Analysing and quantitatively modelling nucleosome binding preferences

Mark Eric Leslie Heron

München 2017

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

Analysing and quantitatively modelling nucleosome binding preferences

Mark Eric Leslie Heron
aus Schwäbisch Hall, Deutschland

2017

Erklärung:

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

Eidesstattliche Versicherung:

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 20.05.2017

Mark Heron

Dissertation eingereicht am 13.02.2017

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

Mündliche Prüfung am 10.05.2017

Summary

The main emphasis of my work as a PhD student was the analysis and prediction of nucleosome positioning, focusing on the role sequence features play.

Part I gives a broad overview of nucleosomes, before defining important technical terms. It continues by describing and reviewing experiments that measure nucleosome positioning and bioinformatic methods that learn the sequence preferences of nucleosomes to predict their positioning.

Part II describes a collaboration project with the Gaul-lab, where I analyzed MNase-Seq measurements of nucleosomes in *Drosophila*. The original intention was to investigate the extent to which experimental biases influence the measurements. We extended the analysis to categorize and explore fragile, average and resistant nucleosome populations. I focused on the relation between nucleosome fragility and the sequence landscape, especially at promoters and enhancers. Analyzing the partial unwrapping of nucleosomes genome-wide, I found that the G+C ratio is a determinant of asymmetric unwrapping. I excluded an analysis of histone modifications from this work, which was part of this collaboration, due to its low relevance to the rest of the presented work.

Part III describes my main project of developing a probabilistic nucleosome-position prediction method. I developed a maximum likelihood approach to learn a biophysical model of nucleosome binding. By including the low positional resolution of MNase-Seq and the sequence bias of CC-Seq into the likelihood, I could separate them from the nucleosome binding preferences and learn highly correlated nucleosome binding energy models. My analysis shows that nucleosomes have a position-specific binding preference and might be uninfluenced by G+C content or even disfavor it – contrary to the consensus in literature.

Part IV describes further analysis I did during my time as a PhD student that are not part of any planned publications. The main topics are: ancillary elements of my main project, unsuccessful attempts to correct experimental biases, analysis of the quality of experimental measurements, and adapting my probabilistic nucleosome-position prediction method to work with occupancy measurements. Lastly, I give a general outlook that reflects on my results and discusses next steps, like ways to improve my method further.

I excluded two collaboration projects I participated in from this thesis, because they are still ongoing: a systematic analysis of how the core promoter sequence influences gene expression in *Drosophila* and the development of an experiment to measure nucleosome occupancy more precisely.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my PhD adviser Johannes Söding, from whom I learned a lot. He helped refine my critical thinking and deepen my understanding of many subjects – foremost probabilistic modeling, by scrutinizing over details he was unsure about and taking time to explain the parts in-depth I was having a hard time grasping. I feel honored that he felt confident in giving me leeway and often let me work more independently. I enjoyed my time as a PhD student in his lab and am grateful he gave me the opportunity to go on this journey. I would also like to thank the other members of my PhD advisory committee: Julien Gagneur, Ulrike Gaul, Philipp Korber, Klaus Förstemann and Karl-Peter Hopfner. In this context, my gratitude goes to all the people who helped me over the last years in big or small ways, in work or personal matters, with bureaucracy or advice.

I have to thank the whole Söding-lab without whom the last years are hard to imagine for me. I count several, especially those who stayed with me in Munich, to my friends and regret not having been able to spend more time with those in Göttingen. I enjoyed all the scientific discussions and personal conversations I had over the years, be it at lunch, conferences or group meetings. First and foremost I would like to thank Anja Kiesel, with whom I relish having worked closely since starting in the lab with my Master thesis. We shared many ups and downs over the years. More than once Anja helped me out of bad moods without knowing, even if I never got that back massage. In no particular order I would also like to thank: Susann Vorberg for jumping out of an airplane with me and continuously reminding me that bioinformatic topics beyond 'transcription' exist; Matthias Siebert for sharing advice like a mentor, without being one; Phillipp Torkler for buying Tablesoccer balls; Markus Meier for driving me around Ireland without crashing; Stefan Seemayer for relieving me of being the geekiest nerd in the group; Jessica Andreani Feuillet for daring me to use the car-parking analogy in my thesis; Holger Hartmann for briefly being my tutor before I started my PhD thesis and asking me every year when I would finally finish; and lastly Verena Friedl and Andrea Kreppel for indulging me as

their tutor. This list is far from complete, but the roles of the other lab members as extras during my time as a PhD student have to remain uncredited for a lack of witty comments I can come up with. I have many fond memories with all of you and hope our paths will cross again in the future!

I would like to thank the Gaul-lab. My ongoing collaborations with them were a substantial part of my work as a PhD student. I have to thank Ulrike Gaul and Ulrich Unnerstall for appreciating my opinion in our discussions, furthering my education by approving every conference I wanted to visit and managing the SysCore project, which funded my work (and in that regards I have to thank the BMBF for financing me via this project). I would also like to thank the other lab members with whom I worked together, mostly on the SysCore project, and those with whom I never worked, but got to know better. I especially have to thank Alessio Renna for the extensive close collaboration and letting my facial hair look tame in comparison.

I would like to thank my collaboration partners in the Korber-lab. I am grateful that Philipp Korber sought me out as a collaboration partner. Our meetings always felt fruitful: I gladly exchanged the newest concerns and their possible solutions. Nils Krietenstein and Elisa Oberbeckmann I would like to thank for indulging my naive questions about minute experimental details and digging deeper into my analyses with the right questions.

I would like to thank the Tresh-lab and the Gagneur-lab for sharing the open office space with our lab. They helped create a great atmosphere that encouraged scientific discussions and work, while keeping plenty of room for personal connections and recreation. The Gagneur-lab alleviated the move of the core Söding-lab to Göttingen for the Munich branch and made us continuously feel welcome. So much that I now count several of them to my friends.

Speaking of which, I would like to thank my friends. They encouraged me, entertained me, and probably saw this day coming before me. Above all I want to thank Matthias Schwienbacher for sharing drinks with me every Friday, before watching a sneak preview at the CINEMA.

Last but not least, I have to thank my family. They were always interested in my work and progress – asking for new developments, but also understanding when I preferred other topics. I have to thank my brothers, Sean and Alex, for our time growing up together, our shared interest in science was a catalyst that sent me on this path. Finally, I have to thank my parents for their unconditional love, unending support and always believing in me.

Part I

Background

1. Introduction

This introduction is aimed at scientists and layman who are lacking background knowledge that is important to understand the rest of the work. I hope this broad introduction with analogies will help them grasp the core concepts. With this intent in mind, I decided to use citations sparingly to improve the readability. Most information is also either basic knowledge that needs no citation, or reiterated in another introductions with a citation.

1.1 What is the genome and DNA?

The genome is the information storage of cells. To store the information four different building blocks are linked to a long chain called DNA. The building blocks are called nucleotides or bases and the four types are abbreviated to A, C, G and T. Genomic DNA consists of two strands i.e. chains (double stranded DNA), which run anti-parallel – one strand runs forward and the other backwards. Together the two strands form a double helix. A double helix consists of two intertwined helices that wrap around each other. An important aspect of the two DNA strands is that the building blocks of either strand complement each other at every position: if one strand has an A the other has a T or vise versa, and equivalently C and G pair.

Cells of higher organisms contain a surprising amount of DNA. The DNA from a single human cell would span about two meters, if it were extracted and spread out. How are strings of such length organized to fit in a cell with a diameter of less than 1 mm? Thinking of the common ways to store yarn: clews and spools, their organization gives access to the ends, but most central regions are inaccessible. In living cells the genomic information is not accessed linearly from the start – like a novel is read – instead the whole genome is accessed more randomly – like a cooking book. The packing has to accommodate this fact in comparison to yarn storage, but the basic idea of looping string is consistent between the two. On the small scale – which this work is about – the DNA wraps roughly 1.7 times around eight histone proteins, which effectively function

as a small bead. Together the eight histones and the wrapped DNA form the structure called nucleosome. On the large scale loops are also formed with the help of proteins, but the proteins do not act as beads. These proteins either stabilize structures between neighboring nucleosomes or clamp the ends of large loops together.

1.2 What are nucleosomes?

Nucleosomes compact the genome and restrict access to the information contained in the wrapped DNA. They are very small in comparison to the whole genome. A nucleosome contains 147 DNA base pairs (bps), while the genome is typically on a scale of mega (10^6) to giga (10^9) bps. This allows nucleosomes to reveal and cover information in a granular fashion, similar to opening a cookbook on a page for a recipe, compared to needing to unfold a whole map to look at a small region. Because many genomic regions are open at any given time, a more precise analogy would be if the pages of the cookbook were glued in the form of a harmonica so cooks could read multiple recipes simultaneously without coming in each others way, while keeping the unused sections folded up.

The focus of this work is understanding and predicting the positioning of nucleosomes along the genomic DNA based on the nucleotide sequence. The four nucleotides are distinct molecules and influence the physical property of DNA locally. The strongest effect on physical properties stems from the pairs of neighboring nucleotides, called dinucleotides. The DNA double helix is ~2 nm wide and a nucleosome (which includes the bound DNA) has a ~11 nm diameter, therefore the DNA has to bend strongly when forming a nucleosome. The double helix rotates fully once every ~10 bp and has a smaller 'minor grove', and wider 'major grove'. Together, this means that a DNA sequence preferentially forms nucleosomes if it switches between favoring expanding or shortening the minor and major groves every 5 bps. On average, this leads to a ~10-bp-periodic enrichment and depletion cycle of dinucleotides along nucleosome bound DNA.

1.3 Why is the positioning of nucleosomes interesting?

Nucleosome positioning controls the accessibility of genomic information. While I primarily discuss the sequence preference of nucleosomes and how this leads to positioning, from a biological standpoint the location of nucleosome-depleted regions is more interesting. Nucleosome-depleted regions are created intrinsically by the DNA sequence and

extrinsically by outside influences, like growth medium in single cell organisms or tissue type in multicellular organisms. Even intrinsic nucleosome-depleted regions only occur in some cells at any given time point due to statistical fluctuations. Consistent access to information needed for fundamental processed can be provided by intrinsic nucleosome depleted regions encoded in the DNA sequence. For example all cells need housekeeping proteins like histones to function and, therefore, the accessibility of their information can be encoded intrinsically. Other information is revealed selectively allowing cells to differ greatly, even though their genomes are identical. By revealing different parts of the genome in every cell type, the same genome can create brain, muscle, skin, etc. tissues.

To understand further details a basic knowledge of the information organization is required. In the genome the primary information units are genes. If the genome were a cookbook, then a gene would be a recipe. The cookbook contains seeming gibberish, dinner plans and meal suggestions between the recipes – I will get to those later. Genomes of higher organisms have more between recipes than genomes of lower organisms. A gene, like a recipe, consists of different sections, for this work the important distinction is between the promoter and the rest of the gene. The promoter is the genes start where the decision is made to transcribe the gene i.e. cook the recipe. The rest of the gene describes how to make the protein. The promoter is like the title or description of the resulting meal, while the rest is like the instructions for preparing the meal.

Averaging over all promoters in yeast a nucleosome depleted region and a DNA sequence signal are visible. In yeast, most promoters belong to housekeeping genes, which have a consistent transcription rate. Intuitively, using static information in the DNA sequence to help keep promoters of housekeeping genes constantly accessible makes sense. While this hypothesis existed for over a decade it was difficult to prove, because high correlations between such features do not imply a causal effect. Recent experiments showed that manipulating nucleosome unfavorable sequences in artificial promoters can increase nucleosome occupancy and decrease gene expression and vice versa, which validates the hypothesis (Raveh-Sadka et al., 2012).

While my work focuses on the intrinsic sequence-dependent nucleosome-positioning signal, extrinsic processes are important in most interesting genomic regions. For example, extrinsic factors dominate at promoters of regulated genes and enhancers. Enhancers regulate the transcription of genes together with promoters, but their location is distinct from the gene. In the cookbook analogy enhancers would be dinner plans or suggestions for combined meals.

1.4 What positions nucleosomes?

Intrinsic and extrinsic nucleosome positioning as described above is difficult to separate in praxis. I will distinguish between sequence preferences of nucleosomes (primarily due to intrinsic factors), steric hindrance (part of both), and other influences (primarily extrinsic factors). To illustrate different aspects I will describe nucleosomes binding DNA with an analogy of cars parked along a sidewalk. Parked cars are a common model for one dimensional gases – a physical model that can describe nucleosomes forming on a stretch of DNA.

The sequence preference of a nucleosome is equivalent to the preference to park cars in front of a shop entrances or home. Steric hindrance matches drivers avoiding collisions while parking their car. Just as only one car can park along a stretch of side walk, only one nucleosome can bind a segment of DNA. Drivers often park their car right next to another parked car, while nucleosomes have no such inclination by themselves, a high density and nucleosome remodelers help achieve this. Preferred spots in combination with a high density of cars will lead to a similar parking pattern occurring on most days. The same happens with nucleosomes around strongly un-/favored DNA sequences.

I will add to the car parking analogy to describe some extrinsic influences. Transcription factors are proteins that bind DNA and regulate gene transcription. Competition between nucleosomes and transcription factors is equivalent to motorbikes parking along the sidewalk and coming in the way of cars trying to park. Transcription factors are smaller than nucleosomes and have a more specific sequence preference, called a motif. Depending on which and in what frequency transcription factors are present in a cell, they compete with nucleosomes at different genomic locations. Most transcription factors have no chance at competing with nucleosomes on their own, which is good – genomic information is inaccessible because transcription factors cannot bind the DNA, amongst other things. Enhancers and promoters generally have several transcription factor binding sites, which collaborate to compete with nucleosomes. Nucleosome remodelers can further aid the transcription factors by moving nucleosomes (remodeling the nucleosome landscape). Remodelers are like tow trucks that remove illegally parked cars (nucleosome eviction) or friendly neighbors that push cars around to free driveways and reduce the distance between parked cars (nucleosome sliding). Often other factors recruit nucleosome remodelers, which then organize the surrounding nucleosomes.

While all these influences are known, an open question is how important their respective roles are in living cells (*in vivo*). Many analysis revolve around the role sequence

preference plays, but reviews agree that its importance is still unknown and depends on the aspects analyzed and definitions used (Kaplan et al., 2010a; Iyer, 2012; Struhl and Segal, 2013) The main arguments for the influence of DNA sequence on nucleosome positioning are: sequence signatures of *in vivo* measured nucleosomes, change of nucleosome occupancy due to sequence modification, the predictability of *in vivo* nucleosomes by sequence based models, and good correlations between nucleosomes measured *in vivo* and *in vitro* ('in glass' i.e. experiments outside of their normal context, in this case genomic DNA purified from cells). The last two points also inform the maximal intrinsic influence the sequence can have and neither fully explain nucleosome positioning *in vivo*. The influences of other factors have been analyzed as well: the average nucleosome distance – which is different between species – depends on the cellular context not the genomic DNA sequence (McManus et al., 1994), and adding whole cell extract and ATP improves the similarity of *in vitro* experiments to *in vivo* due to remodeler activity (Zhang et al., 2011).

As the focus of my work, analysis and discussions of the sequence preference of nucleosomes will appear throughout this thesis. A full discussion of all aspects of nucleosome binding is beyond the scope of this work and I, therefore, refer to existing reviews that cover most (Iyer, 2012; Struhl and Segal, 2013).

1.5 What DNA sequences do nucleosomes prefer?

DNA has to bend strongly to wrap around the histone octamere and form a nucleosome. Therefore, they prefer sequence features that favor such bendability. The two main features that influence the bendability are G+C content and dinucleotide composition. Nucleosomes tend to favor G+C-rich sequences, e.g. the genomic nucleosome occupancy has a high correlation with genomic G+C content. Therefore, higher G+C content appears to correlate with the bendability needed to wrap histones. The DNA double helix structure leads to a ~10-bp-periodic enrichment and depletion of dinucleotides that are preferentially bent in one (WW – W is either A or T) or the other (SS – S is either C or G) direction. Poly(dA:dT) and poly(dC:dG) stretches form nucleosome-depleted regions. The main reason such stretches disfavor nucleosome formation is that they break the preferred dinucleotide periodicity.

1.6 How is nucleosome positioning predicted?

Methods can focus on different aspects when predicting nucleosome positioning. Some methods are developed with the intent of maximizing benchmarking scores. These are typically machine learning algorithms and they bring little insight into the underlying biochemistry. Other methods focus on recreating specific observations, such as the formation of nucleosome arrays over the gene body. The last group of methods approximates the biochemistry of genomic nucleosome positioning with a thermodynamic model. The method I developed and present in Part III belongs to the last group.

The thermodynamic models generally consist of two parts: the binding energy of nucleosomes to different sequences – a quantification of the sequence preference; and the interactions between nucleosomes, transcription factors, and remodelers. The first method with such a model computed the binding energies based on a probabilistic dinucleotide model and only contained the steric hindrance between nucleosomes. Newer methods have replaced either or both parts, usually to approximate the biochemistry more precisely. I went back to the original model and added a third part: measuring the data with experiments. The previous models assumed the measured nucleosome positions reflected the actual nucleosome positioning frequencies. My model contains an intermediary step that can describe uncertainties and biases of the measurements that originate from the experiment protocol.

1.7 Why is my work relevant?

Experiments are imperfect. There is a difference between what you would like to measure and what you actually measure. When measuring nucleosome positions their are two main issues: positional uncertainty and biased frequencies. The positional uncertainty is the deviation between the measured and the actual nucleosome positions. The frequencies are biased if they – the chance of recovering nucleosome positions – systematically deviate from the probabilities of nucleosomes being at those positions.

For the most common experiment to measure nucleosome positioning (MNase-Seq) bias sources are known and have been analyzed. A major bias source depends on the accessibility and fragility of the nucleosomes. In a way, this bias is a biochemical signal that is typically uninteresting and interferes with the signal of interest. With minor adaptations to the experiment, this signal can be extracted and analyzed. I investigated a set of experimental measurements in regards to possible bias sources and the fragility of

nucleosomes. This gained new insights into the role of sequence in nucleosomes fragility and partial unwrapping.

The experimental errors impact models of nucleosome positioning derived from them. The 10-bp-periodic sequence preference of nucleosomes described above is as smooth as a sin curve. I show that this is a result of the positional uncertainty of the experiment and not the true preference of the nucleosomes. The 10-bp-periodic preference was derived from experimental data with about a ± 5 bp positional uncertainty. If unaccounted for, such positional uncertainty smears the preferences out. The smoothness of the previous preference models is therefore expected from the positional resolution and processing of the experimental data.

The novel feature of my nucleosome-position prediction method is modeling experimental errors such as the positional uncertainty. In this way I could learn a high-resolution model of nucleosome preferences from low-resolution data. I also separated out a sequence bias from a dataset measured with another experimental method (CC-Seq), whose issues are distinct of common method. The two models accounting for the experimental errors are more similar to each other than models that ignore the possible errors. The models have similar high-resolution features, but disagree on the preference of G+C-rich sequence and the importance of the sequence on the binding model.

2. Definitions of important terms

Some of the discussions in this work revolve around the precise definition of fundamental terms. The misuse of precisely defined terms occurs too frequently in publications that regard nucleosomes. The meaning of less-precisely defined terms can depend on the context and often creates misleading or wrong statements when cited uncritically. This section first explains the difference between the two types of error – noise and bias. The difference is important to know when discussing limitations of experimental measurements. It then describes the common definition of nucleosome occupancy and positioning, which I use throughout this work.

Noise

In statistics noise describes unreproducible error. The discrepancy between independent measurements of the same thing is noise. If you have a hundred people measure the height of one person the errors between the measurements are noise.

In experiments with a sequencing step, like MNase-Seq, the library preparation and sequencing steps sample a fraction of the DNA fragment population. Every sampling is different and produces noise in the resulting datasets, which is aptly called sampling noise. Another example is positional uncertainty: the distance between the fragment centers measured with MNase-Seq and nucleosome dyads behaves on average like noise. These discrepancies originate from MNase not digesting all the way up to the nucleosome ends. For individual nucleosomes, the distance between the MNase cut sites and the nucleosome ends depends on the local sequence and is partially a bias.

Bias

In statistics bias describes reproducible error. A systematical discrepancy between the measurements and the truth is a bias. If the hundred people use a faulty measuring stick that has a too small scale, the measurements will consistently overestimate the person's height. This overestimation is a bias.

In MNase-Seq, the over digestion of nucleosomes at specific locations reduces their DNA fragments compared to other nucleosomes, which is a bias. Nucleosomes at the same positions will be over digested in repeated experiments. As mentioned above, MNase has a sequence preference, which influences where MNase preferentially cuts around individual nucleosomes based on the local sequence. Systematic deviations of the cut sites from the nucleosome ends that are unsymmetrical produce systematic errors between the center of a nucleosome and the measured fragments. Because these errors are systematic and reproducible for individual nucleosomes they are biases.

Nucleosome occupancy

I use the common definition of nucleosome occupancy described in Kaplan et al. (2010a). In brief, the nucleosome occupancy of a genomic position is the fraction of cells in which a nucleosome covers that position. Another way to describe the occupancy is as the fraction of time a nucleosome covers a position in an individual cell.

Nucleosome coverage derived from nucleosome position measurements are not proportional to the true occupancy and interpreting them as such leads to unrealistic distributions (Section 13.2). Because such issues quickly arise from such an interpretation, I believe calling the derived datasets 'nucleosome occupancy' is a misnomer. That being said, I will use the term when a more precise replacement term would be more confusing than helpful.

Nucleosome positioning

Nucleosome positioning will always refer to absolute nucleosome positioning as defined in Kaplan et al. (2010a). I will never talk about conditional nucleosome positioning, but I might mention rotational positioning. Both describe the local positioning strength: conditional nucleosome positioning compares the positioning over the whole nucleosome region; while rotational position focuses on a smaller region of <50 bps. While bordering nucleosomes can also influence the rotational position, I will solely use the term in the context of the sequence's influence on the positional preference. The nucleosome prediction methods I discuss and compare myself against predict nucleosome positioning. For analyses and validations, these are often transformed into genome-wide nucleosome coverage profiles by 147-bp smoothing.

3. Experimental methods

Different experimental methods exist that map nucleosomes genome-wide. The two main types of datasets analyzed in this work were measured by MNase-Seq and CC-Seq. Both experiments are discussed in detail, including their limitations – dealing with which became a major part of my work. I also analyzed and discuss datasets measured with other experiments, which are described more briefly.

In most cases the concept of the experiment is straight forward, but can become confusing due to scientific jargon. For this reason, I will add a brief explanation of each method based on a variation of the car parking analogy used in the introduction. Instead of cars parked on a road, I use toy cars glued onto cardboard strips (for simplicity sake their position is fixated in the analogy, this is not always done in the experiments). The same general rules apply as before, but now children with tools (enzymes or more generally molecules) can cut the cardboard, select cardboard pieces and tell us about the cardboard pieces they collected. As in real live, the children can have a mind of their own and their actions often deviate from what they were told to do.

Sequencing DNA fragments

Most experiments contain a sequencing step (Seq), which determines the nucleotide sequence of DNA fragments (cardboard pieces). The most common sequencing method of today uses the resynthesis of one of the DNA strands with florescent nucleotides. Each nucleotide is added individually and their color is imaged for millions of DNA fragments in parallel. In the analogy, the cardboard has color stripes of aquamarine (A), cyan (C), green (G) and turquoise (T). We can't distinguish the colors (after all they are all variations of green/green-blue), but children can. Each child has a single piece of cardboard and raises his hand when we call out the color that matches the current stripe. This lets us write down the letters of the DNA sequences. At this point the sequence information is digital, i.e. data on a computer, and there is no need for the 'analog' (and unreliable) children anymore. The next step maps the sequences against the genome to identify

the genomic position the fragment originate from. The computational methods used today are highly optimized, but the basic idea is the same as searching for a word in a text. Depending on the experiment, the measured fragment position represents different information that entails the nucleosome position (e.g. nucleosome borders).

3.1 MNase-Seq

The most common method to measure the positions of nucleosomes genome-wide is MNase-Seq. Children cut the cardboard with scissors (MNase) and once you are happy with the cutting progress, you have them collect car sized cardboard fragments for sequencing. Because the parts of the cardboard glued to a toy car are difficult to cut, children will primarily cut between cars. Therefore, a toy car covers most car sized cardboard fragments the children collect. The experiment has two primary problems: the children have a color preference when cutting the cardboard and they cut cardboard that is more easily accessible more frequently, because they are lazy.

Experimental protocol

MNase stands for micrococcal nuclease, which is a protein that digests DNA. Technically MNase cuts single strand DNA (Cockell et al., 1983). Its ability to cut double stranded DNA relies on single strand cuts i.e. nicks of both strands in proximity. Therefore, MNase digests single strand DNA much quicker than double strand DNA. Wrapping DNA into nucleosomes protects it from MNase digestion. The strength of this protection is an advantage of using MNase to map nucleosome positions over other nucleases from which the DNA is less protected by nucleosomes.

A second advantage is that MNase acts as both an endo- and pseudo-exonuclease. Endonucleases cut anywhere on a DNA fragment, while exonucleases chew off the fragment ends. The reason MNase only has a pseudo-exonuclease ability is because it does not specifically chew of parts of the fragment ends. MNase just preferentially cuts close to a fragment end removing 1-6 bps, which is probably a side effect of its preference to digest single strand DNA. The endonuclease ability is important for the first digestion of chromatin into DNA fragments with a single bound nucleosome. The pseudo-exonuclease ability is important to improve the resolution by trimming the fragment ends down to the nucleosome borders.

After the chromatin was digested with MNase and purified, the DNA fragments are separated by length via a gel electrophoresis. Once separated, the band of mono-

nucleosome length fragments is cut out of the gel and sequenced. The quantification of DNA fragments with sequencing is generally abbreviated with a 'Seq' tag to the experiment name and explained above. The experiment is perfect for high-throughput sequencing: the fragments are short (~147 bp), of similar length, and many need to be sequenced. In recent years, the length separation by gel electrophoresis is being replaced with a beads enrichment for short (e.g. <200 bp) fragments. This has little effect on the measurements, especially if the sequenced fragments are computationally filtered by length. Variations of the experimental protocol that have larger effects on the measurements are discussed below. MNase-ExoIII-Seq is an extension of the MNase-Seq protocol (Section 3.3.1).

Limitations of MNase-Seq

Understanding the limitations of an experimental methods is as important as understanding what it intends to measure. The toy car analogy above mentions the two primary problems: sequence bias of MNase and chromatin accessibility. Further limitations are: positional errors due to MNase not cutting precisely at the nucleosome borders, losing nucleosomes due to nucleosomes not fully protecting the DNA, and no absolute scale due to not quantifying the nucleosome-free DNA. The lack of an absolute scale makes correcting systematic errors created from the other limitations more difficult.

Each fragment represents one bound nucleosome in a single cell of the population. The ratio of fragment counts between positions reflects the ratio of cells in the population that have a nucleosome at these positions. However, we never know what fraction of the population has a nucleosome at a position. In theory, there could be cells that have no nucleosomes whatsoever. They would not influence the measured counts, but would reduce the absolute fraction of cells with a nucleosome at any given position. This fundamental limitation of MNase-Seq's design means that only relative occupancies can be derived from the measurements, and not absolute occupancies.

Most other limitations and issues arise from the digestion being a continuous process and MNase being imperfect. MNase preferentially cuts between TA, therefore the local sequence influences the cut frequency. MNase also has to gain access to the DNA to cut it, which – together with the non-uniform packing of the genome in cells – leads to different digestion speeds between genomic regions. Furthermore, the exonuclease activity of MNase is lower than its endonuclease activity, and MNase can digest nucleosomal DNA – if slower. Based on *in vitro* measurements, MNase may cut preferred sites covered by nucleosomes more frequently than unpreferred open site. In Section 8.1 I analyze some

of these biases and how they affect the measurements. All these effects lead to one important conclusion: experimental biases influence MNase-Seq measurements. A part of this is that the digestion has to be limited and this partial digestion leads to an uncertainty in the measured nucleosome positions. Processing the data can reduce the part of the positional uncertainty that is statistical noise, but the sequence preference of MNase creates systematic positional errors, which are more difficult to correct.

The effects of these biases on the measurements are complex and frequently ignored in analyses. This drastically restricts the expressiveness of the results and has been a major discussion point in literature. The biases also break basic assumptions of my probabilistic model (and others models), which prompted us to try to correct the biases (Section 17). In the end, my efforts in this regard were fruitless and I ended up having to use uncorrected MNase-Seq data.

MNase's sequence preference

As briefly mentioned above, MNase has a sequence preference. It prefers to cut between two 'weak' nucleotides (W = A or T), favoring the TA dinucleotide most (Fan et al., 2010). This preference leads to a 100-fold difference between cut frequencies of naked genomic DNA (*in vitro*). The *in vivo* effect is presumably smaller due to the reduced availability of cut sites, but even a 2-fold difference leads to biases that are stronger than most occupancy variation between nucleosomes. MNase's preference to cut A+T-rich regions may have an evolutionary background: nucleosome linker regions tend to be A+T-rich (also when determined by non-MNase experiments) and the original use of MNase is the digestion of chromatin, which is best achieved by cutting linkers.

Level of MNase digestion

As if it was not bad enough that the factors mentioned above bias the fragment counts, most of their effects depend on the digestion level. The digestion level depends on the MNase concentration and the digestion time, the effects of the two are mostly interchangeable. Matching the digestion level increases the reproducibility of experiments (Rizzo et al., 2012). The continuous digestion provides novel information when measuring several digestion levels (Weiner et al., 2010; Mieczkowski et al., 2016; Chereji et al., 2015). I also analyzed measurements of different MNase digestion levels in my collaboration with the Gaul-lab (Part II). Because different digestion levels produce different nucleosome measurements, of no single digestion level recreates the true nucleosome positioning information.

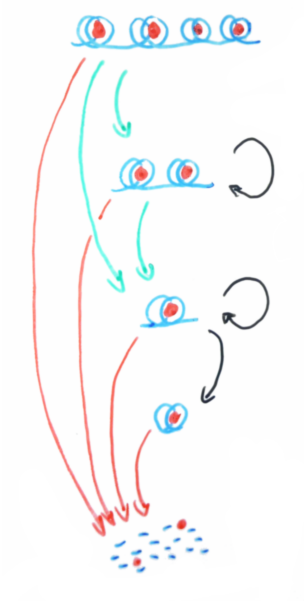

Figure 3.1: **MNase digestion flowchart**: Depicts how the genomic fragments flow between different fragment size groups during the MNase digestion. Green arrows represent MNase cuts in a linker region, red arrows represent full fragment losses (seldom), and black arrows represent MNase's pseudo-exonuclease activity with partial unwrapping for the over-digestion.

As of now, correcting MNase-Seq measurements for their digestion level is impossible, because the digestion process is highly complex and intertwined with chromatin features. Different aspects of the chromatin structure and DNA sequence influence each effect and they affect each other. The cut frequency of the neighboring linkers influences the creation of mono-nucleosome fragments. While MNase's sequence preference could have a simple relationship to the cut frequency, the pseudo-exonuclease activity of MNase makes the rest of the linker DNA influences the cut frequency as well. The effective linker size and its accessibility depends on the chromatin structure and the digestion of proximal linkers affects the accessibility. This reveals a looped interaction that already reaches beyond a nucleosome and its neighboring linkers.

Figure 3.1 depicts the flow diagram of the digestion process. One can imagine how the flow rates depend on the fragment concentrations, local sequence and other features that differ between genome regions and/or change with digestion time. The over digestion of nucleosomal DNA – depleting the mono-nucleosome fragments – depends on the availability of mono-nucleosomes fragments and their DNA sequence. With just one measurement of the central mono-nucleosome fragments the parameters of the in- and outflow are impossible to estimate. With more measurements, some parameters can theoretical

be estimated, but the problem is too complex to solve computational with our limited understanding of the process.

Variations of MNase-Seq Experiments

Some experimental protocols with larger changes are described separately. Here is an incomplete list of more minor variations of MNase-Seq. A larger change I want to mention here is quantifying the DNA fragments with microarray chips (chip) instead of sequencing (Seq). The popularity of microarrays has decreased in recent years, but due to distinct experimental biases and noise they can help distinguish signal from sequencing biases as a sort of control. Section 13.1 contains an in-depth discussion of some insights gained by comparing data measured with different experiments.

Two experiments, which are less a variation and more a different sample, are digesting *in vitro* constituted nucleosomes or naked genomic DNA instead of chromatin. Analyzing the *in vitro* data revealed how much the DNA sequence alone influences nucleosome positioning. The *in vitro* data was used to learn the nucleosome sequence preferences free from other influences (Kaplan et al., 2009). I also used the data to optimize my MNase-Seq based nucleosome energy model. The control experiment with naked genomic DNA revealed some of the limitations described in previously (Locke et al., 2010; Chung et al., 2010).

I previously mentioned measuring different digestion levels to gain novel information. The two main insights of such experiments are the existence of fragile (i.e. weakly bound) nucleosomes and large scale differences in accessibility. Changing the fragment-length selection step and sequencing both shorter and longer fragments provides further information about the MNase digestion process. Depending on the digestion level the shorter fragments are a different mixture of other DNA binding factors and over digested nucleosomes. Longer fragments are DNA sequences bound my multiple nucleosomes, where the linkers are still undigested.

Another variation of the protocol is the addition of a spike-in, which is popular for RNA-Seq experiments. A spike-in is DNA (typically of another species) that is added to the sample during the experiment and helps scale datasets against each other by providing shared reference points. Without a spike-in you cannot tell if a nucleosome is more or less frequent between experiments in absolute terms. You can only tell if a nucleosome is more enriched or depleted compared to the genomic average or another reference point, because of to the issue of relative measurements (Section 3.1). A limitation of spike-ins for MNase-Seq experiments is that they cannot correct for the digestion level, which conditions can

influence. For example, after PolII knockdown the same amount of MNase digests the chromatin noticeably slower (Weiner et al., 2010). If the digestion levels mismatch, the scale of the measurements will still differ after spike-in correction.

The last variation I want to mention is MNase-ChIP-Seq: the merging of MNase-Seq with another popular experimental protocols ChIP-Seq. Chromatin immunoprecipitation (ChIP) is a method to enrich fragments bound by a specific protein in a DNA sample (not to be confused with the microarray chips). A column of protein specific antibodies retains DNA bound to the protein of interest, while unbound DNA is washed away. ChIP-Seq is the primary method to measure *in vivo* transcription factor binding. You can ChIP histone variants, modifications or – less specific – one of the normal histones. In ChIP-Seq sonification or a similar method fragments the genomic DNA before ChIPing, while a MNase digestion replaces the step in MNase-ChIP-Seq. ChIP-exo is another variation of ChIP-Seq that uses a nuclease (Section 3.3.7. The advantage of nuclease digestion over sonification is a higher resolution, the disadvantage is that stronger experimental biases can occur.

3.2 CC-Seq (HC-Seq or chemical map)

A high-resolution method to measure nucleosome positions (Flaus et al., 1996) was adapted for genome-wide application: CC-Seq (chemical cleavage, HC-Seq – hydroxyl cleavage or 'Chemical Map' for the resulting nucleosome map) (Brogaard et al., 2012). Instead of scissors the children use safety knives to cut the cardboard. A cover hides the edge of the safety knives, making them blunt until the knives lock into a notch on the underside of the toy cars, where they can cut the cardboard. After the children cut the cardboard into pieces, you tell them to collect short cardboard pieces that roughly stretch from the center of one toy car to the next. The short fragments span between neighboring nucleosomes, because each cut site or fragment end represents one nucleosome dyad.

The main advantage of CC-Seq is the high-resolution of the cut sites in relation to the cars position. For my later analysis the experimental differences between CC-Seq and MNase-Seq are important: their unique approaches and different cut site locations in relation to the nucleosome lead to distinct biases. Deriving nucleosome energy models from measurements with different biases allow me to analyze where they agree and where they disagree, which points towards unhandled experimental biases.

Experimental protocol

In comparison to MNase-Seq, CC-Seq needs a genetically modified strain. In the strain H4S47C replaces the H4 histone. Covalently binding a copper-chelating label to H4S47C directs the chemical cleavage to the proximity of the nucleosome dyad. Added copper and hydrogen form hydroxyl radicals at the copper-chelating label. Due to the labels position the radicals cleave the DNA backbone close to the dyad, preferring the -1 and +6 positions in relation to the nucleosome dyad on either strand. The digested DNA is purified and the DNA fragments are separated by length via gel electrophoresis, as in MNase-Seq. However, the fragment ends – not the center – represent nucleosome positions in CC-Seq. A gel band is extracted and sequenced that contains fragment lengths that match the distance between neighboring nucleosomes dyads (~125-200 bp).

Advantages of CC-Seq

The main advantage of CC-Seq over MNase-Seq is the improved resolution. High-resolution dyad positions can be deconvoluted from the raw data, even though there are two preferred cut sites per strand. The preference of the two positions is so strong that the deconvolution is straight forward. A second advantage of CC-Seq is the use of small molecules: the label, copper and hydrogen are smaller than MNase and should have an easier time accessing compacted DNA regions.

The reason I focused on CC-Seq over the methods described below is that the protocol has the least similarity to MNase-Seq, while providing high-resolution dyad positions. The alternatives either create nucleosome fragments by digesting linker DNA like MNase-Seq or they measure occupancy values for a subset of positions. In the first case, the linker accessibility biases and the positions influencing the sequence biases of the digestion can overlap. In the second case, the comparison of datasets and adaptation of my method is more difficult and was beyond the scope of this work. I developed the equations for a variation of my method to learn a nucleosome binding energy model from occupancy score measurements (Section 20).

Limitations of CC-Seq

An obvious downside to selecting fragments that span between two neighboring nucleosomes is the enrichment of packed nucleosomes that are more probable to have a neighboring nucleosome close by. The original publication showed this by analyzing nucleosomes bordering nucleosome depleted regions: there is a disparity between the fragments extend-

ing in either direction (Brogaard et al., 2012). They could recover fragments spanning the nucleosome depleted region by extracting a band of longer fragments. To correct the problem without a second measurement, they combined the information of the two strands with a heuristic: ignore the lower if the two strands disagreed to strongly. While an improvement, the heuristic merely patches the most severe phenomena rather than fix the underlying issue.

The original authors observed a second possible bias: the -3 and +3 position close to the dyad have an unexpected A and T enrichment, respectively (Brogaard et al., 2012). To my knowledge, nobody has analyzed if the enrichment is a bias or not, there have only been hypothesis about possible sources (Cole et al., 2015). The original authors claimed in a later publication that it confirms an enrichment seen earlier (Xi et al., 2014), which I strongly disagree with (Section 19.1). In Section 13.4 I investigate this enrichment and confirm it as an experimental bias of CC-Seq.

As with MNase-Seq, the measurements could at best represent relative nucleosome occupancies, but even the interpretation as relative occupancies has issues. The two mentioned biases and possibly still unknown biases distort the occupancy values. Section 13.2 discusses a manifestation of these distortions and Section 16.4 mentions a specific problem it caused for my model.

3.3 Further experiments that measure nucleosome binding

In recent years, further experiments were published that measure nucleosome binding – primarily, or in conjunction with other genomic features. While I only analyzed datasets from some of these experiments, I have looked at the core properties of all the experiments. Here is an overview of how these experimental protocols measure nucleosome binding, their advantages in comparison to MNase-Seq, and their limitations.

3.3.1 MNase-ExoIII-Seq

MNase-ExoIII-Seq is an extension of the MNase-Seq protocol by adding exonulease III (ExoIII) to the digestion (Cole et al., 2015). This is equivalent to giving some children nail trimmers (compound lever style). They can trim cardboard fragments down, but they need the children with scissors (MNase) to cut the genomic cardboard strips into fragments first. The trimmed cardboard fragments match the cars footprint better, im-

proving the resolution of the measurements. The extension does not directly address any of the other issues, but it might indirectly reduce some, such as MNase's sequence bias.

Exonuclease III has a stronger exonuclease activity than MNase (which is low compared to its endonuclease activity) and less of a sequence bias. While exonuclease III only digests one strand away, MNase's high single-strand cleavage rate is expected to trim the other strand down. The increased exonuclease activity leads to a sharper fragment distribution, which reveals more specific preferentially protected fragment lengths. The enriched fragments that are shorter than a nucleosome match those found with MNase-Seq experiments, which are a main topic of Part II. The authors observed an enrichment of fragments 7 and 14 bp longer than the normal 147-bp nucleosome fragments. These proto-chromatosomes are at least partially H1 independent, but still represent a form of linker protection. For details on the proto-chromatosomes I refer to the original publication as these results – while interesting – are irrelevant to this work (Cole et al., 2015).

3.3.2 MPE-Seq

MPE-Seq (Methidiumpropyl-EDTA sequencing) follows the same idea as MNase-Seq (Ishii et al., 2015). In the digestion step Methidiumpropyl-EDTA replaces MNase. Methidiumpropyl and ferrous iron form MPE-Fe(II), which cleaves the DNA in the presence of oxygen. The cleavage occurs primarily in nucleosome linkers with little sequence bias. In my analogy you hand the children surgical scissors that are smaller and sharper than the normal scissors (MNase). The children more frequently cut under a toy car (smaller) and they cut the cardboard independent of its color (sharper – assuming the toughness of the colors differ). To prevent lots of cuts under the cars the children are given a much shorter time to cut the cardboard, which brings its own downsides.

The digestion time is so short that the majority of chromatin is still in >1 kbp long fragments, i.e. fragments of 5+ nucleosomes. Increasing the digestion time reduces the enrichment of preferentially protected fragment lengths. Long before the majority of chromatin is digested to mono-nucleosome size fragments the multi-nucleosome enrichment bands smear out and become indistinguishable from the background. While MPE-Seq improves on the sequence bias of MNase-Seq, the short digestion amplifies the problems related to accessibility and leads to a lower positional resolution, which the missing exonuclease activity reduces further.

3.3.3 NA-Seq

NA-Seq (nuclease-accessible site sequencing) uses restriction enzymes to measure chromatin accessibility (Gargiulo et al., 2009). Restriction enzymes are endonucleases that cut at a specific DNA sequence. The DNA sequence is typically 4 or 6 bps long and a reverse-complement palindrome, i.e. the sequence is identical to its own reverse complement. If a DNA sequence is reverse complement the structure of the molecule is point symmetrical. While the 'MNase' children have a slight preference for cutting sequences with a specific color patterns, 'restriction enzyme' children only cut between a specific color pattern. After letting the children lose on the toy car packed cardboard, the cars are removed and another set of children cut at a different color pattern. The second digestion creates smaller cardboard fragments that are easier to handle. Children are then encouraged to collect fragments with an end that matches the first cut color pattern.

An obvious downside of the method is resolution: the patterns are not uniformly distributed over a genome and a 4-bp pattern occurs on average only every $4^4 = 256$ bps. The original authors used two restriction enzymes for the initial chromatin digestion halving the average distance between cut sites to ~128 bp. The exact frequencies depend on the genomic oligonucleotide frequencies that can vary drastically. A resolution of about one measurement per nucleosome length is too low to extrapolate much information about nucleosome positioning. For other analysis such a low resolution is more than enough, as the authors showed. The underlying idea is also interesting in regards to measuring nucleosome occupancy, because there are fewer moving parts than in MNase-Seq, which improves the chances of correcting problems experimentally or computationally to achieve more quantitative measurements.

3.3.4 RED-Seq

RED-Seq (restriction endonuclease digestion coupled with sequencing) is an improved version of the NA-Seq protocol (Chen et al., 2014). The most prominent difference is the use of sonification instead of a second RE digestion. The children still cut at color patterns for the first fragmentation when the toy cars are present. In the second fragmentation (after removing the cars) the children randomly rip the cardboard into smaller pieces instead of cutting at a second pattern. This reduces issues related to the distances between sequence patterns. After collection, all the children only call out the color sequences from the cut side and never the ripped side. Therefore, the sequenced fragment end matches a cut site. In NA-Seq half of the sequenced ends stem from the

wrong side (second fragmentation) and are uninteresting or have to be assigned in an extra processing step.

The authors repeatedly claim RED-Seq is an 'unbiased' measurement of accessibility. This is false. Their main argument to support the claim is that RED-Seq does not artificially enrich for larger open chromatin regions (e.g. enhancers or promoters) like DNaseI-Seq or FAIRE-Seq. The NA-Seq authors had already mentioned the same about their method. While RED-Seq has a weaker enrichment than DNaseI-Seq and FAIRE-Seq, the authors never prove that RED-Seq has no enrichment at all. This aspect alone reduces the extent of their claim from 'unbiased' to 'less biased' and that is only in comparison to the two distinct methods, not NA-Seq, which they tried to improve upon. They themselves disprove their own claim by showing that RED-Seq measurements of chromatin and naked genomic DNA have a correlation coefficient of 0.376 instead of one close to the aspired 0. In comparison, the NA-Seq authors showed a correlation coefficient of -0.09 between their NA-Seq measurements of chromatin and naked genomic DNA. This means the method they are apparently improving upon appears less biased then the 'unbiased' RED-Seq. The RED-Seq authors also miss the opportunity to decrease the bias by using the information gained from the naked genomic DNA experiment to correct their other measurements.

3.3.5 DNase-FLASH

DNase-FLASH (DNase I-released fragment-length analysis of hypersensitivity) is another method that uses a nuclease (deoxyribonuclease - DNase) to digest the chromatin (Vierstra et al., 2014). DNase is a endonuclease that is frequently used in DNase-Seq to identify open chromatin regions, which are named DHS (DNase hypersensitivity sites) and I use in Part II. Using DNase is like giving the children snips to cut the cardboard. Their larger size makes it harder to cut between two toy cars that are close together, which decreases the frequency of the cutting in linkers. This leads to an enrichment of fragments at larger nucleosome free regions typically found at active enhancers and promoters. DNase can also cut inside nucleosome bound DNA, like MPE mentioned above. One could imagine the snips grasping the cardboard together with a toy car and – given enough force – cutting the cardboard under the car (technically without damaging the car, because DNase cuts the DNA without damaging the histones).

In the common DNase-Seq the frequency of DNase cuts is used to approximate the accessibility. DNase-FLASH adds the fragment-length information: most short fragments (<125 bps) span transcription factors or at least come from within a nucleosome-free

regions; longer fragments (126-185 bps) tend to span a whole nucleosome or at least part of one. The main innovation of DNase-FLASH is that it can simultaneously measure nucleosome and transcription factor binding events. The enrichment of fragments in and around large nucleosome free regions can be an advantage, e.g. by decreasing the needed sequencing depth if such regions are the primary interest. At the same time, the enrichment of course also means that the obtained nucleosome frequencies anything but represent the true nucleosome occupancies.

3.3.6 NucleoATAC

NucleoATAC, like DNase-FLASH, is an extended analyzing of DNA accessibility data – in this case ATAC-Seq – to derive nucleosome positions and occupancy values (Schep et al., 2015). ATAC-Seq (assay for transposase-accessible chromatin using sequencing) uses hyperactive Tn5 transposase loaded with sequencing adapters (needed for sequencing and usually added in a separate step). Transposons are enzymes that insert a DNA fragment into the genome. In nature, the inserted DNA fragment is the gene that encodes the transposon itself, which creates a sort of reproduction cycle of the enzyme and gene. By loading the enzyme with two sequencing adapters, instead of the normal DNA fragment, the enzyme directly tags its cut site for sequencing. Basically, the children use a special pair of scissors that automatically glues red stripes to the two new ends when cutting the cardboard. The red stripes are also added in the other experiments, but in a separate step called library preparation - the description skipped the step for simplicity. ATAC-Seq improves upon DNase-Seq by requiring less sample material to produce data of the same or higher quality. However, other issues of DNase-Seq are present in ATAC-Seq, some of which NucleoATAC tries to address.

NucleoATAC consists of two parts: calling nucleosome positions and computing nucleosome occupancy values. To create the nucleosome position map NucleoATAC scans over the genome with a 2D-footprint pattern that describes the typical distribution of fragment lengths versus central positions in relation to the nucleosome dyad. The footprint pattern contains the information displayed in a V-plot (see Figure 8.6 in Section 8.6 for a V-plot). After computing the cross-correlations between the footprint pattern and the genomic data, NucleoATAC calls the highest local peaks as nucleosome positions. Regarding the nucleosome occupancy, the authors realized that the large scale cross-correlation trends depend on the fragment coverage, which is highly biased between genome regions. To circumvent this issue they compute the nucleosome occupancy independently. NucleoATAC models the occupancy by describing the observed fragment-length distribution as a mix-

ture of nucleosome and nucleosome-free distributions. As for DNase-FLASH, fragments stemming from nucleosomes tend to be longer than fragments from nucleosome-free regions. They define the occupancy as the fraction assigned to the nucleosome distribution for the mixture that maximizes the likelihood of the observed distribution.

A great advantage of using a fraction to estimate the occupancy value is that it bounds the occupancy between 0 and 1 by definition. This limits the effect size biases can have, assuming the measured signal covers most of the range. The distinction between calling nucleosome positions and computing the occupancy values makes it difficult to formulate a probabilistic model of the data for my model to use. For most other analysis this is not an issue or even simplifies them, for example it becomes easier to distinguish between positional shifts and occupancy changes when comparing datasets.

3.3.7 ChIP-exo

ChIP-exo (chromatin immunoprecipitation with exonuclease digestion) is a variation of the ChIP-Seq protocol that improves the positional resolution (Rhee and Pugh, 2011). Figure 3.1 mentions ChIP-Seq and the possibility of using MNase digestion for fragmentation (MNase-ChIP-Seq). ChIP-Seq measures the binding sites of a DNA binding protein genome-wide. It consists of three major steps: fragmenting the chromatin with sonification (after fixating the protein-DNA interactions), enriching the protein of interest together with bound DNA fragments, and lastly sequencing the fragments. The first and third step were covered before and need no further explanation. In the second step, antibodies stuck to beads (or another surface) enrich the protein of interest by binding it, while the rest of the sample is washed away. These antibodies are the type used by the immune system to recognize proteins that do not belong in the organism and are therefor probably pathogens (hence the name immunoprecipitation).

ChIP-exo adds a further step after the second step: a lambda exonuclease digestion. Lambda exonuclease digests the nucleotides of one strand (5'-3') until a protein bound to the DNA blocks its path, leaving the other strand intact (3'-5'). Due to the anti-parallelism of double-stranded DNA, you can think of the two strands being arrows pointing in opposite directions. Lambda exonuclease prunes the origins of the arrows down to the protein binding site, while the arrowheads are left untouched. By measuring the arrow, both location and direction, one can then map the borders of the protein binding site with a high resolution. In comparison, ChIP-Seq measures the location of the full arrows (fragment), without pruning them (making the direction irrelevant). This tells you that the protein was bound somewhere inside the fragment, but not where

exactly the binding site is.

The higher resolution is an obvious advantage of ChIP-exo against ChIP-Seq, but the resolution of measuring nucleosomes was not compared with MNase-Seq. ChIP-exo can enrich individual histones and provide information about the internal structure of the DNA-histone interactions (Rhee et al., 2014). The authors originally neglected possible biases, because they have little impact when identifying transcription factor binding sites, which occur rather infrequently in the genome. In comparison, nucleosomes cover most of the genome, which provides more opportunity for biases to arise. The relative binding frequency between the sites is also more important, the measurement of which is severely affected by these biases. Section 19.3 discusses the published results and touches on this.

3.3.8 EM analysis of DNA molecules

A distinct method of measuring nucleosomes uses electron microscopy (EM) imaging of individual DNA molecules (Brown et al., 2013). The method measures the location of all nucleosomes along a specific stretch of DNA up to 2 kbps long. In a cross-linking step the two strands of nucleosome unoccupied DNA are connected, while nucleosomes protect the DNA from cross-linking. A later denaturing step locally separates the nucleosome bound DNA into its two strands creating so called bubbles. DNA molecules are imaged with EM and the position of the bubbles along the DNA fragment are retrieved.

Until now, the toy car analogy represented DNA as a single piece of cardboard. The explanation of this experiment needs slightly more biological details: the DNA analog consists of two cardboard strands stuck together with Velcro. The children should first staple the two halves together, which they do with the car-free cardboard regions. Next, they remove the cars and pull the two cardboard pieces apart where possible (i.e. where the cars were beforehand). The children cut out and collect a specific region with differently marked start and end. Based on pictures of these cardboard pieces the approximate car positions are figured out. Taking and processing these pictures is much slower than using the children to determine the color stripes.

Compared to all previous methods, this method measures all nucleosome occurrences over a single DNA molecule and not each nucleosome position independently. The main downside is that the authors had to do a lot of the image analysis by hand, or at least double check the results by hand, leading to a much lower throughput. Estimating the position of the bubble along the DNA strand in the image also limits the methods resolution. The authors only distinguished between the -3, -2 and -1 nucleosomes at the promoters without looking at the precise positioning. An upside is that the method

measures both nucleosome-bound and -free DNA, which means it measures the occupancy. Given the low throughput and low resolution there is to little data published to analyze the nucleosome binding behavior based on it. However, nucleosome position predictions can be independently validated with such measurements, especially if they predict the frequencies of individual configurations at promoters.

3.3.9 MTase

DNA methyltransferases (MTases) are used to measure nucleosome occupancy in a variate of experimental protocols (Jessen et al., 2006; Small et al., 2014). MTases adds a methyl group (a single carbon atom with three hydrogens) onto C nucleotides – subject to accessibility and context. Different MTases can methylate Cs in different contexts, for example the common MTases in higher eukaryotes can only methylate CpG (C before a G). When using an MTase to measure nucleosome occupancy, the methylation marks the nucleosome-free DNA due to accessibility of the DNA. For this to work the context of the used MTase cannot be naturally methylated in the species.

As with other experimental protocols, sequencing has replaced other methods of measuring the information. The published protocols do not use high-throughput sequencing (hence the missing Seq suffix), instead they measured single loci (a small genomic region). If the sequencing method cannot detect methylations – most cannot – the sample is bisulfite converted before sequencing. The bisulfite conversion modifies unmethlated Cs into Us, which appear as Ts when sequenced. BS-Seq (bisulfite conversion with sequencing) uses this step without a previous methylation to measure naturally occurring methylation with high-throughput sequencing.

A way to imagine the MTase protocol in the toy car analogy is that children use hole punches (MTase) to punch holes in cyan if green follows it (strand specific). As with the scissors, the children have a difficult time accessing the cardboard covered by toy cars. Because the children do not mention the holes while determining the colors, they have to recolor cyan stripes without holes to turquoise between the two steps. For the sequencing, children extract fragments that cover a specific region instead of fragmenting and collecting everything.

The published MTase based methods have similar advantages and disadvantages as EM analysis described previously. On the one hand, they measure both nucleosome bound and unbound positions for single DNA molecules that are hundreds of basepairs long. On the other hand, the low-throughput sequencing methods restrict the measurements to individual regions and the dependence on Cs in specific contexts limits the resolution.

Nonetheless, looking at one analyzed region shows how unreliable MNase-Seq coverage might represent nucleosome occupancy: while the MNase-Seq coverage changes more than 4-fold between three nucleosome positions, the MTase method measures occupancies of 90%, 96% and 100% (i.e. a 1.1-fold change) for the same nucleosomes (Small et al., 2014).

3.3.10 NOMe-Seq

NOMe-Seq (nucleosome occupancy and methylome sequencing) is an extension of the MTase based nucleosome occupancy measurement described above (Kelly et al., 2012). The two main differences are: genome-wide measurements due to high-throughput sequencing, and the simultaneous measurement of endogenous (i.e. naturally occurring) CpG methylation and nucleosome positions. The second aspect is actually an unavoidable side effect of measuring nucleosome occupancy by methylation in species that have their own MTase. To be able to distinguish the endogenous and experimental methylations, the MTase used for probing the nucleosome occupancy has to methylate Cs in a different contexts than the endogenously methylated CpG. Therefore, NOMe-Seq probes the nucleosome positions with a GpC MTase (M.CviPI).

The analogy of the MTase based method needs little adaptation. Some cyan stripes followed by green already have holes to begin with and the children now punch holes into stripes of cyan preceded by green. The two types of holes are later separated based on the context. The recoloring step is identical and the sequencing step matches the high-throughput sequencing described for the other methods. Mapping the fragments to the genome becomes harder due to the recoloring, but that detail is unimportant for the described analysis.

In principle, NOMe-Seq has the advantages of the MTase protocol with the added advantage of genome-wide measurements. In practice, common high-throughput sequencing methods today produce much shorter reads (50-200 bp) than what the MTase protocol uses. By analyzing the published datasets I found problems that made me cautious to rely on the data. Section 19.2 goes into detail, in brief some control comparisons pointed to high noise and biases. One source of the high noise is the low sequence coverage. This is a common issue when analyzing human cells, whose genome is large compared to yeast – from which most other measurements I used stem. Once these problems are addressed and corrected, I believe that NOMe-Seq (and MTase-Seq for lower eukaryotes) or an experiment based on it will oust MNase-Seq as the method of choice to measure nucleosome positioning.

3.3.11 BunDLE-Seq

BunDLE-Seq (Binding to Designed Library, Extracting, and Sequencing) is an experiment to quantitatively measure the DNA binding properties of a protein to a synthetic short sequence library *in vitro* (Levo et al., 2015). BunDLE-Seq is rather different than all the other methods described earlier. The sequence library contains many copies of thousands of unique short (~200 bp) sequences. The design of the short sequences has no limitations, because the synthesis process can create any sequence. Parts of the library are mixed with the protein of interest at different concentrations *in vitro* (i.e. in a test tube, not living cells). The protein of interest binds the DNA fragments with different preferences based on their sequence. DNA bound by proteins moves slower during a gel electrophoresis, creating distinct bands for unbound DNA and DNA bound by a protein. The amount of proteins bound to a fragment leads to distinct bands, which can be analyzed as well. The bands are extracted and sequenced individually to measure the frequency of every sequence. From the frequencies in the different bands and for the different concentrations the binding preference of the protein to different sequences is derived.

The lab that published BunDLE-Seq had used a similar experimental protocol before on nucleosomes (Kaplan et al., 2009). At the time they quantified the bound DNA sequences with high-throughput sequencing and microarrays. Other experiments measure the binding of transcription factors in similar fashions, but to my knowledge nobody has measured nucleosome binding with them. BunDLE-Seq (and the unnamed predecessor) are the purest methods to measure the sequence preference of *in vitro* nucleosomes. In comparison to *in vitro* measurements of naked genomic DNA, short DNA fragments even prevent steric hindrance between nucleosomes to affect the binding. A significant detail for *in vitro* measurements of nucleosome binding is the nucleosome assembly on the DNA. The common method of salt-gradient dialysis could create a biased preference for higher G+C content compared to the *in vivo* nucleosome assembly (Chung et al., 2010).

4. Nucleosome-position prediction methods

The main focus of my work was to construct a nucleosome-position prediction method that incorporates experimental errors and uncertainties. This section gives a brief overview of other methods that use a thermodynamic model similar to mine. A distinction between nucleosome-position prediction methods is if they try to represent the biochemical mechanics or focus on maximizing benchmarking scores. An important part of representing the biochemical mechanics is a thermodynamic model that treats the nucleosome sequence preferences as binding energies and includes steric hindrance between neighboring nucleosomes. Section 12.1 explains the mathematical details of such a thermodynamic model, because my method includes such a thermodynamic model. The thermodynamic models sometimes have minor variations that are not described here.

4.1 Segal et al.

Segal et al. (2006) were the first to treat the nucleosomes like transcription factors and use a thermodynamic model in combination with dynamic programming – a Forward/Backward algorithm – to predict genome-wide nucleosome positioning with steric hindrance. They represent the sequence preference $P_N(S)$ as a 1st-order Markov chain (even if the did not call it that). In this context a 1st-order Markov chain is more frequently referred to as a 1st-order Markov model (MM) and sometimes as a 1st-order position weight matrix (PWM) They derive the 1st-order MM directly from the dinucleotide frequencies around their measured nucleosome positions. Based on unpublished results that the precise base-pair position has little effect on nucleosome binding energies (see Section 13.3 why this is not true) they smooth their MM over the two neighboring positions (± 1 bp). To compute the free energy of a DNA sequence to bind a nucleosome, the method divides the sequence preference $P_N(S)$ by the background probability of the

sequence $P_B(S)$, which describe the probability of seeing such a sequence by chance.

4.2 Field et al.

Field et al. (2008) published an extension of the model by Segal et al. (2006). They derive new sequence preferences from their new dataset of fully sequenced nucleosome fragments. The majority of the model is identical to Segal et al.'s model described above. The difference between the models is the replacement of the mononucleotide based background probability $P_B(S)$ with a pentanucleotide based linker probability $P_L(S)$. The linker probability $P_L(S)$ has no clear motivation, because it does not represent any specific biochemical mechanic. One could argue that it represents competition with transcription factors and nucleosome remodeler activity, but it merely approximates these mechanics and convolves them with actual nucleosome binding preferences, because linker DNA is also depleted of nucleosome-favoring sequences.

4.3 Kaplan et al.

Kaplan et al. (2009) reused the model of Field et al. (2008), replacing the sequence preference $P_N(S)$. They argue that *in vivo* measurements contain signals other than the pure sequence preference of nucleosomes. They compute new sequence preferences from their *in vitro* nucleosome measurements. Note that my previous attempt at justifying $P_L(S)$ relied on approximating other *in vivo* signals. While the authors motivate the use of *in vitro* measurements with the separation of the nucleosome binding preference from other *in vivo* effects, they appear to be indifferent in representing the nucleosome binding preferences in a way that is biochemically meaningful.

4.4 NuPoP

Xi et al. (2010) extend the model of Field et al. (2008) by three aspects and reused their data to derive new parameters. First, they extend the 1st-order MM (i.e. Markov chain) of the nucleosome sequence preference to a 4th-order MM. Second, to improve predictions for genomes of other species when using the model obtained from yeast, they rescale their model based on the ratio between the nucleotide composition of yeast and the target species. Third, they replace the implicit geometric-duration distribution of linker lengths with an empirical distribution. The geometric-duration distribution is a

result of using a hidden Markov model (HMM) to describes the thermodynamic model, i.e. steric hindrance of nucleosomes, without explicitly modeling lengths.

The lab that had published the three previous methods (Segal et al., Field et al., Kaplan et al.) relaxed the duration distribution of the HMM in their own extension before NuPoP was published (Lubliner and Segal, 2009). They tested different distributions that had two to five parameters and set a maximum linker length of 100 bps. In comparison the empirical distribution of NuPoP is position-specific (one parameter per position) and they used a maximum linker length of 500 bp. To reduce noise in the highly parameterized model, NuPoP smooths the empirical distribution with a gaussian kernel. NuPoP and the method published by Lubliner and Segal (2009) use different strategies to optimize the distribution's parameters.

4.5 NucEnerGen

Locke et al. (2010) devised a new way to extract the nucleosome sequence preference from experimental measurements. Instead of deriving an energy model directly from the nucleotide frequencies, they account for steric hindrance of neighboring nucleosomes when deriving the nucleosome formation energies. They then fit the parameters of the sequence model to approximate these formation energies. They tested and compared different sequence model: e.g. a position-specific 1st-order model, position-unspecific 1st- and 4th-order models. Instead of smoothing the sequence parameters after obtaining them, a preprocessing step smoothed the experimental measurements before deriving the formation energies. They used a basic thermodynamic model to predict nucleosome positions without a special model for the linker length distribution.

4.6 Further methods

Many unmentioned methods exist that predict nucleosome positioning in the broadest sense (Liu et al., 2014; Teif, 2015; Scipioni and Santis, 2011). A comprehensive list is beyond the scope of this work. Here is an incomplete list of the concepts behind the methods: adding transcription-factor competition (Wasson and Hartemink, 2009; Ozonov and van Nimwegen, 2013), adding nucleosome remodelers (Teif and Rippe, 2009), extracting the nucleosome binding energies from crystal structures (Tolstorukov et al., 2008; Minary and Levitt, 2014), and focusing on nucleosome-nucleosome interactions instead of the sequence (Chereji and Morozov, 2011; Möbius et al., 2013; Parmar et al.,

2013).

A comparison between most of these methods and mine is difficult. A good fraction of them predict the binding preference of nucleosomes to short DNA stretches (i.e. ignore steric hindrance) or provide low amounts of information in their prediction (e.g. a nucleosome map of non-overlapping nucleosome positions without occupancy information). Some of the webservers and tools also failed to work or required unreasonable third-party software when I wanted to test them. For these reasons I focused my benchmarks in Part III on a subset of the methods described above.

Part II

Sequence and activity driven nucleosomal features in promoters and enhancers of Drosophila

5. Abstract

The stability of nucleosomes in the genome depends on the G+C content and bendability of the bound DNA. I analyze MNase-Seq measurements of mono-, sub-, and di-nucleosome-length fragments from different digestion levels to investigate this relation and to examine experimental biases of MNase-Seq. I define a fragility score for nucleosomes based on their mono-to-sub-nucleosome coverage ratio, a measure of the nucleosomes' probability to partially unwrap. I explore the fragile, average and resistant nucleosome populations, which have characteristic dinucleotide frequencies, and their placement around promoters, transcription termination sites and enhancers. *Drosophila* has two types of promoters: broad peak for constitutively expressed and narrow peak for inducible genes. I find that in broad-peak promoters of highly expressed genes the nucleosome fragility is carved into the sequence landscape, while in narrow-peak promoters the fragility is more activity-driven. Enhancers show an interesting antagonistic relationship between the G+C landscape favoring resistance and activity-driven fragility. Finally, I analyze the partial unwrapping of nucleosomes genome-wide and find that G+C content is a determinant of asymmetric unwrapping.

6. Introduction

Eukaryotic genomes are packaged into chromatin, whose basic repeating unit is a nucleosome, which consists of a histone octamer wrapped around 147 bps of DNA (Luger et al., 1997). High-resolution genome-wide nucleosome maps have shown that the majority of yeast promoters have a canonical pattern, where well-positioned -1 and +1 nucleosomes flank a nucleosome-depleted region, while inducible promoters that contains a TATA-box have a weaker pattern (Ioshikhes et al., 2006; Lam et al., 2008). Nucleosome-depleted regions also occur in enhancers, and at termination sites (Struhl and Segal, 2013). Controlling DNA accessibility via nucleosomes is an important part of gene regulation and an accurate understanding of the DNA-binding behavior of nucleosomes is vital to quantitatively model gene expression.

Nucleosome positioning depends on multiple factors including DNA sequence, nucleosome remodelers, and competition with other DNA-binding proteins. Nucleosomes have a strong DNA sequence preference *in vitro*, both between competing short DNA fragments and in genome-wide measurements (Thåström et al., 1999; Kaplan et al., 2009; Levo et al., 2015). The influence of this preference on *in vivo* nucleosome positioning is debated (Zhang et al., 2009; Hughes et al., 2012). While G+C content is the best individual predictor of nucleosome affinity, this likely reflects G+C content's correlation with other sequence features that affect structural characteristics of DNA (Tillo and Hughes, 2009). Optimal nucleosome formation occurs when bendable dinucleotides (AT, TA, AA) occur on the face of the helical repeat (~10.5 bp) in proximity of the histones, while stiffer dinucleotides (GC) are located ~5 bp out of phase (Richmond and Davey, 2003). The extended stiffness of the homo-polymeric sequences poly(dG:dC) and poly(dA:dT) disfavor nucleosome formation (Struhl and Segal, 2013). Poly(dA:dT) stretches are more frequent in eukaryotic genomes than expected by the DNA composition (Dechering, 1998). They decrease nucleosome occupancy and increase expression when inserted into yeast promoters (Raveh-Sadka et al., 2012). These DNA sequence features are the primary information used to predict nucleosome positioning computationally (Teif, 2015).

Nucleosomes are highly dynamic structures and transiently partially unwrap and rewrap the DNA in a fraction of a second, which provides opportunities for other factors to access the DNA (Li et al., 2005). In a single molecule experiment the unwrapping direction related to the relative DNA flexibility within the nucleosomal DNA such that the nucleosome preferentially unwraps from the stiffer side (Ngo et al., 2015).

MNase-Seq is the standard experiment to map nucleosomes genome-wide (Section 3.1). MNase is an endo-pseudo-exonuclease, which preferentially cuts and digests naked DNA. MNase first digests linker regions, leaving histone protected DNA fragments, which are then sequenced to obtain nucleosome maps. Such nucleosome maps are affected by MNase's digestion biases. MNase prefers cutting A+T-rich regions (Dingwall et al., 1981) leading to a non-uniform digestion of linkers. Together with the continuous nature of the MNase digestion, this leads to the obtained nucleosome landscapes depending on the digestion level (Weiner et al., 2010; Xi et al., 2011). This information has been used to characterize nucleosomes by their differential susceptibility to MNase titration (Mieczkowski et al., 2016; Chereji et al., 2015)

In this collaboration with the Gaul-lab, I analyze MNase-Seq measurements of *Drosophila* S2 cells. I investigate MNase's sequence preference by exploring the sequence composition of samples with different fragment sizes and digestion levels produced from genomic DNA, native, and cross-linked chromatin. Comparing different size fractions of a digest provides valuable information about nucleosome fragility and unwrapping behavior, without the need for multiple digestions. My fragility score – the ratio between mono- and sub-nucleosomal fragments – is meant to capture the ease of nucleosome eviction and not the accessibility of the nucleosome as a whole. I assign nucleosomes into fragile, average and resistant populations based on this score and explore their sequence features. I analyze these nucleosome populations in the context of promoters, enhancers and transcription termination sites. Finally, my genome-wide analysis of sub-nucleosome fragments confirms asymmetrical unwrapping due to sequence asymmetry *in vivo*, which to date had only been measured in single molecule experiments *in vitro*.

7. Methods

7.1 Experimental Procedures

My collaboration partners from the Gaul-lab performed the experiments outlined here.

MNase digestion was performed on chromatin and naked genomic DNA (gDNA) extracted from S2 cells. To perform the short, typical and long digestion levels the digestion time was varied. Di-, mono- and sub-nucleosome fractions were isolated from an agarose gel and, after purification, libraries were prepared from them. The sample libraries were sequenced on an Illumina GenomeAnalyzer IIx for around 40 million 50-bp paired-end reads.

The same MNase-Seq protocol was used for native chromatin, gDNA, and cross-linked chromatin once the chromatin or gDNA was prepared for MNase digestion. The digestion times were sometimes adapted to the different digestion rates of the conditions. The genomic DNA (gDNA) was isolated with a high-salt procedure. The cross-linking was performed with Formaldehyde.

DNase-Seq was performed as described previously (Vierstra et al., 2014) with minor modifications for S2 cells. After fractionation, an additional size selection was performed during library preparation to enrich for fragments shorter than 150 bp. Sequencing the libraries resulted in 80-100 million 50-bp paired-end reads. The reads were trimmed and mapped using Bowtie2 (Langmead and Salzberg, 2012). Additional filtering was performed using SAMtools (Li et al., 2009). Finally, MACS2 was used to call DHS peaks (Zhang et al., 2008). Processing the DNase-Seq data in this ways was performed by the Gaul-lab.

7.2 Mapping of the MNase-Seq data

Reads were mapped with Bowtie2 (Langmead and Salzberg, 2012) (v2.1.0, parameters: -I 0 -X 1000 -p 4) to the flybase v5.53 *D.melanogaster* genome (Attrill et al., 2015).

Regions where fragments cannot be uniquely mapped to (unmappable) were excluded from all analysis. These unmappable regions were identified by slicing the genome into overlapping 150-bp (roughly one nucleosome length) fragments, creating *in silico* paired-end reads from them, and mapping these against the genome. Excluding all reads that mapped to multiple genomic locations, the coverage was computed and only genome regions with a coverage of 150 (all generated fragments) were kept for the analysis.

Extremely short or long (<30, >400) fragments were excluded from all analysis. For the analysis we tested, repeating them with a stricter *in silico* fragment-size selection did not affect the results. Generally, the nucleosome-dyad position is defined as the fragment center. In the case of di-nucleosome fragments, two dyad positions are estimated from the fragment ends. Each is placed half the average length of a mono-nucleosome fragment (same digestion level) from one end.

The genome-wide coverage tracks were computed by extending the dyad positions with ±73 bps and summing the coverage per base pair. Such coverage tracks are an unscaled approximation of the nucleosome occupancy. The coverage tracks were normalized to a genome-wide average of 1, after doubling the counts on chromosome X, since S2 cells only have one copy due to their sex. Chromosome Y was removed completely, together with the heterochromosome regions of the other chromosomes, due to low mappability.

7.3 Nucleosome calling

For the fragility and resistance analysis, nucleosome positions were called with the R package nucleR (Flores and Orozco, 2011). Peaks were called independently on the mono-nucleosome fractions of the three digestion levels and merged in the final step. Fragments of length 50-200 bp were processed as described in the package vignette. Peaks were called for each dataset with a threshold of 25%. The peaks were filtered by their h-score which describes the height i.e. the amount of count data (>0.55 for the group 'all' used throughout the main figures and >0.9 for the group 'only high'). The filtered peaks of the three datasets were joined and overlapping peaks were merged (<21 bp between the peaks), using the joint center as the called nucleosome position. The filtering and merging was done to reduce error in the analyses.

7.4 Promoters

Gene expression in *Drosophila* is regulated over four orders of magnitude. Two main promoter architectures were originally defined by promoter width, but also show distinct motif composition and expression plasticity (Rach et al., 2009). The broad-peak (BP) promoters, in which the transcription start sites (TSS) are dispersed over tens of bps, are typically found in constitutively expressed genes and have a canonical pattern of nucleosomes (Rach et al., 2011). The narrow peak (NP) promoters, in which the transcription start site (TSS) is sharply defined within a few bps, are typically found in inducible genes with high expression plasticity and their nucleosome patterns are non-canonical (Rach et al., 2011).

I derived 15,971 promoters assigned to 11,536 unique genes from clustering cap analysis of gene expression (CAGE) data (Brown et al., 2014). Transcript annotations were took from the flybase v5.53 (Attrill et al., 2015). NP and BP promoters are distinguished by the dispersion of their transcription initiation. The promoters were classified into 8709 BP and 7262 NP promoters, based on the mean absolute deviation of the CAGE data mapped to the TSS. The promoter list is an updated version of the one used in Siebert and Söding (2016).

Because divergent promoters can share one nucleosome-depleted region, directional and divergent promoters were separated. Divergent promoters were defined to have a distance below 500 bp between TSSs. For NP only the directional promoters were used in the main figures, because divergent promoters could be contaminated with sequence signal of a BP promoter on the opposite strand.

Processed RNA-Seq data was mapped from transcripts that started close ($\pm$50 bps) to the TSS and summed for each promoter. The promoters (BP and NP combined) were separated into quarters based on their assigned expression values. The 4th quarter is marginally extended to contain all unexpressed promoters. Only after the expression based quartering the promoters were further separated based on the other categories (NP/BP and directional/divergent).

7.5 Fragility score and nucleosome populations

The fragility score is defined as the normalized mono-nucleosomal coverage divided by the normalized sub-nucleosomal coverage. Nucleosomes were assigned to the populations

based on the fragility score of the typical digestion:

$$\begin{aligned} \text{fragile} &: \frac{\text{sub-nucleosome}^{\text{typical}}}{\text{mono-nucleosome}^{\text{typical}}} > 2 \\ \text{resistant} &: \frac{\text{mono-nucleosome}^{\text{typical}}}{\text{sub-nucleosome}^{\text{typical}}} > 2 \end{aligned} \tag{7.1}$$

They also needed a minimum value of 0.5 for both coverages to reduce noise. The left panel of Figure 8.4A shows the thresholds as blue and red lines. Nucleosomes that were called neither fragile nor resistant were assigned to the average population.

7.6 Additional details for the figures

Figure 8.1

(D) The Pearson correlations are computed between the genome-wide coverage tracks (excluding heterochromatin and unmappable regions). The full matrix of pairwise Pearson correlations is given in Figure A.4.

Figure 8.2 To analyze the MNase cut site, the genomic contexts were aligned by the upstream fragment ends (lower genomic coordinate).

(A) The PWMs were generated from mono-nucleotide frequencies with a uniform background frequency (A=C=G=T=0.25).

(B) The di-nucleotide fold-changes are the $\log_{10}$ values of their position-specific frequencies divided by their average genomic frequency.

(C) For the profile figures over the whole nucleosome region the fold-changes are smoothed with a 3-bp running window.

Figure 8.3

(A) The coverage tracks are normalized to a genomic average of 1. The gene annotation is taken from flybase.

(B) (C) The di-nucleotide panels are as described in Figure 8.2C.

Figure 8.4

(A) (B) (D) Each dot in the scatter plots is one nucleosome colored by a local density estimate. The x- and y-axis show the normalized coverage, fragility score, or G+C content depending on the subfigure.

(B) For the sequence-feature score distributions a 1st-order Markov model (dinucleotide model) was generated from the typical digestion mono-nucleosome fragments – aligned by their upstream cut site. The outer 15 bps of either side of the 147 bps bound by the nucleosome were excluded to avoid the MNase-sequence bias effecting the results. All fragment-based dyad positions or called nucleosome positions were scored (log-odd score) based on the model, while allowing a ±5 bp shift (using the maximum) to compensate for MNase-Seq's positional uncertainty. Density representations of the distributions were created with R's default values.

(C) The di-nucleosome plots were generated as for Figure 8.2, while aligning the genomic contexts to the called nucleosome dyads. The fold changes were smoothed with a 3-bp running window, as before.

Figure 8.5 The fragile, average, and resistant populations based on the called nucleosome positions were used for the figures. The profiles were smoothed over a window of ±45 bp.

(B) The DHS peaks called by MACS were used and split into quarters based on their fold-change. For Figure A.7 the S2 peaks that overlapped with a OSC cell peak published by Arnold et al. (2013) were grouped in to the S2/OSC shared population and all peaks that overlapped the peaks of the other dataset were removed from the 'only' populations.

Figure 8.6

(C) For the called nucleosome positions the G+C-ratio between the left half (-73:0) and right half (0:73) were computed and used to split the nucleosomes into five categories (only three shown, the mirrored G+C-ratios show mirrored distributions). The V-plots show the normalized dyad summary split by distance to the called nucleosome (x-axis) and fragment length from which the dyad count originated (y-axis). The counts are normalized by the amount of nucleosome positions summed over and the total dyad counts of the dataset so that the genome-wide average of the sum over all fragment lengths is 1.

7.7 Implementation

I did all the computational analysis in R (R Core Team, 2016). I used the bioconductor packages Biostring, rtracklayer, GenomicRanges, and GenomicAlignments to handle the genome and sequencing data (Pages et al., 2016; Lawrence et al., 2009, 2013), the packages ff and ffbase to process many genomic data vectors simultaneously (Adler et al., 2014; de Jonge et al., 2015), and the heatscatter from the LSD package for a visualization (Schwalb et al., 2015). Scons was used to automate the figure generations for TSS, TTS, and DHS datasets (Knight, 2005).

8. Results

For my initial analysis, I looked at three different MNase-digestion times of native chromatin (Figure 8.1A and C): a short 1 minute digestion, a typical 3 minute digestion and a long 15 minute digestion. I analyzed three fragment lengths of the digested chromatin when they were present: sub-, mono- and di-nucleosome fragments. The short digestion results in 10-15% mono-nucleosome fragments, while the remainder of the chromatin is in larger fragments like di-nucleosomes. The typical digestion matches the commonly used digestions by other groups and results in 60-65% of the fragments having mono- or sub-nucleosomal length. The long digestion results in the combined mono- and sub-nucleosome fragments accounting for 85-90%. The short digestion reveals easily accessible chromatin and the long digestion shows resistant nucleosomes and intra-nucleosomal digestions.

The di- and mono-nucleosome fractions of the short digestion are still long enough to contain intact chromatosomes. The typical and long digestion lead to complete digestion of the linker regions. With longer digestion the mono-nucleosomes are further shortened to sub-nucleosome fragment lengths.

To analyze the influence of formaldehyde cross-linking on MNase digestion, I looked at a typical digestion of cross-linked chromatin. The cross-linking slows down the MNase digestion, with around 30% of the chromatin digested into mono-nucleosomal fragments.

To determine MNase's sequence preference independent of nucleosomes, I analyzed a brief digested (30 seconds) and a typical digestion (3 minutes) of naked genomic DNA (gDNA) (Figure 8.1B). The brief digestion produced a broad size distribution of fragments from which ~150 bp long fragments were selected to mimic mono-nucleosomal fragments. The typical digestion produced short fragments ranging from 50 to 75 bp.

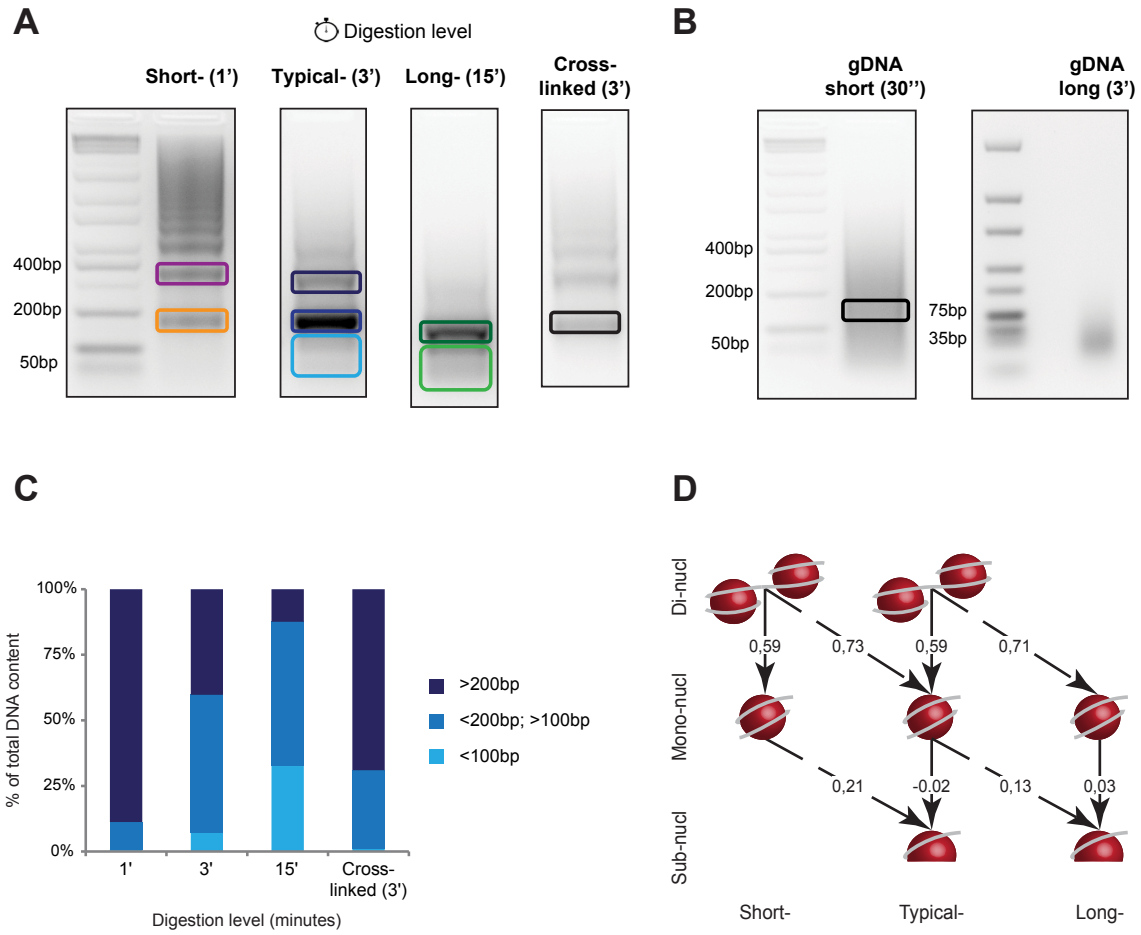

Figure 8.1: **MNase digestion is a continuous process**: (A) Length separation of different MNase-digestion levels of chromatin. The marked boxes indicate the samples that were extracted, sequenced, and analyzed. (B) Length separation of digested naked genomic DNA. (C) Quantified amount of DNA in fragments of mono-nucleosome length (100-200 bp) and longer/shorter fragments for the chromatin digestions. (D) Correlation coefficients between the genome-wide occupancies derived from the MNase-Seq samples marked in (A), the arrangements match.

8.1 Possible experimental biases of MNase-Seq

I analyzed the different MNase-digestion datasets and control experiments to learn more about the biases and errors present in MNase-Seq data. Two types of problems are present: positional errors and quantitative errors. I will ignore quantitative noise, because it does not create a problem for learning a nucleosome binding energy model.

The sequence preference of MNase leads to positional biases depending on the local sequence. My nucleosome-position prediction method handles this issue with a sequence-independent positional uncertainty (Section 12.2). The method can be extended to incorporate the sequence-dependent bias, but a better understanding would help choose the best approach.

As I discuss in Section 13.1 MNase-Seq measurements do not represent quantitative occupancies. I analyzed which biases lead to these errors and if control experiments could reduce them. In a parallel project I tried to correct these biases without further experiments based on the data and sequence information alone (Section 17). After realizing that the biases are difficult to correct, I modeled the MNase digestion of chromatin to gain further insights (Section 18). I mention this here, because the results of these analyses are relevant to the results discussed here.

8.1.1 Sequence bias at the MNase cut site

As mentioned in Section 3.1, MNase has a sequence preference to cut in the center or TA and other WW dinucleotides (Fan et al., 2010). The position weight matrices (PWMs) of the genomic DNA (gDNA) digestion I generated match the described preference (Figure 8.2A). The information content decreases between the short and the typical digestion levels, which is consistent with the notion that MNase first cuts favored sequences.

The sequence bias of MNase is reduced in the digestion of chromatin compared to gDNA (Figure 8.2A). The occlusion of most of the genome by nucleosomes reduces MNase's choices and thereby the influence of its sequence preference. While the consensus sequence matches between all datasets, the relative base contributions change, and the information content decreases as a function of fraction size and digestion time (Figure 8.2A).

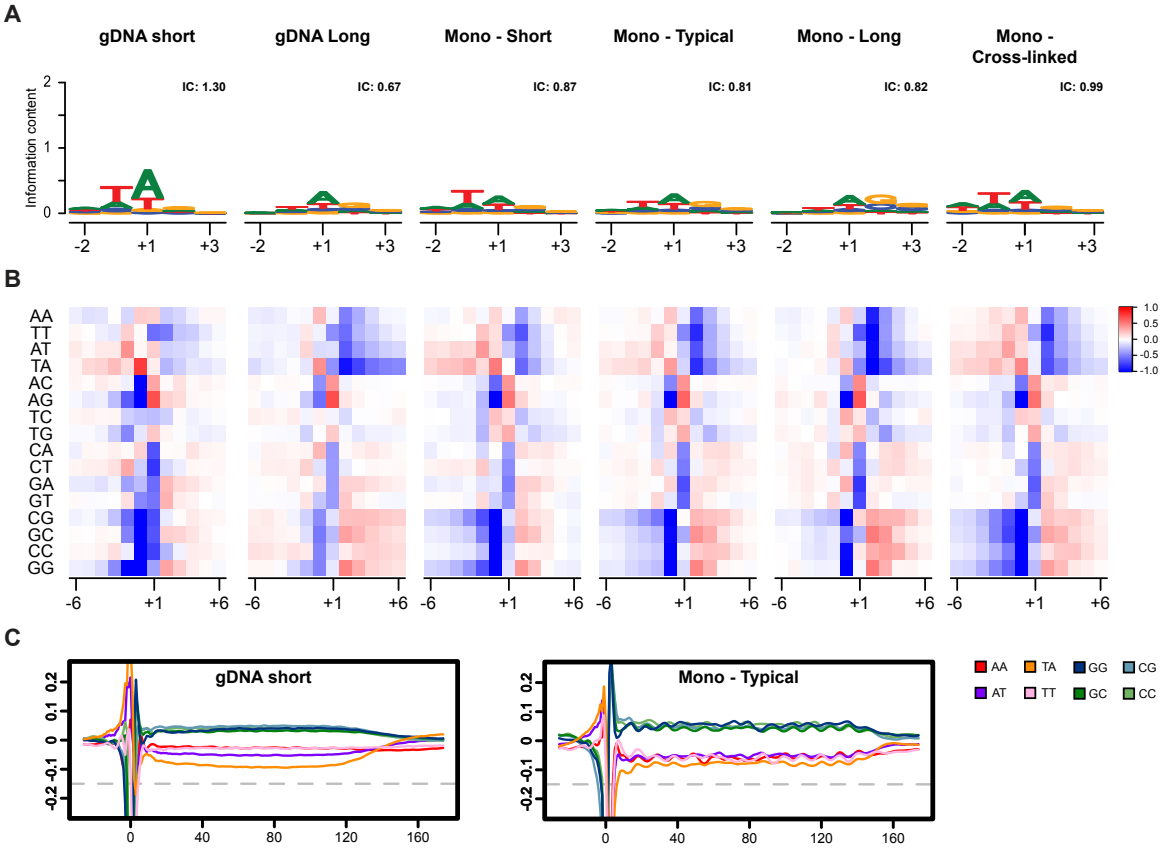

Figure 8.2: **MNase produces sequence biases at the nucleosome borders**: (A) PWMs of the MNase cut site for the different samples. The sequenced fragments begin at position +1. The sum of the depicted PWMs information content (IC) is shown in the corner of the individual panels. (B) Dinucleotide enrichments around the MNase cut sites. The color scale shows the log fold change compared to the average frequency of the dinucleotide genome-wide. (C) Smoothed dinucleotide enrichments over the fragment region aligned by the left MNase cut site – position 0. The left panel shows the gDNA control, which has an enrichment of SS dinucleotides, but no periodicity. The right panel shows the commonly used MNase-Seq sample, it has both an SS enrichment and the typical 10-bp-periodic pattern described for nucleosomes.

8.1.2 Sequence bias around the MNase cut site

Earlier work has shown that the cut frequency of MNase is influenced by a larger sequence content than the neighboring nucleotides (Hörz and Altenburger, 1981; Dingwall et al., 1981). I analyzed the dinucleotide enrichments in the region surrounding the cut sites to gain further insight (Figure 8.2B). The dinucleotide frequencies around the cut site are similar in the gDNA and chromatin digestions. The preferred sequence environment is different upstream and downstream of the cut site and with longer digestion the emphasis changes from upstream to downstream preferences.

Upstream (≤ -2) of the cut site TA, AT and CT are the most enriched dinucleotides and downstream ($\geq +2$) all WW are depleted and SS are enriched (Figure 8.2B). Together with the know pseudo-exonuclease activity of MNase this suggests a continuous digestion of A+T-rich sequences that halts at a G+C-rich border.

8.1.3 Differential digestion of linkers

Digested linkers have a different sequence composition than undigested linkers. The cut and uncut linkers of di-nucleosome fragments have a clear difference of TA dinucleotide frequencies and a weaker difference of all WW and SS nucleotides. This difference is present in both the short and typical digestion (Figure 8.3B).

In the mono-nucleosomal fractions the same reduction of WW (especially TA) and increase of SS is visible with increased digestion levels (Figure 8.3C). Mono-nucleosomes digested from the chromatin first are surrounded by more A+T-rich linkers. This suggests that the sequence bias influences how quickly a nucleosome appears in the mono-nucleosome fraction. Therefore, approximating chromatin accessibility by comparing the mono-nucleosome fractions of different digestions (Chereji et al., 2015; Mieczkowski et al., 2016) is confounded by the sequence bias.

In di-nucleosomal fractions the undigested linkers are flanked by two nucleosome halves with higher G+C content, while the digested linkers are flanked by nucleosome halves with lower G+C content. This suggests that unwrapping of the nucleosome, or something similar, makes linkers more accessible to the MNase digestion (Figure 8.3B).

8.1.4 Nucleosomal DNA is not lost during MNase digestion

A common assumption of analyzing MNase-Seq data is that nucleosome bound chromatin is conserved and only little nucleosomal fragments are digested away. To my knowledge, nobody has tested this assumption extensively. Two main processes could break this

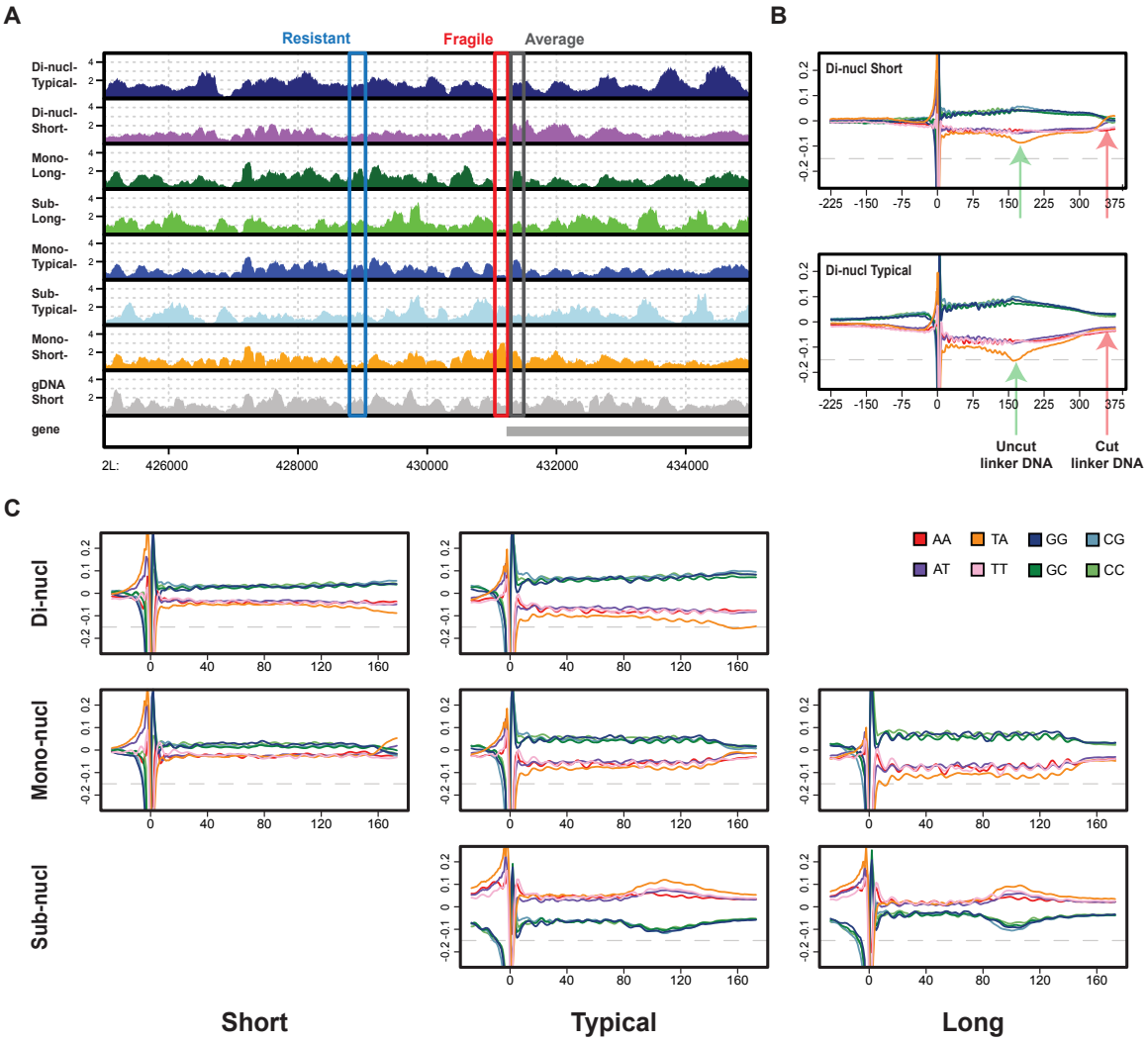

Figure 8.3: **MNase-Seq samples consist of different compositions of nucleosome populations**: (A) Genomic tracks of the measured MNase-Seq samples in an example region of chromosome 2L. Common profiles for resistant, fragile, and average nucleosomes are marked. (B) Smoothed dinucleotide enrichments over the di-nucleosome fragments, showing divergent features between the undigested and digested linker regions. (C) Smoothed dinucleotide enrichments over the nucleosome regions (Figure 8.2C). The panel arrangement matches the sample arrangement in Figure 8.1A. The gray dashed lines are an aid to better visualize the enriched and depleted sequence features, which depend on the digestion level and extracted fragment lengths.

assumption: MNase nicking single strands of nucleosomal DNA with a subsequent loss of such fragments during library preparation, and nucleosomes unwrapping completely during the MNase digestion.

MNase cuts one DNA strand at a time (Cockell et al. (1983) and Section 3.1) and could therefore produces single-strand cuts inside nucleosomal DNA. Based on my concern, my collaboration partners analyzed all chromatin samples in a denaturing alkaline-agarose gel. The DNA doublestrand is separated in the gel, which reveals the presence of single strand nicks as an enrichment of shorter fragments. None of the analyzed samples have a different size distribution under denaturing conditions, proving that nucleosome fragments containing single nicks are rare in chromatin digestions (Figure A.1).

The complete unwrapping of nucleosomes is a concern, because nucleosomes can unwrap *in vivo* and the digestion to mono-nucleosomes could ease unwrapping by allowing the DNA to move more freely. Cross-linking the chromatin should prevent the complete digestion by covalently connecting the histones with the DNA. While the sequence enrichments have minor differences, nothing suggests that the cross-linking prevents large scale unwrapping of whole nucleosomes. Therefore, the effect of whole nucleosomes unwrapping must also be weak in native (not cross-linked) samples.

The most interesting difference is that the WW and SS 10-bp periodicity is more pronounced in native compared to cross-linked chromatin (Figure A.3C). This could result from nucleosomes re-adjusting in native chromatin to find their most favorable rotational position during preparation and MNase digestion. However, the effect can stem from other sources that improve the fragment ends alignment to the nucleosome ends (or a 10-bp shift thereof). To determine the true source further experiments would be needed.

8.1.5 High correlations between chromatin and genomic DNA measurements

Short digestion of naked gDNA with a size selection of ~150 bp correlate highly with mono-nucleosomal fractions, both at the single locus level and genome-wide confirming previous studies (Chung et al., 2010). In Section 13.1 I quantitatively compare the correlations between different yeast chromatin measurements and between them and gDNA measurements. The correlations look similar for yeast and *D.melanogaster*. Cross-linked mono-nucleosomes have the highest correlation with 0.65 and sub-nucleosome of the typical digestion have the lowest correlation with 0.14 (Figure A.4). Genome-wide correlations provide evidence for systematic biases of MNase-Seq, but they can mask important de-

tails. For instance, the nucleosome pattern of BP promoters are phase shifted against the pattern obtained from the short gDNA digestion (Figure A.5). The promoter is depleted of nucleosomes, but enriched for gDNA and conversely the -1 nucleosome peak aligns with a minimum in the gDNA. The amplitude of the downstream nucleosomal pattern is weaker in the gDNA and shifted to nearly represent an anti-correlated pattern.

Both digested chromatin and gDNA correlate with G+C content as I also discuss in Section 13.1. As with the genome-wide correlations, the general trend suggests that biases are involved, but important details are masked. In contrast to chromatin, gDNA shows no 10-bp periodicity in its WW and SS dinucleotide frequencies (Figure 8.2C). The lack of periodicity is not surprising – there is no reason it should be present, but it proves that MNase-Seq measures something different – i.e. nucleosomes – in chromatin compared to gDNA.

8.2 Genome-wide nucleosome populations

The distinct properties of the different fragment lengths (Figures 8.2 and 8.3) reflect the non-homogenous digestion of the chromatin. Based on the Pearson correlation coefficients between the samples, nucleosome populations move to shorter fragment lengths with increasing digestion levels (Figure 8.1D). For instance, the sub-nuclesomal fraction of the typical digestion has a correlation of -0.02 with the mono-nucleosomal fraction of the same digestion, but a correlation of 0.21 with the mono-nucleosome fraction of the short digestion.

Accessible nucleosomes move faster from the oligo-nucleosomes fractions to shorter ones and unstable nucleosomes move faster from the mono-nucleosome fraction to the sub-nucleosome fraction. The movement of the nucleosome populations between fractions during digestion changes their composition in the mono-nucleosome samples. Depending on the level of digestion a different cross-section is measured, which leads to changes in the nucleosome coverage (Figure 8.3A).

To understand which features drive the populations behavior, I first compared sequence features of the fractions. Matching fractions of increasing digestion levels have a stronger 10-bp periodicity and an increased G+C content (Figure 8.3C). Both features are thought to promote nucleosome formation, therefore the nucleosome binding energy appears to partially distinguishes the populations.

The idea of capturing nucleosome populations by comparison of fractions between different digestion levels is not new (Weiner et al., 2010; Xi et al., 2011; Chereji et al.,

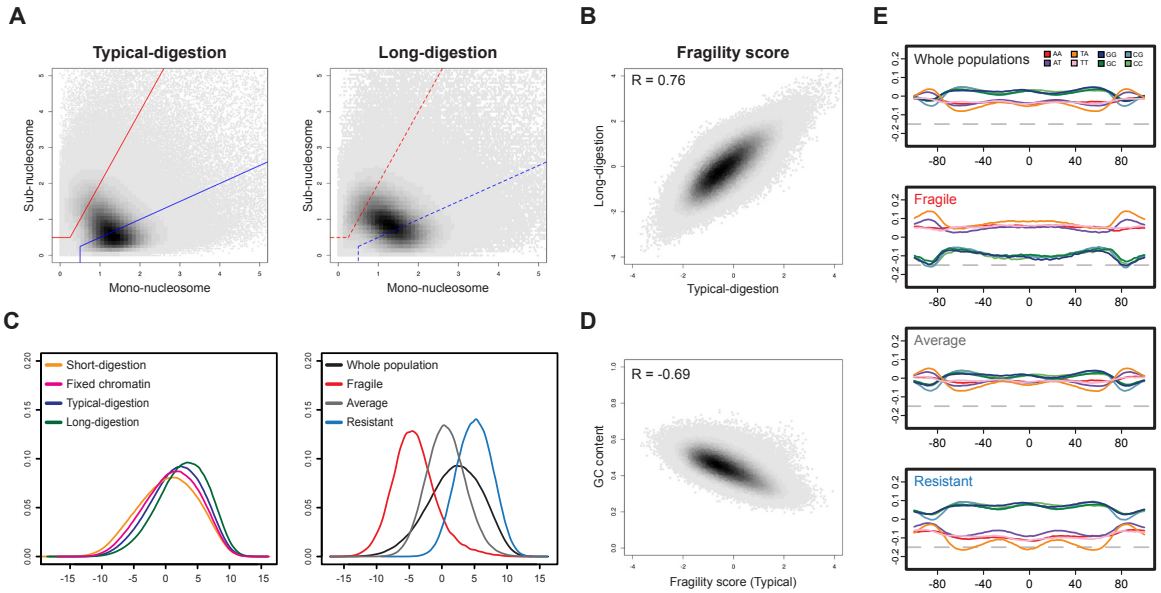

Figure 8.4: **Fragile, average and resistant nucleosome population have different sequence features**: (A) Scatterplot between the mono- and sub-nucleosome coverage of the positioned nucleosomes. The left panel shows the figure for the typical digestion level, which was used to define populations by the red and blue threshold lines. (B) Scatterplot showing the strong correlation between the fragility score based on the typical and long digestion levels. (C) Sequence-feature score (1st-order Markov model) distributions of different nucleosome groups. The left panel shows the distributions for the different mono-nucleosome samples, which are very similar. The right panel shows the distributions for the nucleosome populations separated by their fragility. The nucleosome populations have distinct average scores and therefore sequence features. (D) Anti-correlation between the nucleosome fragility and the average G+C content of the nucleosome region. (E) Smoothed dinucleotide enrichment profiles of the nucleosome populations aligned by the nucleosome dyad.

2015; Mieczkowski et al., 2016). I computed an accessibility score by comparing the mono-nucleosome fractions between digestion levels, but neglect the score in the further analysis, because such scores were the focus of previous studies. I computed further scores by comparing fractions of the same digestion level, one of which I based the definition of the nucleosome populations on. To focus on nucleosome unwrapping – and not accessibility – I compared the mono-nucleosome to sub-nucleosome fractions. With this fragility score, I categorized nucleosomes into three populations: fragile, average and resistant. I analyzed these populations to understand if they have different biological roles and what those could be.

8.2.1 Definition of the populations

For the context of this work, I define nucleosome fragility as the ease with which a nucleosome is over-digested by MNase. This ease of over-digestion is an approximation of the probability for the nucleosome to unwrap. The fragile nucleosome population is the first to be over-digested and is thus enriched in the sub-nucleosome fraction compared to the mono-nucleosome fraction (Figure 8.4A, red border in the left panel). The resistant nucleosomes population is the last to be over-digested and is thus enriched in the mono-nucleosome fraction compared to the sub-nucleosome fraction (Figure 8.4A, blue border in the left panel).

$$\begin{aligned} \text{fragile} &: \frac{\text{sub-nucleosome}^{\text{typical}}}{\text{mono-nucleosome}^{\text{typical}}} > 2 \\ \text{resistant} &: \frac{\text{mono-nucleosome}^{\text{typical}}}{\text{sub-nucleosome}^{\text{typical}}} > 2 \end{aligned} \tag{8.1}$$

I computed the ratios for both the typical and long digestion (Figure 8.4A) and they are similar (R=0.76; Figure 8.4B) goven the distinct digestion levels. Further analyses produced similar results for both. As such, a one-pot reaction of typical digestion with a bead-selection step (<200 bp) is sufficient to compute a fragility score and define the populations. Average nucleosomes are defined by being neither fragile nor resistant. Using these thresholds the analyzed nucleosomes are composed of 7% fragile, 49% average and 44% resistant.

8.2.2 Sequence characteristics of the nucleosome populations

I first characterized the three populations by their sequence features (Figure 8.4E), because the fractions already showed differences. The 10-bp periodicity is not visible due to aligning the nucleosomes based on their called dyad positions instead of the fragment

borders. In fragile nucleosomes WW is enriched and SS is depleted, in resistant nucleosomes the situation is reversed, while the dinucleotide frequencies of average nucleosomes fall between them. To asses the variation within the populations I computed a sequence-feature score for the individual nucleosomes. To score the nucleosomes I used a 1st-order (dinucleotide) Markov model from the mono-nucleosome fragments of the typical digestion. I trimmed 15 bps from either side to avoid including the bias at the cut sites. A high score reflects a higher G+C content and higher SS and WW periodicity.

The score distributions of the populations have distinct averages (Figure 8.4C, right panel), matching the average dinucleotide profiles (Figure 8.4E). The fragile population is markedly shifted towards lower scores, while the resistant population is shifted towards higher scores. The fragile and resistant population have little overlap. In comparison, the score distributions of the mono-nucleosome fractions of different digestion levels strongly overlap, because they contain a mixture of all three populations in slightly different proportions (Figure 8.4C, left panel). Together with the high anti-correlation of the fragility score with the G+C content (Figure 8.4D) the influence of sequence features on nucleosome fragility is undeniable.

The fragility as measured by MNase over-digestion might be a result of sequence biases described above. I could not conclusively disprove this possibility, but the further analyses show that fragile nucleosomes have biological roles, both when correlating and when anti-correlating with their preferred sequence features. This means that a major part of the score and populations is relevant signal, even if the used definition of fragility is probably influenced by MNase's biases.

8.3 Nucleosome populations at promoters

In previous work, I called transcription start sites (TSS) of promoters from CAGE data and distinguished between broad peaked (BP) and narrow peaked (NP) core promoters. BP and NP promoters have different sequence features and nucleosome coverage profiles. In BP promoters the regions surrounding the TSS have a higher WW content, while in comparison NP promoters have lower WW and higher SS content (Figure A.6). I generated composite plots of the nucleosome populations. In BP promoters fragile nucleosomes are enriched in a ~500 bp region upstream of the TSS, while resistant and average nucleosomes dominate the gene body downstream of the TSS, where the G+C content is on average higher. Separating BP promoters into expression quartiles reveals that on average only the higher expressed BP promoters have the pronounced nucleosome array, while

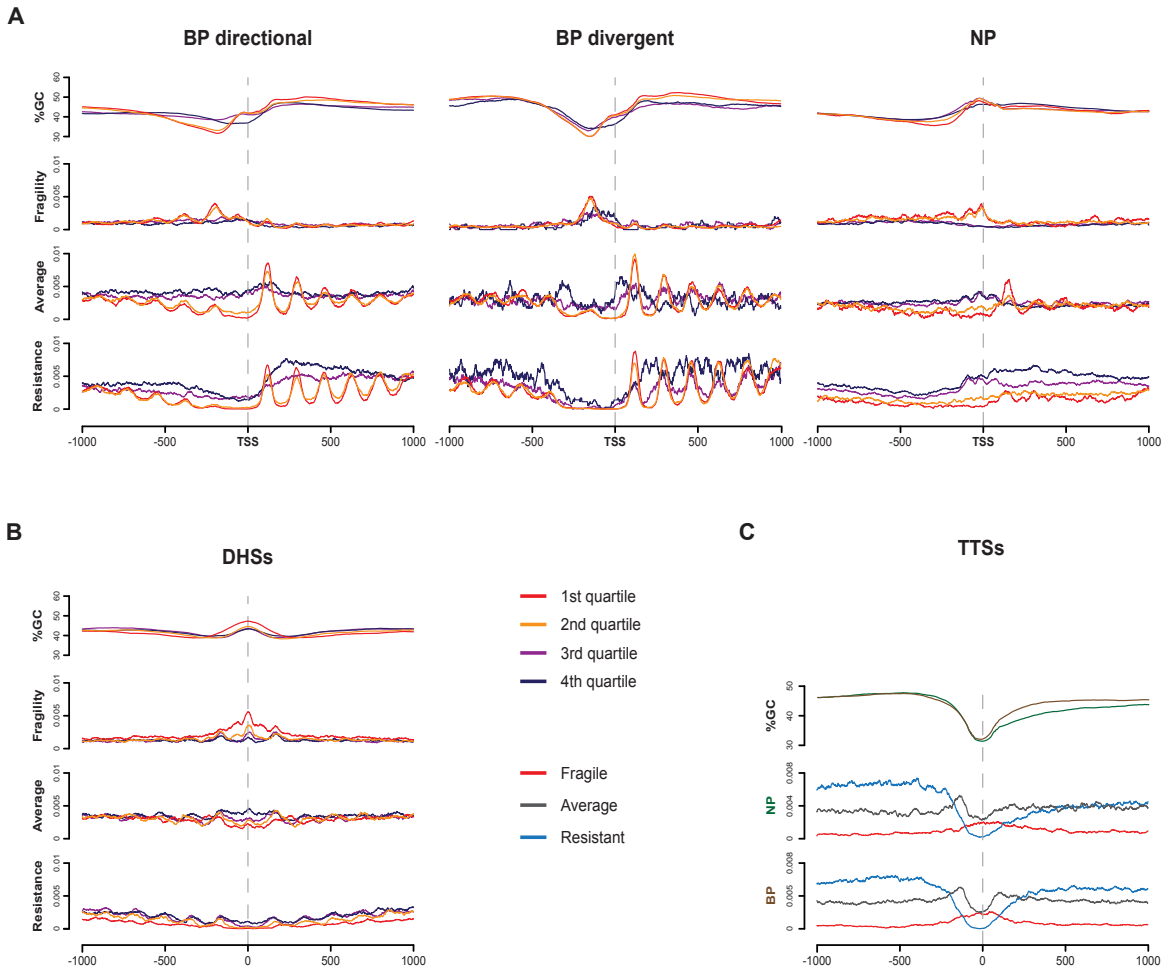

Figure 8.5: **Nucleosome populations around promoters, TTS and enhancers**: (A) Smoothed profiles of the nucleosome populations around BP directional (left), BP divergent (central), and NP promoters (right). The top panel shows the G+C content profile, followed by the profiles of fragile, average and resistant nucleosomes. Promoters belonging to the four gene expression quartiles are shown as individual lines. (B) The same figure as (A) surrounding DHSs and the quartiles are based on the fold-enrichment of the DHS peaks. (C) Smoothed profiles of the G+C content and nucleosome populations around TTS.

the lower expressed promoters have a weak – if any – nucleosomal array (Figure 8.5A).

The G+C content of the directional BP promoters is distinct between the quartiles: the G+C content minimum ~150 bp upstream of the TSS is lower in the first and second quartiles than the third and fourth quartiles (Figure 8.5A, left panel). The G+C content minimum corresponds to the position of the fragile -1 nucleosome in the first and second quartiles. This suggests that the fragility is at least partially encoded in the sequence upstream of the TSS and increases the promoters' expression.

The DNA landscape of the first and second expression quartile of divergent BP promoters is slightly more pronounced than that of the directional BP promoters, as is the fragile -1 nucleosome (Figure 8.5A, middle panel). In the third and fourth quartiles the G+C content minimum is lower and larger in the divergent BP promoters than in the directional BP promoters. This coincides with the presence of a fragile -1 nucleosome in these quartiles.

The fluctuations in G+C content are weaker in NP promoters than in BP promoters (Figure 8.5A, right panel). Promoters of the first and second quartiles have more pronounced G+C content profiles and show an enrichment of fragile nucleosomes upstream of the TSS. Resistant nucleosomes are enriched on and downstream of the TSS in promoters of the third and fourth quartiles.

8.4 Nucleosome populations at TTSs

The local region (±50-100 bp) around the transcription termination sites (TTS) of both BP and NP genes are depleted of G+C content (Figure 8.5C). The NP genes have a flatter curve downstream of the TTS, possibly because NP genes are surrounded by longer intergenic regions on average. The TTS of both BP and NP genes have an increase in fragile and a decrease in resistant nucleosomes a the TTS compared to the genome average (Figure 8.5C).

8.5 Nucleosome populations at enhancers

To analyze nucleosome fragility at enhancers, I define them as open chromatin regions measured by DNase hypersensitivity (DHS) that are not located near TSSs or TTSs (distance >500 bp). I separated the DHSs into three groups: S2 cell specific (active), ovary stem cell (OSC) specific (inactive) and shared (active, possibly constitutive). The nucleosome datasets were measured in S2 cells, which is why active DHSs are those

present in S2 cells. G+C content is enriched at the DHS peaks (position 0) of all three groups, and longer and weaker depletions in G+C content flank the enrichment on either side (Figure A.7). In inactive enhancers (OSC unique) a resistant nucleosome covers the DHS peak, in accordance with the local peak of G+C content. In active enhancers (S2 unique and S2/OSC shared) the nucleosome at the DHS peak is fragile and the flanking nucleosomes are partially fragile.

As an alternative definition of enhancer activity, I separated the DHSs into quartiles based on the accessibility of the DHSs in S2 cells. Higher accessibility leads to a more fragile central nucleosome (Figure 8.5B). The most accessible quartile has a higher G+C content peak, which should lead to more resistant nucleosomes without other influences. The results are all consistent with activity-driven processes being the predominant component to produce fragility at DHSs.

8.6 Partial nucleosome unwrapping

With prolonged digestion times the mono-nucleosome fragments are digested down to sub-nucleosome fragments. Given that this occurs long after gDNA would be digested to sub-nucleosome fragments these fragments mostly stem from nucleosomes. The nucleosomal DNA could be over-digested while bound or while briefly unwrapping. Neither explanation could be fully rejected, but the experimental data described here strongly favors the unwrapping theory.

Probably, MNase is able to cut within the nucleosome due to temporary unwrapping of the DNA (Li et al., 2005) and the small size of the enzyme. To characterize MNase cleavage within the nucleosomal DNA on a genomic level, I analyzed MNase-Seq data without a size selection between mono- and sub-nucleosomes before sequencing. To match the digestion levels, native chromatin was digested for 9 minutes and cross-linked chromatin for 15 minutes.

8.6.1 Sub-nucleosomes do not stem from single strand nicks

My collaboration partners checked if single strand nicks are a common phenomenon by separating the double strand DNA and measuring the fragment lengths via gel electrophoresis. For none of the size selected fractions described above a visible band of shorter fragments appeared. Given that the long digestion produces a visible sub-nucleosome band, single strand nicks must be much less frequent than double strand cuts. This suggests that MNase has access to both strands and nicks them in quick suc-

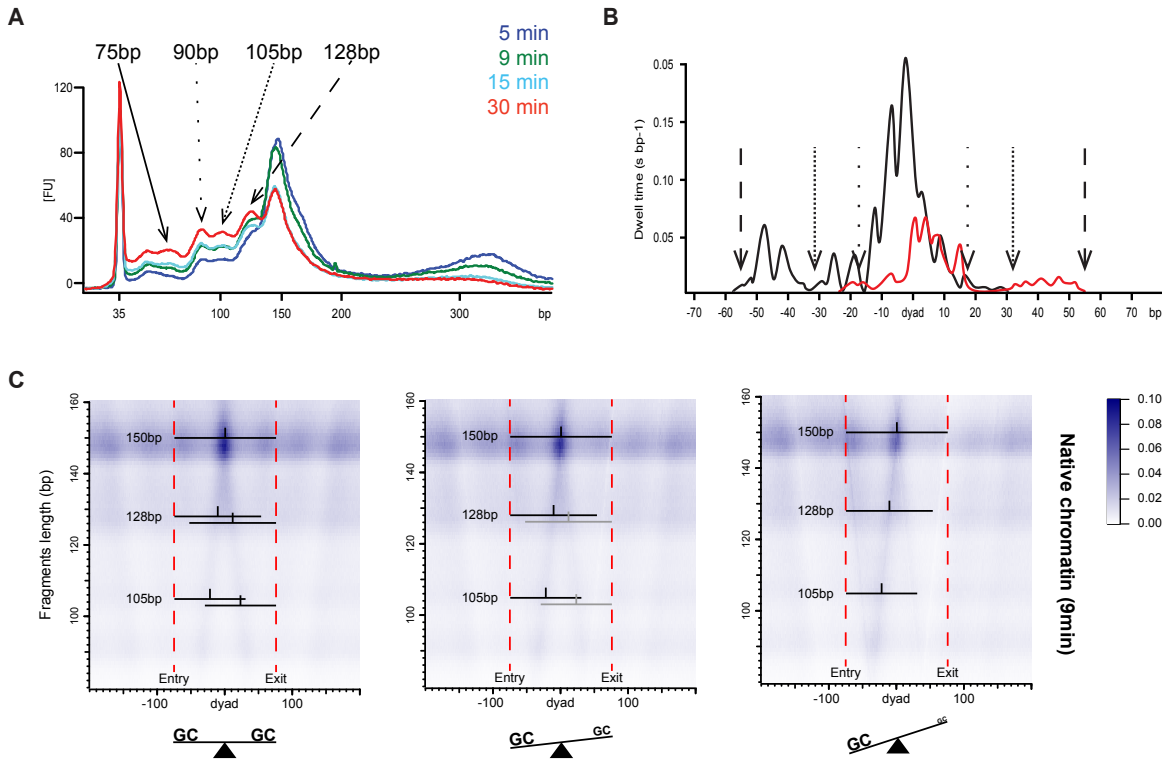

Figure 8.6: **Nucleosomes unwrap asymmetrically**: (A) Bioanalyzer profiles of the fragment lengths for different MNase-digestion levels. (B) Reproduction from a figure by Hall et al. (2009). Profile of the dwell times during mechanical unzipping for the 601-Widom sequence. (C) V-plots showing the fragment-center frequency separated by fragment length and position. For an even G+C-content ratio (left) the nucleosome unwrapping occurs at either side leading to an inverted V pattern. An asymmetric distribution appears with an asymmetry of G+C content (middle) and increases with the G+C asymmetry (right).

cession during the over-digestion, which means that unwrapping or another mechanism must expose the nucleosomal DNA.

8.6.2 Distribution of sub-nucleosomal fragments

The distribution of sub-nucleosomal fragments reveals enriched lengths. There are peaks around fragment lengths of 146, 128, 105 and 90 bps, and a broader peak around 75 bps (Figure 8.6A and Figure A.2B). The pattern is reproducible across different digestion levels, and in cross-linked chromatin (Figure A.8). The ratios of the 128, 105, and 90 bp peaks appear robust across digestion levels in native chromatin. The constant ratio suggests that there is little flow between the peaks and most of these shorter fragments are direct results of cut mono-nucleosome fragments. In cross-linked chromatin the 105-bp peak is higher than the 90-bp peak, which are of equal hight in native chromatin. This suggests that fixation reduces the unwrapping probability close to the nucleosomal dyad.

8.7 Asymmetry of the partial nucleosome unwrapping

The digestion inside the nucleosome proceeds in an asymmetric fashion, in which the nucleosome is digested from one or the other side. This becomes evident by separating the sub-nucleosomal fragments by length and plotting their centers around positioned nucleosomes (Figure 8.6C). Henikoff et al. (2011) visualized digestion fragments around positions of interest with this so-called V-plot. With decreasing fragment size the nucleosome fragments bifurcate into two populations: one in which the entry site remains intact and the exit site is digested, and the other in which the exit site remains intact and the entry site is digested.

Given the strong correlation of the fragility score to G+C content I investigated the relationship further. I analyzed the influence of the G+C content of the two nucleosome halves, and found that the half with the higher G+C content is less likely to unwrap. I divided well positioned nucleosomes by the G+C ratio between their two halves: 1) equal ratio, 2) slightly asymmetric, and 3) highly asymmetric. For each group, I created the V-plot to visualize where the sub-nucleosome fragment centers fall in relationship to the nucleosome dyad. When the G+C content is similar between the two halves no preference is visible between the sides (Figure 8.6C, left panel). When the G+C content is

asymmetric the half with lower G+C content preferentially unwraps and a stronger skew leads to a stronger preference (Figure 8.6C middle and right panels). The preference is weaker when repeating the analysis with absolute G+C content instead of the ratios between the halves.

This proves that the G+C ratio influences the unwrapping and the underlying mechanic might be the ratio between the binding energy of the two nucleosome halves. I tried to further investigate if a role of the 10-bp periodicity could be observed genome-wide, but failed due to interferences of the MNase-sequence bias in the visualizations and analysis. In any case, my findings confirm that asymmetric unwrapping of nucleosomes is a genome-wide phenomenon *in vivo* and suggests that DNA-nucleosome binding energies are one determinant of the asymmetric unwrapping.

9. Discussion

The MNase digestion is a continuous process and MNase-Seq therefore does not capture the whole picture of the nucleosome occupancy landscape. While we are still lacking knowledge to fully model the digestion process, this collaboration investigated several aspects in detail and systematically eliminated possible bias sources. With this work and further ground work of this kind, simulations of the MNase digestion can improve upon my simulation (Section 18).

I analyzed accessibility of nucleosomal DNA and nucleosome-stability populations around genomic features. I computed a fragility score from the mono- and sub-nucleosome fractions, which were originally measured to investigate the MNase-Seq biases. By analyzing nucleosome fragility and resistance around promoters and enhancers, I determined that both are part of the transcription regulation. I revealed that DNA sequence features and active components are used in different contexts to affect nucleosome stability. Their intertwined relationships require further investigation, especially in the context of understanding the regulatory dynamics.

Lastly, I investigated nucleosome unwrapping genome-wide, revealing that asymmetric unwrapping is common and depends on the G+C ratio between the nucleosome halves.

9.1 Nucleosome accessibility and fragility scores

I propose a fragility score that is less influenced by higher-order accessibility than similar scores used in other studies that compare MNase-digestion levels (Kubik et al., 2015; Chereji et al., 2015; Mieczkowski et al., 2016). I defined an accessibility score in a similar fashion: $^{\text{mono-nucleosome}^{\text{short}}}/_{\text{di-nucleosome}^{\text{short}}}$. This ratio indicates how quickly the nucleosomes are isolated from the chromatin and is correlated with the fragility score. There are important differences between the two, such as the accessibility of the +1 nucleosome due to its proximity to the nucleosome-depleted region. Another proposed accessibility score uses mono-nucleosome fractions and produces similar patterns (my version:

$_{\text{mono-nucleosome}^{\text{short}}}/_{\text{mono-nucleosome}^{\text{long}}}$). Chereji et al. (2015) combined the two approaches and compared mono- and di-nucleosomal fractions of different digestion levels. Base on my analysis, all of these scores contain more information about accessibility than my fragility score.

I think a score focusing on accessibility is less interesting when analyzing nucleosomes for two reasons. First, the measured accessibility is not the accessibility of the nucleosomal DNA, but the accessibility of the flanking linkers and the nucleosome as a whole. This is generally less interesting than the accessibility of the DNA to transcription factors and other DNA-binding proteins. DNase-Seq and ATAC-Seq are better methods to measure the accessibility of open chromatin to DNA-binding proteins. Second, assuming one wants to measure this kind of nucleosome accessibility for factors, it is unclear how well the accessibility to MNase digestion represents the accessibility for other factors. For example, different nucleosome remodelers bind different parts of the nucleosome and often only come in contact one of the two linkers. In comparison, MNase has to digest both linkers and does not interact with the nucleosomal DNA at all.

A score focusing on nucleosome fragility is more interesting. Fragility, as defined here, is the probability for a nucleosome to unwrap, which directly represents accessibility of nucleosomal DNA. This also correlates with the nucleosome binding energy and may well influence how easy or frequent the nucleosome shifts or disassociates *in vivo*. Nucleosome turnover rates have been measured directly and comparing them with the fragility would be interesting (Deal et al., 2010). One issue of the fragility score is its contamination with MNase's sequence bias, but so are the accessibility scores described above.

9.2 Nucleosome fragility and resistance at promoters

Constitutive (BP directional, BP divergent) and inducible (NP) genes have different requirements in the regulation of their expression level and use distinct methods to generate nucleosome fragility and resistance. Some previous studies in *D.melanogaster* of fragility and accessibility neglected this important difference and concluded that nucleosome fragility is primarily activity driven at promoters (Chereji et al., 2015; Mieczkowski et al., 2016). I showed that the sequence landscape encodes nucleosome fragility in constitutive promoters, where activity has less effect on fragility.

Most BP promoters have a stable expression, allowing for nucleosomal fragility to be carved into the sequence landscape of the promoter. With higher expression rates

the G+C content upstream of the TSS is lower, which predisposes more fragile and less resistant nucleosomes at the promoter – independent of activity-driven processes. From an evolutionary standpoint different expression rates are unlikely to cause noticeable effects on the DNA sequence, thus the G+C content regulates the basal transcription rate. The fragility is carved more strongly into divergent BP promoters, where two genes share a common -1 nucleosome.

The +1 nucleosome in BP promoters has a built-in asymmetry: the end closer to the TSS is G+C-poor compared to the distal end. The nucleosome is expected to preferably unwrap towards the proximal end due to the sequence. An enrichment of sub-nucleosomes at the distal half of the +1 nucleosome confirms this expected asymmetric unwrapping. These results support the proposition that asymmetric unwrapping might contribute to the transcription directionality by helping the polymerase overcome nucleosomes (Ngo et al., 2015). The +1 nucleosome forms a barrier three times stronger than successive nucleosomes and thus the biggest roadblock (Teves et al., 2014).

NP promoters have a uniform G+C landscape around the TSS, which is similar between expression rates. However, nucleosome fragility and resistance differs between expression rates. The TSS region, which has a local increase of G+C content, is covered by resistant nucleosomes when the transcription rate is low or off. In comparison, for higher transcription rates the TSS region is covered by fragile nucleosomes, which have to predominantly be created by activity-driven processes.

Other organisms have similar differences between constitutive and inducible promoter types. Constitutive promoters in yeast have a nucleosome-depleted region created by nucleosome-positioning signals like poly(dA:dT) stretches, while inducible promoters are depleted upon activation (Field et al., 2008; Cairns, 2009). At mammalian promoters of house-keeping genes, CpG islands and G+C content dictate nucleosome depletion in a transcription-independent manner (Fenouil et al., 2012). Many organisms encode nucleosome depletion or instability with sequences features in constitutively expressed promoters.

9.3 Nucleosome fragility and resistance at enhancers

At enhancers we observe increased G+C content surrounded by A+T-rich regions. This architecture – somewhat representing NP promoters – favors the presence of nucleosomes covering the transcription-factor binding sites (Gaffney et al., 2012), and relies on active

remodeling to open the chromatin. In agreement, more resistant nucleosomes cover inactive enhancers, while active enhancers are covered by fragile nucleosomes, which are less competition for transcription factors. Depending on the enhancer, the fragile nucleosomes can stem from different active mechanisms, such as competing pioneer factors, histone modifications or recruitment of chromatin remodelers.

9.4 Nucleosome unwrapping

Most of the intra-nucleosomal cuts are around the ±55, ±32 and ±17 bp positions from the nucleosome dyad, which roughly matching every other 10-bp period. The same positions were identified in a biophysical experiment in which a nucleosome bound to the 601-Widom sequence was mechanically unzipped (Hall et al., 2009). These positions show a longer dwell times during the unzipping (reproduced in Figure 8.6B), indicating that they are adjacent to regions with stronger DNA-histone interaction, which could function as borders for unwrapping. This affirms that the sub-nucleosome fragments result from unwrapping and hence provide new information about nucleosomes *in vivo*.

By analyzing the sub-nucleosome fragments, I found that asymmetric nucleosome unwrapping is a genome-wide phenomenon *in vivo*. A recent study using single molecule FRET had shown such asymmetric nucleosome unwrapping *in vitro* for a few selected sequences (Ngo et al., 2015). In their measurements, DNA features of the two halves, such as 10-bp periodicity of the TA dinucleotide, play a crucial role in determining the preferred side. Similarly, I found that a G+C asymmetry was a major contributor to asymmetrical unwrapping.

Part III

Learning nucleosome binding energies and modeling nucleosome positioning

10. Abstract

The physical task of wrapping 147 bps around the histone core of a nucleosome leads to characteristic sequence preferences. These preferences can be learned from genome-wide nucleosome position measurements. Unfortunately, sequence biases and positional errors of the experimental techniques can lead to biased and erroneous energy models.

I developed a maximum likelihood approach to learn a biophysical model of nucleosome binding. I maximize the likelihood of measured nucleosome positions under a thermodynamic model with steric hindrance between nucleosomes. By including the low positional resolution of MNase-Seq and the sequence bias of CC-Seq into the likelihood, I can separate them from the nucleosome binding preferences. With this approach I learn highly correlated nucleosome binding energy models from the two measurements despite their distinct experimental biases. My nucleosome-position predictions correlate better than previous predictions to experimental measurements at single-base-pair resolution.

My analysis shows that nucleosomes have a position-specific binding preference, and the described 10-bp-periodic dinucleotide enrichments are a smoothed version of this preference, which is obtained due to the low positional resolution of MNase-Seq. The optimized CC-Seq energy model has a negative correlation with G+C content suggesting that nucleosomes might disregard or disfavor G+C content contrary to the consensus in recent literature. Further evidence supports this possibility and my analysis of published datasets suggests that the common praxis of deriving occupancies from the measurements without correcting for the experimental biases and uncertainties severely affects the analysis. To fully understand what influences nucleosome binding we will have to combine experimental-error models with better thermodynamic models or measure nucleosome occupancy more quantitatively.

11. Introduction

Eukaryotic genomes are packed into chromatin. The basic unit of chromatin is a nucleosome – 147 bps of DNA wrapped around a histone octamere. The positioning of these nucleosomes regulates genome accessibility. Most transcription factors can only bind open regions of the genome, which are unoccupied by nucleosomes. The DNA sequence of promoters regulates gene expression with nucleosome binding preferences and the occurrence of transcription-factor binding sites (Raveh-Sadka et al., 2012). To decode the complex interactions at promoters and quantitatively predict gene expression from sequence alone, we need to understand the sequence features nucleosomes preferentially bind to (Segal and Widom, 2009; Iyer, 2012).

The primary influences of *in vitro* nucleosome formation are the bendability of the DNA sequence and steric hindrance between neighboring nucleosomes. There were disagreements about how much *in vivo* nucleosome formation depends on the DNA sequence and how strong other factors control nucleosome positioning (Segal et al., 2006; Kaplan et al., 2009; Zhang et al., 2009; Tillo and Hughes, 2009; Stein et al., 2010; Locke et al., 2010; Kaplan et al., 2010b; Zhang et al., 2010; Chung et al., 2010). One reason for the alternative interpretation of the same data were the definitions of 'nucleosome code' and 'nucleosome positioning'. The two camps disagreed how much information a code must contain and they interpreted nucleosome positioning in two ways: local rotational positioning (conditional positioning) and genome-wide positioning (absolute positioning – which I use throughout this work, see Section 2).

Later reviews unified the different interpretations by distinguishing between conditional and absolute nucleosome positioning and acknowledging experimental limitations (Kaplan et al., 2010a; Iyer, 2012; Struhl and Segal, 2013). Nucleosome occupancy describes the probability with which a nucleosome covers a genomic position (see Kaplan et al. (2010a) for a formal definition). *In vivo* the nucleosome occupancy results from interactions between DNA sequence, nucleosome remodelers, transcription factors, PolII elongation and steric hindrance. Nucleosomes prefer higher G+C content regions, their

positional phase matches a 10-bp-periodic pattern of SS to WW enrichment (where W is A or T, and S is C or G) and they avoid homo-polymeric sequences (both poly(dA:dT) and poly(dC:dG)). These sequence preferences are frequently trumped – primarily at promoters and enhancers – by nucleosome remodelers and competition with transcription factors.

The most common technique to measure genome-wide nucleosome binding is MNase-Seq. Micrococcal nuclease (MNase) digests the chromatin, and fragments with about the length of a nucleosome (147 bps) are isolated and sequenced. Further experiments exist to measure nucleosome binding genome-wide, e.g. MPE-Seq (Ishii et al., 2015), RED-Seq (Chen et al., 2014), NOMe-Seq (Kelly et al., 2012); at individual loci, e.g. by electron microscopy (Brown et al., 2013), by methylation (Small et al., 2014); or to short DNA fragments *in vitro*, e.g. BunDLE-Seq (Levo et al., 2015). CC-Seq (Chemical Cleavage with sequencing, also known as HC-Seq or Chemical Map) stands out amongst the genome-wide *in vivo* measurements due to its high positional resolution (Brogaard et al., 2012). A copper-chelating label is covalently bound to the genetically modified H4S47C. By adding copper and hydrogen, hydroxyl radicals form in proximity to the nucleosome dyad and cleave the DNA backbone at specific positions. Fragments spanning between two cleavage sites – i.e. nucleosome dyads – are isolated and sequenced. CC-Seq has not yet been studied extensively: it has a high positional resolution, but possible biases have largely been neglected.

MNase-Seq has been studied extensively, revealing several limitations: a low positional resolution (Kaplan et al., 2010a), an unsettling-high correlation to nucleosome-free MNase-Seq experiments (Locke et al., 2010; Chung et al., 2010) and a dependency on the chosen digestion level (Weiner et al., 2010; Rizzo et al., 2012). The low positional resolution has led to a focus on the 10-bp-periodic SS to WW preference for rotational positioning (van der Heijden et al., 2012; Struhl and Segal, 2013), while the analysis of CC-Seq data revealed stronger positional dependencies of the preferences (Brogaard et al., 2012). Correlations to nucleosome-free control experiments must stem from experimental biases. Lastly, the dependency on the digestion level reveals that – even if accounting for the above issues – the measurements cannot reflect the true nucleosome occupancy. The limitations of MNase-Seq have been downplayed for the interpretation of individual results (Kaplan et al., 2010b; Zentner and Henikoff, 2012), but when trying to understand and predict nucleosome positioning in a quantitative fashion such distortions will skew the results (Chung et al., 2010).

A variety of methods exist that predict different aspects of nucleosome formation

(Liu et al., 2014; Teif, 2015). Nucleosome-position prediction methods predict absolute nucleosome-positioning scores for every base pair of the genome. Most of them consist of two steps: learning nucleosome sequence preferences from measured nucleosome positions, and predicting the occupancies for the query sequence using a thermodynamic model that describes effects like steric hindrance. Segal et al. (2006) first published such a predictor. They obtained the sequence preferences by deriving position-specific dinucleotide enrichments from their frequencies. To compute the effects of steric hindrance between nucleosomes they employed a Forward/Backward algorithm, a type of dynamic programming algorithm. Two noteworthy extensions are accounting for competing local nucleosomes when extracting the sequence preferences in the first step (Locke et al., 2010), and adding competition with transcription factors to the thermodynamic model (Wasson and Hartemink, 2009).

Published nucleosome-position predictors are good at reproducing genome-wide MNase-Seq measurements (Kaplan et al., 2009; Tillo and Hughes, 2009). However, how close these predictions and the MNase-Seq data are to the actual nucleosome occupancy is unknown. Their usefulness in predicting individual nucleosome positions, as defined by MNase-Seq, at high resolution is also limited (Locke et al., 2010). This restricts their usefulness in decoding the complex interactions involved in nucleosome formation *in vivo*.

Here I present an improvement upon the two step approach to predicting nucleosome positioning. I extended the model to describe experimental data, distinguishing between nucleosome binding events and measurements of these events. The nucleosome binding process is still described by a thermodynamic model, while an additional layer on top models the experimental biases and positional uncertainty. By maximizing a likelihood of observing experimental data I optimize the nucleosome binding energies, experimental biases, and positional uncertainty in my model, instead of extracting them with an independent method beforehand. This combines benefits of machine learning with the biological interpretability of a probabilistic model.

Thanks to the deconvolution of MNase-Seq's low positional resolution, my method learns a high-resolution nucleosome binding energy model from MNase-Seq data, which bests the competition when validated against CC-Seq data at base-pair resolution (absolute nucleosome positioning). I show that the CC-Seq experiment has a sequence bias and my method can separate this bias from the nucleosome binding energies. These improvements converge the nucleosome binding energy models derived from the two distinct experiments. Both have highly position-specific sequence preferences that correlate strongly with each other. This confirms that the smoothness of the commonly described

10-bp-periodic SS and WW sequence preference stems from the low positional resolution of MNase-Seq. The optimized energy models still have two crucial differences, showing the need for further research and improvements of my method.

My method can be improved by extending the thermodynamic model as others have (e.g. by adding transcription factor competition) and refining the experimental-bias models. Such improvements will not just increase the prediction scores, but will also enhance the obtained energy model. The analysis of nucleosome formation improves with higher resolution and less biased nucleosome binding energies. For example, CC-Seq measurements made it possible to analyze deviations from the 10-bp-periodic pattern (Davey, 2013). As long as no experiments exists that are nearly free of biases and noise, it will be important to separate the relevant signal from experimental errors when extracting nucleosome binding energies. Therefore, I hope that my method can inspire others to construct similar probabilistic models that discern the effects of experimental protocols from the biologically relevant signal contained in nucleosome measurements.

12. Methods

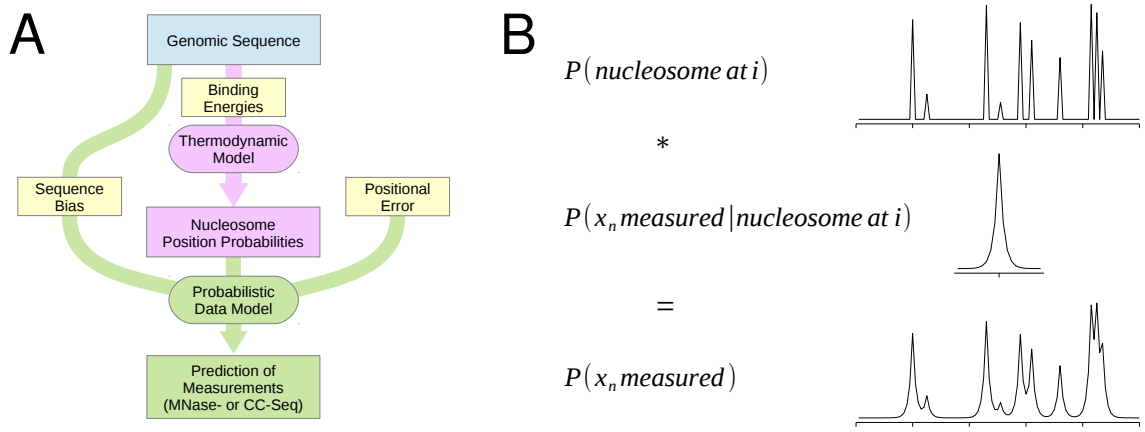

Figure 12.1: **My model distinguishes between a nucleosome position and its experimental measurement**: (A) My probabilistic model consists of a thermodynamic part (purple) that models the nucleosome positioning, and a probabilistic-data part (green) that can model experimental sequence biases (CC-Seq) and positional errors (primarily MNase-Seq) of the measurements. (B) Example of how the positional-convolution function links the nucleosome position and data model probabilities.

When analyzing experimental data of nucleosome positioning, a common implicit assumption is that the measurements have base-pair resolution and no sequence-dependent or other experimental biases. In contrast, my probabilistic approach explicitly models the probability of observing a certain data point given the nucleosome positioning. In my probabilistic model, the measured data, which includes experimental errors, is distinct from the nucleosome positioning, which is the biologically relevant information. Figure 12.1A shows a schematic of my probabilistic method with the nucleosome-positioning model colored in purple and the data model colored in green.

12.1 Thermodynamic model of nucleosome binding

Nucleosome-position prediction methods frequently use similar thermodynamic models. Here the basic concept of them is explained based on the version I implemented. The most important part is the Forward/Backward algorithm. It makes it possible to compute the probabilities of the thermodynamic model in linear time. One issue with the kind of thermodynamic model used is that they assume a thermodynamic equilibrium. This is unlikely given the dynamic properties of nucleosome binding, but a thermodynamic model without this assumption is to complex to apply the Forward/Backward algorithm.

The nucleosome positioning model I use is a thermodynamic model that includes steric hindrance between neighboring nucleosomes. Compared to the thermodynamic models others use, my model describes the probability of a nucleosome dyad occurring at one position compared to other genomic positions (Equation 12.1). We had to reformulate the probabilities in this way to define the measurement-data model. My method optimizes the nucleosome binding energies of the thermodynamic model in the context of the whole probabilistic model, which includes the experimental errors. The method optimizes the model parameters by maximizing the likelihood of observing the training dataset. To maximize the likelihood of the probabilistic model I use a gradient ascent, because the likelihood's derivatives can be calculated for the parameters (Section 12.4).

In my thermodynamic model the probability of a nucleosomes bound at position i is the sum of statistical weights of all configuration that have a nucleosome bound at position i (Z_i^*) normalized that $\sum_{i=1}^{L_S} Z_i^* = 1$ with L_S the length of the sequence S . In all equations the nucleosome positions will be represented by their dyad position.

$$\begin{aligned} \mathrm{P}(\text{nucleosome at i}|S,\epsilon,\mu) &= \frac{Z_i^*}{\sum_{i'=1}^{L_S} Z_{i'}^*} \\ Z_i^* &= F_i^* B_i^* \\ \mathrm{P}(\text{nucleosome at i}|S,\epsilon,\mu) &= \frac{F_i^* B_i^*}{\sum_{i'=1}^{L_S} F_{i'}^* B_{i'}^*} \end{aligned} \tag{12.1}$$

Where ϵ and μ are nucleosome binding energy parameters. Sections 12.1.2 and 12.1.3 describe the nucleosome binding energy model I use.

Computing the statistical weight of all legal configurations individually is impractical, because the time complexity is exponential. The Forward/Backward algorithm – a dynamic programming method – reduces the time complexity down to linearity. The idea behind dynamic programming is to break the problem into smaller parts and save the

relevant results of each part. The results of the previous part are used to solve the next part, without having to iterate over all possibilities of the last part, but only over the relevant results. The approach is popular when working with sequences and is frequently used in sequence-alignment methods. A sequence has a linearity, which provides an obvious way of splitting the problem. The main other aspect needed is a way to save the intermediate solutions compactly.

For the thermodynamic model, every Z_i^* depends on all other Z_i^*. Z_i^* influences the frequency of legal configurations with and without a nucleosome at position i, which in turn affects the neighboring statistical weight and so on. Splitting the Z_i^*s into two parts – no nucleosome overlapping from the left or none from right – also splits the dependency into two parts – nucleosome positions to the left or to the right of i, respectively. This allows the accumulation of the two statistical weight halves by processing the genome sequentially from either side (Forward and Backward). Computing the two halves and combining them has a linear time complexity in regards to the sequence length as described below.

12.1.1 Forward/Backward algorithm

The Forward (F_i^*) and Backward (B_i^*) parts are symmetric, with the exception that F_i^* contains the nucleosome binding energy for position i, while B_i^* does not. The definitions vary between publications: others have made the Forward and Backward parts absolutely symmetrical by splitting out the binding energy of position i into a third part (e.g. Field et al. (2008)). I provide all equations for completeness, but will only explain the Forward equations for briefness. F_i^* is the sum of statistical weights of all legal configurations from the left ($< i$) direction. To compute F_i^* only the information of positions in that direction ($< i$) is needed. The equations also use F_i, the forward sum of statistical weights for all configurations with no nucleosome covering position i (i.e. it being open).

$$\begin{aligned} F_0 &= 1, \quad F_0^* = 0 \quad \text{(Initialization)} \\ F_{i+1} &= F_i + F_{i-D_N}^* \\ F_{i+1}^* &= F_{i-D_N} e^{E_{i+1}-\mu} \end{aligned} \tag{12.2}$$

$$\begin{aligned} B_{L_S+1} &= 1, \quad B_{L_S+1}^* = 0 \quad \text{(Initialization)} \\ B_{i-1} &= B_i + B_{i+D_N}^* e^{E_{i+D_N}-\mu} \\ B_{i-1}^* &= B_{i+D_N} \end{aligned}$$

Where D_N is the half length of nucleosome covered DNA (i.e. 73 bps), E_i is the sequence-specific binding energy of a nucleosome at position i and μ is the sequence-unspecific binding energy of a nucleosome. A position can be open (F_{i+1}), either when the previous position is open (F_i) or when a nucleosome ended on the previous position ($F^*_{i-D_N}$). For a nucleosome dyad to occur (F^*_{i+1}) the position half a nucleosome before has to be open F_{i-D_N}. The statistical weight of the position half a nucleosome earlier being open implies that the positions in between are also open. A nucleosome dyad occurring depends on the nucleosome binding energy of that position ($e^{E_{i+1}-\mu}$).

12.1.2 Sequence-specific binding energy

The common way to describe the sequence binding preference of transcription factors and other DNA-binding factors are position weight matrices (PWMs) (Stormo, 2013). A PWM represents the nucleotide preference at each position assuming independence between all positions. Such PWMs are equivalent to 0th-order Markov chains, which in turn are a type of Markov models (the commonly used term). A higher-order Markov chain loosens the independence assumption by using probabilities that depend on the previous positions. If not specified otherwise, I represent the sequence-specific binding energy with a Markov chain of 1st-order, i.e. the probabilities depend on one previous position. I implemented my method so it can handle higher-order Markov chains and have optimized Markov chains up to 4th-order (Section 16.3). To represent the dyad symmetric nature of nucleosome binding my method uses conditional probabilities that depend on the positions towards the center. The energy terms ϵ that encode the Markov chain represent the logarithm of the conditional probabilities.

$$
\begin{aligned}
E_i &:= \sum_{j=-D_M}^{D_M} \epsilon_j(s_{i+j-k}..s_{i+j}) \\
\epsilon_j(s_{i+j-k}..s_{i+j}) &:= \begin{cases} \ln \frac{p_j(s_{i+j}|s_{i+j-k}..s_{i+j-1})}{p^{\mathrm{bg}}(s_{i+j}|s_{i+j-k}..s_{i+j-1})} & \text{if } j > 0 \\ \ln \frac{p_j(s_{i+j-k}..s_{i+j})}{p^{\mathrm{bg}}(s_{i+j-k}..s_{i+j})} & \text{if } j = 0 \\ \ln \frac{p_j(s_{i+j-k}|s_{i+j-k+1}..s_{i+j})}{p^{\mathrm{bg}}(s_{i+j-k}|s_{i+j-k+1}..s_{i+j})} & \text{if } j < 0 \end{cases}
\end{aligned}
\tag{12.3}
$$

k is the order of the Markov chain and D_M is the half size of the energy model, which is generally $\leq D_N$. The method initializes the parameters ϵ based on Equation 12.3 by estimating the probabilities p and genomic background probabilities p^{bg} from nucleotide frequencies, based on Bolzmann's law. The case separation conserves the dyad symmetry of the energy-model parameters (the cases are shown for a 1st-order model). During

optimization my method removes systematic shifts of the ϵ that represent information contained in neighboring probabilities. This conserves the energy model's representation as a Markov chain.

12.1.3 Sequence-unspecific binding energy

In the model μ represents the sequence-unspecific binding energy. It contains the general binding preference or chemical potential of a nucleosome, and the nucleosome concentration. Because the two cannot be separated without measuring either in an independent experiment, the model represents both with a single parameter.

12.1.4 Occupancy

Section 2 gives the full definition of nucleosome occupancy. In brief, the occupancy is the probability that a position is covered by a nucleosome.

The occupancy is unnecessary to compute the likelihood and optimize the parameters in my method. However, a common validation is to compare predicted and measured occupancies. The occupancy is computed by summing the statistical weights of all configuration where a position is covered by a nucleosome and then dividing it by the sum off all possible configurations. There are two ways to normalize by the sum off all possible configurations:

$$\begin{aligned}
\mathrm{Occ}(i) &= \frac{\sum_{j=i-D_N}^{i+D_N} Z_j^*}{\sum_{j=i-D_N}^{i+D_N} Z_j^* + F_i B_i} \\
&\text{or} \\
\mathrm{Occ}(i) &= \frac{\sum_{j=i-D_N}^{i+D_N} Z_j^*}{F_0 B_0} \\
&\propto \sum_{j=i-D_N}^{i+D_N} Z_j^* \\
&\propto \sum_{j=i-D_N}^{i+D_N} \mathrm{P}(\text{nucleosome at } j | S, \epsilon, \mu)
\end{aligned} \tag{12.4}$$

The local and global normalization approaches have different advantages. In practice I used unnormalized, i.e. scale-free, occupancies based on the last proportional equation, i.e. smoothed dyad probabilities. Because the nucleosome-positioning measurements have no absolute scale, the validation has to mask scaling – including the normalization – anyway. I also used occupancy prediction computed from $\mathrm{P}(x_n\text{measured}|S, \epsilon, \mu, \theta)$, which contains experimental biases (described below). This make sense when comparing the

predictions with data measured by a similar experiment as the training data.

The occupancy has been used to optimize functions that describe dependencies between neighboring nucleosomes (Lubliner and Segal, 2009). In Section 3.2 I mention that I was limited in the experiments my implemented method can model. If the experiment measures nucleosome occupancy at specific positions – not individual nucleosome positions – the likelihood representing the measurements has to reflect this. I did develop a variant of my method that learns the energy model from such occupancy measurements, but I have not implemented it (Section 20).

12.2 Probabilistic data model

This Section introduces the basic version of the probabilistic data model. Section 12.6 contains further variations. Section 12.2.1 describes the likelihood $\mathcal{L}$ and how the thermodynamic model is embedded in it. Section 12.2.2 describes how my method maximizes the likelihood to optimize the parameters. It includes the partial derivatives of the log-likelihood ($\log \mathcal{L}$) and its parts.

12.2.1 Likelihood

$$\mathcal{L} = \prod_{n=1}^{N} \mathrm{P}(x_n \mathrm{measured} | S, \epsilon, \mu, \theta)^{w_n} \tag{12.5}$$

Where n is the index of the measurements, x_n the position, and w_n the summed weight (e.g. counts) of the measurements at position x_n. S is the sequence (e.g. genome), ϵ the sequence-specific binding energies, μ the sequence-unspecific binding energy and θ all other parameters.

The common implicit assumption that the data directly reflects nucleosome positioning is the same as setting $\mathrm{P}(x_n \text{ measured}|...) := \mathrm{P}(\text{nucleosome at } i|...)$. This assumption is an oversimplification, which misses experimental biases and uncertainties. Initially, our main concern for MNase-Seq was the high positional uncertainty compared to CC-Seq (Section 3.1). MNase-Seq has more biases that I ignore in my method. I tried to address and correct these issues independently to no avail (Section 17).

To describe the positional uncertainty the probability of a measurement at position x_n given a nucleosome at i is separated from a nucleosome occurring at position i:

$$\mathcal{L} = \prod_{n=1}^{N} \left[\sum_{i=1}^{N} \mathrm{P}(x_n \text{ measured}|\text{nucleosome at } i, \theta)\, \mathrm{P}(\text{nucleosome at } i|S, \epsilon, \mu) \right]^{w_n} \tag{12.6}$$

The computation of P(nucleosome at $i|S, \epsilon, \mu$) is part of the thermodynamic model (Section 12.1). P(x_n measured|nucleosome at i, θ) can represent different experimental biases and positional errors. Here, I explain the concept based on the positional-uncertainty model I used for the MNase-Seq data. I describe further models, like the one I used for CC-Seq data, in Section 12.6.

The MNase-Seq data model uses a positional-uncertainty point spread function based on a Laplace distribution. The maximal allowed error region is limited to $\pm\delta$, which I set to ± 20 (Section 12.5.2). The model assumes that a measurement cannot be further than δ bps away from the actual nucleosome dyad.

$$\mathrm{P}(x_n \text{ measured}|\text{nucleosome at } i, \theta) = \mathrm{P}_{\mathrm{dist}}(x_n - i|\eta) = A_\eta e^{\frac{-|x_n - i|}{\eta}} \tag{12.7}$$

The parameter η is equivalent to the standard deviation of a Laplace distribution. It controls the width of the point-spread function. A_η is the normalization constant that ensures the position probabilities sum to one:

$$\begin{aligned} \frac{1}{A_\eta} &= \sum_{k=x_n-\delta}^{x_n+\delta} e^{\frac{-|x_n-k|}{\eta}} = \sum_{k=-\delta}^{\delta} (e^{\frac{1}{\eta}})^{-|k|} = \sum_{k=0}^{\delta} (e^{\frac{1}{\eta}})^{-k} + \sum_{k=1}^{\delta} (e^{\frac{1}{\eta}})^{-k} \\ &= \frac{(e^{\frac{1}{\eta}})^{-\delta} - (e^{\frac{1}{\eta}})}{1 - (e^{\frac{1}{\eta}})} + \frac{(e^{\frac{1}{\eta}})^{-\delta} - 1}{1 - (e^{\frac{1}{\eta}})} = \frac{2(e^{\frac{1}{\eta}})^{-\delta} - (e^{\frac{1}{\eta}}) - 1}{1 - (e^{\frac{1}{\eta}})} \end{aligned} \tag{12.8}$$

Inserting Equation 12.7 into Equation 12.6 and taking its logarithm brings the full log-likelihood for the model with Laplace-like positional uncertainty to:

$$\begin{aligned} \log \mathcal{L} &= \sum_{n=1}^{N} \log \left[\sum_{i=x_n-\delta}^{x_n+\delta} A_\eta e^{\frac{-|x_n-i|}{\eta}} \frac{F_i^* B_i^*}{\sum_{i'=1}^{L_S} F_{i'}^* B_{i'}^*} \right] w_n \\ &= \sum_{n=1}^{N} \left(\log \left[\sum_{i=x_n-\delta}^{x_n+\delta} e^{\frac{-|x_n-i|}{\eta}} F_i^* B_i^* \right] - \log \left[\sum_{i=1}^{L_S} F_i^* B_i^* \right] + \log A_\eta \right) w_n \\ &= \sum_{n=1}^{N} \left(\log \left[\sum_{i=x_n-\delta}^{x_n+\delta} e^{\frac{-|x_n-i|}{\eta}} F_i^* B_i^* \right] w_n \right) - \log \left[\sum_{i=1}^{L_S} F_i^* B_i^* \right] \sum_{n=1}^{N} w_n + \log A_\eta \sum_{n=1}^{N} w_n \end{aligned} \tag{12.9}$$

12.2.2 Maximizing the likelihood

The likelihood is maximized with the help of the partial derivatives. Here I describe the partial derivative of the basic log-likelihood (Equation 12.9) for each parameter. The time complexity to compute each partial derivative matches that of the log-likelihood itself.

Section 12.3 discusses the time complexity of the whole model. Maximizing the likelihood over a whole datasets is to time-consuming, instead I used a mini-batch gradient ascent (Section 12.4.1).

Note that to keep the derivatives simpler the log-likelihood use the natural logarithm, i.e. log := ln for all equations. This is common practice. (Using another logarithm would only add a constant factor, which the step size scaling could negate.)

Partial derivative of the log-likelihood

The partial derivative follow by methodical application of the derivative rules to Equation 12.9.

$$\frac{\partial \log \mathcal{L}}{\partial \epsilon, \mu} = \sum_{n=1}^{N} \frac{\sum_{i=x_n-\delta}^{x_n+\delta} e^{\frac{-|x_n-i|}{\eta}} \left(B_i^* \partial F_i^* + F_i^* \partial B_i^*\right)}{\sum_{i=x_n-\delta}^{x_n+\delta} e^{\frac{-|x_n-i|}{\eta}} F_i^* B_i^*} w_n - \frac{\sum_{i=1}^{L_S} \left(B_i^* \partial F_i^* + F_i^* \partial B_i^*\right)}{\sum_{i=1}^{L_S} F_i^* B_i^*} \sum_{n=1}^{N} w_n$$

$$\frac{\partial \log \mathcal{L}}{\partial \eta} = \sum_{n=1}^{N} \frac{\sum_{d=1}^{\delta} \frac{d}{\eta^2} e^{\frac{-d}{\eta}} \left(B_{x_n-d}^* F_{x_n-d}^* + B_{x_n+d}^* F_{x_n+d}^*\right)}{\sum_{i=x_n-\delta}^{x_n+\delta} e^{\frac{-|x_n-i|}{\eta}} F_i^* B_i^*} w_n + \frac{\partial}{\partial \eta} \log A_\eta \sum_{n=1}^{N} w_n \tag{12.10}$$

Partial derivatives of the Forward/Backward algorithm

For the sequence-specific binding energies ϵ and the sequence-unspecific binding energy μ the partial derivatives of the Forward and Backward terms occur in Equation 12.10. These Forward and Backward derivatives follow by methodical application of the derivative rules to Equation 12.2. They are also computed with the Forward/Backward algorithm,

because their dependency structure is identical to the original's.

$$\begin{aligned}
&\frac{\partial F_0}{\partial \epsilon_l(q)} = 0, \quad \frac{\partial F_0^*}{\partial \epsilon_l(q)} = 0 \\
&\frac{\partial F_{i+1}}{\partial \epsilon_l(q)} = \frac{\partial F_i}{\partial \epsilon_l(q)} + \frac{\partial F_{i-D_N}^*}{\partial \epsilon_l(q)} \\
&\frac{\partial F_{i+1}^*}{\partial \epsilon_l(q)} = \left[\frac{\partial F_{i-D_N}}{\partial \epsilon_l(q)} + F_{i-D_N} \, \mathrm{I}(s_{i+1+l-k}, ..., s_{i+1+l} = q)\right] e^{E_{i+1}-\mu}
\end{aligned} \tag{12.11}$$

$$\begin{aligned}
&\frac{\partial B_{L_S+1}}{\partial \epsilon_l(q)} = 0, \quad \frac{\partial B_{L_S+1}^*}{\partial \epsilon_l(q)} = 0 \\
&\frac{\partial B_{i-1}}{\partial \epsilon_l(q)} = \frac{\partial B_i}{\partial \epsilon_l(q)} + \left[\frac{\partial B_{i+D_N}^*}{\partial \epsilon_l(q)} + B_{i+D_N}^* \, \mathrm{I}(s_{i+D_N+l-k}, ..., s_{i+D_N+l} = q)\right] e^{E_{i+D_N}-\mu} \\
&\frac{\partial B_{i-1}^*}{\partial \epsilon_l(q)} = \frac{\partial B_{i+D_N}}{\partial \epsilon_l(q)}
\end{aligned}$$

Where $\epsilon_l(q)$ is the binding-energy parameter of position l for the oligonucleotide q of length $k+1$, i.e. k is the order of the Markov chain that describes the energy model.

$$\begin{aligned}
&-\frac{\partial F_0}{\partial \mu} = 0, \quad -\frac{\partial F_0^*}{\partial \mu} = 0 \\
&-\frac{\partial F_{i+1}}{\partial \mu} = -\frac{\partial F_i}{\partial \mu} - \frac{\partial F_{i-D_N}^*}{\partial \mu} \\
&-\frac{\partial F_{i+1}^*}{\partial \mu} = \left[-\frac{\partial F_{i-D_N}}{\partial \mu} + F_{i-D_N}\right] e^{E_{i+1}-\mu}
\end{aligned} \tag{12.12}$$

$$\begin{aligned}
&-\frac{\partial B_{L_S+1}}{\partial \mu} = 0, \quad -\frac{\partial B_{L_S+1}^*}{\partial \mu} = 0 \\
&-\frac{\partial B_{i-1}}{\partial \mu} = -\frac{\partial B_i}{\partial \mu} + \left[-\frac{\partial B_{i+D_N}^*}{\partial \mu} + B_{i+D_N}^*\right] e^{E_{i+D_N}-\mu} \\
&-\frac{\partial B_{i-1}^*}{\partial \mu} = -\frac{\partial B_{i+D_N}}{\partial \mu}
\end{aligned}$$

The Forward/Backward computations are performed in log-space to cope with large-size increases of the Forward and Backward terms. To perform the computations in log-space the partial derivatives for the sequence-unspecific binding affinity μ have to be negated to make them positive.

Partial derivative of A_η

The partial derivative of η does not contain the derivatives of the Forward and Backward term, because they are independent of η. However, it contains the derivative of the logarithmized normalization term A_η. $\frac{\partial \log A_\eta}{\partial \eta}$ can be derived from the solved or unsolved summation of Equation 12.8:

$$\begin{aligned}\frac{\partial \log A_\eta}{\partial \eta} &= \frac{\partial}{\partial \eta} \log \left(\frac{1 - e^{\frac{1}{\eta}}}{2e^{\frac{-\delta}{\eta}} - e^{\frac{1}{\eta}} - 1} \right) \\ &= \frac{1}{\eta^2} \left(\frac{e^{\frac{1}{\eta}}}{1 - e^{\frac{1}{\eta}}} - \frac{2\delta e^{\frac{-\delta}{\eta}} + e^{\frac{1}{\eta}}}{2e^{\frac{-\delta}{\eta}} - e^{\frac{1}{\eta}} - 1} \right) \end{aligned} \tag{12.13}$$

or:

$$\frac{\partial \log A_\eta}{\partial \eta} = -A_\eta \sum_{k=-\delta}^{\delta} \frac{|k|}{\eta^2} e^{\frac{-|k|}{\eta}}$$

12.3 Time complexity

Table 12.1: Time complexity of individual computation steps. The shown time complexity is for the isolated step that excludes the computation of other steps it relies upon and for individual partial derivatives (i.e. one of the ϵ parameters or μ, not all).

Part	Complexity
E	$O(L_N L_S)$
F^* and B^*	$O(L_S)$
LL	$O(N2\delta + L_S)$
∂F^* and ∂B^*	$O(L_S)$
∂LL	$O(N2\delta + L_S)$

Table 12.1 gives a break down of the time complexity for one iteration. The time complexity is an approximation of the amount of basic operations a computation needs. The combined time complexity for computing the gradient is $O((2 \times 4^{(k+1)} + 1)L_N L_S + 4^{(k+1)} L_N N2\delta))$. The gradient contains all partial derivatives and needs to be computed every iteration.

This is a limiting factor when using a large dataset, i.e. large L_S and N. For one of the datasets I used, the amount of operations is about 4×10^{11} (with $L_N = 140$, $k = 1$,

$L_S = 1.2 \times 10^7$, $\delta = 20$, $N = 4 \times 10^6$). If every operation needs 4×10^{-8} seconds this leads to a computation time of 10^3 seconds $\approx$ 17 minutes. This approximation might be optimistic, because the time needed for one operation could be longer. The given time is based on ~100 float point operation, and the basic operation I used to approximate the time complexity could need more operations (in unoptimized code). An optimization with thousands of iterations would take weeks and be impractical without even using an order (k) higher than 1st, which I wanted to test. Therefore, I optimized the parameters with an algorithm that uses a subset of the data each iteration called mini-batch gradient descent (Section 12.4.1).

To keep the computation time low I implemented the algorithm in C++. To further reduce the runtime I used SSE2 intrinsics, which are parallel operations on the instruction level (SIMD), and parallelized the partial derivative computations with OpenMP (OpenMP Architecture Review Board, 2008), which is a multi-threading framework for single computation nodes (e.g. a computer). Most of the computations are performed in $\log_2$-space, resulting in the frequent use of $\log_2(x)$, 2^x and $\log_2(2^x + 2^y)$ (addition of two values stored in $\log_2$-space). For these operations I implemented fast SSE2 versions that are approximations i.e. the resulting floating-point numbers are not precise to all significands. Using these approximations or even just using float precision leads to significant numerical errors in comparison to double-precision computations for chromosome-length sequences. For this reason and for bug fixing purposes, I maintained two version: a fast low-precision version with the above speed improvements and a slow high-precision version. Being able to use the approximations during the parameter optimization is a secondary benefit of restricting the subset of data used in one iteration to a small genomic region as described below.

12.4 Gradient descent

Gradient descent optimizes a function by following a gradient, as the name implies. Each parameter is updated proportional to the negated partial derivative. Gradient descent is commonly used if the functions gradient can be calculated, but its analytical solution cannot. Descent describes the direction of optimization towards a local minimum. The typical application minimizes the error function of a model. My method maximizes the log-likelihood, technically making it a gradient ascent. Except for the sign, the two are equivalently, so I will continue referring to gradient-descent algorithms for consistency with literature. The standard gradient descent (also called batch gradient descent) com-

putes the gradients for the whole dataset each iteration. Variants reduce the amount of iterations needed to optimize the parameters by using second-order derivatives or the Hessian matrix to fine tune the learning rates and the gradient.

An abundance of data can limit the optimization of the gradient descent by increasing the computation time of every iteration. Stochastic gradient descent computes the gradient for single data points every iteration to update the parameters. This lets it handle large datasets better. Another advantages is the avoidance of local optima. Local optima are unlikely to exist for all data points, allowing the algorithm to find paths out of them. Literature highlights further advantages not mentioned here (Wilson and Martinez, 2003; Bottou, 2012).

The computation time of the stochastic gradient descent can be limiting in a different way. The parameters need more iterations to converge and the afore mentioned tricks to reduce the amount of needed iterations are difficult to apply. One way to speed up the optimization is by computing gradients in parallel and averaging, effectively reducing the parameter update frequency. This is generally revered to as mini-batch gradient descent and described below in detail. Small mini-batches retain the advantages of stochastic gradient descent, while speeding up the optimization. Standard gradient descent and stochastic gradient descent are two opposing extremes with mini-batch gradient descent representing the middle ground.

12.4.1 Mini-batch gradient descent

Even with the implemented speed ups (Section 12.3) computing gradients from genome-wide data in thousands of iterations is to time-consuming. My method uses a mini-batch gradient descent to alleviate this problem. Mini-batch gradient descent uses a subset (mini-batch) of the dataset to compute the gradients every iteration.

I had to adapt the mini-batch approach slightly for my method. As described in the Section 12.3, both large N and L_S lead to high computational times. The normal mini-batch reduces the size of N by selecting a subset of data points. To reduce L_S, which scales the Forward/Backward computation time, I further restricted the subset to one region of the genome. The method creates the mini-batches by separating the genome into windows and only using the data points located in a single window every iteration. The genomic windows have a consistent size (except at the chromosome ends). Therefore, the amount of data points in each mini-batch varies depending on the local density of data. As long as each mini-batch has a reasonable amount of data this has little effect on the optimization. (I remove mini-batches with to little data.)

When using too small region, a phasing effect at the borders, which is created by the Forward/Backward initialization, becomes a problem. On average the phasing effect is a minor problem localized to the first 300 bps, but it can be more severe in individual cases. To prevent such issues, I included buffer regions on either side when computing the Forward/Backward values that are excluded when selecting the data subset. Based on the average phasing effect I chose a generous buffer region size of 1 kbp.

For larger regions the single-precision float points (floats) and approximated operations of my low-precision implementation lead to accumulative errors that overshadow the signal. A simple solution is to use double-precision float points (doubles) and exact operations, but this naturally increases the computation time. For example, I use the high-precision version to compute genome-wide predictions. To keep the computation time of the iterations low, I stuck with mini-batch regions of 25 kbps. This allowed me to compute the gradients with the low-precision version.

I found no downsides of using this combination of length restriction and low-precision computation. From tests the low-precision version is robust in a range of at least 15 to 50 kbps, which splits the yeast genome into 867 to 206 mini-batches.

12.4.2 Convergence of the mini-batch gradient descent

A major downside of the mini-batch and stochastic gradient descent as described above is that they never fully converge. Once the parameters are close to their optima they jiggle around it. Then the average distance between the parameters and their optima depends on the learning rate λ, a hyper parameter also known as step size or step length. Choosing a small learning rate increases the precision of the results, but it also increases the amount of iterations needed and hence run time.

I tested four approaches to resolve the trade-off: decreasing the learning rate, an increasing momentum, ALAP (Almeida et al., 1998), and vSGD-fd (Schaul and LeCun, 2013). With good hyper-parameter choices both decreasing the learning rate and an increasing momentum lead to similar results. For the primary analyzes I primarily relied on the momentum, because it appeared robuster in my tests. I used an initial anti-momentum of ρ_0 of 1.0 (in my implementation I use and decrease the anti-momentum), which I decreases with a factor ϱ of 0.5 every 400 iterations (Table B.3). While ALAP and vSGD-fd improved upon the vanilla version or the other two approaches with a bad choice of hyper-parameters, they were less robust than I had hoped and their results were worse then those of decreasing the learning rate and an increasing momentum with good hyper-parameters.

Decreasing the learning rate

A simple and common way to resolve the trade-off between precision and needed iterations is to decrease the learning rate λ_i with ongoing iterations i. The main downside is the addition of a second hyper-parameter, the decrease rate Λ. To guarantee reaching the proximity of the optimum the learning rate series have to diverge ($\sum_i \lambda_i = \infty$). To guarantee convergence to the optimum from its proximity the individual learning rate terms have to converge to 0 ($\lim_{i\to\infty} \lambda_i = 0$). The two conditions are theoretical boundaries, which are helpful guide lines. In praxis the optimization has to of course finish in a finite amount of iterations (for my method in thousands). I first tried an exponential decay i.e. geometric series ($\lambda_i = \lambda_0(1-\Lambda)^i$), which has the term limit of 0, but does not diverge. Later I also tried a general harmonic series ($\lambda_i = \frac{\lambda_0}{1+\Lambda i}$), which has both the limit of 0 and diverges (slowly). I updated the learning rate every 400 iterations, which is about one cycle through the yeast genome for my mini-batch sizes. In both cases, one has to evaluate test runs to find a good decrease rate Λ.

Increasing momentum

Another way to improve the convergence is to average the new gradient with the gradients of previous iterations. An approach called momentum, due to its similarity to momentum conservation in physics. The formula is $g_i = g_i^*\rho + g_{i-1}(1-\rho)$, given the used gradients g, the new calculated partial derivative g^* and the weight ρ of the new partial derivative compared to the old gradient. This smooths the gradients exponentially. The momentum weights the previous gradients with an exponential decreasing importance, while only needing to store the last. Using the momentum makes the parameter path smoother (like with bigger mini-batches and a smaller learning rate) and can improve the convergence speed (Qian, 1999; Rakhlin et al., 2011).

The weight of the previous gradients can increase with time, similar to decreasing the learning rate. Decreasing the anti-momentum weight ρ_i (increasing $1-\rho_i$) during the optimization improves the final convergence, while adding a second hyper-parameter. The decreasing anti-momentum weight ρ_i results in an increasingly smooth gradient for later iterations, like computing the gradient on larger and larger mini-batches (an alternative approach that is impractical for my method due to restrictions mentioned in Section 12.4.1). An advantage of the momentum approach is that the two boundary cases are the standard gradient descent (low ρ) and mini-batch gradient descent without momentum ($\rho = 1$). The worst case scenarios for the hyper-parameter choice result in behavior of more basic versions of the method. In comparison, when decreasing the learn-

ing rate one boundary case is getting stuck on the way to the proximity of the optimum, because the learning rate deteriorates to fast (slow divergence). The above boundaries of the momentum assume a decrease of ρ where the weight of the new gradient is higher than the effective weight of the last gradient in the new iteration: $\rho_{i+1} \geq \rho_i(1-\rho_{i+1})$. Without this boundary the optimization breaks and the parameters diverge, because information of older partial derivatives dominate the gradient forever. I based the definition of ρ_i and the boundaries of the hyper-parameter ϱ on upholding this relationship:

$$\begin{aligned} \rho_{i+1} &\geq \rho_i(1-\rho_{i+1}) \\ \rho_{i+1} &\geq \rho_i - \rho_i\rho_{i+1} \\ \rho_{i+1} + \rho_i\rho_{i+1} &\geq \rho_i \\ \rho_{i+1}(1+\rho_i) &\geq \rho_i \\ \rho_{i+1} &\geq \frac{\rho_i}{1+\rho_i} \\ \rho_{i+1} &:= \frac{\rho_i}{1+\varrho\rho_i}, \text{ with } 0 < \varrho \leq 1 \end{aligned} \tag{12.14}$$

Local learning rate adaptation

As mentioned above, the tricks of exploiting second-order gradients or Hessian matrices – used in some standard gradient descent variants – are difficult to apply to stochastic or mini-batch gradient descent. When computing the second-order gradients or Hessian matrix on subsets of data the results are approximations that do not guarantee quicker convergence (Bordes et al., 2009). Nonetheless, some variants rely on the same concepts, but in more complex algorithms, to improve the learning rates or gradient in every iteration. Some variants even adapt an individual learning rate per parameter. I implemented and tested two such algorithms: ALAP (Almeida et al., 1998) and vSGD-fd (Schaul and LeCun, 2013). The preliminary results left me unconvinced: while performing better than the two other variations with bad hyper-parameters, the resulting models appeared to be worse than the models produced by the other variations with good hyper-parameters (even without extensively optimizing these good hyper-parameters).

I implemented both algorithms as described in their respective publications and refer the reader to them for further details. I made minor changes to the parameter initialization that should not influence the outcome greatly, because they mostly affect the early optimization. Soon the optimization looks normal and the parameters later converge successfully. The under performance of either method might be due to the mini-batch size, update frequency or high interdependency of the sequence-specific binding energy

parameters (ϵ).

12.5 Hyper-parameters

Hyper-parameters are static i.e. constant throughout the optimization. The user sets them during compilation or when executing the method. I chose their values by combining prior knowledge and leave-out validation. In either case, repeated testing identified some major problems I could resolve, e.g. issues with numerical errors due to the precision of the computations (Section 12.3). I conducted most tests on Kaplan et al.'s *in vitro* dataset (Kaplan et al., 2009).

Here I briefly discuss important hyper-parameters that were not yet described in detail, most of which are mentioned above. Tables B.2, B.3 and B.4 provide an overview of the default values I used for the hyper-parameters and method options. For hyper-parameters and options not described here I refer to the documentation of my implementation.

12.5.1 Nucleosome footprint

The nucleosome footprint $2D_N + 1$ defines the minimal distance between two neighboring nucleosomes in the thermodynamic model. Varying the nucleosome footprint hardly affects the log-likelihood, with 143 bps being the optimal value and a slow drop off for both longer and shorter footprints. The optimized energy models are consistent between varying lengths. Therefore, I stuck with a nucleosome footprint of 147 bps, which is the commonly used nucleosome footprint size.

12.5.2 Maximal positional uncertainty

The majority of MNase-Seq measurements have a positional error of $\leq \pm 5$ bps, so a maximal positional uncertainty δ of 20 bps between the measurement and its nucleosome seemed appropriate. I confirmed this by checking the optimized point-spread function. The weight of the outermost positions is less than 0.2% for the Laplace-like point-spread function after a typical training on the *in vitro* MNase-Seq dataset. The point-spread function for deconvoluted CC-Seq data is even narrower.

The situation is less obvious when using the position-specific deconvolution point-spread function (Section 12.6.2) on the raw CC-Seq data. The outermost positions have higher weights and extending the window leads to more positions with substantial weights.

This appears to stem from a background distribution. I ignored this behavior and set δ to the same value of 20 for all optimizations. The existence of a background distribution is interesting and animated me to model random background measurements (Section 12.6.4). The issue was unchanged when modeling the background measurements, and the origin of this behavior is still unknown to me.

12.5.3 nucleosome binding energy model size

MNase has a cut preference that leads to biased nucleotide frequencies at the nucleosome borders in MNase-Seq data (Section 8.1). Learning and predicting the data with this information has no biological relevance. I restricted the energy model size to ± 50 bps to leave out the biased nucleosome ends, when accounting for the maximal positional uncertainty allowed between the measurements and nucleosomes. I have trained larger CC-Seq energy models, but ultimately decided to use the same reduced model size for consistency.

12.5.4 Stop criteria for the gradient descent

To prevent the optimization from never halting I set the maximal number of allowed iterations to 10000. After that the optimization is terminated. The optimization should converge beforehand and meet a criterion that interrupts it earlier. I implemented two criteria to stop the optimization: convergence of the parameters, and a converged or declining log-likelihood. The data used in every iteration is different in mini-batch gradient descent and, therefore, the optimal parameter set is a moving target. The parameters do not fully converge without outside influence (Section 12.4.1). Stopping based on the log-likelihood also has problems that have to be dealt with, but can additionally prevent overfitting, which is important. I focused on early stopping based on a converged or declining log-likelihood.

With the different data subset each mini-batch has, the log-likelihood of the iteration's data subset is not comparable with the previous iteration's. Instead of comparing the log-likelihood of recent iterations, I compute the difference to the log-likelihood of the last iteration that trained on the same data subset. To ensure consistency I split the data into subsets once and iterated over the subsets in the same order every genomic cycle. The reference log-likelihoods are computed after optimizing on the data subset and stored for the last genomic cycle. They are compared against the new log-likelihoods before the renewed optimization on the same data subset. By comparing the log-likelihoods after

the last and before the new optimization, the computed delta log-likelihood corresponds to the models improvement of describing a subset gained by training on the other subsets. By excluding the subset itself the score reflect a validation on an independent test set better. This helps protect the model from overfitting. With my default parameters, the optimization is interrupted if the average delta log-likelihood of the last 200 iterations drops below a threshold of -0.01.

12.6 Variations of the probabilistic data model

I developed two alternative positional-uncertainty point-spread functions and extensions that model different experimental biases. For the CC-Seq data model shown in Section 13.5, I used a variation with a sequence bias positioned in relation to the nucleosome dyad. The other variations I tested, but ultimately left out of the final analysis.

The first alternative point-spread function replaces the Laplace-like distribution with a Gaussian-like one (Section 12.6.1). The Laplace-like distribution worked better for MNase-Seq in practice, even though my analysis suggested that the positional uncertainty of MNase-Seq data is more Gaussian-like. The second alternative point-spread function is a deconvolution with independent position weights (Section 12.6.2). It allowed my method to optimize a model on raw CC-Seq data, which has a distinct double peak distribution around the nucleosome dyad position (Brogaard et al., 2012). Using the raw data instead of pre-processed data had little effect on the obtained energy model.

The two extensions of the model describe hypothetical error sources of a measurement: random background measurements and a recovery preference based asymmetrically on the sequence around the dyad. I used the asymmetric sequence bias model to optimize the CC-Seq data model and could show that my method can separate the strand dependent sequence bias from the nucleosome binding energy (Section 12.6.3). The random background cut model showed no improvements for CC-Seq, but could still be useful for modeling other experimental measurements (Section 12.6.4). Originally I had planned and developed a third bias model that described a recovery preference based on the sequence around the measurements (Section 12.6.5).

12.6.1 Gaussian-like positional-uncertainty

After observing that the auto-correlation of MNase-Seq data showed a Gaussian-like distribution I developed a matching point-spread function. $\mathrm{P}(x_n\text{measured}|\text{nucleosome at } i)$

is replaced with a Gaussian-like term:

$$\mathrm{P}(x_n\text{measured}|\text{nucleosome at } i,\theta) = \mathrm{P}_{\text{dist}}(x_n - i|\sigma) \propto e^{-\frac{(x_n-i)^2}{\sigma}} \tag{12.15}$$

As for the Laplace-like point-spread function the normalization can be simplified:

$$\begin{aligned} \sum_{i=x_n-\delta}^{x_n+\delta} e^{-\frac{(x_n-i)^2}{\sigma}} &= \sum_{k=-\delta}^{\delta} e^{-\frac{k^2}{\sigma}} = 1 + 2\sum_{k=1}^{\delta} e^{-\frac{k^2}{\sigma}} = \frac{1}{A_\sigma} \\ \mathrm{P}_{\text{dist}}(x_n - i|\sigma) &= A_\sigma e^{-\frac{(x_n-i)^2}{\sigma}} \end{aligned} \tag{12.16}$$

Inserting these terms into the likelihood (Equation 12.6) the derivative of the log-likelihood follows:

$$\frac{\partial \log \mathcal{L}}{\partial \sigma} = \sum_{n=1}^{N} \frac{\sum_{d=-\delta}^{\delta} \frac{-d^2}{-\sigma^2} e^{-\frac{d^2}{\sigma}} B^*_{x_n+d} F^*_{x_n+d}}{\sum_{d=-\delta}^{\delta} e^{-\frac{d^2}{\sigma}} F^*_{x_n+d} B^*_{x_n+d}} w_n + \frac{\partial}{\partial \sigma} \log A_\sigma \sum_{n=1}^{N} w_n$$

with:

$$\begin{aligned} \frac{\partial}{\partial \sigma} \log A_\sigma &= \frac{\partial}{\partial \sigma} \log \left(\frac{1}{\sum_{k=-\delta}^{\delta} e^{-\frac{k^2}{\sigma}}} \right) \\ &= \frac{\partial}{\partial \sigma} - \log \left(\sum_{k=-\delta}^{\delta} e^{-\frac{k^2}{\sigma}} \right) \\ &= -A_\sigma \left(\sum_{k=-\delta}^{\delta} \frac{k^2}{\sigma^2} e^{-\frac{k^2}{\sigma}} \right) = -A_\sigma 2 \left(\sum_{k=1}^{\delta} \frac{k^2}{\sigma^2} e^{-\frac{k^2}{\sigma}} \right) \end{aligned} \tag{12.17}$$

All other equations remain unchanged except for the replaced $\mathrm{P}_{\text{dist}}(x_n - i|\theta)$. The time complexity is identical to the Laplace-like positional-uncertainty.

12.6.2 Position-specific deconvolution as positional-uncertainty

I developed a position-specific deconvolution model that can describe any form of positional-uncertainty distribution. It uses independent parameters ζ_k to represent the probabilities $\mathrm{P}(x_n\text{measured}|\text{nucleosome at } i,\theta)$ for each distance k in the allowed uncertainty region ($\pm\delta$ bps) between the nucleosome and measurement:

$$\mathrm{P}(x_n\text{measured}|\text{nucleosome at } i,\theta) = \mathrm{P}_{\text{dist}}(x_n - i|\zeta) \propto e^{\zeta_{x_n-i}} \tag{12.18}$$

With the normalization term:

$$\begin{aligned} \sum_{k=-\delta}^{+\delta} e^{\zeta_k} &= \frac{1}{A_\zeta} \\ \mathrm{P}_{\mathrm{dist}}(x_n - i|\delta) &= A_\zeta e^{\zeta_{x_n - i}} \end{aligned} \tag{12.19}$$

Inserting these terms into the likelihood (Equation 12.6) the derivative of the log-likelihood follows:

$$\begin{aligned} \frac{\partial \log \mathcal{L}}{\partial \zeta_j} &= \sum_{n=1}^{N} \frac{e^{\zeta_j} B^*_{x_n+j} F^*_{x_n+j}}{\sum_{i=x_n-\delta}^{x_n+\delta} e^{\zeta_{x_n-i}} F^*_i B^*_i} w_n + \frac{\partial}{\partial \zeta_j} \log A_\zeta \sum_{n=1}^{N} w_n \\ with: & \\ \frac{\partial}{\partial \zeta_j} \log A_\zeta &= \frac{\partial}{\partial \zeta_j} \log \left(\frac{1}{\sum_{k=-\delta}^{+\delta} e^{\zeta_k}} \right) \\ &= \sum_{k=-\delta}^{+\delta} e^{\zeta_k} \frac{-1}{(\sum_{k=-\delta}^{+\delta} e^{\zeta_k})^2} e^{\zeta_j} \\ &= -e^{\zeta_j} A_\zeta \end{aligned} \tag{12.20}$$

Again, the other equations remain unchanged except for the replaced $\mathrm{P}_{\mathrm{dist}}(x_n - i|\zeta)$. The optimization of all ζ_j adds time complexity in theory, but because $2\delta + 1 \ll (2 * 4^k) L_N$ the actual time increase is negligible.

12.6.3 Sequence-dependent retrieval bias

CC-Seq data has a sequence bias close to the dyad position (Section 13.4). The bias source is still unclear and 'retrieval' is used as a catch all phrase for the chance of observing any measurement assuming a nucleosome is present. I extended my method to include a model variant that describes such a bias explicitly and separates it from the nucleosome binding model. Specifically, the model uses a energy model positioned in relation to the dyad position to describe the retrieval bias. I split the definition of $\mathrm{P}(x_n \text{measured}|S, \text{nucleosome at } i, \theta)$ into the positional-uncertainty and a sequence-dependent probability to retrieve a measurement from a nucleosome:

$$\begin{aligned} \mathrm{P}(x_n|S, i, \theta) &= \mathrm{P}_{\mathrm{dist}}(x_n - i|\mathrm{retrieved}, \theta)\, \mathrm{P}_{\mathrm{seq}}(\mathrm{retrieved}|S, i, \theta) \\ &\quad + \mathrm{P}_{\mathrm{dist}}(x_n - i|\overline{\mathrm{retrieved}}, \theta)\, \mathrm{P}_{\mathrm{seq}}(\overline{\mathrm{retrieved}}|S, i, \theta) \\ &= \mathrm{P}_{\mathrm{dist}}(x_n - i|\mathrm{retrieved}, \theta)\, \mathrm{P}_{\mathrm{seq}}(\mathrm{retrieved}|S, i, \theta) \end{aligned} \tag{12.21}$$

The second term drops out, because if no measurement is retrieved than no distance is measured:
$\mathrm{P}_{\mathrm{dist}}(x_n - i|\overline{\text{retrieved}}, \theta) = 0$. Inserted into the likelihood (Equation 12.6) we get:

$$\mathcal{L} = \prod_{n=1}^{N} \left[\sum_{i=x_n-\delta}^{x_n+\delta} \mathrm{P}_{\mathrm{dist}}(x_n - i|\text{retrieved}, \theta)\, \mathrm{P}_{\mathrm{seq}}(\text{retrieved}|S, i, \theta)\, \mathrm{P}(i|S, \epsilon, \mu) \right]^{w_n} \tag{12.22}$$

Where I use a Markov chain (C) to describe $\mathrm{P}_{\mathrm{seq}}(\text{retrieved}|i, S, \theta)$. The Markov chain has the weight $\psi_j(s)$ for position j and oligonucleotide s:

$$\mathrm{P}_{\mathrm{seq}}(\text{retrieved}|i, S, \theta) = C_{S_i} = e^{\sum_{j=-D_C}^{D_C} \psi_j(s_{i+j})} \tag{12.23}$$

Inserted into Equation 12.22 and logarithmized this leads to:

$$\begin{aligned}
\log \mathcal{L} &= \sum_{n=1}^{N} \log \left[\sum_{i=x_n-\delta}^{x_n+\delta} A_\eta e^{\frac{-|x_n-i|}{\eta}}\, \mathrm{P}_{\mathrm{seq}}(\text{retrieved}|i, S)\, \mathrm{P}(i|S, \epsilon, \mu) \right] w_n \\
&= \sum_{n=1}^{N} \left(\log \left[\sum_{i=x_n-\delta}^{x_n+\delta} e^{\frac{-|x_n-i|}{\eta}} C_{S_i} F_i^* B_i^* \right] - \log \left[\sum_{i=1}^{L_S} C_{S_i} F_i^* B_i^* \right] + \log A_\eta \right) w_n \\
&= \sum_{n=1}^{N} \log \left[\sum_{i=x_n-\delta}^{x_n+\delta} e^{\frac{-|x_n-i|}{\eta}} C_{S_i} F_i^* B_i^* \right] w_n - \log \left[\sum_{i=1}^{L_S} C_{S_i} F_i^* B_i^* \right] \sum_{n=1}^{N} w_n + \log A_\eta \sum_{n=1}^{N} w_n
\end{aligned} \tag{12.24}$$

And the derivative:

$$\begin{aligned}
\frac{\partial \log \mathcal{L}}{\partial \psi_{j'(a')}} &= \sum_{n=1}^{N} \frac{\sum_{i=x_n-\delta}^{x_n+\delta} e^{\frac{-|x_n-i|}{\eta}} B_i^* F_i^* \partial C_{S_i}}{\sum_{i=x_n-\delta}^{x_n+\delta} e^{\frac{-|x_n-i|}{\eta}} F_i^* B_i^* C_{S_i}} w_n - \frac{\sum_{i=1}^{L_S} F_i^* B_i^* \partial C_{S_i}}{\sum_{i=1}^{L_S} F_i^* B_i^* C_{S_i}} \sum_{n=1}^{N} w_n \\
\partial C_{S_i} &= \frac{\partial C_{S_i}}{\partial \psi_{j'(a')}} = e^{\sum_{j=-D_C}^{D_C} \psi_j(s_{i+j})}\, \mathrm{I}(S_{i+j'} == a') \\
\frac{\partial \log \mathcal{L}}{\partial \psi_{j'(a')}} &= \sum_{n=1}^{N} \frac{\sum_{i=x_n-\delta}^{x_n+\delta} e^{\frac{-|x_n-i|}{\eta}} B_i^* F_i^* e^{\sum_{j=-D_C}^{D_C} \psi_j(s_{i+j})}\, \mathrm{I}(S_{i+j'} == a')}{\sum_{i=x_n-\delta}^{x_n+\delta} e^{\frac{-|x_n-i|}{\eta}} F_i^* B_i^* e^{\sum_{j=-D_C}^{D_C} \psi_j(s_{i+j})}} w_n \\
&\quad - \frac{\sum_{i=1}^{L_S} F_i^* B_i^* e^{\sum_{j=-D_C}^{D_C} \psi_j(s_{i+j})}\, \mathrm{I}(S_{i+j'} == a')}{\sum_{i=1}^{L_S} F_i^* B_i^* e^{\sum_{j=-D_C}^{D_C} \psi_j(s_{i+j})}} \sum_{n=1}^{N} w_n
\end{aligned} \tag{12.25}$$

C_{S_i} is also added in the other derivatives as a constant factor, which leaves the time complexity of their computation unchanged. If $D_C < D_M$ optimizing the sequence-bias model parameters ψ leads to less than a 2-fold increase of the total runtime. I used $D_C = 10$, while $D_M = 50$, leading to a ~20% increase in computation time.

12.6.4 Background measurements for CC-Seq

Before I showed that the sequence bias close to the dyad position in CC-Seq is strand specific (Section 13.4), I had a second hypothesis that nucleosome-independent cuts produced these biases. I developed a probabilistic data model that described the possibility of nucleosome-independent measurements. In comparison to the previous model, the sequence bias of these background measurements is positioned in relation to the cut site – not the dyad position. I use 'real' as the opposite of a background measurement in the equations below.

$$\begin{aligned} \mathrm{P}(x_n\text{measured}|\mathrm{S},\,\text{parameter}) &= \mathrm{P}(x_n\text{measured}|\overline{\text{real}})\,\mathrm{P}(\overline{\text{real}}) \\ &\quad + \mathrm{P}(x_n\text{measured}|\text{real})\,\mathrm{P}(\text{real}) \\ &= P_{x_n|\overline{\text{real}}}q + P_{x_n|\text{real}}(1-q) \end{aligned} \tag{12.26}$$

Inserting these new definitions into Equation 12.6 and logarithmizing the result we get:

$$\begin{aligned} \log\mathcal{L} &= \sum_{n=1}^{N}\log\left[qP_{x_n|\overline{\text{real}}} + (1-q)P_{x_n|\text{real}}\right]w_n \\ &= \sum_{n=1}^{N}\log\left[qP_{x_n|\overline{\text{real}}} + \frac{(1-q)A_\zeta}{\sum_{i=1}^{L_S}F_i^*B_i^*}\sum_{d=-\delta}^{\delta}e^{\zeta_d}B^*_{x_n+d}F^*_{x_n+d}\right]w_n \end{aligned} \tag{12.27}$$

Because the logarithm now surrounds an addition not a multiplication in, the terms cannot be separated and derivated independently. This had ramifications for the implementation, and the computation time increased, as described below. In Equation 12.27 and the following equations I used the position-specific point-spread function, replacing it with the other point-spreads functions is rather trivial and excluded for brevity. I describe the derivatives for the parameters of the three point-spreads in Equations 12.30, 12.30, and 12.30, because the lack of separating out the normalization terms affects them.

$$\text{Normalization Term} = \left[qP_{x_n|\overline{\text{real}}} + \frac{(1-q)A_\zeta}{\sum_{i=1}^{L_S}F_i^*B_i^*}\sum_{d=-\delta}^{\delta}e^{\zeta_d}B^*_{x_n+d}F^*_{x_n+d}\right] \tag{12.28}$$

$$\frac{\partial \log \mathcal{L}}{\partial \epsilon, \mu} = \sum_{n=1}^{N} \frac{(1-q)A_\zeta \sum_{i=x_n-\delta}^{x_n+\delta} e^{\zeta_{i-x_n}} \frac{\left(B^*_i \partial F^*_i + F^*_i \partial B^*_i\right)\left(\sum_{j=0}^{L_S} F^*_j B^*_j\right) - F^*_i B^*_i \left(\sum_{j=0}^{L_S} (F^*_j \partial B^*_j + \partial F^*_j B^*_j)\right)}{\left(\sum_{j=0}^{L_S} F^*_j B^*_j\right)^2}}{\text{Normalization Term}} w_n \tag{12.29}$$

$$\frac{\partial \log \mathcal{L}}{\partial \eta} = \sum_{n=1}^{N} \frac{(1-q)\sum_{d=-\delta}^{\delta} \frac{B^*_{x_n+d} F^*_{x_n+d}}{\sum_{i=1}^{L_S} F^*_i B^*_i} e^{-\frac{|d|}{\eta}} \left(A_\eta \frac{|d|}{\eta^2} - A^2_\eta \sum_{k=-\delta}^{\delta} \frac{|k|}{\eta^2} e^{-\frac{|k|}{\eta}}\right)}{\text{Normalization Term}} w_n \tag{12.30}$$

$$\frac{\partial \log \mathcal{L}}{\partial \sigma} = \sum_{n=1}^{N} \frac{(1-q)\sum_{d=-\delta}^{\delta} \frac{B^*_{x_n+d} F^*_{x_n+d}}{\sum_{i=1}^{L_S} F^*_i B^*_i} e^{-\frac{d^2}{\sigma}} \left(A_\sigma \frac{d^2}{\sigma^2} - A^2_\sigma \sum_{k=-\delta}^{\delta} \frac{k^2}{\sigma^2} e^{-\frac{k^2}{\sigma}}\right)}{\text{Normalization Term}} w_n \tag{12.31}$$

$$\frac{\partial \log \mathcal{L}}{\partial \zeta_k} = \sum_{n=1}^{N} \frac{(1-q)\left(\frac{B^*_{x_n+k} F^*_{x_n+k}}{\sum_{i=1}^{L_S} F^*_i B^*_i} A_\zeta - \sum_{d=-\delta}^{\delta} \frac{B^*_{x_n+d} F^*_{x_n+d}}{\sum_{i=1}^{L_S} F^*_i B^*_i} A^2_\zeta e^{\zeta_d}\right)}{\text{Normalization Term}} w_n \tag{12.32}$$

The probabilistic data model with background measurements has two new parameters, which can be optimized with their partial derivatives. The background measurement ratio q and the weights m of the sequence-bias Markov chain model of the background measurements:

$$\frac{\partial \log \mathcal{L}}{\partial q} = \sum_{n=1}^{N} \frac{P_{x_n|\overline{\text{real}}} - \frac{A_\zeta}{\sum_{i=1}^{L_S} F^*_i B^*_i} \sum_{d=-\delta}^{\delta} e^{\zeta_d} B^*_{x_n+d} F^*_{x_n+d}}{\text{Norm.Term}} w_n \tag{12.33}$$

$$\begin{aligned} P_{x_n|\overline{\text{real}}} &= \frac{\prod_{j=0}^{J} e^{m_{j,s_{x_n+j}}}}{\sum_{i=0}^{L_C} \prod_{j=0}^{J} e^{m_{j,s_{i+j}}}} \\ \frac{\partial \log \mathcal{L}}{\partial m_{j,s}} &= \sum_{n=1}^{N} \frac{q \frac{\prod_{j=0}^{J} e^{m_{j,s_{x_n+j}}}}{\sum_{i=0}^{L_C} \prod_{j=0}^{J} e^{m_{j,s_{i+j}}}} \left(\mathrm{I}(s_{x_n+j} = s) - \frac{\sum_{l=0}^{L_S} \mathrm{I}(s_{l+j}=s) \prod_{j=0}^{J} e^{m_{j,s_{x_n+j}}}}{\sum_{i=0}^{L_C} \prod_{j=0}^{J} e^{m_{j,s_{i+j}}}}\right)}{\text{Norm.Term}} w_n \end{aligned} \tag{12.34}$$

Being unable to split the log-likelihood makes the equations look more complex, but has little impact on the time complexity. I had to cautiously rearrange the computation to handle the different structure and conserve computational precision. From my tests,

using single floating point precision (float) produced numerical errors in the optimization that corrupted the gradient ascent. For this reason, I had to run the model in the high-precision mode (double or long double) increasing the computation time and reducing the amount of tests I could reasonably execute.

On paper, I combined the background-measurement and the sequence-dependent retrieval bias probabilistic data models. I never implemented the combined model and do not present the equations here, because the background-measurement probabilistic data model never produced significant improvements by itself. I leave the formulation of the log-likelihood and the calculation of the partial derivatives as an exercise to the reader.

12.6.5 Sequence-dependent retrieval bias relative to the measurement

I had intended to develop a third variant of the probabilistic data model that describes a sequence bias in relation to the measurement position – not the dyad position as the model described in Section 12.6.3. After working out the equations I realized that I had unintentionally approximated the normalization term in a bad way. The concept of the model is similar to the standard sequence-dependent retrieval bias (Section 12.6.3):

$$\mathrm{P}(x_n|S,i,\theta) = \mathrm{P}_{\mathrm{dist}}(x_n - i)|\mathrm{retrieved},i,\theta)\,\mathrm{P}_{\mathrm{seq}}(\mathrm{retrieved}|S,x_n,\theta) \tag{12.35}$$

The only difference is the dependence of the sequence bias on x_n instead of i. Inserting this into the likelihood (Equation 12.6) leads to:

$$\mathcal{L} = \prod_{n=1}^{N}\left[\sum_{i=x_n-\delta}^{x_n+\delta} \mathrm{P}_{\mathrm{dist}}(x_n - i)|\mathrm{retrieved},i,\theta)\,\mathrm{P}_{\mathrm{seq}}(\mathrm{retrieved}|S,x_n,\theta)\,\mathrm{P}(i|S,\epsilon,\mu)\right]^{w_n} \tag{12.36}$$

I neglected to mention it in Section 12.2.1, but the separate normalization of the positional uncertainty and the nucleosome occurence relies on the fact that the sum can be separated into independent factors of those terms. That trick cannot be applied here, making the independent normalization terms a bad approximation of the real normalization term. In brief, the approximation ignores the unequal sequence distribution between positions in proximity of nucleosome dyads and others. As this whole work is about learning sequence-preferences of nucleosomes that is an unlikely assumption.

I would have been able to develop the correct model with the exact normalization term and implement it, but I gave it little thought and did not even bother finding out

how strongly the time complexity would be affected. I first implemented the other two versions and after finding out the other sequence-dependent retrieval bias (Section 12.6.3) was promising, there was no need for this model.

12.6.6 Competing DNA-binding factors

Extensions of the thermodynamic model with competing DNA-binding factors have been published (Wasson and Hartemink, 2009). I developed an equivalent extension for the thermodynamic model of my method. The gradient ascent can optimize the parameters of the new DNA-binding factors in a similar way as the nucleosome binding parameters. I added F^t and B^t terms to the Forward/Backward algorithm that describe the Forward and Backward statistical weights of the DNA-binding factor t:

$$\mathrm{P}(i|S,\epsilon,\mu) = \frac{Z_i}{\sum_{i'=1}^{L_S} Z_{i'}}$$

$$\begin{aligned} F_{i+1} &= F_i + F^*_{i-D_N} + \sum_{t\in T} F^t_{i-D_t} \\ F^*_{i+1} &= F_{i-D_N} e^{E_{i+1}-\mu} \\ F^t_{i+1} &= F_{i-D_t} e^{E^t_{i+1}-\mu^t} \end{aligned} \tag{12.37}$$

$$\begin{aligned} B_{i-1} &= B_i + B^*_{i+D_N} e^{E_{i+D_N}-\mu} + \sum_{t\in T} B^t_{i+D_t} e^{E^t_{i+D_t}-\mu^t} \\ B^*_{i-1} &= B_{i+D_N} \\ B^t_{i-1} &= B_{i+D_t} \end{aligned}$$

The partial derivatives of Forward/Backward (only showing the changes):

$$\begin{aligned}
\frac{\partial F_{i+1}}{\partial \epsilon_l(q)} &= \frac{\partial F_i}{\partial \epsilon_l(q)} + \frac{\partial F^*_{i-D_N}}{\partial \epsilon_l(q)} + \sum_{t \in T} \frac{\partial F^t_{i-D_t}}{\partial \epsilon_l(q)} \\
\frac{\partial F^t_{i+1}}{\partial \epsilon_l(q)} &= \frac{\partial F_{i-D_t}}{\partial \epsilon_l(q)} e^{E^t_{i+1} - \mu^t}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial B_{i-1}}{\partial \epsilon_l(q)} &= \frac{\partial B_i}{\partial \epsilon_l(q)} + \left[\frac{\partial B^*_{i+D_N}}{\partial \epsilon_l(q)} + B^*_{i+D_N} \, \mathrm{I}(s_{i+D_N+l-k}, ..., s_{i+D_N+l} = q) \right] e^{E_{i+D_N} - \mu} \\
&\quad + \sum_{t \in T} \frac{\partial B^t_{i+D_t}}{\partial \epsilon_l(q)} e^{E^t_{i+D_t} - \mu^t} \\
\frac{\partial B^t_{i-1}}{\partial \epsilon_l(q)} &= \frac{\partial B_{i+D_t}}{\partial \epsilon_l(q)}
\end{aligned} \tag{12.38}$$

$$\begin{aligned}
-\frac{\partial F_{i+1}}{\partial \mu} &= -\frac{\partial F_i}{\partial \mu} - \frac{\partial F^*_{i-D_N}}{\partial \mu} - \sum_{t \in T} \frac{\partial F^t_{i-D_t}}{\partial \mu} \\
-\frac{\partial F^t_{i+1}}{\partial \mu} &= -\frac{\partial F_{i-D_t}}{\partial \mu} e^{E^t_{i+1} - \mu^t} \\
-\frac{\partial B_{i-1}}{\partial \mu} &= -\frac{\partial B_i}{\partial \mu} + \left[-\frac{\partial B^*_{i+D_N}}{\partial \mu} + B^*_{i+D_N} \right] e^{E_{i+D_N} - \mu} - \sum_{t \in T} \frac{\partial B^t_{i+D_t}}{\partial \mu} e^{E^t_{i+D_t} - \mu^t} \\
-\frac{\partial B^t_{i-1}}{\partial \mu} &= -\frac{\partial B_{i+D_t}}{\partial \mu}
\end{aligned} \tag{12.39}$$

The extension does not affect the log-likelihood of the nucleosome measurements. The partial derivatives of ϵ^t and μ^t are equivalent to those of ϵ and μ with $*$ and specific t swapped, except – important – for the log-likelihood where F^* and B^* remain. They reflect that the probabilistic model describes the measurement of nucleosomes, not the other DNA-binding factors.

To include measurements of other factors the log-likelihood could be extended, but this goes beyond the scope of this work. I never went beyond testing the probabilistic model with competing factors, because addressing the issues of the quantitative interpretation of the nucleosome measurements was more crucial.

12.6.7 Modeling CC-Seq fragments

CC-Seq fragments span between two neighboring nucleosomes centers as described in Section 3.2. The probability of seeing a fragment end for a nucleosome depends on the frequency with which its neighboring nucleosomes are bound. Similarly, possible sequence biases at either end influence the measurements of both nucleosomes. Paired-end sequencing identifies both fragment ends and I developed the equations of probabilistic data model that uses this information. However, I never got around to implementing this version, because it requires restructuring of some basic elements.

In this probabilistic data model the measurement x_n describes a fragment from $x_{n,1}$ to $x_{n,2}$ that stems from two neighboring nucleosomes.

$$\begin{aligned}\mathcal{L} &= \prod_{n=1}^{N} \mathrm{P}(x_n\text{measured}|S,\epsilon,\mu)^{w_n} \\ &= \prod_{n=1}^{N} \left[\sum_{i=x_{n,1}-\delta_1}^{x_{n,1}+\delta_2} \sum_{j=x_{n,2}+\delta_1}^{x_{n,2}-\delta_2} \mathrm{P}(x_n\text{measured}|\text{nucleosomes at i,j})\,\mathrm{P}(i,j|S,\epsilon,\mu) \right]^{w_n}\end{aligned} \tag{12.40}$$

Let me first show $\mathrm{P}(x_n\text{measured}|\text{nucleosome at } i,j)$, which is the paired positional uncertainty, using the position-specific deconvolution:

$$\begin{aligned}\mathrm{P}(x_n\text{measured}|\text{nucleosomes at } i,j) &= \mathrm{P}(x_{n,1}-i|\delta)\,\mathrm{P}(-(x_{n,2}-j)|\delta) \\ &\propto \zeta_{x_{n,1}-i}\zeta_{-(x_{n,2}-j)} \\ \sum_{k=-\delta_1}^{+\delta_2}\sum_{m=-\delta_1}^{+\delta_2}\zeta_k\zeta_m &= \frac{1}{A_\zeta} \\ \mathrm{P}(x_n\text{measured}|\text{nucleosomes at } i,j) &= A_\zeta\zeta_{x_{n,1}-i}\zeta_{-(x_{n,2}-j)}\end{aligned} \tag{12.41}$$

The paired positional uncertainty is a straight forward combination of two position-specific deconvolutions. This assumes that the fragment length has no bias, i.e. the probability to measure a fragment is independent of its length. I would reduce length biases by computationally restricting the fragment sizes more strictly than the experimental protocol. In CC-Seq fragments of an approximate length are extracted from after a gel electrophoresis separation. The imprecision of the experimental method leads to the underrepresentation of lengths closer to the extraction borders. Choosing conservative $L_{F_{\min}}$ and $L_{F_{\max}}$ length restrictions for the fragment sizes could remove the most biased regions in proximity of the borders. The time complexity increase depends on the range of allowed sizes, making a narrower range of modeled fragment sizes advantageous.

Next we come to the probability of seeing the two nucleosomes:

$$\mathrm{P}(i,j|S,\epsilon,\mu) = \frac{Z_{i,j}}{\sum_{i'=1}^{L_S}\sum_{j'=i'+L_{F_{\min}}}^{i'+L_{F_{\max}}} Z_{i',j'}}$$

$$\begin{aligned} Z_{i,j} &= F_i^* B_j^* \\ \mathrm{P}(i,j|S,\epsilon,\mu) &= \frac{F_i^* B_j^*}{\sum_{i'=1}^{L_S}\sum_{j'=i'+L_{F_{\min}}}^{i'+L_{F_{\max}}} F_{i'}^* B_{j'}^*} \end{aligned}$$

$$\begin{aligned} F_0 &= 1, \quad F_0^* = 0 \quad \text{(Initialization)} \\ F_{i+1} &= F_i + F_{i-D_N}^* \\ F_{i+1}^* &= F_{i-D_N} e^{E_{i+1}-\mu} \end{aligned} \tag{12.42}$$

$$\begin{aligned} B_{L_S+1} &= 1, \quad B_{L_S+1}^* = 0 \quad \text{(Initialization)} \\ B_{i-1} &= B_i + B_{i+D_N}^* \\ B_{i-1}^* &= B_{i+D_N} e^{E_{i-1}-\mu} \end{aligned}$$

Note that the Backward statistical weights B_i^* now include the binding energy of the nucleosome at position i, making them exact reverse versions of the Forward weights.

The above equations assumes the two nucleosomes are neighbors and, therefore, only works for a range of distances. The two nucleosomes are not allowed to be so close they overlap nor so far apart further nucleosomes could bind between them. The reason for this restriction is that $Z_{i,j} = F_i^* B_j^*$ relies on $B_{i+D_N|\text{nucleosome at } j} = B_j^*$ (or vice versa for F_i^*) to be valid. To model longer fragments, which allow for extra nucleosomes between the two, one would have to compute the Forward or Backward stretch between the nucleosome positions given one of the nucleosome positions. I believe the benefit of modeling longer fragments does not warrant the accompanying increase in computation time. A similar calculation of the Forward and Backward algorithm with additional conditions is described in Section 20.1 in the context of methylation data.

Inserting Equations 12.41 and 12.42 into Equation 12.40 and logarithmizing the result

brings us to the log-likelihood:

$$\begin{aligned}
\log \mathcal{L} &= \sum_{n=1}^{N} \log \left[\sum_{i=x_{n,1}-\delta_1}^{x_{n,1}+\delta_2} \sum_{j=x_{n,2}+\delta_1}^{x_{n,2}-\delta_2} A_\zeta \zeta_i \zeta_j \, \mathrm{P}(i,j|S,\epsilon,\mu) \right] w_n \\
&= \sum_{n=1}^{N} \left(\log \left[\sum_{i=x_{n,1}-\delta_1}^{x_{n,1}+\delta_2} \sum_{j=x_{n,2}+\delta_1}^{x_{n,2}-\delta_2} \zeta_i \zeta_j F_i^* B_j^* \right] \right. \\
&\quad \left. - \log \left[\sum_{i=1}^{L_S} \sum_{j=i+L_{F_{\min}}}^{i+L_{F_{\max}}} F_i^* B_j^* \right] + \log A_\zeta \right) w_n \\
&= \sum_{n=1}^{N} \log \left[\sum_{i=x_{n,1}-\delta_1}^{x_{n,1}+\delta_2} \sum_{j=x_{n,2}+\delta_1}^{x_{n,2}-\delta_2} \zeta_i \zeta_j F_i^* B_j^* \right] w_n \\
&\quad - \log \left[\sum_{i=1}^{L_S} \sum_{j=i+L_{F_{\min}}}^{i+L_{F_{\max}}} F_i^* B_j^* \right] \sum_{n=1}^{N} w_n + \log A_\zeta \sum_{n=1}^{N} w_n
\end{aligned} \tag{12.43}$$

The partial derivatives of the Backward terms change and are now the reverse of the partial derivatives of the Forward terms:

$$\begin{aligned}
\frac{\partial B_{L_S+1}}{\partial \epsilon_l(q)} &= 0, \quad \frac{\partial B^*_{L_S+1}}{\partial \epsilon_l(q)} = 0 \\
\frac{\partial B_{i-1}}{\partial \epsilon_l(q)} &= \frac{\partial B_i}{\partial \epsilon_l(q)} + \frac{\partial B^*_{i+D_N}}{\partial \epsilon_l(q)} \\
\frac{\partial B^*_{i-1}}{\partial \epsilon_l(q)} &= \left[\frac{\partial B_{i+D_N}}{\partial \epsilon_l(q)} + B^*_{i+D_N} \, \mathrm{I}(s_{i-1+l-k}, ..., s_{i-1+l} = q) \right] e^{E_{i-1}-\mu}
\end{aligned} \tag{12.44}$$

$$\begin{aligned}
-\frac{\partial B_{L_S+1}}{\partial \mu} &= 0, \quad -\frac{\partial B^*_{L_S+1}}{\partial \mu} = 0 \\
-\frac{\partial B_{i-1}}{\partial \mu} &= -\frac{\partial B_i}{\partial \mu} - \frac{\partial B^*_{i+D_N}}{\partial \mu} \\
-\frac{\partial B^*_{i-1}}{\partial \mu} &= \left[-\frac{\partial B_{i+D_N}}{\partial \mu} + B_{i+D_N} \right] e^{E_{i-1}-\mu}
\end{aligned} \tag{12.45}$$

For the partial derivatives of the log-likelihood I introduce $f^{\zeta*}_{x_n}$, $b^{\zeta*}_{x_n}$, f^*_i and b^*_i as short hands for sums. Note that the sums these variables describe have different index ranges. While they hopefully increase the readability of the equation, their main purpose is to

highlight terms that can be precomputed once for all partial derivatives.

$$\begin{aligned}
\frac{\partial \log \mathcal{L}}{\partial \epsilon, \mu} &= \sum_{n=1}^{N} \frac{\sum_{i=x_{n,1}-\delta_1}^{x_{n,1}+\delta_2} \sum_{j=x_{n,2}+\delta_1}^{x_{n,2}-\delta_2} \zeta_i \zeta_j \left(B_j^* \partial F_i^* + F_i^* \partial B_j^* \right)}{\sum_{i=x_{n,1}-\delta_1}^{x_{n,1}+\delta_2} \sum_{j=x_{n,2}+\delta_1}^{x_{n,2}-\delta_2} \zeta_i \zeta_j F_i^* B_j^*} w_n \\
&\quad - \frac{\sum_{i=1}^{L_S} \sum_{j=i+L_{F_{\min}}}^{i+L_{F_{\max}}} \left(B_j^* \partial F_i^* + F_i^* \partial B_j^* \right)}{\sum_{i=1}^{L_S} \sum_{j=i+L_{F_{\min}}}^{i+L_{F_{\max}}} F_i^* B_j^*} \sum_{n=1}^{N} w_n \\
&= \sum_{n=1}^{N} \frac{\sum_{i=x_{n,1}-\delta_1}^{x_{n,1}+\delta_2} \zeta_i \partial F_i^* \sum_{j=x_{n,2}+\delta_1}^{x_{n,2}-\delta_2} \zeta_j B_j^* + \sum_{j=x_{n,2}+\delta_1}^{x_{n,2}-\delta_2} \zeta_j \partial B_j^* \sum_{i=x_{n,1}-\delta_1}^{x_{n,1}+\delta_2} \zeta_i F_i^*}{\sum_{i=x_{n,1}-\delta}^{x_{n,1}+\delta} \zeta_i F_i^* \sum_{j=x_{n,2}+\delta_1}^{x_{n,2}-\delta_2} \zeta_j B_j^*} w_n \\
&\quad - \frac{\sum_{i=1}^{L_S} \partial F_i^* \sum_{j=i+L_{F_{\min}}}^{i+L_{F_{\max}}} B_j^* + \sum_{j=1}^{L_S} \partial B_j^* \sum_{i=j+L_{F_{\min}}}^{j+L_{F_{\max}}} F_i^*}{\sum_{i=1}^{L_S} F_i^* \sum_{j=i+L_{F_{\min}}}^{i+L_{F_{\max}}} B_j^*} \sum_{n=1}^{N} w_n \\
&= \sum_{n=1}^{N} \frac{\sum_{i=x_{n,1}-\delta_1}^{x_{n,1}+\delta_2} \zeta_i \partial F_i^* b_{x_n}^{\zeta *} + \sum_{j=x_{n,2}+\delta_1}^{x_{n,2}-\delta_2} \zeta_j \partial B_j^* f_{x_n}^{\zeta *}}{f_{x_n}^{\zeta *} b_{x_n}^{\zeta *}} w_n \\
&\quad - \frac{\sum_{i=1}^{L_S} \partial F_i^* b_i^* + \sum_{j=1}^{L_S} \partial B_j^* f_j^*}{\sum_{i=1}^{L_S} F_i^* b_i^*} \sum_{n=1}^{N} w_n
\end{aligned} \tag{12.46}$$

$$\begin{aligned}
\frac{\partial \log \mathcal{L}}{\partial \zeta_l} &= \sum_{n=1}^{N} \frac{\sum_{i=x_{n,1}-\delta_1}^{x_{n,1}+\delta_2} \zeta_i F_i^* B_l^* + \sum_{j=x_{n,2}+\delta_1}^{x_{n,2}-\delta_2} \zeta_j F_l^* B_j^*}{\sum_{i=x_{n,1}-\delta_1}^{x_{n,1}+\delta_2} \sum_{j=x_{n,2}+\delta_1}^{x_{n,2}-\delta_2} \zeta_i \zeta_j F_i^* B_i^*} w_n + \frac{\partial}{\partial \zeta_l} \log A_\zeta \sum_{n=1}^{N} w_n \\
&= \sum_{n=1}^{N} \frac{f_{x_n}^{\zeta *} B_l^* + F_l^* b_{x_n}^{\zeta *}}{f_{x_n}^{\zeta *} b_{x_n}^{\zeta *}} w_n + \frac{\partial}{\partial \zeta_l} \log A_\zeta \sum_{n=1}^{N} w_n
\end{aligned}$$

with:

$$\begin{aligned}
\frac{\partial}{\partial \zeta_l} \log A_\zeta &= \frac{\partial}{\partial \zeta_l} \log \frac{1}{\sum_{k=-\delta_1}^{+\delta_2} \sum_{m=-\delta_1}^{+\delta_2} \zeta_k \zeta_m} \\
&= -\frac{\sum_{k=-\delta_1}^{+\delta_2} \sum_{m=-\delta_1}^{+\delta_2} \zeta_k \zeta_m}{(\sum_{k=-\delta_1}^{+\delta_2} \sum_{m=-\delta_1}^{+\delta_2} \zeta_k \zeta_m)^2} \left(\sum_{k=-\delta_1}^{+\delta_2} \zeta_k + \sum_{m=-\delta_1}^{+\delta_2} \zeta_m \right) \\
&= -2 A_\zeta \sum_{k=-\delta_1}^{+\delta_2} \zeta_k
\end{aligned} \tag{12.47}$$

While the equations are more complex, the time complexity is less affected then I initially expected. As mentioned above, $f_{x_n}^{\zeta *}$, $b_{x_n}^{\zeta *}$, f_i^* and b_i^* do not rely on the partial derivative and just need to be computed once per iteration. Therefore, the time complexity of computing all log-likelihood partial derivatives only increases by a constant factor, instead of multiplicative factors of $\delta_1 + \delta_2 + 1$ or $L_{F_{\max}} - L_{F_{\min}}$. This assumes that $\delta_1 + \delta_2 + 1$ and $L_{F_{\max}} - L_{F_{\min}}$ are $\ll 4^k L_N$, otherwise the precomputed terms start having a noticeable

time complexity.

12.6.8 Training on more than one dataset simultaneously

In Section 12.6.6 I mention the possibility of applying the probabilistic model to learn parameters of other DNA-binding factors. This is indirectly possible via the log-likelihood of nucleosome measurements or directly by formulating the likelihood for measurements of the other DNA-binding factors. To optimize such likelihoods in parallel with the nucleosome likelihood the weights w_n of the datasets have to be on the same scale, but otherwise there are no issues. I feel that such a method would be a niche application and would need greatly improved measurements to deliver improved results. However, using the same idea to optimize a nucleosome binding model simultaneously on distinct nucleosome measurements sounds promising.

Using my method to optimize energy models on distinct nucleosome measurements individually produces high-resolution models that are similar (Section 13.5). Combining the optimizations would force the convergence on a single energy model and help the probabilistic data model separate this nucleosome energy model from sequence biases of the different measurements. Data with high-positional resolution could also provide a sharp energy model for data with low-positional resolution – still a frequent issue for MNase-Seq models (Section 16.2).

For the simplest case without sequence biases the joint likelihood is:

$$\mathcal{L} = \prod_{y \in Y} \prod_{n=1}^{N} \left[\sum_{i=1}^{N} \mathrm{P}(x_{y,n}\text{measured}|\text{nucleosome at } i, \theta_y)\, \mathrm{P}(i|S, \epsilon, \mu) \right]^{w_{y,n}} \tag{12.48}$$

Where y is the dataset from which the measurement comes. P(nucleosome at $i|S, \epsilon, \mu$) is independent of the dataset y, while P($x_{y,n}$measured|nucleosome at i, θ_y) depends on the dataset. Each dataset has its own probabilistic data model to describe its experimental errors. The probabilistic data models are not restricted to just having unique parameters θ_y, but can have distinct positional-uncertainty and sequence-bias models.

Extending my method to allow such a simultaneous optimization on multiple datasets requires drastic code restructuring. This went beyond what was possible during my PhD studies. Without improved measurements with less sequence biases I believe this will be an important step moving forward.

12.7 Estimate of the reasonable average genomic nucleosomes occupancy

I estimated a reasonable average genomic nucleosome occupancy based on a few basic assumption, like that most of the genome is covered by nucleosomes. Such an approximation was used by Segal et al. (2006) to estimate a 75-90% average occupancy. They used linker lengths of 10-50-bps measured in nucleosome arrays (van Holde, 1989). I improved the approximation by using *in vivo* measured linker lengths, including nucleosome-free regions, and allowing for partially absent nucleosomes.

The average linker length (l_{link}) of 30 bps is based on CC-Seq data that measures dyad to dyad distances directly (Brogaard et al., 2012). I rounded the average linker length upwards, because I am more interested in the lower boundary of the approximated average occupancy. Based on yeast, I use a genome length (l_{genome}) of 12,156,677 bps with 6,275 genes (n_{genes}). As nucleosome length (l_{nuc}) I use the conventional 147 bps.

The average occupancy (occ) of 'fully' covered regions runs to ~83% and a fully covered genome would be bound by 68682 nucleosomes (n_{nuc}).

$$\begin{aligned} \text{occ} &= \frac{l_{\text{nuc}}}{l_{\text{nuc}} + l_{\text{link}}} = \frac{147}{147 + 30} \approx 0.83 \\ n_{\text{nuc}} &= \frac{l_{\text{genome}}}{l_{\text{nuc}} + l_{\text{link}}} = \frac{12156677}{177} \approx 68682 \end{aligned} \tag{12.49}$$

Assuming one nucleosome-free region (0% occupancy) per gene gives an average occupancy of ~75%.

$$\text{occ} = \frac{(n_{\text{nuc}} - n_{\text{genes}})l_{\text{nuc}}}{l_{\text{genome}}} = \frac{(68682 - 6275) * 147}{12156677} \approx 0.75 \tag{12.50}$$

Further assuming that every nucleosome position has a 90% chance of a nucleosome being bound (i.e. 10% absence) gives an average occupancy of ~68%.

$$\text{occ} = \frac{0.9(n_{\text{nuc}} - n_{\text{genes}})l_{\text{nuc}}}{l_{\text{genome}}} = \frac{0.9(68682 - 6275) * 147}{12156677} \approx 0.68 \tag{12.51}$$

The 90% chance of the average nucleosome to be bound has not been direct measurements for a large sample of nucleosomes. (If that were the case this approximation would probably be unnecessary). MTase measurements at the Pho5 promoter measured frequencies above 90% for three nucleosomes (Small et al., 2014). I chose the 90% based on this

and the fact that chromatin-accessibility assays suggest the genome is mostly covered by nucleosomes and inaccessible.

12.8 Additional Details for the figures

Figure 13.1

The Pearson correlation coefficients are computed between normalized apparent nucleosome occupancies, i.e. derived from the measurements assuming the recovery frequency reflected the nucleosome occurrence frequency. Subfigure A is produced with an adapted version of the corrplot function (Wei, 2013).

For the MNase-Seq measurements the fragment centers are approximated from the sequenced fragment ends and used as dyad positions, if no processed dyad positions were published. For the CC-Seq data the published nucleosome scores were used as dyad frequencies. Other datasets from experiments that sequenced the fragments were treated like the MNase-Seq measurements. The dyad frequencies are smoothed over a 147-bp window to approximate nucleosome coverage. For the chip experiments, whose measurements already represent coverage, the data was not smoothed. All datasets were normalized to a genome-wide average value of 1.

The preprocessed data is used from the publications and transfered (via the lift-over tool) to the newest yeast genome (sacCer3). The mapping of the raw reads was not repeated to conserve effects different data processing could have. Unmappable genomic regions were excluded from all analysis.

For the G+C content I smoothed the C+G frequency with a 147-bp window twice. The first smoothing produces the G+C content of the nucleosome covered region for each dyad position, and the second smoothing matches the coverage smoothing of the other datasets.

Figure 13.2

Histograms of the normalized apparent nucleosome occupancies. The datasets were processed as described above for Figure 13.1.

The Poission noise distribution was estimated based on the average coverage. This assumes the positions are all independent, which they are not due to the smoothing. However, tests with Poission sampling and smoothing revealed that sampling coverage values directly or smoothing sampled dyads have similar distribution widths. Therefore,

the approximation is good and the conclusions hold even if the noise distribution were modeled more precisely.

Figure 13.3

Pearson correlation coefficients between dyad probability predictions and dyad position frequencies without smoothing. The dyad position frequencies were derived from the measurements as described for Figure 13.1 (without smoothing).

The method of Kaplan et al. (2009) was downloaded from their website and ran on the sacCer3 genome with their default parameters. No cross-validation was performed for their method.

My prediction method was trained on the *in vitro* dataset of Kaplan et al. (2009) in the genome direction without symmetrizing the energy model parameters. The predictions where then performed on the reverse complement genome, which is practically independent. To make sure the method was not overfitting, I left one chromosome out of the optimization as an independent validation and its results confirmed that there is no issue.

Figure 13.4

The reads mapped to either strand and the deconvolution tool were taken from Brogaard et al. (2012). For this figure the single template model was used to deconvolution the data. For Figure B.3 the four template model was used, which has distinct cut distributions based on the presence and absence of A at the -3 and T at the +3 position in regards to the dyad. The genomic positions are weighted by the computed nucleosome score to compute the nucleotide frequencies, as one would to generate a PWM from numeric measurements. No cutoff or other processing step was applied. As expected, the difference is the frequencies derived from the Crick strand data subtracted from those derived from the Watson strand data.

Figure 13.5

As described in Section 12.1.2, the model parameters ϵ are defined in a way to conserve the expected dyad symmetry. For $i < 0$ the conditional probability is from the left nucleotide to the right nucleotide, while for $i > 0$ the conditional probability is mirrored. In subfigures E, F, and G the model parameters are grouped by representing the same conditional probability, i.e. $\mathrm{P}(s_i = a|s_j = b)$ where j is $i + 1$ for $i < 0$ and $i - 1$ for $i > 0$.

12.9 Implementation

I originally implemented the core of my method in C and later converted most of it to C++. For time critical parts I maintained two versions, one that uses normal instructions and one that uses parallel operations on the instruction level (SIMD). The normal version was used for validation and bugfixing. The precision of the calculation can be set between single precision (float), double precision (double) and double extended (long double) to check for numerical errors. The SIMD version uses float SSE2 intrinsics to speed up the calculations as much as possible. I further parallelized the partial derivative computations, which are independent of each other, with OpenMP – a multi-threading framework for single computation nodes (OpenMP Architecture Review Board, 2008). Thanks to the rise of multi-core CPUs and hundreds of partial derivatives that need to be computed, this can decrease the runtime by an order of magnitude.

For the other analysis and figure generation I used R (R Core Team, 2016). I used a wrapper R script to call my method for convenience and ran jobs on our cluster via the slurm framework (Jette et al., 2002). I analyzed the data in R with the help of several packages, primarily: Biostrings (Pages et al., 2016), ff (Adler et al., 2014), ffbase (de Jonge et al., 2015), and packages I developed myself.

13. Results

13.1 Strong biases dominate genome-wide nucleosome occupancy measurements

I systematically compared nucleosome datasets of several publications with Pearson's correlation coefficient between the derived genomic occupancy vectors, as proposed by Kaplan et al. (2010a). The nucleosome occupancy is the probability that a position is wrapped in a nucleosome. I selected datasets that use distinct methods to measure nucleosomes to analyze a broad spectrum of experiments and added MNase-Seq datasets with conditions expected to affect genome-wide nucleosome positioning to compare the bias and signal strengths (Brogaard et al., 2012; Kaplan et al., 2009; Field et al., 2008; Shivaswamy et al., 2008; Gossett and Lieb, 2012; Celona et al., 2011; Mavrich et al., 2008; Lee et al., 2007; Fan et al., 2010; Zhang et al., 2009).

The correlation coefficients between genomic nucleosome occupancies derived from MNase-Seq measurements of untreated wild-type cells are generally low (median of 0.51; green triangle in Figure 13.1A) given that they effectively represent replicate measurements (Table B.1). The replicates of Kaplan et al. (2009) have higher correlations (median of 0.88; Figure 13.1C, which also compares the framed correlations of Figure 13.1A)). This suggests that the low correlations are not due to experimental noise, but technical details that strongly influence the measurements creating biases. The correlations between measurements of MNase-Seq and other experimental protocols (median of 0.29; brown rectangle in Figure 13.1A) are even lower than those between the MNase-Seq measurements, and lower than correlations between wild-type and *in vitro* MNase-Seq measurements (median of 0.46, orange frame in Figure 13.1A). Varying experimental biases due to technical details are probably the reason for the low correlations between measurements of untreated wild-type cells.

Correlations between datasets of untreated wild-type and heat shock or knock-down cells from the same publication are among the highest (compare cyan to the other values

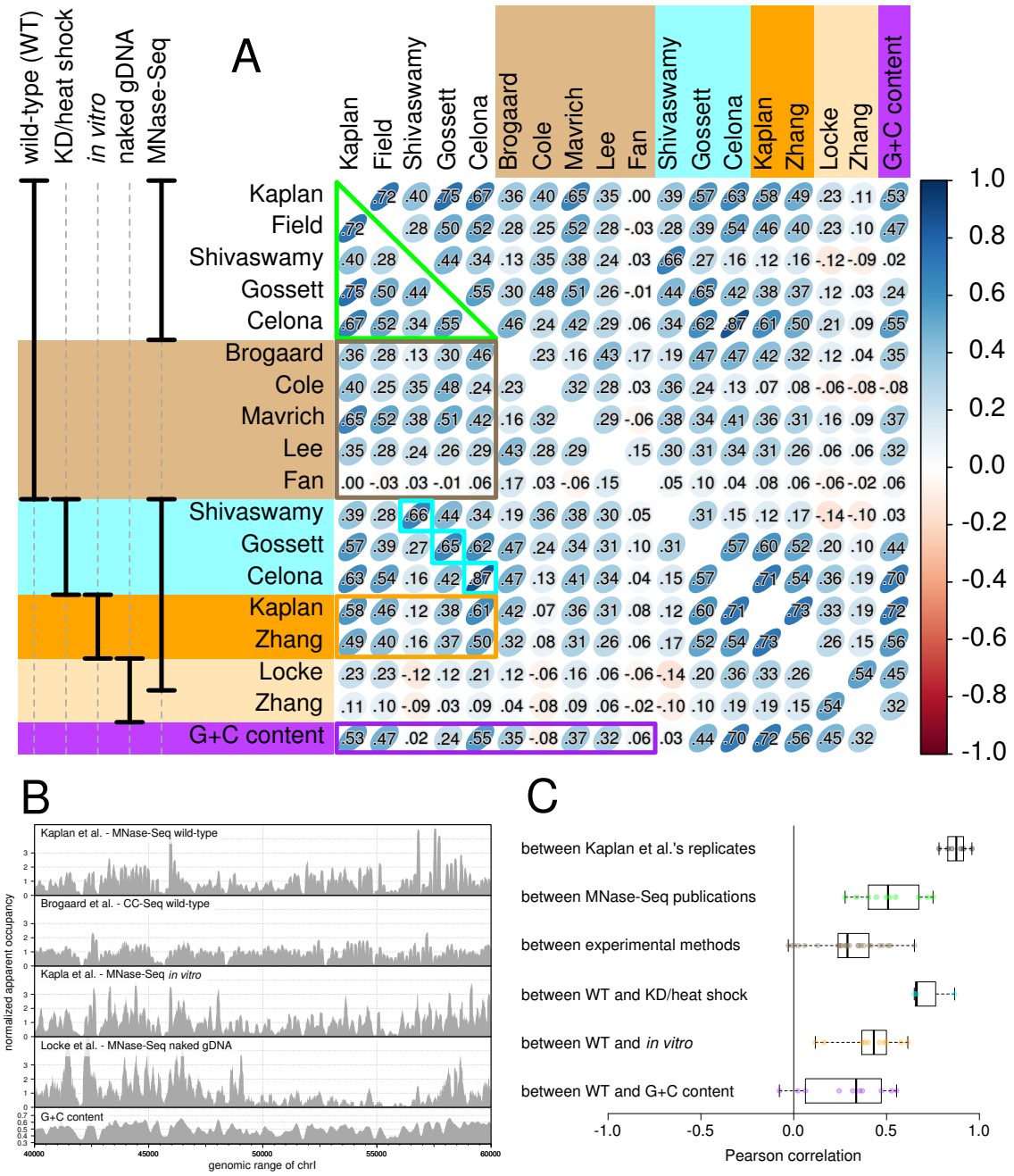

Figure 13.1: **Correlations between experimental nucleosome data reveal strong biases**: (A) Matrix of Pearson's correlations coefficients between occupancies derived from nucleosome measurements, control experiments and G+C content for *S.cerevisia*. The annotation on the left describes shared features of the used experiments. (B) Individual occupancy tracks for a subset of the measurements and G+C content. (C) Boxplots of grouped comparisons outlined by the same color in (A).

in Figure 13.1A,C), even though the knocked-down chromosomal proteins (H3 and nhp6) take part in the nucleosome formation (Gossett and Lieb, 2012; Celona et al., 2011). Batch effects, which lead to increased correlations in datasets produced in one batch or from one laboratory, are known from gene expression measurements (Leek et al., 2010). These results show that they affect MNase-Seq measurements as well. The batch effects are probably also partially responsible for the stronger correlations between the replicate measurements than between the MNase-Seq measurements from different publications.

The correlation of wild-type nucleosome measurements to G+C content varies strongly (purple in Figure 13.1A,C), and the measurements therefore disagree on the importance of the most basic sequence feature. Three datasets have reasonable correlations ($r = 0.53$, 0.47, and 0.55) suggesting the G+C content is a major determining feature of the nucleosome positioning. However, three datasets have barely any or negative correlations ($r = 0.02$, -0.08, and 0.06), which would suggest that nucleosomes are positioned independent of G+C content. This contradiction demonstrates that the available nucleosome measurements form no consensus on the importance of sequence features for nucleosome positioning.

Our analysis shows that the experimental biases are of the same order of magnitude or larger than the observed difference between wild-type and knock-down, heat shock or *in vitro* measurements. Any information gained from the measurements without separating out the biases contains at least as much bias as signal. One illustration of this issue is the strong variation of the measurement's correlations to G+C content, which range from $r = -0.08$ to 0.55.

13.2 Genome-wide measurements of nucleosome positions indicate an unrealistically-low nucleosome occupancy

To investigate other problems with the interpretation of nucleosome measurements, I analyzed the genome-wide distribution of signals commonly interpreted as nucleosome occupancies, measured by MNase-Seq and CC-Seq (Figure 13.2A,B). The common assumption is that these measurements are proportional to nucleosome occupancy, i.e. that the number of measured fragments at a genomic location is proportional to the fraction of cells in which this genomic location is covered by a nucleosome. Segal et al. (2006) estimated the average nucleosome occupancy in the range of 75-90% and I estimate it in

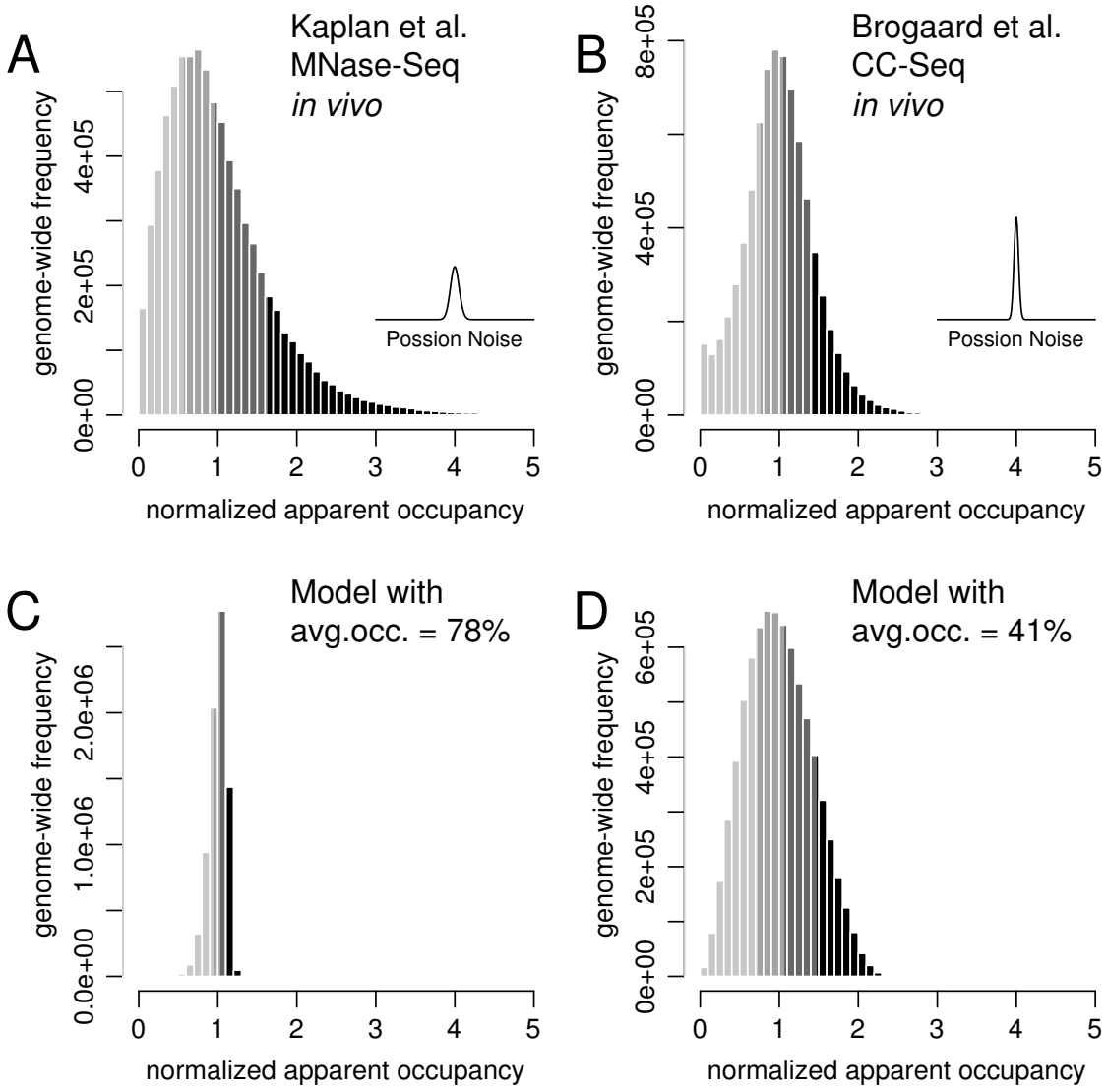

Figure 13.2: **Nucleosome occupancies deduced from MNase-Seq and CC-Seq have an unrealistic distribution**: Histograms of nucleosome occupancies deduced from measured fragment centers (A,B) and model predictions (C,D). The occupancies are normalized to a mean of 1 and the four shades depict the quartiles. Deriving occupancies directly from the MNase-Seq (A) and CC-Seq (B) measurements would imply an unrealistically-low nucleosome occupancy. (A,B) The inserts depict the expected Poission noise at 1 using the same x-axis scale. (C,D) As comparison I show predictions of our method without added Possion noise. The method was set to produce a realistic 78% (C) and an unrealistically-low 41% (D) occupancy.

the range of 68-83% (Section 12.7). Such average occupancies set a hard upper limit for fully occupied DNA at 1.2-1.5 times the average, which is missing for the measurement based distributions. All datasets I analyzed have heavy tails at high values (Figure B.1), which are not explained by sampling noise (line inserts in Figure 13.2A,B).

To get a better understanding what occupancy distribution is expected, I computed distributions predicted by my model, which includes steric exclusion and nucleosome sequence preferences. I chose a realistic average occupancy of 78% and to match the observed distribution better I also chose an unrealistically-low average occupancy of 41% (Figure 13.2C,D). The hard upper limit for high occupancies is obvious for my prediction with the average occupancy based on my estimate (Figure 13.2C). The distributions of all experiments are at least as wide as that of the unrealistically-low occupancy prediction and most have a heavier tail at high-occupancy values.

There are two possible explanations for the discrepancy between the expected and observed distributions: either the published and my genomic occupancy estimates are off by more than a factor of two, or the signals measured by MNase-Seq, CC-Seq, etc. are not unscaled measurements of nucleosome occupancy. The latter possibility is more probable given that my analysis above showed that the signal and bias of the measurements are of the same order of magnitude. While MNase-Seq measurements are typically still implicitly assumed to be proportional occupancy measurements to derive nucleosome energies models from them, it has been suggested before that read frequencies of MNase-Seq are not quantitative representations of nucleosome frequencies (Zhang et al., 2009).

Data with strong biases can make benchmarks of prediction methods misleading, because the biased data may lead to artificially high scores that represent the prediction of the biases, instead of the score reflecting the true performance of predicting the signal of interest. This issue occurs if the training data, on which the method parameters are optimized, contains the same bias as the test data, which is used to estimate the methods performance. To mitigate this problem the training and test datasets should be measured with two methods that have different technical biases. For this reason, I benchmark energy models obtained from MNase-Seq data against CC-Seq data. While CC-Seq and MNase-Seq both have a sequencing step, which can produce sequence-dependent biases (Harismendy et al., 2009), CC-Seq is otherwise distinct from MNase-Seq: instead of MNase digesting linker DNA the nucleosomal DNA is cleaved close to the dyad.

13.3 A high-resolution nucleosome energy model from MNase-Seq data

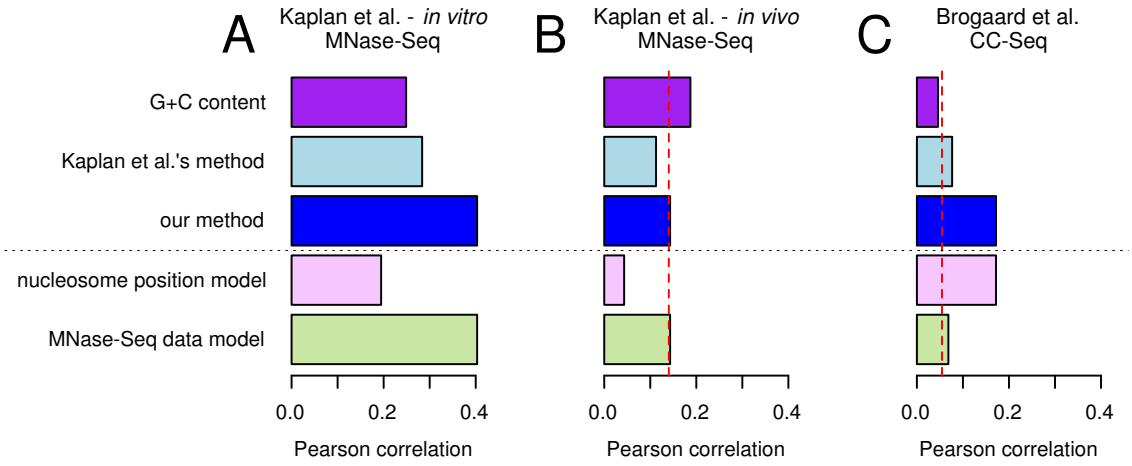

Figure 13.3: **My method performs well at predicting nucleosome positioning in base-pair resolution**: Pearson's correlation coefficients between prediction methods (rows) and experimental measurements (columns – subfigures A, B, C). As Kaplan et al.'s method, my method was trained on the *in vitro* dataset of Kaplan et al. (A) and the correlation of this dataset to the other two is shown by the dashed red lines (B,C). 'our method' (blue) is the relevant prediction of my model for the given test set. Below the dashed line I show both predictions of my model: the MNase-Seq data prediction (green) and nucleosome-position predictions (pink).

My method models both the nucleosome binding and experimental errors jointly to separate out positional uncertainty of the experiment. The first part of my method computes the probability of each genome position to be covered by a nucleosome dyad using a thermodynamic model (orange part in Figure 12.1A). The method further computes the probability of measuring a data point at a genomic position – here MNase-Seq reads (green part in Figure 12.1A). These MNase-Seq model probabilities contain the positional uncertainty of MNase-Seq and are therefore equivalent to the predictions of other methods, like the one from Kaplan et al., that are trained on MNase-Seq data and do not distinguish between a nucleosome position and its measurement. The distinction between the positioning of nucleosomes and their experimental measurement allows me to obtain a position-specific energy model from the low-resolution MNase-Seq data. My method iteratively learns both the nucleosome binding energy model and the positional-uncertainty

model by maximizing a log-likelihood (Section 12).

To compare my method to Kaplan et al.'s method (Kaplan et al., 2009) I trained my model parameters on the same *in vitro*, MNase-Seq dataset (Kaplan et al., 2009) as they used to train theirs. I compared the methods by computing Pearson's correlation coefficients between the predicted probabilities and several test datasets (Figures 13.3 and B.2). I used base-pair resolution, i.e. dyad measurements and predictions without smoothing – not occupancies, to highlight the resolution of the predictions. In this comparison our method has higher correlations than Kaplan et al.'s method to most datasets and equal correlations otherwise. NuPoP (Xi et al., 2010), another nucleosome prediction method I tested, has very low correlations in my comparison (Figure B.2).

Both the method of Kaplan et al. and mine perform best on their *in vitro* dataset (Figure 13.3A), which is expected given that the energy models of both methods were trained on this dataset. The correlation pattern of the two methods to MNase-Seq datasets are similar and match the correlation pattern of G+C content to the datasets. This suggests that even at single-base-pair resolution the test sets' correlation with G+C content is a major determinant of the methods' performance.

As mentioned earlier, MNase-Seq measurements share common biases that lead to higher correlations for prediction methods that predict these MNase-Seq biases. This flaw is revealed when comparing the two predictions of my model trained on MNase-Seq data with the MNase-Seq test sets: the correlations of my MNase-Seq-data model probabilities are more than double the matching correlations of my nucleosome-position probabilities. This means that adding the positional uncertainty – the only difference between the two probabilities – improves the correlation coefficients while decreasing the information content. The energy-model parameters of the method used by Kaplan et al. were average over three neighboring positions leading to a similar effect.

I used the *in vivo* CC-Seq dataset by Brogaard et al. (2012) as a test set, whose experimental protocol has little similarity to that of the training set (Figure 13.3C). Due to CC-Seq's high resolution in comparison to MNase-Seq, my nucleosome-position prediction doubles the correlation coefficients of the predictions of Kaplan et al. and my MNase-Seq-data model, which both contain the positional uncertainty of MNase-Seq. This confirms that my approach improves the nucleosome binding energy model by explicitly modeling the positional uncertainty.

13.4 CC-Seq has a strand-specific bias close to the dyad position

While the CC-Seq dataset published by Brogaard et al. has a higher positional resolution than MNase-Seq data, I showed with Figure 13.2 that neither have genome-wide distributions consistent with unbiased relative nucleosome occupancies. In comparison to nucleosomes measured with MNase-Seq, CC-Seq produces an A and T enrichment at the -3 and $+3$ positions from the nucleosome dyad, respectively. The authors originally conceded this could be a bias, but later claimed it validated an enrichment seen in a previous study (Brogaard et al., 2012; Xi et al., 2014). Cole et al. (2015) hypothesized several bias sources that could cause these A and T enrichments. I show that the enrichments are indeed a bias.

CC-Seq measures nucleosome positioning by cleaving the DNA close to the dyad, sequencing the produced fragment ends, and deconvolving these genome-wide counts based on the distribution of cuts around the nucleosome dyad. I ran the published sequence-independent deconvolution tool (single template model from Brogaard et al. (2012)) on the Watson- and Crick-strand fragment ends separately, i.e. distinguishing between which side of the cut site the fragment lies on. The nucleotide-frequency profiles surrounding the two dyad datasets reveal an asymmetric preference that must be an experimental bias, because of the nucleosome's point symmetry around the dyad position (Figure 13.4). While this asymmetry proves the presence of a sequence bias, it does not reveal how the bias is distributed between the sides and to what it is positioned.

I extended my method with two variants that describe sequence biases: positioned in regards to the measured nucleosome dyad, or independent measurements unrelated to nucleosomes (Sections 12.6.3 and 12.6.4). The first variant, where the sequence-bias model depends on the dyad position, separates the sequence bias from the nucleosome binding preference. The correlations between the variant and CC-Seq data are not meaningfully improved compared to those of the basic model, but the variant learns a cleaner energy model and shows that the enrichment of A and T – not its absence – is the bias.

The bias seems to originate from histone-DNA-enzyme interactions that influence the probability of cutting the DNA at the preferred -1 and $+6$ sites. Different cut-site distributions and increased overall cut frequencies influence the deconvolution score of a nucleosome. The authors observed that the cut-site distribution depended on the occurrence or absence of A at the -3 and T at the $+3$ positions and used four separate cut-site distributions to deconvolve the data with the intention of reducing possible biases.

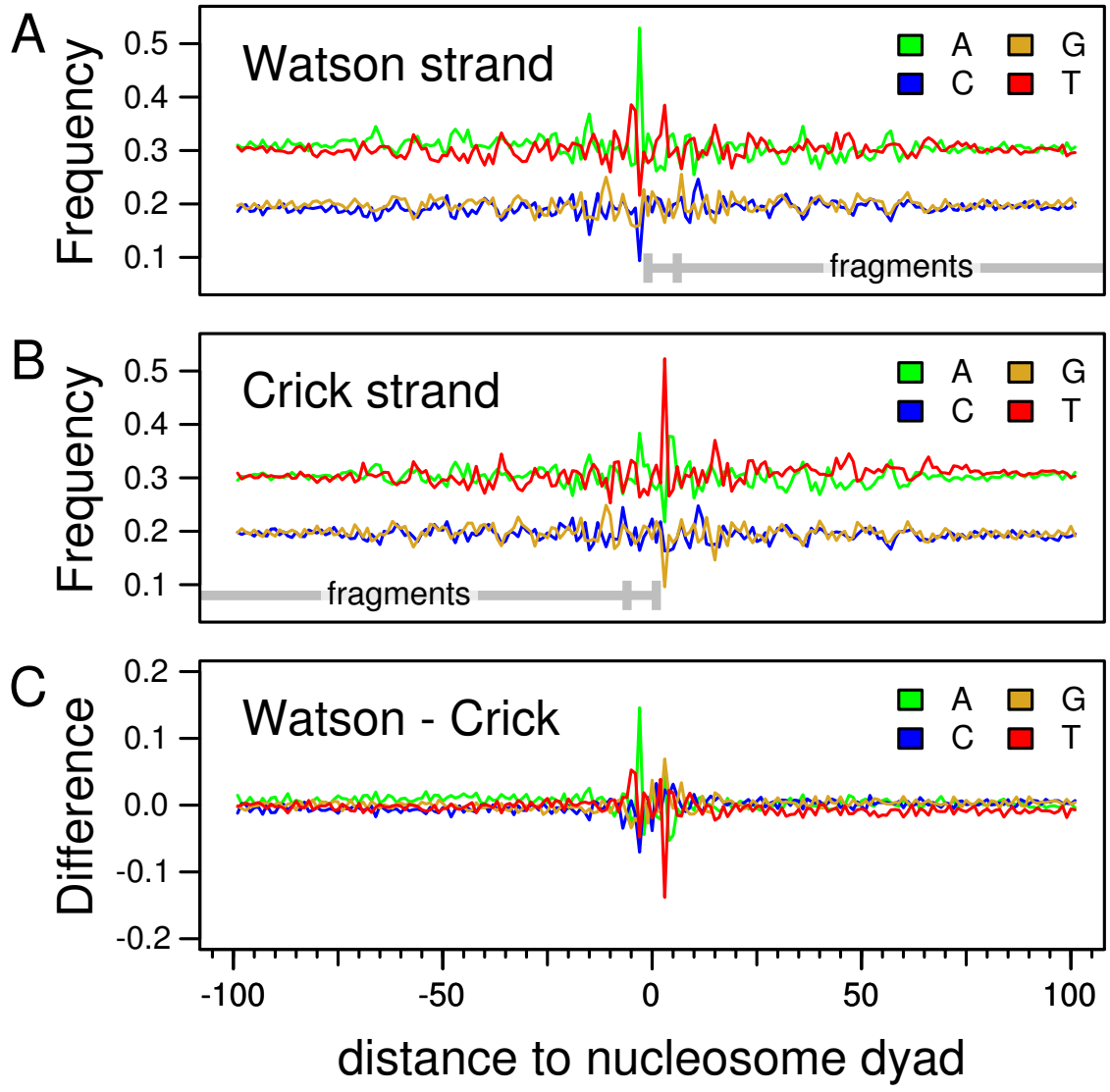

Figure 13.4: **CC-Seq has a strand-specific sequence bias**: Comparison of the nucleotide frequencies along the central nucleosome region for reads mapped to the Watson- and Crick-strand, i.e. the fragment's start and end with regards to the genome direction (depicted in gray below the subfigures). To obtain the strand-specific dyad position scores I separately deconvoluted the two strand datasets with the tool Brogaard et al. (2012) provided. (A, B) Frequency profiles for the Watson- and Crick-strand datasets. (C) Difference between the frequency profiles of (A) and (B). The peaks around the nucleosome dyads reveal strong sequence biases close to the cut site. There is also a minor general depletion of T vs A, C, and G over the region of the sequenced fragment.

While this improves the similarity between the strand data, the enrichments are still present (Figure B.3). Their sequence-dependent deconvolution apparently equalizes the bias of either strand instead of removing it.

13.5 Comparisons of energy models obtained from MNase-Seq and CC-Seq

Assuming that the data has position-specific resolution and the nucleosome binding preference only depends on the wrapped sequence, the nucleosome binding energy can be derived from the binding frequencies following Bolzmann's law. The parameters $\epsilon_i^{\mathrm{MNase}}(ab)$ of the 1st-order nucleosome binding energy model are initialized as the log ratio between the conditional frequency of dinucleotide ab at position i and the genomic background frequency (Section 12.1.2). The initial parameters of the models derived from the dinucleotide frequency around the MNase-Seq and CC-Seq nucleosome measurements differ strongly (Figure 13.5A, C). The amplitudes of the MNase-Seq model approximate a smoothed version of the CC-Seq model, which is expected given the lower resolution of MNase-Seq measurements. The energy model of CC-Seq contains the biased A enrichment at -3, which is absent in the MNase-Seq model. Together this leads to low Pearson's correlation coefficients of 0.30 between all nucleosome binding model parameters and the correlation coefficients between parameters of individual conditional probabilities along the 100 model positions span a low range between -0.07 and 0.47 (Figure 13.5F).

In contrast, after training my models the energy parameters obtained from the two datasets are more similar to each other (Figure 13.5B, D). The amplitude of the MNase-Seq energy model becomes more pronounced and jagged. This reveals that the commonly described 10-bp-periodic pattern, which is observed due to the low-positional resolution of MNase-Seq, is a smoothed version of the position-specific nucleosome preference. In the CC-Seq model the enrichment of A at the -3 positions is separated out of the nucleosome-energy parameters and described with the sequence-bias parameters, which further confirms that the enrichment is an experimental bias. The correlation coefficient between the two models improves to 0.51 and the correlations for individual conditional probabilities improve even more: falling into the range of 0.54 to 0.86 (Figure 13.5F).

While the two energy models learned by my method agree much better than the frequency derived models, there are still systematic differences between them (Figure 13.5B, D). One of the two main differences is the tendency of the MNase-Seq model to have higher parameters (binding preference) that represent conditional probabilities of G or C, while

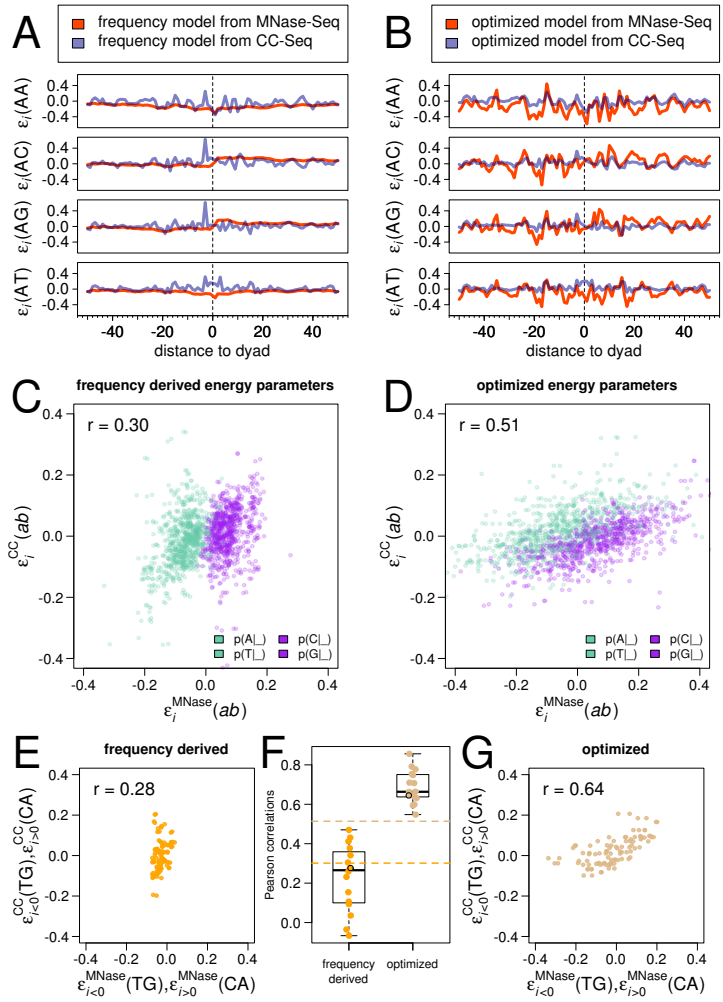

Figure 13.5: **The energy models optimized with my method from MNase-Seq and CC-Seq data have more in common than frequency-derived models.**: Comparison of the energy models obtained directly from dinucleotide frequencies (A, C) and by my method (B, D) from MNase-Seq and CC-Seq data. (A, B) Profiles of a subset of the energy parameters over the nucleosome region. (C, D) Scatterplot of all energy terms obtained from the two datasets with a color separation based on the conditional probability the parameters represent, to visualize the different preferences of G+C content. (F) Correlations between the energy parameters of the MNase-Seq and CC-Seq model for the energy models obtained directly from the frequencies (orange) and optimized by my method (brown). The boxplots and dots show the correlations between the energy parameters that represent a single conditional probability. The dashed line show the correlation between all parameters. (E, G) Scatterplot of one conditional probability marked by a black circle in (F). The energy parameters ϵ that describe one conditional probability in either nucleosome half are the reverse complement of each other because of the dyad symmetry of nucleosomes.

the CC-Seq model has the opposite tendency. The different preferences of G+C vs A+T content reflect the datasets' different correlations to G+C content shown in Figure 13.1A. I cannot learn such general G+C biases well without independent information, because such a bias is practically indistinguishable from the nucleosome binding preferences. The second main difference between the models is the roughly 3-fold larger amplitude of the MNase-Seq model's energy parameters. This larger amplitude might reflect the nucleosome binding preference being more pronounced *in vitro* than *in vivo*. Other disagreements between the two energy models might also originate from the MNase-Seq data being measured *in vitro*, while the CC-Seq data was measured *in vivo*.

14. Discussion

14.1 The average genomic nucleosome occupancy has not been measured experimentally

The nucleosome occupancy at a genomic position is the fraction of cells in which the position is covered by a nucleosome (Struhl and Segal, 2013). I showed that occupancies cannot be derived from genome-wide nucleosome measurements (Figure 13.2). In fact, the average genomic nucleosome occupancy is still unknown, while approximations have been given in literature. An 80% coverage of the genome by called nucleosome positions has been given (Lee et al., 2007; Shivaswamy et al., 2008) and has been called occupancy (Tillo and Hughes, 2009; Ozonov and van Nimwegen, 2013). However, this coverage value misses a lot of information the true occupancy contains: the coverage is based on a limited set of called nucleosomes that are assumed to always be present, while the occupancy contains all weaker nucleosomes and includes the probability of their presence.

An average genomic nucleosome occupancy of 75-90% has been given (Field et al., 2008; Segal and Widom, 2009; Locke et al., 2010; Liu et al., 2014; Chereji and Morozov, 2014), citing the *Chromatin* book of 1989 by van Holde (van Holde, 1989). However, the book never mentions such a range nor an experiment that measured the average genomic nucleosome occupancy at all. The 75-90% range first appears as an approximation given by Segal et al. (2006), who cite van Holde (1989) for data their approximation relies on – not a direct measurement. Their approximation assumes the complete genome is covered by nucleosomes spaced by linkers, which have a length of 10-50 bps based on *in vitro* nucleosome array formation (van Holde, 1989). To improve this approximation I extended the back-of-an-envelope calculation to include nucleosome-depleted regions, nucleosome binding frequencies, and *in vivo*-measured linker lengths (Section 12.7). I estimate an average genomic nucleosome occupancy of 68-83%. Knowing the average nucleosome occupancy is fundamental for the field and an experiment to measure the genomic nucleosome occupancy directly is long overdue. Without knowing the average

occupancy we cannot learn precise sequence preferences of nucleosomes.

14.2 Do nucleosomes really prefer G+C-rich DNA?

Reviews agree that G+C-rich DNA is preferred by nucleosomes, but to what degree is still unclear (Iyer, 2012; Struhl and Segal, 2013). *In vivo* nucleosome measurements in *S.cerevisia* disagree on their correlation to local G+C content (Figure 13.1). Many datasets show a G+C enrichment in nucleosome-covered sequences, which suggests a strong preference, but at least for MNase-Seq datasets the signal may partially come from sequence biases, because nucleosome-free MNase-Seq measurements show similar G+C enrichments (Figure 13.1) (Locke et al., 2010; Chung et al., 2010). Similarly, the high correlations between *in vitro* and *in vivo* MNase-Seq datasets suggest that the DNA sequence is important for nucleosome positioning (Kaplan et al., 2009), but could also stem from the sequence bias of MNase-Seq (Locke et al., 2010). An enrichment of G+C-rich sequences is also seen in MNase-independent, *in vitro* measurements of competition between synthetic oligonucleotides to bind histones (Kaplan et al., 2009; Levo et al., 2015). These measurements may have G+C biases of their own, e.g. from the salt-gradient dialysis (Chung et al., 2010).

While the majority of measurements suggest that G+C content influences nucleosome occupancy *in vivo*, several measurements show low or negative correlations between nucleosome occupancy and G+C content. Three of the *S.cerevisia* datasets I analyzed have low or negative correlation coefficients with G+C content (Shivaswamy: 0.02, Cole: -0.08, Fan: 0.06; Figure 13.1). CC-Seq measurements in *S.pombe* anti-correlate with G+C content (Moyle-Heyrman et al., 2013). However, these low and anti-correlations with G+C content could be effects of experimental sequence biases, in the same way the high correlations with G+C content otherwise have to be.

While the *S.cerevisia* CC-Seq dataset has a correlation coefficient of 0.35 to G+C content, the predicted occupancies of our optimized CC-Seq model have a strongly negative correlation coefficient of -0.47 to G+C content. The preference of the CC-Seq model for A+T-rich sequences is also seen in the model's parameters (Figure 13.5). This probably originates from the thermodynamic model capping the maximal possible occupancy. The method is forced to learn an energy model that is positionally precise instead of replicating the unrealistic large-scale variations that correlate with G+C content. Comparing the model's predictions with the CC-Seq dataset confirm this: the correlation coefficient is 0.37 at single-base-pair resolution and 0.04 at 147-bp resolution (i.e. comparing oc-

cupancies). That our CC-Seq model prefers A+T content, even though the data has a G+C enrichment is an indication that the nucleosomes do not strongly prefer high G+C content in *S.cerevisia*.

Nucleosomes might not prefer G+C-rich sequences, even if the nucleosome occupancy correlates with G+C content. The G+C content of nucleosomal DNA was shown to be enriched by different nucleotide-to-nucleotide mutation rates of DNA bound by nucleosomes and unbound DNA (Chen et al., 2012). Further support for this hypothesis is a low correlation of G+C content to *in vitro* MNase-Seq measurements of nucleosomes assembled on the genome of *E.coli*, which natively has no nucleosomes (Xing and He, 2015). The G+C content also acts as a proxy for other correlated, biologically-relevant sequence features like the physical properties of dinucleotides (Tillo and Hughes, 2009). Together, the preference of nucleosomes for G+C-rich sequences is unlikely to be the cause for high correlations between nucleosome measurements and G+C content. This means that even a nucleosome model whose predictions correlate strongly with measurements that are free of bias might misrepresent the binding preferences of nucleosomes and be misleading when analyzing the mechanisms of nucleosome positioning *in vivo*.

14.3 Position-specific nucleosome binding preferences

The nucleosomes rotational positioning is influenced by a 10-bp-periodic pattern of WW/SS dinucleotide enrichment (Struhl and Segal, 2013). A more jagged pattern was observed with CC-Seq (Brogaard et al., 2012). I showed that, with the exception of the peaks at the positions ±3 bps from the dyad, this jagged pattern is closer to the truth than the 10-bp-periodic pattern. The high-resolution energy model my method learned from MNase-Seq data shows a similar jagged pattern as the energy model learned from CC-Seq data (Figure 13.5), revealing that the smooth 10-bp-periodic pattern is an oversimplification obtained due to the low resolution of MNase-Seq. The jagged pattern follows the 10-bp-periodic pattern, but the deviations are not periodic themselves, therefore the nucleosome-bound sequence is important for the single-base-pair rotational positioning.

The low-resolution energy models derived by others (Kaplan et al., 2009; Lubliner and Segal, 2009; Xi et al., 2010; Locke et al., 2010) – often smoothed on purpose – miss the above-mentioned details and lead to predictions with low positional resolution. My method can learn a high-resolution energy model leading to nucleosome-position predictions with high positional resolution from low-resolution measurements. A nucleosome

binding energy model with a higher resolution improves the precision of the thermodynamic model predictions and reduces the errors of the binding energies. Therefore, I hope that my method and similar approaches will reinvigorate quantitative modeling of the competitive binding of nucleosomes and transcription factors at promoters and enhancers to predict gene expression.

14.4 How well can we predict nucleosome positioning

Measurements of different experimental protocols disagree on the nucleosome occupancy and binding preferences (Figure 13.1). A reason for this disagreement are the different experimental biases. Similar biases in measurements performed with the same experimental protocol or by the same laboratory inflate the correlations between these measurements. Biases shared by the training and test datasets are also responsible for misleading benchmarks of nucleosome-prediction methods. Strong correlations between predictions and measurements of nucleosome occupancies have been published (Kaplan et al., 2009; Tillo and Hughes, 2009), but the accuracy of such predictions are inconsistent when evaluated on other datasets as a result of the low agreements between some measurements (Figure B.2).

Low-resolution predictions can outperform high-resolution predictions when benchmarked against low-resolution measurements even though the high-resolution models are more precise. This issue is amplified when comparing smoothed predictions and measurements. Therefore, I use single-base-pair resolution – without smoothing – to benchmark the prediction methods (Figure 13.3). Often the predictions and dyad measurements are smoothed over a 147 bp running window, which corresponds to comparing occupancies. This smoothing deemphasizes high-precision modeling and instead emphasizes large-scale feature, which our analysis showed contain strong biases and represent unrealistic genomic occupancy distribution (Figures 13.1 and 13.2).

Together this means that individual high correlations of prediction methods to measurements may feign a better understanding of nucleosome positioning than we have. The methods can be good at predicting specific experimental datasets, but this data is not necessarily a good representation of the biological signal we are interested in. I believe the primary goal of nucleosome-position prediction is to understand what defines the nucleosome positioning *in vivo* and ideally create a quantitative model of it. While experimental biases are widely discussed and accepted in the field, they have been over-

looked when formulating and learning such models. These fundamental issues still need to be resolved before we can confidently claim to have a good understand of the sequence preferences of nucleosomes.

15. Conclusion

I showed that genome-wide nucleosome measurements cannot quantitatively represent nucleosome occupancies due to their genomic distributions (Figure 13.2). The experimental biases that are to blame for this have effect sizes comparable to those of experimental conditions such as knock-downs of chromosomal proteins or even removal of the whole cellular context (Figure 13.1). Due to these shortcomings of measurement-derived occupancies, I decided to focus on obtaining a positionally-high-resolution nucleosome energy model instead of optimizing the model's reproduction of such occupancies. My method can separate out the strand-dependent sequence bias of CC-Seq and can learn a position-specific energy model from low-resolution MNase-Seq measurements. The two energy models my method learned from CC-Seq and MNase-Seq data agree better than models derived directly from dinucleotide frequencies. The optimized models both contain the jagged dinucleotide pattern already visible in the CC-Seq frequency-derived model, with exception of the sequence bias. This means the true nucleosome sequence is probably more jagged than the smooth 10-bp-periodic pattern that has been derived from MNase-Seq measurements (Kaplan et al., 2009).

My two optimized models still disagree in two vital aspects: the amplitude of the pattern and the influence of the average G+C content. To settle this disagreement the method would need to model the experimental bias more precisely. The sequence preference (Fan et al., 2010) and continuous nature (Weiner et al., 2010) of the MNase digestion would need to be modeled for MNase-Seq data, and whole fragments, instead of independently processed fragment ends, would need to be modeled for CC-Seq. A single nucleosome binding energy model could also be learned from distinct experiments simultaneously, optimizing the experimental bias and positional error models for each experiment in parallel.

I believe that the next major improvements in modeling nucleosome positioning will come from explicitly including experimental biases to distinguish them from the nucleosome binding energy model. My method is a first step in that direction. To fully

understand what influences nucleosome binding *in vivo* we will have to improve the experimental bias models and integrate published extensions of the thermodynamic model into a method like mine. Alternatively, more quantitative nucleosome occupancy measurements would solve a lot of the issues I described here. Determining the average genomic nucleosome occupancy would already set vital boundaries for nucleosome-position prediction methods.

Part IV

Ancillary analyses and methods

16. Analyses left out of Part III

Part III contains a refined story of my main PhD project. Here I discuss aspects of this project that I left out of Part III to streamline the story for publication. In Section 16.1 I describe alternative benchmarks and why I ended up leaving them out of the story. Section 16.2 discusses issues with learning high-resolution model from most MNase-Seq datasets. I learned up to 4th-order energy models with my method and present the results in Section 16.3. Finally, Section 16.4 describes issues with learning the sequence-unspecific binding energy.

16.1 Alternative benchmarks

I skipped more common quality-based benchmark of prediction methods, because they are not well suited for nucleosome predictions. Because they are frequently used by others, I discuss these issues here. Other benchmarks I tested, but the results were of limited interest. I describe the most interesting case: validation with *in vivo* competition assays.

16.1.1 Quality-based benchmarks

The most common scores to evaluate prediction methods (e.g. accuracy, precision, and recall) all rely on qualitative predictions and only include quantitative information by shifting a threshold (ROC and AUC). For many prediction methods this is fine because there is an obvious score to apply the threshold on and the qualitative prediction is of interest. These scores have been used to benchmark nucleosome-position prediction methods (Peckham et al., 2007; Reynolds et al., 2010; Moser and Gupta, 2012; Ozonov and van Nimwegen, 2013; Guo et al., 2014). There are flaws in this approach and I believe it is a partial reason that the scientific field is overstating how well nucleosomes can be predicted.

The first issue is the definition of the nucleosome map that is used as ground truth. For MNase-Seq data, nucleosome positions are often selected from the highest to lowest signal

disallowing overlap until the whole genome is covered or the signal is below a threshold. While this does enforce a uniform coverage of the genome, the definition of overlap is arbitrary and both covering the whole genome or using a threshold to stop calling nucleosomes have issues. Covering the whole genome will probably place nucleosomes in several nucleosome-depleted regions (NDRs) removing an important distinction the prediction methods should make. It generally makes sense to use a threshold to define the ground truth, the issue here is that there is no clear border between signal and noise to orient the threshold by. As shown in Figure 13.2 the genomic occupancy values are a single continuous distribution and the same is true for smaller smoothing windows.

The second issue arises from defining when a prediction is correct. MNase-Seq and most other experimental protocols only have a resolution of several base pairs. Adding the fuzziness with which nucleosomes are frequently positioned, this leads to an expected positional uncertainty of 10-20 bps. At the same time, the whole genome is covered by nucleosomes. Depending on the overlap definition used nucleosomes will occur every 100-200 bps. As a result the benchmark is about predicting the rough position of nucleosomes, but ignores the exact rotational positioning.

The third issue arises from defining when a prediction is false. Instead of using all genome positions some benchmarks define a set of false regions. One such benchmark defined gaps between called nucleosomes as linkers or nucleosome-free regions. These regions do not have to have a low occupancy as measured by the same experiment, actually the occupancy distribution of these regions and the distribution of called nucleosomes overlapped strongly. Similar issues are hard to circumvent when using a map of non-overlapping nucleosomes.

The fourth issue is that the benchmark validates an aspect that is of little interest. The second issue plays into this. Being able to predict the rough position of nucleosomes without any quantitative information about their frequency is rather useless. The predictions are unhelpful in understanding and modeling any thermodynamic processes.

I came about the best summary of the issues with qualitative benchmarks for nucleosome-position prediction methods by chance. I benchmarked experimental measurements (i.e. data not predictions) against a canonical nucleosome map based on several datasets (Jiang and Pugh, 2009), including ones I was testing. I used one of the benchmarks described in Ozonov and van Nimwegen (2013). The experimental measurements performed poorly: they beat random guesses, but were unable to distinguish between a substantial fraction of nucleosome and linker regions. This bad performance would be acceptable if the benchmark removed biases in the test data, but there is no reason why

the experimental biases should be reduced.

16.1.2 Validation with *in vitro* competition assays

BunDLE-Seq (Section 3.3.11) measures the ratio of bound to unbound DNA fragments: $\frac{[TF*S]}{[S]}$. Where S denotes the DNA fragment or sequence, and TF denotes transcription factors (or other DNA-binding factors). In my case, the nucleosomes are the DNA-binding factors described with TF. From thermodynamics it follows that:

$$\begin{aligned} \frac{[TF][S]}{[TF*S]} &= c_0 e^{E_S} \\ \frac{[S]}{[TF*S]} &\propto e^{E_S} \\ \frac{[TF*S]}{[S]} &\propto e^{-E_S} \end{aligned} \tag{16.1}$$

Assuming the factors bind independently, i.e. the weak-binding approximation:

$$\begin{aligned} \frac{[TF*S]}{[S]} &\propto e^{-\sum_i E_{S_i}} \\ \ln\left(\frac{[TF*S]}{[S]}\right) &\propto -\sum_i E_{S_i} \end{aligned} \tag{16.2}$$

Accounting for the factors getting into each others way:

$$\begin{aligned} \frac{[TF*S]}{[S]} &\propto \sum_i e^{-E_{S_i}} \\ \log\left(\frac{[TF*S]}{[S]}\right) &\propto \log\left(\sum_i e^{-E_{S_i}}\right) \end{aligned} \tag{16.3}$$

Where E_S is the binding energy to the sequence S and E_{S_i} the binding energy to position i of sequence S. Ignoring $[TF]$ and c_0 in Equation 16.1 removes the need to know the concentration of unbound factors ($[TF]$) and sequence-unspecific binding preference of the factor (c_0). Both of these values are difficult to measure and BunDLE-Seq does not measure them. The proportionality or unknown scale (offset after taking the logarithm) can limit the further analysis. Such a limitation is that validating the log predictions has to be offset independent, e.g. with Pearson's correlation coefficient.

Comparing the performance of my MNase-Seq model against Kaplan et al.'s model showed little difference on the data from Levo et al. (2015). The predictions of both methods had Pearson's correlations coefficients of nearly 80%. The G+C content already

captures a large part of the signal with a correlation of nearly 70%. Given the CC-Seq model lacks the strong correlation to G+C content it performs far worse with correlations below 10%. Because this measurement largely reflects the correlation to G+C content and my MNase-Seq model performed similar to Kaplan et al.'s, I decided to leave these results out of the main story in Part III.

The high correlation of the BunDLE-Seq measurements to G+C content suggest that nucleosomes prefer G+C content. This contradicts the results of other analyses (Section 14.2). The high correlation could also stem from experimental biases of the salt-gradient dialysis used to form the nucleosomes (Chung et al., 2010). Nucleosomes could also favor G+C-rich sequences at lower concentrations and in the *in vitro* environment, but *in vivo* the cellular context prohibits them from manifesting this preference.

16.2 Optimizing a high-resolution model from low-resolution data is difficult

In Section 13.3 I describe a high-resolution nucleosome energy model that I learned on low-resolution MNase-Seq measurements. I could consistently learn such a high-resolution model on the *in vitro* dataset of Kaplan et al. (2009). This worked with all three deconvolution point-spread functions, as long as the initialization was not to sharp.

I optimized an energy model for most other MNase-Seq datasets I analyzed, but these energy models always had a low resolution. Especially for the *in vivo* measurements of Kaplan et al., I tried everything I could think of and nothing helped optimize a high-resolution model from the data. To identify possible reasons I optimized the energy model on individual replicates of Kaplan et al. instead of the merged datasets. My method only learns a high-resolution model from one of the two *in vitro* replicates. The other optimization behaved like the optimizations on the *in vivo* datasets.

The main difference between the *in vitro* replicate that leads to a high-resolution model and the other replicate, as well as the *in vivo* datasets, is the average distance distribution between the strand data, i.e. the fragment starts and ends. I suspect that this difference stems from experimental discrepancies in the digestion level or the band extraction from the gel electrophoresis. In either case, the distribution of fragment lengths influences the positional uncertainty of the measurements, which in turn affects the optimization.

Based on these findings I optimized models with the position-specific deconvolution on the individual replicates. Most of the optimized point-spread function had additional peaks at roughly ± 9 bps from the dyad. This suggests many fragments miss two helical

turn on one end, which produces an offset dyad estimate. In my analysis of partial unwrapping I saw an enrichment of such ~128-bp fragments (Section 8.6). Even with this more specific deconvolution the optimized models were low resolutions.

The only way I managed to achieve higher-resolution models from these datasets was to initiate the model from a high-resolution model instead of the dinucleotide frequencies. Even then, the models seemed to drift slightly towards a lower resolution, but they soon reached a performance plateau and the optimization was interrupted.

16.3 Higher-order Markov models

As described in Section 12.1.2, I implemented my nucleosome-position prediction method to handle 1st-order energy models and higher. The optimization runtime increases with higher orders, while the prediction time is hardly affected. With increasing order the signal-to-noise ratio decreases, because there are less measurements per alternative possibility. This is one reason most others have used positioned 1st-order or unpositioned 4th-order models.

I trained up to 4th-order energy models. The Markov chain describing my energy model is position-specific (positioned). Both frequency-derived and optimized 2nd- and 3rd-order models each capture more information than the equivalent model of the previous order. Their predictions for a chromosome left out of the training set have higher correlation coefficients to the data. The frequency-derived 4th-order model had a lower correlation than the 3rd-order model, but after the optimization the models performed similar. For the 4th-order model the counts are apparently so low that the noise outweighs the benefits of the more detailed model when deriving the model from the frequencies.

Analogous to an approach a colleague used in BAMM!motif to learn motifs of DNA-binding factors, I tested interpolating between the orders of the model to cope with the noise issue (Siebert and Söding, 2016). BAMM!motif uses an interpolated inhomogeneous Markov model, which uses the model's probabilities of one order below as pseudo counts in each order. My implementation of this concept uses a penalty score that pulls the probabilities towards the lower-order probability. The end result is similar, with the main difference being that my method does not explicitly learn the lower-order models in parallel, but derives them from the higher-order model. This ended up working, in so far as that the 4th-order model improved slightly above the 3rd-order model. However, there were no strong improvements for the 1st-, 2nd-, and 3rd-order models.

In the end, I sticked with a 1st-order energy model without the need for such an

interpolation, even though the higher-order models correlated better with the left-out chromosome. The over all improvements of the higher-order models were unclear, and they were more difficult to work with. The improvements might be limited to validations with similar measurement, because of the issues with experimental biases. Higher-order models are also hard to visualize and comprehend. To analyze such models in the future I propose adapting and refining BAMM!motif's logo visualization of its motif models to work better for larger models.

16.4 Sequence-unspecific binding energy of nucleosomes

As described in Section 12.1.3, μ describes the sequence-unspecific binding energy in my thermodynamic model. The parameter indirectly sets the average genomic occupancy, which I used to generate the realistic and low occupancy predictions in Figure 13.2.

The partial derivative of the log-likelihood can be computed for μ as described throughout Section 12. However, the issues with nucleosome occupancies and frequencies derived from the experimental measurements (Section 13.2) create problems when optimizing μ. The optimization reduces the μ parameter until the average genomic occupancy is far below any realistic value. This reflects that the method is trying to reproduce the data, which represent occupancies with an unrealistically-low genomic average.

Making it possible to learn the sequence-unspecific binding energy μ was one of the main reasons I tried to correct the data to be more quantitative (Section 17). After all the failed attempts I had to accept that the method would either learn an unrealistic binding energy μ or I had to set it without optimization. I refuse to produce completely unrealistic predictions, even if their benchmarking results are superior. Such models and predictions provide no understanding of the underlying processes and can even lead to flawed analysis. Therefore, I deactivated the optimization of μ and set it to a value that produced an average genomic occupancy slightly below my estimated range (Section 12.7). This is a deliberate trade-off between realistic predictions and a sensible optimization by approximating the datasets properties.

17. Correcting nucleosome measurements

As discussed in Section 13.1, MNase-Seq has severe biases. This breaks the assumption that occupancies can be derived from such nucleosome-positioning data. Many nucleosome prediction methods and their validations rely on this assumption. The main issue is that some regions have an over 2-fold higher coverage than the genome-wide average, which leads to an unrealistically-low average occupancy (Section 13.2).

I attempted to correct the deviation between the expected distribution and that of the MNase-Seq measurements to make the data more quantitative. I also tried to correct the CC-Seq measurements, which have the same issue, with most of the approaches. Less is known about the biases of CC-Seq than of MNase-Seq, but my analysis in Section 13.4 shows that CC-Seq has at least one sequence bias as well.

17.1 Sequence-dependent correction factor

I will refer to frequently refer to the same terms in the next sections:

$$\begin{aligned} z_i &:= \text{Measured nucleosome-dyad frequency at position } i \\ C_i &:= \text{Correction factor for sequence bias at position } i \\ y_i &:= \text{Corrected nucleosome occupancy at position } i \\ &= \frac{1}{147} \sum_{j=i-73}^{i+73} \frac{z_j}{C_j} \end{aligned} \tag{17.1}$$

The basic concept is to divide the measurements by the correction factor to make the data more quantitative. I attempted different approaches to optimize the model of the correction factor. Because a large part of the MNase-Seq bias appears to be sequence

driven, my first attempts use a correction factor that depends on the local sequence:

$$\begin{aligned} C_i &:= e^{\sum_{k=-73}^{+73} \sum_{a,b=1}^{4} \beta_k(ab)\, \mathrm{I}(s_{i+k}=ab)} \\ \beta_k(ab) &:= \text{Correction parameters, dinucleotide } (ab) \text{ and position } (k) \text{ specific} \end{aligned} \tag{17.2}$$

A penalty score between neighboring βs can be added to these sequence-dependent correction methods to reduce the noise in the correction parameters. I performed the corrections on single-end sequencing data and made a stranded model, for which the MNase cut site is precisely defined. I experimented with reducing the penalty at the MNase cut site, because I expected a more position-specific bias there. I either used a Laplace or a Gaussian distribution to penalize the deviation between the neighboring βs. Given a set of βs the optimal scale parameters (ρ or σ) can be calculated analytically for both penalty scores. Alternatively, the scale parameters can be set to allow more variation or force higher similarity between neighboring βs.

17.1.1 Laplace penalty score

$$\mathcal{L}(y) = [...] * \prod_{a,b=1}^{4} \prod_{k=-72}^{+73} \frac{1}{2\rho} e^{\frac{-|\beta_k(ab)-\beta_{k-1}(ac)|}{\rho}}$$

$$\begin{aligned} \log \mathcal{L}(y) &= [...] + \sum_{a,b=1}^{4} \sum_{k=-72}^{+73} \log(\frac{1}{2\rho}) + \frac{-|\beta_k(ab) - \beta_{k-1}(ab)|}{\rho} \\ &= [...] - 4^2 146 \log(2\rho) + \frac{1}{\rho} \sum_{a,b=1}^{4} \sum_{k=-72}^{+73} -|\beta_k(ab) - \beta_{k-1}(ab)| \end{aligned} \tag{17.3}$$

$$\frac{\partial \log \mathcal{L}}{\partial \beta_k(ab)} = [...] + \frac{1}{\rho} \left(-\mathrm{sign}(\beta_k(ab) - \beta_{k-1}(ab)) + \mathrm{sign}(\beta_{k+1}(ab) - \beta_k(ab))\right)$$

$$\begin{aligned} \frac{\partial \log \mathcal{L}}{\partial \rho} &= -4^2 146 \frac{2}{2\rho} - \frac{1}{\rho^2} \sum_{a,b=1}^{4} \sum_{k=-72}^{+73} -|\beta_k(ab) - \beta_{k-1}(ab)| \\ 0 &= -\frac{4^2 146}{\rho} - \frac{1}{\rho^2} \sum_{a,b=1}^{4} \sum_{k=-72}^{+73} -|\beta_k(ab) - \beta_{k-1}(ab)| \\ \rho &= \frac{\sum_{a,b=1}^{4} \sum_{k=-72}^{+73} |\beta_k(ab) - \beta_{k-1}(ab)|}{4^2 146} \end{aligned} \tag{17.4}$$

The Laplace penalty score is linear and its partial derivatives therefore constant. If a parameter value lies between the values of its two neighbors, the two penalty terms

negate each other. Therefore, this penalty scores only penalizes parameters that are higher or lower than both of its neighbors.

17.1.2 Gaussian penalty score

$$\mathcal{L}(y) = [...] * \prod_{a,b=1}^{4} \prod_{k=-72}^{+73} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(\beta_k(ab)-\beta_{k-1}(ab))^2}{2\sigma^2}}$$

$$\begin{aligned} \log \mathcal{L}(y) &= [...] + \sum_{a,b=1}^{4} \sum_{k=-72}^{+73} \log(\frac{1}{\sqrt{2\pi}\sigma}) + \frac{-(\beta_k(ab) - \beta_{k-1}(ab)^2}{2\sigma^2} \\ &= [...] - 4^2 146 \log(\sqrt{2\pi}\sigma) + \frac{1}{2\sigma^2} \sum_{a,b=1}^{4} \sum_{k=-72}^{+73} -(\beta_k(ab) - \beta_{k-1}(ab))^2 \end{aligned} \tag{17.5}$$

$$\begin{aligned} \frac{\partial \log \mathcal{L}}{\partial \beta_k(ab)} &= [...] + \frac{1}{2\sigma^2} \left(-2(\beta_k(ab) - \beta_{k-1}(ab)) + 2(\beta_{k+1}(ab) - \beta_k(ab))\right) \\ &= [...] + \frac{1}{\sigma^2} \left(\beta_{k-1}(ab) - 2\beta_k(ab) + \beta_{k+1}(ab)\right) \end{aligned}$$

$$\begin{aligned} \frac{\partial \log \mathcal{L}}{\partial \sigma} &= -4^2 146 \frac{\sqrt{2\pi}}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^3} \sum_{a,b=1}^{4} \sum_{k=-72}^{+73} -(\beta_k(ab) - \beta_{k-1}(ab))^2 \\ 0 &= -\frac{4^2 146}{\sigma} - \frac{1}{\sigma^3} \sum_{a,b=1}^{4} \sum_{k=-72}^{+73} -(\beta_k(ab) - \beta_{k-1}(ab))^2 \\ \sigma &= \sqrt{\frac{\sum_{a,b=1}^{4} \sum_{k=-72}^{+73} (\beta_k(ab) - \beta_{k-1}(ab))^2}{4^2 146}} \end{aligned} \tag{17.6}$$

The quadratic nature of the Gaussian penalty score leads to smooth β profiles. This penalty score pulls the value of every correction parameter towards the center of its two neighbors. The partial derivatives is linear in the distances to the two neighboring parameters and therefore zero if they are equidistant.

17.2 Minimizing the variation of binned occupancies

Nucleosomes are expected to cover most of the genome, and nucleosome-depleted regions – like promoters – are randomly distributed across the genome. My first correction method is based on the assumption that the average nucleosome occupancy of larger

regions (>1 kbp) are similar. I optimized a sequence-dependent correction factor that reduces the variation between the average occupancies of genomic regions.

$$
\begin{aligned}
b &= \text{bin size } (:= 2000) \\
E_1 &= \sum_{i=0}^{L/b} \left| \sum_{j=0}^{L} \frac{y_i}{L/b} - \sum_{j=ib}^{ib+b} y_j \right| \\
E_2 &= \sum_{i=0}^{L/b} \left| \sum_{j=0}^{L} \frac{z_j}{L/b} - \sum_{j=ib}^{ib+b} y_j \right|
\end{aligned}
\tag{17.7}
$$

My first attempt was to minimize the error function E_1, the L_1 norm of the corrected data. I quickly realized that increasing all βs decreases the error function, which is a trivial and useless solution. In the error function E_2, I substituted the average reference point with the uncorrected data average.

In principle, this approach works with the error function E_2. I could roughly half the variance of the binned nucleosome occupancies. However, major issues still remained: the genomic profiles looked problematic, the similarity between the strand specific data hardly improved, and my nucleosome-position prediction method could not learn a realistic sequence-unspecific nucleosome binding energy from the corrected data. After several failed attempts I moved on to our next idea.

17.3 Minimizing the high-occupancy tail

Most of the genome is covered by nucleosomes, which creates a hard upper limit of 100% occupancy (Section 13.2). No peaks should stand out compared to a realistic genome-wide average occupancy of 68-83% (Section 12.7). Without better genome-wide nucleosome measurements the true occupancy distribution is unknown. We assume that we know nothing about the occupancy distribution under a set threshold, while the occupancies above should be distributed like a Gaussian (half a Gaussian to be precise). With enough observations the Gaussian distribution is a good approximation of the Poission distribution, which is the expected distribution based on the fragment sampling that happens in the experiment. A uniform distribution models the lack of knowledge below the threshold. I developed a probabilistic model for this combined distribution and optimize the βs

to maximize the models likelihood.

$$\begin{aligned} t &:= \text{threshold} = \text{mean} y_i \\ c &:= \text{fraction of data below } t = 0.5 \\ \sigma &= \frac{t}{c\sqrt{2\pi}} \end{aligned} \tag{17.8}$$

The definition of σ follows from the hight of the Gaussian and uniform distributions matching, i.e. $\frac{1}{\sigma\sqrt{2\pi}} = \frac{c}{t}$.

$$\begin{aligned} \mathcal{L}(y) &= \prod_{i=1}^{L} \left[\mathrm{I}(0 < y_i < t)\frac{c}{t} + \mathrm{I}(t < y_i)\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-t)^2}{2\sigma^2}} \right] \\ \log \mathcal{L}(y) &= \sum_{i=1}^{L} \left[\mathrm{I}(0 < y_i < t)(\log(c) - \log(t)) + \mathrm{I}(t < y_i)(-\frac{(y_i-t)^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi})) \right] \\ &= L(\log(c) - \log(t)) + \sum_{i=1}^{L} \left[\mathrm{I}(t < y_i)(-\frac{(y_i-t)^2}{2\sigma^2}) \right] \end{aligned} \tag{17.9}$$

The logarithm can be pulled into the sum, because only one of the identities is true at a time. While the formula contains the addition, for all i only a single term exists. Therefore, the logarithm can be applied to the two terms individually.

Learning from the previous failure it was clear that a constant threshold would simply lead to large correction factors, which shift the whole dataset below the threshold. Therefore, I set the threshold to the average corrected occupancy, which turned out to have the same basic problem.

$$\begin{aligned} t &:= \frac{\sum_{j=1}^{L} y_j}{L} \\ \log \mathcal{L}(y) &= L(\log(c) - \log(\frac{\sum y_j}{L})) + \sum_{i=1}^{L} \left[\mathrm{I}(\frac{\sum y_j}{L} < y_i)(-\frac{(y_i - \frac{\sum y_j}{L})^2}{\frac{(\frac{\sum y_j}{L})^2}{c^2\pi}}) \right] \\ &= L\log(c) - L\log(\frac{\sum y_j}{L}) - \sum_{i=1}^{L} \left[\mathrm{I}(\frac{\sum y_j}{L} < y_i)\frac{(y_i - \frac{\sum y_j}{L})^2}{\frac{(\sum y_j)^2}{c^2\pi L^2}} \right] \end{aligned} \tag{17.10}$$

$$
\begin{aligned}
\frac{\partial \log \mathcal{L}}{\partial \beta_k(ab)} = & -\frac{L}{\frac{\sum y_j}{L}} \frac{1}{L} \sum \frac{\partial y_j}{\partial \beta} - c^2 \pi L^2 \sum_{i=1}^{L} \Bigg[\mathrm{I}(\frac{\sum y_j}{L} < y_i) \\
& \left(2(y_i - \frac{\sum y_j}{L})(\frac{\partial y_i}{\partial \beta} - \frac{1}{L} \sum \frac{\partial y_j}{\partial \beta}) \frac{1}{(\sum y_j)^2} + (y_i - \frac{\sum y_j}{L})^2 \frac{-2 \sum \frac{\partial y_j}{\partial \beta}}{(\sum y_j)^3} \right) \Bigg] \\
= & -\frac{\sum \frac{\partial y_j}{\partial \beta}}{\frac{\sum y_j}{L}} - 2c^2 \pi L^2 \sum_{i=1}^{L} \Bigg[\mathrm{I}(\frac{\sum y_j}{L} < y_i) \\
& \frac{(y_i - \frac{\sum y_j}{L})}{(\sum y_j)^2} \left((\frac{\partial y_i}{\partial \beta} - \frac{1}{L} \sum \frac{\partial y_j}{\partial \beta}) - (y_i - \frac{\sum y_j}{L}) \frac{\sum \frac{\partial y_j}{\partial \beta}}{\sum y_j} \right) \Bigg]
\end{aligned}
$$

$$
\frac{\partial y_i}{\partial \beta_k(ab)} = -\sum_{j=i-73}^{i+73} \frac{z_j \, \mathrm{I}(s_{j+k} = ab)}{147 C_j} \tag{17.11}
$$

The probability of each data point is higher for narrower distributions. Downscaling all occupancy values increases the likelihood by decreasing the threshold and narrowing the distribution. Once again the first attempt produced a trivial and useless solution.

To avoid this issue I redefined the corrected nucleosome occupancy, which made the derivations more complex. I also noticed that the definition of σ relies on $c = 0.5$, because the area under the half-Gaussian and uniform distributions otherwise does not equal one. Because I also wanted to test other values for c, I normalized the two distributions together (with h), which changed the dependency of σ on c and t.

$$
\begin{aligned}
y_i &:= \sqrt[L]{\prod_{l=1}^{L} C_l} \frac{1}{147} \sum_{j=i-73}^{i+73} \frac{z_j}{C_j} \\
\sigma &:= \alpha t \\
c &= \frac{ht}{ht(1 + \alpha\sqrt{\frac{\pi}{2}})} = \frac{1}{1 + \alpha\sqrt{\frac{\pi}{2}}} \rightarrow \alpha = (\frac{1}{c} - 1)\sqrt{\frac{2}{\pi}} \\
\rightarrow \sigma &= (\frac{1}{c} - 1)\sqrt{\frac{2}{\pi}} t \\
1 &= h(t + \sigma\sqrt{\frac{\pi}{2}}) \rightarrow 1 = ht(1 + \alpha\sqrt{\frac{\pi}{2}}) \rightarrow h = \frac{1}{t(1 + (\frac{1}{c} - 1))} \\
\rightarrow h &= \frac{1}{t\frac{1}{c}} = \frac{c}{t}
\end{aligned} \tag{17.12}
$$

$$
\begin{aligned}
\mathcal{L}(y) &= \prod_{i=1}^{L} h\left[\mathrm{I}(0 < y_i \leq t) + \mathrm{I}(t < y_i) e^{-\frac{(y_i - t)^2}{2\sigma^2}}\right] \\
\log \mathcal{L}(y) &= \sum_{i=1}^{L} log(h) + \left[0 + \mathrm{I}(t < y_i)(-\frac{(y_i - t)^2}{2\sigma^2}))\right] \\
&= L \log(\frac{c}{t}) + \sum_{i=1}^{L}\left[\mathrm{I}(t < y_i)(-\frac{(y_i - t)^2}{2((\frac{1}{c} - 1)\sqrt{\frac{2}{\pi}}t)^2})\right] \\
&= L \log(\frac{c}{t}) - \frac{1}{2((\frac{1}{c} - 1)\sqrt{\frac{2}{\pi}}t)^2} \sum_{i=1}^{L}\left[\mathrm{I}(t < y_i)(y_i - t)^2\right]
\end{aligned}
\tag{17.13}
$$

$$
\begin{aligned}
\frac{\partial \log \mathcal{L}}{\partial \beta_k(ab)} &= -\frac{1}{2(\frac{1}{c} - 1)^2 \frac{2}{\pi} t^2} \sum_{i=1}^{L}\left[\mathrm{I}(t < y_i)\frac{\partial (y_i - t)^2}{\partial \beta}\right] \\
&= -\frac{\pi}{4}(\frac{1}{c} - 1)^{-2} t^{-2} \sum_{i=1}^{L}\left[\mathrm{I}(t < y_i) 2(y_i - t)\frac{\partial y_i}{\partial \beta}\right]
\end{aligned}
\tag{17.14}
$$

$$
\begin{aligned}
\frac{\partial y_i}{\partial \beta_k(ab)} &= \frac{1}{147}\left(\frac{\partial \sqrt[L]{\prod_{l=1}^{L} C_l}}{\partial \beta_k(ab)} \sum_{j=i-73}^{i+73} \frac{z_j}{C_j} - \sqrt[L]{\prod_{l=1}^{L} C_l} \sum_{j=i-73}^{i+73} \frac{z_j \,\mathrm{I}(s_{j+k} = ab)}{C_j}\right) \\
&= \frac{1}{147}\left(e^{\frac{\sum_{l=1}^{L} \log C_l}{L}}\left(\sum_{l=1}^{L} \frac{\mathrm{I}(s_{l+k} = ab)}{L}\right) \sum_{j=i-73}^{i+73} \frac{z_j}{C_j}\right. \\
&\qquad \left. - e^{\frac{\sum_{l=1}^{L} log C_l}{L}} \sum_{j=i-73}^{i+73} \frac{z_j \,\mathrm{I}(s_{j+k} = ab)}{C_j}\right) \\
&= \frac{1}{147} e^{\frac{\sum_{l=1}^{L} \log C_l}{L}}\left(\left(\sum_{l=1}^{L} \frac{\mathrm{I}(s_{l+k} = ab)}{L}\right) \sum_{j=i-73}^{i+73} \frac{z_j}{C_j} - \sum_{j=i-73}^{i+73} \frac{z_j \,\mathrm{I}(s_{j+k} = ab)}{C_j}\right)
\end{aligned}
\tag{17.15}
$$

While less obvious, using the geometrical mean to normalize the corrected occupancies has a similar issue. Even with these constraint the optimization of the βs can decrease the average corrected occupancy. My final solution was to force a constant average corrected

occupancy of one, by dividing the corrected occupancies by their average.

$$y_i := \frac{L}{\sum_{l=1}^{L} \frac{z_l}{C_l} 147} \sum_{j=i-73}^{i+73} \frac{z_j}{C_j}$$

$$\begin{aligned}
\frac{\partial y_i}{\partial \beta_k(ab)} &= \frac{L}{147} \left(-\frac{\partial \sum_{l=1}^{L} \frac{z_l}{C_l}}{\partial \beta_k(ab)} \sum_{j=i-73}^{i+73} \frac{z_j}{C_j} + \sum_{l=1}^{L} \frac{z_l}{C_l} \frac{\partial \sum_{j=i-73}^{i+73} \frac{z_j}{C_j}}{\partial \beta_k(ab)} \right) / \left(\sum_{l=1}^{L} \frac{z_j}{C_j} \right)^2 \\
&= \frac{L}{147 \left(\sum_{l=1}^{L} \frac{z_j}{C_j} \right)^2} \left(\sum_{l=1}^{L} \frac{z_l \, \mathrm{I}(s_{l+k} = ab)}{C_l} \sum_{j=i-73}^{i+73} \frac{z_j}{C_j} - \sum_{l=1}^{L} \frac{z_l}{C_l} \sum_{j=i-73}^{i+73} \frac{z_j \, \mathrm{I}(s_{j+k} = ab)}{C_j} \right) \\
&= \frac{L}{147 \left(\sum_{l=1}^{L} \frac{z_j}{C_j} \right)^2} \sum_{l=1}^{L} \sum_{j=i-73}^{i+73} \left(\frac{z_l \, \mathrm{I}(s_{l+k} = ab)}{C_l} \frac{z_j}{C_j} - \frac{z_l}{C_l} \frac{z_j \, \mathrm{I}(s_{j+k} = ab)}{C_j} \right) \\
&= \frac{L}{147 \left(\sum_{l=1}^{L} \frac{z_j}{C_j} \right)^2} \sum_{l=1}^{L} \sum_{j=i-73}^{i+73} \frac{z_l z_j}{C_l C_j} \left(\mathrm{I}(s_{l+k} = ab) - \mathrm{I}(s_{j+k} = ab) \right)
\end{aligned} \tag{17.16}$$

This version works as intended and dampens most high-occupancy outliers. However, the genomic profiles were barely improved otherwise. The correlations between occupancies derived from either strand and corrected separately only show minor (<0.02) improvements compared to the uncorrected occupancies and are still far below (r<0.60) the expected correlations (r>0.85) based on the sampling variance of replicates.

17.4 Minimizing the difference between data of either strand

While I validated the previous correction method by comparing the data from the two strands, I realized that one could exploit this information to correct the measurements. After contemplating the possibility, I decided that such an approach was unlikely to work significantly better than the previous ones and is not universally applicable. The correction is limited to single-strand data and strand-specific biases. Strand-specific biases are only part of the problem, because paired-end data have the same issues. Therefore, the correction would still leave other biases that cause issues when deriving occupancies from the measurements.

17.5 Separating the G+C signal from the nucleosome binding energy

An obvious difference between the nucleosome measurements is the correlation between the derived occupancies and G+C content (Section 13.1). After the previous attempts of correcting out the sequence biases had failed, we still wanted to validate my nucleosome-position prediction method by comparing the MNase-Seq and CC-Seq models. I tested if separating the G+C signal from the data would improve the comparison. The idea is to remove the G+C signal, optimize my model on the cleaned data, and then apply the G+C signal back onto my model's predictions. Our hope was that without the G+C signal the datasets are more quantitative. The validation between datasets should also improve further, by adding the G+C signal of another measurement back onto the predictions.

The G+C content vs occupancy scatter plots show that a linear model describes the relationship reasonably – hence the good correlation, but the linearity loosens at the extremes. I tried a 2nd- and 3rd-order polynomial function to fit the raw or the log-normalized occupancies. I divided the nucleosome-dyad position data (i.e. not smeared like occupancies) by the fitted models to remove the G+C signal and produce cleaned datasets. After optimizing my nucleosome-position prediction model on these datasets and predicting the nucleosome positioning, I applied the G+C signal back onto the data by multiplying the predictions with the G+C model.

The signal separation has little effect on the validation against the same dataset, and for the MNase-Seq model the validation against CC-Seq data is also similar. The correlation of the CC-Seq model to the MNase-Seq data improved with the G+C signal separation. However, only the correlation between occupancies improves and there the raw G+C content beats the CC-Seq model – both with and without the G+C signal separation. The minor improvements, together with the complexity of removing and adding the G+C signal, led me to discard this approach. The lack of improvement suggests that the base assumption of the approach – the G+C signal is primarily experimental bias – is wrong or at least a gross simplification of the truth.

17.6 Applying a Fast Fourier Transformation (FFT) band filter

I abandon further sequence based corrections even though I am sure sequence-based experimental biases are part of the underlying issue, because the above approaches failed and the MNase digestion has an inherent complexity (Section 18). I tested two sequence-independent corrections, the first being a Fast Fourier Transformation (FFT) band filter. An illustrative application of FFT is the analysis of sound. FFT separates the different frequencies (pitch) revealing the individual amplitudes (loudness).

A band filter removes all signal components that have a frequency larger than a threshold or smaller than another threshold. In other words, a band filter retains a band of frequencies. Band filters can easily be applied to data using FFT. Our idea was to filter out low frequencies that cover many nucleosomes and are probably biases. At the same time, filtering out high frequencies with a periodicity of only a few base pairs might reduce the data's noise. As with the other correction methods, the model optimization is based on the occupancies, but than applied to dyad-position data.

My initial tests on real datasets lacked a significant improvement. Tests on artificial data then revealed that the sharp borders of positioned nucleosomes could create artifacts. To maintain the valleys of linkers and nucleosome depleted regions between positioned nucleosomes the neighboring signal is boosted by the FFT band filter. These setbacks led me to abandon this approach before extensive testing and switch to my final attempt to correct nucleosome measurements.

17.7 Processing the data with a thermodynamic model

My last attempt to correct the nucleosome occupancies uses the thermodynamic model (Forward/Backward algorithm), which is part of my nucleosome-position prediction method. The experimental dyad frequencies are used as nucleosome binding energies in the thermodynamic model. This enforces a realistic nucleosome occupancy distribution. An issue of this approach is that the measurements contain more than just the sequence preference of nucleosomes. For example, they also contain steric hindrance, which the thermodynamic model then applies again.

Transforming the measurements into nucleosome binding energies has two free pa-

rameters – scale (β) and offset (α). Two restrictions will define two free parameters. As restrictions, I used an average occupancy of 75% based on my estimate (Section 12.7, but any value could be used), and a maximization of the correlation between the corrected and uncorrected data. At first glance it might seem easy to fixate the average occupancy with the offset α, but due to running the produced binding energies through the thermodynamic model, the offset α has a non-linear affect. I had to optimize the scale β iteratively, while re-adjusting the offset α without shifting the scaling center of β. This process is best explained in pseudo code:

$$\begin{aligned}
z_i &:= \log(\text{ measured nucleosome dyads at position } i + \text{pseudo-counts }) \\
E(i) &:= \text{Inferred binding energy at position } i \\
&= (G(i) - \mu)/k_B T \\
&= \beta(z_i - \alpha) \\
p(i) &:= \text{Corrected dyad probability (i.e. Forward/Backward probabilities)} \\
y(i) &:= \text{Corrected nucleosome occupancy at position } i \\
&= \sum_{j=i-73}^{i+73} p(j) \\
y_{aim} &:= \text{Expected average occupancy, I chose 0.75} \\
p'(i), \beta' &:= \text{Values from the previous iteration}
\end{aligned} \tag{17.17}$$

Iterate Steps 1-4 until α and β converge:

1. Optimize α keeping β constant to match the average occupancy to the target value:

$$\min_{\alpha} \left| y_{aim} - \sum_{i=1}^{L} \frac{y(i)}{L} \right| \tag{17.18}$$

2. Compute the expected occupancy shift from the new α:

$$z_{avg} := \frac{\sum_i z_i \left[p(i) - p'(i)\right]}{\sum_i \left[p(i) - p'(i)\right]} \tag{17.19}$$

3. Optimize β keeping α constant:

$$E(i) = \beta(z_i - z_{avg}) + \beta'(z_{avg} - \alpha) \max_{\beta} \text{ cor}\left(y(i), \sum_{j=i-73}^{i+73} e^{z_j}\right) \tag{17.20}$$

4. update α for the new β and update the binding energies:

$$\begin{aligned} \alpha &= z_{avg} + \frac{\beta'}{\beta}(z_{avg} - \alpha) \\ E(i) &= \beta(z_i - \alpha) \end{aligned} \tag{17.21}$$

The complexity of steps 2 and 3 stem from attempting to minimize how the β optimization affects the average occupancy. Without this, the two parameter optimizations work against each other. This can lead to the parameter cycling between values instead of converging. For example, β_1 is optimal for α_1, but α_2 is then needed to shift the average occupancy back to the mean, which makes β_2 optimal. If optimizing α for β_2 leads back to α_1 an infinite loop is created. Such infinite loops appeared repeatedly before adjusted the method to attempt to conserve the average occupancy.

After resolving this issue I applied the thermodynamic model with the obtained parameters and the resulting genomic occupancy tracks looked promising. However, neither the models optimized on the corrected MNase-Seq and CC-Seq data nor the corrected data themselves showed improved similarities. One explanation for the lower correlations could be the reduction of common biases in both datasets. This is in accord with the correlation between both datasets and G+C content decreasing after the correction.

While the correction does what it should, the corrected datasets still disagree. Processing the data with the thermodynamic model might do more harm than good. Furthermore, my uncorrected MNase-Seq model outperforms the corrected model on the corrected data. This is counter intuitive and my best guess is that an increase of noise in the corrected data decreases the quality of the optimized energy model more than the model benefits from learning on the corrected data. Given these results I decided that correcting experimental biases in nucleosome data is a futile task. Instead I focused on describing biases in my probabilistic model and analyzing the results.

18. Simulating the MNase digestion

After repeatedly failing to correct MNase-Seq's biases, I wondered how complex the MNase digestion process is and how MNase's known sequence preference affects the experimental measurements. To investigate these questions, I developed a simulation of the MNase-Seq experiment. The simulation relies on stochastic processes (sampling), because I wanted to observe effects – not enforce assumptions through a probabilistic model. I later found out that Rizzo et al. (2012) published a similar simulation.

The first effect I analyzed was how the size-dependent accessibility of linkers and nucleosome-free regions affects the recovery frequency of neighboring nucleosomes in the mono-nucleosome fragments. Later I extended the simulation to include MNase's sequence bias and its periodic accessibility pattern of nucleosome-bound DNA to analyze more phenomena. Due to runtime restrictions I ignored MNase's pseudo-exonuclease activity and could only model partial nucleosome unwrapping as a constant higher digestion probability close to the nucleosome borders.

18.1 Simulation Steps

The simulation consists of four steps that can be replaced and extended independently:

1. Sample the nucleosome positions for each simulated genome
2. Create the digestion profile of each sampled nucleosome arrangement
3. Sample MNase cut sites based on the digestion profile and level
4. Extract mono-nucleosome fragments and compute the coverage profiles

1. Sample the nucleosome positions for each simulated genome

The first step samples the nucleosome arrangement hundreds of times for a single DNA region. I refer to this DNA region as 'genome' for simplicity. I tested artificial nucleosome

arrays and positioning the nucleosomes based on experimental measurements.

For my first artificial test I defined a perfect nucleosome array, where each nucleosome had a single-base-pair position. The nucleosome frequencies were above 90% with minor variations, and the linker lengths were based on measured linker lengths. I removed nucleosome positions from the array to create nucleosome free regions. Later I added fuzziness to the nucleosome positions by replacing the precise array positions with an exponentiated sin curve. I then enforced a minimal distance between the sampled nucleosomes to stop them from overlapping.

I also used this approach to sample nucleosomes based on experimental data. First I derived dyad frequencies from the experimental measurements to sample the nucleosome arrangements. Then I sampled the nucleosomes sequentially without overlap, preventing issues arising from the derived nucleosome-dyad frequencies leading to unrealistic occupancies (Section 13.2).

2. Create the digestion profile of each sampled nucleosome arrangement

The second step generates a digestion profile of MNase's relative probability to cut every genome position. In the simplest version each nucleosome dyad is extended to a 147-bp stretch of lower (e.g. 2%) cut probability. Other versions change the nucleosome protection footprint to have a periodic fluctuation, a decreasing cutting chance towards the dyad or both. I also tested increasing the size of the nucleosome-protected footprint to approximate a steric clash between the histones and MNase at the nucleosome's border.

The sequence preference of MNase can be added by multiplying the relative probability of every position with a MNase-preference score. To model MNase's sequence preference I derived a PWM from MNase-Seq control experiments on naked genomic DNA. I defined MNase's preference as the PWM score of the local sequence.

3. Sample MNase cut sites based on the digestion profile and level

The third step samples the MNase cut sites using the normalized digestion profiles. The samples are chosen from the combined profiles of all sampled nucleosome arrangements. The simulated digestion level is regulated by linearly scaling the amount of sampled cut sites.

I sample the cut sites with replacement to reduce the computation time, which allows repeated sampling of the same site. This reduces MNase's efficiency with the increasing level of digestion, because the probability of sampling a previous cut site increases. Sampling the cut sites independent of previous cuts means that the method cannot model

several aspects of MNase digestion – primarily pseudo-exonuclease activity and the effects of the digestion on nucleosome unwrapping.

A possible idea I had to include these aspects was to split the total digestion time into rounds: after every round the digestion profiles are updated based on the previous cut occurrences. I never implemented this approach, and I am unsure how well it would simulate such aspects without increasing the computational time too much.

4. Extract mono-nucleosome fragments and compute the coverage profiles

The final step filters the digested genome fragments by length and computes their coverage. In my tests I used fragment filters of mono-nucleosome length. I visualized the coverage like usual MNase-Seq measurements.

Before filtering, the fragment length distribution can also be extracted. I visualized this distribution the way measurements of Bioanalyzer are. This allowed me to compare my simulated distribution with the real one.

18.2 Discussion

The published simulation primarily distinguishes between nucleosomal and open DNA (Rizzo et al., 2012). They decreased the MNase-digestion probability towards the nucleosome dyad, but this had little effect on their result. With their analysis they claimed that higher digestion levels, where more fragments have the length of mono-nucleosomes, lead to less biases. Based on their simulation a high digestion level with results in 100% mono-nucleosome fragments is ideal. Differences in the accessibility then lead to no meaningful difference between nucleosome retrieval frequencies. However, in praxis such long digestions show an enrichment of sub-nucleosomal fragments (Part II), an aspect their simulation fails to replicate.

For this reason, I added MNase's sequence-dependent cut bias to my model. The most preferred cut sites that are covered by nucleosomes are cut more frequent than the most disliked cut sites in open linkers. While partial nucleosome unwrapping is the most probable origin of nucleosome internal cuts (Section 8.6), a realistic simulation of the unwrapping process would increase the computational complexity of the whole simulation drastically. Instead, I approximated the effect by decreasing the average difference in cut frequency between covered and open DNA. Together with the inclusion of MNase's sequence preference, this change led to nucleosomes being digested away at different rates. This decreases the uniformity of the simulated measurements further and the proposed

100% mono-nucleosome digestion is impossible to achieve due to over-digestion.

A result that surprised me, was how strong minor differences in the nucleosome frequencies could be amplified in the measured data. Together with the different accessibility and over-digestion rates of nucleosomes, it explained why I had such a difficult time trying to correct experimental MNase-Seq measurements. All the effects interlock and create a complex non-linear system, which my sequence-bias models are unable to capture. A single aspect of the biases cannot easily be extracted and modeled without considering the other parts.

I briefly experimented with how well my simulation could recreate MNase-Seq measurements, if I used CC-Seq data as the ground truth of nucleosome positioning. My simulated MNase digestion data have little in common with the experimental MNase-Seq measurements. Given that most of my assumptions rely on few – if any – experimental measurements and my simulation is vastly simplified, my expectations were low to begin with. A solid approximation of the MNase digestion would help focus the development of nucleosome-position prediction methods that model experimental biases and errors.

19. Analyses of experimental measurements

Attempting to learn a more quantitative nucleosome binding energy model I was interested in new experimental protocols to measure nucleosomes genome wide. Section 3 gives brief descriptions, which contain some limitations of the methods. Reading the publications and investigating the datasets I found several issues. The obvious issues I discussed in their description, here I describe some further issues in more depth. These issues might appear to be nitpicking on my part, but I believe critically revising published results is vital to the integrity of the scientific process. It serves no-one (except possibly the authors) if a claim of "unbiased chromatin accessibility profiling" (Chen et al., 2014) is not backed up by data (Section 3.3.4).

19.1 CC-Seq confirms previous findings, or does it?

Nucleosomes measured with CC-Seq have an A and T enrichment at the -3 and +3 positions from the nucleosome dyad, respectively (Section 3.2). I showed that these enrichments are strand-specific biases of CC-Seq (Section 13.4). In the original publication, the authors conceded that the enrichments could be an experimental bias (Brogaard et al., 2012). In a later publication, extending the deconvolution model, they claimed the enrichments validated previous results (Xi et al., 2014). Specifically they say: "The most striking pattern is that about 60% of the unique nucleosomes have a nucleotide 'A' at the -3 position of dyad (or 'T' at +3 position by symmetry), which confirms the finding of a previously published *in vitro* MNase study (Thåström et al., 2004)".

Checking the cited study, they indeed see an enrichment of A at the -3 position: 18 of 34 sequences (53%) have an A at that position. This is a weaker enrichment and does not stand out compared to some other positions in the study, e.g. the -5 position has an A in 20 of the 37 sequences (54%). The CC-Seq data also lacks the strongest enrichment in

the *in vitro* MNase study: the -14 position has a 84% frequency of A (35 of 37 sequences).

The amount of sequences varies between the positions, because unknown bases (Ns) are ignored and they are frequent in the dataset. This brings me to my next point. I find it quite dubious that in this day and age, where millions of nucleosome-bound sequences are measured regularly and BunDLE-Seq has been used to measure the *in vitro* nucleosome preference of thousands of unique short fragments (Levo et al., 2015), a decade-old study based on 19 fragments is cited to confirm findings.

Developing a new protocol without investigating possible experimental biases properly is bad enough. It moves academic research away from scientific rigor towards sensationalism. Validating the possible biases as signal by cherry-picking previous studies is worse.

19.2 NOMe-Seq has a strand-specific bias and published datasets are under-sequenced

In Section 20.1 I describe a variation of my method that could learn nucleosome binding preferences from NOMe-Seq data. Before fully developing the method, I analyzed the NOMe-Seq measurements to estimate the quality of the datasets. The error rates of the methylation and bisulfite conversation can be estimated from the raw data, but I did not investigate them and relied on the published, pre-processed data. I analyzed the noise and bias by comparing the measured occupancy values between strands and published datasets.

As Section 3.3.10 describes in detail, NOMe-Seq measures nucleosome occupancy via methylation of the C at GpC positions. A sequenced read captures the methylation status of all Cs for one strand. For each GpC the occupancy measurement of C on the positive strand and of the C on the negative strand (G on the positive strand) are independent, i.e. the two occupancy values are averages over disjoint sets of reads. This provides an easy way to estimate the experimental noise, because the two independent measurements are one base-pair apart and should have near identical occupancies.

For the datasets I analyzed, Pearson's correlation coefficients for this comparison are in the range of 0.04 to 0.06. This is extremely low for replicates, and the data even stem from the same experiment. The main reason for these low correlations is a low sequence coverage. Reads with a combined length of between 11.8 and 36.4 gigabases were used to construct the dataset (Kelly et al., 2012), which is equivalent to a 3.5- to 11-fold coverage. Because each read only measures the methylation for the Cs on one strand, that equates to an average of 1.8 to 5.5 data points per position. Similar genome-wide

human datasets are also frequently under-sequenced, due to the large genome size and the cost of sequencing.

Comparing the occupancy measurements of proximal positions on the same strand produces reasonable correlations of ~0.6. This value suggests that the measurements do contain consistent signals and the low correlations between strands stems from high noise. To interpret the correlation coefficient further, the frequency with which a nucleosome-linker border is between the position pairs would have to be accounted for based on their distance. The correlation, as I computed it, is a mixture of proximal measurements from the same fragment (dependent) and from different fragments (independent), which should be separated of also accounted for.

Last, I correlated the occupancy measurements of the same positions between experiments. The Pearson's correlation coefficients range from 0.07 to 0.12, which is higher than the correlations comparing the two GpC positions between the strands. The experimental conditions might have the same nucleosome arrangements, therefore I assumed that they represented replicates for this analysis. Higher correlations between replicate measurements on the same strand compared to measurements between the strands suggests that the experimental protocol has a strand-specific bias. Examining the experimental protocol I surmised that the source of such an experimental bias would probably depend on the local sequence. However, the nucleotide frequencies around sites with different methylation combinations of the two GpC positions match and cannot be the source. While I could guess, I have no solid hypothesis for the origin or behavior of this bias.

19.3 Does ChIP-exo measure half-nucleosomes?

ChIP-exo has a higher resolution than ChIP-Seq or MNase-Seq (Section 3.3.7). The resolution is high enough to separate the positions of the two histone copies of positioned nucleosomes (Rhee et al., 2014). Based on these measurements the authors made a bold claim: "We detect widespread subnucleosomal structures in dynamic chromatin, including what appear to be half-nucleosomes consisting of one copy of each histone". If this were true, subnucleosomal structures should be included in a thermodynamic model of nucleosomes.

Reading the paper, the evidence that supports this claim in comparison to alternative explanations appeared lacking. To be clear, this is not a criticism of the quality of their data or most of the analysis they did. I simple believe that extraordinary claims require extraordinary proof. I decided to investigated the study in detail, partially because

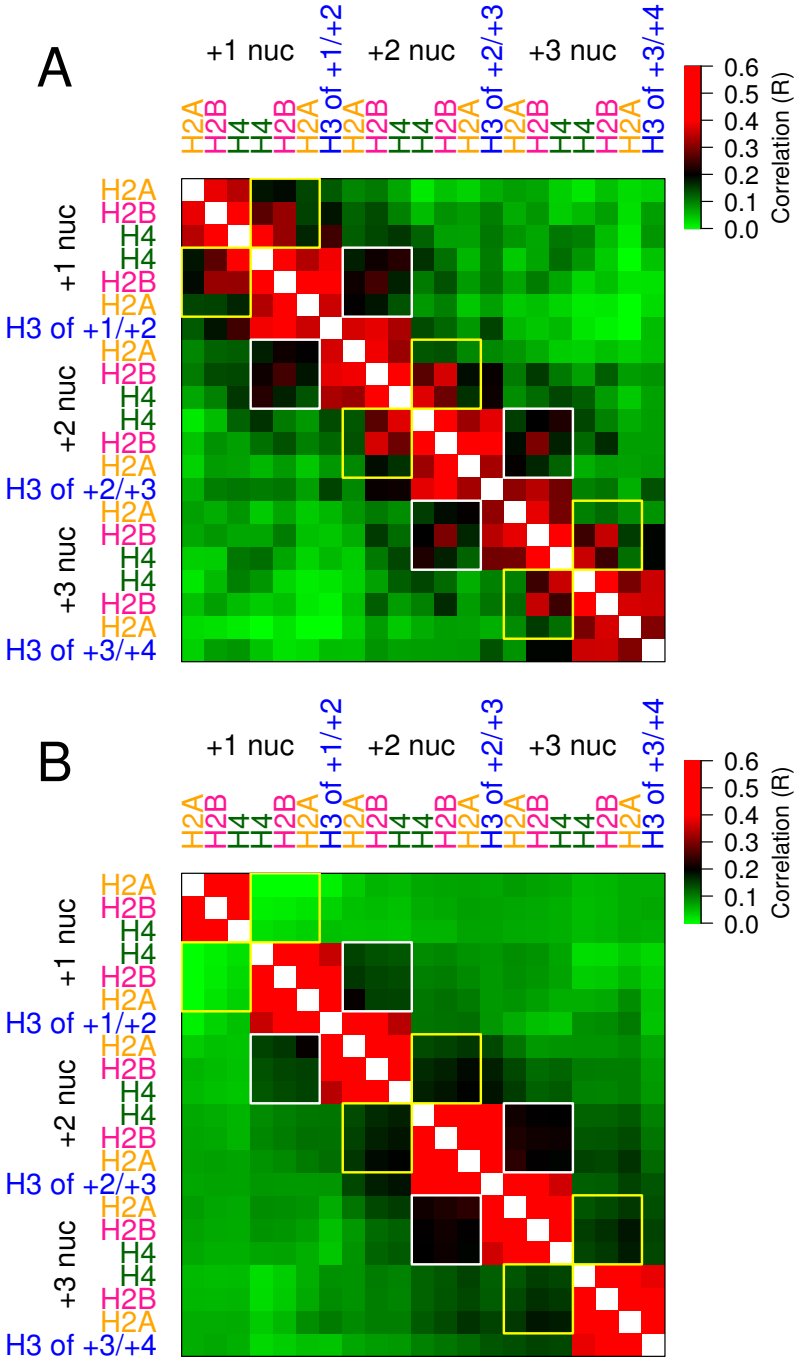

Figure 19.1: **G+C content could explain the ChIP-exo signal**: (A) Reproduction of Figure 2A from Rhee et al. (2014). Pearson's correlation coefficients between ChIP-exo counts ($\log_{10}$) in bins representing the binding regions of individual histones. (B) Same as (A) but using G+C counts instead of experimental measurements.

another publication by the laboratory using ChIP-exo data was retracted (Venters and Pugh, 2013). An analysis by our laboratory had revealed that the main claims of the publication were false (Siebert and Söding, 2014).

The first assumption anybody should have when analyzing a novel dataset, is that it contains a bias – the question is whether the signal is sufficiently stronger or not. The authors mention control analysis to exclude possible biases, which is good, but they do not go into details. The data has a strong correlation with G+C content, which is visible in several of their figures. Based on my analysis of other nucleosome measurements, I had experience with G+C biases that are hard to understand and cause issues. Therefore, I investigated what for results I would expect if the measurements consisted solely of a G+C bias.

Figure 19.1A is a recreation of *Figure 2A* in their publication (Rhee et al., 2014). The low correlations between intra-nucleosome halves (yellow boxes) suggest that the two halves can bind independently. I did an analogous analysis using G+C content instead of the ChIP-exo measurements (Figure 19.1B). The correlations between G+C content of the histones in one half are high, because the regions of the histones strongly overlap. The G+C content of the intra- and inter-nucleosome halves have much less in common. The expected effects of a sequence bias matches the observations they use as evidence for half-nucleosomes.

Assuming the measurements are bias-free and their interpretation is correct, what else would follow from the data? This line of thought matches how other experimental protocols imply an unrealistically-low average genomic occupancy (Section 13.2). To produce the lower correlations between the two nucleosome halves (Figure 19.1A) nearly as many cells have to have half-nucleosomes bound at these sites as cells with nucleosomes bound at the sites. A similar conclusion follows from about 50% of the analyzed nucleosomes having a >2-fold difference between the H2B counts of the two halves. If a significant fractions of the cells have half-nucleosomes at every analyzed nucleosome position, then a large fraction of the genome is open to other DNA binding factors, which is not measured in chromatin accessibility experiments. A possibility, might be a higher-order structure or another factor protects the frequent ~95- and ~170-bp gaps, but the larger gaps are of similar size to the nucleosome-depleted regions of promoters. The above estimate of the half-nucleosome frequency assumes that they strongly prefer binding to one side or the other. Assuming a weaker or no preference can lead to hardly any full-nucleosomes based on this interpretation of the measurements.

In their *Figure 3A* they use CC-Seq data as an independent validation. They show

a correlation between the ChIP-exo H4 count ratio and CC-Seq's strand signal for the +1 nucleosome. The nucleosome half with higher H4 signal coincides with more CC-Seq fragments going in that direction. The way I understand their argument is that a H4 imbalance means that during the CC-Seq protocol there is less chance to cut at the site, because the chemical cleavage relies on a modified H4. They further claim CC-Seq measures the occupancy of the histones pointing towards the included linker – not the occupancy of the whole nucleosomes. However, they themselves earlier correctly state that both strands are cut during CC-Seq. This means that even for a half-nucleosome the chemical cleavage produces fragments in both directions. Therefore, a parable shaped relationship is expected for both CC-Seq strand data, not anti-correlated linear relationships.

The study also includes good independent validations: MNase-ChIP-Seq and MNase-Seq measurements with low digestion levels. In both cases, short nucleosome fragments that cover half a nucleosome are recovered. At low digestion levels, over-digested nucleosomes are rare. Without inspecting the data and its processing in detail, I cannot discern how frequent the sub-nucleosome fragments are compared to mono-nucleosome fragments in these datasets. Based on the *D.melanogaster* data I analyzed, they are very rare in comparison (Figure 8.1). Their frequency increases once over-digestion becomes common due to the digestion level.

This brings me to my final two points, I believe many of the phenomena could be explained by partial nucleosome unwrapping. The only assumption needed is that partial unwrapping reduces the cross-linking frequency of ChIP-exo noticeably. While discussing the subnucleosomal structures the possibility of partially-wrapped nucleosomes is never mentioned by the authors. A second explanation could be non-canonical arrangements of the nucleosomes in a subset of cells. The ChIP-exo coverage conforms with this possibility: the linker valleys are at least half the hight of the peaks at the canonical histone positions, even if the data is aligned by the +1 nucleosome.

In summary: a strong G+C bias would lead to similar results; if the data is unbiased it implies low nucleosome occupancies; CC-Seq measurements do not validate the claim; and partial nucleosome unwrapping or non-canonical nucleosome arrangements might explain the data. While half-nucleosomes are possible, they are unlikely to be a widespread phenomenon. Therefore, I neglected the possibility and did not develop a variant of my nucleosome-position prediction method that includes them.

20. Optimizing a nucleosome energy model on occupancy measurements

Experiments like NA-Seq and NOMe-Seq measure the nucleosome occupancy at given positions instead of measuring the positions of nucleosomes like most other methods (Sections 3.3.3 and 3.3.10). The way the probabilistic data model of my method represents the data points does not reflect such occupancy measurements. I developed a model variant for such data that uses an alternative likelihood based on the occupancy equation of the thermodynamic model (Equation 12.4).

$$\mathcal{L} = \prod_{n=1}^{N} \mathrm{P}(x_n\text{occupied}|S,\epsilon,\mu,\theta)^{o_n}\,\mathrm{P}(x_n\text{unoccupied}|S,\epsilon,\mu,\theta)^{u_n} \tag{20.1}$$

With o_n and u_n being the counts of occupied and unoccupied measurements for position x_n. Below I describe a variant with experimental error terms, which are similar to the positional-uncertainty terms of my usual probabilistic model for nucleosome-position data. Before coming to this more complex model, I will explain the concept with the simple version without error terms:

$$\begin{aligned}\mathrm{P}(x_n\text{occupied}|S,\epsilon,\mu,\theta) &= \mathrm{Occ}(x_n)\\ &= \frac{\sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^*}{\sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + F_{x_n} B_{x_n}}\end{aligned} \tag{20.2}$$

$$\begin{aligned}\mathrm{P}(x_n\text{unoccupied}|S,\epsilon,\mu,\theta) &= 1 - \mathrm{P}(x_n\text{occupied}|S,\epsilon,\mu,\theta)\\ &= \frac{F_{x_n} B_{x_n}}{\sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + F_{x_n} B_{x_n}}\end{aligned}$$

The thermodynamic model and Forward/Backward computations are unchanged, as are their partial derivatives. The partial derivation of the log-likelihood is a simple exercise:

$$\begin{aligned}\frac{\partial \log \mathcal{L}}{\partial \epsilon, \mu} = &\sum_{n=1}^{N} o_n \frac{\sum_{i=x_n-D_N}^{x_n+D_N} (B_i^* \partial F_i^* + F_i^* \partial B_i^*)}{\sum_{i=x_n-D_N}^{x_n+D_N} F_i^* B_i^*} + \sum_{n=1}^{N} u_n \frac{B_{x_n} \partial F_{x_n} + F_{x_n} \partial B_{x_n}}{F_{x_n} B_{x_n}} \\ &- \sum_{n=1}^{N} (o_n + u_n) \frac{\sum_{i=x_n-D_N}^{x_n+D_N} (B_i^* \partial F_i^* + F_i^* \partial B_i^*) + B_{x_n} \partial F_{x_n} + F_{x_n} \partial B_{x_n}}{\sum_{i=x_n-D_N}^{x_n+D_N} F_i^* B_i^* + F_{x_n} B_{x_n}}\end{aligned} \tag{20.3}$$

With experimental error rates

The experimental errors of these measurements have yet to be as extensively analyzed as MNase-Seq's. Their error rates are probably high enough to cause problems for a model that assumes perfect measurements. I added basic error rates that are independent of the local sequence to the model. Nothing prevents a variation that uses sequence-dependent error rates (i.e. biases) from working. The sequence-dependent bias could be encoded in the same way it is for CC-Seq data (Section 12.6.3).

Using $\mathrm{P}(x_n = o)$ as the probability of measuring $x_n = o$, $\mathrm{P}(x_n\text{occupied})$ as the probability of the position x_n actually being occupied, and likewise for u, the probabilities with errors are:

$$\begin{aligned}\mathrm{P}(o|S, \epsilon, \mu, \theta) &= \mathrm{P}(o|x_n\text{occupied}, \theta)\, \mathrm{P}(x_n\text{occupied}|S, \epsilon, \mu, \theta) \\ &\quad + \mathrm{P}(o|x_n\text{unoccupied}, \theta)\, \mathrm{P}(x_n\text{unoccupied}|S, \epsilon, \mu, \theta) \\ &= q_o \frac{\sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^*}{\sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + F_{x_n} B_{x_n}} + (1 - q_u) \frac{F_{x_n} B_{x_n}}{\sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + F_{x_n} B_{x_n}} \\ \mathrm{P}(u|S, \epsilon, \mu, \theta) &= \mathrm{P}(u|x_n\text{occupied}, \theta)\, \mathrm{P}(x_n\text{occupied}|S, \epsilon, \mu, \theta) \\ &\quad + \mathrm{P}(u|x_n\text{unoccupied}, \theta)\, \mathrm{P}(x_n\text{unoccupied}|S, \epsilon, \mu, \theta) \\ &= (1 - q_o) \frac{\sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^*}{\sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + F_{x_n} B_{x_n}} + q_u \frac{F_{x_n} B_{x_n}}{\sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + F_{x_n} B_{x_n}}\end{aligned}$$

with:

$$\begin{aligned}q_o &= \mathrm{P}(o|x_n\text{occupied}) \;=\; \text{positive predictive value} \\ q_u &= \mathrm{P}(u|x_n\text{unoccupied}) \;=\; \text{negative predictive value}\end{aligned} \tag{20.4}$$

Which results in the likelihood:

$$\begin{aligned}
\mathcal{L} = \prod_{n=1}^{N} & \left(\frac{q_o \sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + (1-q_u) F_{x_n} B_{x_n}}{\sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + F_{x_n} B_{x_n}} \right)^{o_n} \\
& \left(\frac{(1-q_o) \sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + q_u F_{x_n} B_{x_n}}{\sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + F_{x_n} B_{x_n}} \right)^{u_n} \\
\log \mathcal{L} = & \sum_{n=1}^{N} o_n \log \left(q_o \sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + (1-q_u) F_{x_n} B_{x_n} \right) \\
& + \sum_{n=1}^{N} u_n \log \left((1-q_o) \sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + q_u F_{x_n} B_{x_n} \right) \\
& - \sum_{n=1}^{N} (o_n + u_n) \log \left(\sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + F_{x_n} B_{x_n} \right)
\end{aligned} \tag{20.5}$$

Calculating the partial derivatives follows the same scheme as the partial derivatives for the model without error terms. The mini-batch gradient ascent can also optimize the error rates q_o and q_u based on their partial derivatives:

$$\begin{aligned}
\frac{\partial \log \mathcal{L}}{\partial q_o} = & \sum_{n=1}^{N} o_n \frac{\sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^*}{\left(q_o \sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + (1-q_u) F_{x_n} B_{x_n} \right)} \\
& - \sum_{n=1}^{N} u_n \frac{\sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^*}{\left((1-q_o) \sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + q_u F_{x_n} B_{x_n} \right)} \\
\frac{\partial \log \mathcal{L}}{\partial q_u} = & - \sum_{n=1}^{N} o_n \frac{F_{x_n} B_{x_n}}{\left(q_o \sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + (1-q_u) F_{x_n} B_{x_n} \right)} \\
& + \sum_{n=1}^{N} u_n \frac{F_{x_n} B_{x_n}}{\left((1-q_o) \sum_{j=x_n-D_N}^{x_n+D_N} F_j^* B_j^* + q_u F_{x_n} B_{x_n} \right)}
\end{aligned} \tag{20.6}$$

20.1 Optimizing on NOMe-Seq measurements

The above model for nucleosome occupancy data assumes that each measurement is independent. In NOMe-Seq most sequenced fragments contain two or more GpC positions that can be methylated. The occupancy values measured with one fragment depend on each other and break the above assumed independence. Due to the short read length used in the published datasets, the average number of measurements on one fragment is low, which limits the relevance of this dependence. However, nothing in the protocol should prevent the use of longer reads to increase the information gained from modeling this

dependence. I developed a probabilistic model for fragment measurements that includes the dependency between the measurements of each fragment.

The underlying idea is to compute the probability of observing the set of measured occupancies of a fragment, given a fragment was recovered that covers the region. After coming up with several versions that had artificial limits on the possible fragment sizes, due to time complexity constraints, I found a solution that works independent of the fragment size. The statistical weight $\tilde{\mathrm{P}}$ given the measurements of a fragment is computed with a conditional Forward/Backward algorithm (below I show that computing one direction is enough). To normalize these statistical weights, they are divided by the sum of statistical weights for all possible measurements the fragment could have produced. I describe the conditional Forward/Backward algorithm below, its time complexity is linear as that of the normal Forward/Backward algorithm.

Computing the conditional Forward/Backward weights for all possibilities would be time intensive for long fragments, because the amount of possibilities is exponential in the amount of measurements the fragment contains. However, I found a much simpler solution that seems trivial in hindsight: the case of allowing all possible measurements is identical to the basic Forward/Backward algorithm once the equations are grouped and simplified. The sum of all possible fragment measurements is identical to the occupancy normalization above: the sum off all dyad positions that lead to nucleosome coverage, plus the open case for any given position.

For simplicity I chose a position k that lies to the side of the fragment, D_N after the fragment to be precise. The method computes the conditional Forward statistical weights F^f for the fragment region plus D_N. The probability of observing the fragments measurements is then the fraction of the conditional sum over the unconditional sum for position k.

Before describing the model equations we need a new encoding for the NOMe-Seq measurements: x_f represents all measurements of fragment f, i.e. a vector containing the individual position measurements $x_{f,j}$, which each represents a measurement of fragment f at position j. $x_{f,j}$ is u if unmethylated, m if methylated, and $\varnothing$ if the position cannot be methylated or is beyond the fragment. l_f is the left fragment border and r_f the right

border.

$$\begin{aligned}
\mathrm{P}(x_f|S,\theta) &= \frac{\tilde{\mathrm{P}}(\text{all nucleosome configurations}, x_f|S,\theta)}{\tilde{\mathrm{P}}(\text{all nucleosome configurations}|S,\theta)} \\
\forall k: & \\
\mathrm{P}(x_f|S,\theta) &= \frac{\tilde{\mathrm{P}}(\text{all Paths through } k, x_f|S,\theta)}{\tilde{\mathrm{P}}(\text{all Paths through } k|S,\theta)} \\
&= \frac{Z_k^f + \sum_{d=-D_N}^{D_N} Z_{k+d}^{f*}}{Z_k + \sum_{d=-D_N}^{D_N} Z_{k+d}^{*}}
\end{aligned} \tag{20.7}$$

for convenience:

$$k := r_f + D_N$$

$$\begin{aligned}
Z_i &= F_i B_i \\
Z_i^* &= F_i^* B_i^* \\
Z_i^f &= F_i^f B_i^f \\
Z_i^{f*} &= F_i^{f*} B_i^{f*} \\
F_{i+1}^f &= \left(F_i^f + F_{i-D_N}^{f*}\right) \mathrm{P}(x_{f,i+1}|\text{open}) \\
F_{i+1}^{f*} &= F_{i-D_N}^f e^{E_{i+1}-\mu} \mathrm{P}(x_{f,i+1-D_N}, ..., x_{f,i+1+D_N}|\overline{\text{open}}) \\
B_{i-1}^f &= B_i^f + B_{i+D_N}^{f*} e^{E_{i+D_N}-\mu} \mathrm{P}(x_{f,i}, ..., x_{f,i+2D_N}|\overline{\text{open}}) \\
B_{i-1}^{f*} &= B_{i+D_N}^f \mathrm{P}(x_{f,i+D_N}|\text{open})
\end{aligned} \tag{20.8}$$

As with the energy term, the conditionals of the Backward terms have to lag behind. I use $\mathrm{P}(\varnothing|...) = 1$ as a convenience for simplicity, technically the probability is conditioned on the position being methylatable or not. In other contexts this does theoretically break mathematical axioms of probabilities, but I am simply using it as a shorthand here. Together with $\mathrm{P}(\mathrm{m}|\text{open}) = 1$ and $\mathrm{P}(\mathrm{u}|\overline{\text{open}}) = 1$ the conditional equations break down to:

$$\begin{aligned}
F_i^f = F_i \;;\; F_i^{f*} = F_i^* \qquad & \forall i < l_f \qquad \text{i.e. } i \text{ before any measured position} \\
B_i^f = B_i \;;\; B_i^{f*} = B_i^* \qquad & \forall i > r_f \qquad \text{i.e. } i \text{ after any measured position} \\
& \qquad\qquad\qquad\qquad\quad \text{(before when going backwards)}
\end{aligned} \tag{20.9}$$

And otherwise:

$$
\begin{aligned}
F^{f}_{i+1} &= \left(F^{f}_{i} + F^{f*}_{i-D_N}\right) \mathrm{I}(x_{f,i+1} \neq \mathrm{u}) \\
F^{f*}_{i+1} &= F^{f}_{i-D_N} e^{E_{i+1}-\mu} \mathrm{I}(\sum_{j=i+1-D_N}^{i+1+D_N} \mathrm{I}(x_{f,j} = \mathrm{m}) = 0) \\
&= F^{f}_{i-D_N} e^{E_{i+1}-\mu} \prod_{j=i+1-D_N}^{i+1+D_N} \mathrm{I}(x_{f,j} \neq \mathrm{m}) \\
B^{f}_{i-1} &= B^{f}_{i} + B^{f*}_{i+D_N} e^{E_{i+D_N}-\mu} \prod_{j=i}^{i+2D_N} \mathrm{I}(x_{f,j} \neq \mathrm{m}) \\
B^{f*}_{i-1} &= B^{f}_{i+D_N} \mathrm{I}(x_{f,i+D_N} \neq \mathrm{u})
\end{aligned}
\tag{20.10}
$$

The log-likelihood is:

$$
\begin{aligned}
\log \mathcal{L} &= \sum_{f=1}^{N} \left(\log(Z^{f}_{k} + \sum_{d=-D_N}^{D_N} Z^{f*}_{k+d}) - \log(Z_k + \sum_{d=-D_N}^{D_N} Z^{*}_{k+d}) \right) \\
&= \sum_{f=1}^{N} \left(\log(Z^{f}_{k} + \sum_{d=-D_N}^{D_N} Z^{f*}_{k+d}) \right) - N \log(Z_k + \sum_{d=-D_N}^{D_N} Z^{*}_{k+d})
\end{aligned}
\tag{20.11}
$$

Note that k can depend on the fragment, i.e. k_f, but does not have to. In theory the normalization terms are independent of the chosen k and can be factored out even if k depends on the fragment. In practice this might not hold due to numerical errors accumulated in the computation, which is why I suggest normalizing each fragment individually. Other parts of the computation dwarf the extra computation time needed, given that the partial derivatives share the same normalization terms.

Partial Derivatives

The partial derivatives of the conditional Forward equations are identical to the unconditional Forward equations before the fragment begins and the same is true for the Backward equations coming from the other side. Between this and computing the Forward weights for the whole fragment, the conditional Backward equations and their partial derivatives are unnecessary. I describe their equations nonetheless for completeness sake and in case

they have a use I missed.

$$\begin{aligned}\frac{\partial \log \mathcal{L}}{\partial \epsilon, \mu} &= \sum_{f=1}^{N} \left(\frac{\frac{\partial Z_k^f}{\partial \epsilon,\mu} + \sum_{d=-D_N}^{D_N} \frac{\partial Z_{k+d}^{f*}}{\partial \epsilon,\mu}}{Z_k^f + \sum_{d=-D_N}^{D_N} Z_{k+d}^{f*}} - \frac{\frac{\partial Z_k}{\partial \epsilon,\mu} + \sum_{d=-D_N}^{D_N} \frac{\partial Z_{k+d}^{*}}{\partial \epsilon,\mu}}{Z_k + \sum_{d=-D_N}^{D_N} Z_{k+d}^{*}} \right) \\ &= \sum_{f=1}^{N} \left(\frac{\frac{\partial F_k^f}{\partial \epsilon,\mu} B_k^f + F_k^f \frac{\partial B_k^f}{\partial \epsilon,\mu} + \sum_{d=-D_N}^{D_N} \left(\frac{\partial F_{k+d}^{f*}}{\partial \epsilon,\mu} B_{k+d}^{f*} + F_{k+d}^{f*} \frac{\partial B_{k+d}^{f*}}{\partial \epsilon,\mu} \right)}{F_k^f B_k^f + \sum_{d=-D_N}^{D_N} F_{k+d}^{f*} B_{k+d}^{f*}} \right. \\ &\quad \left. - \frac{\frac{\partial F_k}{\partial \epsilon,\mu} B_k + F_k \frac{\partial B_k}{\partial \epsilon,\mu} + \sum_{d=-D_N}^{D_N} \left(\frac{\partial F_{k+d}^{*}}{\partial \epsilon,\mu} B_{k+d}^{*} + F_{k+d}^{*} \frac{\partial B_{k+d}^{*}}{\partial \epsilon,\mu} \right)}{F_k B_k + \sum_{d=-D_N}^{D_N} F_{k+d}^{*} B_{k+d}^{*}} \right)\end{aligned} \tag{20.12}$$

$$\begin{aligned}\frac{\partial F_{i+1}^f}{\partial \epsilon_l(q)} &= \left(\frac{\partial F_i^f}{\partial \epsilon_l(q)} + \frac{\partial F_{i-D_N}^{f*}}{\partial \epsilon_l(q)} \right) \mathrm{I}(x_{f,i+1} \neq \mathrm{u}) \\ \frac{\partial F_{i+1}^{f*}}{\partial \epsilon_l(q)} &= \left[\frac{\partial F_{i-D_N}^f}{\partial \epsilon_l(q)} + F_{i-D_N}^f \mathrm{I}(s_{i+1\ldots} = q) \right] e^{E_{i+1}-\mu} \prod_{j=i+1-D_N}^{i+1+D_N} \mathrm{I}(x_{f,j} \neq \mathrm{m})\end{aligned} \tag{20.13}$$

$$\begin{aligned}\frac{\partial B_{i-1}^f}{\partial \epsilon_l(q)} &= \frac{\partial B_i^f}{\partial \epsilon_l(q)} + \left[\frac{\partial B_{i+D_N}^{f*}}{\partial \epsilon_l(q)} + B_{i+D_N}^{f*} \mathrm{I}(s_{i+D_N\ldots} = q) \right] e^{E_{i+D_N}-\mu} \prod_{j=i}^{i+2D_N} \mathrm{I}(x_{f,j} \neq \mathrm{m}) \\ \frac{\partial B_{i-1}^{f*}}{\partial \epsilon_l(q)} &= \frac{\partial B_{i+D_N}^f}{\partial \epsilon_l(q)} \mathrm{I}(x_{f,i+D_N} \neq \mathrm{u})\end{aligned} \tag{20.14}$$

$$\begin{aligned}-\frac{\partial F_{i+1}^f}{\partial \mu} &= \left(-\frac{\partial F_i^f}{\partial \mu} - \frac{\partial F_{i-D_N}^{f*}}{\partial \mu} \right) \mathrm{I}(x_{f,i+1} \neq \mathrm{u}) \\ -\frac{\partial F_{i+1}^{f*}}{\partial \mu} &= \left[-\frac{\partial F_{i-D_N}^f}{\partial \mu} + F_{i-D_N}^f \right] e^{E_{i+1}-\mu} \prod_{j=i+1-D_N}^{i+1+D_N} \mathrm{I}(x_{f,j} \neq \mathrm{m})\end{aligned} \tag{20.15}$$

$$\begin{aligned}-\frac{\partial B_{i-1}^f}{\partial \mu} &= -\frac{\partial B_i^f}{\partial \mu} + \left[-\frac{\partial B_{i+D_N}^{f*}}{\partial \mu} + B_{i+D_N}^{f*} \right] e^{E_{i+D_N}-\mu} \prod_{j=i}^{i+2D_N} \mathrm{I}(x_{f,j} \neq \mathrm{m}) \\ -\frac{\partial B_{i-1}^{f*}}{\partial \mu} &= -\frac{\partial B_{i+D_N}^f}{\partial \mu} \mathrm{I}(x_{f,i+D_N} \neq \mathrm{u})\end{aligned} \tag{20.16}$$

With experimental error rates

Based on my analysis NOMe-Seq has a strand-specific bias (Section 19.2). The published datasets also have a high noise rate, due to low genome coverage. With these issues I

was not even tempted to implement and test this model.

Even low error rates could be a problem for the basic model. Taking erroneous measurements at face value can easily lead to fragments that are impossible in the probabilistic model. For example, any occupied measurement surrounded by two open measurements that are less than 147 bps apart would lead to a fragment probability of zero. Modeling DNA-binding factors that occupy smaller footprint would reduce the problem, but masking experimental errors in such a way is inelegant. The probabilistic model should try to describe the root cause of experimental errors to improve our understanding of them. In the theme of adding experimental error to my probabilistic data models, I developed an extended version that includes error rates.

For the model with errors rates the conditional Forward equations get more complicated, but can be simplified. I exclude the conditional Backward terms, because they are unneeded.

$$\begin{aligned}
F^{f}_{i+1} &= \left(F^{f}_{i} + F^{f*}_{i-D_N}\right)\left(\mathrm{I}(x_{f,i+1} = \varnothing) + \mathrm{I}(x_{f,i+1} = \mathrm{m})\,\mathrm{P}(\mathrm{m}|\mathrm{open})\right. \\
&\quad \left. + \mathrm{I}(x_{f,i+1} = \mathrm{u})\,\mathrm{P}(\mathrm{u}|\mathrm{open})\right) \\
&= \left(F^{f}_{i} + F^{f*}_{i-D_N}\right)\mathrm{P}(\mathrm{m}|\mathrm{open})^{\mathrm{I}(x_{f,i+1}=\mathrm{m})}\,\mathrm{P}(\mathrm{u}|\mathrm{open})^{\mathrm{I}(x_{f,i+1}=\mathrm{u})} \\
F^{f*}_{i+1} &= \left(F^{f}_{i-D_N} e^{E_{i+1}-\mu}\right)\prod_{j=i+1-D_N}^{i+1+D_N}\left(\mathrm{I}(x_{f,j} = \varnothing) + \mathrm{I}(x_{f,j} = \mathrm{u})\,\mathrm{P}(\mathrm{u}|\overline{\mathrm{open}})\right. \\
&\quad \left. + \mathrm{I}(x_{f,j} = \mathrm{m})\,\mathrm{P}(\mathrm{m}|\overline{\mathrm{open}})\right) \\
&= F^{f}_{i-D_N} e^{E_{i+1}-\mu}\,\mathrm{P}(\mathrm{u}|\overline{\mathrm{open}})^{\sum_{j=i+1-D_N}^{i+1+D_N}\mathrm{I}(x_{f,j}=\mathrm{u})}\,\mathrm{P}(\mathrm{m}|\overline{\mathrm{open}})^{\sum_{j=i+1-D_N}^{i+1+D_N}\mathrm{I}(x_{f,j}=\mathrm{m})}
\end{aligned} \tag{20.17}$$

Naturally with $\mathrm{P}(\mathrm{u}|\mathrm{open}) = 1-\mathrm{P}(\mathrm{m}|\mathrm{open})$ and $\mathrm{P}(\mathrm{u}|\overline{\mathrm{open}}) = 1-\mathrm{P}(\mathrm{m}|\overline{\mathrm{open}})$. $\mathrm{I}(x_{f,i+1} = \varnothing)$ has no $\mathrm{P}(\varnothing|...)$ factor, because the probability of seeing no data is 1 for a position without a measurement.

The log-likelihood is identical to the model without error terms.

Partial derivatives for the error rates

The partial derivatives for ϵ and μ have the identity in the Forward/Backward equations replaced with the more extensive version containing error rates. I omit their equations, because they are otherwise identical. For the error rates $q_o = \mathrm{P}(\mathrm{m}|\mathrm{open})$ and $q_{\overline{o}} = \mathrm{P}(\mathrm{m}|\overline{\mathrm{open}})$ the partial derivative of the log-likelihood and conditional Forward equations

are:

$$\begin{aligned}\frac{\partial \log \mathcal{L}}{\partial q_o, q_{\overline{o}}} &= \sum_{f=1}^{N} \frac{\frac{\partial Z_k^f}{\partial q_o, q_{\overline{o}}} + \sum_{d=-D_N}^{D_N} \frac{\partial Z_{k+d}^{f*}}{\partial q_o, q_{\overline{o}}}}{Z_k^f + \sum_{d=-D_N}^{D_N} Z_{k+d}^{f*}} \\ &= \sum_{f=1}^{N} \frac{\frac{\partial F_k^f}{\partial q_o, q_{\overline{o}}} B_k^f + F_k^f \frac{\partial B_k^f}{\partial q_o, q_{\overline{o}}} + \sum_{d=-D_N}^{D_N} \left(\frac{\partial F_{k+d}^{f*}}{\partial q_o, q_{\overline{o}}} B_{k+d}^{f*} + F_{k+d}^{f*} \frac{\partial B_{k+d}^{f*}}{\partial \epsilon, \mu} \right)}{F_k^f B_k^f + \sum_{d=-D_N}^{D_N} F_{k+d}^{f*} B_{k+d}^{f*}}\end{aligned} \tag{20.18}$$

$$\begin{aligned}\frac{\partial F_{i+1}^f}{\partial q_o} &= \left(\frac{\partial F_i^f}{\partial q_o} + \frac{\partial F_{i-D_N}^{f*}}{\partial q_o} \right) q_o^{\mathrm{I}(x_{f,i+1}=\mathrm{m})} (1-q_o)^{\mathrm{I}(x_{f,i+1}=\mathrm{u})} \\ &\quad + \left(F_i^f + F_{i-D_N}^{f*} \right) \\ &\qquad \left(\mathrm{I}(x_{f,i+1}=\mathrm{m})(1-q_o)^{\mathrm{I}(x_{f,i+1}=\mathrm{u})} + q_o^{\mathrm{I}(x_{f,i+1}=\mathrm{m})} \mathrm{I}(x_{f,i+1}=\mathrm{u}) \right) \\ &= \left(\frac{\partial F_i^f}{\partial q_o} + \frac{\partial F_{i-D_N}^{f*}}{\partial q_o} \right) q_o^{\mathrm{I}(x_{f,i+1}=\mathrm{m})} (1-q_o)^{\mathrm{I}(x_{f,i+1}=\mathrm{u})} \\ &\quad + \left(F_i^f + F_{i-D_N}^{f*} \right) \left(\mathrm{I}(x_{f,i+1}=\mathrm{m}) + \mathrm{I}(x_{f,i+1}=\mathrm{u}) \right)\end{aligned} \tag{20.19}$$

$$\begin{aligned}\frac{\partial F_{i+1}^{f*}}{\partial q_o} &= \frac{\partial F_{i-D_N}^f}{\partial q_o} e^{E_{i+1}-\mu} (1-q_{\overline{o}})^{\sum_{j=i+1-D_N}^{i+1+D_N} \mathrm{I}(x_{f,j}=\mathrm{u})} q_{\overline{o}}^{\sum_{j=i+1-D_N}^{i+1+D_N} \mathrm{I}(x_{f,j}=\mathrm{m})} \\ &= \frac{\partial F_{i-D_N}^f}{\partial q_o} e^{E_{i+1}-\mu} (1-q_{\overline{o}})^{I_{f,i+1}(\mathrm{u})} q_{\overline{o}}^{I_{f,i}(\mathrm{m})}\end{aligned}$$

With:

$$I_{f,i}(\mathrm{u}) = \sum_{j=i-D_N}^{i+D_N} \mathrm{I}(x_{f,j}=\mathrm{u}) \qquad ; \qquad I_{f,i}(\mathrm{m}) = \sum_{j=i-D_N}^{i+D_N} \mathrm{I}(x_{f,j}=\mathrm{m}) \tag{20.20}$$

$$\frac{\partial F_{i+1}^f}{\partial q_{\overline{o}}} = \left(\frac{\partial F_i^f}{\partial q_{\overline{o}}} + \frac{\partial F_{i-D_N}^{f*}}{\partial q_{\overline{o}}} \right) q_o^{\mathrm{I}(x_{f,i+1}=\mathrm{m})} (1-q_o)^{\mathrm{I}(x_{f,i+1}=\mathrm{u})}$$

$$\begin{aligned}\frac{\partial F_{i+1}^{f*}}{\partial q_{\overline{o}}} &= \frac{\partial F_{i-D_N}^f}{\partial q_{\overline{o}}} e^{E_{i+1}-\mu} (1-q_{\overline{o}})^{I_{f,i}(\mathrm{u})} q_{\overline{o}}^{I_{f,i}(\mathrm{m})} \\ &\quad + F_{i-D_N}^f e^{E_{i+1}-\mu} \\ &\qquad \left((1-q_{\overline{o}})^{I_{f,i}(\mathrm{u})} I_{f,i}(\mathrm{m}) q_{\overline{o}}^{I_{f,i}(\mathrm{m})-1} - I_{f,i}(\mathrm{u})(1-q_{\overline{o}})^{I_{f,i}(\mathrm{u})-1} q_{\overline{o}}^{I_{f,i}(\mathrm{m})} \right)\end{aligned} \tag{20.21}$$

21. Outlook

MNase-Seq and other experiments that measure nucleosome-positioning genome-wide have biases. For MNase-Seq the experimental biases have been repeatedly investigated. I collaborated with the Gaul-lab to perform a comprehensive analysis of the effects the MNase-digestion level and fragment-length selection have on MNase-Seq data (Section 8.1). I gained insights into the MNase-digestion process through this analysis, which helped me simulate the process (Section 18). Further work is needed to correct the experimental measurements based on such analyses. As a first step, I found that including fragments of sub- and di-nucleosome length reduces – but does not remove – problems with the data. The mono-nucleosome fragments must not be viewed in isolation, but evaluated in the context of the MNase digestion, if one wants to derive realistic occupancies. Occupancy distributions derived solely from mono-nucleosome fragments imply an unrealistically-low average genomic occupancy (Section 13.2). This causes issues when optimizing the sequence-unspecific binding energy (Section 16.4) and probably influences the obtained sequence-specific binding-energy model. More quantitative measurements are needed to learn and model all aspects of the nucleosome binding behavior.

Experimental biases also affect similarity scores when comparing datasets (Section 13.1). The similarity is over- or underestimated if the experiments share common or have distinct biases, respectively. Such issues are frequently ignored, even though they can have severe effects and invalidate the results. I tried different approaches to correct such biases out of the data without success (Section 17). More complex and specific correction schemes could be developed to correct the biases for each experimental method. However, until we have a solid understanding of both the bias sources and their effects, I believe such attempts will be futile.

The experimental biases must be addressed nonetheless, even if they cannot be removed. Publications have to be examined thoroughly and their results and claims viewed cautiously (Section 19). Common assumptions and facts should be challenged when new evidence contradicts them. A handful of analyses suggest that nucleosomes might not

prefer G+C-rich sequences, or at least to a lesser degree than previous analyses had shown (Section 14.2). To resolve these contradictions the underlying data should be evaluated independently and these new analyses should address the experimental biases more rigorously.

With my nucleosome-position prediction method I took a step in that direction by bringing nucleosome binding energy models derived from two published datasets more in line with each other. The method separates out the strand-dependent sequence bias of CC-Seq and learns a position-specific energy model from low-resolution MNase-Seq measurements. The two energy models my method learns from CC-Seq and MNase-Seq data agree better than models derived without accounting for CC-Seq's bias or MNase-Seq's low positional resolution. However, the models still disagree on two vital aspects: the MNase-Seq energy model prefers G+C, while the CC-Seq model prefers A+T; and the position-specific preferences are three times as strong in the MNase-Seq model.

These disagreements will decrease by further improving the method's model of the experimental measurements. The likelihood can use the probability of observing the sequenced fragments instead of the processed dyad counts or scores (e.g. Section 12.6.7). A model of MNase's sequence preference at the cut sites could then be added. The theory behind these method variations is straight forward, the main work would be in the implementation: accommodating changes in the data structure and covering the edge cases. Including the continuous nature of the MNase digestion is more difficult and could fail at the theory, because the computation might be impossible in a reasonable time complexity. Based on my attempts at correcting the sequence biases, simple bias approximations might be incapable of fully converging the nucleosome models of the distinct experimental protocols.

Novel experimental protocols, like NOMe-Seq, that measure nucleosome occupancy directly and have different (and hopefully less) biases will improve our understanding further. I am collaborating with the Korber-lab, who are establish such an experiment. Some of my initial analysis look promising, but further validations are needed and being worked on. Measuring nucleosome occupancy at genomic positions, as NOMe-Seq does, provides distinct information compared to experiments that measure nucleosome-positions. Deriving high-resolution nucleosome energy models from such data is more difficult, because the measurements have a low positional resolution. I developed the theory for a variation of my nucleosome-position prediction method that can learn a nucleosome energy model from such data (Section 20). The method might be unable to learn a high-resolution model, but should be able to learn the sequence-unspecific binding

energy of nucleosomes.

Independent of which type of data the nucleosome model is learned from, the thermodynamic model of my method will have to be extended to fully describe what affects nucleosome binding *in vivo*. Published extensions of the thermodynamic model used by other methods can be incorporated into my method (e.g. Section 12.6.6). There are ample opportunities to improve the nucleosome-position prediction methods and gain more biological insights.

The endeavor to understand nucleosome positioning will also require further biochemical studies. The average genomic nucleosome occupancy is vital to model nucleosomes, but still unknown (Section 14.1). Other aspects have also yet to be analyzed in a genome-wide fashion. For example, my collaboration with the Gaul-lab was the first to investigate genome-wide partial nucleosome unwrapping in greater detail (Section 8.6). Partial nucleosome unwrapping probably plays a role in nucleosome fragility and might be linked with nucleosome remodeling.

The cumulative knowledge about nucleosomes will improve our understanding of other cell processes, such as gene regulation. The underlying motivation for my work was to improve gene-expression predictions with a more quantitative nucleosome binding energy model. I am part of a larger collaboration project with the Gaul-lab to study the effect of a promoter's sequence features on gene expression. At promoters and enhancers nucleosomes compete against transcription factors, which can hinder or promote gene expression. The binding preference of nucleosomes is therefore a vital part of a full quantitative model of gene regulation.

Part V

Appendix

A. Supplemental figures (Part II)

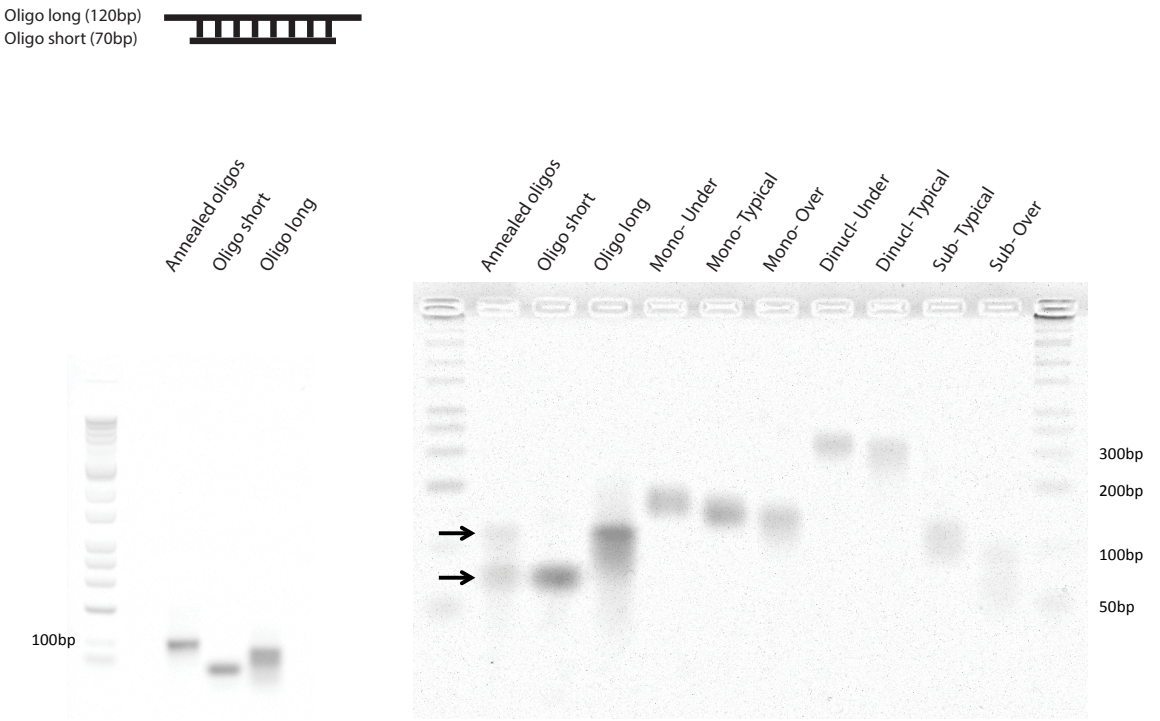

Figure A.1: **Sample fragments are free of single strand nicks**: A denaturing gel electrophoresis experiment to check if the MNase-Seq samples contain single strand nicks. Left panel: control experiment of the annealed oligo in a normal gel electrophoresis. Right panel: denaturing alkaline agarose gel of the control annealed oligos and the samples.

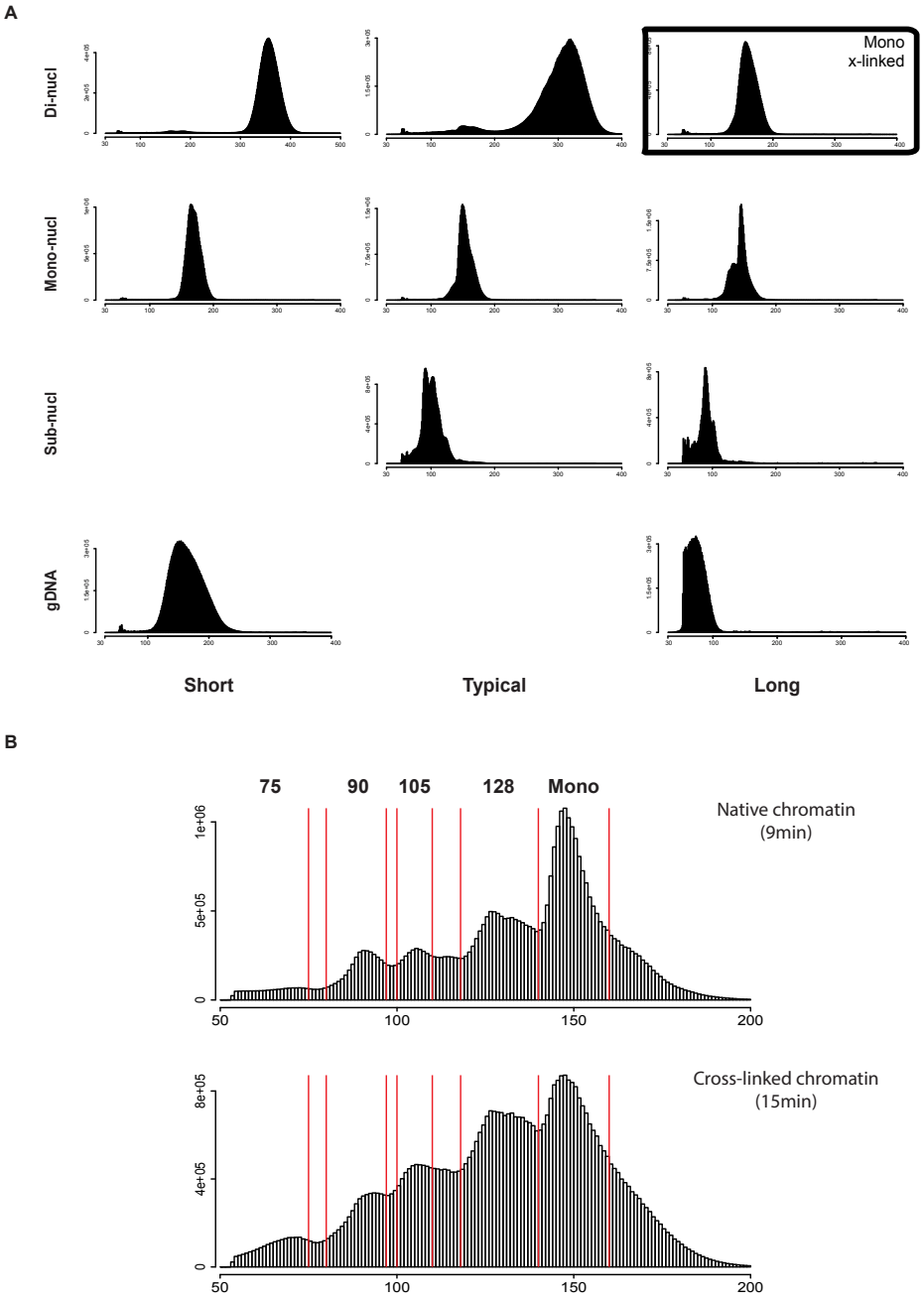

Figure A.2: **Fragment length distributions**: (A) Fragment length distributions of the samples extracted from the electrophoresis gel. (B) Fragment length distributions of the one pot experiments.

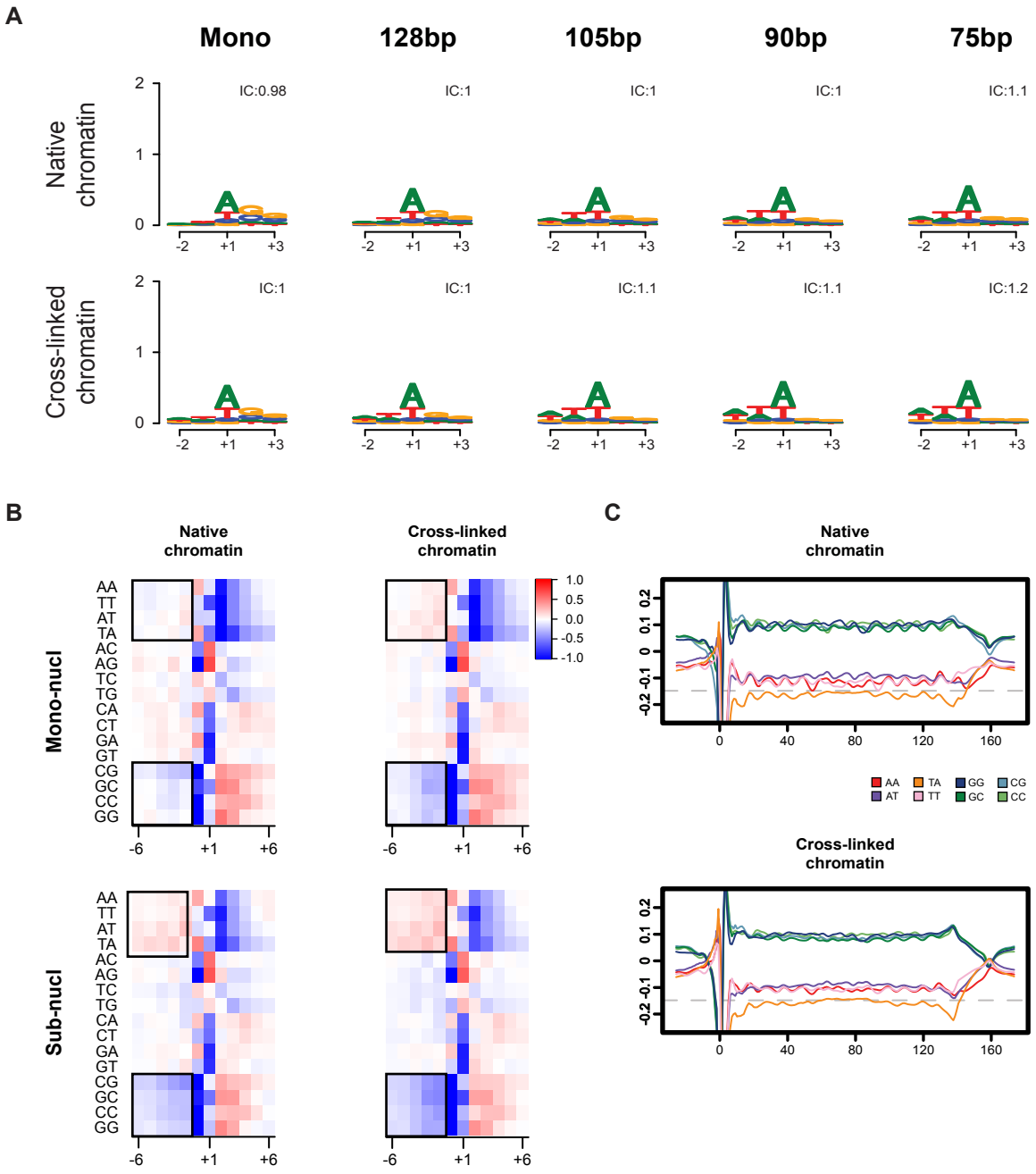

Figure A.3: **Sequence feature comparison between digested native and cross-linked chromatin**: (A,B,C) Subfigures matching those of Figure 8.2, but comparing matched MNase digestion levels of native and cross-linked chromatin.

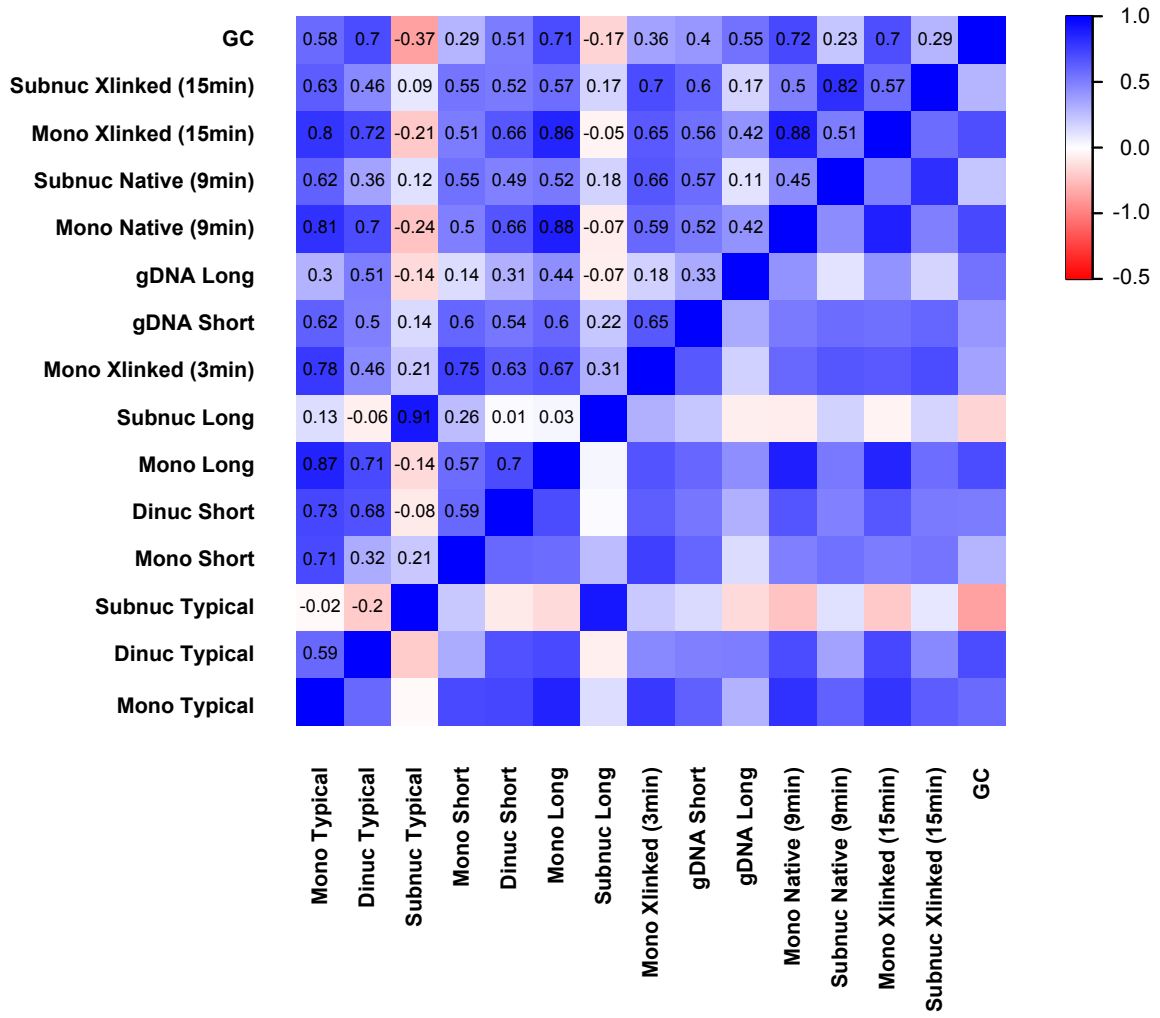

Figure A.4: **Pairwise correlations between the MNase-Seq samples**: Pairwise Pearson's correlation coefficients between the genome-wide nucleosome coverage of the different MNase-Seq samples I analyzed.

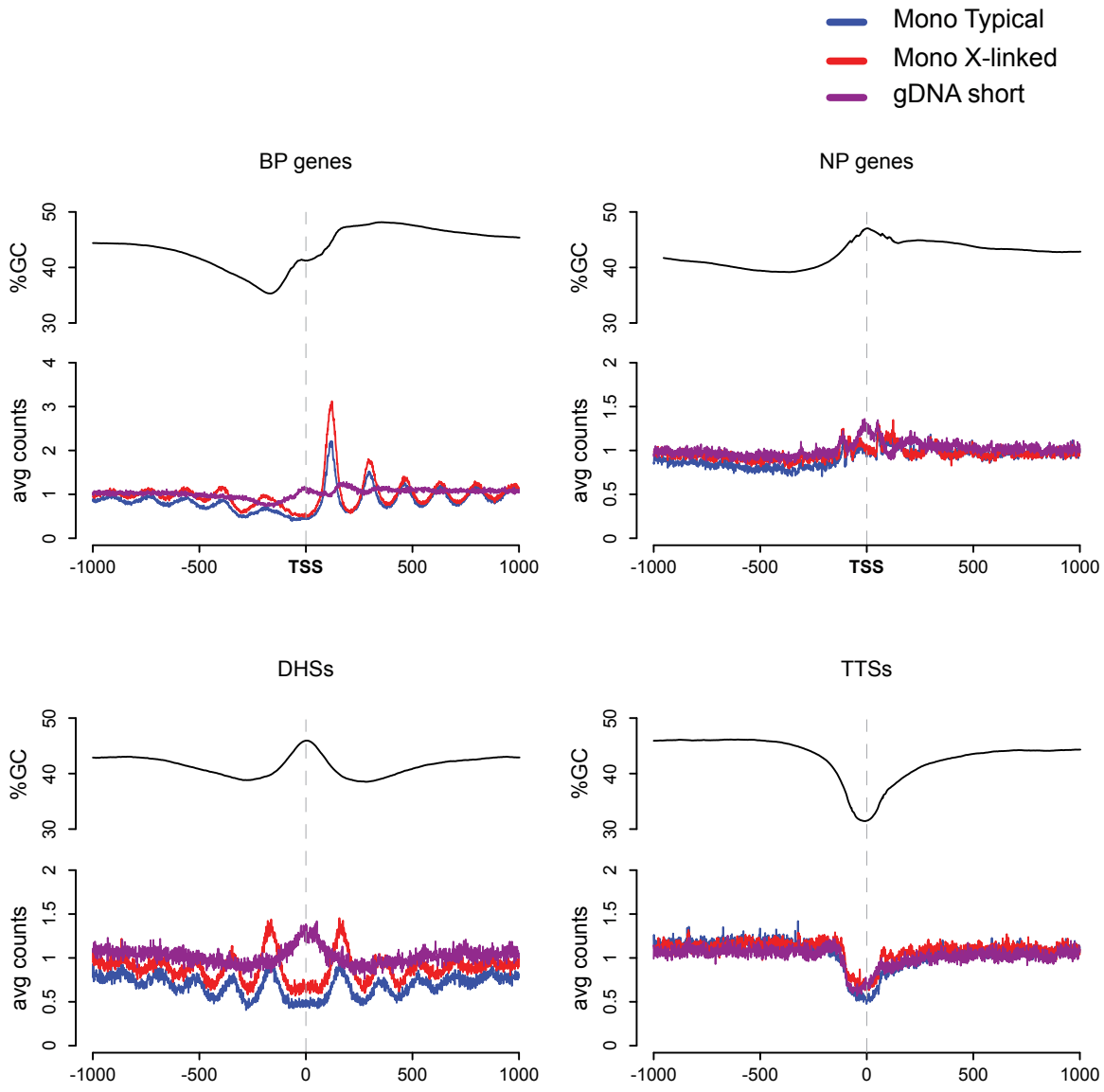

Figure A.5: **Mono-nucleosome coverage profiles around genomic features**: Smoothed profiles around BP promoters (top left), NP promoters (top right), DHSs (bottom left), and TTS (borrom right).

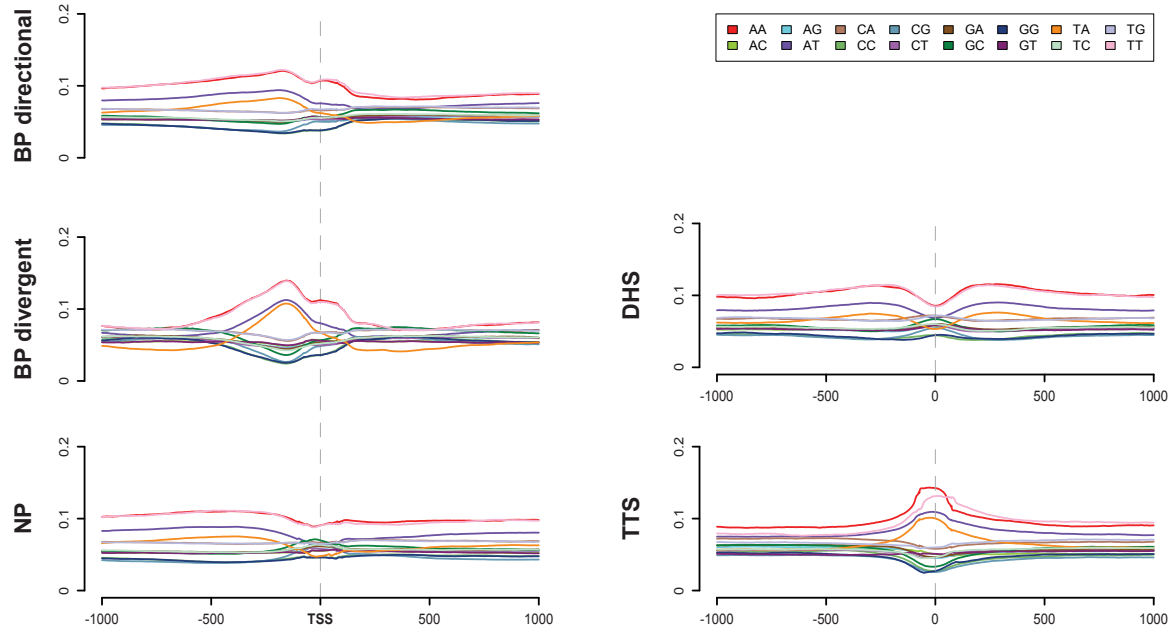

Figure A.6: **Dinucleotide profiles around genomic features**: Smoothed dinucleotide frequency profiles around different genomic features.

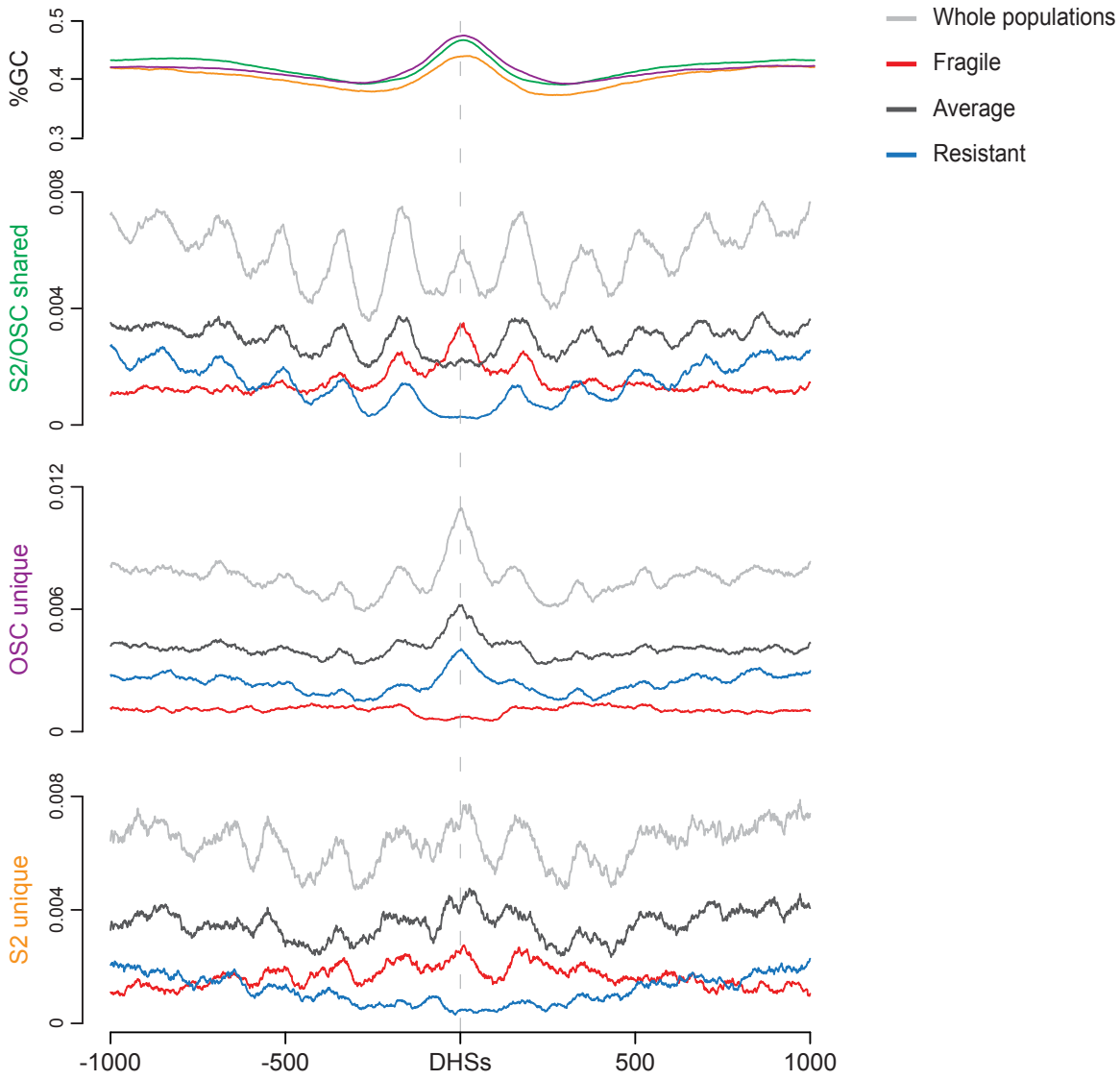

Figure A.7: **Nucleosome populations around cell line un-/specific enhancers**: Smoothed profiles of the nucleosome populations around DHSs sites. The top panel shows the G+C content profiles, followed by the population profiles for DHSs present in: both, only OSC, and only S2 cell lines.

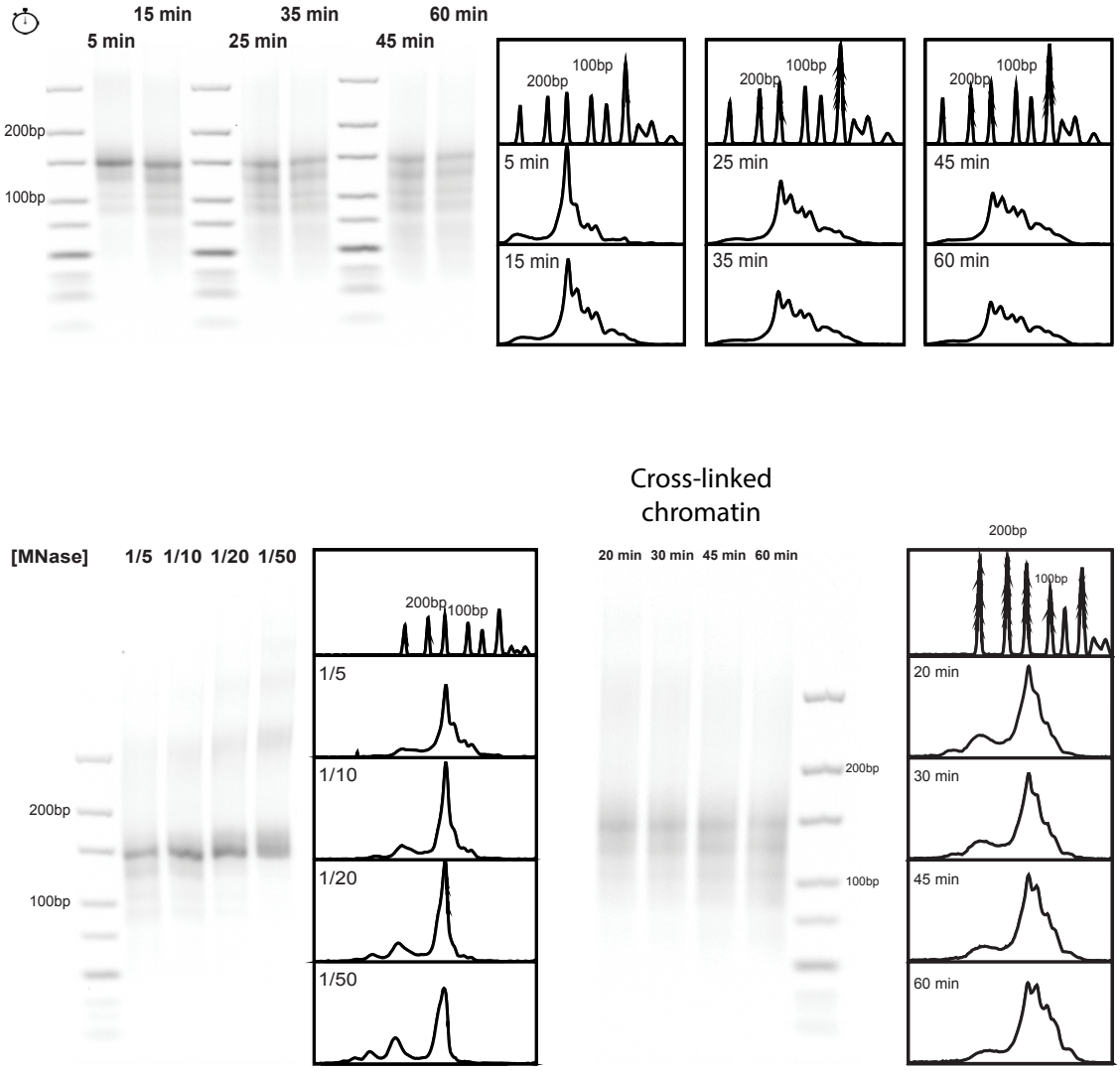

Figure A.8: **Fragment length distributions**: Gel images and ImageJ readouts of the fragment length distributions for a variety of experimental controls and tests.

B. Supplemental figures (Part III)

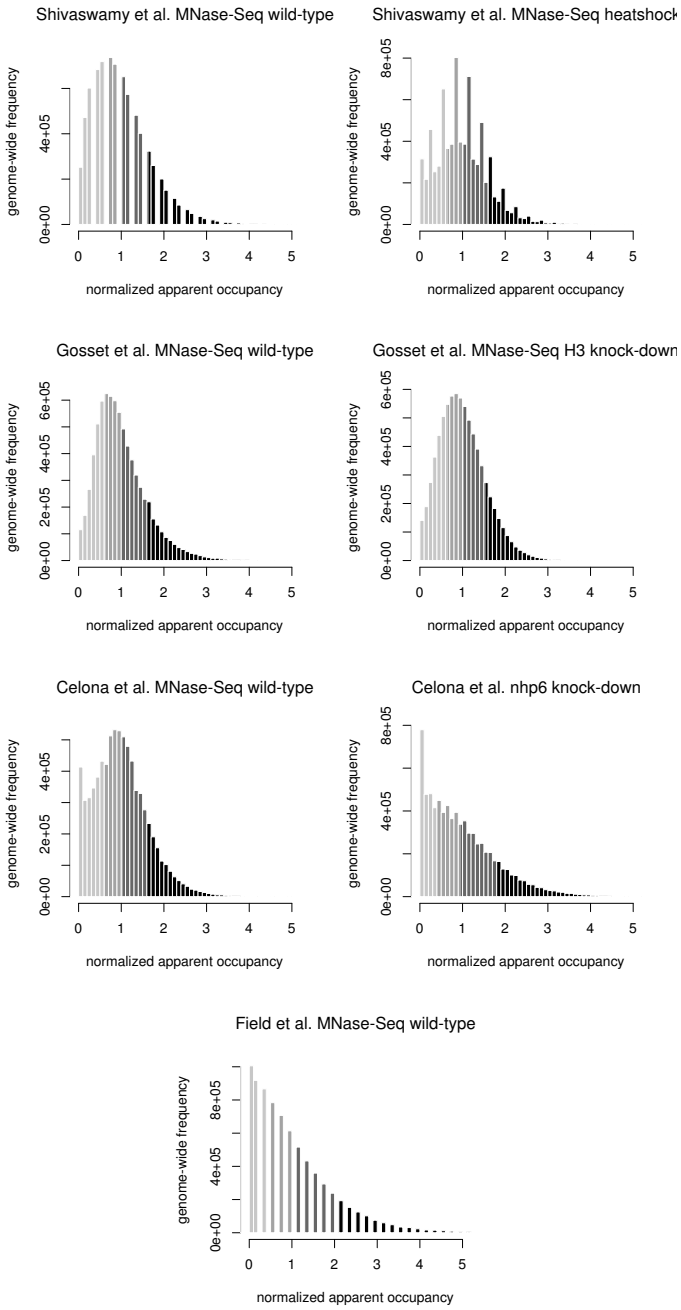

Figure B.1: continued on next page

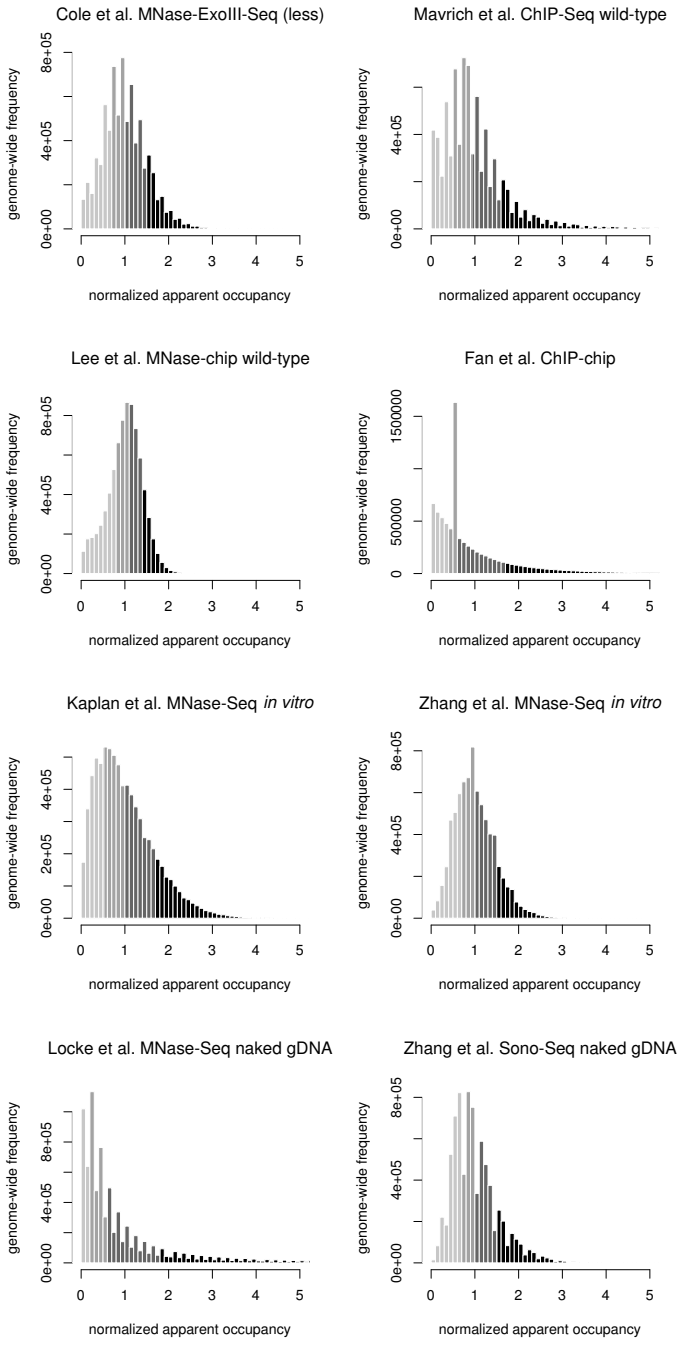

Figure B.1: **Nucleosome occupancies distribution of measurements**: Histograms of nucleosome occupancies deduced from measured fragment centers as show in Figure 13.2A,B for further datasets.

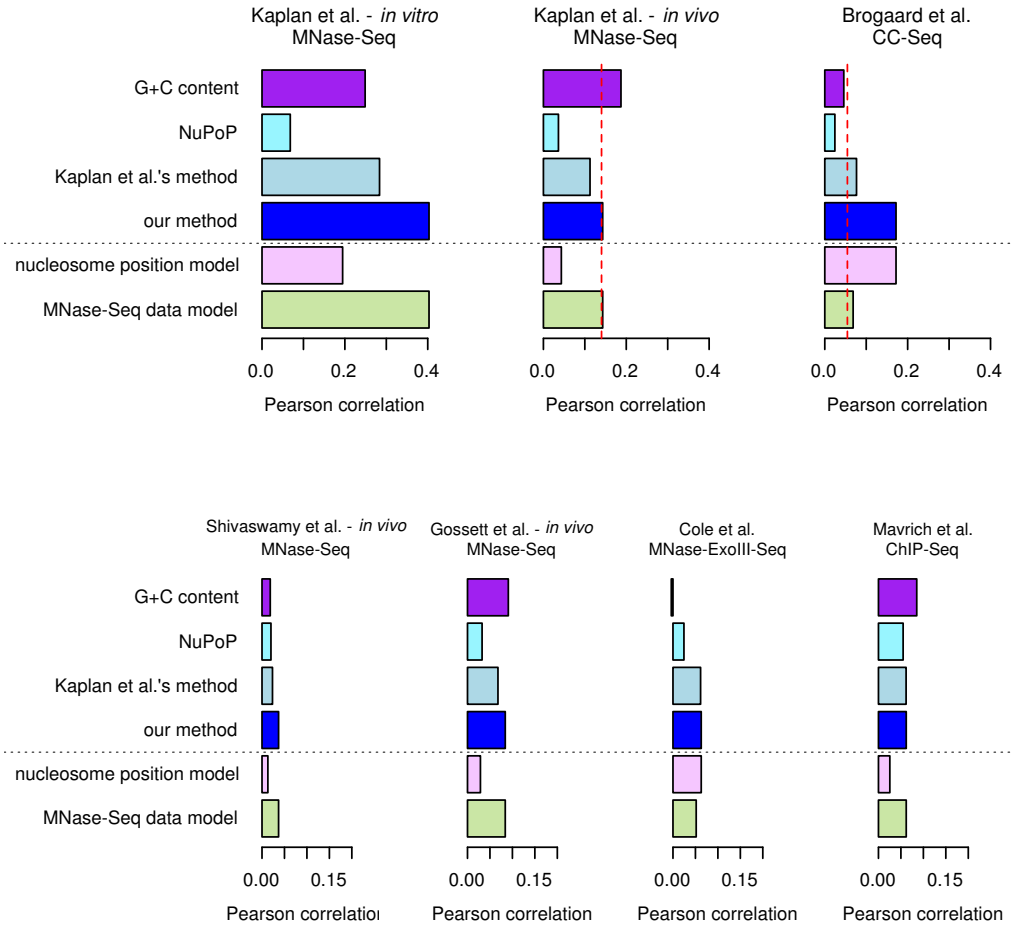

Figure B.2: **Further method performances**

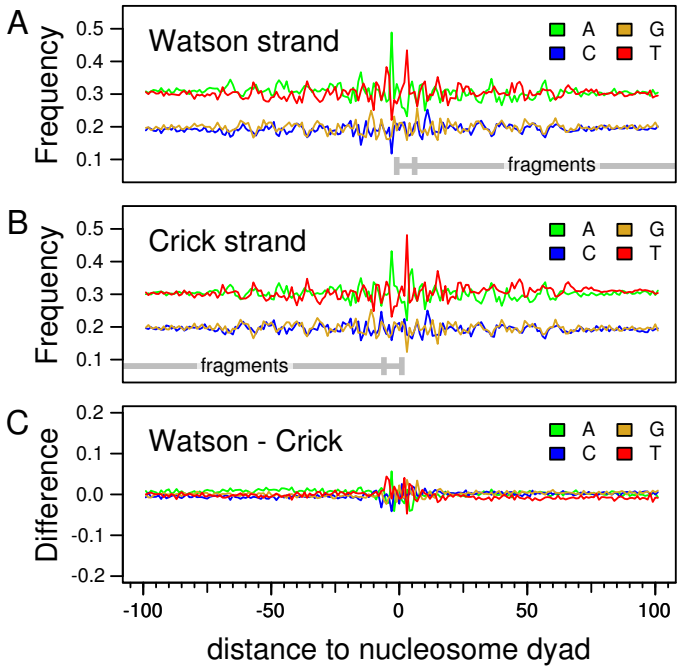

Figure B.3: **CC-Seq's sequence bias with the four template model**: Equivalent to Figure B.3, but using the four template model to deconvolute the data insted of the single template model. The difference in (C) is decreased, because the bias in (A) and (B) becomes more symmetrical.

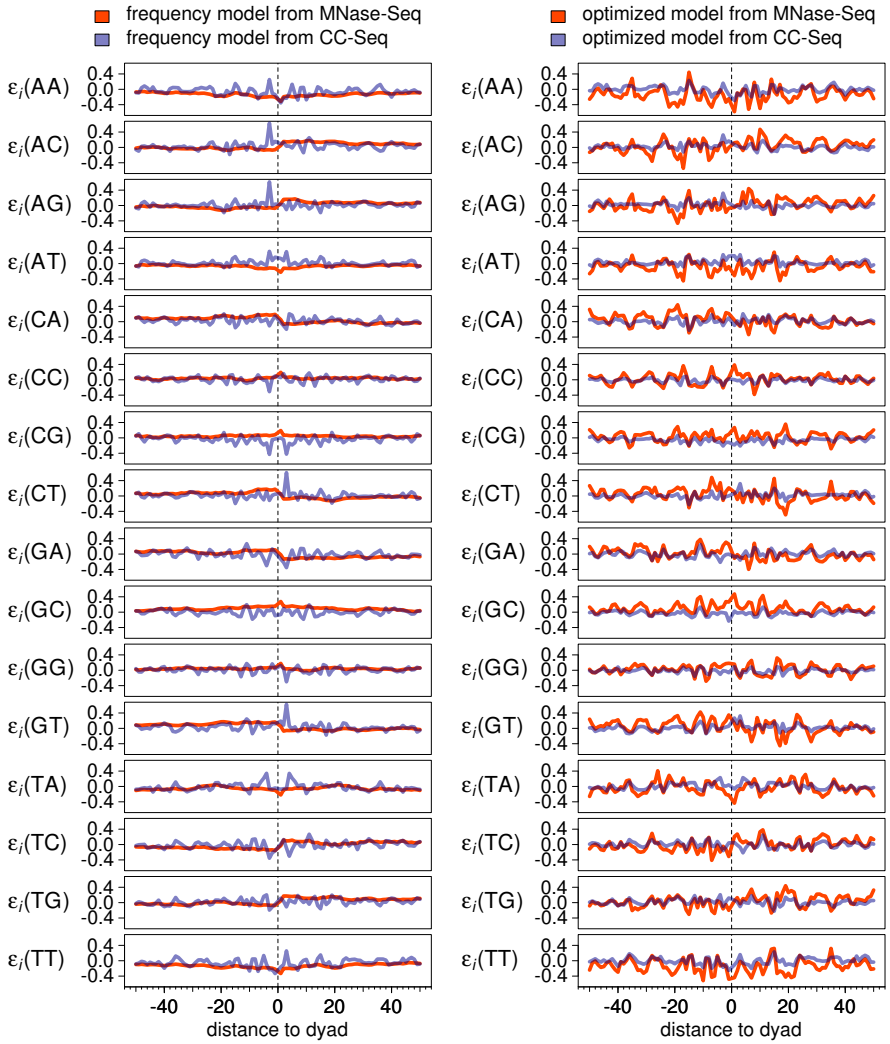

Figure B.4: **Full comparison of my energy models.**: All profiles from which a sample is shown in Figure 13.5A, B.

Table B.1: Overview of the published nucleosome measurements used in this work. For further details of the experimental protocols I refer to the original publications.

publication	condition	method	medium	growth	WT strain	cross-linking	measurement
Lee et al. (2007)	in vivo	MNase-chip	YPD	log-phase	BY4741	yes	Affymetrix tiling microarray
Shivaswamy et al. (2008)	in vivo / heat shock	MNase-Seq	rich medium	?	S288C	yes	Solexa sequencing
Brogaard et al. (2012)	in vivo	CC-Seq	YPD	log-phase	BY4741	no	ABI SOLID sequencing
Celona et al. (2011)	in vivo / nhp6 KD	MNase-Seq	YPD	?	S288C	no	Illumina GA IIx
Kaplan et al. (2009)	in vivo / in vitro	MNase-Seq	YPD	log-phase	YLC8	no	Illumina sequencer
Gossett and Lieb (2012)	in vivo / H3 KD	MNase-Seq	YPD	log-phase	YEF473A	yes	Illumina GA IIx
Zhang et al. (2009)	in vitro	MNase-Seq	-	-	-	no	Illumina GA
Cole et al. (2015)	in vivo	MNase-ExoII-Seq	?	log-phase	YVN381	no	Illumina sequencer
Mavrich et al. (2008)	in vivo	Mnase-ChIP-Seq	rich medium	log phase	BY4741	yes	454 pyrosequencing
Field et al. (2008)	in vivo	MNase-Seq	?	log-phase	?	no	454 pyrosequencing
Fan et al. (2010)	in vivo	ChIP-chip	YPD	log-phase	BY4741	?	Affymetrix tiling microarray
Locke et al. (2010)	naked DNA	MNase-Seq / Sono-Seq	-	-	-	no	Illumina GA

Table B.2: Optimization parameters

Parameter	Denomination	Value
	nr_of_iterations	10000
	fragment_length	25000
	fragment_side_buffer	1000
	numeric_double	0
	numeric_double_long	0
	use_SSE	1
	stopcriterion_deltaLL	-0.01
	stopcriterion_amount	200
	stopcriterion_min_parameter_change	0.00

Table B.3: Learning-rate parameters

Parameters	Denomination	Value
λ	steplength_general	0.10
λ^{ϵ}	steplength_epsilon	1.00
	steplength_epsilon_lower	1.00
λ^{μ}	steplength_mu	15.00
λ^{η}	steplength_eta	10.00
	steplength_decay_frequency	400
Λ	steplength_decay_speed	0.10
$1 - \rho_0$	exponential_average_old_weight	0.00
ϱ	exponential_average_increase_speed	0.50

Table B.4: Model parameters

Parameters	Denomination	Value
$2D_M + 1$	binders_motif_length	100
$2D_N + 1$	binders_footprint	147
μ	initial_mu	4.00
δ	DELTA	20
$k + 1$	mer_length	2

Bibliography

Adler D., Gläser C., Nenadic O., Oehlschlägel J., and Zucchini W. *ff: memory-efficient storage of large data on disk and fast access functions*, **2014**. R package version 2.2-13.

Almeida L.B., Langlois T., Amaral J.D., and Plakhov A. Parameter adaptation in stochastic optimization. In *On-line learning in neural networks*. Cambridge University Press, **1998**; pages 111–134.

Arnold C.D., Gerlach D., Stelzer C., Boryń Ł.M., Rath M., and Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **2013**; *339*(6123):1074–1077. doi:10.1126/science.1232542.

Attrill H., Falls K., Goodman J.L., Millburn G.H., Antonazzo G., Rey A.J., Marygold S.J., and . FlyBase: establishing a Gene Group resource for Drosophila melanogaster. *Nucleic Acids Research* **2015**; *44*(D1):D786. doi:10.1093/nar/gkv1046.

Bordes A., Bottou L., and Gallinari P. SGD-QN: Careful Quasi-Newton Stochastic Gradient Descent. *J. Mach. Learn. Res.* **2009**; *10*:1737–1754. ISSN 1532-4435.

Bottou L. *Stochastic Gradient Descent Tricks*, pages 421–436. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-642-35289-8, **2012**; doi:10.1007/978-3-642-35289-8_25.

Brogaard K., Xi L., Wang J.P., and Widom J. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **2012**; *486*(7404):496–501. doi:10.1038/nature11142.

Brown C.R., Mao C., Falkovskaia E., Jurica M.S., and Boeger H. Linking stochastic fluctuations in chromatin structure and gene expression. *PLoS Biol.* **2013**; *11*(8):e1001621. doi:10.1371/journal.pbio.1001621. Experiment.

Brown J.B., Boley N., Eisman R., May G.E., Stoiber M.H., Duff M.O., Booth B.W., Wen J., Park S., Suzuki A.M., Wan K.H., Yu C., Zhang D., Carlson J.W., Cherbas L., Eads B.D., Miller D., Mockaitis K., Roberts J., Davis C.A., Frise E., Hammonds A.S., Olson S., Shenker S., Sturgill D., Samsonova A.A., Weiszmann R., Robinson G., Hernandez J., Andrews J., Bickel P.J., Carninci P., Cherbas P., Gingeras T.R., Hoskins R.A., Kaufman T.C., Lai E.C., Oliver B., Perrimon N., Graveley B.R., and Celniker S.E. Diversity and dynamics of the Drosophila transcriptome. *Nature* **2014**; *512*(7515):393–399. doi:10.1038/nature12962.

Cairns B.R. The logic of chromatin architecture and remodelling at promoters. *Nature* **2009**; *461*(7261):193–198. doi:10.1038/nature08450.

Celona B., Weiner A., Felice F.D., Mancuso F.M., Cesarini E., Rossi R.L., Gregory L., Baban D., Rossetti G., Grianti P., Pagani M., Bonaldi T., Ragoussis J., Friedman N., Camilloni G., Bianchi M.E., and Agresti A. Substantial histone reduction modulates genomewide nucleosomal occupancy and global transcriptional output. *PLoS Biol.* **2011**; *9*(6):e1001086. doi:10.1371/journal.pbio.1001086. Experiment.

Chen P.B., Zhu L.J., Hainer S.J., McCannell K.N., and Fazzio T.G. Unbiased chromatin accessibility profiling by RED-seq uncovers unique features of nucleosome variants in vivo. *BMC Genomics* **2014**; *15*:1104. doi:10.1186/1471-2164-15-1104. Experiment.

Chen X., Chen Z., Chen H., Su Z., Yang J., Lin F., Shi S., and He X. Nucleosomes Suppress Spontaneous Mutations Base-Specifically in Eukaryotes. *Science* **2012**; *335*(6073):1235–1238. doi:10.1126/science.1217580.

Chereji R.V., Kan T.W., Grudniewska M.K., Romashchenko A.V., Berezikov E., Zhimulev I.F., Guryev V., Morozov A.V., and Moshkin Y.M. Genome-wide profiling of nucleosome sensitivity and chromatin accessibility in Drosophila melanogaster. *Nucleic Acids Research* **2015**; *44*(3):1036–1051. doi:10.1093/nar/gkv978.

Chereji R.V. and Morozov A.V. Statistical Mechanics of Nucleosomes Constrained by Higher-Order Chromatin Structure. *J Stat Phys* **2011**; *144*(2):379–404. doi:10.1007/s10955-011-0214-y.

Chereji R.V. and Morozov A.V. Ubiquitous nucleosome crowding in the yeast genome. *Proc. Natl. Acad. Sci. USA* **2014**; *111*(14):5236–5241. doi:10.1073/pnas.1321001111.

Chung H.R., Dunkel I., Heise F., Linke C., Krobitsch S., Ehrenhofer-Murray A.E., Sperling S.R., and Vingron M. The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS ONE* **2010**; *5*(12):e15754. doi:10.1371/journal.pone.0015754. Experiment.

Cockell M., Rhodes D., and Klug A. Location of the primary sites of micrococcal nuclease cleavage on the nucleosome core. *J. Mol. Biol.* **1983**; *170*(2):423–446.

Cole H.A., Cui F., Ocampo J., Burke T.L., Nikitina T., Nagarajavel V., Kotomura N., Zhurkin V.B., and Clark D.J. Novel nucleosomal particles containing core histones and linker DNA but no histone H1. *Nucleic Acids Res.* **2015**; *44*(2):573–581. doi:10.1093/nar/gkv943. Experiment.

Davey C.A. Does the nucleosome break its own rules? *Current Opinion in Structural Biology* **2013**; *23*(2):311–313. doi:10.1016/j.sbi.2013.01.011. Analysis.

de Jonge E., Wijffels J., and van der Laan J. *ffbase: Basic Statistical Functions for Package 'ff'*, **2015**. R package version 0.12.1.

Deal R.B., Henikoff J.G., and Henikoff S. Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones. *Science* **2010**; *328*(5982):1161–1164. doi:10.1126/science.1186777.

Dechering K. Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Research* **1998**; *26*(17):4056–4062. doi:10.1093/nar/26.17.4056.

Dingwall C., Lomonossoff G.P., and Laskey R.A. High sequence specificity of micrococcal nuclease. *Nucleic Acids Res.* **1981**; *9*(12):2659–2674.

Fan X., Moqtaderi Z., Jin Y., Zhang Y., Liu X.S., and Struhl K. Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3'-end formation. *Proc. Natl. Acad. Sci. USA* **2010**; *107*(42):17945–17950. doi:10.1073/pnas.1012674107.

Fenouil R., Cauchy P., Koch F., Descostes N., Cabeza J.Z., Innocenti C., Ferrier P., Spicuglia S., Gut M., Gut I., and Andrau J.C. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Research* **2012**; *22*(12):2399–2408. doi:10.1101/gr.138776.112.

Field Y., Kaplan N., Fondufe-Mittendorf Y., Moore I.K., Sharon E., Lubling Y., Widom J., and Segal E. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.* **2008**; *4*(11):e1000216. doi:10.1371/journal.pcbi.1000216.

Flaus A., Luger K., Tan S., and Richmond T.J. Mapping nucleosome position at single base-pair resolution by using site-directed hydroxyl radicals. *Proceedings of the National Academy of Sciences* **1996**; *93*(4):1370–1375.

Flores O. and Orozco M. NucleR: A package for non-parametric nucleosome positioning. *Bioinformatics* **2011**; *27*(15):2149–2150.

Gaffney D.J., McVicker G., Pai A.A., Fondufe-Mittendorf Y.N., Lewellen N., Michelini K., Widom J., Gilad Y., and Pritchard J.K. Controls of nucleosome positioning in the human genome. *PLoS Genet.* **2012**; *8*(11):e1003036. doi:10.1371/journal.pgen.1003036.

Gargiulo G., Levy S., Bucci G., Romanenghi M., Fornasari L., Beeson K.Y., Goldberg S.M., Cesaroni M., Ballarini M., Santoro F., Bezman N., Frigè G., Gregory P.D., Holmes M.C., Strausberg R.L., Pelicci P.G., Urnov F.D., and Minucci S. NA-Seq: A discovery tool for the analysis of chromatin structure and dynamics during differentiation. *Dev. Cell* **2009**; *16*(3):466–481. doi:10.1016/j.devcel.2009.02.002. Experiment.

Gossett A.J. and Lieb J.D. In vivo effects of histone H3 depletion on nucleosome occupancy and position in Saccharomyces cerevisiae. *PLoS Genet.* **2012**; *8*(6):e1002771. doi:10.1371/journal.pgen.1002771. Experiment.

Guo S.H., Deng E.Z., Xu L.Q., Ding H., Lin H., Chen W., and Chou K.C. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics* **2014**; *30*(11):1522. doi: 10.1093/bioinformatics/btu083.

Hall M.A., Shundrovsky A., Bai L., Fulbright R.M., Lis J.T., and Wang M.D. High-resolution dynamic mapping of histone-DNA interactions in a nucleosome. *Nat. Struct. Mol. Biol.* **2009**; *16*(2):124–129. doi:10.1038/nsmb.1526.

Harismendy O., Ng P.C., Strausberg R.L., Wang X., Stockwell T.B., Beeson K.Y., Schork N.J., Murray S.S., Topol E.J., Levy S., and Frazer K.A. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* **2009**; *10*(3):R32. doi:10.1186/gb-2009-10-3-r32.

Henikoff J.G., Belsky J.A., Krassovsky K., MacAlpine D.M., and Henikoff S. Epigenome characterization at single base-pair resolution. *Proc. Natl. Acad. Sci. USA* **2011**; *108*(45):18318–18323. doi:10.1073/pnas.1110731108.

Hughes A.L., Jin Y., Rando O.J., and Struhl K. A functional evolutionary approach to identify determinants of nucleosome positioning: A unifying model for establishing the genome-wide pattern. *Mol. Cell* **2012**; *48*(1):5–15. doi:10.1016/j.molcel.2012.07.003.

Hörz W. and Altenburger W. Sequence specific cleavage of DNA by micrococcal nuclease. *Nucleic Acids Research* **1981**; *9*(12):2643–2658. doi:10.1093/nar/9.12.2643.

Ioshikhes I.P., Albert I., Zanton S.J., and Pugh B.F. Nucleosome positions predicted through comparative genomics. *Nat. Genet.* **2006**; *38*(10):1210–1215. doi:10.1038/ng1878.

Ishii H., Kadonaga J.T., and Ren B. MPE-seq, a new method for the genome-wide analysis of chromatin structure. *Proc. Natl. Acad. Sci. USA* **2015**; *112*(27):E3457–65. doi:10.1073/pnas.1424804112. Experimental.

Iyer V.R. Nucleosome positioning: Bringing order to the eukaryotic genome. *Trends Cell Biol.* **2012**; *22*(5):250–256. doi:10.1016/j.tcb.2012.02.004.

Jessen W.J., Hoose S.A., Kilgore J.A., and Kladde M.P. Active PHO5 chromatin encompasses variable numbers of nucleosomes at individual promoters. *Nat. Struct. Mol. Biol.* **2006**; *13*(3):256–263. doi:10.1038/nsmb1062.

Jette M.A., Yoo A.B., and Grondona M. SLURM: Simple Linux Utility for Resource Management. In *In Lecture Notes in Computer Science: Proceedings of Job Scheduling Strategies for Parallel Processing (JSSPP) 2003*. Springer-Verlag, **2002**; pages 44–60.

Jiang C. and Pugh B.F. A compiled and systematic reference map of nucleosome positions across the Saccharomyces cerevisiae genomes. *Genome Biology* **2009**; *10*(10):R109. ISSN 1474-760X. doi:10.1186/gb-2009-10-10-r109.

Kaplan N., Hughes T.R., Lieb J.D., Widom J., and Segal E. Contribution of histone sequence preferences to nucleosome organization: proposed definitions and methodology. *Genome Biol* **2010**a; *11*(11):140. doi:10.1186/gb-2010-11-11-140.

Kaplan N., Moore I., Fondufe-Mittendorf Y., Gossett A.J., Tillo D., Field Y., Hughes T.R., Lieb J.D., Widom J., and Segal E. Nucleosome sequence preferences influence in vivo nucleosome organization. *Nat Struct Mol Biol* **2010**b; *17*(8):918–920. doi:10.1038/nsmb0810-918.

Kaplan N., Moore I.K., Fondufe-Mittendorf Y., Gossett A.J., Tillo D., Field Y., LeProust E.M., Hughes T.R., Lieb J.D., Widom J., and Segal E. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **2009**; *458*(7236):362–366. doi:10.1038/nature07667. Nucleosomes - main.

Kelly T.K., Liu Y., Lay F.D., Liang G., Berman B.P., and Jones P.A. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* **2012**; *22*(12):2497–2506. doi:10.1101/gr.143008.112. Experiment.

Knight S. Building software with SCons. *Computing in Science & Engineering* **2005**; *7*(1):79–88.

Kubik S., Bruzzone M.J., Jacquet P., Falcone J.L., Rougemont J., and Shore D. Nucleosome Stability Distinguishes Two Different Promoter Types at All Protein-Coding Genes in Yeast. *Mol. Cell* **2015**; *60*(3):422–434. doi:10.1016/j.molcel.2015.10.002.

Lam F.H., Steger D.J., and O'Shea E.K. Chromatin decouples promoter threshold from dynamic range. *Nature* **2008**; *453*(7192):246–250. doi:10.1038/nature06867.

Langmead B. and Salzberg S.L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **2012**; *9*(4):357–359. doi:10.1038/nmeth.1923.

Lawrence M., Gentleman R., and Carey V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **2009**; *25*:1841–1842. doi:10.1093/bioinformatics/btp328.

Lawrence M., Huber W., Pagès H., Aboyoun P., Carlson M., Gentleman R., Morgan M., and Carey V. Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology* **2013**; *9*. doi:10.1371/journal.pcbi.1003118.

Lee W., Tillo D., Bray N., Morse R.H., Davis R.W., Hughes T.R., and Nislow C. A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.* **2007**; *39*(10):1235–1244. doi:10.1038/ng2117. Experiment.

Leek J.T., Scharpf R.B., Bravo H.C., Simcha D., Langmead B., Johnson W.E., Geman D., Baggerly K., and Irizarry R.A. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **2010**; *11*(10):733–739. doi:10.1038/nrg2825.

Levo M., Zalckvar E., Sharon E., Machado A.C.D., Kalma Y., Lotam-Pompan M., Weinberger A., Yakhini Z., Rohs R., and Segal E. Unraveling determinants of transcription factor binding outside the core binding site. *Genome Research* **2015**; *25*(7):1018–1029. doi:10.1101/gr.185033.114.

Li G., Levitus M., Bustamante C., and Widom J. Rapid spontaneous accessibility of nucleosomal DNA. *Nat. Struct. Mol. Biol.* **2005**; *12*(1):46–53. doi:10.1038/nsmb869.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., and and R.D. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**; *25*(16):2078–2079. doi:10.1093/bioinformatics/btp352.

Liu H., Zhang R., Xiong W., Guan J., Zhuang Z., and Zhou S. A comparative evaluation on prediction methods of nucleosome positioning. *Brief. Bioinformatics* **2014**; *15*(6):1014–1027. Prediction.

Locke G., Tolkunov D., Moqtaderi Z., Struhl K., and Morozov A.V. High-throughput sequencing reveals a simple model of nucleosome energetics. *Proceedings of the National Academy of Sciences* **2010**; *107*(49):20998 – 21003. doi:10.1073/pnas.1003838107.

Lubliner S. and Segal E. Modeling interactions between adjacent nucleosomes improves genome-wide predictions of nucleosome occupancy. *Bioinformatics* **2009**; *25*(12):i348–55.

Luger K., Mäder A.W., Richmond R.K., Sargent D.F., and Richmond T.J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **1997**; *389*(6648):251–260.

Mavrich T.N., Ioshikhes I.P., Venters B.J., Jiang C., Tomsho L.P., Qi J., Schuster S.C., Albert I., and Pugh B.F. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* **2008**; *18*(7):1073–1083. doi:10.1101/gr.078261.108.

McManus J., Perry P., Sumner A., Wright D., Thomson E., Allshire R., Hastie N., and Bickmore W. Unusual chromosome structure of fission yeast DNA in mouse cells. *Journal of Cell Science* **1994**; *107*(3):469–486. ISSN 0021-9533.

Mieczkowski J., Cook A., Bowman S.K., Mueller B., Alver B.H., Kundu S., Deaton A.M., Urban J.A., Larschan E., Park P.J., Kingston R.E., and Tolstorukov M.Y. MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nature Communications* **2016**; *7*:11485. doi:10.1038/ncomms11485.

Minary P. and Levitt M. Training-free atomistic prediction of nucleosome occupancy. *Proc. Natl. Acad. Sci. USA* **2014**; *111*(17):6293–6298. doi:10.1073/pnas.1404475111.

Moser C. and Gupta M. A generalized hidden Markov model for determining sequence-based predictors of nucleosome positioning. *Statistical applications in genetics and molecular biology* **2012**; *11*(2). ISSN 2194-6302. doi:10.2202/1544-6115.1707.

Moyle-Heyrman G., Zaichuk T., Xi L., Zhang Q., Uhlenbeck O.C., Holmgren R., Widom J., and Wang J.P. Chemical map of Schizosaccharomyces pombe reveals species-specific features in nucleosome positioning. *Proc. Natl. Acad. Sci. USA* **2013**; *110*(50):20158–20163. doi:10.1073/pnas.1315809110.

Möbius W., Osberg B., Tsankov A.M., Rando O.J., and Gerland U. Toward a unified physical model of nucleosome patterns flanking transcription start sites. *Proc. Natl. Acad. Sci. USA* **2013**; *110*(14):5719–5724. doi:10.1073/pnas.1214048110.

Ngo T.T.M., Zhang Q., Zhou R., Yodh J.G., and Ha T. Asymmetric unwrapping of nucleosomes under tension directed by DNA local flexibility. *Cell* **2015**; *160*(6):1135–1144. doi:10.1016/j.cell.2015.02.001.

OpenMP Architecture Review Board. OpenMP Application Program Interface Version 3.0. **2008**.

Ozonov E.A. and van Nimwegen E. Nucleosome free regions in yeast promoters result from competitive binding of transcription factors that interact with chromatin modifiers. *PLoS Comput. Biol.* **2013**; *9*(8):e1003181. doi:10.1371/journal.pcbi.1003181.

Pages H., Aboyoun P., Gentleman R., and DebRoy S. *Biostrings: String objects representing biological sequences, and matching algorithms*, **2016**. R package version 2.32.1.

Parmar J.J., Marko J.F., and Padinhateeri R. Nucleosome positioning and kinetics near transcription-start-site barriers are controlled by interplay between active remodeling and DNA sequence. *Nucleic Acids Research* **2013**; *42*(1):128–136. doi:10.1093/nar/gkt854.

Peckham H.E., Thurman R.E., Fu Y., Stamatoyannopoulos J.A., Noble W.S., Struhl K., and Weng Z. Nucleosome positioning signals in genomic DNA. *Genome Res.* **2007**; *17*(8):1170–1177. doi:10.1101/gr.6101007.

Qian N. On the momentum term in gradient descent learning algorithms. *Neural Networks* **1999**; *12*(1):145 – 151. ISSN 0893-6080. doi:10.1016/S0893-6080(98)00116-6.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, **2016**.

Rach E.A., Winter D.R., Benjamin A.M., Corcoran D.L., Ni T., Zhu J., and Ohler U. Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genet.* **2011**; *7*(1):e1001274. doi:10.1371/journal.pgen.1001274.

Rach E.A., Yuan H.Y., Majoros W.H., Tomancak P., and Ohler U. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. *Genome Biology* **2009**; *10*(7):R73. ISSN 1474-760X. doi:10.1186/gb-2009-10-7-r73.

Rakhlin A., Shamir O., and Sridharan K. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647* **2011**; .

Raveh-Sadka T., Levo M., Shabi U., Shany B., Keren L., Lotan-Pompan M., Zeevi D., Sharon E., Weinberger A., and Segal E. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Genet.* **2012**; *44*(7):743–750. doi:10.1038/ng.2305. Experiment.

Reynolds S.M., Bilmes J.A., and Noble W.S. Learning a weighted sequence model of the nucleosome core and linker yields more accurate predictions in Saccharomyces cerevisiae and Homo sapiens. *PLoS Comput. Biol.* **2010**; *6*(7):e1000834. doi:10.1371/journal.pcbi.1000834.

Rhee H.S., Bataille A.R., Zhang L., and Pugh B.F. Subnucleosomal structures and nucleosome asymmetry across a genome. *Cell* **2014**; *159*(6):1377–1388. doi:10.1016/j.cell.2014.10.054.

Rhee H.S. and Pugh B.F. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell* **2011**; *147*(6):1408–1419. doi:10.1016/j.cell.2011.11.013. Not printed.

Richmond T.J. and Davey C.A. The structure of DNA in the nucleosome core. *Nature* **2003**; *423*(6936):145–150. doi:10.1038/nature01595.

Rizzo J.M., Bard J.E., and Buck M.J. Standardized collection of MNase-seq experiments enables unbiased dataset comparisons. *BMC Mol Biol* **2012**; *13*(1):15. doi:10.1186/1471-2199-13-15. Analysis.

Schaul T. and LeCun Y. Adaptive learning rates and parallelization for stochastic, sparse, non-smooth gradients. *arXiv preprint arXiv:1301.3764* **2013**; .

Schep A.N., Buenrostro J.D., Denny S.K., Schwartz K., Sherlock G., and Greenleaf W.J. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res.* **2015**; *25*(11):1757–1770. doi:10.1101/gr.192294.115. Experiment.

Schwalb B., Tresch A., Torkler P., Duemcke S., and Demel C. *LSD: Lots of Superior Depictions*, **2015**. R package version 3.0.

Scipioni A. and Santis P.D. Predicting nucleosome positioning in genomes: Physical and bioinformatic approaches. *Biophys. Chem.* **2011**; *155*(2 - 3):53–64. doi:10.1016/j.bpc.2011.03.006.

Segal E., Fondufe-Mittendorf Y., Chen L., Thåström A., Field Y., Moore I.K., Wang J.P.Z., and Widom J. A genomic code for nucleosome positioning. *Nature* **2006**; *442*(7104):772–778. doi:10.1038/nature04979. Nucleosomes - main.

Segal E. and Widom J. What controls nucleosome positions? *Trends in Genetics* **2009**; *25*(8):335–343. doi:10.1016/j.tig.2009.06.002.

Shivaswamy S., Bhinge A., Zhao Y., Jones S., Hirst M., and Iyer V.R. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.* **2008**; *6*(3):e65. doi:10.1371/journal.pbio.0060065.

Siebert M. and Söding J. Universality of core promoter elements? *Nature* **2014**; *511*(7510):E11–E12. doi:10.1038/nature13587.

Siebert M. and Söding J. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Research* **2016**; doi:10.1093/nar/gkw521.

Small E.C., Xi L., Wang J.P., Widom J., and Licht J.D. Single-cell nucleosome mapping reveals the molecular basis of gene expression heterogeneity. *Proc. Natl. Acad. Sci. USA* **2014**; *111*(24):E2462–71. doi:10.1073/pnas.1400517111. Experiment.

Stein A., Takasuka T.E., and Collings C.K. Are nucleosome positions in vivo primarily determined by histone-DNA sequence preferences? *Nucleic Acids Res.* **2010**; *38*(3):709–719. doi:10.1093/nar/gkp1043.

Stormo G.D. Modeling the specificity of protein-DNA interactions. *Quantitative Biology* **2013**; *1*(2):115–130. ISSN 2095-4697. doi:10.1007/s40484-013-0012-4.

Struhl K. and Segal E. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.* **2013**; *20*(3):267–273. doi:10.1038/nsmb.2506. Review.

Teif V.B. Nucleosome positioning: Resources and tools online. *Brief. Bioinformatics* **2015**; Review.

Teif V.B. and Rippe K. Predicting nucleosome positions on the DNA: Combining intrinsic sequence preferences and remodeler activities. *Nucleic Acids Res.* **2009**; *37*(17):5641–5655.

Teves S.S., Weber C.M., and Henikoff S. Transcribing through the nucleosome. *Trends in Biochemical Sciences* **2014**; *39*(12):577 – 586. ISSN 0968-0004. doi:10.1016/j.tibs.2014.10.004.

Thåström A., Bingham L., and Widom J. Nucleosomal Locations of Dominant DNA Sequence Motifs for Histone–DNA Interactions and Nucleosome Positioning. *Journal of Molecular Biology* **2004**; *338*(4):695 – 709. ISSN 0022-2836. doi:10.1016/j.jmb.2004.03.032.

Thåström A., Lowary P., Widlund H., Cao H., Kubista M., and Widom J. Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *Journal of Molecular Biology* **1999**; *288*(2):213 – 229. ISSN 0022-2836. doi:10.1006/jmbi.1999.2686.

Tillo D. and Hughes T.R. G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics* **2009**; *10*:442. doi:10.1186/1471-2105-10-442. Prediction.

Tolstorukov M.Y., Choudhary V., Olson W.K., Zhurkin V.B., and Park P.J. NuScore: A web-interface for nucleosome positioning predictions. *Bioinformatics* **2008**; *24*(12):1456–1458. doi:10.1093/bioinformatics.

van der Heijden T., van Vugt J.J.F.A., Logie C., and van Noort J. Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy. *Proc. Natl. Acad. Sci. USA* **2012**; *109*(38):E2514–22. doi:10.1073/pnas.1205659109.

van Holde K.E. *Chromatin*. Springer New York, **1989**. doi:10.1007/978-1-4612-3490-6.

Venters B.J. and Pugh B.F. Genomic organization of human transcription initiation complexes. *Nature* **2013**; *502*(7469):53–58. doi:10.1038/nature12535.

Vierstra J., Wang H., John S., Sandstrom R., and Stamatoyannopoulos J.A. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nat. Methods* **2014**; *11*(1):66–72. doi:10.1038/nmeth.2713. Experiment.

Wasson T. and Hartemink A.J. An ensemble model of competitive multi-factor binding of the genome. *Genome Res.* **2009**; *19*(11):2101–2112. doi:10.1101/gr.093450.109.

Wei T. *corrplot: Visualization of a correlation matrix*, **2013**. R package version 0.73.

Weiner A., Hughes A., Yassour M., Rando O.J., and Friedman N. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res.* **2010**; *20*(1):90–100. doi:10.1101/gr.098509.109. Experiment.

Wilson D. and Martinez T.R. The general inefficiency of batch training for gradient descent learning. *Neural Networks* **2003**; *16*(10):1429 – 1451. ISSN 0893-6080. doi:10.1016/S0893-6080(03)00138-2.

Xi L., Brogaard K., Zhang Q., Lindsay B., Widom J., and Wang J.P. A locally convoluted cluster model for nucleosome positioning signals in chemical map. *J Am Stat Assoc* **2014**; *109*(505):48–62. doi:10.1080/01621459.2013.862169.

Xi L., Fondufe-Mittendorf Y., Xia L., Flatow J., Widom J., and Wang J.P. Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC Bioinformatics* **2010**; *11*:346. doi:10.1186/1471-2105-11-346.

Xi Y., Yao J., Chen R., Li W., and He X. Nucleosome fragility reveals novel functional states of chromatin and poises genes for activation. *Genome Res.* **2011**; *21*(5):718–724. doi:10.1101/gr.117101.110.

Xing K. and He X. Mutation bias, rather than binding preference, underlies the nucleosome-associated G+C% variation in eukaryotes. *Genome Biol Evol* **2015**; *7*(4):1033–1038. doi:10.1093/gbe/evv053. Analysis.

Zentner G.E. and Henikoff S. Surveying the epigenomic landscape, one base at a time. *Genome Biol.* **2012**; *13*(10):250. doi:10.1186/gb4051. Review.

Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., Bernstein B.E., Nussbaum C., Myers R.M., Brown M., Li W., and Liu X.S. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **2008**; *9*(9):R137. doi:10.1186/gb-2008-9-9-r137.

Zhang Y., Moqtaderi Z., Rattner B.P., Euskirchen G., Snyder M., Kadonaga J.T., Liu X.S., and Struhl K. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nat. Struct. Mol. Biol.* **2009**; *16*(8):847–852. doi:10.1038/nsmb.1636.

Zhang Y., Moqtaderi Z., Rattner B.P., Euskirchen G., Snyder M., Kadonaga J.T., Liu X.S., and Struhl K. Reply to "Evidence against a genomic code for nucleosome positioning". *Nat Struct Mol Biol* **2010**; *17*(8):920–923. doi:10.1038/nsmb0810-920.

Zhang Z., Wippo C.J., Wal M., Ward E., Korber P., and Pugh B.F. A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science* **2011**; *332*(6032):977–980. doi:10.1126/science.1200508.