Understanding the role of nucleosome positioning in gene regulation by leveraging deep learning models

A Thesis

submitted to

Indian Institute of Science Education and Research Pune in partial fulfilment of the requirements for the BS-MS Dual Degree Programme

by

Grishma Mehta

Indian Institute of Science Education and Research Pune

Dr. Homi Bhabha Road,

Pashan, Pune 411008, INDIA.

April, 2025

Under the guidance of,

Supervisor : Dr. Julia Zeitlinger

Investigator

Stowers Institute for Medical Research

From June 2024 to March 2025

Certificate

This is to certify that this dissertation entitled **'Understanding the role of nucleosome positioning in gene regulation by leveraging deep learning models'** towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Grishma Mehta at Stowers Institute for Medical Research, USA under the supervision of **Dr. Julia Zeitlinger,** Investigator, during the academic year 2024-2025.

Julia Zeitlinger

Dr. Julia Zeitlinger

Investigator

Stowers Institute for Medical Research

Dr. Krishanpal Karmodiya

Associate Professor

Indian Institute of Science Education and Research, Pune

This thesis is dedicated to my mother….

Declaration

I hereby declare that the matter embodied in the report entitled **“Understanding the role of nucleosome positioning in gene regulation by leveraging deep learning models”** are the results of the work carried out by me at the Stowers Institute for Medical Research, USA, under the supervision of **Dr. Julia Zeitlinger**, Investigator, and the same has not been submitted elsewhere for any other degree. Wherever others contribute, every effort is made to indicate this clearly, with due reference to the literature and acknowledgement of collaborative research and discussions

Grishma Mehta

20201056

Table of contents

Abbreviations

NDR	Nucleosome Depleted Region
GRF	General Regulatory Factors
MCP	MS2 bacteriophage Coat Protein

List of Tables

List of Figures

Abstract

Chromatin is packed into basic repeating units called nucleosomes, but how exactly nucleosomes influence gene regulation is not clear. *S.cerevisiae* has well-positioned nucleosomes throughout its genome, giving us an opportunity to study what positions them and how this regulates gene expression. Previously, the exact relationship between genomic sequence and nucleosome positioning has been hard to interpret given the complex nature by which nucleosomes are regulated by sequence features and chromatin remodelers. Sequence-to-function deep learning models have recently been used to identify complex non-linear patterns, making this a promising approach for learning sequence rules that position nucleosomes. This project leverages one such sequence-to-function deep learning model, BPReveal, to learn the sequence rules underlying genome-wide MNase-seq data. We show that BPReveal correctly learned important nucleosome-positioning sequences without prior knowledge. Since BPReveal has the ability to accurately predict genome-wide MNase-seq data, this study also shows that BPReveal can be used as a tool to design synthetic sequences such that alter nucleosome positioning at a specific locus in a desired fashion. We validated some of these designs experimentally and started to characterise the effect they have on gene expression by employing MS2-MCP based live imaging to detect single mRNAs across many cells. Overall this work is a proof-of-principle study that deep learning models can be used to better understand how DNA sequences position nucleosomes and thereby influence gene regulation.

Acknowledgements

Julia, I am really grateful to you for giving me the opportunity to be a part of this lab. Her patience and unwavering support helped me in believing myself and keep going despite all the problems I faced during my thesis. Thank you for trying your best to help me deal with different aspects of my project like planning experiments, connecting with different Core facility members and making decisions in time crunch situations.

Charles, thanks for generating amazing deep model predictions which allowed me to learn a lot of cool techniques. It has been a pleasure working with you on this project. Tom, thanks for helping me with all the Microscopy experiments. Performing those experiments was really challenging but you were always ready to try new ways to make the experiments work, so thank you for your support and patience throughout. Jennifer, thank you for making a bulk of the strains I worked on, it allowed me to start my experiments as soon as I joined the lab. Cathy, thank you for generating the pipeline for analysing my imaging data. I was really scared about having to count all the spots on my images manually, your code was a lifesaver.

I also want to thank all the Zeitlinger lab members for always listening to my problems and trying their best to help me. You have always been around and I don't think I will ever be able to forget our lunch time conversations.

I want to thank Stowers Institute for Medical Research for always trying to make sure that everyone can work on their science comfortably. The support from the core facilities enabled me to try experiments which are not routinely done in the lab. I want to specifically thank the Sequencing and Discovery Genomics Core for doing the library preparation of my samples and ensuring that all the samples are processed on time. I also want to thank Alexis and Pooja from Media Prep who made all of the custom media and plates which were used in this project. It would not have been possible for me to perform my experiments without your support.

Contributions

Contributor name	Contributor role
Dr. Julia Zeitlinger, Grishma Mehta, Dr. Charles McAnany	Conceptualization Ideas
Grishma Mehta	Methodology
Dr. Charles McAnany	Software
Grishma Mehta	Validation
Grishma Mehta, Dr. Cathy McKinney	Formal analysis
Grishma Mehta, Dr. Tom Kleist	Investigation
Zeitlinger Lab, Dr.Jennifer Gardner	Resources
Zeitlinger Lab	Data Curation
Grishma Mehta	Writing - original draft preparation
Grishma Mehta, Dr. Charles McAnany, Dr. Julia Zeitlinger	Writing - review and editing
Grishma Mehta	Visualization
Dr. Julia Zeitlinger	Supervision
Dr. Julia Zeitlinger	Project administration
Stowers Institute for Medical Research	Funding acquisition

Chapter 1
Introduction

1.1 Nucleosome positioning and occupancy in *S.cerevisiae*

In eukaryotes, the evolution of the nucleus generated a need for developing packaging strategies in order to accommodate the enormous amount of linear DNA into tiny volumes of a nucleus. This is not just a simple packaging problem as certain parts of the DNA need to be accessible for maintaining cellular processes like replication, gene regulation and transcription. In order to engineer this, cells coil negatively charged DNA around basic histone proteins to form a DNA-protein complex called the nucleosome (Kornberg., 1974, Jansen *et al.,* 2011; Oudet *et al.,* 1975). A nucleosome is formed when 147 bp of DNA is wrapped around a histone octamer consisting of two copies each of H2A, H2B, H3 and H4 histones (*Luger et al.,* 1997). The DNA between two nucleosomes is called the linker region and the length of this linker is variable across cell types. This “beads on a string model” is present throughout the genome and seemed like a non specific coating in the beginning but further research indicated that nucleosomes positioning is not uniform across the genome, we see some patterns in nucleosome positioning which play a key role in regulating gene expression (Thoma *et al.,* 1979).

The structure of a nucleosome gives rise to differential accessibility of DNA as the DNA between nucleosomes, i.e the linker region, is more accessible as compared to DNA in contact with the histones (Anderson *et al.,* 2000). Apart from this, it is often seen that the functionally important regions like promoters or enhancers are depleted of nucleosomes, making the regions more accessible towards the binding of transcription factors and the transcriptional machinery. These regions are called Nucleosome Depleted Regions or NDR, and are usually surrounded by two very well positioned nucleosomes, the +1 and the -1 nucleosomes.(Kornberg and Stryer, 1988). The downstream border of this NDR is formed by the +1 nucleosome which is usually placed at a canonical distance downstream of the transcriptional start site and the upstream border is formed by the -1 nucleosome. In *S. Cerevisiae*, there is an array of nucleosomes placed at a defined interval of around 165bp (dyad to dyad) downstream of the +1 nucleosome and into the gene body. The strength of positioning decreases as we move downstream of the +1 nucleosome as shown in Figure 1.1 (Mavrich *et al.,* 2008). This overall arrangement of nucleosomes across the gene body is highly dynamic and is crucial in the context of gene regulation. (Cui *et al*., 2012; Jiang, 2009)

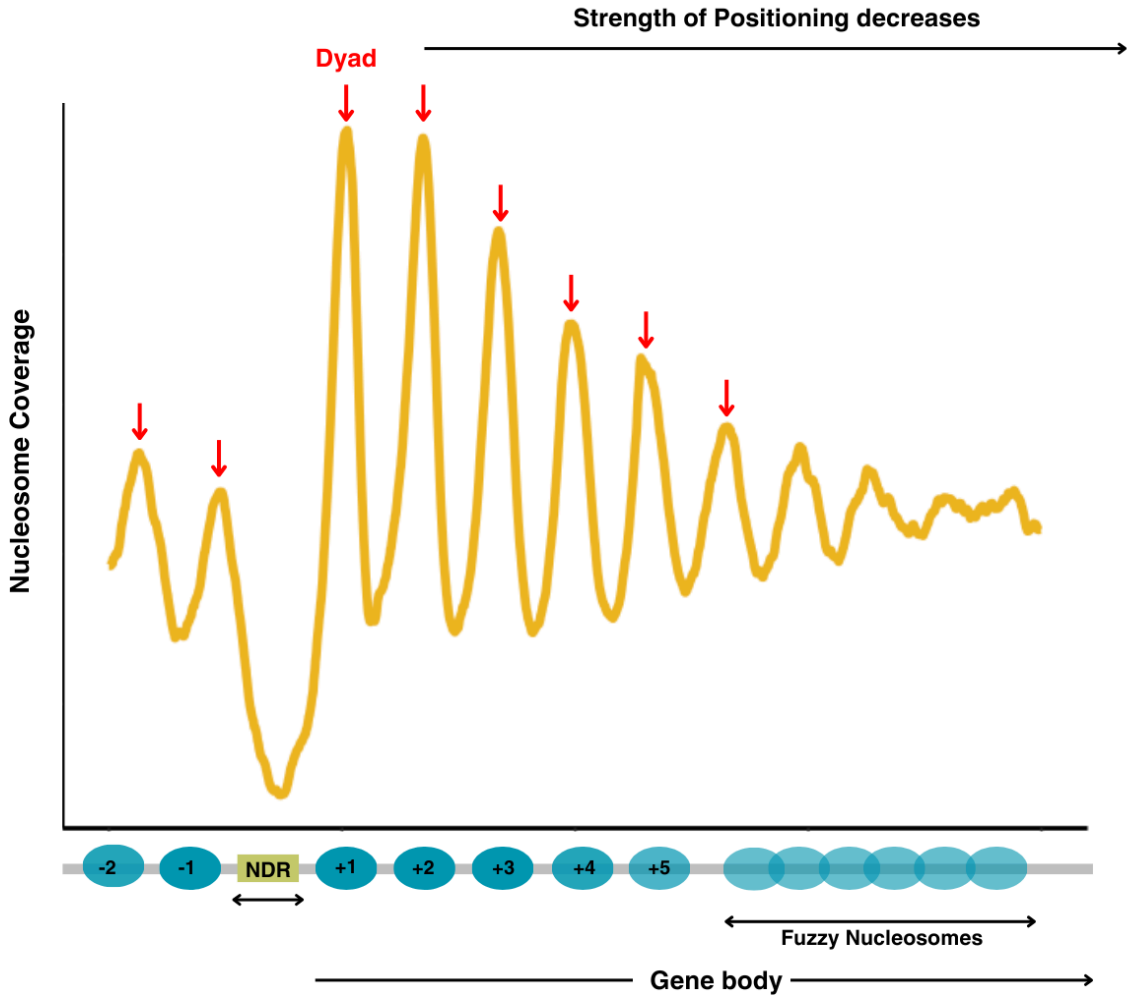


Figure 1.1| Figure 1.1 shows the nucleosome coverage and nucleosome positioning which is typically seen at gene loci. The red arrows indicate the position of the Nucleosome dyads. This Figure has been adapted from Lai and Pugh *et al.,* 2017.

1.2 Factors affecting genome-wide nucleosome positioning in *S. cerevisiae*

The precise arrangement of nucleosomes across compact genomes (like in the case of *S. cerevisiae*) led to multiple studies aimed at trying to decode the mechanisms which result in creating this intricate pattern of nucleosome positioning in a genome-wide fashion. There are three major factors known to drive nucleosome positioning: 1) Intrinsic sequence 2) ATP-dependent chromatin remodelers and 3) General Regulatory Factors (GRFs) (Struhl and Segal, 2013). The roles of these factors are summarised in Figure 1.2.1 and will be elaborated in the following sections.

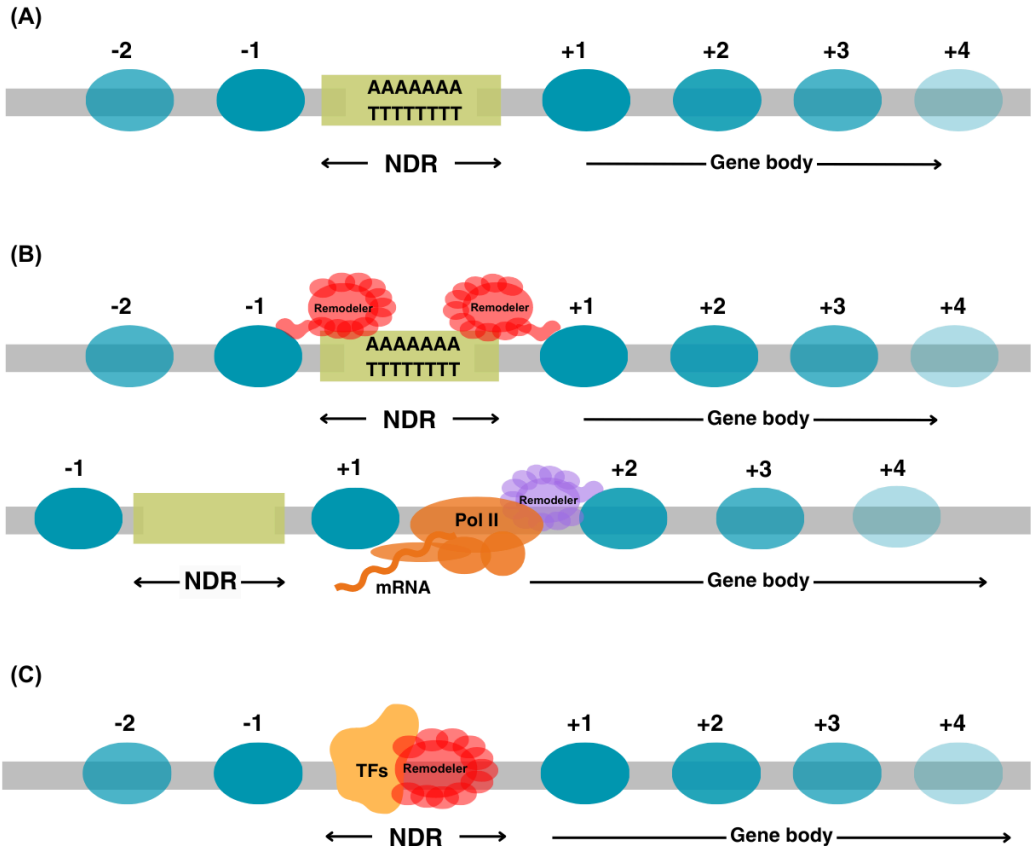


Figure 1.2.1| Figure 1.2 is a schematic diagram representing different cis and trans factors determining nucleosome positioning. Figure 1.2 A represents the role of polyA sequences in NDR formation. Figure 1.2 B and 1.2 C show the role of chromatin remodelers (in red) in establishing the NDR and positioning the NDR flanking nucleosomes. It also shows the interaction between remodelers (in purple) and Pol II (in orange)which affects the positioning of the downstream nucleosomes. Figure 1.2 C represents the role of GRFs or Transcription factors in establishing NDR. These TFs often interact with remodelers as shown in the figure. This schematic has been adapted from Struhl and Segal, 2013, the remodeler design has been inspired from Prajapati *et al*., 2020.

1.2.1 Intrinsic sequence plays a role in positioning nucleosomes

The stability of a nucleosome depends on the interaction of wrapped DNA with the histone core. Since DNA is bent around the core formed by histones, the flexibility of the DNA itself is important as more bendable sequences would favour nucleosomes whereas stiffer sequences would negatively impact the formation of a nucleosome. It was observed that AT-rich dinucleotides were more prevalent in the regions where the minor groove was facing the nucleosome core and this has been mainly attributed to the narrower minor grooves of AT-rich sequences (Drew *et al.,* 1985). This was further validated by precise mapping of nucleosomes by a technique called chemical mapping, which revealed that the nucleosomes of all the classes (+1, -1 and the downstream nucleosomes) show a strong 10 bp dinucleotide periodicity of AT-rich sequences like AT/TT/AT/TA, which may play a role in determining nucleosome positioning in the protein coding regions (Brogaard *et al.,* 2012). Along with the role of AT-rich sequence periodicity, it has also been observed that polyA sequences are stiff and can hence act as a barrier in nucleosome formation as shown in Figure 1.2 A. This intrinsic nature of polyA sequences has been shown to be crucial in maintaining the degree of nucleosome depletion at promoters of many eukaryotes, and is best demonstrated in *S. cerevisiae* (Anderson, 2001; Raveh-Sadka *et al.,* 2012).

While sequences like AT-rich dinucleotides and polyA sequences are important in genome-wide nucleosome organization, a study showed that around 50% of the nucleosome organization observed *in vivo* can be explained by considering different kinds of sequence preferences of the nucleosomes (Segal *et al.*, 2006). This is mainly because a lot of *in vitro* studies have failed in reconstituting the strong positioning of +1 nucleosomes, which often serve as anchor points in establishing the downstream nucleosomal arrays. However, when ATP and crude whole cell extracts containing a mix of proteins were added to purified histones and DNA sequences, the positioning of the +1 nucleosome could be recapitulated to some extent.(Zhang *et al.,* 2011; Krietenstein *et al.,* 2016). This indicates that there are some other factors which can override intrinsic sequence preference in order to establish the *in vivo* nucleosomal pattern. ATP-dependent chromatin remodelers and GRFs are two such trans factors known to influence genome-wide nucleosome positioning.

1.2.2 Chromatin remodelers play a role in positioning nucleosomes

Chromatin remodelers are multi-subunit proteins which contain an ATPase domain that hydrolyses ATP to catalyse chromatin remodeling. Based on the structure of their ATPase, chromatin remodelers are often classified into four major families: SWI/SNF, ISWI, CHD and INO80 as shown in Figure 1.2.2. These remodelers regulate different aspects of genome-wide nucleosome positioning (Kingston *et al.,* 1999). Remodelers from the CHD and ISWI families are often referred to as nucleosome sliders as they slide nucleosomes in place, establishing an array of well-spaced nucleosomes. Remodelers from the INO80 family are involved in nucleosome positioning and histone variant exchange. Swr1 is known to be involved in the exchange of H2A and H2AZ histones in the nucleosomes of intergenic and coding regions, which leads to a change in transcription dynamics (Mizuguchi *et al.,* 2004). SWI/SNF family remodelers have been known to be involved in NDR formation and positioning of the +1 and -1 nucleosomes. When RSC, an essential chromatin remodeler, is depleted, the NDR of some genes shrinks, accompanied with a shift of the nucleosome array towards the NDR (Hartley and Madhani, 2009; Ganguly and Chereji *et al.,* 2014).

For these remodelers to function, they need to be recruited to a genomic locus and this recruitment can happen in a couple of ways. GRFs like Reb1 and Abf1 are often hypothesized to recruit RSC, as shown in Figure 1.2.1 C (Kubik *et al.*, 2015). The Pol II transcriptional machinery is also known to recruit remodelers like Isw1 and Chd1 for positioning nucleosomes within the coding regions as shown in Figure 1.2.1 B. In addition to these mechanisms, some ChIP studies have shown that two RSC subunits: Rsc3 and Rsc30 can bind CG-rich sequences and can thus target RSC to specific genes (Zhu *et al.,* 2009). It has also been shown that even though polyA sequences are stiff and hence destabilise nucleosome formation, the extent of nucleosome depletion *in vivo* is higher than the depletion observed *in vitro*. Some gene coding regions also have polyA tracts but nucleosome depletion is not seen in those regions. These observations led to the hypothesis that there are active mechanisms like the involvement of remodelers, which could explain the higher extent of nucleosome depletion seen *in vivo*. There have been a few *in vitro* studies which have shown that polyA sequences stimulate the activity of RSC and Chd1 remodelers but the exact mechanism is not very well understood (Lorch *et al.,* 2014; Winger *et al.,* 2017).

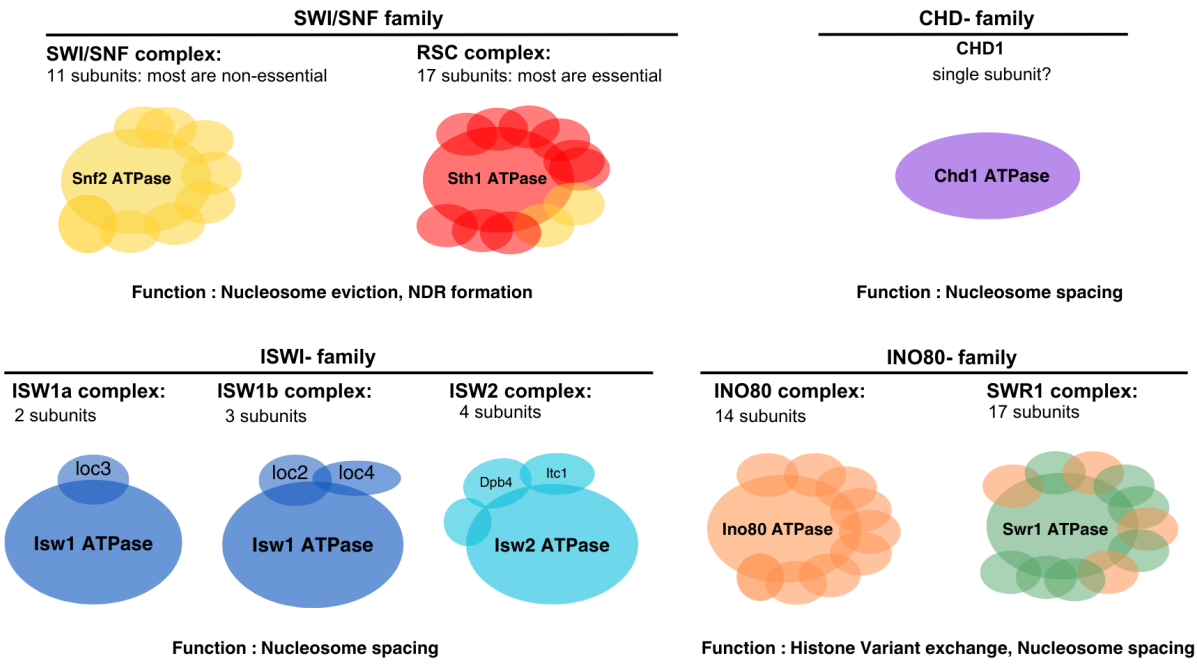


Figure 1.2.2| This figure shows different families of ATP-dependent Chromatin remodelers. The general function associated with each remodeler family has also been summarised. This schematic has been adapted from Prajapati *et al.,* 2020.

1.2.3 Transcription factors play a role in positioning nucleosomes

The degree of nucleosome depletion seen at promoters *in vivo* is not entirely recapitulated by intrinsic factors and the activity of remodelers. This is because there is a special family of transcription factors called the GRFs (General Regulatory Factors) which establish the NDRs and position nucleosomes at a subset of *S. cerevisiae* genes. Transcription factors like Abf1, Rep1 and Rap1 belong to this family of GRFs, and their conditional knockdown leads to inadequate depletion of the NDR, demonstrating their importance (Ganapathi *et al.,* 2011)

1.3 *PHO5* locus as a model locus to understand the relationship between well positioned nucleosomes and transcriptional plasticity

The *PHO5* gene in *S. cerevisiae* is a part of the PHO regulon, which consists of a group of around 20 genes regulated by the availability of phosphate in their environment (Kaneko *et al.,* 1982). Pho4 is a basic helix-turn-helix protein which is instrumental in regulating the expression of these PHO genes by binding to its motif (CACGTG) located in the upstream regions of the Pho4-regulated genes (Vogel *et al.*, 1989). When the cells are grown in rich media, Pho4 is phosphorylated and is mostly present in the cytoplasm and is not bound to its motif as shown in Figure 1.3 A. When cells are grown in phosphate-free or low-phosphate media, Pho4 is unphosphorylated and transported to the nucleus, where it binds its motifs near its target genes and induces PHO genes, which code for phosphatases among other proteins (Komeili *et al.,* 1999; O'Neill *et al.,* 1996). These phosphatases scavenge phosphate from extracellular substrates, enabling the cells to survive under phosphate starved conditions. *PHO5* is one such gene which codes for acid phosphatase. Its regulation has been very well studied due to the exemplary nature by which the gene is activated and undergoes a chromatin transition.

The *PHO5* regulatory region has 4 well-positioned nucleosomes along with two *Pho4* motifs (also referred to as Upstream Activating Sequence phosphate or the UASp): a high-affinity motif (CACGTG) and a low-affinity motif (CACGTT). The precise arrangement of all of these components is crucial in determining *PHO5* expression (Rudolph and Hinen, 1987) and is shown in figure 1.3 B. As indicated, the low affinity *Pho4* motif is exposed, whereas the high affinity *Pho4* motif is covered by the -2 nucleosome (Venter *et al.,* 1994). Upon phosphate starvation, Pho4 first binds its exposed low-affinity site, which leads to chromatin remodelling and eviction of the -2 nucleosome, exposing the previously covered high-affinity *Pho4* site. This leads to the binding of Pho4 to this high affinity site, causing further chromatin remodelling, which ultimately makes the entire region accessible and leads to *PHO5* expression as demonstrated in Figure 1.3 B. (Almer *et al.,* 1986, Bergman et al., 1983).

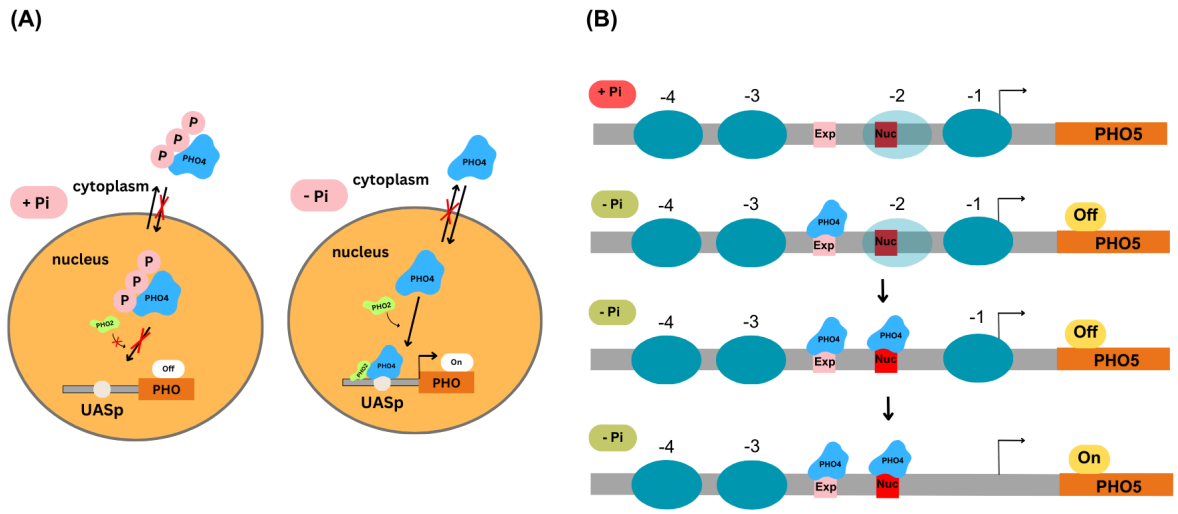


Figure 1.3| The schematic in Figure 1.3 A represents the phosphorylation and localisation of Pho4 in phosphate rich and phosphate free conditions. This has been adapted from Korber *et al.,* 2014. The schematic in Figure 1.3 B represents the changes in the chromatin state which take place when *PHO5* is induced under phosphate starved conditions. This schematic design has been inspired from Rajkumar *et al.,* 2013.

1.4 Sequence-to-function deep learning models can be used to study rules underlying genome-wide nucleosome positioning

The mechanism of *PHO5* regulation clearly demonstrates the critical role of well-positioned nucleosomes in influencing gene expression. *The PHO5* locus has been studied since the 1980s, yet the mechanism establishing the positioning of nucleosomes -1 to -4 is not known. One of the main reasons for this is the lack of understanding of the intrinsic role of DNA sequence, apart from polyA sequences or the 10bp periodicity of AT-rich dinucleotides. Intrinsic sequences could be playing a context-dependent role in determining regional nucleosome positioning, but this is hard to characterize, especially in the case of MNase-seq data, as the readout is a continuous signal across the genome. With such data, one cannot compare a region which is completely free of nucleosomes to another one which has nucleosomes to identify potential sequence rules.

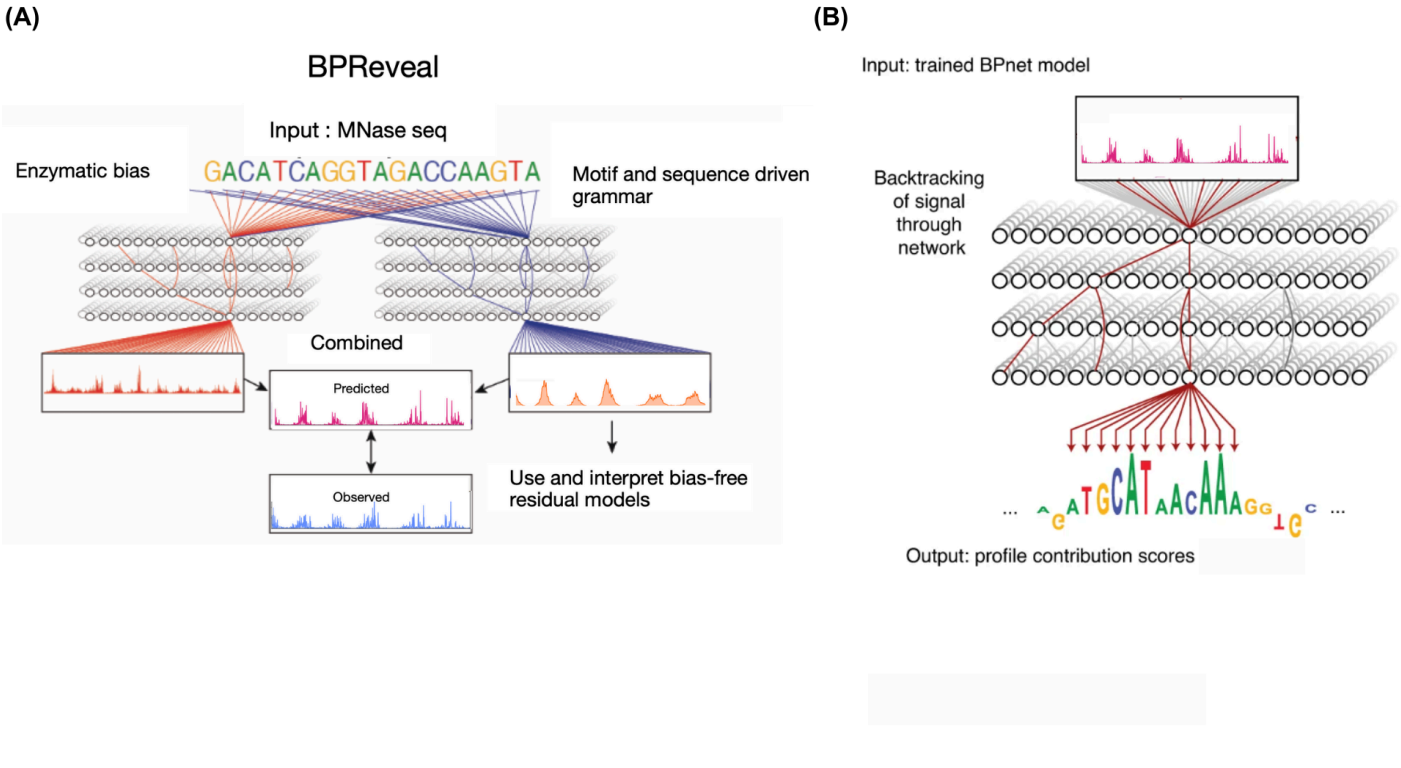
Convolutional neural networks have been recently used to predict different kinds of genomics data sets as these networks can detect non-linear patterns, and capture the sequence context without making any biological assumptions (Avsec *et al.,* 2021). This project leverages the use of deep learning models in order to understand the contribution of both the global and the local features in establishing genome-wide nucleosome positioning.

BPReveal is a sequence-to-function deep learning model which can be trained on different data sets like ChIP-seq, ATAC-seq and MNase-seq amongst others. When this model is trained on these genomic data sets, it learns the underlying sequence rules giving rise to the experimental readouts and it can thus accurately predict the experimental genome-wide readout. These models are usually trained on a part of the genome and the accuracy of these models is then assessed by analysing its predictions across the regions that it has seen during the training and the regions that were withheld. Accurate predictions across both types of regions indicate that the learned rules are general, and not just memorized based on what it has seen. When BPReveal was trained on *S.cerevisiae* MNase-seq data, it predicted the experimental data with high accuracy in both the trained and withheld regions, and it also outcompeted existing models predicting nucleosome profiles.

It has been extensively shown in the past that the enzyme MNase (Micrococcal Nuclease) used in MNase-seq experiments has a strong AT bias. It has been shown that MNase cleaves DNA upstream of an AT-rich sequence almost 30 times faster than DNA upstream of a GC-rich sequence (Dingwell *et al.,* 1981). This results in enrichment of nucleosome bound fragments from the regions which are more accessible to the

enzyme and underrepresentation of nucleosomes from the less accessible regions (Mieczkowski *et al.,* 2016; Chereji *et al.,* 2016). BPReveal can correct this enzymatic bias and generate bias-free data, which simplifies the interpretation of the sequence rules learned by the model (*Pampari et al.,* 2025). In figure 1.4 B, the experimental and the predicted tracks appear spiky due to the enzymatic bias involved. Post BPReveal's bias correction, these tracks are smoothened, enabling us to correctly visualize the nucleosome dyads.

Convolutional Neural Networks (CNNs) are often considered as black boxes as the rules learned by these models are difficult to interpret. There are however post-hoc interpretation tools which can be used to extract the learned sequence rules. BPReveal uses deepLIFT (Shrikumar *et al.,* 2017) to quantify the effect of a base pair in determining the experimental readout by assigning contribution scores as shown in figure 1.4 B. These contribution scores are then consolidated into motifs, which in this context would be sequences important for establishing a particular nucleosomal profile. BPReveal *de novo* discovers TF motifs like *Abf1, Reb1 and Rap1* as shown in figure 1.4 C, which are known to be important in NDR formation. BPReveal also discovered sequences like CGCG (CG-rich sequence) and polyA sequences which are also important in the context of nucleosome positioning.



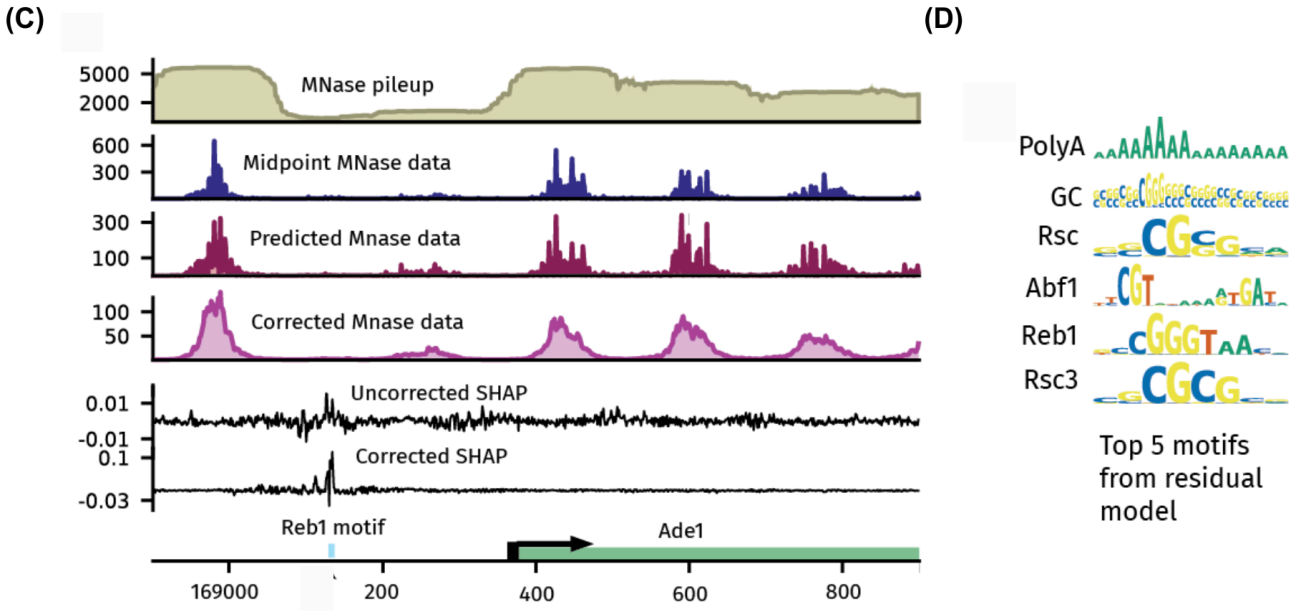


Figure 1.4| Figure 1.4 A represents different features of BPReveal, a sequence to function deep learning model developed by Charles McAnany. This schematic has been adapted from Brennan et al., 2023. In Figure 1.4 B, the schematic shows the generation of contribution scores. This figure has been adapted from Avsec et al., 2021. iN Figure 1.4 C Nucleosome coverage has been shown on the y-axis and the genomic coordinates have been shown on the x-axis. The plot shows experimental (in blue) and model predicted (in maroon) MNase-seq profile at the *ADE1* locus. It also shows the model generated bias corrected profile (in pink). Figure 1.4 D represents the top 6 motif which were returned by BPReveal. The data shown in Figure 1.4 B and C has been generated by Charles McAnany.

1.5 Objectives of this project

The first objective is to understand the biological relevance of the polyA sequences returned by the model by looking at the potential trans factors which could be involved in recognising these sequences. Previous studies have shown that polyA sequences are stiff and hence do not favour nucleosome formation. A study has also shown that 7bp long polyA sequences can influence the activity of the essential chromatin remodeling enzyme RSC in determining NDR length *in vivo* (Kubik *et al.,* 2018; Kubik *et al*., 2015). Studying the relationship between polyA sequences and remodelers has been difficult because the genome often has a lot of AT-rich regions with polyAs of different lengths. Most studies focus on polyA sequences of a certain length as looking at all possible lengths is difficult due to their high prevalence. The dynamic nature of the remodelers also makes it hard to carry out studies like ChIP-seq to understand their binding. Given the complexity of this problem, convolutional neural networks might be more useful in this case as the model can identify sequences important for nucleosome positioning based on contribution scores and is hence not limited to looking at polyA sequences of a certain length. It can go through all possible permutations of polyAs throughout the genome and return the ones which are important. When BPReveal was trained on wild-type yeast MNase-seq data, it returned close to 12500 polyA sequences that could be important in determining nucleosome positioning. The first objective of this project is to look at the potential chromatin remodelers whose activity could be influenced by these tracks by performing MNase seq experiments on different chromatin remodeler mutant strains and analysing this data across model-returned polyA sequences to understand the biological context of the model-mapped sequences.

The second objective is to leverage BPReveal to understand the complex relationship between nucleosome positioning in gene regulatory regions and its effect on the gene's transcriptional plasticity. This is because the cis (intrinsic sequence) and trans (chromatin remodelers and GRFs) factors discussed before explain some general features of global nucleosome positioning, but these factors often fail to explain the differential nucleosome positioning seen across the regulatory regions of different genes. The extent of nucleosome depletion and the length of the NDR is quite variable across different gene loci in *S. cerevisiae*. Some studies have shown that this differential nucleosome positioning is not just important in the context of variability in expression seen across genes, but also in the capability of a particular gene to alter its own expression based on rapidly changing environmental conditions (Tirosh *et al.,* 2008). Since these principles are not only true in the case of *S. cerevisiae,* but are often seen in higher eukaryotes as well, especially in the context of cell fate specification, the second aim of this project is to employ some newer approaches in order to try and

understand this complicated relationship between precise nucleosome positioning and transcriptional plasticity in *S.cerevisiae*.

In order to do this, we focus on the *PHO5* locus since the relationship between nucleosome positioning and gene expression is known to some extent, making it a good starting point to test out different approaches that can be used to tackle this question. As described before, the regulatory region of the *PHO5* locus has 4 well-positioned nucleosomes and two *Pho4* binding motifs, one with high affinity and one with low affinity. The high-affinity motif is covered by the -2 nucleosome, whereas the low-affinity motif is exposed. Studies where these motifs were swapped by mutating the sequence or completely mutated have shown that this precise arrangement is important but there are no studies where the nucleosome itself is shifted such that both the sites are either exposed or covered (Rajkumar *et al.,* 2013; Lam *et al.,* 2008). This is because perturbing the position of a particular nucleosome in a directed manner is not a trivial problem and would require a system where the role of every single base towards establishing the nucleosome pattern is known. Since deep learning models learn the underlying rules of nucleosome positioning when they are trained, these models can be used to design sequences with desired nucleosomal conformations. A system where the nucleosome positioning itself is mutated (and not the sequence) can be instrumental in understanding the precise role of a particular nucleosome in regulating different gene expression. This part of the project will focus on establishing two scenarios:
1) the -2 nucleosome is shifted such that both *Pho4* binding motifs are exposed
2) the -2 nucleosome is shifted such that both *Pho4* binding motifs are now covered. The goal is then to characterise the effects of these perturbations on gene expression by employing MS2-MCP based single mRNA detection system.

Chapter 2
Materials and Methods

2.1 Materials

All the yeast media YPD, Modified Complete (-Leu +10 mM Pi), Modified Complete (-Leu - Pi), YPD plates, G418 YPD plates, SD Leu dropout plates, 5M NaCl, Tris pH 7.5, 10% SDS, 0.5 M EDTA, 10xTE, 50xTAE, 30%Glycerol were purchased from the Stowers Institute Media prep. Micrococcal Nuclease (Cat. No: M0247S), Monarch® Spin PCR & DNA Cleanup Kit (5 µg) (Cat. No: T1130S), Monarch® Spin DNA Gel Extraction Kit (5 µg) (Cat No. T1120S) were purchased from NEB. Zymolase ® 100 T (Arthrobacter luteus) (Cat No. 120493-1) was purchased from AMS Bio. D-Sorbitol (Lot No. 21E1056857) was purchased from VWR. 0.5 M EGTA (Cat No. 40520008-1), 0.1 M Spremidine Solution (Cat No. 05292-1ML-F), 1M aq. Calcium Chloride Solution (Cat No. J63122.AE), UltraPure Agarose (Cat No. 16500500), RNase A (10mg/mL) (Cat No. EN0531), Proteinase K Solution (20 mg/ml) (Cat No. 25530049), Ultrapure Salmon Sperm DNA Solution (Cat No. 15632011) were purchased from Thermo Fisher Scientific. 2 Mercaptoethanol (Cat No. M6250), Nuclease Free Water (Cat No. W4502), Formaldehyde solution (47608) were purchased from Millipore Sigma. MasterPure Yeast dna Purification Kit (Cat No. MPY80200) was purchased from LGC Biosearch Technologies. CloneAmp™ HiFi PCR Premix (Cat No. 639298), GoTaq® Green Master Mix (Cat No. M7122) were purchased from Promega. pET264-pUC 24xMS2V6 Loxp KANr Loxp (Cat No. 104393), pET296-YcpLac111 CYC1p-MCP-NLS-2xyeGFP(Cat No. 104394) were purchased from Addgene.

2.2 Yeast strains and Growth conditions

Strain	Background	Description	Media
Wt BY4741	BY4741	MATa;his3Δ1;leu2Δ0;met15Δ0; ura3Δ0 background	YPD
isw1Δ	BY4741	BY4741 background ISW1::HYGr	YPD
chd1Δ	BY4741	BY4741 background CHD1::HYGr	YPD
swr1Δ	BY4741	BY4741 background SWR1::HYGr	YPD
Suppress 1	BY4741	BY4741 background, 3 point mutations in the PHO5 regulatory locus	YPD
Coverall 1	BY4741	BY4741 background, 10 point mutations in the PHO5 regulatory locus	YPD
Coverall 2	BY4741	BY4741 background, 5 point mutations in the PHO5 regulatory locus	YPD
Wt BY4741 PHO5 24xMS2V6	BY4741	BY4741 background, PHO5::PHO5-24XMS2V6U-variant KANr-	YPD
Suppress 1 PHO5 24xMS2V6	BY4741	Suppress 1 background PHO5::PHO5-24XMS2V6U-variant KANr-	YPD
Wt BY4741 PHO5 24xMS2V6 MCPNLS	BY4741	BY4741 background, PHO5::PHO5-24XMS2V6U-variant KANr-; <Ycp Lac111 CYC1p MCP-NLS-2xyeGFP>	SC Leu DO
Suppress 1 PHO5 24xMS2V6 MCPNLS	BY4741	Suppress 1 background, PHO5::PHO5-24XMS2V6U-variant KANr-; <Ycp Lac111 CYC1p MCP-NLS-2xyeGFP>	SC Leu DO

Table 1: This table shows the strains used in the project and the media in which the strains were cultured.

2.3 Methods

2.3.1 MNase seq

MNase-seq experiments were performed by following the protocol mentioned in Mcknight *et al.,* 2016 study. Yeast cultures were grown at 30°C in YPD to OD_{600} ~ 0.8-1 and crosslinking was performed with 1% formaldehyde for 15 minutes at room temperature. 125mM glycine was added to quench the reaction. Cell pellets were resuspended in spheroplasting buffer (1M Sorbitol, 5 mM β-mercaptoethanol, 50mM Tris pH 7.5, 2 mg/ml zymolyase (1 mL of buffer per 20 mL of cell culture) and incubated for 15 minutes at room temperature to break the cell wall. Spheroplasts derived from the cultures were digested using 100U MNase for 30-40 min at 37°C. A mix of EDTA pH 8.0 (50 mM final conc.) and EGTA pH 8.0 (50 mM final conc.) were added to stop the reaction. Samples were then incubated with RNase A (final conc. 0.2 mg/ml) at 42°C for 30 min to digest RNA. SDS and proteinase K (1mg/ml final concentration) were added and the samples were incubated at 65°C for 45 min to reverse the crosslinking. DNA extraction was done using the Monarch PCR & DNA cleanup kit. Samples were resolved on 1% agarose gel to evaluate the digestion. Mononucleosome-sized bands were extracted and libraries were prepared from 10ng purified DNA using the Watchmaker DNA Library Prep kit from Watchmaker Genomics. Two experimental replicates were generated. Paired-end sequencing was performed on AVITI (2x 75bp cycles).

MNase-seq data processing analysis:

MNase-seq paired-end sequencing reads were aligned using bowtie2 to sacCer3 genome. MNase-seq coverage was RPM normalized in RStudio. Nucleosome dyads were determined by resizing each fragment to its midpoint. MNase seq data analysis and plotting was done in RStudio.

2.3.2 Synthetic sequence design

To design novel sequences with a desired profile, a genetic algorithm (GA) was implemented.This GA designs small sets of mutations that can be applied to an initial sequence in order to maximize a user-defined property of the prediction.To design the mutations where both the Pho4 sites get covered, we used the GA to maximize the nucleosome density over two windows: 431195-431211 and 431305-431321 (all with respect to sacCer3 chrII). To design the mutant where both the Pho4 sites are exposed, we used the GA to minimize nucleosome density in a window from 431150-431250, which corresponds to the nucleosome around the high-affinity *Pho4* motif. We disallowed mutations inside the *Pho5* gene body, on the *Fkh2* motif, or on either of the

Pho4 motifs. Of the 51 runs (one for each PAM site), we manually selected a design that was predicted to alter the nucleosomes of interest while leaving the other nucleosomes undisturbed. The source code for the genetic algorithm is available at https://github.com/mmtrebuchet/bpreveal

2.3.3 Mutant Nucleosome profile prediction

A BPNet-style model on MNase-seq data from Begley et al., 2019, specifically the wild-type experiments SRR12073988 and SRR12073989 from GSE153035. Paired-end reads were aligned against the sacCer3 genome using bowtie2. Aligned fragments that spanned more than 1 kb were eliminated, but no other size selection was performed. The BPnet-style model used 9 dilational layers, 96 filters, a 1,000 bp output window, and a counts loss fraction of 0.1. ChrII, which contains the *Pho5* locus, was not included in the training data for the model. The source code for model training is available at https://github.com/mmtrebuchet/bpreveal.
We used this model to generate the tracks shown in figures 3.3.1, 3.3.2, 3.3.3, and we also used it with the GA to design our mutations at the *Pho5* locus.

2.3.4 Tagging *PHO5* gene with the MS2 cassette

PCR Amplification of the MS2 cassette

Endogenous tagging of the *PHO5* gene with MS2 cassette was done as according to the protocol described in Tutucci *et al.,* 2018. The MS2 cassette containing 24xMS2V6-*loxP*-KANr-*loxP* was amplified from the pET264 vector using CloneAmp ™ HiFi PCR premix and primers containing sequences homologous to the PHO5 locus. A touchdown PCR was done for the amplification. The PCR program used is as follows :

Cycle Number	Denaturation	Annealing	Polymerization	Hold
1	2 mins at 98°C	---	---	---
2-17	10 secs at 98°C	10 secs, 68-60°C gradient (-0.5 °C	2 mins at 72°C	---
18-38	10 secs at 98°C	10 secs at 60°C	2 mins at 72°C	---
39			2 mins at 72°C	4°C

Table 2| This table shows the PCR conditions used for amplification

The PCR product was run on a 1% gel to visualise the size of the product. The product was then purified using Monarch® Spin PCR & DNA Cleanup Kit. The next step was to transform wild type and CRISPR mutants with the PCR product.

Transformation with the MS2 cassette

Yeast strains were grown overnight in YPD at 26°C. A 6 mL secondary culture was grown until the OD_{600} of 0.6-0.8 was achieved. The culture was centrifuged and washed with Li-TE buffer and the cells were resuspended in 100 uL of Li-TE buffer. Cells were then added to a mix containing Li-TE-PEG, 5 uL of 10 mg/ml sterile and denatured salmon sperm DNA and 3 ug of purified PCR product and incubated for 30 minutes at room temperature. This was followed by incubation at 42°C for 20 minutes. Cells were then centrifuged and resuspended in DDW. Cells were centrifuged again and resuspended in 1 mL YPD. The tube containing the resuspended cells was locked , parafilmed and incubated overnight for homologous recombination to take place. The cells were centrifuged, resuspended in 100 uL YPD and plated on selective G418-YPD plates followed by incubation at 26°C for 3 days.

Genotyping for positive clone selection

6 transformants which grew on G418 plate were then restreaked for genotyping. Genomic DNA was isolated from these transformants using the MasterPure Yeast DNA Purification Kit. 3 different types of strategies were used to identify the positive clones.

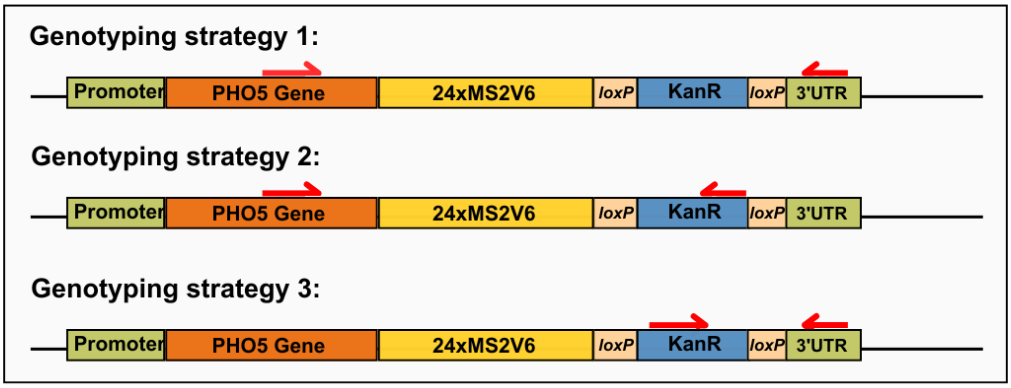


Figure 2.1| This represents the three genotyping strategies which were designed for selecting positive clones

As shown in figure 2.1, in the first genotyping strategy, oligos complimentary to the PHO5 coding region and 3'UTR were design. In this PCR reaction both positive and negative transformants are expected to form am amplicon but the sizes of the amplicon would be different. In the second Genotyping strategy, oligos were designed such that they are complimentary to the PHO5 coding region and KANr coding region. Here only the positive transformants would have an amplicon. In the third Genotyping strategy, oligos were designed such that they are complementary to the KANr coding region and the 3'UTR. PCR reactions were performed with the extracted genomic DNA as template and different combination of oligos described above using the GoTaq® Green Master Mix. PCR products were then run on a 0.8% gel to visualise the product. Products of the

transformant with expected amplicon lengths in all three PCRs were sent for sanger sequencing. Positive transformants were identified post sequencing.

2.3.5 Transformation with MCP-NLS-2xyeGFP plasmid

Yeast strains where MS2 was tagged to PHO5 were grown overnight in YPD at 26°C. A 6 mL secondary culture was grown until the OD_{600} of 0.6-0.8 was achieved. The culture was centrifuged and washed with Li-TE buffer and the cells were resuspended in 100 uL of Li-TE buffer. Cells were then added to a mix containing Li-TE-PEG, 5 uL of 10 mg/ml sterile and denatured salmon sperm DNA and 1 ug of purified pET296 vector with MCP-NLS-2xyeGFP construct. This mix was incubated for 30 minutes at room temperature. This was followed by incubation at 42°C for 20 minutes. Cells were then centrifuged and resuspended in DDW. The transformation was then plated on selective LEU dropout plates and incubated at 26°C for 3 days.

2.3.6 Live imaging of MS2 tagged strains to visualise the *PHO5* mRNA

CellASIC ONIX system was used to culture the cells during the live imaging. Yeast cells were incubated overnight at 30°C in custom made synthetic phosphate rich leucine dropout media. The culture was diluted to OD_{600} of 1 and 50 uL of this culture was loaded to the cell inlet well of the CellASIC ONIX gradient plate. Custom made phosphate rich and phosphate free media were added to the solution inlet wells. Cells were loaded to the growth chamber and allowed to grow for 3-4 hours in phosphate rich media. 5 hour time lapse was then recorded where cells were grown in hosphate rich media for 60 minutes followed by growth in phosphate free media. Images were taken every 15 minutes. To cover the entire cell volume, 21 Z-stacks were acquired. The MS2-MCP-NLS-2xyeGFP labelled mRNAs were visualised with a 488 nm laser.

2.3.7 Quantification of MS2-MCP punctae in wild type and suppress 1 mutant strains

The first step of the analysis was to obtain max projection for all the time points of the time lapse. The next step was to identify individual cells in each frame. This was doen using the Cellpose package in python. The next step was to quantify the number of punctae in each cell. This was done by establishing a baseline intensity of the diffused GFP signal and setting up a threshold intensity to ensure that punctae are not detected in cells with extremely bright GFP signal. This intensity based thresholding enabled us to detect the number of punctae in each cell.

Chapter 3
Results

3.1 Nucleosome profile in wild-type vs chromatin remodeler mutants across BPReveal-mapped polyA sequences

When BPReveal was trained on MNase-seq data from wild-type (wt) BY4741 strain, the model *de novo* discovered GRF motifs like *Abf1* and *Reb1* which are known to be important in NDR formation. The model also mapped close to 12,500 polyA sequences with different polyA sequences containing different lengths of consecutive As. In order to understand the role of model mapped polyA sequences in determining nucleosome positioning, we performed MNase-seq experiments in wt BY4741 and chromatin remodeler mutant strains Isw1Δ, Chd1Δ and Swr1Δ, which belong to different families of chromatin remodelers. The RSC remodeler mutant strain was not available in the lab so published MNase-seq data from GSE73337 were analysed in the same way as the lab generated MNase-seq data. Since Rsc is an essential remodeler, it cannot be deleted so the published study used a conditional mutant where the ATPase unit of RSC, Sth1, would be depleted by FRB protein regulated by rapamycin-based induction.

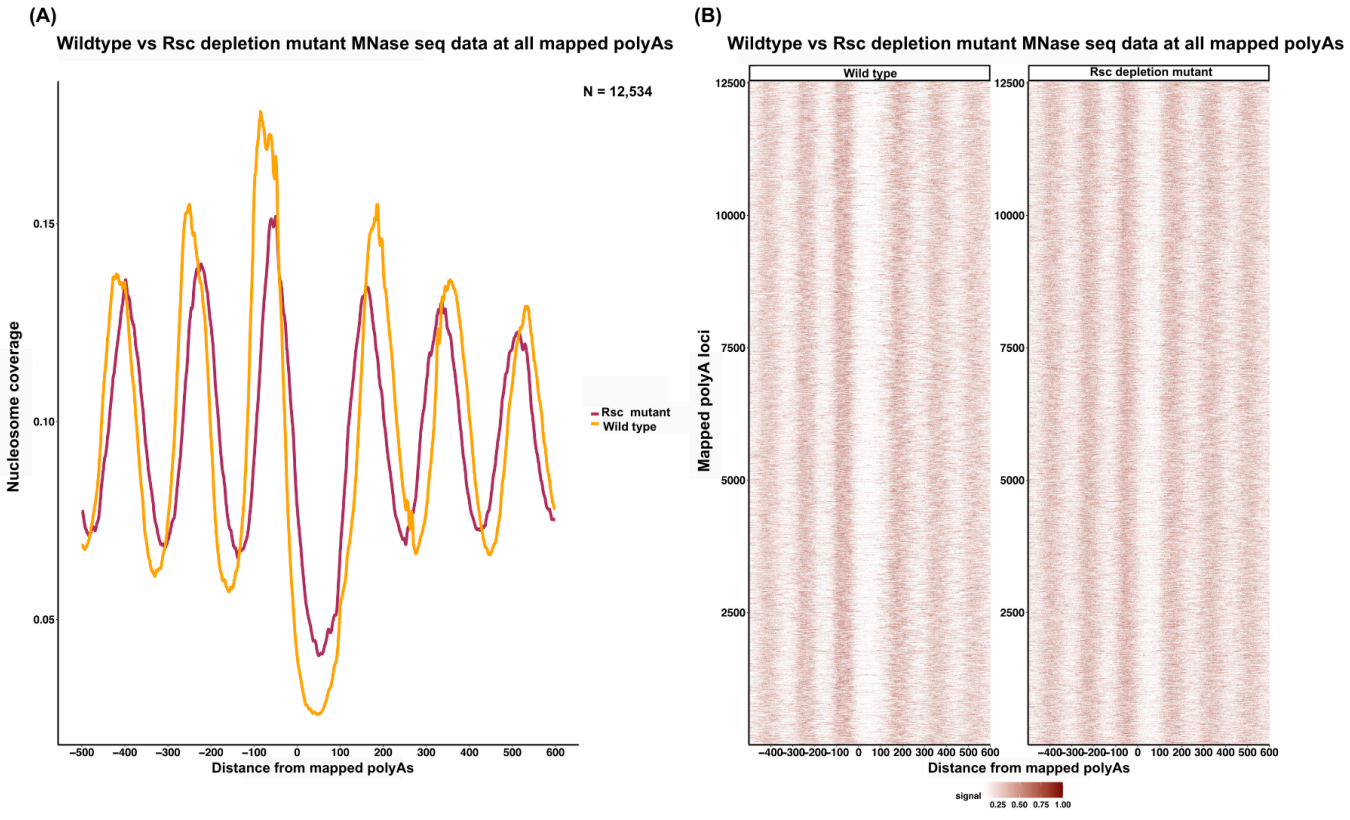
In Figure 3.1. MNase-seq data from wild-type and different remodeler mutants has been shown in two forms, a metapeak and a heatmap, centered at model-mapped 12,534 polyAs loci. A metapeak is a line graph representing the average MNase-seq signal calculated across different polyA loci, whereas a heatmap represents the MNase-seq signal at the individual locus. Figure 3.1 A shows that in RSC depleted mutants, the overall nucleosome positioning pattern in the 1.1kb region is shifted towards the NDR present near the model-mapped polyA sequences, causing the NDR to shrink. This result is consistent with the Kubik *et al.,* 2015 study, where they show that the NDR near some polyA sequences with seven As shrinks in RSC-depleted strains. The Kubik *et al.,* 2018 study also used ChEC-seq data to suggest that this NDR shrinkage is due to the binding of RSC to the polyA sequences. The studies done previously were limited to polyA sequences of 7bp length or some other fixed length, whereas the model mapped polyAs consist of different lengths of A. This shows that polyA-mediated RSC activity might not be limited to polyA sequences of a certain length, demonstrating the model's capability of capturing important sequences throughout the genome in an unconstrained manner.

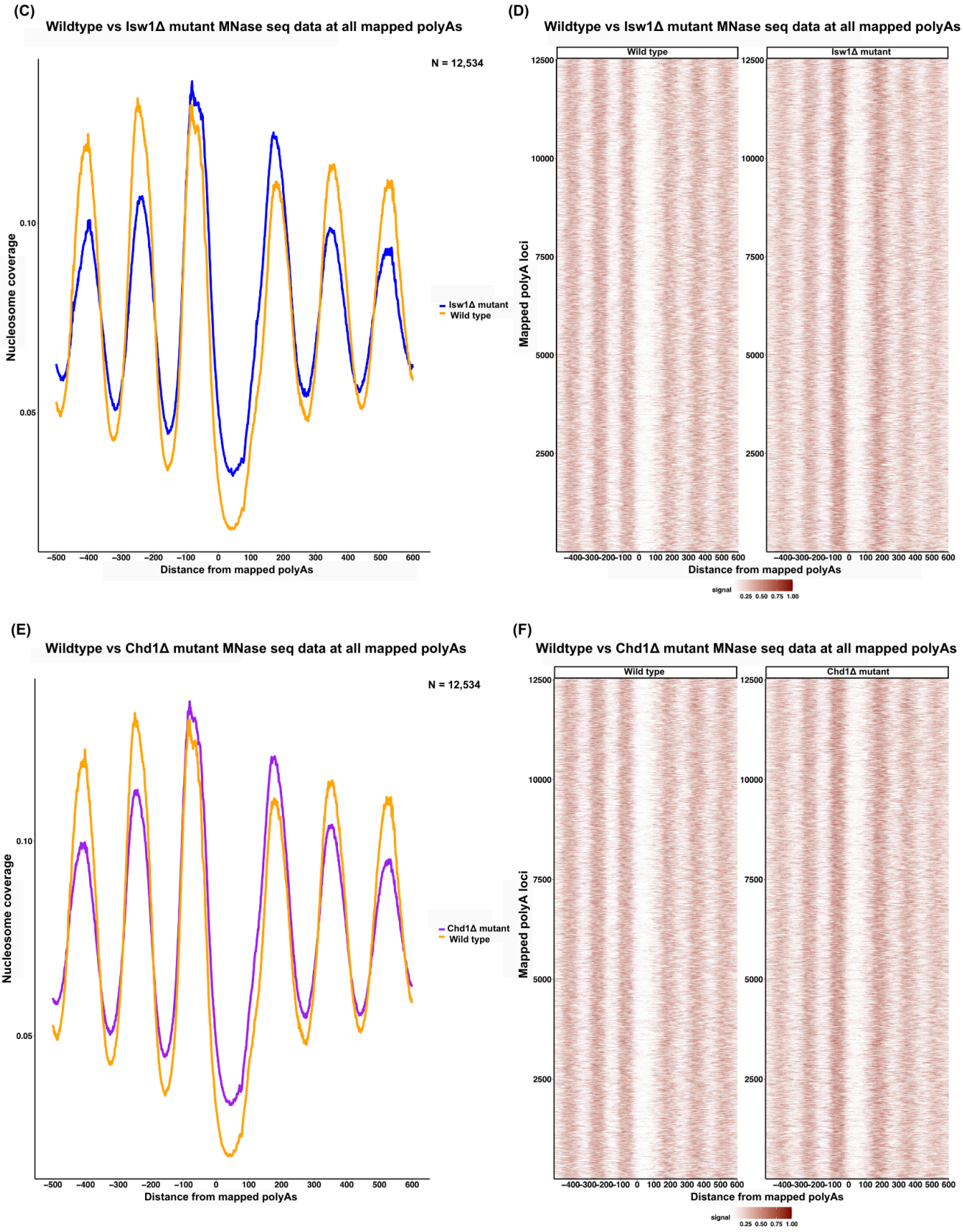
In figure 3.1 C,D the nucleosome coverage in a 1.1 kb window centered at model-mapped polyA sequences in wild-type vs Isw1Δ mutant strains has been shown. The change in nucleosome positioning in this mutant is different from the change seen

in the RSCdepleted mutant. In the Isw1Δ mutants, the signal at the nucleosomes not flanking the NDR is much lower as compared to the flanking nucleosomes as shown in figure 3.1 C,D demonstrating that isw1 deletion leads to weaker nucleosomal phasing of the non flanking nucleosomes. This is consistent with the previous literature showing that isw1 is involved in nucleosome sliding and its deletion leads to weaker nucleosomal phasing.

In figure 3.1 E,F the nucleosome coverage in a 1.1 kb window centered at model mapped polyA sequences in wild type vs Chd1Δ mutant is shown. The change in nucleosome positioning in this mutant is similar to the change seen in the Isw1Δ mutant. This is consistent with the nucleosome sliding role of the Isw1 and Chd1 remodelers reported in the past (Gkikopoulos *et al.,* 2011, Ocampo *et al.*, 2016; Ocampo and Chereji *et al.*, 2019).

In figure 3.1 G,H the nucleosome coverage in a 1.1 kb window centered at model-mapped polyA sequences in wild-type vs Swr1Δ mutant is shown. The nucleosome positioning in this mutant is quite similar to the wild type. The occupancy of the nucleosomes flanking the NDR seems higher in the Swr1Δ mutant as compared to wild-type from figure 3.1 G. The positioning and occupancy of the non-flanking nucleosomes in the Swr1Δ mutant is similar to that of wild-type. The cause for this increase in the occupancy of the flanking nucleosomes is unclear as Swr1 is known for its role in histone variant exchange. Overall the deletion Swr1 does not seem to be majorly disrupting the nucleosome positioning in the 1.1 kb region surrounding the model-mapped polyA regions.




(C)
Wildtype vs Isw1Δ mutant MNase seq data at all mapped polyAs
N = 12,534
Nucleosome coverage
0.10
0.05
−500 −400 −300 −200 −100 0 100 200 300 400 500 600
Distance from mapped polyAs
Isw1Δ mutant
Wild type
(D)
Wildtype vs Isw1Δ mutant MNase seq data at all mapped polyAs
Wild type
Isw1Δ mutant
Mapped polyA loci
12500
10000
7500
5000
2500
−400-300-200-100 0 100 200 300 400 500 600
Distance from mapped polyAs
signal
0.25 0.50 0.75 1.00
(E)
Wildtype vs Chd1Δ mutant MNase seq data at all mapped polyAs
N = 12,534
Nucleosome coverage
0.10
0.05
−500 −400 −300 −200 −100 0 100 200 300 400 500 600
Distance from mapped polyAs
Chd1Δ mutant
Wild type
(F)
Wildtype vs Chd1Δ mutant MNase seq data at all mapped polyAs
Wild type
Chd1Δ mutant
Mapped polyA loci
12500
10000
7500
5000
2500
−400-300-200-100 0 100 200 300 400 500 600
Distance from mapped polyAs
signal
0.25 0.50 0.75 1.00

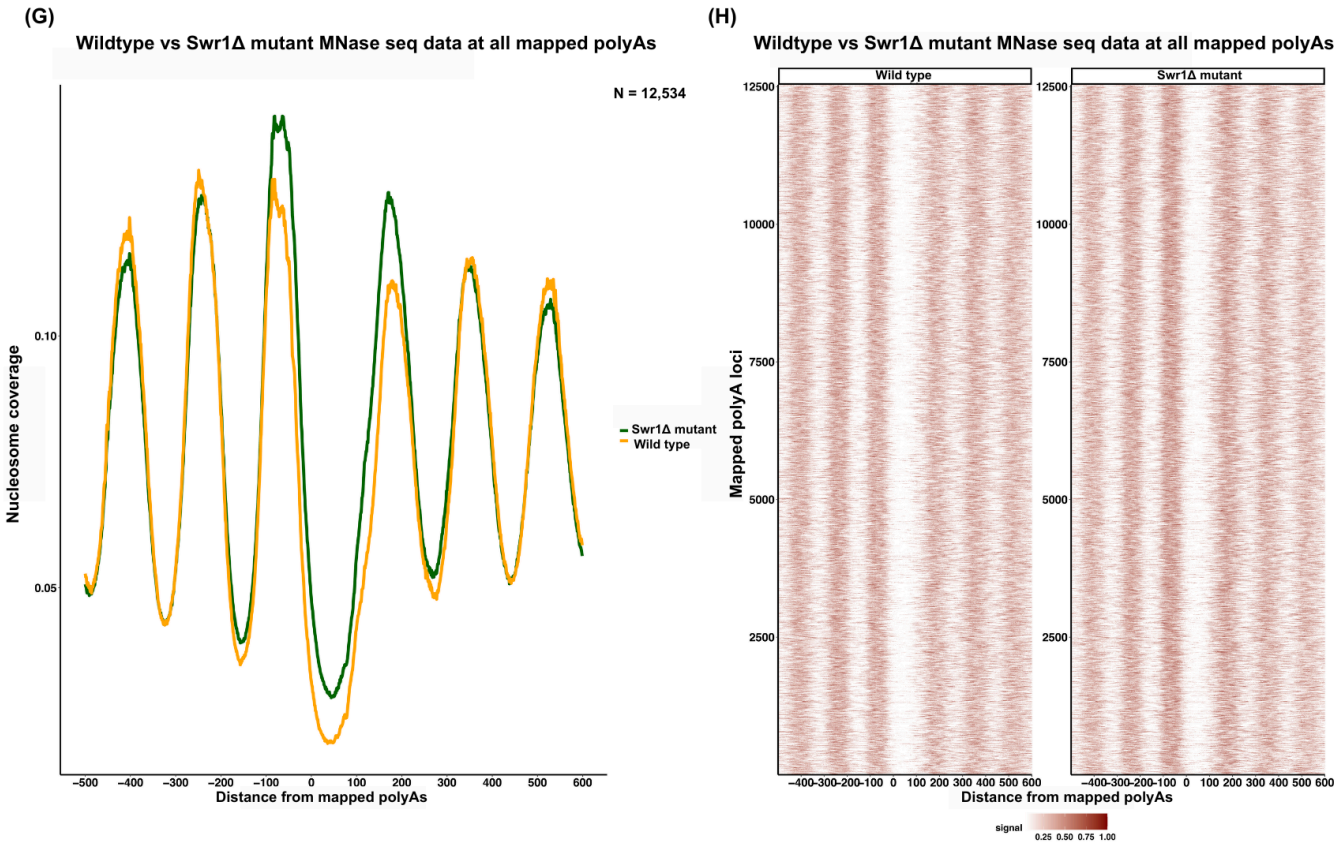


Figure 3.1| In figure 3.1 A,C,E,G the y-axis in all plot represents RPM normalised MNase-seq signal and the x-axis represents the distance from model mapped polyA sequences. These plots show the average MNase-seq profile which has been calculated in a 1.1 kb window surrounding 12534 model mapped polyA sequences from wild type and different remodeler mutants. In Figure 3.1 B,D,F,H; each row of the heatmap is an individual polyA locus and the x-axis shows the distance from the poly A loci. In these plots, MNase seq signal in 1.1 kb window centered at model mapped polyA loci, from wild type and different remodeler mutants has been plot in the form of a heatmap.

3.2 BPReveal can accurately predict differential NDR formation potentially mediated by RSC chromatin remodeler

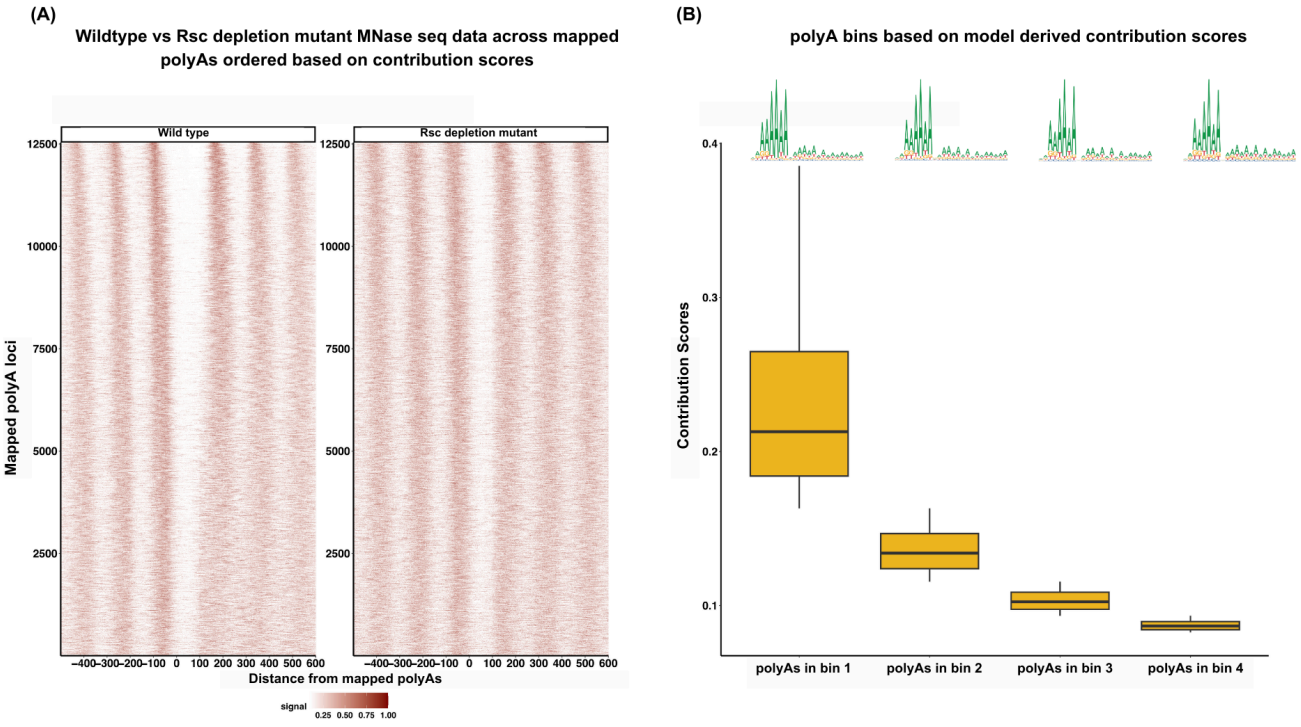
Based on figure 3.1 and previously reported literature, the RSC chromatin remodeler seems to be involved in NDR formation near the model-mapped polyA sequences. The shrinkage of the NDR in the RSCdepleted mutants is clear in the metapeak shown in figure 3.1 A, but it is not very clear in the heatmap. This is because the overall pattern of a heatmap often depends on the way it is ordered. All the heatmaps in figure 3.1 have been ordered based on the genomic positions of the model-mapped polyA sequences. In order to further explore RSC-mediated NDR formation, we decided to employ some of the model interpretation tools. DeepLIFT assigns a contribution score to every base pair, which represents the importance of sequences in predicting the nucleosomal profile in its vicinity. In order to test whether the model-derived contribution scores are indeed accurate in predicting the effect of a sequence or motif, we decided to order the heatmap based on contribution scores. In figure 3.2 A, the heatmap is ordered such that the contribution score of the model-mapped polyA sequences increases as we move from bottom to top. In the RSCdepleted mutant, we can now clearly see the shrinkage of the NDR caused by all the nucleosomes moving closer to the NDR.

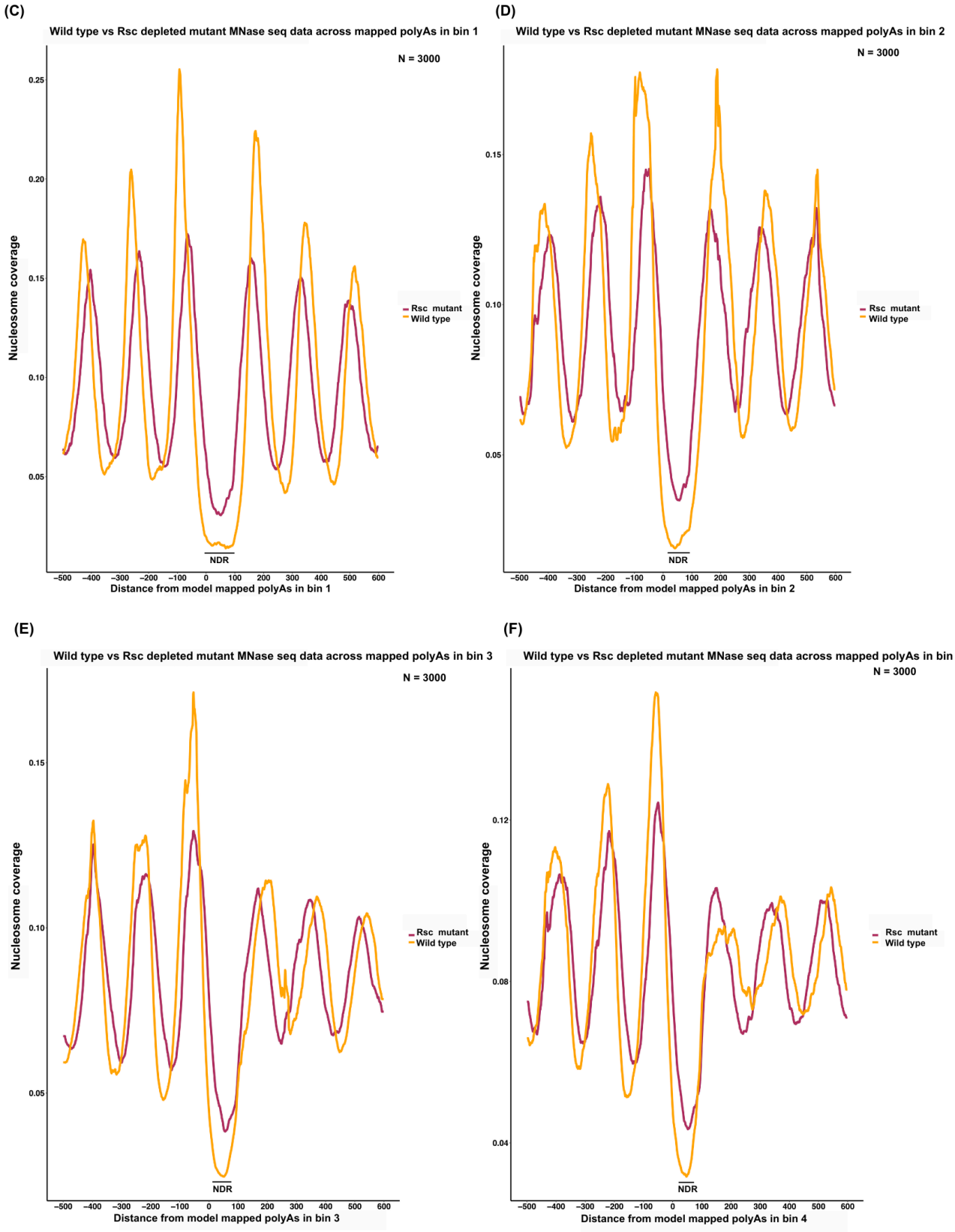
In order to take a closer look at this, we binned mapped polyA sequences to 4 quartiles, each with 3000 model-mapped polyAs. This binning was done using the model-generated contribution scores. From figure 3.2 B, we can see that the first bin has polyAs with high contribution scores, and the contribution score decreases across the subsequent bins. This kind of binning enabled us to visualize the extent of nucleosome positioning changes across loci with varying contribution scores. The sequence logos at the top of the plot represent the average composition of the sequences (PWMs) belonging to each bin. Based on the logos, it is clear that the polyA sequences in each bin are not very different in their composition, such as the number of As or some other factor.

The next step was to look at MNase-seq data in wild type vs RSCdepleted mutant. In figure 3.2 C,D,E,F, the MNas-seq data have been plotted across polyA sequences belonging to different bins. In figure 3.2 C, MNase-seq data have been plotted across polyA sequences belonging to bin 1.The average width of the NDR near the mapped model-mapped polyA sequences is close to 150 bp in the wild-type strain as shown in figure 3.2 C. In the mutant, the occupancy of all the nucleosomes in the 1.1 kb window has gone down and the nucleosomes have shifted towards the NDR, causing the NDR to shrink to around ~ 100 bp. In figure 3.2 D,E,F we see a similar trend but the extent of change in nucleosome positioning in wild-type vs RSCdepleted mutant is decreasing as we move from figure 3.2 C to figure 3.2 F. The change in NDR width across different

polyA bins is also quite striking. We see a steady decrease in the width of NDR in the wild-type strain as we move across the 4 polyA bins. The width ranges from close to 150 bp for the first polyA bin to around 75 bp in the last polyA bin (shown in Figure 3.2 F). The extent of change in nucleosome positioning has been quantified in figure 3.2 G. This change was quantified by first calculating the difference between the MNase-seq signal in wild-type vs RSC-depleted mutant for every base in the 1.1 kb window centered at the model-mapped polyA sequences. The difference for every single base was added for 1 locus and every point in the boxplot represents this sum for 3000 polyA loci in each bin. Figure 3.2 G clearly shows that the contribution scores correlate very well with the extent of disruption in nucleosome positioning caused by RSC depletion as the change between wild-type and RSC-depleted mutant decreases consistently across the polyA bins made based on the contribution scores. The strong correlation suggests that the model has correctly learned which polyA sequences have a strong effect on nucleosome positioning.

In order to look for potential causes that led to the change in the NDR width across different polyA bins, we decided to look at the 400 bp window surrounding model-mapped polyA sequences. We found that polyAs in the first bin are surrounded by a higher number of other model mapped polyA sequences as compared to bin 4. This trend is clear from figure 3.2 H, where we see that the average number of model-mapped polyA sequences in a 400 bp window (centered at polyA from different bins) decreases as we move across the bins. This might indicate that polyA sequences act in a cooperative manner, but further investigation is needed to prove this.




(C)
Wild type vs Rsc depleted mutant MNase seq data across mapped polyAs in bin 1
N = 3000
Nucleosome coverage
Rsc mutant
Wild type
NDR
Distance from model mapped polyAs in bin 1
(D)
Wild type vs Rsc depleted mutant MNase seq data across mapped polyAs in bin 2
N = 3000
Nucleosome coverage
Rsc mutant
Wild type
NDR
Distance from model mapped polyAs in bin 2
(E)
Wild type vs Rsc depleted mutant MNase seq data across mapped polyAs in bin 3
N = 3000
Nucleosome coverage
Rsc mutant
Wild type
NDR
Distance from model mapped polyAs in bin 3
(F)
Wild type vs Rsc depleted mutant MNase seq data across mapped polyAs in bin 4
N = 3000
Nucleosome coverage
Rsc mutant
Wild type
NDR
Distance from model mapped polyAs in bin 4

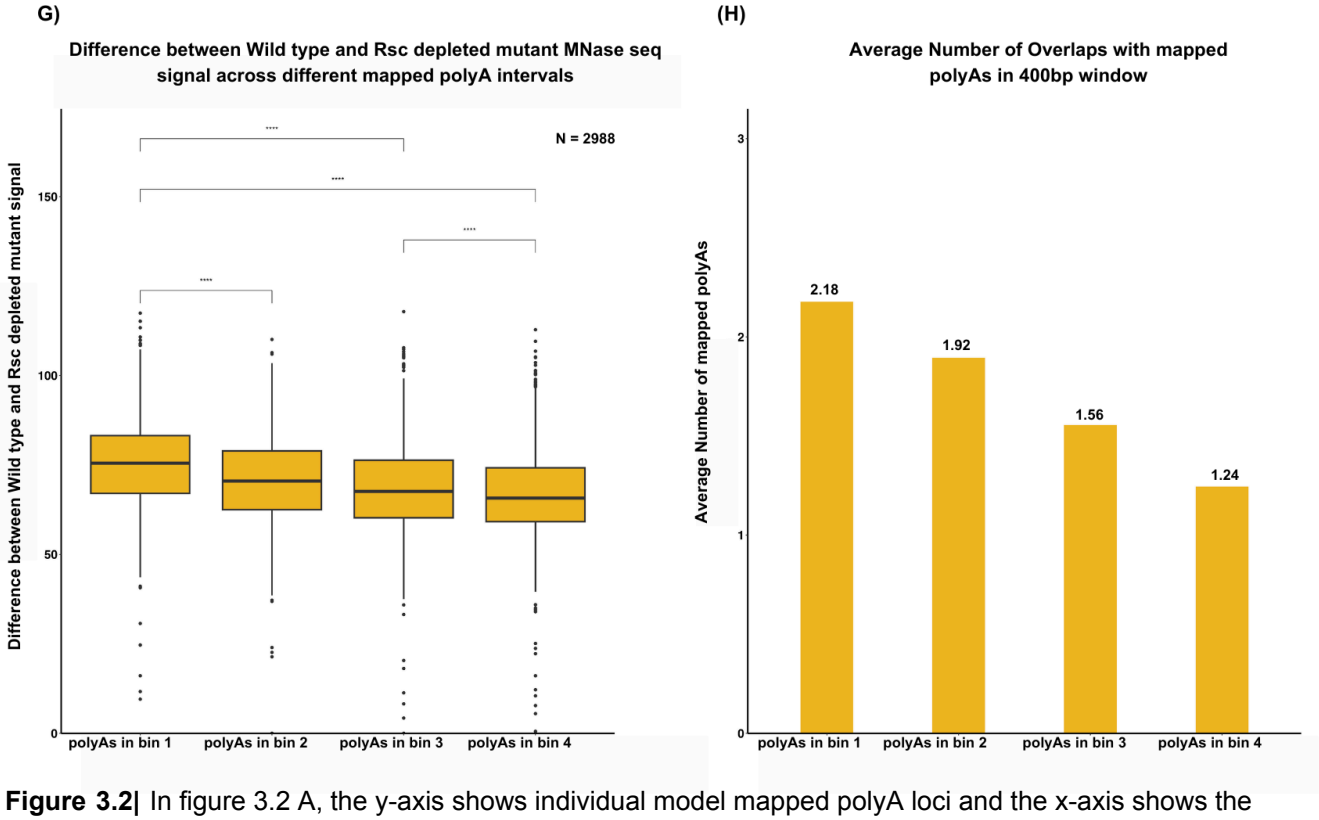


Figure 3.2| In figure 3.2 A, the y-axis shows individual model mapped polyA loci and the x-axis shows the distance from the mapped polyA loci. The plot represents MNase-seq data from wild type and Rsc depleted mutant in the 1.1 kb window centered at model mapped polyA sequences in the form of a heatmap. In this figure the polyA loci are arranged in the order of increasing contribution scores. In figure 3.2 B, the model determined contribution scores have been shown on the y-axis and the polyA bins have been shown on the x-axis. In this plot, the range of contribution scores associated with each polyA bin has been shown as a box plot. For the sake of clarity, outliers have been removed. The logos in the plot are PWM representations of the polyA sequences in each bin. In figure 3.2 C,D,E,F normalised nucleosome coverage has been shown on the y-axis and the distance from mapped polyA loci in the 1.1 kb window has been shown on the x-axis. These plots show the average MNase-seq profile for wild type(in yellow) and Rsc depleted mutant (maroon) which has been calculated in a 1.1 kb window centered at 3000 model mapped polyA sequences belonging to different bins described in figure 3.2 B. In Figure 3.2 G, the y-axis represents the sum of difference between wild type and Rsc depleted MNase-seq signal at each loci in a 1.1 kb window surrounding polyAs in a particular bin. The x-axis shows different polyA bins. Each point in the box plot represents this sum of difference for individual polyA loci in a particular bin. Some loci were removed from this analysis as their values were very high likely due to some sequencing artifacts. A t-test was performed to calculate statistical significance. In figure 3.2 H, the average number of polyA sequences found in 400bp window centered at polyA sequences in particular bins has been shown on the y-axis.The x-axis shows different polyA bins.

3.3 BPReveal can be used as a tool to design mutations to perturb nucleosome positioning in a desired manner

The results in section 3.1 and 3.2 show that BPReveal has learnt complex sequence rules directing genome-wide nucleosome positioning. In this part of the thesis, we focussed on leveraging the model to design synthetic sequences such that nucleosomes can be perturbed in a controlled manner. We decided to focus on the *PHO5* locus because the well-positioned nucleosomes in its regulatory region play a crucial role in its gene regulation, making it an excellent model locus to study the relationship between precise nucleosome positioning and gene regulation. The *PHO5* regulatory region has 4 well positioned nucleosomes, and 2 *Pho4* binding motifs, a low-affinity one (CACGTT) and a high-affinity one (CACGTG). The nucleosomes are positioned such that the low-affinity site is exposed, whereas the high-affinity site is covered by the -2 nucleosome as shown in figure 3.3.1 A. When cells are grown in phosphate-free media, Pho4 (a transcription factor) first binds its exposed low-affinity site, which leads to chromatin remodelling and eviction of the -2 nucleosome, exposing the previously covered high-affinity *Pho4* site. This leads to the binding of Pho4 to this high-affinity site, triggering further chromatin remodelling that ultimately makes the entire region accessible and leads to the expression of the *PHO5* gene.

Based on previous studies, we know the consequences of mutating the *Pho4* motifs, but the effect of perturbing a nucleosome while keeping the motifs intact is not known. This is because perturbing the positioning of one nucleosome in a controlled manner is not a trivial problem. Since our model can accurately predict MNase-seq data, we decided to leverage it to design sequences such that the nucleosome positioning can be altered in a desired manner. We aimed at generating two configurations, one where both the *Pho4* sites are exposed and the other where both the *Pho4* sites are covered. We tested one design where both the *Pho4* motifs are expected to be exposed and two designs where both the *Pho4* motifs are expected to be covered

3.3.1 BPReveal designed sequences can form a new NDR

We leveraged the genetic algorithm of the model to design a synthetic sequence such that both the *Pho4* motifs are covered by nucleosomes. We tested two synthetic designs generated by the model, which were expected to change the nucleosome profile such that both *Pho4* sites are covered ("coverall") as shown in Figure 3.3.1 A. In this section, I will focus on the coverall 1 mutant, where 10 point mutations were made such that an *Abf1* motif is created around 200 bp away from the *Pho4*-exposed binding site along with a CG-rich sequence. Abf1 is a GRF known to create NDRs. The mutations made in this strain have been shown in Figure 3.3.1 B. The y axis represents the model-determined contribution scores in wild-type and the mutant strain. We can see that the 10 point mutations suggested by the model leads to an increase in the contribution score of the new *Abf1* motif along with the CG rich sequence.

Figure 3.3.3 C shows the predicted nucleosomal profile in wild-type and the coverall 1 mutant strain. The predicted profile suggests that a new NDR will be created due to the introduction of an *Abf1* motif resulting in a shift in the -3 nucleosome such that both the sites are being covered to some extent. Both the *Pho4* sites cannot be fully covered as the two *Pho4* binding sites are around ~110 bp away so it's physically not possible to cover both the sites with high occupancy as one nucleosome is associated with at least 150 bp of DNA. In the predicted mutant profile we can see that the previously exposed site is now covered but the nucleosomal occupancy of the previously covered site is actually decreasing as it seems to be closer to the nucleosome linker.

In order to validate this prediction, we performed MNase-seq experiments with the wild-type and the coverall 1 mutant strains. The results are shown in figure 3.3.1 D. In the mutant profile, we can see that a new 200 bp long NDR has been created near the *Abf1* motif (shown using the red box in the graph) as predicted. The nucleosomes next to the NDR seem fuzzier than predicted as the start and end of nucleosomes is not clear in the mutant. This fuzziness could be due to differential occupancy of the nucleosomes across the population so the well-positioned nature of the nucleosomes has been disrupted. The mutant MNase-seq profile also shows that the occupancy of the nucleosomes at the previously exposed *Pho4* site has gone up but the fuzziness of the profile makes it difficult to conclude whether both the sites are indeed covered.

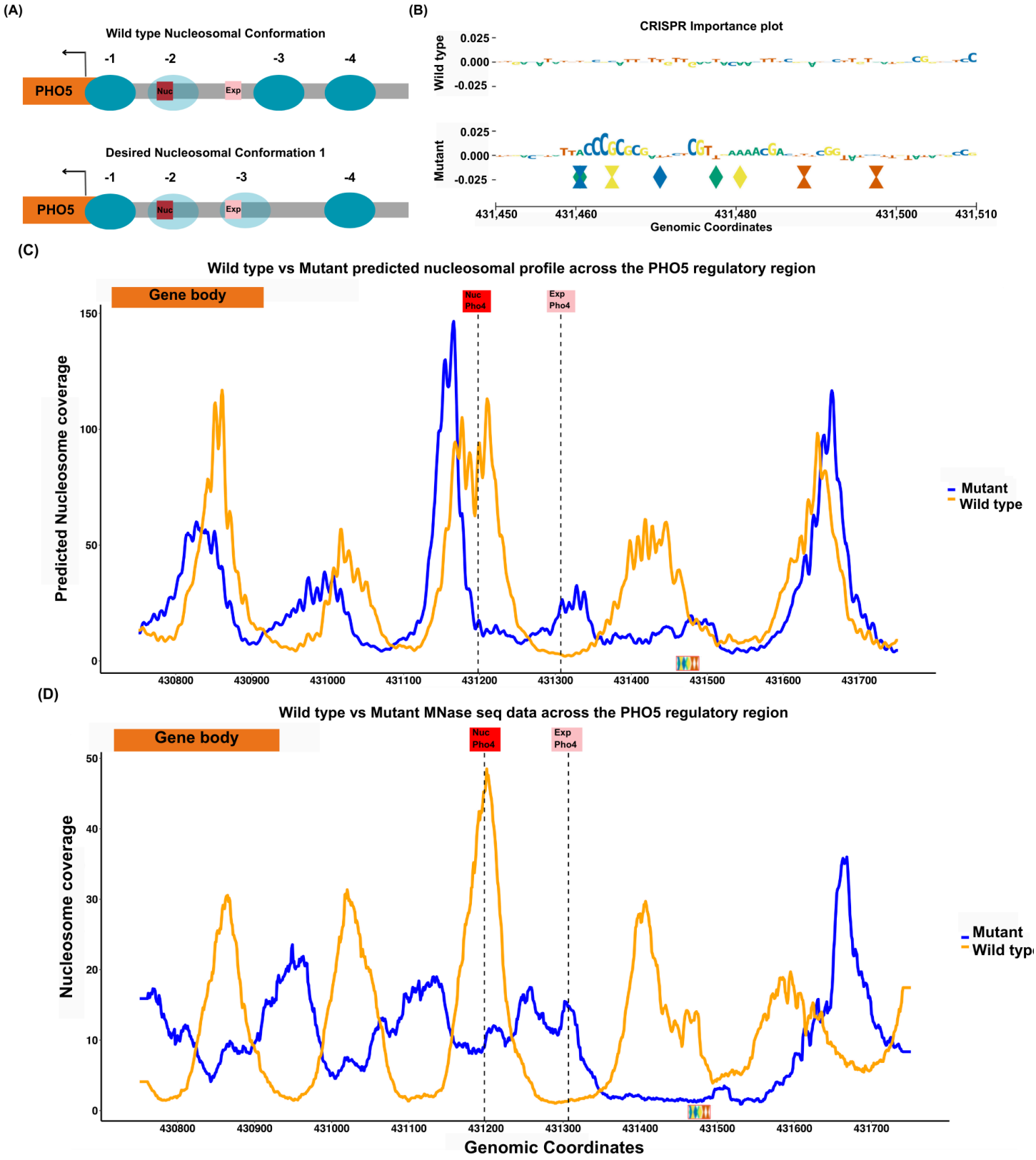

(A)
Wild type Nucleosomal Conformation
-1
-2
-3
-4
PHO5
Nuc
Exp
Desired Nucleosomal Conformation 1
(B)
CRISPR Importance plot
Wild type
Mutant
0.025
0.000
-0.025
431,450
431,460
431,480
431,500
431,510
Genomic Coordinates
(C)
Wild type vs Mutant predicted nucleosomal profile across the PHO5 regulatory region
Gene body
Nuc Pho4
Exp Pho4
Predicted Nucleosome coverage
150
100
50
0
430800
430900
431000
431100
431200
431300
431400
431500
431600
431700
Mutant
Wild type
(D)
Wild type vs Mutant MNase seq data across the PHO5 regulatory region
Gene body
Nuc Pho4
Exp Pho4
Nucleosome coverage
50
40
30
20
10
0
Genomic Coordinates
Mutant
Wild typ

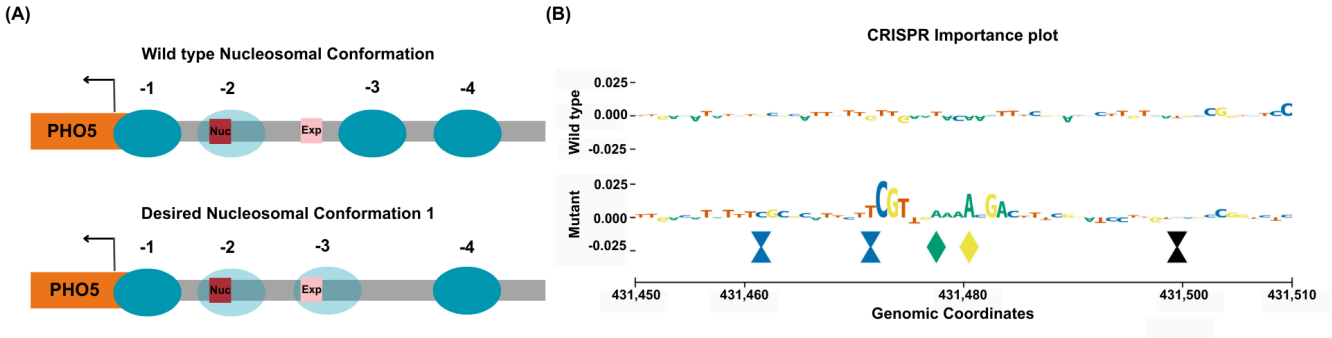
Figure 3.3.1| Figure 3.3.1 A is an illustration representing the wild type and desired nucleosomal conformations in the *PHO5* regulatory region adapted from Rajkumar *et al.,* 2013. In figure 3.3.1 B the model-determined contribution scores have been shown on the y-axis and genomic co-ordinates have been shown on the x-axis.The plot shows the point mutations proposed by the model and the contribution score associated with these mutations. The colored hourglass in the plot represent insertions, color coded according to the base being added. The wedges represents a base substitution, color coded according to the base formed post substitution. In figure 3.3.1 C, the predicted nucleosome coverage has been shown on the y-axis and the genomic coordinates have been shown in the x-axis. The plot shows the model predicted MNase-seq profile from wild type and the coverall 1 mutant strains, in a 900bp window near the *PHO5* locus. The CRISPR importance score and model prediction data has been generated by Charles McAnany. In figure 3.3.1 D, the MNase-seq nucleosome coverage has been shown on the y-axis and the genomic coordinates have been shown on the x-axis. The plot shows the MNase-seq profile from wild type and coverall 1 mutant strains, in a 900bp window near the *PHO5* locus.

3.3.2 BPReveal designed sequences can perturb nucleosome positioning such that both *Pho4* motifs are covered

In this section, I will focus on the second mutant design for achieving the nucleosomal profile such that both *Pho4* sites are covered. The coverall 2 mutant design involved making 5 point mutations such that an *Abf1* motif is formed around 200 bp away from the exposed binding. The mutations in this mutant are slightly different from the ones described in figure 3.3.1 B. The 5 point mutations, as shown in figure 3.3.2 B, result in the creation of a new *Abf1* motif, which can be seen by the resulting increase in contribution scores.

Figure 3.3.2 C shows the predicted nucleosomal profile in wild-type and the mutant strain. The predicted profile suggests that a new NDR will be created due to the introduction of an *Abf1* motif. In the predicted profile, the occupancy of the -3 nucleosome goes down and the dyad has been shifted closer to the previously exposed motif.

In order to validate this prediction, we performed MNase-seq experiments with wild-type and the coverall 2 mutant strain. The results are shown in figure 3.3.2 D. In the mutant profile, we can see that a new 200 bp long NDR is being created near the *Abf1* motif (shown using the red box in the graph), and the -3 nucleosome shifts such that the previously exposed *Pho4* binding site is now being covered. The nucleosome positioning in this mutant is a lot better in this mutant as compared to the one described in the previous section as we can clearly see nucleosome peaks. In the mutant profile we see an overall shift in the positioning of the nucleosomes towards the left and the extent of this shift is a little higher as compared to the predictions (figure 3.3.2 C). Here, we achieved the desired nucleosomal profile where both the Pho4 sites are covered to some extent but the other nucleosomes in the window are also perturbed.



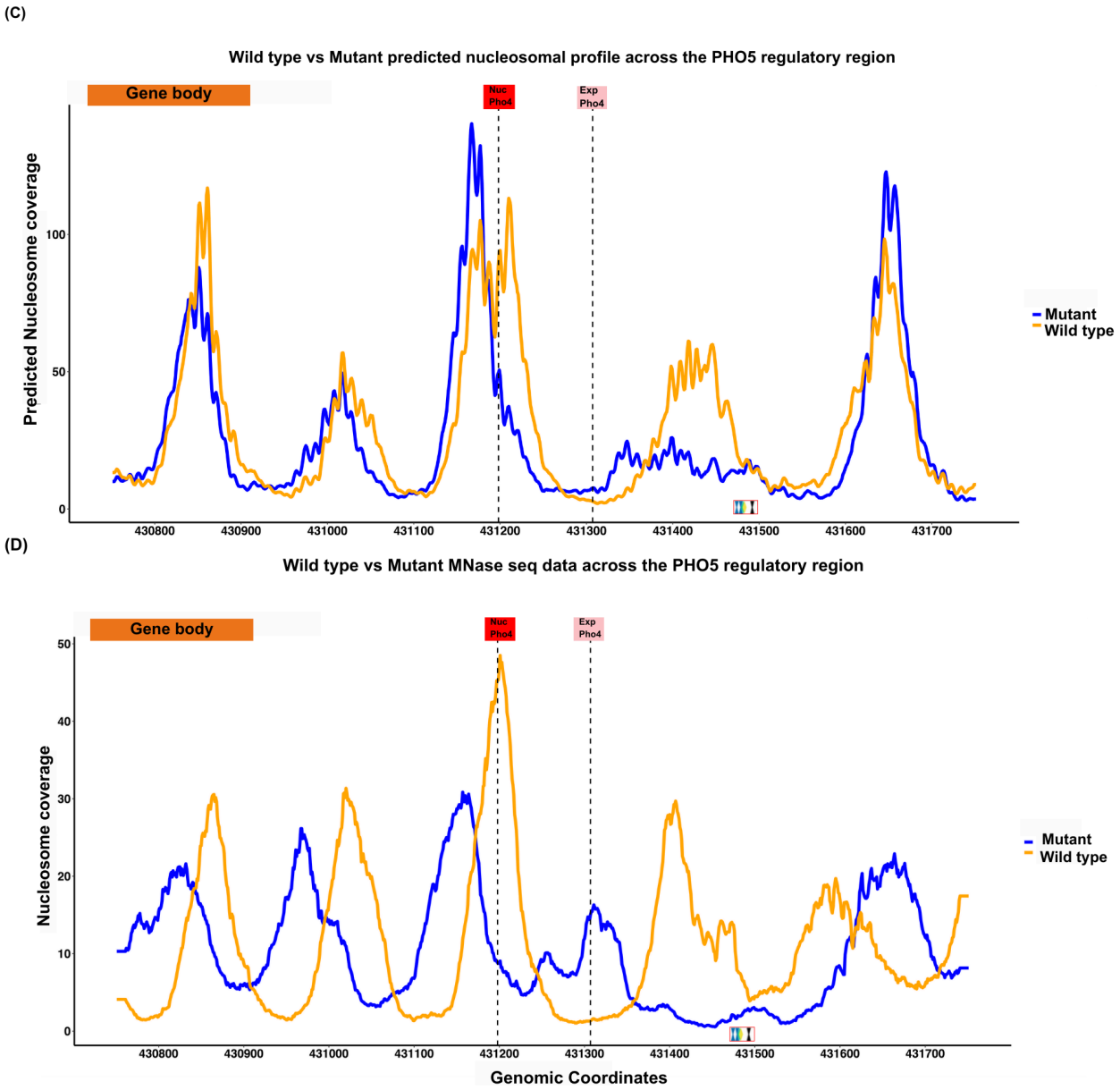
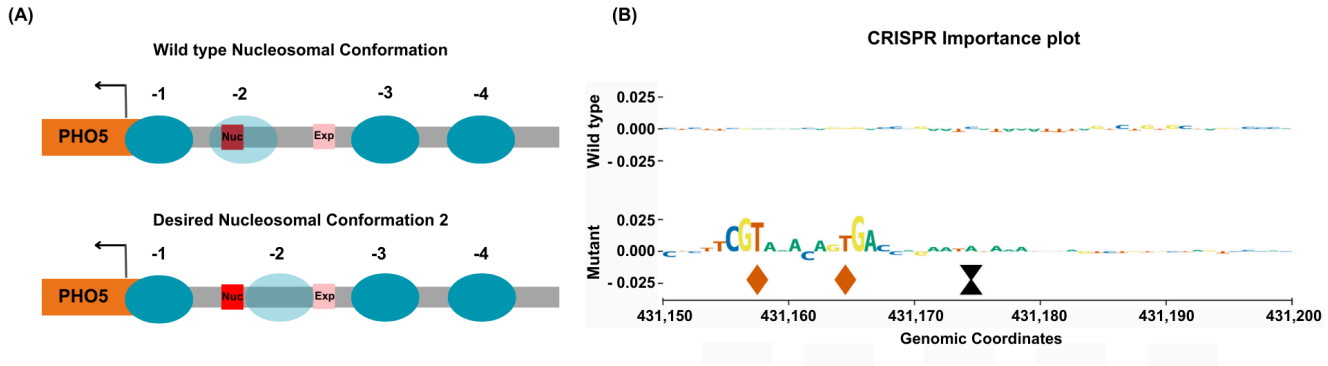


Figure 3.3.2| Figure 3.3.2 A is an illustration showing the wild type and desired nucleosomal conformations in the *PHO5* regulatory region adapted from Rajkumar *et al.,* 2013. In figure *3.3.2 B* the y-axis shows the model-determined contribution scores and the x-axis shows genomic coordinates. The plot shows the point mutations proposed by the model and the contribution score associated with these mutations. The colored hourglass in the plot represent insertion, color coded according to the base being added. The wedges represents a base substitution, color coded according to the base formed post substitution. The black hourglass represents a deletion. In figure 3.3.2 C, the y-axis shows the predicted nucleosome coverage and the x-axis shows the genomic coordinates. The plot represents the model predicted MNase-seq profile from wild type and the coverall 2 mutant strains, in a 900 bp window near the *PHO5* locus. The CRISPR importance score and model prediction data has been generated by Charles McAnany. In figure 3.3.2 D, the y-axis shows the nucleosome coverage and the x-axis shows the genomic coordinates. The plot shows the MNase-seq profile from wild type and coverall 2 mutant strains, in a 900 bp window near the *PHO5 locus.*

3.3.3 BPReveal designed sequences perturb nucleosome positioning such that both *Pho4* motifs are exposed

We also asked the model to generate a synthetic sequence such that both the *Pho4* motifs will be exposed as shown in Figure 3.3.3 A. The model came up with a solution involving three point mutations such that a new *Abf1* motif is created around 30 bp away from the nucleosomal *Pho4* binding site. The plot in figure 3.3.3 B shows the mutations made in this strain. We call this strain as “suppress 1” since the goal was to suppress the formation of nucleosomes over the nucleosomal *Pho4* site. In the wild-type strain, the contribution scores of all the bases in the window are quite low, but when we make two point mutations, the contribution score increases as the mutation results in the formation of a new *Abf1* motif. The third mutation also leads to a slight increase in the contribution score.

Figure 3.3.1 C shows the model-predicted nucleosome profile in the wild-type and the suppress 1 mutant strain. The predicted profile from figure 3.3.1 C suggests that the mutations will result in the downstream shift of the -2 nucleosome along with a decrease in occupancy, exposing the previously covered *Pho4* binding site. In order to test this prediction, we performed MNase-seq experiment with wild-type and the suppress 1 mutant strain Figure 3.3.1 D shows the results, which show that the -2 nucleosome has indeed shifted to the right. The occupancy of the nucleosome is not going down as predicted, but we achieved our desired conformation where both the *Pho4* sites are exposed without perturbing any other neighbouring nucleosomes.



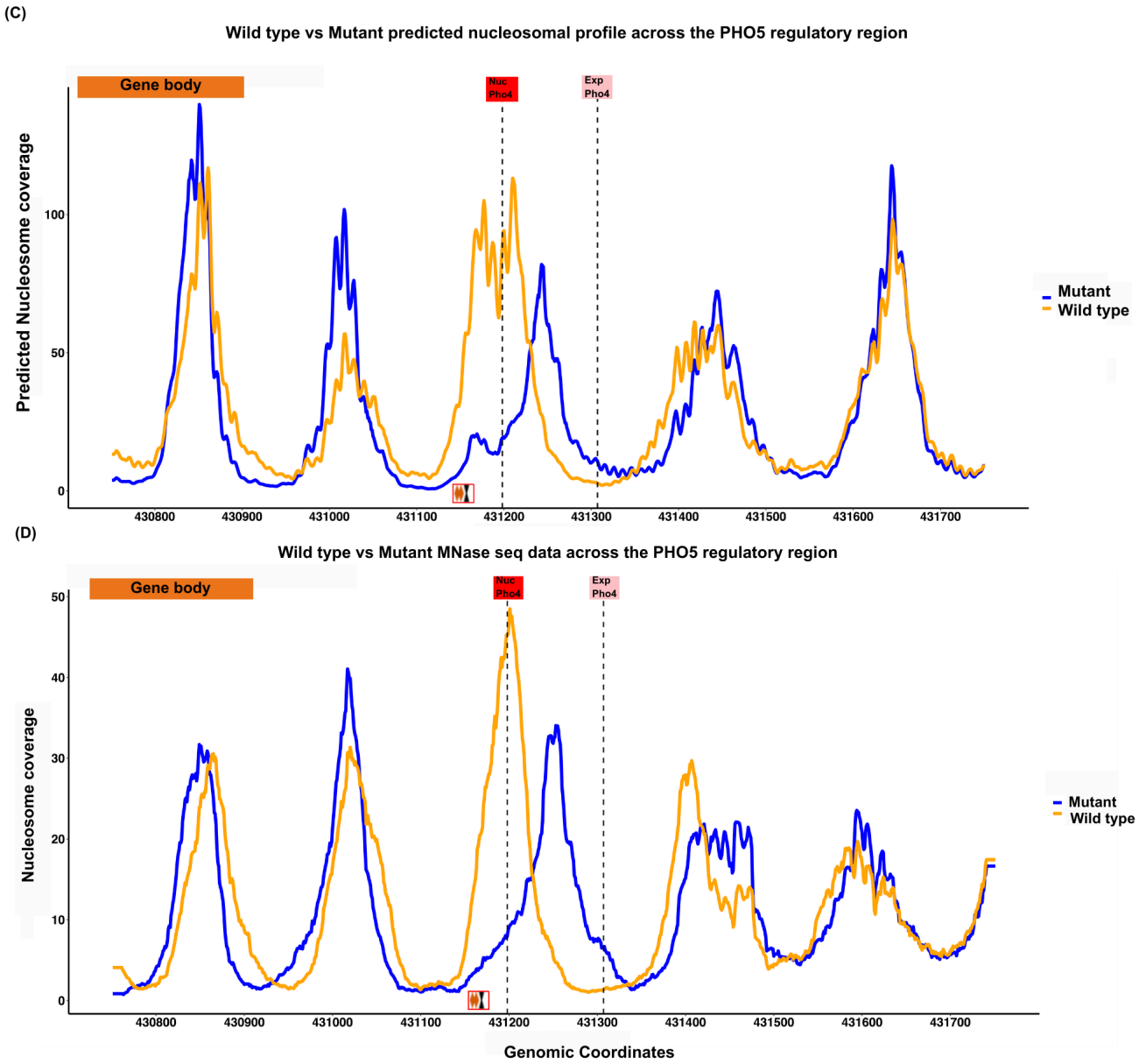


Figure 3.3.3| Figure 3.3.3 A is an illustration demonstrating the wild type and desired nucleosomal conformations in the *PHO5* regulatory region adapted from Rajkumar *et al.,* 2013. In figure 3.3.3 B the y-axis shows the model-determined contribution scores whereas the x-axis shows the genomic coordinates. The plot shows the point mutations proposed by the model and the contribution score associated with these mutations.The wedges represents a base substitution, color coded according to the base formed post substitution. The black hourglass represents a deletion. In figure 3.3.3 C, the y-axis shows the nucleosome coverage and the x-axis shows the genomic coordinates. The plot shows the model predicted MNase-seq profile from wild type and mutant strains, in a 900 bp window near the *PHO5* locus. The CRISPR importance score and model prediction data has been generated by Charles McAnany. In figure 3.3.3 D, the y-axis shows the nucleosome coverage and the x-axis shows the genomic coordinates. Plot shows the MNase-seq profile from wild type and mutant strains, in a 900 bp window near the *PHO5* locus.

3.4 MS2-MCP based tagging can be used to visualise and quantify *PHO5* expression

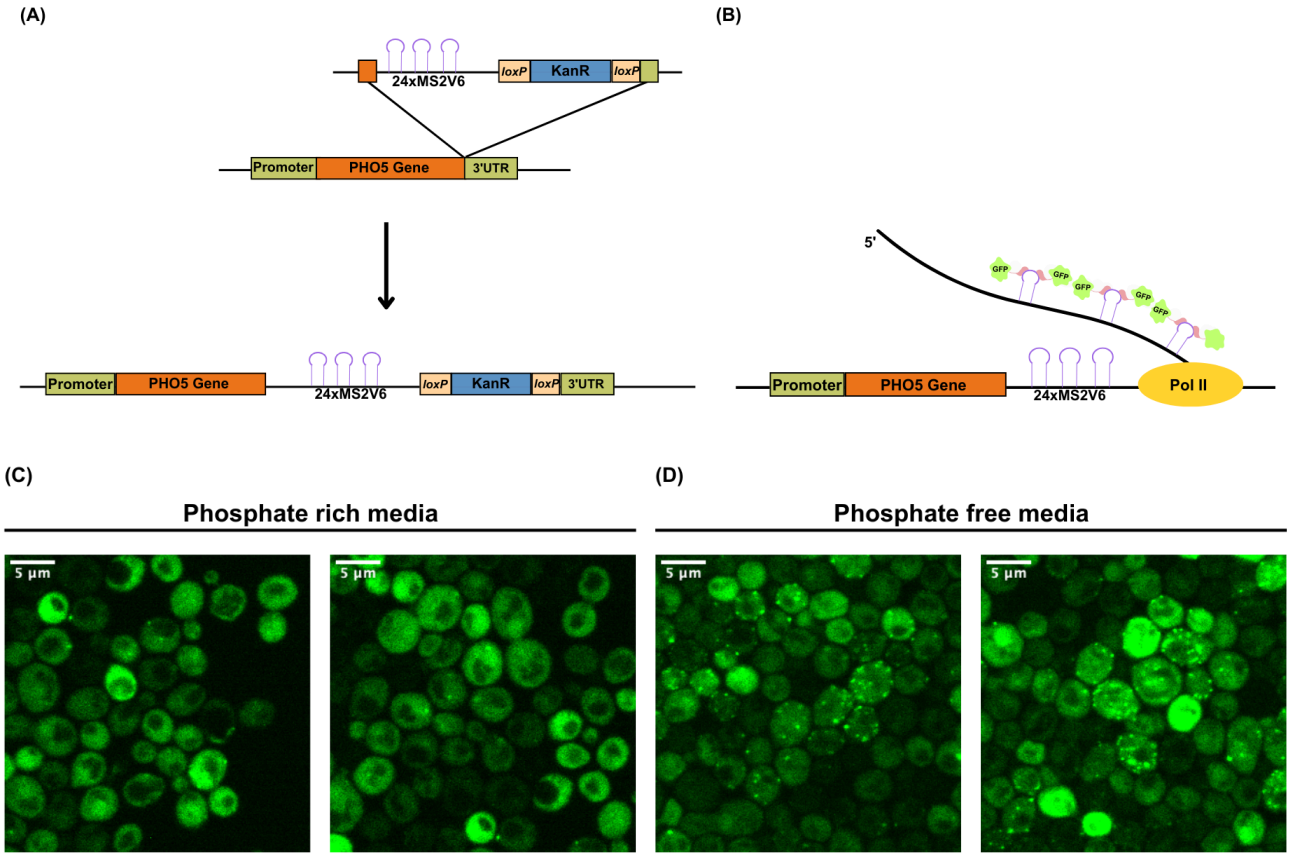
As shown in section 3.3, BPReveal can be used to design synthetic sequences in order to perturb nucleosomes in a desired manner. These mutants give us an opportunity to look at the specific role of nucleosomes in regulating different aspects of gene expression since the *Pho4* binding motifs are intact in all the mutants discussed above. In order to look at the impact of nucleosome perturbation on the expression of the *PHO5* gene, we decided to employ MS2-MCP based endogenous tagging. In this system, the gene of interest is tagged with MS2 bacteriophage derived RNA loops. When the gene is transcribed, the mRNA forms these stem loops leading to the recruitment of the MS2 bacteriophage Coat Protein or MCP, which is co-expressed in the cells. This coat protein is usually fused with fluorescent proteins so the recruitment of MCP to the MS2 stem loops enables us to visualise single mRNA molecules in real time.

MS2 stem loops can be inserted in 5' or the 3'UTR of the gene of interest. Here, we inserted these stem loops near the 3' UTR in order to ensure that the nucleosome positioning in the promoter region is not affected by the stem loops structure. The MS2 loops were introduced through homologous recombination. As shown in figure 3.4 B, a cassette containing 24 MS2 loops and a kanamycin resistance gene was introduced in the wild-type and nucleosome-perturbed mutant strains discussed in the previous sections. This resulted in tagging of the *PHO5* gene with the MS2 loops. The wild type and mutant strains tagged with MS2 loops were then transformed with the MCP-GFP plasmid in order to visualise *PHO5* mRNA. With this system in place, MCP-GFP was expected to bind the MS2 stem loops when PHO5 is expressed as shown in figure 3.4 B. This binding would lead to the formation of puncta, which can be quantified to investigate changes in gene expression in the wild-type and nucleosome-perturbed mutant strains.

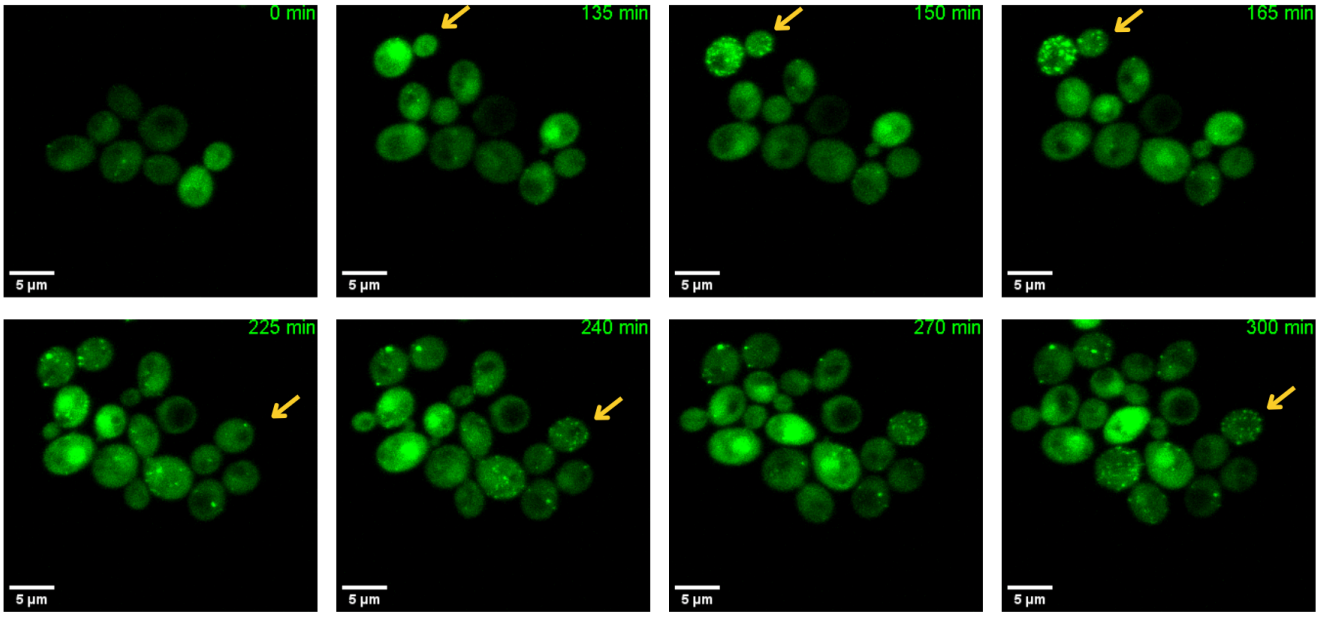
As described in the introduction, *PHO5* is a stress response gene which is expressed when cells are grown under phosphate-starved conditions. In order to check whether the MS2-MCP system is working as expected, we cultured wild-type MS2-tagged strains in phosphate-rich and phosphate-free media. The images in figure 3.4 C show that when cells are grown in phosphate-rich media, the GFP signal is diffused and we see around 1-2 puncta. When the cells are cultured in phosphate-free media, we start seeing multiple puncta representing *PHO5* mRNA molecules, as shown in Figure 3.4 D. This indicates that the MS2-based tagging approach is working since *PHO5* is supposed to be induced on phosphate starvation, resulting in more *PHO5* transcripts and hence more GFP puncta.

In order to look at this in more detail, a time lapse experiment was performed where the cells were grown in selective phosphate-rich media for 60 minutes, followed by growth in phosphate-free media for 4 hours. Images were acquired every 15 minutes. Based on the time lapse data, we first start seeing the appearance of the puncta around 1.5 hours post starvation, as shown in figure 3.4 E. The figure also shows that the expression of *PHO5* is variable as different cells are undergoing the transcription burst at different time points. In figure 3.4 F, the number of MS2-MCP puncta detected per cell across different time points is shown, revealing a gradual increase of puncta detected over time. Figure 3.4 G shows the mean number of puncta detected for every time point, once again revealing a gradual increase in puncta. This is in agreement with the general expression pattern of *PHO5,* where we see a gradual increase in *PHO5* expression with time upon phosphate starvation due to an increase in the concentration of nuclear Pho4 and the resultant chromatin remodeling.

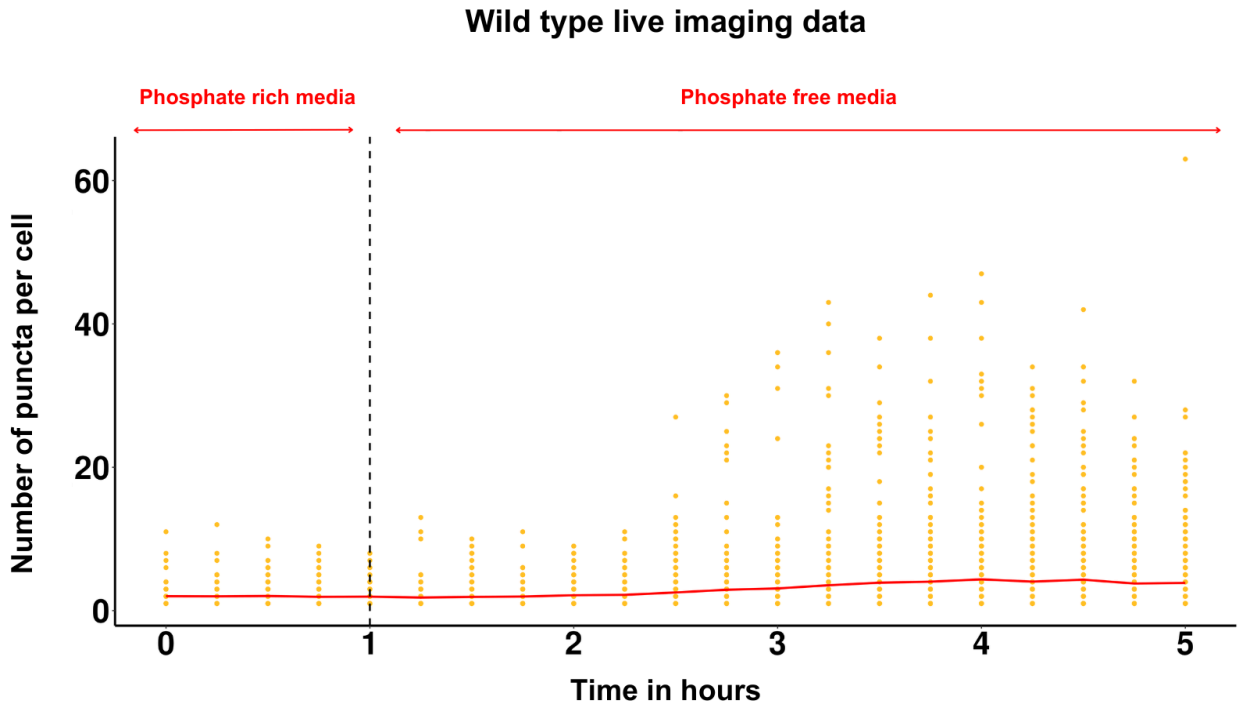
Overall, figure 3.4 shows that the quantification of *PHO5* expression using the MS2-MCP system is capturing the expected trend of *PHO5* expression. Since the system is working, we can use it to characterise changes in *PHO5* expression upon perturbing nucleosomes.



(E)


0 min
135 min
150 min
165 min
225 min
240 min
270 min
300 min
5 µm

(F)


Wild type live imaging data
Phosphate rich media
Phosphate free media
Number of puncta per cell
0
20
40
60
0
1
2
3
4
5
Time in hours

(G)

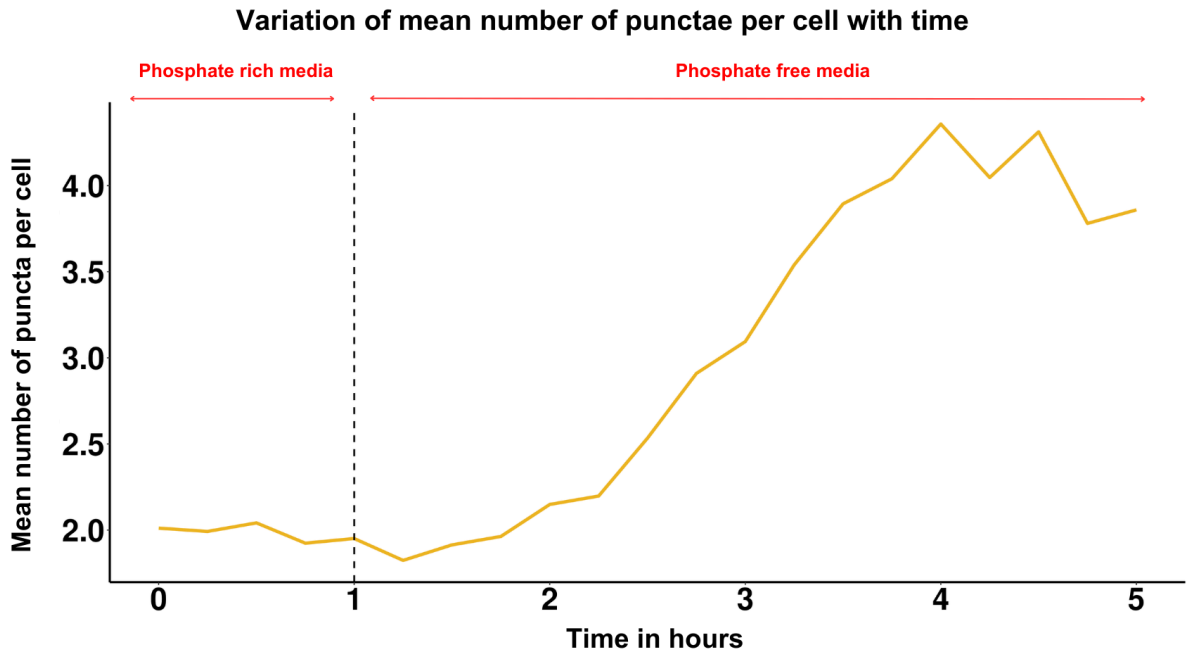


Figure 3.4| Figure 3.4 A is an illustration showing the insertion of the MS2 cassette at the *PHO5* locus. This schematic has been adapted from Tutucci *et al.,* 2018. Figure 3.4 B is a schematic which shows the interaction between MS2 loops and MCP-GFP protein when PHO5 is transcribed. This schematic has been adapted from Hoppe *et al.,* 2021. The panel in figure 3.4 C,D shows the images of the wild type strains cultured in phosphate rich and phosphate free media respectively. Z-stacks were acquired in each case and the panel is a max projection derived from the Z-stacks. The green signal is from the expression of MCP-GFP protein and the punctas in figure 3.4 D represent the transcribed *PHO5* mRNAs. The panel in figure 3.4 E represents some time points from a 5 hour time lapse. Individual time points have been mentioned on each image. The green signal is from MCP-GFP whereas the punctas represent transcribed *PHO5* mRNAs.The data shown in figure 3.4 C and D was generated with Tom Kleist. The dot plot in figure 3.4 F shows the live imaging data from the time lapse. Each dot in the plot represents a cell. The x-axis shows the time in hours and the y-axis shows the number of punctae detected per cell. The cells which did not show any puncta have not been shown in the plot. The line plot in figure 3.4 G, represents the mean number of punctae detected per cell for a particular time point. The x-axis shows the time in hours and the y-axis shows the mean number of punctae detected per cell. The code which has been used for quantifying the punctae has been developed by Cathy McKinney.

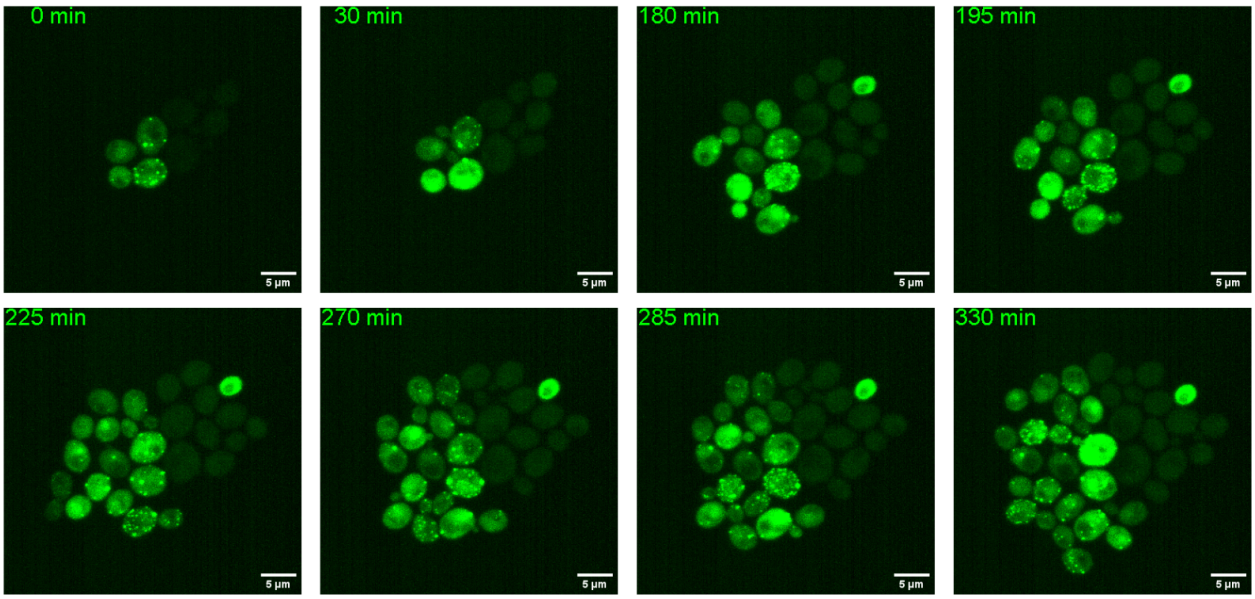
3.5 Nucleosome perturbation in the suppress 1 mutant leads to changes in *PHO5* expression levels

As shown in section 3.4, MS2-based tagging of the *PHO5* gene can be used to visualise and quantify the expression of *PHO5* in real time. Since this system worked well in the wild-type strain, we decided to use this system to characterize the changes in *PHO5* expression in the mutants where the nucleosomes are perturbed. We first decided to look at the suppress 1 mutant described in section 3.3.3. The strain was grown in phosphate-rich media for 1 hour, followed by growth in phosphate-free media for 5 hours, while time lapse imaging was performed by acquiring images every 15 minutes.

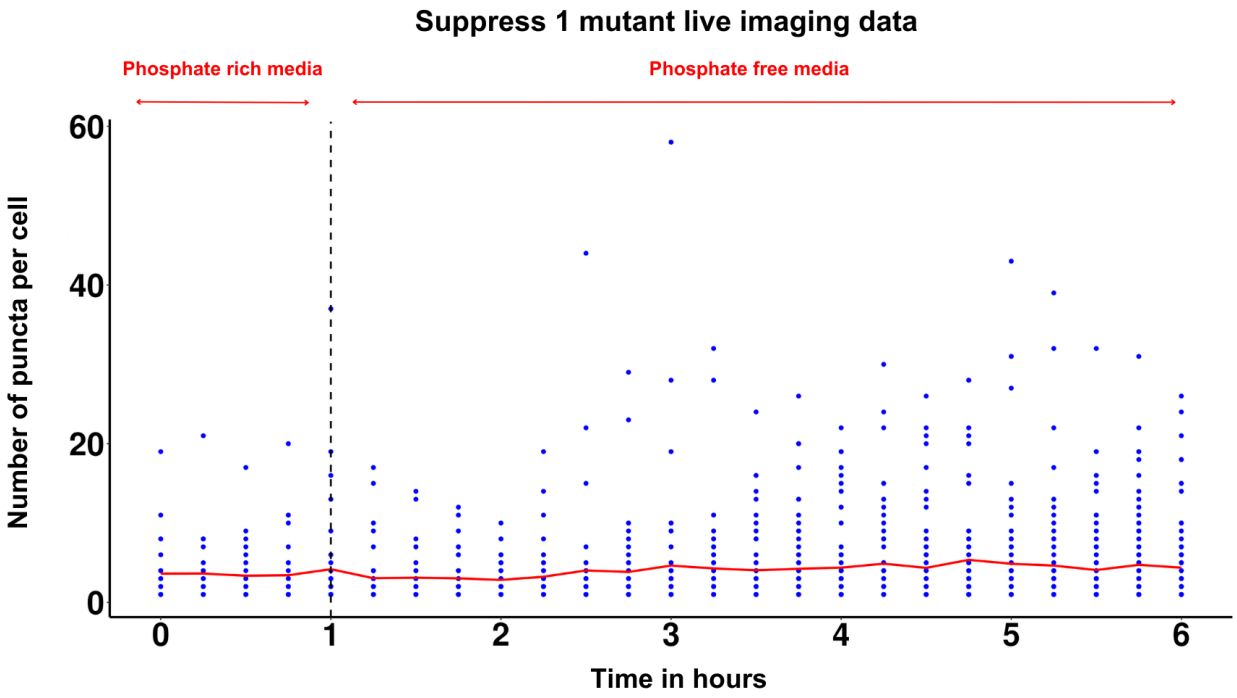
As shown in Figure 3.5 A, we detected some MS2-MCP puncta in the mutant strain even when the cells were grown in phosphate-rich media. This suggests that the exposure of both the binding sites might cause an increase in leaky expression of *PHO5*. Figure 3.5 B also shows that in the case of the mutant, we are starting from a different baseline as we can detect MS2-MCP puncta even in the absence of phosphate starvation. The figures also show that when the cells are grown under phosphate-starved conditions, there is a further increase in the levels of *PHO5* expression. This can be attributed to the increase in the nuclear localisation of Pho4 on phosphate starvation.

Figure 3.5 C, also confirm an increase in the mean number of MS2-MCP puncta per cell when cells are grown in phosphate starved conditions.The line plot in figure 3.5 C is a bit noisy as the number of imaged cells from the mutant strain were lower than those from wild-type. Imaging data from more replicate experiments is likely to reduce the noise, revealing the trend in PHO5 expression more clearly. When we compare the expression of *PHO5* in wild-type and suppress 1 mutant as shown in figure 3.5 D, we can see that there is a gradual increase in the levels of *PHO5* over time in the wild-type, whereas the increase in the expression levels seems more sudden in the mutant. This could be because the exposure of both *Pho4* binding sites might make it easier for Pho4 to bind its site upon nuclear localisation, leading to an increase in the expression of *PHO5*.

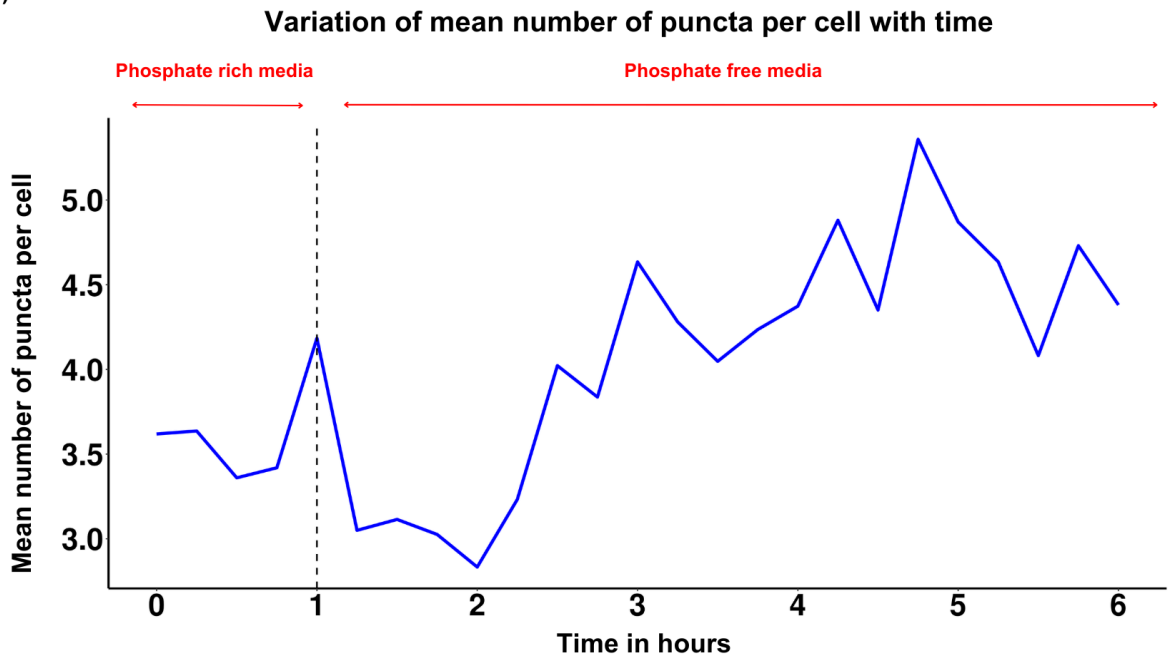
(A)


0 min
30 min
180 min
195 min
225 min
270 min
285 min
330 min
5 µm

(B)


Suppress 1 mutant live imaging data
Phosphate rich media
Phosphate free media
Number of puncta per cell
Time in hours
0
20
40
60
0
1
2
3
4
5
6

(C)


Variation of mean number of puncta per cell with time
Phosphate rich media
Phosphate free media
Mean number of puncta per cell
5.0
4.5
4.0
3.5
3.0
0
1
2
3
4
5
6
Time in hours

(D)

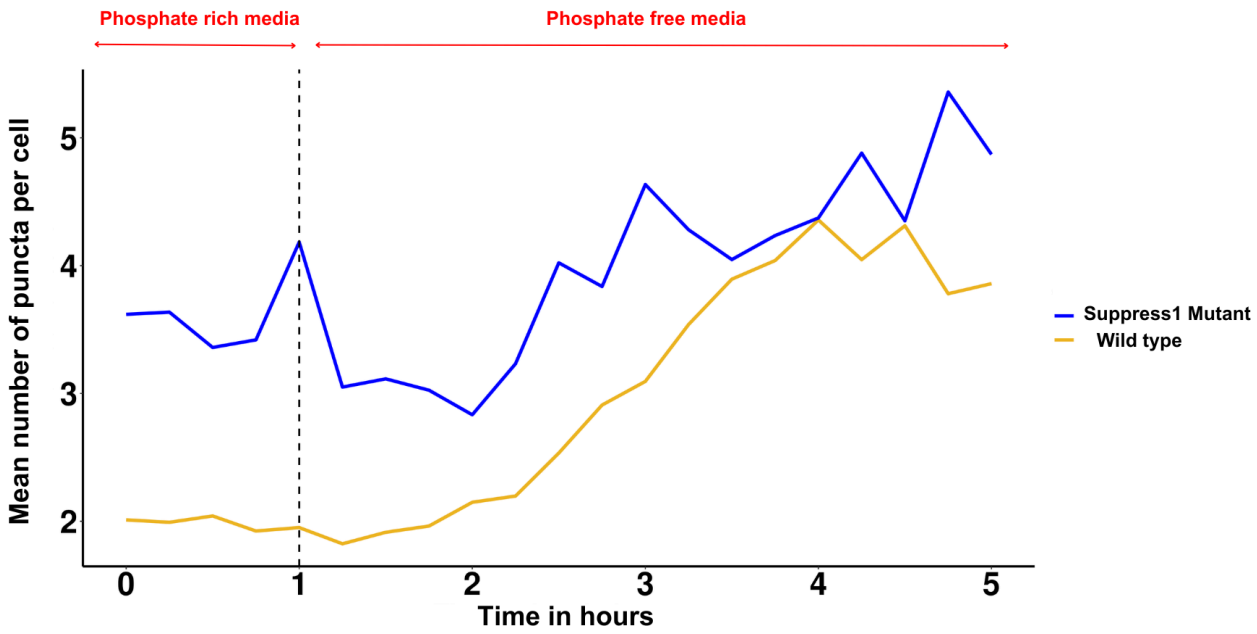

Variation of mean number of puncta per cell with time in Wild type vs Suppress 1
Phosphate rich media
Phosphate free media
Mean number of puncta per cell
5
4
3
2
0
1
2
3
4
5
Time in hours
Suppress1 Mutant
Wild type

Figure 3.5|The panel in figure 3.5 A shows the images of the suppress 1 mutant strain cultured in phosphate rich and phosphate free media respectively. Z-stacks were acquired in each case and the panel is a max projection derived from the Z-stacks. The green signal is from the expression of MCP-GFP protein and the punctae represent the transcribed *PHO5* mRNAs. The panel in figure 3.5 B represents some time points from a 6 hour time lapse. Individual time points have been mentioned on each image. The green signal is from MCP-GFP whereas the punctae represent transcribed *PHO5* mRNAs.The data shown in figure 3.5 A and B was generated with Tom Kleist. The dot plot in figure 3.5 C shows the live imaging data from the time lapse. Each dot in the plot represents a cell. The x-axis shows the time in hours and the y-axis shows the number of punctae detected per cell. The cells which did not show any punctae have not been shown in the plot. The line plot in figure 3.4 D, represents the mean number of punctae detected per cell for a particular time point. The x-axis shows the time in hours and the y-axis shows the mean number of punctae detected per cell. The line plot in figure 3.5 E, shows the mean number of punctae per cell for a particular time point in the wild type and the suppress 1 mutant strain. The yellow line represents the data from wild type and the blue line represents the data from the suppress 1 mutant strain. The code which has been used for quantifying the punctae has been developed by Cathy McKinney.

Chapter 4
Discussion

In this study we showed that BPReveal, a convolutional neural network, can be used to learn the sequences important for determining nucleosome positioning in *S. cerevisiae*. Studies in the past have shown that polyA sequences are important in creating NDRs but these studies were usually limited to polyA sequences of a certain length (consecutive number of Adenine bases). The high prevalence of polyA sequences in the genome, combined with multiple permutations and combinations with respect to the number of consecutive Adenine bases, makes it difficult to consider all possible sequences.The deep learning approach used in this study helped solve this problem as we were not restricted to polyA sequences of a certain length. This enabled us to identify and map polyAs that play a crucial role in determining nucleosome positioning genome- wide and in an unconstrained and unbiased manner.

PolyA sequences are known to affect nucleosome positioning due to their biophysical properties and interaction with chromatin remodelers. Here we looked at 4 chromatin remodelers Isw1, Chd1, Swr1 and Rsc. As shown in figure 1.1 B and 1.2 A, the NDR length decreases in 1.1 kb window across mapped polyA motifs in the RSC mutant. We also saw that contribution scores of the polyA motifs generated by BPReveal correlate with the extent of perturbation in nucleosome positioning as summarised in figure 1.2 G. This shows that the model indeed learns the rules guiding nucleosome positioning and can hence be used to study nucleosome positioning in *S.cerevisiae* and possibly other model systems where these rules are not well understood. We also saw that the regions with wider NDRs and higher extent of nucleosome perturbation in the RSC mutant often seem to have more number of model mapped polyA sequences in the 400 bp window as shown in figure 1.2 H. This suggests that there could be some cooperativity between multiple polyA sequences which leads to different NDR widths and differential RSC activity.

Figure 1.2 D, F showed that there is weaker phasing in the Isw1Δ and Chd1Δ mutants in the 1.1 kb window across mapped polyA motifs. Weaker nucleosome positioning is a global phenomenon which has been reported for Isw1Δ and Chd1Δ mutants in the past. Since the nucleosomes flanking the NDR or the NDR itself are not affected, it is difficult to comment on whether the activity of these remodelers is associated with polyA sequences or not. The correlation between contribution scores and extent of nucleosome perturbation, which was seen in the case of RSC mutant, was not seen in the case of Isw1Δ and Chd1Δ, further suggesting that their activity is probably not dependent on polyA sequences. However, BPReveal models can be trained on Isw1Δ

and Chd1Δ mutant MNase-seq data to discover other sequences which could be mediating the roles of these remodelers.

Previous studies have shown that nucleosome positioning plays an important role in regulating different aspects of gene expression. In the case of some special loci, nucleosome positioning is seen to influence the capability of a gene to alter its expression based on changing environmental conditions. The role of nucleosome positioning in regulating gene expression has been best demonstrated at the *PHO5* locus where 4 well positioned nucleosomes have been shown to be instrumental, but there are very few such loci where this has been characterised. This is because perturbing the position of 1 or 2 nucleosomes at a locus in a controlled manner is complicated. A nucleosome is a complex associated with 147 bp of wrapped DNA and variable length of linker DNA, so it is difficult to predict the effect of mutating DNA sequences on the positioning of that nucleosome and other nucleosomes in the vicinity. Since BPreveal can accurately predict MNase-seq profiles in a genome wide fashion, we decided to use it as a tool to design synthetic sequences in order to perturb nucleosome positioning in a directed manner.

In order to test whether BPReveal can be used to design sequences which lead to a desired nucleosome perturbation, we decided to use *PHO5* as the model locus to achieve two configurations, 1) both *Pho4* binding motifs are exposed and 2) both *Pho4* binding motifs are covered. As shown in sections 3.3.1 and 3.3.2, one out of the two sequences which were tested achieved the desired nucleosomal conformation where both *Pho4* motifs are covered. However, the perturbation in the positioning of the nucleosomes in the entire 1kb window near the *PHO5* gene body was higher than expected. As shown in section 3.3.3, the sequence tested to achieve the nucleosomal conformation where both *Pho4* motifs are exposed worked very well. The resultant MNase-seq profile showed that there was a shift in the -2 nucleosome such that both the *Pho4* motifs were exposed. The positioning of all the surrounding nucleosomes was almost unperturbed and this was achieved by just making 3 point mutations. The results of the three tested sequences hint that the model-designed sequences work better when the site of mutation is closer to the nucleosome we want to perturb. They also suggest that lowering the number of allowed point mutations can help in making sure that the positioning of the other nucleosomes is not affected.

In order to look at the effect of nucleosome perturbations on the expression of the *PHO5* gene, we decided to employ the MS2-MCP based mRNA detection system. This system can allow us to look at different aspects of gene expression like induction kinetics, transcription burst, steady state expression levels and so on. When we imaged

the wt BY4741 PHO5 24xMS2V6 MCPNLS strain in phosphate-rich and phosphate-free conditions, we saw an increase in the number of puncta appearing over time,as shown in figure 3.4 E. We could also visualise transcriptional bursts where a lot of puncta would appear at certain time points. The quantification of the signal showed that the *PHO5* mRNAs quantified using this approach follows similar trends as other reporter assays, indicating that this system can be used to further characterize *PHO5* expression in mutants where nucleosomes are perturbed.

While the overall trend in expression was as expected, there are a few unexpected observations which need to be addressed. Ideally, there should not be any puncta when the cells are grown in phosphate-rich media, but during our imaging, we saw that cells tend to have around 2-3 puncta even when grown in phosphate-rich conditions. Based on the RNA-seq data from untagged wild-type strain cultured in YPD, it looks like there is indeed a basal level of expression in the absence of phosphate starvation. Such basal levels of expression are consistent with the puncta that we saw in cells grown in phosphate-rich media.

Since the trend of *PHO5* expression in wild-type was as expected, the next step was to characterise the effect of perturbing the nucleosomes on *PHO5* expression. We decided to start with the suppress 1 mutant, where both *Pho4* motifs are exposed. When we imaged the cells growing in phosphate rich conditions, we saw that these cells had more puncta as compared to the wild-type. This suggests that exposure of both Pho4 sites might be causing an increase in the leakiness of *PHO5* expression in phosphate-rich conditions. The time lapse data in figure 3.5 C,D show that there is a sudden increase in the levels of *PHO5* expression on phosphate starvation in the suppress 1 mutants. Figure 3.5 E also shows that the levels of *PHO5* increase gradually in the wildtype strain, whereas in the suppress 1 mutant strain, the increase is more sudden. The pattern of *PHO5* expression in the mutant suggests that on phosphate starvation, the exposure of both sites is making it easier for Pho4 to bind its sites, resulting in increased *PHO5* expression.

Another possible factor which could be contributing towards the increase in *PHO5* expression in the suppress 1 mutant is the introduction of the *Abf1* motif. However, Abf1 is not a classical transcriptional activator, but is involved in regulating chromatin architecture by creating a NDR and positioning nucleosomes. Previous studies have associated Abf1 with both repression and activation in a context dependent manner (Miyake *et al.,* 2004). This suggests that the introduction of an *Abf1* motif in the coverall and the suppress mutants is not activating per se but that its effect may depend on whether it covers or exposes a Pho4 site. Analyzing these mutants in more detail could

therefore provide an opportunity to corroborate such nucleosome-mediated mechanisms of gene regulation.

Previous studies and our results in section 3.1, 3.2 show that polyAs are also involved in creating NDRs and positioning nucleosomes. This similarity in function with Abf1 suggests that polyA sequences could also play a similar role in gene regulation, which has been missed due to their complex nature. We can leverage BPReveal to design mutations such that a polyA is replaced by an *Abf1* motif and vice versa to further explore their role in gene regulation.

The coverall1 mutant described in section 3.3.1 also gives us a chance to understand how *PHO5* expression gets affected when the strength of nucleosome positioning decreases. This is particularly important from a mammalian perspective, as mammals exhibit tight gene regulation despite the absence of well positioned nucleosomes. This clear gap in understanding presents an opportunity to characterise previously unknown mechanisms by which fuzzy nucleosomes could be involved in gene regulation.

Chapter 5
References

Almer,A., Rudolph,H., Hinnen,A. and Ho ¨rz,W. (1986) Removal of positioned nucleosomes from the yeast *PHO5* promoter upon *PHO5* induction releases additional upstream activating DNA elements. *EMBO J.*, 5, 2689–2696.

Anderson JD, Widom J. Poly(dA-dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites. Mol Cell Biol. 2001 Jun;21(11):3830-9.

Anderson JD, Widom J. Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites. J Mol Biol. 2000 Mar 3;296(4):979-87.

Avsec, Ž., Weilert, M., Shrikumar, A. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* 53, 354–366 (2021)

Begley, V. et al. Xrn1 influence on gene transcription results from the combination of general effects on elongating RNA pol II and gene-specific chromatin configuration. RNA Biology 18, 1310–1323 (2021).

Bergman,L.W. and Kramer,R.A. (1983) Modulation of chromatin structure associated with derepression of the acid phosphatase gene of Saccharomyces cerevisiae. *J. Biol. Chem.*, 258, 7223–7227.

Brennan KJ, Weilert M, Krueger S, Pampari A, Liu HY, Yang AWH, Morrison JA, Hughes TR, Rushlow CA, Kundaje A, Zeitlinger J. Chromatin accessibility in the Drosophila embryo is determined by transcription factor pioneering and enhancer activation. Dev Cell. 2023 Oct 9;58(19):1898-1916.e9.

Brogaard, K., Xi, L., Wang, JP. *et al.* A map of nucleosome positions in yeast at base-pair resolution. *Nature* 486, 496–501 (2012).

Chereji RV, Ocampo J, Clark DJ. MNase-Sensitive Complexes in Yeast: Nucleosomes and Non-histone Barriers. Mol Cell. 2017 Feb 2;65(3):565-577.e3.

Cui F, Cole HA, Clark DJ, Zhurkin VB. Transcriptional activation of yeast genes disrupts intragenic nucleosome phasing. Nucleic Acids Res. 2012 Nov;40(21):10753-64.

Dingwall C, Lomonossoff GP, Laskey RA. High sequence specificity of micrococcal nuclease. Nucleic Acids Res. 1981 Jun 25;9(12):2659-73.

Drew HR, Travers AA. DNA bending and its relation to nucleosome positioning. J Mol Biol. 1985 Dec 20;186(4):773-90.

Ganapathi M, Palumbo MJ, Ansari SA, He Q, Tsui K, Nislow C, Morse RH. Extensive role of the general regulatory factors, Abf1 and Rap1, in determining genome-wide chromatin structure in budding yeast. Nucleic Acids Res. 2011 Mar;39(6):2032-44

Ganguli D, Chereji RV, Iben JR, Cole HA, Clark DJ. RSC-dependent constructive and destructive interference between opposing arrays of phased nucleosomes in yeast. Genome Res. 2014 Oct;24(10):1637-49.

Gkikopoulos T, Schofield P, Singh V, Pinskaya M, Mellor J, Smolle M, Workman JL, Barton GJ, Owen-Hughes T. A role for Snf2-related nucleosome-spacing enzymes in genome-wide nucleosome organization. Science. 2011 Sep 23;333(6050):1758-60.

Hartley PD, Madhani HD. Mechanisms that specify promoter nucleosome location and identity. Cell. 2009 May 1;137(3):445-58.

Hoppe C, Ashe HL. Live imaging and quantitation of nascent transcription using the MS2/MCP system in the *Drosophila* embryo. STAR Protoc. 2021 Mar 18;2(1):100379.

Hughes AL, Rando OJ. Mechanisms underlying nucleosome positioning in vivo. Annu Rev Biophys. 2014;43:41-63.

Lai WKM, Pugh BF. Understanding nucleosome dynamics and their links to gene expression and DNA replication. Nat Rev Mol Cell Biol. 2017 Sep;18(9):548-562.

Jansen A, Verstrepen KJ.2011.Nucleosome Positioning in Saccharomyces cerevisiae. Microbiol Mol Biol Rev75.

Jiang C, Pugh BF. A compiled and systematic reference map of nucleosome positions across the Saccharomyces cerevisiae genome. Genome Biol. 2009;10(10):R109.

Kaneko Y, Toh-e A, Oshima Y (1982) Identification of the genetic locus for the structural gene and a new regulatory gene for the synthesis of repressible alkaline phosphatase in Saccharomyces cerevisiae. Mol Cell Biol 2:127–137

Kingston RE, Narlikar GJ. ATP-dependent remodeling and acetylation as regulators of chromatin fluidity. Genes Dev. 1999 Sep 15;13(18):2339-52.

Komeili A, O'Shea EK. Roles of phosphorylation sites in regulating activity of the transcription factor Pho4. Science. 1999 May 7;284(5416):977-80.

Korber P, Barbaric S. The yeast PHO5 promoter: from single locus to systems biology of a paradigm for gene regulation through chromatin. Nucleic Acids Res. 2014;42(17):10888-902.

Kornberg RD, Stryer L. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. Nucleic Acids Res. 1988 Jul 25;16(14A):6677-90.

Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. Science. 1974 May 24;184(4139):868-71.

Krietenstein N, Wal M, Watanabe S, Park B, Peterson CL, Pugh BF, Korber P. Genomic Nucleosome Organization Reconstituted with Pure Proteins. Cell. 2016 Oct 20;167(3):709-721.e12

Kubik S, Bruzzone MJ, Jacquet P, Falcone JL, Rougemont J, Shore D. Nucleosome Stability Distinguishes Two Different Promoter Types at All Protein-Coding Genes in Yeast. Mol Cell. 2015 Nov 5;60(3):422-34.

Kubik, S. et al. Sequence-directed action of RSC remodeler and general regulatory factors modulates +1 nucleosome position to facilitate transcription. *Mol. Cell* 71, 89–102.e5 (2018)

Lam FH, Steger DJ, O'Shea EK. Chromatin decouples promoter threshold from dynamic range. Nature. 2008 May 8;453(7192):246-50

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

Luger, K., Mäder, A., Richmond, R. *et al.* Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389, 251–260 (1997).

Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res. 2008 Jul;18(7):1073-83.

McKnight LE, Crandall JG, Bailey TB, Banks OGB, Orlandi KN, Truong VN, Donovan DA, Waddell GL, Wiles ET, Hansen SD, Selker EU, McKnight JN. Rapid and inexpensive preparation of genome-wide nucleosome footprints from model and non-model organisms. STAR Protoc. 2021 May 18;2(2):100486.

Mieczkowski, J., Cook, A., Bowman, S. *et al.* MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nat Commun* 7, 11485 (2016).

Miyake T, Reese J, Loch CM, Auble DT, Li R. Genome-wide analysis of ARS (autonomously replicating sequence) binding factor 1 (Abf1p)-mediated transcriptional regulation in Saccharomyces cerevisiae. J Biol Chem. 2004 Aug 13;279(33):34865-72

Mizuguchi G, Shen X, Landry J, Wu WH, Sen S, Wu C. ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. Science. 2004 Jan 16;303(5656):343-8.

O'Neill EM, Kaffman A, Jolly ER, O'Shea EK. Regulation of PHO4 nuclear localization by the PHO80-PHO85 cyclin-CDK complex. Science. 1996 Jan 12;271(5246):209-12.

Ocampo J, Chereji RV, Eriksson PR, Clark DJ. Contrasting roles of the RSC and ISW1/CHD1 chromatin remodelers in RNA polymerase II elongation and termination. Genome Res. 2019 Mar;29(3):407-417.

Ocampo J, Chereji RV, Eriksson PR, Clark DJ. The ISW1 and CHD1 ATP-dependent chromatin remodelers compete to set nucleosome spacing in vivo. Nucleic Acids Res. 2016 Jun 2;44(10):4625-35.

Oudet P, Gross-Bellard M, Chambon P. Electron microscopic and biochemical evidence that chromatin structure is a repeating unit. Cell. 1975 Apr;4(4):281-300.

Pampari, A. et al. ChromBPNet: bias factor-ized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor foot-prints and regulatory variants. BioRxiv(2025).

Prajapati HK, Ocampo J, Clark DJ. Interplay among ATP-Dependent Chromatin Remodelers Determines Chromatin Organisation in Yeast. Biology (Basel). 2020 Jul 25;9(8):190.

Rajkumar AS, Dénervaud N, Maerkl SJ. Mapping the fine structure of a eukaryotic promoter input-output function. Nat Genet. 2013 Oct;45(10):1207-15.

Raveh-Sadka, T., Levo, M., Shabi, U. *et al.* Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat Genet* 44, 743–750 (2012)

Rudolph,H. and Hinnen,A. (1987) Proc. Natl Acad. Sci. USA, 84,1340-1344.

Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang JP, Widom J. A genomic code for nucleosome positioning. Nature. 2006 Aug 17;442(7104):772-8.

Shrikumar, A., Greenside, P. & Kundaje, A.Learning important features through propagating activation differences. arXiv (2017).

Struhl K, Segal E. Determinants of nucleosome positioning. Nat Struct Mol Biol. 2013 Mar;20(3):267-73.

Thoma F, Koller T, Klug A. Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. J Cell Biol. 1979 Nov;83(2 Pt 1):403-27.

Tirosh I, Barkai N. Two strategies for gene regulation by promoter nucleosomes. Genome Res. 2008 Jul;18(7):1084-91

Tutucci, E., Vera, M. & Singer, R.H. Single-mRNA detection in living *S. cerevisiae* using a re-engineered MS2 system. *Nat Protoc* 13, 2268–2296 (2018)

Tutucci, E., Vera, M., Biswas, J. *et al.* An improved MS2 system for accurate reporting of the mRNA life cycle. *Nat Methods* 15, 81–89 (2018).

Venter, U., Svaren, J., Schmitz, J., Schmid, A. & Hörz, W. A nucleosome precludes binding of the transcription factor Pho4 *in vivo* to a critical target site in the PHO5 promoter. *EMBO J.* 13, 8 (1994).

Vogel K, Hörz W, Hinnen A. The two positively acting regulatory proteins PHO2 and PHO4 physically interact with PHO5 upstream activation regions. Mol Cell Biol. 1989 May;9(5):2050-7.

Winger J, Bowman GD. The Sequence of Nucleosomal DNA Modulates Sliding by the Chd1 Chromatin Remodeler. J Mol Biol. 2017 Mar 24;429(6):808-822.

Zhang Z, Wippo CJ, Wal M, Ward E, Korber P, Pugh BF. A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. Science. 2011 May 20;332(6032):977-80. doi: 10.1126/science

Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, Newburger DE, Saulrieta K, Smith Z, Shah MV, Radhakrishnan M, Philippakis AA, Hu Y, De Masi F, Pacek M, Rolfs A, Murthy T, Labaer J, Bulyk ML. High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res. 2009 Apr;19(4):556-66.