However, they had very different properties, as the first was regulated by global assembly and used a sliding process while the second was regulated by global disassembly and had no sliding at all.

Out of 68 145 tested models only 7 models fulfilled our threshold criterion after fitting all four data sets. All 7 satisfactory models used sliding away from the N-2 position, but towards the N-2 only from the N-3 position. Models without any sliding and equilibrium models within the tested class did not fit all four data sets. Since we made all sliding processes optional, this showed that sliding is crucial and has a net directionality. Configuration-specific processes, i.e. processes representing only a single reaction, were used by all but two of the 7 models, including the model with the highest likelihood. This also means that in the two cases without configuration-specific processes, the coupling between positions introduced by sliding was sufficient to reproduce the experiments.

Incorporating dynamical Flag-/Myc-tagged histone exchange data sets was essential to set the time scale for each model, as this can not be done by using steady state data alone. Since this was a completely new use case for these data sets, we took care not to over-interpret them and performed a thorough sensitivity analysis. Some models had a rather "sloppy" time scale, i.e. the time scale could vary quite strongly without notably decreasing the fit quality. One possible reason is that the ratios at different times of Flag over Myc amount at N-1 needed to be shifted by their mean to take into account the high fit error of the absolute values [12]. The approach of [13] did not have this problem, but unfortunately provided measurements at only one time point, not restraining the slope of the dynamics of the ratio of N-1 and N-2 Flag amounts Figure 4.9.

Our models were fitted to four independent data sets derived from three orthogonal experimental approaches. Importantly, after knowing the time scales, we could calculate approximate lower boundaries on how fast the satisfactory models could switch from a closed chromatin state to an open state and vice versa. Taking into account the time scale errors, the effective chromatin opening rates were compatible with experimental values. Additionally the comparison between effective chromatin opening versus closing rates were only compatible with the regulated assembly rather than disassembly models already after the first stage just using the data from Brown et al. [34]. Thus a fourth type of orthogonal data confirms regulation by assembly.

Most findings, like the identification of essential and directional sliding during *PHO5* promoter chromatin opening, the estimated time scales and the central role for the N-2 nucleosome in all seven models matched expectations based on earlier studies. However, we were surprised that only a global or site-specific (N-2) regulated assembly, but not a regulated disassembly process was compatible with the data which would be commonly expected. In equilibrium systems, the same chromatin opening could result from more disassembly or from less assembly, but as noted above, our selected models are all non-equilibrium models and as such do not shared this symmetry. Models with regulated disassembly could result in satisfactory fits to all data sets after we allowed one additional fit parameter, but these were still a very small minority among all satisfactory models, with best maximum likelihood drastically lower (less than 1/20) than the best assembly regulated models (data not shown). However, increasing the number of processes and parameters makes the modeling approach less distinctive and a large number of different mechanisms could be modeled and fitted successfully. Thus, at first sight, regulation by assembly challenges the common view on promoter chromatin opening mechanisms as derived from pioneering studies at the yeast *PHO5* [33] or other, like the *HO* [135] promoter, but also at mammalian promoters like the glucocorticoid-regulated MMTV promoter [136, 137]. According to this view, the opening of chromatin is triggered by binding of a (transcription) factor that locally recruits a chromatin remodeler, either directly or via histone modifications like acetylation [138]. This remodeler then mediates nucleosome removal either by sliding and/or by disassembly [139, 140, 141, 142, 143]. This view is mainly based on the following points: i) experiments showing physical interactions between transcription factors and remodelers [144, 145, 146] or between remodelers and modified chromatin [138], ii) on chromatin immunoprecipitation or microscopy data showing that transcription factors or histone modifiers are recruited to the promoter upon promoter activation [135, 147, 148, 149, 137], and iii) on in vitro assays that exposed nucleosome sliding and disassembly activities for chromatin remodelers [150, 151, 152, 153]. However, we note that remodelers were equally shown to mediate nucleosome assembly in vitro [151, 154] and we therefore wonder if the well-documented remodeler recruitment at promoters upon promoter activation may also cause downregulation of nucleosome assembly rather than upregulation of nucleosome disassembly.

We propose to reconsider the common view of regulation by disassembly in the case of the *PHO5* promoter given our results and in the light of a long standing and recently revived debate regarding the role of binding competition in nucleosome remodeling. Before ATP dependent remodeling enzymes and their active nucleosome displacement activities were discovered and fully recognized, it was proposed that binding competition between the histone octamer and sequence specific DNA binders like transcription factors was a major mechanism for nucleosome removal. For instance, if a nucleosome was assembled over Gal4 binding sites in the absence of Gal4 it could be displaced by adding Gal4 in vitro [155] or inducing Gal4 in vivo [156]. More recently, the class of pioneer factors and general regulatory factors that have important roles in opening chromatin or keeping chromatin open, were also suggested to displace nucleosomes by binding competition [157, 158, 159]. Regarding the *PHO5* promoter, this mechanism seemed appealing because its transactivator Pho4 indeed competes with nucleosome N-2 during binding to the intranucleosomal UASp2 (Upstream Activating Sequence phosphate regulated 2) site and a *PHO5* promoter lacking UASp2, i.e. left only with the constitutively accessible UASp1 between N-2 and N-3, was not induced during phosphate starvation in vivo [160]. However, early studies ruled out an essential role for binding competition at the *PHO5* promoter as i) it was possible to open the coregulated *PHO8* promoter without an intranucleosomal UASp [66], ii) even the ΔUASp2 *PHO5* promoter mutant could be opened if *PHO4* was overexpressed [160] and iii) even an overexpressed Pho4 version containing a functional DNA-binding but no transactivation domain was not able not displace N-2 and trigger chromatin opening although just the Gal4 DNA binding domain could displace nucleosomes in other contexts [155, 156]. However again, even though the binding competition at UASp2 in N-2 was not essential if the binding to UASp1 was boosted by increasing its affinity or by *PHO4* overexpression, it did have a critical role in *PHO5* promoter chromatin remodeling for wild type UASp1 and *PHO4* expression levels [125]. Therefore, we propose for the wild type version of the *PHO5* promoter, that the competition between Pho4 binding to UASp elements at the promoter and nucleosome assembly at the promoter corresponds to the downregulated assembly process upon promoter activation in our satisfactory regulated on-off-slide models. The downregulated N-2 assembly processes in our models 27443 and 27448 (Figure 4.11) then could directly correspond to inhibiting the nucleosome assembly at N-2 by Pho4 binding at UASp2. Pho4 binding to both UASp2 and UASp1 may correspond to decreased global assembly in the other models. Such a promoter chromatin opening mechanism does not preclude that Pho4 recruits chromatin remodelers which are important for the nucleosome dynamics during the chromatin transition, as posed by the common view, but it shifts the interpretation of previous data away from a focus on a regulated disassembly to a regulated assembly process. Thus, our modeling results are not countered by any existing evidence, but may lead us to reconsider the nucleosome dynamics in vivo, which so far could not be investigated by any approach. In hindsight, although it was possible to open the *PHO5* promoter by *PHO4* overexpression in the absence of UASp2 resulting in the same open promoter state [160], different effective nucleosome dynamics may govern this mutant compared to the wild type promoter transition. If the same kind of single molecule nucleosome configuration data as in [34] and other data used in this study were available for this ΔUASp2 *PHO5* promoter chromatin transition, our modeling method could test this and investigate especially if a regulated assembly process was coupled to binding competition at N-2.

Additionally long read single molecule sequencing with reliable nucleosome calling methods could render the very elaborate and time consuming measurements as in [34] obsolete and provide nucleosome configurations data of several promoters at the same time. The here presented modeling approach could then be adapted and used for further nucleosome dynamics studies in the future.

# A Absolute Occupancy by ORE-seq and ODM-seq - Supplement [1]

## A.1 Experimental materials and methods

Please see our paper [28] for information on yeast strains and media, isolation of yeast nuclei, DNA methylation in chromatin, restriction enzyme digestion of chromatin, calibration samples for restriction enzyme digests, DNA methylation and restriction enzyme digestion for in vitro-reconstituted chromatin, Illumina library construction and sequencing, Oxford Nanopore library construction and sequencing, bioinformatics and data access.

The following strains and restriction enzymes (Figure 2.5) were not presented in [28]. The *nhp6a nhp6b* mutant (Y869, MATa *ura3-52 trp1-289 his3-D1 leu2-3 112 gal2 gal10 nhp6A-D3::URA3 nhp6B-D3::HIS3* [161, 162, 76] and the corresponding wild type (Y865) were provided by Alessandra Agresti and grown as the BY4741 strain in [28]. The histone depletion strain RMY102a [75] (MATa *ade2-101 his3-Δ200 lys2-801 trp1Δ901 ura3-52 hht1 hhf1::LEU2 hht2 hhf2::HIS3 plus pRM102 [CEN4 ARS1 URA3 P(GAL10)-HHT2 P(GAL1)-HHF2]*) was a kind gift of Michael Grunstein. The strain was grown in YPA/2% Galactose to mid log phase, harvested, washed twice in water and then resuspended and incubated in YPDA for 2 to 3 h to induce histone depletion. HindIII-HF and HhaI were incubated in 1x CutSmart Buffer (NEB) and KpnI+BamHI-HF in 1x NEB 1.1 buffer and otherwise treated as the described restriction enzymes [28].

## A.2 Calculation of cut and uncut fragment count

The following steps describe the analysis of fastq Illumina sequencing files to obtain cut and uncut fragment counts for ORE-seq to be used in the cut-uncut and the cut-all cut version. The source code can be downloaded from https://github.com/gerland-group/absolute-occupancy-analysis.

### Bioinformatic preparation steps

1. Cut read ends according to basecalling quality with FastqFilter and quality threshold 10.
2. Map reads to the joined *S. cerevisiae* and *S. pombe* reference genome with BWA [163].

[1] Large parts of this supplemental chapter are adapted from our publication [28] under CC BY-NC 4.0 license.

3. Ignore read pairs with unreasonable bam flags using the rules of readGAlignmentPairs.
4. In the following we need the paired-end read information: chromosome, start, end, as well as strand.
5. Discard fragments longer than 500 bp as well as fragments on the loci of rDNA genes: *S. cer.* chr. 12: 45100 to 495000, *S. pom.* chr. 3: 0 to 30000, *S. pom.* chr. 3: 2430000 to 2452883.

## Fragment count at cut sites

1. Count the starting/ending fragments on plus and minus strand $c_\tau(x)$ for each genomic position $x$ with $\tau = 1, 2, 3, 4$ denoting starts on plus, starts on minus, ends on plus and ends on minus strands, respectively. For starting reads, we count the position of the first base pair, for ending reads we count the position after the last base pair (i.e. end positions are shifted by $+1$ bp). We use the notation $c^1_\tau(x)$ and $c^2_\tau(x)$ for the sample without 2nd RE digest and the sample with 2nd RE digest, respectively. For later modeling we assume that one single given fragment with cut or sheared fragment start or end at $x$ supplied to PCR and Illumina sequencing will on average yield $p^x_\tau$ counts.
2. For the cut-uncut method, we need the uncut fragments for fixed genomic positions $x$, i.e. fragments that start before $x - d$ and end after $x + d$ (where we shifted end positions as before by $+1$ bp) in the sample without 2nd RE digest. The extension by $d$ is needed due to the fact that not all RE cut both strands at the same position, as explained later. We denote this number of uncut fragments with $u^1_\tau(x, d)$, also using the index $\tau$ to differentiate between plus ($\tau = 1$ or 3) and minus strand ($\tau = 2$ or 4). We assume that one such uncut fragment at $x$ supplied to PCR and Illumina sequencing will on average results in $q^x_\tau$ counts.
3. We determine the cut site positions, depending on the recognition motif of the RE, on both genomes, taking into account the actual DNA ends generated by end polishing in the following way. Let $x^i$ be the position of the first base pair of the recognition motif of cut site $i$ plus half the length of the recognition motif (they are always even in length). As as example: HindIII, with '||' indicating the cut positions:

   $x^i$

   ```
   + strand:  5'-...A||A G C T T...-3'
   - strand:  3'-...T T C G A||A...-5'
   ```

   In case the 3' end of a fragment is shorter than the 5' end after cutting, the 3' end is elongated to match the 5' end by polymerase. In case the 3' end is longer than the 5' end, the 3' end is digested to match the 5' end. For an exemplary HindIII cut site, we obtain the following double stranded fragments:

   $x^i$ $x^i$

   ```
   + strand:  ending:  5'-...A A G C T-3' and starting:  5'-A G C T T...-3'
   - strand:  ending:  3'-...T T C G A-5' and starting:  3'-T C G A A...-5'
   ```

   Let $\Delta s$ be the shift length from the pattern center to the cut position of the + strand in upstream direction, which corresponds to the half the length of the 5' overhang of

| RE | recognition site | shift length |
|---|---|---|
| AluI | AGCT | 0 |
| BamHI | GGATCC | 2 |
| HindIII | AAGCTT | 2 |
| HhaI | GCGC | -1 |
| KpnI | GGTACC | -2 |

**Table A.1:** Restriction enzyme recognition sites with shift length $\Delta s$. BamHI-HF and HindIII-HF have the same recognition sites and length shifts as the non-HF versions.

the cleavage product in bp. For HindIII, $\Delta s = +2$, and other used REs in Table A.1. $\Delta s = 0$ for blunt end cutting RE whereas in case of an RE with 3' overhangs, $\Delta s$ is negative. Assuming proper end polishing of cut fragments as described above, we have the following counts for site $i$:

| Read type | cut sample | all cut sample |
|---|---|---|
| Starting read on + strand | $c^1_1(x^i - \Delta s)$ | $c^2_1(x^i - \Delta s)$ |
| Starting read on − strand | $c^1_2(x^i - \Delta s)$ | $c^2_2(x^i - \Delta s)$ |
| Ending read on + strand | $c^1_3(x^i + \Delta s)$ | $c^2_3(x^i + \Delta s)$ |
| Ending read on − strand | $c^1_4(x^i + \Delta s)$ | $c^2_4(x^i + \Delta s)$ |
| Uncut read on + strand | $u^1_1(x^i, \Delta s)$ | |
| Uncut read on − strand | $u^1_2(x^i, \Delta s)$ | |

To obtain the number of fragments not cut by the RE at a given site, we count all fragments that start before $x^i - \Delta s$ and end after $x^i + \Delta s$, yielding $u^1_\tau(x^i, \Delta s)$. For easier notation we set $x^i_1 = x^i_2 = x^i - \Delta s$ and $x^i_3 = x^i_4 = x^i + \Delta s$, yielding the cuts at site $i$ as $c^1_\tau(x^i_\tau)$, $c^2_\tau(x^i_\tau)$, $\tau = 1, 2, 3, 4$.

4. Fragments with length below 100 bp are very unlikely to be amplified and then sequenced, thus sites with close neighbors may be biased. Furthermore uncut fragments counts are increased at cut sites with any neighbor within approx. 150 bp (Figure A.1, lowest two rows). In the following, we ignore cut sites with one neighbor less than 200 bp away or both neighbors less than 300 bp away. Denote the set of left over sites with $I$ and $J$, for the *S. cerevisiae* and the *S. pombe* genome, respectively.

5. We often saw dependencies between the fragment counts $C^i_\tau$ and $A^i_\tau$ (defined below) and the distance to the next neighboring site, ranging up to 250 bp, e.g. for starting reads and the downstream distance to the next neighbor (Figure A.1, row 2 and 3). Thus we ignore start/end cut counts of a cut site and near the cut site, when the next cut site downstream/upstream is closer than 300 bp, respectively.

## Treatment of endogenous exonuclease activity

Due to endogenous exonucleases that may be present in the chromatin preparations and trim DNA ends after restriction enzyme cleavage, some fragments ends do not match the cut site positions any more, even though they were generated by restriction enzyme. Thus we need to count the starting and ending fragments not only at the exact cut positions, but also at some distance from it. The amount of strand resection varies between samples, so its correction needs to be tailored to each pair of samples without and with 2nd RE digest.

We define count windows for each fragment type: For read starts, $W_1 = W_2 = \{0, 1, 2, ..., w\}$ to apply a window in downstream direction and for read ends, $W_3 = W_4 = \{-w, -w+1, ..., 0\}$ to apply a window in upstream direction. The algorithm to find the optimal value for $w$ is described at the end of this step. $C^i_\tau$ denotes the number of cut fragments in the sample without 2nd RE digest ("cut sample") and $A^i_\tau$ denotes the number of cut fragments in the sample with 2nd RE digest ("all cut sample"):

$$C^i_\tau \coloneqq \sum_{a \in W_\tau} c^1_\tau(x^i_\tau + a) \quad \text{and} \quad A^i_\tau \coloneqq \sum_{a \in W_\tau} c^2_\tau(x^i_\tau + a) \tag{A.1}$$

$w$ is calculated using the sample without 2nd RE digest and the sites on the *S. cerevisiae* genome and then the same value is applied to the sample with 2nd RE digest and the *S. pombe* genome. For completely uncut samples (test samples without RE), we set $w = 5$ to average over fluctuations in the very low cut counts at a single position. In the case of ignored start counts of the previous step, we set $C^i_\tau = \text{NA}$ and $A^i_\tau = \text{NA}$ for $\tau = 1, 2$ and the same for $\tau = 3, 4$ in the case of ignored end counts. Figure A.2 shows the histograms of different cut and uncut counts cut and all cut samples with good sequencing coverage.

For normal samples, we use the following algorithm, which makes sure that increasing $w$ by 1,2,3,4 or 5 bp does not increase the summed counts until $w$ by more than 1%, after correcting for cut counts from shearing.

Calculate the mean counts (averaged over all cut sites) at each position $-200$ bp to 200 bp away from the cut sites for fragment starts and ends and both strands. These cut counts near the average cut site usually show a single peak at 0, but depending on the conditions there also is a decreasing shoulder downstream/upstream for starts/ends (Figure 2.2A, WT3 samples with high units). Averaging the different types and strands (end counts need to be mirrored at 0 first) yields $m(d)$, $d$ being the distance to the average cut site. The cut counts need to be corrected by the average shearing cut counts, which we obtain 100 to 200 bp away from the cut site: $m^c(d) = m(d) - \langle m(d) \rangle_{d=100,...,200}$ ($\langle ... \rangle$ indicating the average). We define the cumulative sum of counts by $S(d) = \sum_{l=0}^{d} m^c(d)$. Finally we set $w$ equal to the first integer starting from 0 such that for all $n \in \{1, 2, 3, 4, 5\}$, the sum of the counts of the next $n$ positions, $S(w + n) - S(w)$, is lesser than 1% of $S(w)$. In our samples typical values for $w$ ranged from 0 to 20, going up to 40 for samples with very strong resection.

Uncut fragment counts at any RE cut site are not influenced by endogenous exonucleases as they are still occupied by a nucleosome or other protein that blocked the RE. For easier notation we define the uncut counts at site $i$ by

$$U^i_\tau \coloneqq u^1_\tau(x^i, \Delta s). \tag{A.2}$$

The mean resection length is defined as $\sum_{d=0}^{w} m^c(d)$ (Figure 2.2B and Figure 2.5E). The values of $w$ and the mean resection length of different samples are illustrated in Figure A.3.

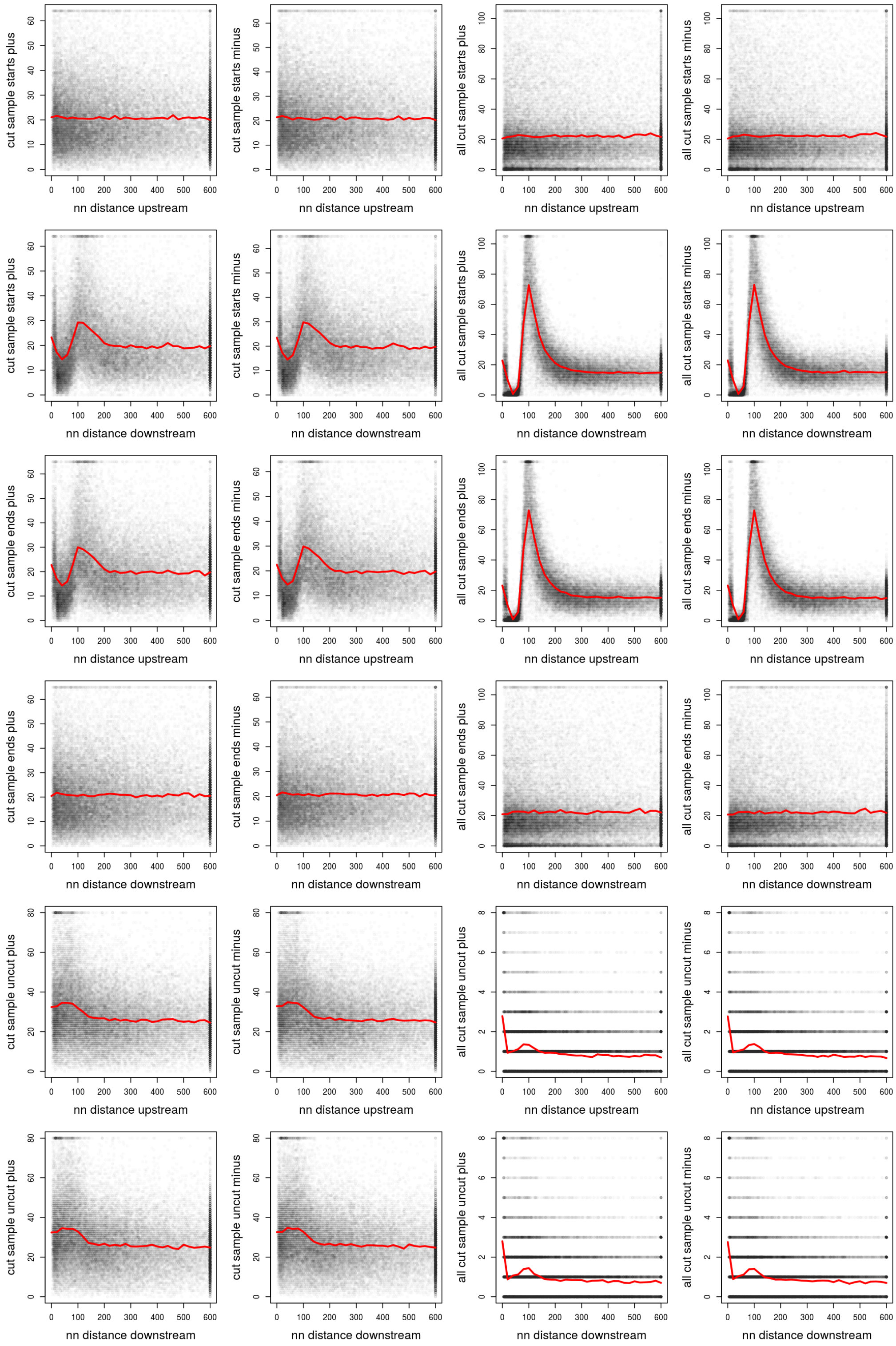


**Figure A.1:** Different fragment counts at the RE sites on the *S. cerevisiae* genome of the WT4 AluI high units cut sample (first and second column) and the WT4 AluI high units all cut sample (third and forth column) plotted against the distance to the next neighboring site upstream (odd rows) or downstream (even rows).

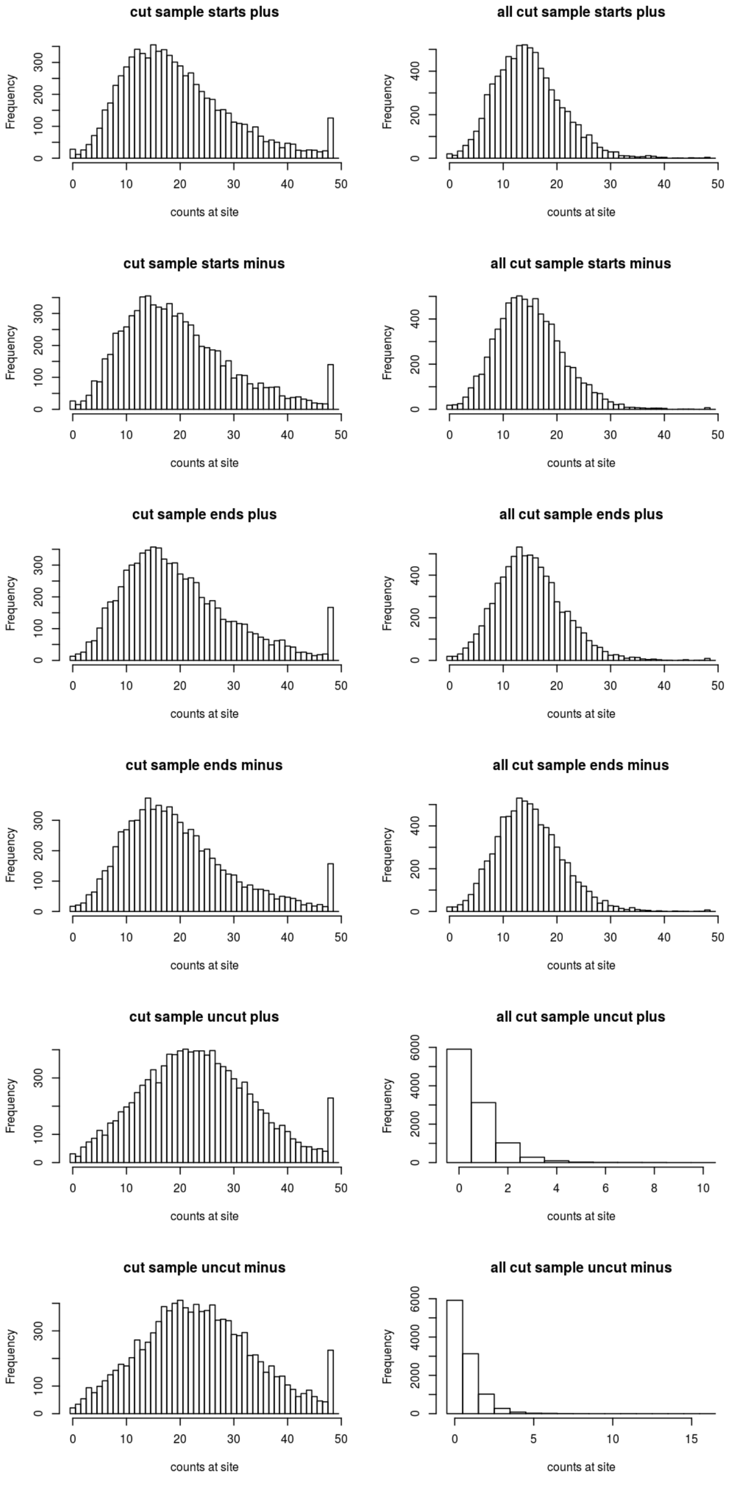


**Figure A.2:** Histograms of fragment counts at each site on the *S. cerevisiae* genome, from the top: fragment starts, fragment ends and uncut fragments on plus and minus strand, respectively, for the WT4 AluI high units cut sample (left column) and the WT4 AluI high units all cut sample (right column). Close sites are already filtered out, but counts are not yet normalized with respect to the *S. pombe* counts.

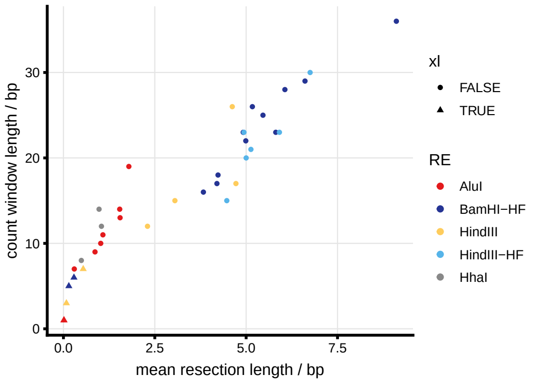


**Figure A.3:** Count window length vs. mean resection length for samples with different REs with and without crosslinking.

## A.3 Mathematical model for occupancy estimation by cut-all cut method

We seek to estimate the real accessibility $\alpha_i$ at cut site $i$ using the cut counts of the cut and all cut samples taking into account a bias due to sheared fragment ends and effective sequencing probabilities. We begin with viewing $C^i_\tau$ and $A^i_\tau$ as random variables with the expectation values

$$E[C^i_\tau] = N_C \bar{p}^i_\tau \mu^i \quad \text{with} \quad \mu^i = \alpha^i + (1-\alpha^i)s \tag{A.3}$$

$$E[A^i_\tau] = N_A \bar{p}^i_\tau \tag{A.4}$$

where $N_C$ and $N_A$ are the number of cell cores in the samples without and with 2nd RE digest, respectively, and $\bar{p}^i_\tau$ is a factor that combines the sequencing probabilities and the PCR multiplication of fragments of type $\tau$ in the window $W_\tau$ at cut site $i$ and is an effective average of the $p^x_\tau$ with $x \in x^i_\tau + W_\tau$ described earlier. The probability that a given (longer) fragment will be cut by shearing within a fixed region of length w+1 within the fragment is denoted by $s$.

Since the restriction enzymes act before the shearing step, only the fraction that has not been cut by the enzyme can be cut in the shearing step in the chromatin sample, leading to $\mu^i = \alpha^i + (1-\alpha^i)s$. For the all cut sample, we assume that all counts near a cut site came from a cut of the restriction enzyme and all counts far away from cut sites occurred due to shearing.

We define these four estimators for $\mu^i$ and $\alpha^i$:

$$\hat{\mu}^i_\tau := \frac{C^i_\tau}{A^i_\tau}\frac{N_A}{N_C} \tag{A.5}$$

$$\hat{\alpha}^i_\tau := \frac{\hat{\mu}^i_\tau - s}{1-s} \tag{A.6}$$

The estimators for $\alpha^i$ are approximately unbiased, as

$$E[\hat{\alpha}^i_\tau] = \frac{\hat{\mu}^i_\tau - s}{1-s} = \frac{1}{1-s}\left(E[C^i_\tau]E\left[\frac{1}{A^i_\tau}\right]\frac{N_A}{N_C} - s\right) \approx \alpha_i,$$

since the $C^i_\tau$ and $A^i_\tau$ are statistically independent as they originate from different samples and $E\left[\frac{1}{A^i_\tau}\right] \approx \frac{1}{E[A^i_\tau]}$. The two sets $\{C^i_\tau\}$ and $\{A^i_\tau\}$ within themselves, however, are statistically dependent.

If $A^i = 0$ or $A^i = \mathrm{NA}$ (because of a close neighbor in direction of $\tau$) we set $\hat{\alpha}^i_\tau = \mathrm{NA}$. We use the *S. pombe* spike-in sites, which are completely cut in both samples, to estimate $N_A/N_C$:

$$\frac{N_A}{N_C} = \frac{\langle A^i_\tau \rangle_{i\in J,\tau}}{\langle C^i_\tau \rangle_{i\in J,\tau}} \tag{A.7}$$

with $\langle X^i_\tau \rangle_{i\in J,\tau}$ denoting the average of $X^i_\tau$ over $i \in J$ and $\tau$, discarding NA values. The ratio of the number of sequenced *S. pombe* reads could also be used, but gave slightly worse results in our calibration runs.

To estimate the probability $s$, we consider the set $Z$ of all genomic positions in *S. cerevisiae* that are further away than 300 bp from any cut site (including those with a close neighbor). At these positions all counted fragment starts and ends originate from shearing. Consequently, $\langle c^1_\tau(x) \rangle_{x\in Z,\tau}$ is an estimator for $N_C s_1 \langle p^x_\tau \rangle_{x\in Z,\tau}$, with $s_1$ being the probability of shearing a long fragment at one fixed position. As $\bar{p}^i_\tau$ are effective averages of $p^x_\tau$ in the count window of site $x^i$, their averages over two different large regions of the genome are with good approximation the same:

$$\langle p^x_\tau \rangle_{x\in Z,\tau} \approx \langle \bar{p}^i_\tau \rangle_{i\in I,\tau} = \frac{1}{N_A} \langle E[A^i_\tau] \rangle_{i\in I,\tau} \approx \frac{1}{N_A} \langle A^i_\tau \rangle_{i\in I,\tau}$$

where we approximated the average of $E[A^i_\tau]$ over $i$ and $\tau$ by the average of $A^i_\tau$, yielding

$$s_1 = \frac{N_A}{N_C} \frac{\langle c^1_\tau(x) \rangle_{x\in Z,\tau}}{\langle A^i_\tau \rangle_{i\in I,\tau}}. \tag{A.8}$$

Thus, $s_1$ is given by the normalized ratio of the average fragment number at genomic positions where cuts can happen only by shearing (counts in the chromatin sample away from cut sites) and the average fragment number at genomic positions where cuts have to happen by the restriction enzyme (counts in the all cut sample at the cut sites). We can now approximate the probability that a fragment is sheared at least once within a fixed window of length $w+1$: $s = (w+1)s_1$. If a fragment is sheared more than once within a window of length $w+1$, the new fragments within the window will be too small and filtered out before the PCR and sequencing steps.

The stochasticity in the values for $C^i_\tau$ and $A^i_\tau$ for fixed $i$ and $\tau$ can cause the estimators $\hat{\alpha}^i_\tau$ to be smaller than 0 or larger than 1, even though the values they estimate, i.e. $\alpha^i$, are between 0 and 1. As very large outliers influence the mean very strongly, we cap the values for $\hat{\alpha}^i_\tau$ at 1.5 when averaging over $\tau$,

$$\hat{\alpha}^i = \left\langle \min\left(\hat{\alpha}^i_\tau, 1.5\right) \right\rangle_\tau ,$$

yielding one accessibility estimate for each cut site $i$. If $\hat{\alpha}^i_\tau = \mathrm{NA}$, it is ignored during the averaging step. We calculate the global accessibility by averaging over all sites:

$$\hat{\alpha} = \langle \hat{\alpha}^i \rangle_{i\in I}$$

It is useful to further restrict the accessibility values of individual sites, $\alpha^i$, to $[0;1]$, since this gives the best estimate when comparing the accessibility values of individual sites with the measured values from other assays, for example ODM-seq.

## A.4 Mathematical model for occupancy estimation by cut-uncut method

In the cut-uncut method, we only use data from the cut sample, i.e. the sample without 2nd RE digest to estimate the accessibility. We assume all fragments cut near cut sites have the same sequencing probability $p$ and all uncut fragments have sequencing probability $q$. We sum over the different fragment types for cut and uncut counts,

$$C^i := C_1^i + C_2^i + C_3^i + C_4^i \tag{A.9}$$

$$U^i := 2(U_1^i + U_2^i) \tag{A.10}$$

for sites without any neighbor within 300 bp and

$$C^i := C_1^i + C_2^i \quad \text{or} \quad C^i := C_3^i + C_4^i \tag{A.11}$$

$$U^i := U_1^i + U_2^i \tag{A.12}$$

for sites with one upstream/downstream neighbor within 300bp, respectively. Then define the ratio of cut and uncut fragments,

$$\hat{\kappa}^i := \frac{C^i}{U^i}. \tag{A.13}$$

If the denominator is zero, we set $\hat{\kappa}^i = \infty$, which will lead to an accessibility of 1 in the following steps. Similar to the previous section we have $E[C^i] = 4N_C p(\alpha^i + (1-\alpha^i)s_1(w+1))$ with $s_1$ being the shearing probability per base pair, but now calculated only using the cut sample, i. e. the ratio of all cut counts away from sites and the sum of cut and uncut fragment counts away from cut sites. For $U^i$ we assume that the uncut fragment counts are given by fragments that have not been cut by the restriction enzyme at $x_\tau^i$ and after that also not been cut by shearing at $x_\tau^i$. The generally very low sequencing probabilities justify the assumption that $C^i$ and $U^i$ are "independent enough" to make the following approximation:

$$E[\hat{\kappa}^i] \approx \frac{E[C^i]}{E[U^i]} = \frac{4N_C p}{4N_C q} \frac{(\alpha^i + (1-\alpha^i)s_1(w+1)}{(1-(\alpha^i + (1-\alpha^i)s_1))} \tag{A.14}$$

The "uncut correction factor" is defined as the ratio of sequencing probabilities of cut and uncut fragments, $\gamma = \frac{p}{q}$, and fitted to the calibration samples as described in the section below. We now obtain the following estimator for $\alpha^i$:

$$\hat{\alpha}^i := 1 - \frac{1+\sigma}{\frac{\hat{\kappa}^i}{\gamma} + 1 - \sigma w} \tag{A.15}$$

$$= \frac{C^i - \sigma(w+1)U^i\gamma}{C^i - \sigma(w+1)U^i\gamma + (1+\sigma)U^i\gamma} \tag{A.16}$$

$$= \frac{C_{eff}^i}{C_{eff}^i + U_{eff}^i} \tag{A.17}$$

with $\sigma := \frac{s_1}{1-s_1} = \frac{1}{\gamma} \frac{\langle c_\tau^1(z) \rangle_{z \in Z, \tau}}{\langle u_\tau^1(z) \rangle_{z \in Z, \tau}}$ being the corrected ratio of all cut counts away from all sites and all uncut fragment counts away from all sites.

$$C_{eff}^i = C^i - \sigma(w+1)U^i\gamma \tag{A.18}$$

$$U_{eff}^i = (1+\sigma)U^i\gamma \tag{A.19}$$

| Enzyme | AluI | BamHI | HindIII | combined |
|---|---|---|---|---|
| $\gamma_{min}$ | 1.555 | 1.699 | 1.680 | 1.642 |

**Table A.2:** ORE-seq uncut correction factor values used in the cut-uncut method.

| map name | mean abs. occ. | mean st. dev. | replicates | sites |
|---|---|---|---|---|
| ORE-seq + ODM-seq | 77.6 | 6.0 | 17 | 1348985 |
| ORE-seq (AluI + BamHI-HF + HindIII) | 71.5 | 5.4 | 12 | 13807 |
| ODM-seq (CpG + GpC) | 77.7 | 6.0 | 5 | 1345945 |
| AluI | 70.3 | 5.3 | 3 | 9603 |
| BamHI-HF | 77.2 | 5.8 | 7 | 1547 |
| HindIII | 72.6 | 4.9 | 2 | 3757 |
| CpG | 73.9 | 6.6 | 3 | 661758 |
| GpC | 80.7 | 5.4 | 2 | 807851 |

**Table A.3:** Properties of maps using different enzymes. Mean absolute occupancies (in %) and mean (over sites) of the occupancy standard deviations over replicates (in %) for all used enzymes (ignoring REs without individual calibration), either alone or combined to one map as indicated. Column 4 contains the number of replicates, column 5 the number of methylation or restriction enzyme sites.

are the effective counts of cut and uncut fragments, respectively, both corrected for cuts in the shearing step and different sequencing probabilities of cut and uncut fragments. $C^i_{eff} + U^i_{eff}$ then defines an "effective coverage" of cut and uncut fragments at the site $i$ and we decided to ignore sites with an effective coverage below 40. As in the section before, the genome-wide average accessibility is $\hat{\alpha} = \langle\hat{\alpha}^i\rangle_{i\in I}$

We fitted the uncut correction factor $\gamma$ using prepared calibration samples for the restriction enzymes AluI, BamHI and HindIII (Figure 2.3A). For each RE and each calibration sample $s$ with 0%, 10%, 30%, 50%, 70%, 90% and 100% prepared fraction of uncut DNA molecules, i.e. prepared occupancy $\omega_s = 1 - \alpha_s$, we calculated the estimated genome-wide average occupancy $\hat{\omega}_s(\gamma) = 1 - \hat{\alpha}_s(\gamma)$, varying $\gamma$ with the aim to minimize $\langle(\omega_s - \hat{\omega}_s(\gamma))^2\rangle_s$ (Table A.2). We also did a combined fit, averaging the error over the three restriction enzymes to use the resulting value of $\gamma$ for enzymes without specific calibration samples. The dependency of the relative fit error $\left(\langle(\omega_s - \hat{\omega}_s(\gamma))^2\rangle_s / \langle(\omega_s - \hat{\omega}_s(1))^2\rangle_s\right)^{1/2}$ from $\gamma$ for AluI, BamHI and HindIII is shown in Figure 2.3B.

## A.5 ODM-seq analysis

### BS-seq and EM-seq

The analysis of BS-seq and EM-seq data is identical, since they differ only in the conversion method. Paired-end reads were mapped with BS-Seeker2 (version 2.1.8, [78]). Different motifs/patterns were analyzed: GCH, HCG, GCG and HCG, where "H" stands for any base except guanine. The average conversion ratio of a pattern along the reads is given by the number of converted cytosines in the pattern at a fixed position on the reads divided by the number of all reads with this pattern at this position. The average conversion ratio usually

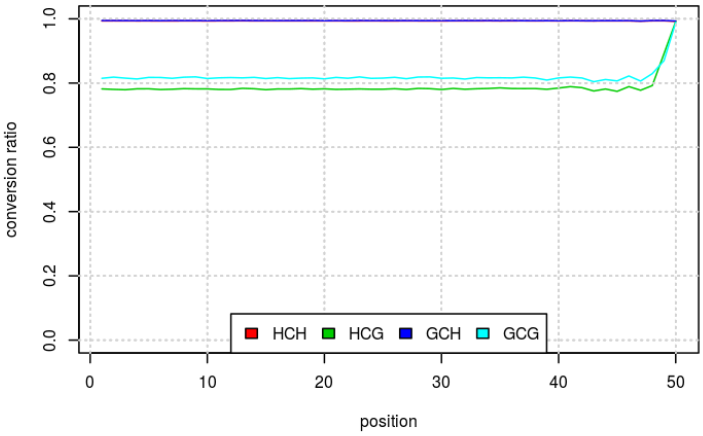


**Figure A.4:** ODM-seq bisulfite conversion ratio along reads which were mapped to the plus strand. Read position 50 marks the real 3' fragment end with an increase in conversion ratio due to end repair.

shows an increase or decrease at the read ends due to end repair (Figure A.4). We thus trimmed the fragments at the ends by 5 bp to 10 bp to achieve a constant conversion ratio along reads. Reads on the loci of rDNA genes were ignored as well as reads with an unconverted HCH motif, which is a indicator for incomplete conversion. For CpG/GpC DNA methyltransferases the GCH/HCG pattern, respectively, should also not be converted, and we named it “anti pattern”. If the average anti pattern conversion ratio among all left-over reads was less than 0.98, the sample was discarded. Finally, the ratio of the converted reads over the number of analyzed reads at each genomic CpG/GpC methylation site estimates the real absolute occupancy. Methylation sites with a coverage less than 20 were ignored as then the error of the occupancy estimation becomes too large. Source code can be downloaded from https://github.com/gerland-group/absolute-occupancy-analysis.

The conversion of the different patterns can be nicely observed for the 601 25mer plasmid (Figure A.5A), where for a CpG sample, the HCH and GCH patterns show conversion ratios close to one, whereas the HCG and GCG pattern conversion decreases towards 10% in the linker region. Each CpG site on the plus strand has a corresponding CpG site on the minus strand, with the position of the cytosine shifted by 1 bp. Comparing the data for plus and minus strand, we find very good agreement (Figure A.5B and Figure A.6). Note that the bisulfite mapper can distinguish between reads coming from original plus strands and reads from original minus strand fragments, even though there are several PCR amplification steps. This is because a conversion of a C on the plus strand will yield a T after the PCR (C converted to U and paired with A in PCR) on the forward read (read that maps to the original strand, here the plus strand) and an A on the reverse read (read that maps to the other strand) but the minus strand just had a G at the same position, so after PCR there will still be a G on the forward read and C on the reverse read [77]. Since HCH motifs are never methylated and quite common, there usually is at least one converted C on each read.

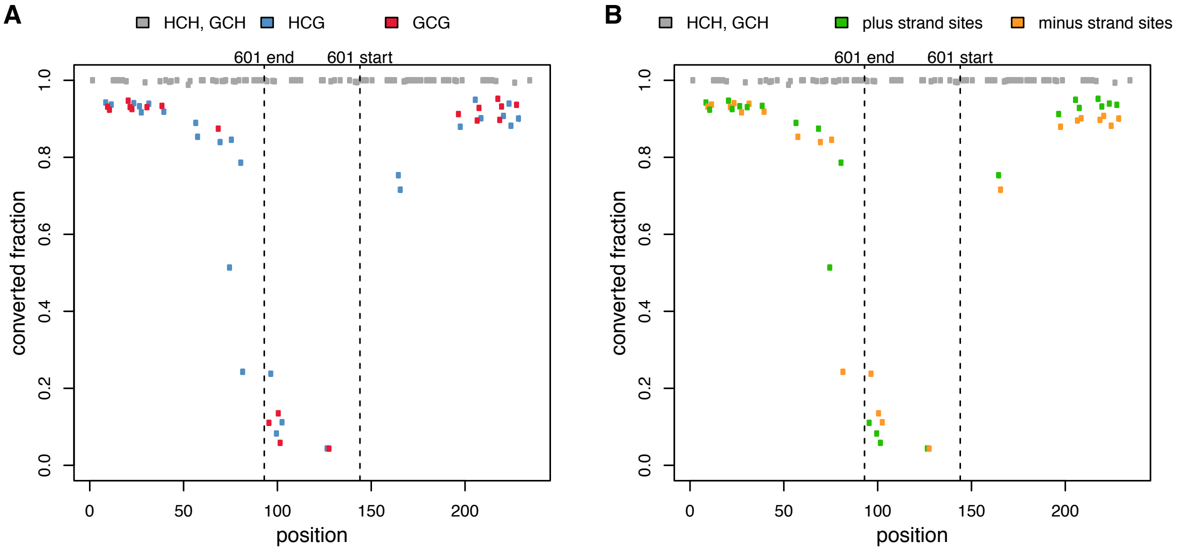


**Figure A.5:** Converted fraction of cytosines (C) on the 601 25mer spike-in exemplary for the WT5 xl CpG 180min ODM-seq bisulfite sample. Fragments were mapped to a generic 601-linker-601 reference sequence, since the exact position of short reads on the highly repetitive 25mer sequence could not be determined. (**A**) Cs colored by pattern: HCH, GCH, HCG and GCG, with G = guanine and H = any base except guanine. The converted fractions at HCG and GCG sites define the absolute occupancy at these sites for a CpG ODM-seq bisulfite sample. (**B**) As in A, but Cs colored by strand, if they belonged to HCG or GCG patterns.

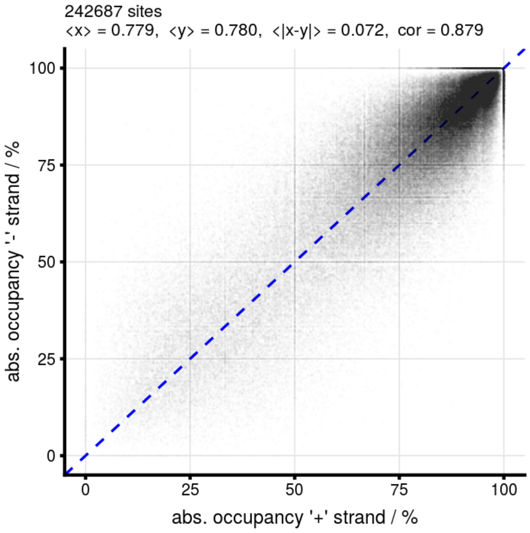


**Figure A.6:** Absolute occupancy on plus vs. minus strand exemplary for the WT5 xl CpG 180min ODM-seq bisulfite sample.

### Nanopore-seq

The first step was base calling with Albacore (Oxford Nanopore Technologies, version 2.3.3, for WT4 samples) or the successor software Guppy (v3.3.3, for WT5 samples) followed by mapping the reads with minimap2 (version 2.14-r892-dirty, [164]). For methylation calling Nanopolish (version 0.11.0, [80]) was used. At each CpG methylation site, the occupancy is estimated by the ratio of unmethylated reads over the sum of methylated and unmethylated reads, thus ignoring reads where the site has been called "ambiguous" by Nanopolish. Again, methylation sites with a coverage less than 20 were ignored. Currently, Nanopolish is not able to resolve each CpG site: neighboring sites closer than 11 bp are grouped together into one site.

# B Chromatin Remodeling Simulations - Supplement

## B.1 One-dimensional grand canonical systems

### B.1.1 Basics

#### System definition

A grand canonical system allows exchange of energy and particles with a reservoir at temperature $T$ and chemical potential $\mu$. We look at 1-dimensional non-uniform systems, i.e. with external potential $u(x)$ and we distinguish arbitrary pair interaction and next-neighbor interaction. In both cases, the interaction strength is given by $\phi(x, y)$. It is possible to think of $\phi(x, y) = \phi(y-x)$, but this will not simplify calculations at this stage, since we need to keep track of positions anyway, because of the external potential. We assume that $\phi(x, y) = \infty$ for $x \geq y$ such that particle can not pass each other and thus have a fixed order. The partition sums are

$$\Xi_{ap} = \sum_{N=0}^{\infty} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} dx_1 ... dx_N \exp\left( -\beta \sum_{i=1}^{N} (u(x_i) - \mu) - \beta \sum_{i=1}^{N} \sum_{j=i+1}^{N} \phi(x_i, x_j) \right) \tag{B.1}$$

$$\Xi_{nn} = \sum_{N=0}^{\infty} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} dx_1 ... dx_N \exp\left( -\beta \sum_{i=1}^{N} (u(x_i) - \mu) - \beta \sum_{i=1}^{N-1} \phi(x_i, x_{i+1}) \right). \tag{B.2}$$

In each case we calculate macroscopic observables by taking the average $\langle ... \rangle$ with respect to one of the grand canonical phase space densities

$$\rho_{ap}(N, x_1, ..., x_N) = \frac{1}{\Xi_{ap}} \exp\left( -\beta \sum_{i=1}^{N} (u(x_i) - \mu) - \beta \sum_{i=1}^{N} \sum_{j=i+1}^{N} \phi(x_i, x_j) \right) \tag{B.3}$$

$$\rho_{nn}(N, x_1, ..., x_N) = \frac{1}{\Xi_{nn}} \exp\left( -\beta \sum_{i=1}^{N} (u(x_i) - \mu) - \beta \sum_{i=1}^{N-1} \phi(x_i, x_{i+1}) \right). \tag{B.4}$$

For example, the mean particle number is denoted by $\langle N \rangle$ and the mean particle density by

$$n(x) = \Big\langle \sum_{i=1}^{N} \delta(x - x_i) \Big\rangle. \tag{B.5}$$

We then also have $\langle N \rangle = \int n(x)dx$. Of course, the phase space density also defines the probability P(...) for events and states.

It is important to understand that for general $\phi$ the difference between systems with arbitrary pair interaction and systems with next neighbor interaction lies in the definition of the Hamiltonian used in $\Xi$ and $\rho$, because the terms in $\Xi$ determine which particles interact with each other and $\phi$ determines "how strong".

### Densities and probabilities

We now define the mean pair density $n_2(x, y)$ and the mean next neighbour density $\overline{n}_2(x, y)$ by

$$n_2(x,y) = \big\langle \sum_{i=1}^{N} \sum_{j=1}^{N} \delta(x - x_i)\delta(y - x_j) \big\rangle \tag{B.6}$$

$$\overline{n}_2(x,y) = \big\langle \sum_{i=1}^{N-1} \delta(x - x_i)\delta(y - x_{i+1}) \big\rangle. \tag{B.7}$$

These definitions are the same as in the canonical case with fixed particle number, where the average is taken with respect to the canonical phase space density. However, in the grand canonical case not only the different $x_i$, but also $N$ is a random variable. From the canonical case we are used to the relations $\int \int n_2(x, y)dxdy = N(N - 1)$ and $\int \int \overline{n}_2(x, y)dxdy = N - 1$. These do not hold in the grand canonical system, even if we replace $N$ by its average $\langle N \rangle$. This is related to the fact that also the state with $N = 0$ is possible and its probability is $\mathrm{P}(N = 0) = 1/\Xi$. Instead, the following holds

$$\int \int \overline{n}_2(x,y)dxdy - \int n(x)dx + 1 = \mathrm{P}(N = 0), \tag{B.8}$$

because

$$\begin{aligned} &\int \int \overline{n}_2(x,y)dxdy - \int n(x)dx \\ &\quad = \sum_{k=2}^{\infty} (k-1)\mathrm{P}(N=k) - \sum_{k=1}^{\infty} k\mathrm{P}(N=k) \\ &\quad = -\sum_{k=1}^{\infty} \mathrm{P}(N=k) \\ &\quad = \mathrm{P}(N=0) - 1. \end{aligned}$$

Thus, it is possible to express $\Xi$ as a functional of $n$ and $\overline{n}_2$, whereas above, it was a functional of $u$ and $\phi$,

$$\Xi = \Big(\int \int \overline{n}_2(x,y)dxdy - \int n(x)dx + 1\Big)^{-1}. \tag{B.9}$$

Since $\Xi$ is usually huge, $\int \int \overline{n}_2(x, y)dxdy - \int n(x)dx + 1$ is extremely small, but still larger than zero.

We now define $n^k(x) = \langle \delta(x - x_k) \rangle$, which looks similar to the "probability density of finding the $k$-th particle at $x$", but with the caveat that there might not be a $k$-th particle in the system at all. Thus, $n^k$ is not normalized to 1. However, we can write

$$\int_a^b n^k(x)dx = \langle \theta_{a \leq x_k < b} \rangle = \mathrm{P}(N \geq k, a \leq x_k < b), \tag{B.10}$$

since the average of an indicator function $\theta$ of an event is the probability of this event. The $n^k$ densities allow us to phrase the dependencies between the probability measure P(...) and the mean particle and neighbour-pair densities $n$, $\overline{n}_2$. With the definition of $n(x) = \langle \sum_{i=1}^{N} \delta(x - x_i) \rangle$ we can write

$$\int_a^b n(x)dx = \langle \sum_{k=1}^{N} \theta_{a \leq x_k < b} \rangle = \sum_{k=1}^{\infty} \mathrm{P}(N \geq k, a \leq x_k < b). \tag{B.11}$$

If $b - a$ becomes small enough such that only one particle fits between $a$ and $b$ (for example for particles with hard cores) then the sum of probabilities can be written as a probability of the union of disjoint[1] events

$$n(x)dx = \mathrm{P}(N \geq 1, x \leq x_k < x + dx \text{ for some } k), \tag{B.12}$$

i.e. $n(x)dx$ corresponds to the probability of finding a particle between $x$ and $x + dx$. This relationship becomes exact on a lattice, where we can set $dx = 1$. A similar argument can be made for the position of neighboring pairs, such that $\overline{n}_2(x, y)dxdy$ corresponds to the probability of finding a neighboring pair in $[x, x + dx)$ and $[y, y + dy)$.

$$\overline{n}_2(x, y)dxdy = \mathrm{P}(N \geq 2, x \leq x_k < x + dx, y \leq x_{k+1} < y + dy \text{ for some } k) \tag{B.13}$$

In the following we omit the explicit minimal values for $N$, when it can be inferred by the rest of the event description.

$$\left( \int \overline{n}_2(x, y)dy \right) dx = \mathrm{P}(x \leq x_k < x + dx \text{ for some } k < N) \tag{B.14}$$

$$\left( \int \overline{n}_2(y, x)dy \right) dx = \mathrm{P}(x \leq x_k < x + dx \text{ for some } k > 1) \tag{B.15}$$

$$\left( n(x) - \int \overline{n}_2(x, y)dy \right) dx = \mathrm{P}(x \leq x_N < x + dx) \tag{B.16}$$

$$\left( n(x) - \int \overline{n}_2(y, x)dy \right) dx = \mathrm{P}(x \leq x_1 < x + dx) \tag{B.17}$$

The last two equations are quite useful to understand the inverse formulas derived in [165] as outlined in the following section.

[1] The summation over $k$ corresponds to different particles lying between $a$ and $b$. These events become disjoint if only one particle fits.

### B.1.2 Inverse problems

In inverse problems, the external potential or interaction potential or both is unknown but the corresponding densities in the grand canonical ensemble are given.

#### Inferring external and interaction potential from one and two-particle density

The following solution to the inverse problem, where both potentials are to be determined, is based on the mathematical treatment of the problem by Percus in [165]. Assuming a grand canonical ensemble with nearest neighbor interaction and given the density $n(x)$ and the nearest-neighbour pair distribution $\overline{n}_2(x,y)$ (or the normal pair distribution $n_2(x,y)$) we want to calculate the underlying nearest-neighbor interaction $\phi(x,y)$ and the external potential $u(x)$.

Following the notation of [165], numbers in function arguments encode the position of the specific particle, i.e. $f(1)$ stands for $f$ at the position of particle 1 (not necessarily the first particle), i.e. $x_1$ and $\langle x|A|y\rangle$ denotes operator $A$ applied to $x$ and $y$. The missing arguments in $\langle A|y\rangle$, $\langle x|A\rangle$ or $\langle A\rangle$ indicate integration over the missing arguments.

Given $z(1) := e^{\beta(\mu-u(1))}$ and $\langle 1|w|2\rangle := e^{-\beta\phi(1,2)}\theta(x_2-x_1)$ the one-particle and the next-neighbor densities can be calculated as follows, effectively rewriting (B.1), (B.5) and (B.7):

$$\Xi[\mu-u,\phi] = 1+\langle (I-zw)^{-1}\rangle \tag{B.18}$$

$$n(1) = \frac{1}{\Xi}\langle (I-zw)^{-1}|1\rangle z(1)\langle 1|(I-wz)^{-1}\rangle \tag{B.19}$$

$$\overline{n}_2(1,2) = \frac{1}{\Xi}\langle (I-zw)^{-1}|1\rangle z(1)\langle 1|w|2\rangle z(2)\langle 2|(I-wz)^{-1}\rangle . \tag{B.20}$$

In the inverse setting, the partition sum and the two potentials can be calculated [165] with

$$\Xi[n,\overline{n}_2] = \frac{1}{\langle \overline{n}_2\rangle - (\langle n\rangle - 1)} \tag{B.21}$$

$$z(1) = \frac{\langle I-\overline{n}_2 n^{-1}|1\rangle n(1)\langle 1|I-n^{-1}\overline{n}_2\rangle}{1-\langle (I-\overline{n}_2 n^{-1})n\rangle} \tag{B.22}$$

$$\langle 1|w|2\rangle = \frac{\langle 1|n^{-1}\overline{n}_2 n^{1-}|2\rangle\left[1-\langle (I-\overline{n}_2 n^{-1})n\rangle\right]}{\langle 1|I-n^{-1}\overline{n}_2\rangle\langle I-\overline{n}_2 n^{-1}|2\rangle}. \tag{B.23}$$

MATLAB code implementing the forward and inverse equations for a discrete grid can be found below. Already small perturbations in the densities have strong effects on the potentials and the numerical implementations of these formulas can become unstable, since for example $\langle \overline{n}_2\rangle - (\langle n\rangle - 1)$ is a difference of two very similarly large quantities. Thus, even a numerical forward followed by a numerical inverse calculation can become quickly unstable with increasing system size and average particle number.

In the previous notation, (B.22) and (B.23) can be written as

$$\begin{aligned}
e^{\beta(\mu-u(x))} &= \Xi\frac{\left(n(x)-\int\overline{n}_2(y,x)dy\right)\left(n(x)-\int\overline{n}_2(x,y)dy\right)}{n(x)}\\
&= \Xi\frac{\mathrm{P}(x_N=x)\mathrm{P}(x_1=x)}{\mathrm{P}(x_k=x\text{ for some }k)} \qquad\qquad\qquad\qquad (\mathrm{B.24})\\
e^{-\beta\phi(x,y)} &= \frac{1}{\Xi}\frac{\overline{n}_2(x,y)}{\left(n(x)-\int\overline{n}_2(x,y)dy\right)\left(n(y)-\int\overline{n}_2(x,y)dx\right)}\\
&= \frac{1}{\Xi}\frac{\mathrm{P}(x_k=x,x_{k+1}=y\text{ for some }k)}{\mathrm{P}(x_N=x)\mathrm{P}(x_1=y)} \qquad\qquad\qquad\qquad (\mathrm{B.25})
\end{aligned}$$

where the last steps use the relations between densities and probabilities in the case of a lattice (B.14), (B.15), (B.16) and (B.17), where we can set $dx=1=dy$. An alternative derivation of these formulas (where we also use the convention that $\phi(x,y)=\infty$ if $x\geq y$) follows now.

$$\begin{aligned}
&\Xi\mathrm{P}(x_N=x)\mathrm{P}(x_1=x)\\
&=\frac{1}{\Xi}\left(\sum_{N=1}^{\infty}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}dx_1...dx_N\delta(x_N-x)e^{-\beta\sum_{i=1}^{N}(u(x_i)-\mu)-\beta\sum_{i=1}^{N-1}\phi(x_i,x_{i+1})}\right)\\
&\quad\cdot\left(\sum_{N=1}^{\infty}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}dx_1...dx_N\delta(x_1-x)e^{-\beta\sum_{i=1}^{N}(u(x_i)-\mu)-\beta\sum_{i=1}^{N-1}\phi(x_i,x_{i+1})}\right)\\
&=\frac{e^{\beta(\mu-u(x))}}{\Xi}\sum_{N=1}^{\infty}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}dx_1...dx_N\left(\sum_{k=1}^{N}\delta(x_k-x)\right)e^{-\beta\sum_{i=1}^{N}(u(x_i)-\mu)-\beta\sum_{i=1}^{N-1}\phi(x_i,x_{i+1})}\\
&=e^{\beta(\mu-u(x))}\mathrm{P}(x_k=x\text{ for some }k)
\end{aligned}$$

together with

$$\begin{aligned}
&\Xi\mathrm{P}(x_N=x)\mathrm{P}(x_1=y)\\
&=\frac{1}{\Xi}\left(\sum_{N=1}^{\infty}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}dx_1...dx_N\delta(x_N-x)e^{-\beta\sum_{i=1}^{N}(u(x_i)-\mu)-\beta\sum_{i=1}^{N-1}\phi(x_i,x_{i+1})}\right)\\
&\quad\cdot\left(\sum_{N=1}^{\infty}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}dx_1...dx_N\delta(x_1-y)e^{-\beta\sum_{i=1}^{N}(u(x_i)-\mu)-\beta\sum_{i=1}^{N-1}\phi(x_i,x_{i+1})}\right)\\
&=\frac{1}{\Xi e^{-\beta\phi(x,y)}}\sum_{N=2}^{\infty}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}dx_1...dx_N\left(\sum_{k=1}^{N-1}\delta(x_k-x)\delta(x_{k+1}-y)\right)\\
&\qquad\qquad\qquad\qquad\cdot e^{-\beta\sum_{i=1}^{N}(u(x_i)-\mu)-\beta\sum_{i=1}^{N-1}\phi(x_i,x_{i+1})}\\
&=e^{\beta\phi(x,y)}\mathrm{P}(x_k=x,x_{k+1}=y\text{ for some }k)
\end{aligned}$$

gives the desired relationships.

The following MATLAB function implements (B.19) and (B.20) for a discrete system.

```
function [n, n2] = pots2dens(u, phi)
% Calculates one— and two—body densities from external potatial u and interaction
    potential phi (u and phi are vectors with length of the system)

    if size(u,1)==1
        u = u';
    end
    L = length(u);
    z_vec = exp(—u);
    z = diag(z_vec);
    w_vec = exp(—phi);
    w = zeros(L);
    for i = 1:L
        j = i+1:L;
        w(i,j)=w_vec(j—i);
    end
    id = eye(L);
    m_zw = inv(id—z*w);
    m_wz = inv(id—w*z);
    Xi = 1+sum(sum(m_zw*z));
    n = sum(m_zw,1)'.*z_vec.*sum(m_wz,2)./Xi;
    n2 = zeros(L);
    for i=1:L
        j=i+1:L;
        n2(i,j)=sum(m_zw(:,i),1)'*z_vec(i)*w(i,j)'.*z_vec(j).*sum(m_wz(j,:),2)./Xi;
    end
end
```

The following MATLAB function implements (B.22) and (B.23) for a discrete system.

```
function [u, phi] = dens2pots(n, n2)
% Calculates the external potatial u and interaction potential phi from one— and two—
    body densities (n vector, n2 matrix)

    Xi = 1/(sum(n2(:))+1—sum(n));
    z_vec = Xi./n.*transpose(sum(diag(n)—n2,1)).*sum(diag(n)—n2,2);
    w = n2./(sum(diag(n)—n2,2)*sum(diag(n)—n2,1)*Xi);
    u = —log(z_vec);
    phi = —log(w(1,:));
end
```

### Inferring the external potential from one-particle density for given interaction

Percus also found mathematical solutions for the inverse problem of finding the external potential from the one-particle density for a given interaction potential. An explicit analytic solution is known for the hard core interaction and the sticky core interaction [166, 167], but for arbitrary interaction the analytic solution can only be given implicitly by a functional differential equation [167]. An example of a simple numerical solution to this problem is the calculation of a boundary neutralizing external potential in the following Section B.2.

Furthermore this problem can be treated numerically with the amoeba method, as has been done in [168].

### Inferring the interaction potential from next-neighbor density in a homogeneous system

If the system is homogenous, i.e. the external potential is constant, and the interaction is known to have a hard core, the interaction potential can be calculated from the next-neighbor pair density [169]. Assuming $u(x) = 0$ and $\phi(x, y) = \phi(|x-y|)$, a hard core of size $c$ and a next-neighbor interaction with range less then $2c$, a relatively simple formula for the dyad-to-dyad distance distribution $\hat{n}_2(r) = \int \overline{n}_2(x, x+r)dx$ can be derived [169]:

$$\hat{n}_2(r) = \begin{cases} 0 & \text{for } 0 \leq r < c, \\ \exp\left[-\beta pr - \beta\phi(r)\right]/\Omega(\beta p) & \text{for } c \leq r \leq 2c \end{cases} \tag{B.26}$$

with $p$ being the pressure and $\Omega(s) = \int_0^\infty \exp\left[-sr - \beta\phi(r)\right] dr$. Then, $\beta p$ and $\Omega(\beta p)$ can be obtained from a linear fit to $\ln(\hat{n}_2(r))$ for $r \lesssim 2c$, where $\phi(r)$ vanishes. After that, $\phi(r)$ for $c < r < 2c$ follows from (B.26). Even though the main assumptions are not valid, this methods has been applied to the next-neighbor distance data of the chemical cleavage method of Brogaard et al. [29] for *S. cer.* and Moyle-Heyrman et al. [170] for *S. pombe* in [44] (Supplement) and the resulting interaction potentials agree well with a fit of the same data to the soft nucleosome interaction model motivated by nucleosome breathing designed in [41].

## B.2 Boundary neutralizing external potential

With the typical nucleosome array pattern being so robust under several different types of perturbations of external potential and active remodeling [44], we wondered under what circumstances the array pattern vanishes completely in a system with boundaries, resulting in a constant non-zero particle density without changing the nucleosome interaction potential $\phi$. This is the inverse problem from the previous section, where the external potential is sought after, given a known interaction. Such an external potential could be used in non-periodic systems, where one does not want the boundaries to influence the results of other mechanisms. In these cases, usually only the center part of the system away from the boundaries is considered, where the oscillations due to the boundary are dampened enough, leading to a larger systems size and longer calculation times.

In the following we present a simple algorithm that numerically calculates a “continuous” optimal external potential that is non-constant only within distance $d$ of both boundaries and results in a constant non-zero density $\rho$ throughout the system of length $L$. Starting with an initial, possibly constant, potential guess $u_0$, the update formula for the $k + 1$th guess is

$$u_{k+1}(x) = u_k(x) + \gamma_k \left(n_k(x) - \rho\right), \tag{B.27}$$

with $x$ being the positions within distance $d$ of the boundaries, $n_k$ the dyad density resulting from external potential $u_k$ and $\gamma_k$ a factor determining the correction given with density difference. In the region not near the boundary, i.e. $x > d$ or $x < L - d$, we set $u(x) = u(d)$,

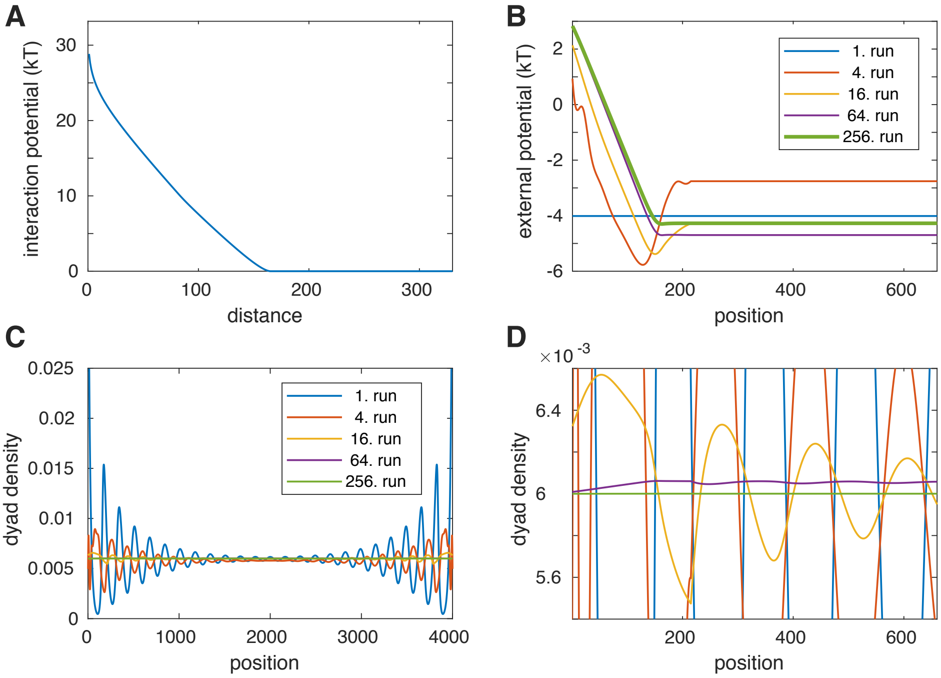


**Figure B.1:** Boundary neutralizing external potential in the soft core gas model. (**A**) Nucleosome interaction potential used for the soft nucleosome gas model with parameters $\varepsilon = 0.152\,\mathrm{k_B T}$ and $w = 82$. (**B**) Iteratively optimized external potentials already include the chemical potential and are only allowed to be non-constant until $d = 215$ away from the boundaries. The optimal potential after 256 iterations (green) has a global minimum around 160. (**C**) Dyad densities for the iterative runs. In this optimization the density goal is 0.006. (**D**) As panel C but zoomed in.

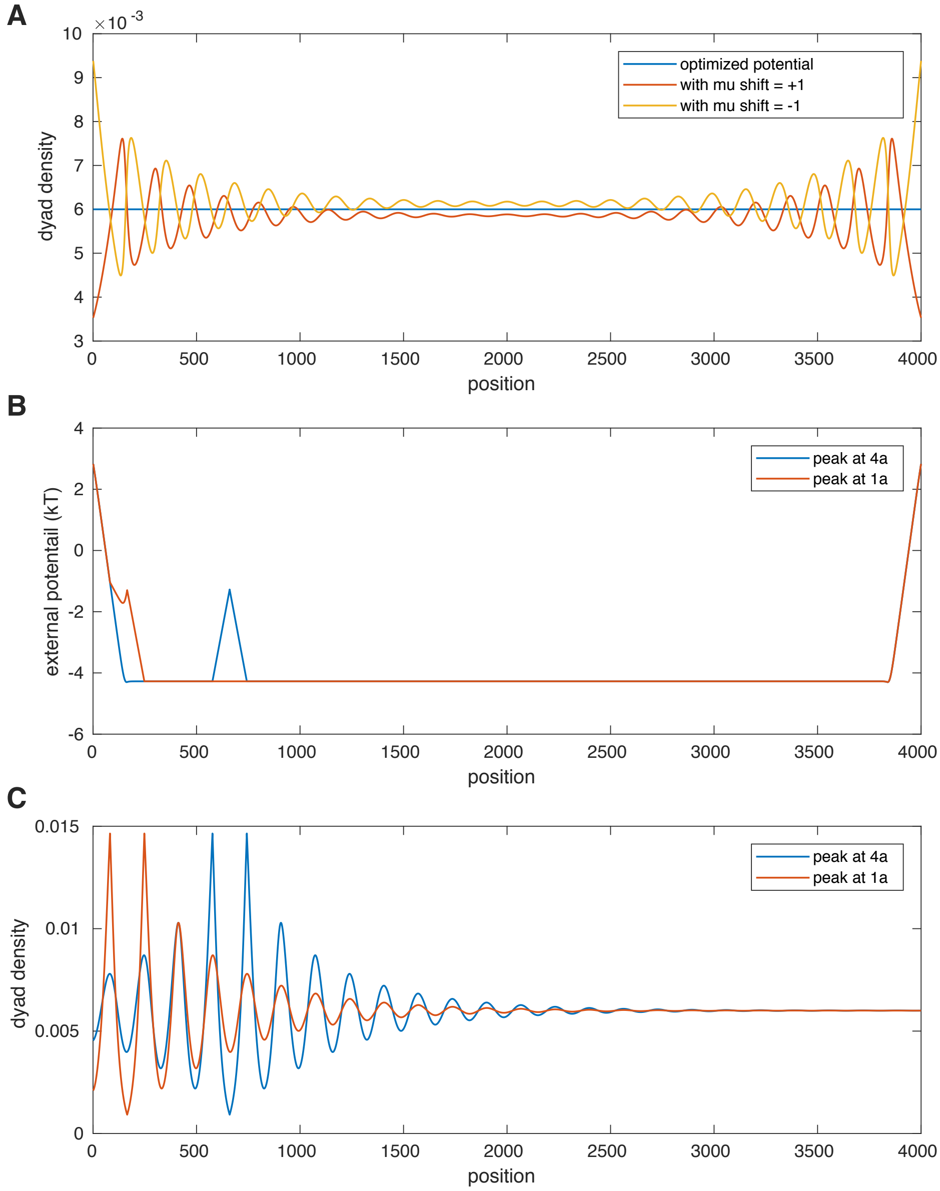


**Figure B.2:** Robustness of the boundary neutralizing external potential. (**A**) Dyad densities resulting from shifts of the optimized external potential (density goal 0.006) by $\Delta\mu = \pm 1\,\mathrm{k_B T}$. (**B**) Optimized external potential modified with triangular peaks at $4a$ and $1a$ ($a = 2w + 1$). (**C**) Dyad densities resulting from the external potentials in panel B.

which is equal to $u(L-d)$ due to the system's symmetry. $\gamma_k = 1\,\mathrm{k_B T\,bp}$ worked well in most tested cases. In the following length 1 corresponds to 1 bp. $n_k$ can be calculated from $u_k$ and $\phi$ using (B.19) and its implementation in MATLAB.

Using the soft core nucleosome gas interaction with parameters $\varepsilon = 0.152\,\mathrm{k_B T}$ and $w = 82$ (Figure B.1A) and a density goal of $\rho = 0.006$, corresponding to 88.2% absolute occupancy (147 bp nucleosome footprint), we were able to find an optimal external potential (Figure B.1B-D) after 256 iterations with the length of the non-constant part being as small as $d = 215$. Note that the non-constant part of $u_{256}$ is non-monotonous with a minimum around $x = 160$.

Further testing showed that the form of the non-constant part depends on the density goal $\rho$, as simply shifting the optimized potential by $\Delta\mu = \pm 1\,\mathrm{k_B T}$ does not result in constant densities anymore (Figure B.2A). We also investigated local potential perturbations like triangular peaks very close and further away from the boundary (Figure B.2B). Here we did not expect any constant density, but rather wanted to check if perturbations near and further away from the boundary result in approximately the same density fluctuations, which we found to be valid as long as the perturbations are not too strong (Figure B.2C).

## B.3 Remodeler simulation details

We use a rejection-free kinetic Monte-Carlo algorithm to simulate the remodeled nucleosome systems, in some cases with a periodic boundary condition, in others with a nucleosome fixed at position 0 and N+1, just outside the simulated region interacting with the particles inside. The following sections show a simple algorithm to calculate the soft core interaction potential for nucleosome and DNA-binding remodeler complexes and then define the reaction rates needed for the Monte-Carlo algorithm.

### B.3.1 Interaction potential calculation

All particle interactions are next-neighbor interactions, defined by an extension of the soft core nucleosome gas model [41], now allowing different binding energies and unwrapping lengths for the two interacting particles. The different unwrapping lengths are used for DNA-binding remodelers, treating the remodeler complex as one particle with the same dyad as the nucleosome, but longer unwrapping lengths towards the side where the remodeler binds the DNA with the same DNA binding energies per bp as the nucleosome. This a strong assumption that is probably not justified, but is a simple available model for DNA-binding remodeler interaction and most likely gives the same qualitative results as other alternatives.

The following MATLAB function efficiently calculates the interaction potential of effective unwrapping of two particles with possibly different wrapping lengths and energies by first calculating the sum of the Boltzmann weights for fixed total numbers of wrapped sites and then adding up all accessible configurations at a fixed dyad to dyad distance.

```
function phi = unwrapping_potential(w1, w2, E1, E2)
% Generate the effective interaction potential of site wrapping and unwrapping for
    particle interaction
% w1 and w2 = number of sites the particles can unwrap on the side facing the other
    particle
% E1 and E2 = binding energies of these sites
% The order of the particles does not matter.
% If one of w1 or w2 is zero, the corresponding particle is a hard particle that does
    not unwrap at all.

    d = w1 + w2;  % max number of unwrapped sites
    y = zeros(d+1,1);
    % y(n+1) = sum of boltzmann weights of wrapping configurations with exactly n
        bound/wrapped sites

    for i = 0:w1
        for j = 0:w2
            n = i+j;
            y(n+1) = y(n+1) + exp(i*E1)*exp(j*E2);
        end
    end
    phi = -log(cumsum(y(1:d))) + log(sum(y));
    % cumsum adds up all wrapping states that are possible at the corresponding dyad
        distance
    % y(1:d) because "d+1" gives 0 interaction energy at distance d+1 anyway
end
```

### B.3.2 Rate calculation

We use the chemical potentials $\mu_N$ for nucleosomes, $\mu_R$ for remodelers and $\mu_C$ for remodeler nucleosome complexes and apply them fully to the adsorption rate. In simulations with several remodeler types, e.g. upstream and downstream sliding remodelers, $\mu_R$ and $\mu_C$ and all other rate calculations apply to each type individually. The time scales of nucleosome, remodeler and complex adsorption and desorption are defined by $r_N$, $r_R$ and $r_C$, respectively. $\Delta\phi_{+N}$, $\Delta\phi_{NC}$ and $\Delta\phi_{+C}$ denote the changes in the total interaction energies due to nucleosome adsorption, remodeler binding to a nucleosome and complex binding to DNA, respectively, at position $x$. With the Boltzmann weight parameter $B = 1/2$, changes in the interaction energies apply equally to adsorption and desorption rates. The binding and unbinding rates for nucleosomes are:

$$r_+^N = r_N \exp[\mu_N - B\Delta\phi_{+N}] \tag{B.28}$$

$$r_-^N = r_N \exp[(1-B)\Delta\phi_{+N}] \tag{B.29}$$

$$\Rightarrow \frac{r_+^N}{r_-^N} = \exp[\mu_N - \Delta\phi_{+N}] \tag{B.30}$$

The binding and unbinding rates for remodelers are:

$$\begin{aligned} r_+^R &= r_R \exp[\mu_R - B\Delta\phi_{NC}] && \text{(B.31)}\\ r_-^R &= r_R \exp[(1-B)\Delta\phi_{NC}] && \text{(B.32)}\\ \Rightarrow \frac{r_+^R}{r_-^R} &= \exp[\mu_R - \Delta\phi_{NC}] && \text{(B.33)} \end{aligned}$$

The binding and unbinding rates for complexes are:

$$\begin{aligned} r_+^C &= r_C \exp[\mu_C - B\Delta\phi_{+C}] && \text{(B.34)}\\ r_-^C &= r_C \exp[(1-B)\Delta\phi_{+C}] && \text{(B.35)}\\ \Rightarrow \frac{r_+^C}{r_-^C} &= \exp[\mu_C - \Delta\phi_{+C}] && \text{(B.36)} \end{aligned}$$

For non-DNA-binding remodelers, the interaction between all particles is the same, only depends on the dyad to dyad distance $d$ and is denoted by $\phi(d)$. Given position $x$, we denote the left and right neighbor particle positions with $x_L < x$ and $x_R > x$. We then have the following changes in interaction energy:

$$\begin{aligned} \Delta\phi_{+N} =& \phi(x - x_L) + \phi(x_R - x) - \phi(x_R - x_L) && \text{(B.37)}\\ \Delta\phi_{+C} =& \Delta\phi_{+N} && \text{(B.38)}\\ \Delta\phi_{NC} =& 0 && \text{(B.39)}\\ \Delta\phi_s =& \phi(x - s - x_L) + \phi(x_R - x + s)\\ &- \phi(x - x_L) - \phi(x_R - x), && \text{(B.40)} \end{aligned}$$

with $\Delta\phi_s$ being the change in interaction energy due to sliding by $s$ (usually 1 bp) in upstream direction. Using the bias parameter $\gamma$, the upstream sliding rate $r_u$, i.e. the rate of the sliding reaction from $x$ to $x - s$, and the downstream sliding rate $r_d$, i.e. the rate of the sliding reaction from $x - s$ to $x$, can be calculated with

$$\begin{aligned} r_u &= r \exp[\pm\gamma/2 - \Delta\phi_s/2] && \text{(B.41)}\\ r_d &= r \exp[\mp\gamma/2 + \Delta\phi_s/2] && \text{(B.42)}\\ \Rightarrow \frac{r_u}{r_d} &= \exp[\pm\gamma - \Delta\phi_s], && \text{(B.43)} \end{aligned}$$

where the top sign applies for upstream movers and the bottom sign for downstream movers.

Note that if $\mu_C = \mu_N + \mu_R$, detailed balance holds for pure adsorption and desorption processes, which is preferable, since we only want to break detailed balance by the action of the remodelers. For the eviction remodelers, we choose $r_-^C = (r_C + r_e)\exp[(1-B)\Delta\phi_{+C}]$, increasing the normal complex disassembly rate by $r_e \exp[(1-B)\Delta\phi_{+C}]$ and thus breaking detailed balance. If $\gamma = 0$, detailed balance also holds for sliding remodelers, since sliding complexes then perform a symmetric random walk.

For DNA-binding remodelers, the interactions with the neighbors depend on the type and the orientation of all involved particles and the expressions for the changes in interaction energy

$\Delta\phi_{+N}$, $\Delta\phi_{NC}$, $\Delta\phi_{+C}$ and $\Delta\phi_s$ need to take this into account. As before, the position of a nucleosome remodeler complex is still defined by the nucleosome dyad. The DNA-binding by the remodeler on one side of the nucleosome leads to different interactions towards the different directions (see Section B.3.1). In our DNA-binding remodeler simulations we assume, that the remodeler binds an extra $w_R = w = 82\,\text{bp}$ of DNA directly next to the nucleosome and that this extra DNA unwraps the same way from the remodeler as from the nucleosome (same binding energy per base pair), leading to twice the amount of "unwrappable" base pairs on this side of the nucleosome dyad. Let $\tau_x^R$ and $\tau_x^L$ encode the relevant interaction parameters, in our case the number of unwrappable base pairs, for the interaction of the (possibly new) particle at $x$ towards the right and the left direction, respectively. For instance, for a remodeler complex with remodeler DNA binding in downstream direction, $\tau_x^R = w + w_R$ and $\tau_x^L = w$. Furthermore, let $\tau_{x_L}^R$ and $\tau_{x_R}^L$ encode the unwrapping length parameters for the interaction of the neighbors at $x_L$ and $x_R$ towards $x$. These are then used in the soft core interaction potential $\phi(d, w_1, w_2)$ which is calculated with the algorithm in Section B.3.1. We then have

$$\Delta\phi_{+N} = \phi(x - x_L, \tau_{x_L}^R, w) + \phi(x_R - x, w, \tau_{x_R}^L) - \phi(x_R - x_L, \tau_{x_L}^R, \tau_{x_R}^L) \tag{B.44}$$

$$\Delta\phi_{+C} = \phi(x - x_L, \tau_{x_L}^R, \tau_x^L) + \phi(x_R - x, \tau_x^R, \tau_{x_R}^L) - \phi(x_R - x_L, \tau_{x_L}^R, \tau_{x_R}^L) \tag{B.45}$$

$$\Delta\phi_{NC} = \Delta\phi_{+C} - \Delta\phi_{+N} \tag{B.46}$$

$$\begin{aligned}\Delta\phi_s = &\phi(x - s - x_L, \tau_{x_L}^R, \tau_x^L) + \phi(x_R - x + s, \tau_x^R, \tau_{x_R}^L)\\ &- \phi(x - x_L, \tau_{x_L}^R, \tau_x^L) - \phi(x_R - x, \tau_x^R, \tau_{x_R}^L).\end{aligned} \tag{B.47}$$

The expression for $\Delta\phi_{+C}$ thus depends on the orientation of the new remodeler complex at $x$ as well as the type and orientation of the neighbors. Sliding does of course not change the type and orientation of particles, but only the position and in the expression of $\Delta\phi_s$, $\tau_x^L$ and $\tau_x^R$ describe the properties of the sliding particle starting at $x$.

| parameter | description | value |
|---|---|---|
| $\mu_N$ | nucleosome chemical potential | 4.012 |
| $\mu_R$ | remodeler chemical potential | $-\ln(2)$ |
| $\mu_C$ | complex chemical potential | $\mu_N + \mu_R$ |
| $r_N$ | nucleosome ads./des. timescale factor | 1 |
| $r_C$ | complex ads./des. timescale factor | 1 |
| $r_R$ | remodeler ads./des. timescale factor | 10 |
| $B$ | Boltzmann weight distribution parameter | 1/2 |
| $L$ | system size | 5028 |
| $T$ | simulated time | 10 |
| $N$ | number of realizations | 5000 |
| $r$ | sliding rate parameter | $\frac{w}{2}\frac{r_R}{8\varepsilon}$ |
| $\gamma$ | bias parameter | $8\varepsilon \cdot (0, \frac{1}{2^4}, \frac{1}{2^3}, \frac{1}{2^2}, \frac{1}{2}, 1)$ |
| $s$ | jump size | 1 |
| $p$ | calculated processivity | [0, 2.3, 4.7, 9.4, 19, 40] |
| $r$ | sliding rate parameter | $\frac{w}{2}\frac{r_R}{\varepsilon} \cdot (\frac{1}{2^5}, \frac{1}{2^4}, \frac{1}{2^3}, \frac{1}{2^2}, \frac{1}{2}, 1, 2)$ |
| $\gamma$ | bias parameter | $\varepsilon$ |
| $s$ | jump size | 1 |
| $p$ | calculated processivity | (1.2, 2.3, 4.7, 9.3, 19, 37, 75) |

**Table B.1:** Simulation parameters for directional remodelers near a boundary. The nucleosome properties are chosen from the fit to *S. cerevisiae* data in [44], namely $\varepsilon = 0.152$ and $w = 82$. $\mathrm{k_B T}$ as well as 1 bp is set to 1. $\mu_N$ and $\mu_C$ correspond to the chemical potential of nucleosomes and complexes per lattice site (here 1 bp), respectively, $\mu_R$ to the chemical potential of remodelers per accessible nucleosome. The lower two blocks contain the remodeler action parameters for the parameters sweep in $\gamma$ as well as $r$ used in Figure 3.2E and Figure 3.3B.

| parameter | description | value |
|---|---|---|
| $\mu_N$ | nucleosome chemical potential | $4.012 - \ln(2w+1)$ |
| $r_N$ | nucleosome binding/unbinding timescale factor | 1 |
| $r_C$ | complex binding/unbinding timescale factor | 0 (no complexes from bulk) |
| $r_R$ | remodeler binding/unbinding timescale factor | 50 |
| $B$ | Boltzmann weight distribution parameter | 1/2 |
| $L$ | system size | 3000 |
| $T$ | simulated time | 10 |
| $N$ | number of realizations | 20000 |
| $\mu_R$ | remodeler chemical potential | $\ln(0.6/0.4)$ |
| $r_e$ | eviction rate parameter | 60 |
| $\mu_R$ | remodeler chemical potential | $\ln(0.3/0.4)$ |
| $r$ | sliding rate parameter | $0.1\frac{w}{2}\frac{r_R}{\varepsilon}$ |
| $\gamma$ | bias parameter | $10\varepsilon$ |
| $s$ | jump size | 10 |

**Table B.2:** Simulation parameters for nucleosome eviction mechanisms used in Figure 3.4. The nucleosome properties are chosen from the fit to *S. cerevisiae* data in [44], namely $\varepsilon = 0.152$ and $w = 82$. $\mathrm{k_B T}$ as well as 1 bp is set to 1. $\mu_N$ and $\mu_C$ correspond to the chemical potential of nucleosomes and complexes per lattice site (1 bp, with dependence on $w$ to enable course-graining), respectively, $\mu_R$ to the chemical potential of remodelers per accessible nucleosome. The two lower blocks contain the parameters for the eviction remodeler simulation (Figure 3.4A) and the sliding remodeler simulation (Figure 3.4C). The chemical potentials are tuned such that if the recruitment was one everywhere and the remodelers performed no actions, 40% of all particles would be nucleosomes and the rest complexes. All remodeler adsorption rates are scaled with the Gaussian recruitment peak (centered at 1500 with standard deviation of 165). The last block applies to upstream and downstream remodelers individually.

| parameter | description | value |
|---|---|---|
| $\mu_R$ | remodeler chemical potential | $-\ln(2)$ |
| $r_R$ | remodeler binding/unbinding timescale factor | 1 |
| $B$ | Boltzmann weight distribution parameter | 1/2 |
| $L$ | system size | 7410 |
| $T$ | simulated time | 1000 |
| $N$ | number of realizations | 1000 |
| $r$ | sliding rate parameter | $w$ |
| $\gamma$ | bias parameter | 0 (A, B) and $\varepsilon/2$ (C, D) |
| $s$ | jump size | 1 |

**Table B.3:** Simulation parameters for DNA binding remodelers used in Figure 3.6. The nucleosome properties are chosen from the fit to *S. cerevisiae* data in [44], namely $\varepsilon = 0.152$ and $w = 82$. $\mathrm{k_B T}$ as well as 1 bp is set to 1. The number of nucleosomes was fixed to 30, without the possibility of nucleosome adsorption/desorption.

# C Effective Dynamics of Nucleosome Configurations - Supplement [1]

## C.1 Materials and methods

We implemented the maximum likelihood fit procedure (Figure 4.4D) in MATLAB and made the source code available at `https://github.com/gerland-group/PHO5_on-off-slide_models`.

### C.1.1 Sticky N-3 experiments

The sticky N-3 measurements were done in the lab of Philipp Korber by Andrea Schmidt and Philipp Korber. The strains "sticky N-3 mutant 1" and "sticky N-3 mutant 2" used for restriction enzyme accessibility assays were generated by transformation of linear fragments of plasmids ECS53 and ECS56, respectively, into the wild type strain BY4741 as described for the "periodicity mutants" in [69]. For the sticky N-3 mutant 1, the sequence GTTTTCTCATGTAAGCGGACGTCGTC inside the *PHO5* promoter was replaced with GTTTTCTTATGTAAGCTTACGTCGTC and for the sticky N-3 mutant 2, GCGCAAATATGTCAACGTATTTGGAAG was replaced with GCGCAAATATGTCAAAGTATTTGGAAG. Strains grew in YPDA medium to logarithmic phase for repressive (+Pi) and then shifted from logarithmic phase to phosphate-free YNB medium (Formedia) for inducing (-Pi) conditions over night. The nuclei preparation, restriction enzyme digestion, DNA purification, secondary digest, agarose gel electrophoresis, Southern blotting, hybridization and Phosphorimager analysis were done as in [115]. Secondary digest was performed with HaeIII for both ClaI and HhaI digests probing N-2 or N-3, respectively and the probe for both ClaI and HhaI digests corresponded to the ApaI-BamHI restriction fragment upstream of N-3.

## C.2 Supplementary figures

[1] This supplemental chapter is adapted from our manuscript [112] under CC BY 4.0 license.

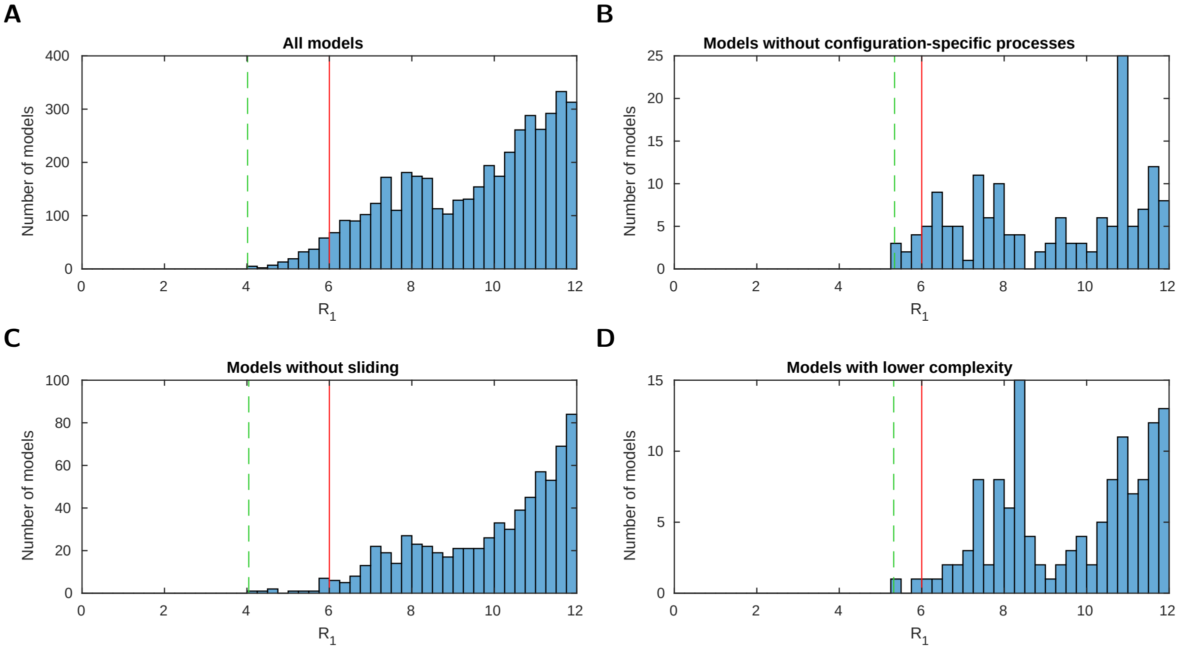


**Figure C.1:** Histograms of the logarithmic likelihood ratio with respect to the perfect fit likelihood in stage 1 ($R_1$), i.e. using the configurational data of [34] only. (**A - D**) For all models and three model subsets. 0 on the x-axis corresponds to a perfect fit. Dashed green line: value of the best model (within the subset). Red line: threshold for a satisfactory fit of $R_{max} = 6$.

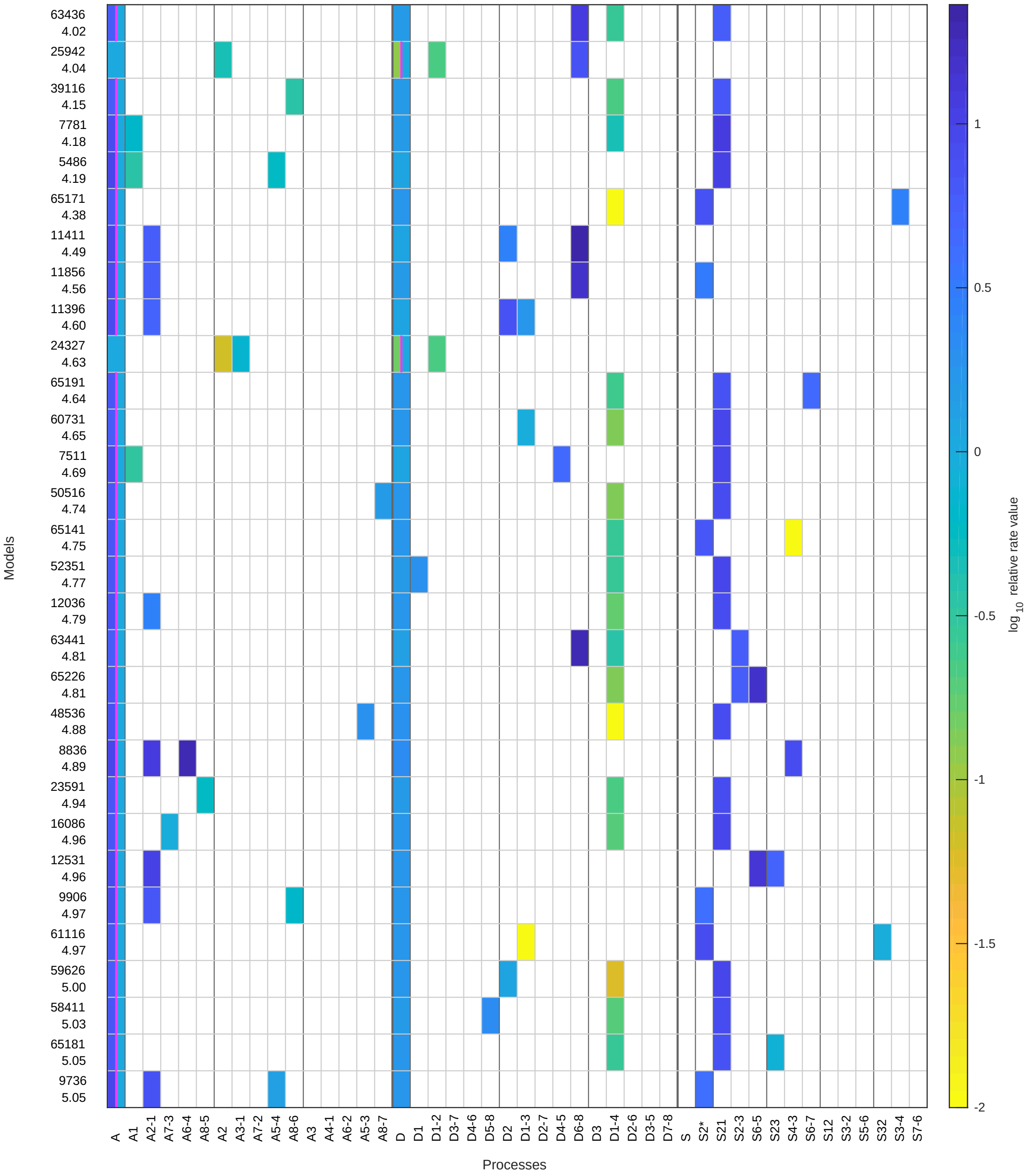


**Figure C.2:** The top 30 models with likelihood above the threshold after the first stage. The x-axis lists all possible processes and the colored boxes in each row show each model's processes with their rate values. White boxes indicate the absence of a process in a model. Regulated processes are separated into two differently colored boxes for repressed (left half) and activated (right half) promoter state. Weakly activated rate values are not presented here. On the left side are the model number and the log10 likelihood ratio $R_1$.

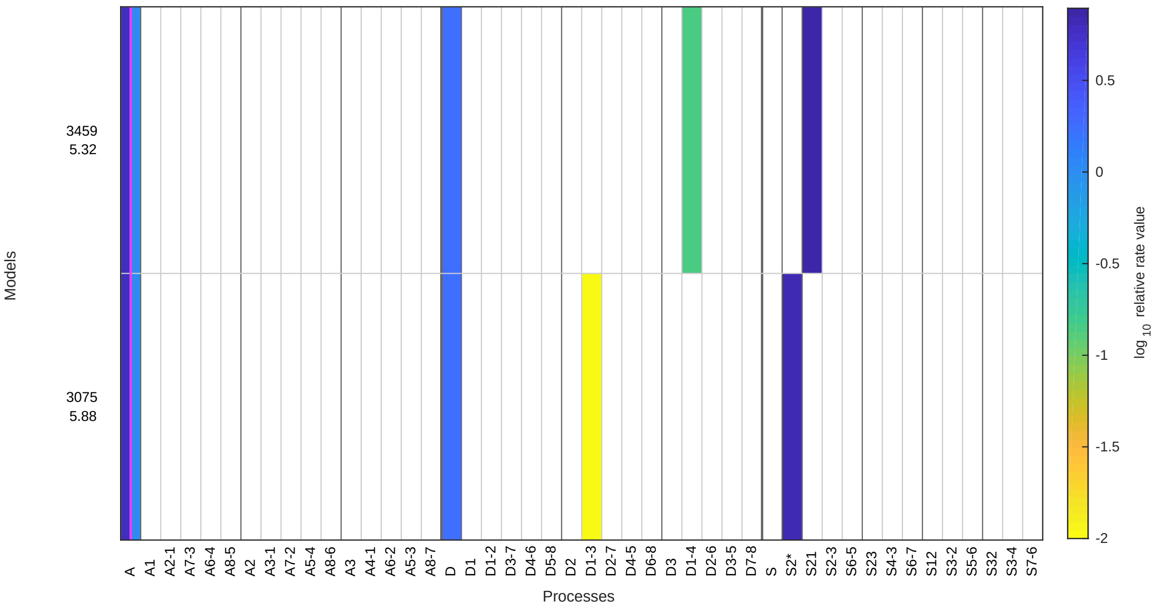


**Figure C.3:** Same as in Figure C.2, but showing satisfactory regulated on-off-slide models after stage 1 with up to only 6 fit parameters.

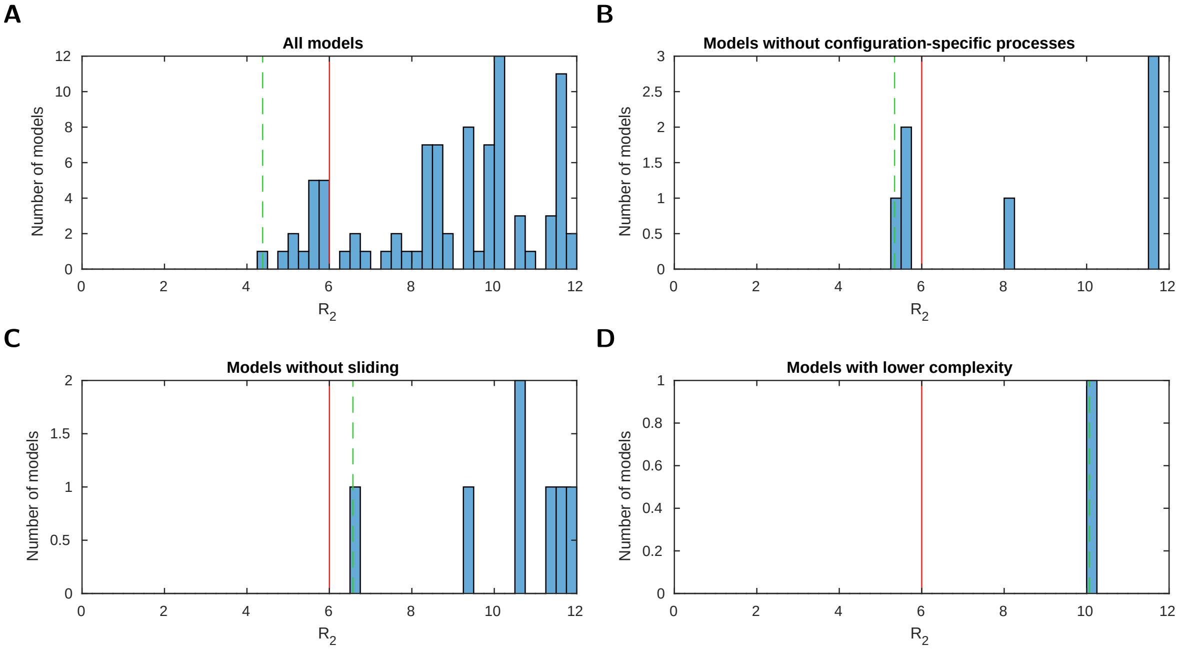


**Figure C.4:** Histograms of the logarithmic likelihood ratio with respect to the perfect fit likelihood in stage 2, $R_2$, i.e. using the configurational data of [34] and the sticky N-3 accessibility data (Table 4.2). (**A - D**) For all models and three model subsets. 0 on the x-axis corresponds to a perfect fit. Dashed green line: value of the best model (within the subset). Red line: threshold for a "good" fit of $R_{max} = 6$.

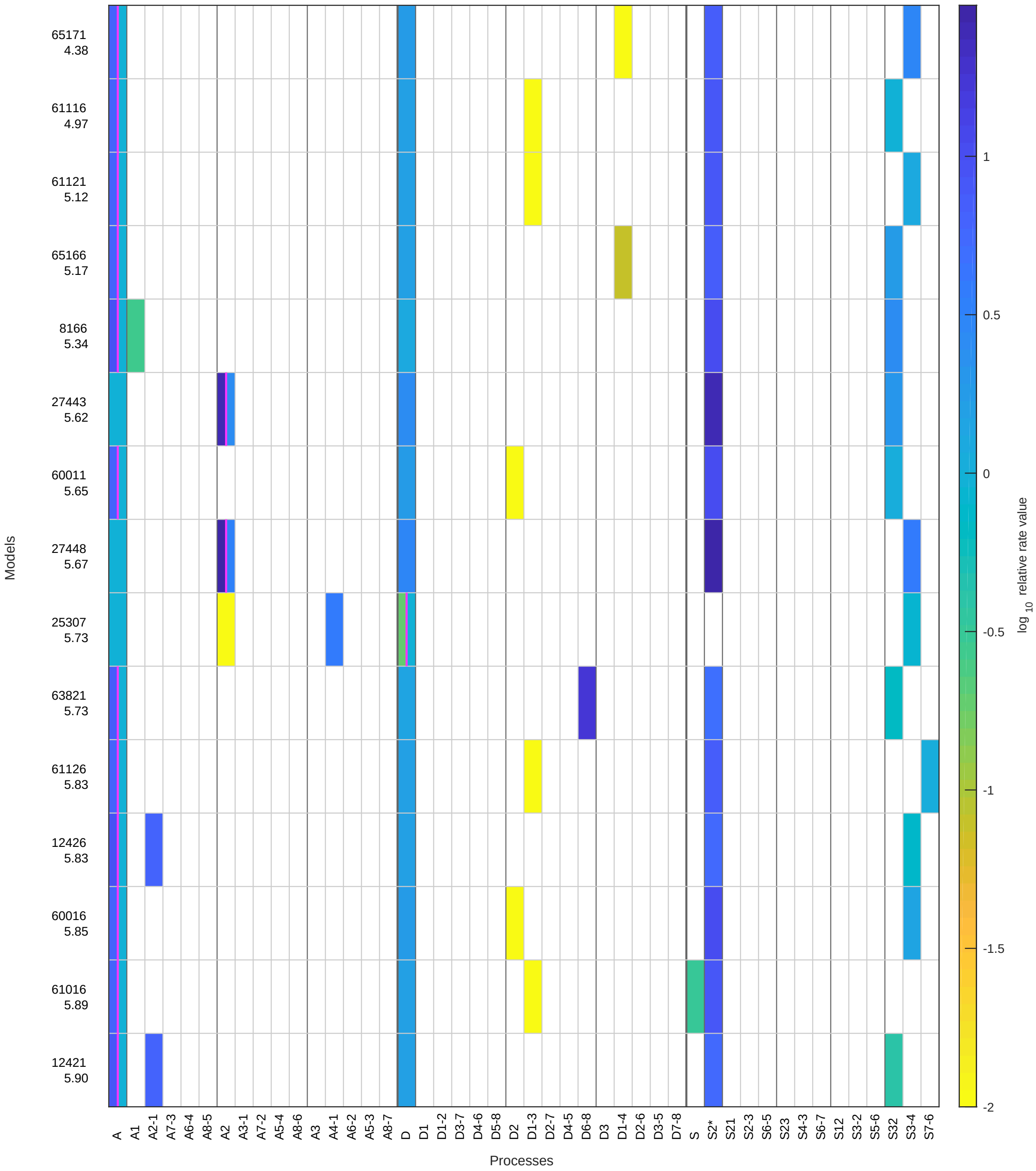


**Figure C.5:** Same as in Figure C.3, but showing the regulated on-off-slide models with likelihood above the threshold with up to 7 parameters after stage 2, with the stage 2 log10 likelihood ratios, $R_2$, written below the model numbers.

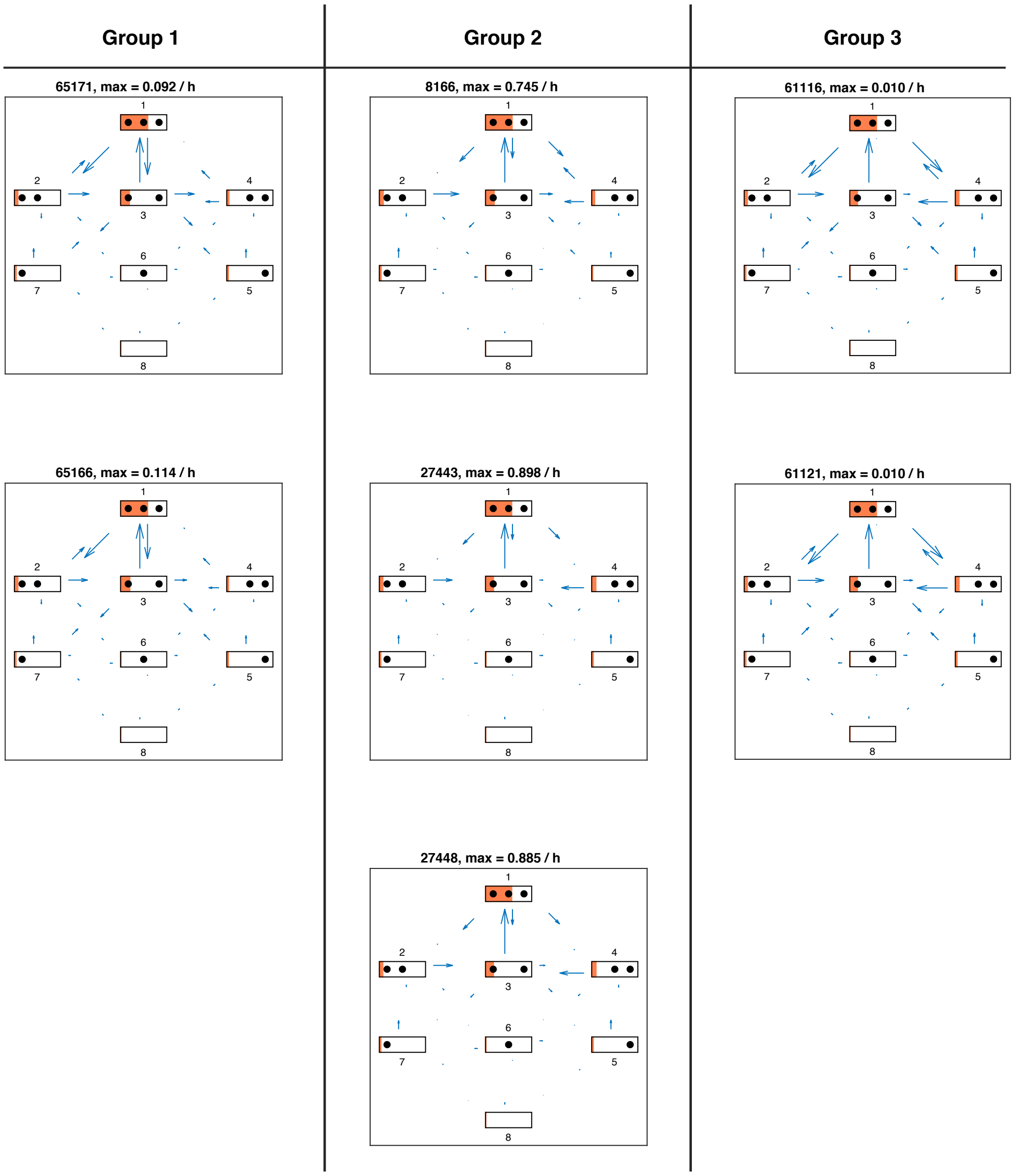


**Figure C.6:** Directional fluxes in repressed promoter state for each satisfactory model in stage 3. The length of the flux arrows indicates the amount of net flux with respect to the maximum value for each model stated above. The orange filling of each state symbol shows the steady state probabilities with a full rectangle corresponding to probability 1. The models are grouped with respect to similarities in the site-centric net fluxes (Figure C.10).

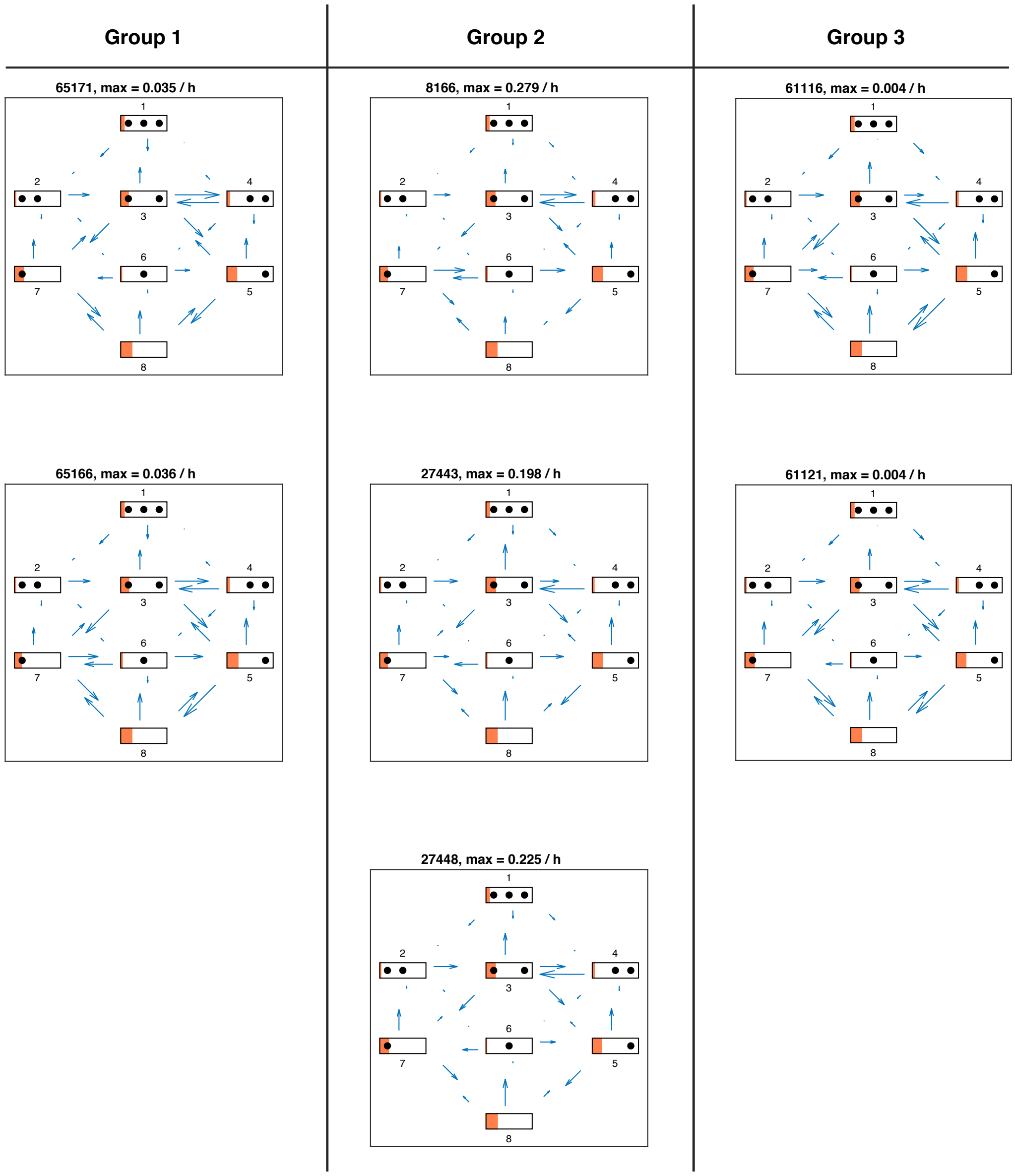


**Figure C.7:** Directional fluxes in activated promoter state for each satisfactory model in stage 3. Otherwise as Figure C.6.

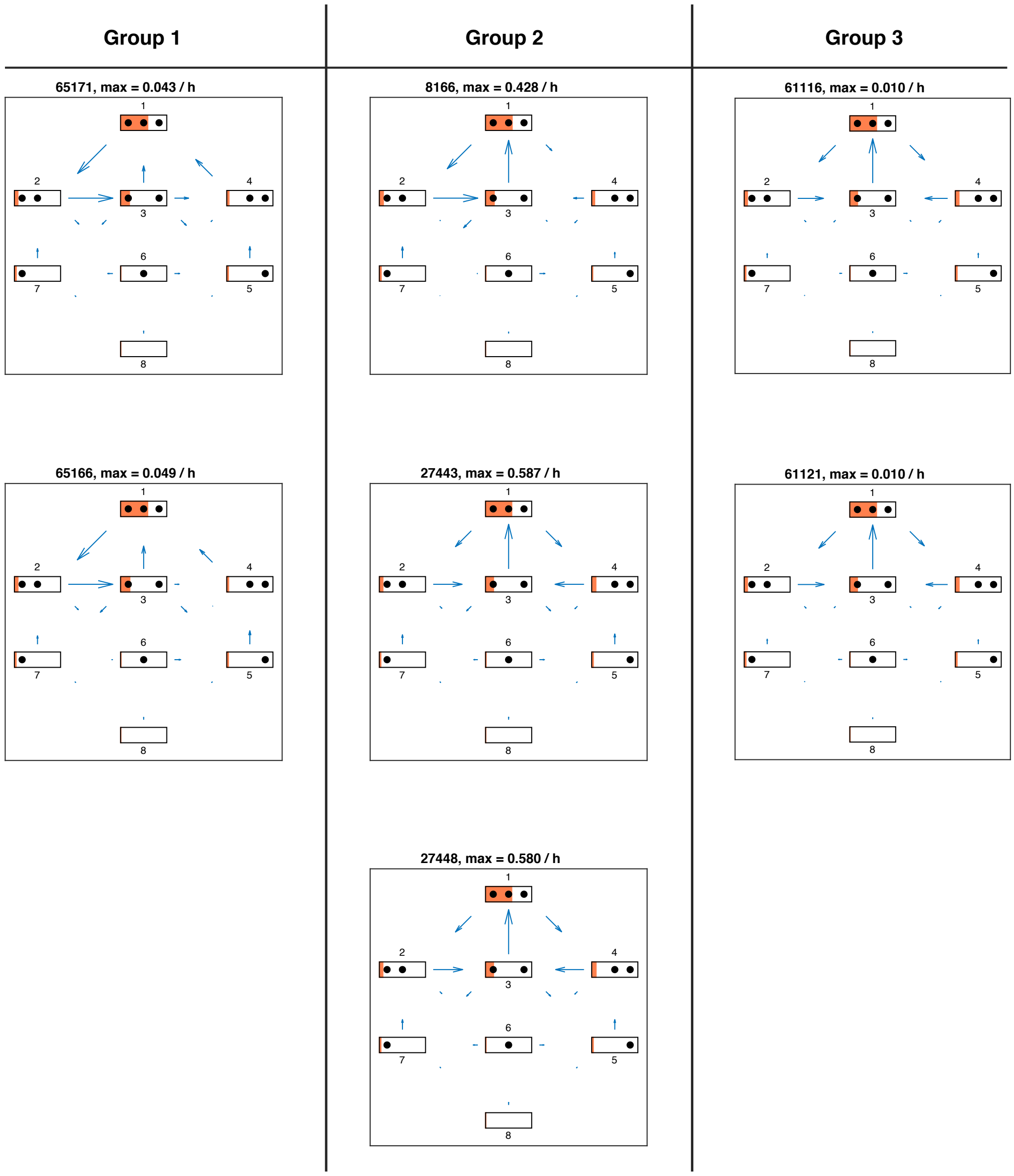


**Figure C.8:** Net fluxes in repressed promoter state for each satisfactory model in stage 3. The length of the flux arrows indicates the amount of net flux with respect to the maximum value for each model stated above. The orange filling of each state symbol shows the steady state probabilities with a full rectangle corresponding to probability 1. The models are grouped with respect to similarities in the site-centric net fluxes (Figure C.10).

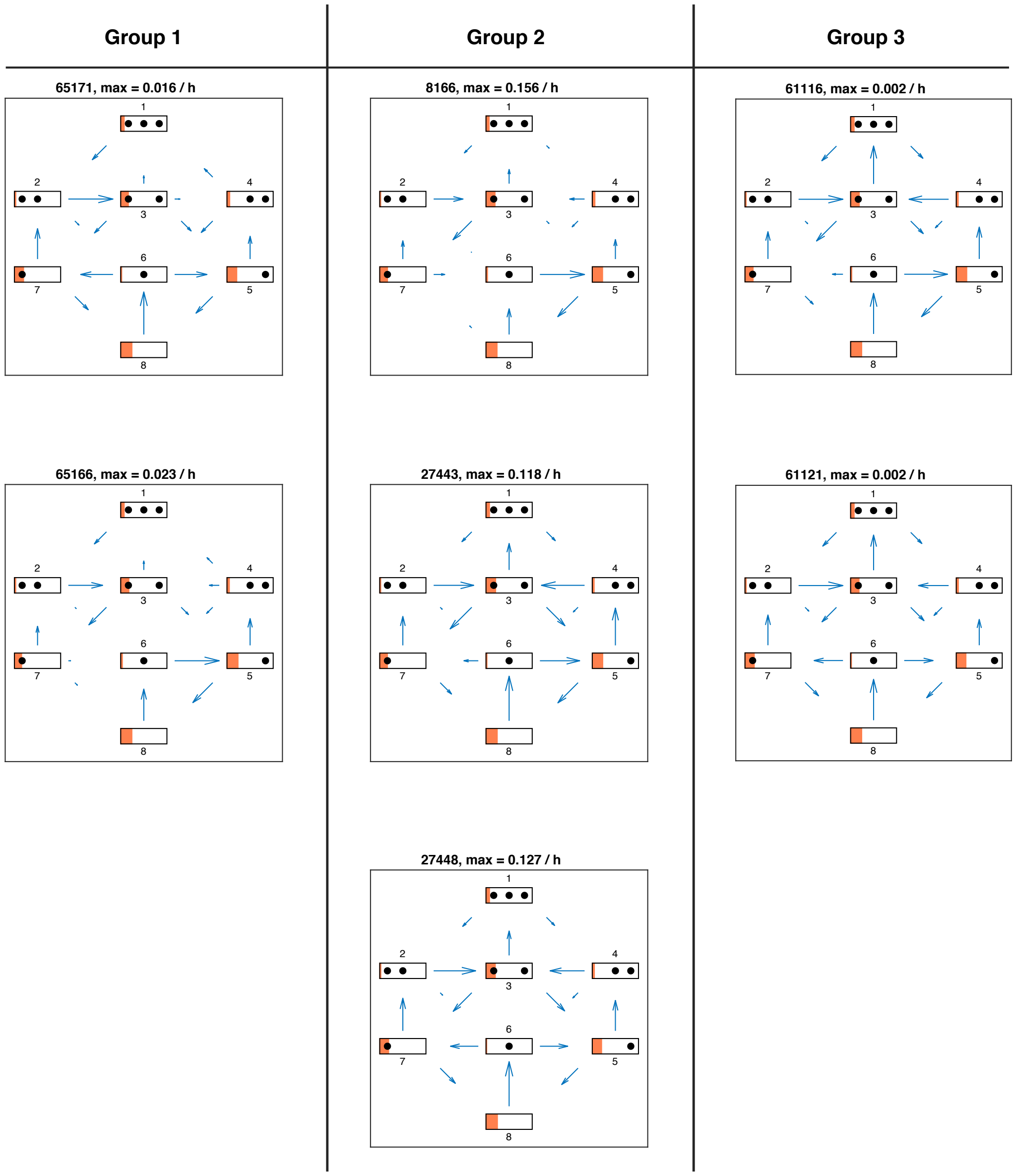


**Figure C.9:** Net fluxes in activated promoter state for each satisfactory model in stage 3. Otherwise as Figure C.8.

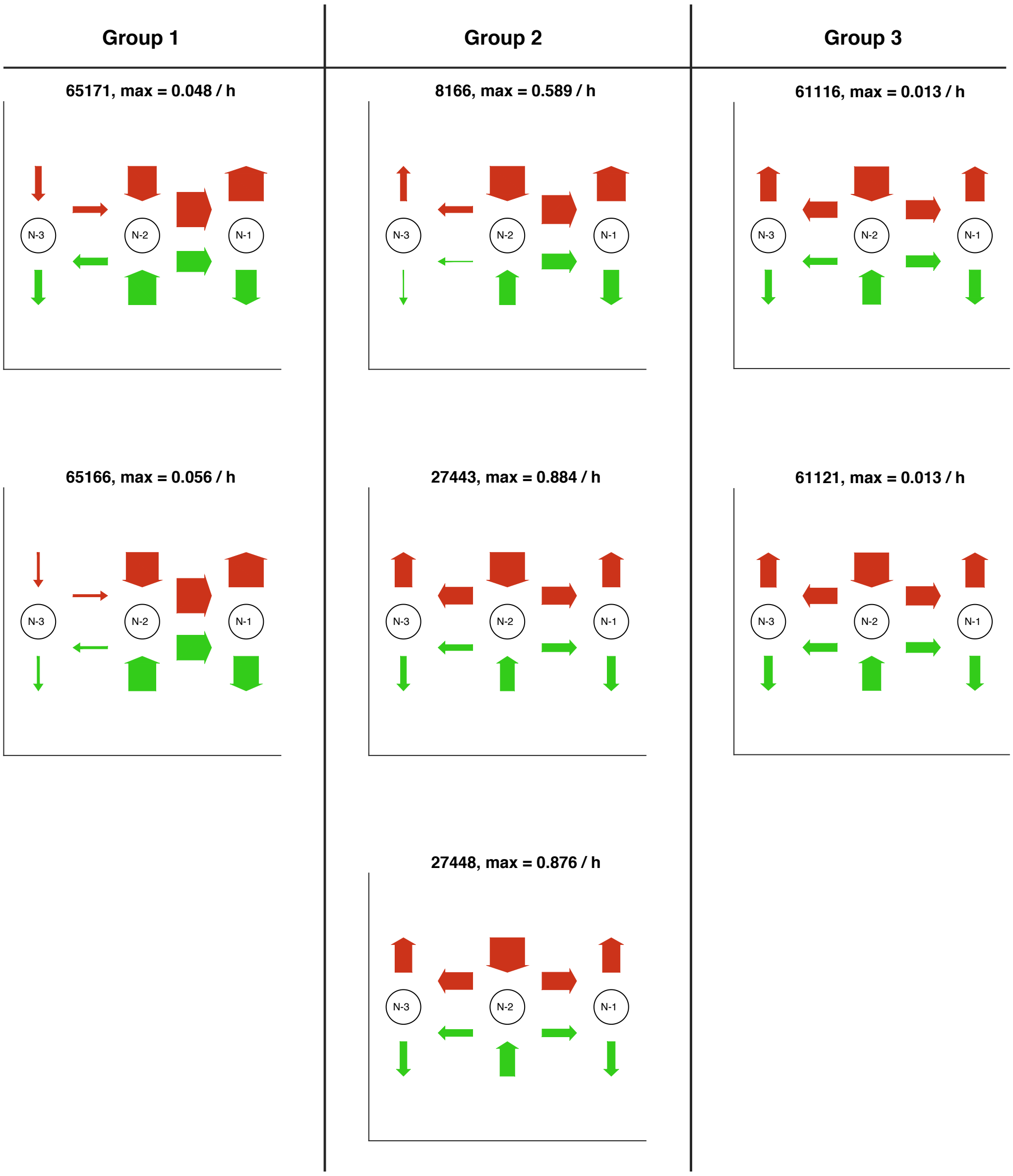


**Figure C.10:** Site-centric net fluxes in active (green) and repressed (red) promoter state for each satisfactory model in stage 3. Obtained by summing all assembly/disassembly net fluxes at each site and sliding net fluxes between N-1 and N-2 as well as N-2 and N-3. The arrow thickness indicates the amount of flux with the maximum value stated above. The models are grouped with respect to similarities in the site-centric net fluxes.

# Bibliography

[1] N. Paweletz. Walther Flemming: Pioneer of mitosis research. *Nature Reviews Molecular Cell Biology*, 2(1):72–75, 2001.

[2] Boyer. The chromatin signature of pluripotent cells. *StemBook*, 2009.

[3] R. D. Kornberg. Chromatin Structure: A Repeating Unit of Histones and DNA. *Science*, 184(4139):868–871, 1974.

[4] T. J. Richmond, J. T. Finch, B. Rushton, D. Rhodes, and A. Klug. Structure of the nucleosome core particle at 7 Å resolution. *Nature*, 311(5986):532–537, 1984.

[5] 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 Å resolution. *Nature*, 389(6648):251–260, 1997.

[6] U. Venter, J. Svaren, J. Schmitz, A. Schmid, and W. Hörz. A nucleosome precludes binding of the transcription factor Pho4 in vivo to a critical target site in the PHO5 promoter. *The EMBO Journal*, 13(20):4848–4855, 1994.

[7] O. Bell, V. K. Tiwari, N. H. Thomä, and D. Schübeler. Determinants and dynamics of genome accessibility. *Nature Reviews Genetics*, 12(8):554–564, 2011.

[8] H.-W. Chang, M. Pandey, O. I. Kulaeva, S. S. Patel, and V. M. Studitsky. Overcoming a nucleosomal barrier to replication. *Science Advances*, 2(11):e1601865, 2016.

[9] M. H. Hauer and S. M. Gasser. Chromatin and nucleosome dynamics in DNA damage and repair. *Genes & Development*, 31(22):2204–2221, 2017.

[10] A. Seeber, M. H. Hauer, and S. M. Gasser. Chromosome Dynamics in Response to DNA Damage. *Annual Review of Genetics*, 52(1):295–319, 2018.

[11] D. M. MacAlpine and G. Almouzni. Chromatin and DNA replication. *Cold Spring Harbor Perspectives in Biology*, 5(8):a010207–a010207, 2013.

[12] M. F. Dion, T. Kaplan, M. Kim, S. Buratowski, N. Friedman, and O. J. Rando. Dynamics of Replication-Independent Histone Turnover in Budding Yeast. *Science*, 315(5817):1405–1408, 2007.

[13] A. Rufiange, P.-É. Jacques, W. Bhat, F. Robert, and A. Nourani. Genome-Wide Replication-Independent Histone H3 Exchange Occurs Predominantly at Promoters and Implicates H3 K56 Acetylation and Asf1. *Molecular Cell*, 27(3):393–405, 2007.

[14] J. D. Anderson, A. Thastrom, and J. Widom. Spontaneous access of proteins to buried nucleosomal DNA target sites occurs via a mechanism that is distinct from nucleosome translocation. *Molecular and Cellular Biology*, 22(20):7147–7157, 2002.

[15] G. Li and J. Widom. Nucleosomes facilitate their own invasion. *Nature Structural & Molecular Biology*, 11(8):763–769, 2004.

[16] W. J. A. Koopmans, R. Buning, T. Schmidt, and J. van Noort. spFRET using alternating excitation and FCS reveals progressive DNA unwrapping in nucleosomes. *Biophysical Journal*, 97(1):195–204, 2009.

[17] K. Zhou, G. Gaullier, and K. Luger. Nucleosome structure and dynamics are coming of age. *Nature Structural & Molecular Biology*, 26(1):3–13, 2019.

[18] M. Engeholm, M. de Jager, A. Flaus, R. Brenk, J. van Noort, and T. Owen-Hughes. Nucleosomes can invade DNA territories occupied by their neighbors. *Nature Structural & Molecular Biology*, 16(2):151–158, 2009.

[19] H. S. Tims, K. Gurunathan, M. Levitus, and J. Widom. Dynamics of nucleosome invasion by DNA binding proteins. *Journal of Molecular Biology*, 411(2):430–448, 2011.

[20] S. Baldi, P. Korber, and P. B. Becker. Beads on a string—nucleosome array arrangements and folding of the chromatin fiber. *Nature Structural & Molecular Biology*, 27(2):109–118, 2020.

[21] R. V. Chereji and D. J. Clark. Major Determinants of Nucleosome Positioning. *Biophysical Journal*, 114(10):2279–2289, 2018.

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

[23] J. D. True, J. J. Muldoon, M. N. Carver, K. Poorey, S. J. Shetty, S. Bekiranov, and D. T. Auble. The Modifier of Transcription 1 (Mot1) ATPase and Spt16 Histone Chaperone Co-regulate Transcription through Preinitiation Complex Assembly and Nucleosome Organization. *Journal of Biological Chemistry*, 291(29):15307–15319, 2016.

[24] Z. Zhang, C. J. Wippo, M. Wal, E. Ward, P. Korber, and B. F. Pugh. A Packing Mechanism for Nucleosome. *Science*, 332:977–980, 2011.

[25] R. V. Chereji, T. D. Bryson, and S. Henikoff. Quantitative MNase-seq accurately maps nucleosome occupancy levels. *Genome Biology*, 20(1), 2019.

[26] R. V. Chereji, S. Ramachandran, T. D. Bryson, and S. Henikoff. Precise genome-wide mapping of single nucleosomes and linkers in vivo. *Genome Biology*, 19(1), 2018.

[27] K. Chen, Y. Xi, X. Pan, Z. Li, K. Kaestner, J. Tyler, S. Dent, X. He, and W. Li. DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Research*, 23(2):341–351, 2013.

[28] E. Oberbeckmann, M. Wolff, N. Krietenstein, M. Heron, J. L. Ellins, A. Schmid, S. Krebs, H. Blum, U. Gerland, and P. Korber. Absolute nucleosome occupancy map for the Saccharomyces cerevisiae genome. *Genome Research*, 29(12):1996–2009, 2019.

[29] K. Brogaard, L. Xi, J.-P. Wang, and J. Widom. A map of nucleosome positions in yeast at base-pair resolution. *Nature*, 486(7404):496–501, 2012.

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

[31] T. N. Mavrich, C. Jiang, I. P. Ioshikhes, X. Li, B. J. Venters, S. J. Zanton, L. P. Tomsho, J. Qi, R. L. Glaser, S. C. Schuster, D. S. Gilmour, I. Albert, and B. F. Pugh. Nucleosome organization in the Drosophila genome. *Nature*, 453(7193):358–362, 2008.

[32] D. E. Schones, K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, 2008.

[33] P. Korber and S. Barbaric. The yeast PHO5 promoter: From single locus to systems biology of a paradigm for gene regulation through chromatin. *Nucleic Acids Research*, 42(17):10888–10902, 2014.

[34] C. R. Brown, C. Mao, E. Falkovskaia, M. S. Jurica, and H. Boeger. Linking Stochastic Fluctuations in Chromatin Structure and Gene Expression. *PLOS Biology*, 2013.

[35] R. D. Kornberg and L. Stryer. Statistical distributions of nucleosomes: Nonrandom locations by a stochastic mechanism. *Nucleic Acids Research*, 16(14):6677–6690, 1988.

[36] T. N. Mavrich, I. P. Ioshikhes, B. J. Venters, C. Jiang, L. P. Tomsho, J. Qi, S. C. Schuster, I. Albert, and B. F. Pugh. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Research*, 18(7):1073–1083, 2008.

[37] L. Tonks. The complete equation of state of one, two and three-dimensional gases of hard elastic spheres. *Phys. Rev.*, 50(10):955–963, 1936.

[38] J. Widom. Role of DNA sequence in nucleosome stability and dynamics. *Quarterly Reviews of Biophysics*, 34(3):269–324, 2001.

[39] P. D. Partensky and G. J. Narlikar. Chromatin Remodelers Act Globally, Sequence Positions Nucleosomes Locally. *Journal of Molecular Biology*, 391(1):12–25, 2009.

[40] W. Möbius and U. Gerland. Quantitative test of the barrier nucleosome model for statistical positioning of nucleosomes up- and downstream of transcription start sites. *PLOS Computational Biology*, 6(8), 2010.

[41] W. Möbius, B. Osberg, A. M. Tsankov, O. J. Rando, and U. Gerland. Toward a unified physical model of nucleosome patterns flanking transcription start sites. *Proceedings of the National Academy of Sciences*, 110(14):5719–24, 2013.

[42] B. Osberg, J. Nuebler, P. Korber, and U. Gerland. Replication-guided nucleosome packing and nucleosome breathing expedite the formation of dense arrays. *Nucleic Acids Research*, 49(89):1–13, 2014.

[43] B. Osberg, J. Nuebler, and U. Gerland. Adsorption-Desorption Kinetics of Soft Particles. *Physical Review Letters*, 115(088301), 2015.

[44] J. Nuebler, M. Wolff, B. Obermayer, W. Möbius, and U. Gerland. Emergence of robust nucleosome patterns from an interplay of positioning mechanisms. *bioRxiv*, 2018.

[45] C. L. Peterson. Salt Gradient Dialysis Reconstitution of Nucleosomes. *Cold Spring Harbor Protocols*, 2008(12):pdb.prot5113, 2008.

[46] F. Mueller-Planitz, H. Klinker, and P. B. Becker. Nucleosome sliding mechanisms: New twists in a looped history. *Nature Structural & Molecular Biology*, 20(9):1026–32, 2013.

[47] B. Bartholomew. Regulating the Chromatin Landscape: Structural and Mechanistic Perspectives. *Annual Review of Biochemistry*, 83(1):671–696, 2014.

[48] G. J. Narlikar, R. Sundaramoorthy, and T. Owen-Hughes. Mechanisms and functions of ATP-dependent chromatin-remodeling enzymes. *Cell*, 154(3):490–503, 2013.

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

[50] C. Jiang and B. F. Pugh. Nucleosome positioning and gene regulation: Advances through genomics. *Nature Reviews Genetics*, 10(3):161–172, 2009.

[51] X. Liu, M. Li, X. Xia, X. Li, and Z. Chen. Mechanism of chromatin remodelling revealed by the Snf2-nucleosome structure. *Nature*, 544(7651):440–445, 2017.

[52] N. Krietenstein, M. Wal, S. Watanabe, B. Park, C. L. Peterson, B. F. Pugh, and P. Korber. Genomic Nucleosome Organization Reconstituted with Pure Proteins. *Cell*, 167(3):709–721.e12, 2016.

[53] T. Gkikopoulos, P. Schofield, V. Singh, M. Pinskaya, J. Mellor, M. Smolle, J. L. Workman, G. J. Barton, and T. Owen-Hughes. A Role for Snf2-Related Nucleosome-Spacing Enzymes in Genome-Wide Nucleosome Organization. *Science*, 333(6050):1758–1760, 2011.

[54] J. Ocampo, R. V. Chereji, P. R. Eriksson, and D. J. Clark. The ISW1 and CHD1 ATP-dependent chromatin remodelers compete to set nucleosome spacing in vivo. *Nucleic Acids Research*, 2016.

[55] C. Lieleg, P. Ketterer, J. Nuebler, J. Ludwigsen, U. Gerland, H. Dietz, F. Mueller-Planitz, and P. Korber. Nucleosome Spacing Generated by ISWI and CHD1 Remodelers Is Constant Regardless of Nucleosome Density. *Molecular and Cellular Biology*, 35(9):1588–1605, 2015.

[56] L. R. Racki and G. J. Narlikar. ATP-dependent chromatin remodeling enzymes: Two heads are not better, just different. *Current Opinion in Genetics and Development*, 2008.

[57] Y. Lorch, B. Maier-Davis, and R. D. Kornberg. Role of DNA sequence in chromatin remodeling and the formation of nucleosome-free regions. *Genes & Development*, 28(22):2492–2497, 2014.

[58] G. Sirinakis, C. R. Clapier, Y. Gao, R. Viswanathan, B. R. Cairns, and Y. Zhang. The RSC chromatin remodelling ATPase translocates DNA with high force and small step size. *The EMBO Journal*, 30:2364–2372, 2011.

[59] B. Bartholomew. ISWI chromatin remodeling: One primary actor or a coordinated effort? *Current Opinion in Structural Biology*, 2014.

[60] V. K. Gangaraju and B. Bartholomew. Dependency of ISW1a chromatin remodeling on extranucleosomal DNA. *Molecular and Cellular Biology*, 27(8):3217–3225, 2007.

[61] K. Yamada, T. D. Frouws, B. Angst, D. J. Fitzgerald, C. DeLuca, K. Schimmele, D. F. Sargent, and T. J. Richmond. Structure and mechanism of the chromatin remodelling factor ISW1a. *Nature*, 472(7344):448–453, 2011.

[62] M. N. Kagalwala, B. J. Glaus, W. Dang, M. Zofall, and B. Bartholomew. Topography of the ISW2-nucleosome complex: Insights into nucleosome spacing and chromatin remodeling. *The EMBO Journal*, 23(10):2092–2104, 2004.

[63] S. Deindl, W. L. Hwang, S. K. Hota, T. R. Blosser, P. Prasad, B. Bartholomew, and X. Zhuang. ISWI remodelers slide nucleosomes with coordinated multi-base-pair entry steps and single-base-pair exit steps. *Cell*, 2013.

[64] T. R. Blosser, J. G. Yang, M. D. Stone, G. J. Narlikar, and X. Zhuang. Dynamics of nucleosome remodelling by individual ACF complexes. *Nature*, 462(7276):1022–1027, 2009.

[65] A. Almer and W. Hörz. Nuclease hypersensitive regions with adjacent positioned nucleosomes mark the gene boundaries of the PHO5/PHO3 locus in yeast. *The EMBO Journal*, 5(10):2681, 1986.

[66] S. Barbarić, K. D. Fascher, and W. Hörz. Activation of the weakly regulated PHO8 promoter in S. cerevisiae: Chromatin transition and binding sites for the positive regulatory protein PHO4. *Nucleic Acids Research*, 20(5):1031–1038, 1992.

[67] W. J. Jessen, A. Dhasarathy, S. A. Hoose, C. D. Carvin, A. L. Risinger, and M. P. Kladde. Mapping chromatin structure in vivo using DNA methyltransferases. *Methods (San Diego, Calif.)*, 33(1):68–80, 2004.

[68] J. A. Kilgore, S. A. Hoose, T. L. Gustafson, W. Porter, and M. P. Kladde. Single-molecule and population probing of chromatin structure using DNA methyltransferases. *Methods (San Diego, Calif.)*, 41(3):320–332, 2007.

[69] E. C. Small, L. Xi, J.-P. Wang, J. Widom, and J. D. Licht. Single-cell nucleosome mapping reveals the molecular basis of gene expression heterogeneity. *Proceedings of the National Academy of Sciences*, 111(24):E2462–2471, 2014.

[70] N. Korolev, O. V. Vorontsova, and L. Nordenskiöld. Physicochemical analysis of electrostatic foundation for DNA–protein interactions in chromatin transformations. *Progress in Biophysics and Molecular Biology*, 95(1):23–49, 2007.

[71] C. Y. Zhou, S. L. Johnson, N. I. Gamarra, and G. J. Narlikar. Mechanisms of ATP-Dependent Chromatin Remodeling Motors. *Annual Review of Biophysics*, 45(1):153–181, 2016.

[72] G. Gargiulo, S. Levy, G. Bucci, M. Romanenghi, L. Fornasari, K. Y. Beeson, S. M. Goldberg, M. Cesaroni, M. Ballarini, F. Santoro, N. Bezman, G. Frigè, P. D. Gregory, M. C. Holmes, R. L. Strausberg, P. G. Pelicci, F. D. Urnov, and S. Minucci. NA-Seq: A Discovery Tool for the Analysis of Chromatin Structure and Dynamics during Differentiation. *Developmental Cell*, 16(3):466–481, 2009.

[73] P. B. Chen, L. J. Zhu, S. J. Hainer, K. N. McCannell, and T. G. Fazzio. Unbiased chromatin accessibility profiling by RED-seq uncovers unique features of nucleosome variants in vivo. *BMC Genomics*, 15:1104, 2014.

[74] T. K. Kelly, Y. Liu, F. D. Lay, G. Liang, B. P. Berman, and P. A. Jones. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Research*, 22(12):2497–2506, 2012.

[75] R. K. Mann and M. Grunstein. Histone H3 N-terminal mutations allow hyperactivation of the yeast GAL1 gene in vivo. *The EMBO Journal*, 11(9):3297–3306, 1992.

[76] B. Celona, A. Weiner, F. Di Felice, F. M. Mancuso, E. Cesarini, R. L. Rossi, L. Gregory, D. Baban, G. Rossetti, P. Grianti, M. Pagani, T. Bonaldi, J. Ragoussis, N. Friedman, G. Camilloni, M. E. Bianchi, and A. Agresti. Substantial histone reduction modulates genomewide nucleosomal occupancy and global transcriptional output. *PLOS Biology*, 9(6):e1001086, 2011.

[77] P.-Y. Chen, S. J. Cokus, and M. Pellegrini. BS Seeker: Precise mapping for bisulfite sequencing. *BMC Bioinformatics*, 11:203, 2010.

[78] W. Guo, P. Fiziev, W. Yan, S. Cokus, X. Sun, M. Q. Zhang, P.-Y. Chen, and M. Pellegrini. BS-Seeker2: A versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics*, 14(1):774, 2013.

[79] K. Y. Y. Huang, Y.-J. Huang, and P.-Y. Chen. BS-Seeker3: Ultrafast pipeline for bisulfite sequencing. *BMC Bioinformatics*, 19(1), 2018.

[80] J. T. Simpson, R. E. Workman, P. C. Zuzarte, M. David, L. J. Dursi, and W. Timp. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, 14(4):407–410, 2017.

[81] N. Krietenstein, C. J. Wippo, C. Lieleg, and P. Korber. Genome-wide in vitro reconstitution of yeast chromatin with in vivo-like nucleosome positioning. *Methods in Enzymology*, 513:205–232, 2012.

[82] R. P. Darst, N. H. Nabilsi, C. E. Pardo, A. Riva, and M. P. Kladde. DNA methyltransferase accessibility protocol for individual templates by deep sequencing. *Methods in Enzymology*, 513:185–204, 2012.

[83] P. T. Lowary and J. Widom. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal of Molecular Biology*, 276(1):19–42, 1998.

[84] C. L. White, R. K. Suto, and K. Luger. Structure of the yeast nucleosome core particle reveals fundamental changes in internucleosome interactions. *The EMBO Journal*, 20(18):5207–5218, 2001.

[85] R. V. Chereji, J. Ocampo, and D. J. Clark. MNase-Sensitive Complexes in Yeast: Nucleosomes and Non-histone Barriers. *Molecular Cell*, 65(3):565–577.e3, 2017.

[86] J. Gutin, R. Sadeh, N. Bodenheimer, D. Joseph-Strauss, A. Klein-Brill, A. Alajem, O. Ram, and N. Friedman. Fine-Resolution Mapping of TF Binding and Chromatin Interactions. *Cell Reports*, 22(10):2797–2807, 2018.

[87] G. Badis, E. T. Chan, H. van Bakel, L. Pena-Castillo, D. Tillo, K. Tsui, C. D. Carlson, A. J. Gossett, M. J. Hasinoff, C. L. Warren, M. Gebbia, S. Talukder, A. Yang, S. Mnaimneh, D. Terterov, D. Coburn, A. Li Yeo, Z. X. Yeo, N. D. Clarke, J. D. Lieb, A. Z. Ansari, C. Nislow, and T. R. Hughes. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Molecular Cell*, 32(6):878–887, 2008.

[88] L. S. Churchman and J. S. Weissman. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, 469(7330):368–373, 2011.

[89] Y. Xu, C. Bernecky, C.-T. Lee, K. C. Maier, B. Schwalb, D. Tegunov, J. M. Plitzko, H. Urlaub, and P. Cramer. Architecture of the RNA polymerase II-Paf1C-TFIIS transcription elongation complex. *Nature Communications*, 8(1):15741, 2017.

[90] S. Kubik, E. O'Duibhir, W. J. de Jonge, S. Mattarocci, B. Albert, J.-L. Falcone, M. J. Bruzzone, F. C. P. Holstege, and D. Shore. Sequence-Directed Action of RSC Remodeler and General Regulatory Factors Modulates +1 Nucleosome Position to Facilitate Transcription. *Molecular Cell*, 71(1):89–102.e5, 2018.

[91] S. Brahma and S. Henikoff. RSC-Associated Subnucleosomes Define MNase-Sensitive Promoters in Yeast. *Molecular Cell*, 73(2):238–249.e3, 2019.

[92] T. J. Parnell, A. Schlichter, B. G. Wilson, and B. R. Cairns. The chromatin remodelers RSC and ISW1 display functional and chromatin-based promoter antagonism. *eLife*, 4:e06073, 2015.

[93] J. Zhao, J. Herrera-Diaz, and D. S. Gross. Domain-Wide Displacement of Histones by Activated Heat Shock Factor Occurs Independently of Swi/Snf and Is Not Correlated with RNA Polymerase II Density. *Molecular and Cellular Biology*, 25(20):8985–8999, 2005.

[94] B. R. Cairns, Y. Lorch, Y. Li, M. Zhang, L. Lacomis, H. Erdjument-Bromage, P. Tempst, J. Du, B. Laurent, and R. D. Kornberg. RSC, an Essential, Abundant Chromatin-Remodeling Complex. *Cell*, 87(7):1249–1260, 1996.

[95] P. D. Hartley and H. D. Madhani. Mechanisms that specify promoter nucleosome location and identity. *Cell*, 137(3):445–458, 2009.

[96] K. Yen, V. Vinayachandran, K. Batta, R. T. Koerber, and B. F. Pugh. Genome-wide nucleosome specificity and directionality of chromatin remodelers. *Cell*, 149(7):1461–1473, 2012.

[97] F. Kraft and I. Kurth. Long-read sequencing in human genetics. *medizinische genetik*, 31(2):198–204, 2019.

[98] D. Deamer, M. Akeson, and D. Branton. Three decades of nanopore sequencing. *Nature Biotechnology*, 34(5):518–524, 2016.

[99] N. Kono and K. Arakawa. Nanopore sequencing: Review of potential applications in functional genomics. *Development, Growth & Differentiation*, 61(5):316–326, 2019.

[100] V. Boža, B. Brejová, and T. Vinař. DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. *PLOS ONE*, 12(6):e0178751, 2017.

[101] H. Teng, M. D. Cao, M. B. Hall, T. Duarte, S. Wang, and L. J. M. Coin. Chiron: Translating nanopore raw signal directly into nucleotide sequence using deep learning. *GigaScience*, 7(5), 2018.

[102] F. J. Rang, W. P. Kloosterman, and J. de Ridder. From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, 19(1):90, 2018.

[103] N. J. Loman, J. Quick, and J. T. Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, 12(8):733–735, 2015.

[104] A. C. Rand, M. Jain, J. M. Eizenga, A. Musselman-Brown, H. E. Olsen, M. Akeson, and B. Paten. Mapping DNA methylation with high-throughput nanopore sequencing. *Nature Methods*, 14(4):411–413, 2017.

[105] Y. Wang, A. Wang, Z. Liu, A. L. Thurman, L. S. Powers, M. Zou, Y. Zhao, A. Hefel, Y. Li, J. Zabner, and K. F. Au. Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Research*, 29(8):1329–1342, 2019.

[106] Q. Liu, L. Fang, G. Yu, D. Wang, C.-L. Xiao, and K. Wang. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nature Communications*, 10(1), 2019.

[107] Z. Shipony, G. K. Marinov, M. P. Swaffer, N. A. Sinnott-Armstrong, J. M. Skotheim, A. Kundaje, and W. J. Greenleaf. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nature Methods*, 17(3):319–327, 2020.

[108] A. B. Stergachis, B. M. Debo, E. Haugen, L. S. Churchman, and J. A. Stamatoyannopoulos. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science*, 368(6498):1449–1454, 2020.

[109] I. Lee, R. Razaghi, T. Gilpatrick, N. Sadowski, F. Sedlazeck, and W. Timp. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *bioRxiv*, 2018.

[110] U. J. Schermer, P. Korber, and W. Hörz. Histones are incorporated in trans during reassembly of the yeast PHO5 promoter. *Molecular Cell*, 19(2):279–285, 2005.

[111] L. R. Racki, J. G. Yang, N. Naber, P. D. Partensky, A. Acevedo, T. J. Purcell, R. Cooke, Y. Cheng, and G. J. Narlikar. The chromatin remodeller ACF acts as a dimeric motor to space nucleosomes. *Nature*, 462(7276):1016–21, 2009.

[112] M. R. Wolff, A. Schmid, P. Korber, and U. Gerland. Effective Dynamics of Nucleosome Configurations at the Yeast *PHO5* Promoter. *bioRxiv*, 2020.

[113] E. M. O'Neill, A. Kaffman, E. R. Jolly, and E. K. O'Shea. Regulation of PHO4 Nuclear Localization by the PHO80-PHO85 Cyclin-CDK Complex. *Science*, 271(5246):209–212, 1996.

[114] A. Komeili. Roles of Phosphorylation Sites in Regulating Activity of the Transcription Factor Pho4. *Science*, 284(5416):977–980, 1999.

[115] S. Musladin, N. Krietenstein, P. Korber, and S. Barbaric. The RSC chromatin remodeling complex has a crucial role in the complete remodeler set for yeast PHO5 promoter opening. *Nucleic Acids Research*, 42(7):4270–4282, 2014.

[116] K.-D. Fascher, J. Schmitz, and W. Hörz. Structural and functional requirements for the chromatin transition at the PHO5 promoter in saccharomyces cerevisiae upon PHO5 activation. *Journal of Molecular Biology*, 231(3):658–667, 1993.

[117] C. R. Brown and H. Boeger. Nucleosomal promoter variation generates gene expression noise. *Proceedings of the National Academy of Sciences*, 111(50):17893–17898, 2014.

[118] H. Kharerin, P. J. Bhat, J. F. Marko, and R. Padinhateeri. Role of transcription factor-mediated nucleosome disassembly in PHO5 gene expression. *Scientific Reports*, 6:20319, 2016.

[119] T. Blasi, C. Feller, J. Feigelman, J. Hasenauer, A. Imhof, F. J. Theis, P. B. Becker, and C. Marr. Combinatorial Histone Acetylation Patterns Are Generated by Motif-Specific Reactions. *Cell Systems*, 2(1):49–58, 2016.

[120] T. J. Sheskin. Technical Note —A Markov Chain Partitioning Algorithm for Computing Steady State Probabilities. *Operations Research*, 33(1):228–235, 1985.

[121] W. K. Grassmann, M. I. Taksar, and D. P. Heyman. Regenerative analysis and steady state distributions for markov chains. *Operations Research*, 33(5):1107–1116, 1985.

[122] D. N. Shanbhag and C. R. Rao. *Stochastic Processes: Modelling and Simulation*. Number 21 in Handbook of Statistics. Elsevier, 2003.

[123] S. C. Satchwell, H. R. Drew, and A. A. Travers. Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology*, 191(4):659–675, 1986.

[124] P. D. Gregory, S. Barbaric, and W. Hörz. Restriction nucleases as probes for chromatin structure. *Methods in Molecular Biology (Clifton, N.J.)*, 119:417–425, 1999.

[125] F. Ertel, A. B. Dirac-Svejstrup, C. B. Hertel, D. Blaschke, J. Q. Svejstrup, and P. Korber. In vitro reconstitution of PHO5 promoter chromatin remodeling points to a role for activator-nucleosome competition in vivo. *Molecular and Cellular Biology*, 30(16):4060–4076, 2010.

[126] Y. Zhang, C. L. Smith, A. Saha, S. W. Grill, S. Mihardja, S. B. Smith, B. R. Cairns, C. L. Peterson, and C. Bustamante. DNA Translocation and Loop Formation Mechanism of Chromatin Remodeling by SWI/SNF and RSC. *Molecular Cell*, 24(4):559–568, 2006.

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

[128] S. Barbaric, J. Walker, A. Schmid, J. Q. Svejstrup, and W. Hörz. Increasing the rate of chromatin remodeling and gene activation–a novel role for the histone acetyltransferase Gcn5. *The EMBO Journal*, 20(17):4944–4951, 2001.

[129] A. Schmid, K. D. Fascher, and W. Hörz. Nucleosome disruption at the yeast PHO5 promoter upon PHO5 induction occurs in the absence of DNA replication. *Cell*, 71(5):853–864, 1992.

[130] P. Korber, S. Barbaric, T. Luckenbach, A. Schmid, U. J. Schermer, D. Blaschke, and W. Hörz. The histone chaperone Asf1 increases the rate of histone eviction at the yeast PHO5 and PHO8 promoters. *The Journal of Biological Chemistry*, 281(9):5539–5545, 2006.

[131] S. Barbaric, T. Luckenbach, A. Schmid, D. Blaschke, W. Hörz, and P. Korber. Redundancy of chromatin remodeling pathways for the induction of the yeast PHO5 promoter in vivo. *The Journal of Biological Chemistry*, 282(38):27610–27621, 2007.

[132] A. Almer, H. Rudolph, A. Hinnen, and W. Hörz. Removal of positioned nucleosomes from the yeast PHO5 promoter upon PHO5 induction releases additional upstream activating DNA elements. *The EMBO Journal*, 5(10):2689–2696, 1986.

[133] C. J. Wippo, B. S. Krstulovic, F. Ertel, S. Musladin, D. Blaschke, S. Stürzl, G.-C. Yuan, W. Hörz, P. Korber, and S. Barbaric. Differential Cofactor Requirements for Histone Eviction from Two Nucleosomes at the Yeast PHO84 Promoter Are Determined by Intrinsic Nucleosome Stability. *Molecular and Cellular Biology*, 29(11):2960–2981, 2009.

[134] J. Winger and G. D. Bowman. The Sequence of Nucleosomal DNA Modulates Sliding by the Chd1 Chromatin Remodeler. *Journal of Molecular Biology*, 429(6):808–822, 2017.

[135] M. P. Cosma, T. Tanaka, and K. Nasmyth. Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. *Cell*, 97(3):299–311, 1999.

[136] B. J. Deroo and T. K. Archer. Glucocorticoid receptor-mediated chromatin remodeling in vivo. *Oncogene*, 20(24):3039–3046, 2001.

[137] T. A. Johnson, C. Elbi, B. S. Parekh, G. L. Hager, and S. John. Chromatin remodeling complexes interact dynamically with a glucocorticoid receptor-regulated promoter. *Molecular Biology of the Cell*, 19(8):3308–3322, 2008.

[138] A. H. Hassan, K. E. Neely, and J. L. Workman. Histone acetyltransferase complexes stabilize swi/snf binding to promoter nucleosomes. *Cell*, 104(6):817–827, 2001.

[139] P. Sudarsanam and F. Winston. The Swi/Snf family nucleosome-remodeling complexes and transcriptional control. *Trends in Genetics*, 16(8):345–351, 2000.

[140] M. Vignali, A. H. Hassan, K. E. Neely, and J. L. Workman. ATP-dependent chromatin-remodeling complexes. *Molecular and Cellular Biology*, 20(6):1899–1910, 2000.

[141] J. L. Workman. Nucleosome displacement in transcription. *Genes & Development*, 20(15):2009–2017, 2006.

[142] P. B. Becker and W. Hörz. ATP-dependent nucleosome remodeling. *Annual Review of Biochemistry*, 71:247–273, 2002.

[143] G. J. Narlikar, H.-Y. Fan, and R. E. Kingston. Cooperation between complexes that regulate chromatin structure and transcription. *Cell*, 108(4):475–487, 2002.

[144] K. E. Neely, A. H. Hassan, A. E. Wallberg, D. J. Steger, B. R. Cairns, A. P. Wright, and J. L. Workman. Activation domain-mediated targeting of the SWI/SNF complex to promoters stimulates transcription from nucleosome arrays. *Molecular Cell*, 4(4):649–655, 1999.

[145] N. Yudkovsky, C. Logie, S. Hahn, and C. L. Peterson. Recruitment of the SWI/SNF chromatin remodeling complex by transcriptional activators. *Genes & Development*, 13(18):2369–2374, 1999.

[146] K. Natarajan, B. M. Jackson, H. Zhou, F. Winston, and A. G. Hinnebusch. Transcriptional activation by Gcn4p involves independent interactions with the SWI/SNF complex and the SRB/mediator. *Molecular Cell*, 4(4):657–664, 1999.

[147] E. Kowenz-Leutz and A. Leutz. A C/EBP beta isoform recruits the SWI/SNF complex to activate myeloid genes. *Molecular Cell*, 4(5):735–743, 1999.

[148] S. Barbaric, H. Reinke, and W. Hörz. Multiple mechanistically distinct functions of SAGA at the PHO5 promoter. *Molecular and Cellular Biology*, 23(10):3468–3476, 2003.

[149] A. Dhasarathy and M. P. Kladde. Promoter occupancy is a major determinant of chromatin remodeling enzyme requirements. *Molecular and Cellular Biology*, 25(7):2698–2707, 2005.

[150] T. Tsukiyama, P. B. Becker, and C. Wu. ATP-dependent nucleosome disruption at a heat-shock promoter mediated by binding of GAGA transcription factor. *Nature*, 367(6463):525–532, 1994.

[151] Y. Lorch, M. Zhang, and R. D. Kornberg. Histone octamer transfer by a chromatin-remodeling complex. *Cell*, 96(3):389–392, 1999.

[152] G. Längst, E. J. Bonte, D. F. Corona, and P. B. Becker. Nucleosome movement by CHRAC and ISWI without disruption or trans-displacement of the histone octamer. *Cell*, 97(7):843–852, 1999.

[153] A. Hamiche, R. Sandaltzopoulos, D. A. Gdula, and C. Wu. ATP-dependent histone octamer sliding mediated by the chromatin remodeling complex NURF. *Cell*, 97(7):833–842, 1999.

[154] K. A. Haushalter and J. T. Kadonaga. Chromatin assembly by DNA-translocating motors. *Nature Reviews. Molecular Cell Biology*, 4(8):613–620, 2003.

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

[156] R. H. Morse. Nucleosome disruption by transcription factor binding in yeast. *Science*, 262(5139):1563–1566, 1993.

[157] C. Yan, H. Chen, and L. Bai. Systematic Study of Nucleosome-Displacing Factors in Budding Yeast. *Molecular Cell*, 71(2):294–305.e4, 2018.

[158] B. T. Donovan, H. Chen, C. Jipa, L. Bai, and M. G. Poirier. Dissociation rate compensation mechanism for budding yeast pioneer transcription factors. *eLife*, 8, 2019.

[159] M. Iwafuchi, I. Cuesta, G. Donahue, N. Takenaka, A. B. Osipovich, M. A. Magnuson, H. Roder, S. H. Seeholzer, P. Santisteban, and K. S. Zaret. Gene network transitions in embryos depend upon interactions between a pioneer transcription factor and core histones. *Nature Genetics*, 52(4):418–427, 2020.

[160] K. D. Fascher, J. Schmitz, and W. Hörz. Role of trans-activating proteins in the generation of active chromatin at the PHO5 promoter in S. cerevisiae. *The EMBO Journal*, 9(8):2523–2528, 1990.

[161] C. Costigan, D. Kolodrubetz, and M. Snyder. NHP6A and NHP6B, which encode HMG1-like proteins, are candidates for downstream components of the yeast SLT2 mitogen-activated protein kinase pathway. *Molecular and Cellular Biology*, 14(4):2391–2403, 1994.

[162] S. Giavara, E. Kosmidou, M. P. Hande, M. E. Bianchi, A. Morgan, F. d'Adda di Fagagna, and S. P. Jackson. Yeast Nhp6A/B and mammalian Hmgb1 facilitate the maintenance of genome stability. *Current Biology*, 15(1):68–72, 2005.

[163] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[164] H. Li. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[165] J. K. Percus. Entropy of a non-uniform one-dimensional fluid. *Journal of Physics: Condensed Matter*, 1(17):2911–2922, 1989.

[166] J. K. Percus. Equilibrium state of a classical fluid of hard rods in an external field. *Journal of Statistical Physics*, 15(6):505–511, 1976.

[167] J. K. Percus. One-dimensional classical fluid with nearest-neighbor interaction in arbitrary external field. *Journal of Statistical Physics*, 28(1):67–81, 1982.

[168] B. Osberg. *One-Dimensional Lattice Gasses With Soft Interaction.* PhD thesis, Ludwig-Maximilians-Universität München, 2015.

[169] H. Hansen-Goos, C. Lutz, C. Bechinger, and R. Roth. From pair correlations to pair interactions: An exact relation in one-dimensional systems. *Europhysics Letters (EPL)*, 74(April):8–14, 2007.

[170] G. Moyle-Heyrman, T. Zaichuk, L. Xi, Q. Zhang, O. C. Uhlenbeck, R. Holmgren, J. Widom, and J.-P. Wang. Chemical map of Schizosaccharomyces pombe reveals species-specific features in nucleosome positioning. *Proceedings of the National Academy of Sciences*, 110(50):20158–20163, 2013.

# Acknowledgments

When I started my PhD, I did not foresee the path that it would finally take me. As I realized, research can be unpredictable and frustrating at times, which makes it even more rewarding when projects come to a successful end. I am very glad that my PhD took this path, because it lead me to working with wonderful people and getting to know not only them, but also myself better. This work would not have been possible without the help and support of many colleagues, friends and my family, who motivated and encouraged me.

Especially, I like to thank my supervisor Ulrich Gerland for the opportunity to pursue a PhD in this group, for his guidance and patient support, all the discussions and the freedom needed to become a more and more independent researcher.

Special thanks goes to Elisa Oberbeckmann and Philipp Korber at the Biomedical Center in Munich (BMC), who I really enjoyed working with on the absolute occupancy project for more than three years. Elisa was performing most of the experiments, glad to help anytime with biochemistry questions and always a cheerful discussion partner. I also like to thank Mark Heron for introducing me to the analysis of Elisa's data at the time I joined the project.

I am also very happy to be working with Daan Verhagen and Felix Müller-Planitz, together with Elisa and Philipp, from the BMC as well as Maryam Kathami and Matthias Hanke from the Gerland group in our project on nucleosome detection using nanopore data. Furthermore, I like to thank Johannes Nübler for the support at the beginning of my PhD and all other former and current members of the Gerland group for the discussions on various topics, research as well as non-research related, and the supporting and friendly group atmosphere.

I especially want to thank Nanni, whose door was always open for a friendly chat and for, together with Linda, taking care of the coffee machine even though only rarely drinking coffee himself, Mareike and Tobi for organizing amazing winter group retreats at the Albert-Link-Hütte in Spitzingsee, Stephan for his lessons in cross country skating there, Mareike, my cheerful office colleague, for all the backcountry skiing and mountainbiking tips and trips together, Bernhard for the very helpful discussions and Elena for sharing the ups and downs during our PhD time.

I am very grateful for the support by the graduate school of Quantitative Biosciences Munich (QBM), their lectures on biochemistry, all other courses and yearly research retreats with all QBM members as well as for the feedback of my QBM thesis advisory committee consisting of Ulrich Gerland, Erwin Frey, Julian Gagneur and Philipp Korber. Philipp's help, support and knowledge was indispensable for the projects in this thesis.

Finally, I like to thank my amazing friends, among others, Christian, Markus, Line, Sarah and especially my lovely girlfriend Ming and my family for all their support and love.