Computational Study of Nucleosome Positioning Sequence Patterns and the Effects of the Nucleosome Positioning on the Availability of the Transcription Factor Binding Sites in Study Systems

by

Doo Seok Yang

Dissertation

Presented to the

Faculty of the Graduate and Postdoctoral Studies

in Partial Fulfilment

of the Requirements

for the Degree of

Doctor of Philosophy

Department of Biochemistry, Microbiology & Immunology

Faculty of Medicine

University of Ottawa

Dedication

To my grandmother, who have done everything

and would’ve done anything for me.

I miss your lunch.

Abstract

Nucleosomes, the primary unit of chromatin structure, are positioned either statistically or specifically. The statistical positioning denotes the arbitrary positioning of nucleosomes on DNA agreeing with the nucleosome's broad coverage of the genome—however, there is evidence that nucleosomes are also positioned specifically at controlled positions. DNA sequences determine the specific nucleosome positions, and the presence or depletion of nucleosomes affects the availability of the DNA region to other proteins. The DNA sequences of H2A and H2A.Z nucleosomes in Drosophila were analysed in search of nucleosome positioning patterns. Dinucleotide patterns with 10 bp periodicity were identified from the DNA sequences of H2A nucleosomes. Compared with the yeast patterns, the Drosophila patterns had the same periodicity but different dinucleotides near the dyad, which was related to the different H3 structure between them. The nucleosome positioning patterns from the H2A.Z nucleosomes implied the specific histone-DNA interaction as a result of the deviations of the patterns where the different amino acids of H2A and H2A.Z interact with the DNA. The Ly49 gene cluster was selected as a model system to study the interplay between nucleosomes and transcription factors. Ly49 proteins, the surface receptors on NK cells, display variegated expression patterns, and the bidirectional promoter Pro-1 is known as a key determinant of the stochastic expression of each Ly49 gene. The systematic analysis of nucleosome positions based on the genome sequences in the Ly49 gene cluster revealed that the repressing Pro-1 reverse promoters are open, while the activating forward Pro-1 promoters were covered by nucleosomes. Furthermore, specific nucleosome positions covered transcription factor

binding sites. The covered factor binding sites were further examined by their periodic appearances on the nucleosome-covered sequences, which revealed the accessibility to the sites. The sequence analysis predicted that the regulation by the transcription factor AML-1 would be sensitive to the nucleosome coverage; the prediction was confirmed by cell line experiments. The 10 bp periodic nucleosome positioning patterns interact with histones specifically. The long nucleosome positioning patterns coexist with the short sequence motifs for transcription factor binding sites adding another layer of the control to the transcriptional regulation.

Acknowledgements

I would like to thank my supervisor, Dr Ilya Ioshikhes for accepting me to study in his lab. He guided me with patience so that I could finish this journey successfully.

I also thank Dr Andrew Makrigiannis, who encouraged me when I was almost worn out and was willing to offer possible help. I wish all the best where you are.

I would like to thank my lab colleagues: Dr Soumyadeep Nandi and Dr Sergey Hosid. You are encouragements to me. I could free my mind and refresh during the otherwise isolated times. Other people who have worked together in the lab, I remember, and you were refreshing air to me. Moreover, late Mr Slav, I could've learned more from you if we could hang around little longer. The Kefir starter always reminds me of you.

My fellow students and BMIGSA, who have gone through this exhausting process, yet always encouraging and challenging to me. Andrew, who is a trustworthy, smart, and language-loving friend. I could not finish this study without your help. Megan, who is a gentle yet strong lady, who lead the BMIGSA with pleasure. I will not forget the famous BMI burger.

I appreciate all the University staff for me to go through this process. Fay, you are a life saver. How many times you guided me and saved me through the technicalities and policies, with much patience. Victoria, you are kind and resourceful to offer me many bits of help.

I could not emphasise more of the support of my family, who are the reasons and the strength I started and finished this journey. My wife Kyeng Whoi, and my beloved children, Eumin and Tiffany. I wanted you to be proud of me. Also, thinking of my family who supports me in full even though live far far away. Mum and Dad, your prayers and sacrifices are the beginning of my journey. Also, mother-in-law and father-

in-law, I could not express enough my thankfulness. Your supports and prayer in these long years were the strength for my family and me. My brother Yoonseok, my sister Yeonwook, and my brother Seungwhan, you have always been on my side. Even though I am away from you, we are always one. Moreover, all my Aunts and Uncles, you were a great support and root for me always.

Finally but not lastly, my LORD, I love you: You strengthened me when I was weak, comforted me when I was distressed, and guided me when I was lost.

Call to me and I will answer you, and will tell you great
and hidden things that you have not known.
Jeremiah 33:3

Table of Contents

List of Abbreviations

AML-1	Acute myeloid leukemia gene, Runt related transcription factor 1, Runx
AP-1	Activator protein 1
ATF-2	Activating transcription factor 2
bp	Base pair
C/EBP	CCAAT-enhancer binding protein
c-ETS-1	E26 avian leukemia oncogene 1
ChIP-Seq	Chromatin immunoprecipitation sequencing
FDR	False discovery rate
GATA-3	GATA binding protein 3
GEO	Gene Expression Omnibus
H2A	Histone H2A
H2A.Z	Histone H2A.Z
HMM	Hidden Markov model
Ik-3	CDK5 and ABL1 enzyme substrate 1
LHS	Left-hand side of an association rule
Lyf-1	IKAROS family zinc finger 1
MZF-1	Myeloid zinc finger 1
NF-AT	Nuclear factor of activated T-cells
NF-kB	Nuclear factor of kappa light polypeptide gene enhancer in B cells 1
NK cell	Natural killer cell
NKT	Natural killer T
NPS	Nucleosome positioning sequence
Oct-1	POU domain, class 2, transcription factor 1

PWM	Position weight matrix
RHS	Right-hand side of an association rule
RMA	MHC class I-positive tumor cell line
Sp1	Trans-acting transcription factor 1
STAT3	Signal transducer and activator of transcription 3
SVM	Support vector machine
Tal-1a	T-cell acute lymphocytic leukemia 1
TATA	TATA box
TCF-1	T cell factor 1
TCRβ	T cell receptor beta chain
TF	Transcription factor
TSS	Transcription start site
ITIM	Immunoreceptor tyrosine-based inhibitory motif
SH2	SRC homology 2
SHP-1	SH2 domain-containing protein tyrosine phosphatase 1
NFR	Nucleosome free region
NDR	Nucleosome depletion region
RMA	RBL-5 (Rauscher virus-induced lymphoma), mutagenized
Klra	Killer cell lectin-like receptor subfamily A
Ly49	T lymphocyte activation marker, or Klra
uNK	Uterine natural killer
KIR	Killer cell Ig-like
NK	Natural killer
MHC-I	Major histocompatibility class I

List of Figures

List of Tables

GENERAL INTRODUCTION

Nucleosome positioning sequences of Drosophila

CHROMATIN STRUCTURE

The wonder of DNA structure is putting the long DNA into a confined space of a nucleus. For example, every human cell contains about 2 m of linear DNA in total, whereas the size of the nucleus is only 5 to 20 μm in diameter (Teif and Bohinc, 2011). Moreover, the folded DNA must remain accessible for various cellular processes, such as replication and transcription. DNA forms into a compact structure called chromatin by wrapping DNA around histones, highly basic proteins that favour binding to the negatively charged DNA. The chromatin structure is established on a repeating unit of the histones and DNA, which had been assumed of eight histone molecules and 200 DNA base pairs (Kornberg, 1974). X-ray structure of the repeating unit, termed nucleosome, has been solved later. The nucleosome is indeed a particle in which ~146 bp of DNA was wrapped in a left-handed superhelix around the octamer of core histones, consisting of two molecules of each histone H2A, H2B, H3, and H4 (Arents and Moudrianakis, 1993; Arents et al., 1991; Harp et al., 2000; Kornberg and Thomas, 1974; Luger et al., 1997).

The chromatin structure can be compacted further from an extended conformation of approximately 10 nm in size into an array of nucleosomes to form a more compact chromatin fibre (Robinson et al., 2006; Wong et al., 2007). The extended structure of 10 nm chromatin fibre, appearing at low ionic strength, is often called as a “beads-on-a-string” because it looks like a string by stretches of linker DNA between the DNA-histone structures under electron microscopy. At the higher ionic strength, the chromatin fibre forms more compact structure, which is often referred to as the 30 nm fibre (Olins and Olins, 1974).

During the cellular processes such as replication, transcription, repair, and recombination, the structure of chromatin is required to be dynamically and reversibly altered to provide access to the underlying DNA template for proteins. As a fundamental unit of the dynamic chromatin, the nucleosome unites previously disparate observations of gene activation and repression into transcription units, and to whole chromosomal domains (Lewin, 1994).

REGULATORY ROLE OF NUCLEOSOME POSITIONING IN GENOME

Nucleosomes were assumed to be evenly spaced at every ~200 bp of DNA with no specific preference to DNA sequences, emphasising the structural role they play (Kornberg and Thomas, 1974). However, it has been known that many nucleosomes are specifically positioned within the genome: either by the action of chromatin remodelling enzymes (Ito et al., 1997) or by the basal affinity of DNA for histones (Peckham et al., 2007; Segal et al., 2006; Tillo et al., 2010).

In more specific terms, nucleosomes are perceived to be positioned either statistically or specifically (Mavrich et al., 2008a). The statistical positioning means the nucleosome position is limited by the adjacent nucleosomes functioning as a barrier; if there is no adjacent nucleosome or protein, the statistically positioned nucleosome could reside on any sequence of DNA. This positioning explains that nucleosome can cover the majority of the chromosome. On the other hand, specific nucleosome positioning means that the nucleosome position is closely regulated by the underlying DNA sequences (Collings et al., 2010; Ioshikhes et al., 2006; Peckham et al., 2007; Segal et al., 2006) or by nucleosome remodelling factors (Chandy et al., 2006; Ito et al., 1997; Sadeh and Allis, 2011)

The specific positioning of nucleosomes is often found in promoter or enhancer regions. In *Saccharomyces cerevisiae*, a 150 bp nucleosome-free region (NFR) or sometimes called as a nucleosome-depleted region (NDR) at the 5' of the transcription start site (TSS) is surrounded by the highly localised nucleosomes containing the histone variant H2A.Z. The nucleosome at the 5' of the NFR is often designated as the -1 nucleosome and the one at the 3'is designated as the +1 nucleosome (Yuan et al., 2005). In Drosophila, the similar NFR is observed, but the -1 nucleosome is missing (Mavrich et al., 2008b).

The position of the +1 nucleosome relative to the TSS varies between organisms. In *S. cerevisiae*, the TSS is overlapped at 10 bp into the +1 nucleosome (Albert et al., 2007; Mavrich et al., 2008a; Montgomery et al., 2001; Yuan et al., 2005), while in metazoans, the upstream border of the +1 nucleosome is located ~ 60 bp downstream of the TSS (Barski et al., 2007; Mavrich et al., 2008b). Even though the NFR is commonly seen among eukaryotes, there are subtle differences regarding the locations and the surrounding nucleosome distributions.

Beyond the impressive packing and organising DNA, nucleosomes are believed to occupy a central role in the epigenetic regulation of gene transcription (Bai and Morozov, 2010; Wyrick et al., 1999). Histone octamers must be assembled on the parental and daughter strands of DNA during S phase (Tagami et al., 2004). Once assembled on DNA, nucleosomes provide a barrier to the transcriptional machinery due to its strong binding to DNA. Therefore, the association of DNA with histone octamers must be altered to facilitate the binding of transcription factors and travel of the RNA polymerase complex. Consequently, specific sites within genes that are actively transcribed or poised for transcription in response to environmental signals can be identified by their accessibility to endonuclease cleavage by DNase I, so called DNase hypersensitivity sites. Often, but

not always, these sites correlate with regulatory sequences such as promoters, enhancers, and locus control regions.

The tails of histones in the nucleosome are some of the most heavily targeted proteins for post-translational modification. These modifications, or histone marks, encode the epigenetic signal: a highly plastic set of expression instructions that are responsible for the diverse cellular morphology and function that can arise from a single genome (Kwon and Wagner, 2007; Roy et al., 2010; Talbert and Henikoff, 2010). The role of histone modifications in regulating gene expression has been receiving extensive attention since the discovery of the epigenome. Besides these histone modifications affecting the chromatin state, DNA positioning on a specific DNA region is sufficient to regulate gene expression simply by their steric effects on the accessibility of a given location of DNA (Lickwar et al., 2009).

On the one hand, nucleosomes can prevent protein binding to the locations occupied by the nucleosome. On the other hand, nucleosomes can potentiate protein binding to sites exposed in linkers or on the surface of the nucleosome or can facilitate the interaction between proteins juxtaposed by DNA coiling around the nucleosome (Lu et al., 1995). Therefore, it is important to know how this positioning is established (Bai and Morozov, 2010).

NUCLEOSOME POSITIONING SEQUENCES

Nucleosomes often appear at specific positions in the proximity of promoters, regulatory elements, and other special sites in DNA (Thoma and Acta, 1992). The locations of nucleosome are constrained either by adjacent barriers such as sequence specific DNA-binding proteins on the DNA or an adjacent nucleosome or by more subtle effects such as DNA bendability based on the DNA sequence (Drew and Travers, 1985).

A quick glance at nucleosome positioning maps, however, does not reveal a clear consensus DNA sequence within a nucleosome or a linker region. The lack of consensus DNA sequence for nucleosome positioning is anticipated because of the almost universal binding of histones to DNA covering the whole chromosome. Hence, more detailed analyses needed to uncover possible correlations between the DNA sequence and nucleosome positions.

The DNA sequences controlling nucleosome positions may have motifs either occurring within the nucleosome or motifs not occurring within the nucleosome. The position dependent motifs were initially characterised as particular dinucleotides that tend to occur periodically through the nucleosome, with a 10 bp periodicity (Ioshikhes et al., 1996). Both *in vivo* and *in vitro*, nucleosomal DNA showed similar 10 – 11 bp periodicities of dinucleotide distributions, although the amplitude of the periodic changes was more prominent in the latter case (Kaplan et al., 2009). These observations and the enrichment or depletion of nucleosomes at specific locations such as promoters or enhancers lead to the proposal that nucleosome positions are controlled primarily by DNA sequence in living cells (Field et al., 2008; Gabdank et al., 2009; Ioshikhes et al., 1996; Kaplan et al., 2009; Segal et al., 2006).

Histones do not directly interact with the bases in the DNA as other DNA binding proteins do. The lack of direct interaction between histones and the bases puzzled the idea of specific sequences for nucleosome positioning. Alternatively, the bendability or flexibility of DNA is considered as one of the physical properties that contribute to the specific interaction between histones and DNA. DNA is severely bent in the nucleosome structure. DNA flexibility strongly affects the intrinsic histone-DNA affinity (Morozov et al., 2009). GC rich sequences are believed to facilitate nucleosome formation by increasing DNA flexibility (Chung and Vingron, 2009; Peckham et al., 2007; Tillo and

Hughes, 2009), whereas relatively rigid poly-A:T sequences disfavour nucleosome assembly (Field et al., 2008; Kaplan et al., 2009; Mavrich et al., 2008a).

There is evidence that DNA sequence can position nucleosomes, at least *in vitro*, both translationally and rotationally. Translational positioning refers to the 146 bp sequence covered by a histone octamer, and rotational positioning refers to the 10 – 11 bp periodic orientation of the DNA helix in the histone-DNA complex. AA/TT/TA dinucleotides occur preferentially where the minor groove faces the histone octamer, whereas GC/CC/GG dinucleotides tend to occur where the minor groove points away (Travers et al., 2009). The rotational positioning affects the histone-DNA affinity and is closely related to other regulatory factors (Ioshikhes et al., 2011).

Nucleosome positioning by DNA sequence *in vivo*, however, is debatable. Kaplan et al., (2009) showed that the intrinsic DNA sequence preference of nucleosome was a major determinant of organising nucleosomes. They compared the *in vitro* nucleosome map with the *in vivo* map by measuring the nucleosome occupancy on purified yeast genomic DNA. From the fact that the *in vitro* map was comparable with the *in vivo* map and the identification of the nucleosome preferred sequences from the nucleosome-bound sequences, they concluded that the intrinsic DNA sequence preference was the major determinant organising nucleosomes *in vivo* as well as *in vitro*. However, there is another view on the role of DNA sequences in determining the organization of nucleosomes *in vivo*. Zhang et al., (2009) claimed that the intrinsic DNA sequence was a determinant of the rotational positioning of nucleosomes rather than the translational positioning (Zhang et al., 2009). They claimed the translational positioning was determined by the barriers such as the nucleosome-disfavouring sequences at the promoters and terminators, or transcription factor binding sites. Even though both groups agreed that the *in vitro* nucleosome positions were determined by intrinsic DNA sequences, their conclusions

about the role of DNA sequences *in vivo* are different. One probable reason for the discrepancy is the differences in the definitions of nucleosome occupancy and nucleosome position and the methods the two groups used to evaluate the predicted nucleosome positions *in vivo*. Kaplan et al. determined the nucleosome occupancy as the most probable position from the set of short reads, while Zhang et al. determined the nucleosome position from each short read.

Even though they did not agree on the role of intrinsic DNA sequences on the translational positioning of nucleosomes *in vivo*, both groups showed that nucleosomes assemble on purified yeast genomic DNA (Kaplan et al., 2009) and nucleosomes preferentially form on yeast DNA than *Escherichia coli* DNA (Zhang et al., 2009). Even though the nucleosome positions can be shifted *in vivo* by such factors as the transcription factor binding and remodellers, the intrinsic DNA sequence determines the organization of nucleosomes for their "default" positions without the effects of the other factors. The well-positioned nucleosomes downstream of transcription start sites, which serve as an anchor for the subsequent nucleosome positions, are good samples to examine the sequence preference for the nucleosome formation.

Various nucleosome positioning sequence (NPS) patterns in eukaryotes have been proposed. Most of the NPS patterns are characterised by repeating dinucleotide sequences as in yeast (Albert et al., 2007; Ioshikhes et al., 2006; Segal et al., 2006), *Caenorhabditis elegans* (Salih et al., 2008), and humans (Barski et al., 2007; Ozsolak et al., 2007; Schones et al., 2008). Specifically, the periodic appearance of dinucleotides in every 10 bp favours the histone binding by allowing the tight bending of DNA around the histone octamer core (Fraser et al., 2009; Ioshikhes et al., 2011; Richmond and Davey, 2003).

Many statistical methods and machine learning methods were applied to predict the nucleosome positions. First of all, the dinucleotide periodicities such as repeating

AA/TT or GG/CC dinucleotides in 10 bp interval are frequent parameters in the predictions (Ioshikhes et al., 1996, 2006; Lowary and Widom, 1998). Other predictions incorporated various parameters: free energies associated with DNA bending based on structural DNA of crystallography or NMR (Miele et al., 2008; Morozov et al., 2009), a thermodynamic model including interactions between adjacent nucleosomes (Lubliner and Segal, 2009), or DNA deformation energy (Roy et al., 2010). In addition to the genomic sequences, the *in vivo* cellular status was integrated into the prediction model (Kaplan et al., 2009; Segal et al., 2008). Peckham et al. used Support Vector Machine (SVM) classifier to find the favouring and disfavouring sequences: AT-rich *k*-mer oligonucleotides disfavoured histone binding, whereas GC-rich sequences enhanced nucleosome formation (Peckham et al., 2007). Most of the prediction models used the yeast sequences, and a Hidden Markov Model was used in various eukaryotes beyond yeast to predict nucleosome positions with the underlying genomic sequences (Xi et al., 2010).

NUCLEOSOME WITH HISTONE VARIANT H2A.Z

Since the nonallelic variants of histones can be incorporated in chromatin replacing the S-phase histones throughout the cell cycle independent of the replication, they are also known as replacement variant. Two major H3 variants, the centromere specific protein CenH3 and H3.3, have been extensively studied. Also, the number of H2A variants is large and include H2A.Z, H2A.X, H2A-Bbd, and macroH2A (Zlatanova and Thakar, 2008) .

H2A.Z nucleosome has been involved in diverse biological processes, such as gene activation, chromosome segregation, and blocking heterochromatin silencing and progression through the cell cycle. H2A.Z is essential for viability in Drosophila

(Clarkson et al., 1999; van Daal and Elgin, 1992). The mutation of the H2A.Z in mice is also lethal so that no homozygous H2A.Z $^{-/-}$ mice could be obtained (Faast et al., 2001). H2A.Z histones are highly conserved with ~ 90% sequence conservation between yeast and human proteins (Iouzalen et al., 1996). Its sequence identity to the canonical H2A is, however, only 60%, which suggests unique and essential functions for H2A.Z (Jackson and Gorovsky, 2000). Overexpression of H2A.Z was observed in several major types of malignancies especially at the metastatic stage (Rhodes et al., 2004; Zucchi et al., 2004).

OBJECTIVES

NPS patterns were proposed in some eukaryotes including yeast (Albert et al., 2007; Ioshikhes et al., 2006; Segal et al., 2006), *C. elegans* (Salih et al., 2008), and humans (Barski et al., 2007; Ozsolak et al., 2007; Schones et al., 2008). However, Drosophila patterns are not fully investigated. I hypothesized that Drosophila NPS patterns might well be distinct from the yeast NPS patterns because the locations of the enriched nucleosomes at 5' of genes are different between yeast and Drosophila. I also hypothesised that the histone variant H2A.Z nucleosome may have different patterns considering its low sequence homology with the H2A and distinct biological roles.

I examined the NPS patterns of Drosophila nucleosomes following identifying nucleosome positions in Drosophila by aligning the high-throughput sequencing data from ChIP-Seq experiments. The DNA sequences of the identified H2A and H2A.Z nucleosomes were examined for the NPS, especially in terms of the dinucleotide patterns. The analysis revealed the common and unique properties of the Drosophila NPS patterns.

As previously described, H2A.Z nucleosomes have distinct cellular roles from the H2A nucleosomes. I also hypothesized that the positioning the H2A and H2A.Z nucleosomes are related to the cellular processes, and the underlying NPS is designed to

control the gene expression through nucleosomes as transcription factor binding sites direct the transcription factors. I examined the enriched biological functions of the selected genes based on the enriched nucleosome and the NPS at the promoters and found relationships.

Ly49 gene clusters as a model system

The role of the NPS in regulating gene expression and the interplay of nucleosomes and protein factors guided by underlying genomic sequences was examined using Ly49 gene cluster as a model system. I chose the Ly49 gene cluster as a model system to investigate the effects of nucleosome positions on transcriptional regulation. The Ly49 genes, which code surface receptor proteins of NK cells, has several advantages to being an ideal model, besides their importance in innate immunity, for our investigation of transcriptional regulation by nucleosomes.

The Ly49 genes exist as a cluster in a 650 kb region of chromosome 6 in mouse and have similar requirements for transcription factors (Carlyle et al., 2008). The cluster comprises of several Ly49 genes, often called Ly49 gene family. They are polymorphic and polygenic: different mouse strains have different alleles of Ly49 genes. However, expression of an individual Ly49 gene is stochastic, such that each NK cell acquires a unique repertoire of Ly49 receptors sequentially during development and then maintains this repertoire throughout its life (Kubota et al., 1999b; Ortaldo et al., 1999; Pascal et al., 2006).

The clustered organisation and stochastic expression of Ly49 genes provide a good condition as a model system: multiple copies of the genes under the similar transcriptional control make it possible to analyse the results quantitatively using statistical methods. Besides, the relatively small number of the genes and the

chromosome size make it feasible to examine and verify the procedure manually, if necessary, before applying complex computational methods. Additionally, a common NKT cell line—called RMA (RBL-5, mutagenized but not selected)—naturally expresses only Ly49A and none of the other Ly49 proteins, providing a convenient model for verifying this computational study. The positions of the nucleosomes and transcription factors were determined based on the DNA sequences. The goal has been to study how changes to this "default", sequence-determined landscapes of nucleosomes and transcription factors correlate with gene expression. Therefore, the effects of nucleosome positioning on the expression of a family of immune genes, the Ly49 receptors, were investigated.

Ly49 gene family

LY49 RECEPTORS

Introduction and nomenclature

NK cells are bone marrow-derived lymphocytes. They were characterised by their large, granular morphology and cytolysis activity against a variety of tumour targets and infected cells (Trinchieri, 1989). Ly49 receptors are surface proteins expressed on NK cells and other immune cells of mice. The receptors are homodimeric type II C-type lectin-like membrane glycoproteins. The genes encoding the receptors are located on chromosome 6 as a cluster. Ly49 gene family is also called as Killer Cell Lectin-Like Receptor Subfamily A (Klra) (Schenkel et al., 2013). In this nomenclature, the genes are indicated by a number, e.g. Klra1 or Klra2, while the Ly49 nomenclature uses the subscript indicating the strain allelic variants, e.g. $Ly49A^{C57BL/6}$. The Ly49 nomenclature was followed in this study.

The highly polymorphic and polygenic Ly49 gene was first identified as T lymphocyte activation marker (Chan and Takei, 1989; Yokoyama et al., 1989). The Ly49 gene product was first identified as an inhibitor of NK lymphocyte killing tumour cell lines that express MHC-I. The eradication of NK cells was inhibited by Ly49A's recognition of H2-Dd on the tumour cells (Yokoyama and Seaman, 1993). The recognition of MHC-I by Ly49 receptors is a key to killing by NK cells. The abhorrent expression of MHC-I on tumours or infected cells triggers the killing by NK cells.

The Ly49 genes were encoded in rodents and cattle, while humans have the Killer Immunoglobulin-like Receptor (KIR) fulfilling the same role, instead. They are not genetically homologous to the Ly49 receptors, but rather have a parallel function (Middleton and Gonzelez, 2010). KIRs are also highly polymorphic and contain activating and inhibitory receptors. Ly49 receptors and KIR are functionally similar but structurally distinct: the former are disulfide-linked homodimers homologous to type II C-type lectins, while the latter are type I monomeric Ig-like receptors. However, both bind directly to MHC-I molecules (Boyington et al., 2000; Wagtmann et al., 1995).

Role of Ly49 on various immune cells

The Ly49 expression is not limited to the NK cells but broad in the adaptive immune system as well. Invariant NK T (iNKT) lymphocytes, intestinal epithelial lymphocytes (IELs), NKT cells, uterine NK (uNK) cells, $CD8^+$ T regulatory (Treg) cells, $CD8^+$ T cells, $CD3^+$ cells, $NK1.1^+$ γ/δ T lymphocytes also express Ly49 receptors (Denning et al., 2007; Gays et al., 2006; Hara et al., 2001; Kamogawa-Schifter et al., 2005; Maeda et al., 2001; Nikolova et al., 2011; Ortaldo et al., 1998; To et al., 2009; Toyama-Sorimachi et al., 2004; Yadi et al., 2008). Ly49Q is found in myeloid cells and plasmacytoid dendritic cell (pDCs) (Kamogawa-Schifter et al., 2005).

The cells expressing the Ly49 or KIR receptors show various roles in immunity. For example, uNK cells play a significant role during mouse pregnancy (Yadi et al., 2008), and Human uNK cells have been shown to protect congenital viral infection by human cytomegalovirus (HCMV) displaying cytotoxic effector function upon recognition HCMV-infected cells (Siewiera et al., 2013).

Both Ly49 receptors and T cell receptors recognise MHC-I molecules expressed on the cells. However, the mode of recognition varies between them. T cell receptors (TC) have specificity to the antigenic peptide bound to MHC-I (Garboczi et al., 1996). On the other hand, the degree of MHC-I binding of Ly49 receptors is related to the extent of functional inhibition, and neither the peptide specificity nor co-receptors are necessary for MHC-I binding to Ly49 receptors in most cases (Correa and Raulet, 1995; Raulet et al., 1997).

Inhibitory and stimulatory Ly49 receptors

Ly49 receptors have distinct roles: many of them are inhibitory, while some are stimulatory. Both the inhibitory and stimulatory functions are essential in the education of NK cells. The inhibitory Ly49 receptors bind to MHC-I complex presented on target cells recognising MHC-I molecule H2-D, H2-I, and H2-K. The inhibitory Ly49 receptors prevent killing the target cells upon recognition of their cognate ligand. The absence of these ligands on target cells combined with other activating signals to the NK cell triggers a cytolytic response by the NK cell against the target cells (Jonsson et al., 2010; Schenkel et al., 2013).

The killing or tolerance of NK cells of the target cells upon the recognition of the signal is explained by the missing-self hypothesis (Kärre et al., 1986). NK cells survey MHC-I expression on cells with which they come into contact. To avoid the detection

and possible killing by cytotoxic T cells, abhorrent or infected cells often down-regulate expression of MHC-I to avoid detection. However, this down-regulation is recognised by NK cells as "missing-self", and functions as a kill signal for NK cells: the cells missing β_2m (MHC-I light chain) is killed. The reintroduction of the light chain by transgene restores the resistance to NK cell killing (Liao et al., 1991). Interestingly, NK cells from $\beta_2m^{-/-}$ mice exhibited a diminished ability to kill traditional NK cell targets as well, which suggests that the MHC-I interaction is not only inhibiting the cytolytic activity of NK cells but is also necessary to the activity of NK cells during the development.

Besides the inhibiting Ly49 receptors, many activating Ly49 receptors also bind to MHC-I molecules (Dimasi and Biassoni, 2005). But unlike the inhibitor receptors, the binding of the activating Ly49 receptors is the recognition of alleles from different strains or altered MHC molecules. It has been discovered that disruption of Ly49O unintentionally lowered transcription of all Ly49 genes (Belanger et al., 2008). The Ly49 receptors have not only inhibitory activities but play a role in stimulating the NK cell killing. Under some conditions, NK cells attack MHC-I^+ tumour cells. Many activating receptors on NK cells have been identified. Some are specific for ligands that are upregulated on tumour cells and stressed cells, and others apparently specific for ligands on normal cells. This disparate recognition system can be understood in a model in which NK cells are regulated by the balance of signalling via stimulatory receptors, specific for diverse ligands, and inhibitory receptors, specific for MHC-I molecules (Raulet and Vance, 2006). More than 16 different Ly49 receptors have been identified in CL57BL/6 mice, though just 8 of these are of the inhibitory type (McQueen et al., 1999; Smith et al., 1994). Some Ly49 genes inhibited killing, but others activated killing both via interactions with MHC-I (Makrigiannis and Anderson, 2000).

Ly49 inhibitory receptors are important for the prevention of autoimmunity by suppressing NK cell activation. The acquisition of inhibitory Ly49 for self-MHC-I is a key step in the "licensing" of developing NK cells to avoid hyper-responsive state (Kim et al., 2005b). At the same time, the activating Ly49 receptors recognise ligands that are expressed on abnormal or infected cells. The recognition activates cytokine production and cellular cytotoxicity by NK cells.

Receptor structure and signal transduction

NK cell signals transduce through the cascades of phosphorylation or dephosphorylation. The Ly49 receptors have motifs in their cytoplasmic domains, called immunoreceptor tyrosine-based inhibitory motifs (ITIM) (Burshtyn et al., 1997). Upon the recognition of the MHC-I, the tyrosine in the ITIM is phosphorylated. The tyrosine phosphorylation is restricted to the inhibitory Ly49 molecules such as Ly49 A, C/I and G2, while the activating Ly49D is not phosphorylated. Phosphorylated ITIM is responsible for the recruitment and activation of tyrosine phosphatases such as SRC homology 2 (SH2) domain-containing protein tyrosine phosphatases 1 (SHP-1) and possibly SHP-2 (Binstadt et al., 1996; Burshtyn et al., 1996; Campbell et al., 1996; Olcese et al., 1996; Vély et al., 1997). The phosphatases probably interfere with the subsequent signals suppressing the tyrosine phosphorylation-based signals downstream of NK cell stimulation pathways. The activating receptors play a role in the signal transduction as the phosphorylation of ITIM by engagement of the inhibitory receptors is enhanced by cross-linking with activating receptors.

The activating Ly49 receptors transduce signals through different path using small transmembrane adapter proteins such as killer cell activating receptor-associated protein (KARAP)/DNAX activating protein of 12 kDa (DAP12), and DAP10 that transmits

activation signals (Gosselin et al., 1999; Lanier et al., 1998; Olcese et al., 1997). DAP12 is important for signalling, but the requirement of DAP10 for signalling is potentially minimal (Tassi et al., 2009).

Education and immunosurveillance of NK cells

NK cells acquire the tolerance toward the self-MHC-I during the differentiation through a process called "education". The inhibitory Ly49 receptors are important factors in the NK cell education. NK cells must undergo a MHC-I-dependent "licensing" process in order to be "self-tolerant", or be functional in which self-specific Ly49 receptor interactions with self-MHC-I (Kim et al., 2005a). The "licensing" process was believed finished during the early development of NK cells, but recently it was reported that the education does not occur only during development of immature NK cells but is actually a dynamic and reversible process (Elliott et al., 2010; Joncker et al., 2010).

The "arming" and "disarming" models are the widely-accepted model for the licensing. In the arming model, NK cells need to express an inhibitory receptor for a self MHC-I to turn on their stimulatory signalling pathways, or "armed". Otherwise, they might never be "armed". In the disarming model, NK cells that fail to express an inhibitory receptor for a self MHC-I molecule might be actively rendered hypo responsive, or "disarmed", by downregulation of stimulatory signalling pathways (Raulet and Vance, 2006). Whether the education is through the *cis*- interaction with MHC-I expressed on its own or *trans*-interaction with the MHC-I expressed on another cell is unknown. Loss of MHC-I expression along with up-regulation of ligands for activating NK cell receptors on tumour cells results in their recognition and elimination of NK cells (Beĺanger et al., 2012).

LY49 EXPRESSION ON NK CELLS

Role of NK cells

Unlike the adaptive immunity, NK cells need neither prior sensitization nor proliferation to kill the target cells. The recombination-activating gene (RAG)-dependent antigen receptors are not expressed in the NK cells, either. These traits distinguish NK cells from the adaptive immune system to be considered important in innate immunity (Raulet, 2004). NK cells can also secrete cytokines (especially IFN-gamma and TNF-α) and chemokines (MIP-1 family members and RANTES). In contrast to T and B cell responses to antigen, which typically require a proliferation phase, the NK cell response is immediate, implying that NK cells are involved in curbing pathogens during the initial several days of infection contributing to the defense against intracellular bacteria (Unanue, 1997), parasites (Scharton-Kersten and Sher, 1997), and that they are critical for controlling several types of viral infection (Biron et al., 1989, 1999). The anti-tumour activities of NK cells are well described *in vitro* (Kiessling et al., 1975; Nunn et al., 1976) and in certain *in vivo* models (Beĺanger et al., 2012), yet their roles in the defence against spontaneous neoplastic transformation remains less well established.

The sensitivity to recognise MHC-I expression levels is critical for NK cells to distinguish healthy cells from cancer or virus-infected cells (Hanke et al., 1999; Kim et al., 2005a). These receptors may regulate not only the immune response to tumours but also viruses and other intracellular pathogens as well. Lack of NK cell function, even benign viruses can be severe (Orange, 2002). Many viruses inhibit expression of MHC-I molecules on the cell surface as a way of evading conventional cytotoxic T lymphocytes (Loch and Tampé, 2005). Although this allows the virus to hide in the cell from these classical T lymphocytes, the virus-infected cells are vulnerable to killing by NK cells.

Viruses, however, developed a way to evade the NK cells by expressing viral proteins on the infected cells that resemble the host MHC-I molecules or by down-regulating the activating ligands on the surface of infected cells. The recognition of the viral MHC-I-like molecules through interaction with inhibitory receptors inhibits NK cells (Rajagopalan and Long, 2005). MCMV expresses MHC-I-like viral protein to evade the NK cell killing (Arase, 2002).

The paradigm of inhibiting NK cells upon the recognition of MHC-I by the inhibitor receptors shifts recently because unlicensed NK cells are better at killing neuroblastoma tumours (Tarek et al., 2012). Unlike the other tumours reducing the MHC-I levels on the surface to avoid detection from T cells, the levels of MHC-I is high in neuroblastoma cells, which can inhibit licensed NK cells. However, unlicensed NK cells are not inhibited by MHC-I on neuroblastoma cells and effectively kill neuroblastoma tumours.

Thanks to the ability of NK cells killing aberrant cells through recognising with Ly49 receptors, there are attempts to use the NK cells to treat diseases. Manipulation of human KIR signalling has been proposed as a potential cancer therapeutics (Benson et al., 2012; Koh et al., 2001; Romagné et al., 2011; Vey et al., 2012). Or blocking inhibitory receptors on NK cells may have beneficial therapeutic effect in certain viral infections (Benson et al., 2012; Koh et al., 2001; Romagné et al., 2011; Vey et al., 2012).

NK cells go through five stages of development assessed by the surface expression of the markers (Kim et al., 2002). Immature NK cells undergo extensive cell division at stage V. Expression of the Ly49 family are increasing in the developing NK cells during the first 2 – 3 weeks and reach optimal within 6 – 8 weeks after birth. With the one possible exception of Ly49E that is expressed in fetal cells, most Ly49 receptors

are expressed after birth in mice (Van Beneden et al., 2001; Dorfman and Raulet, 1998; Sivakumar et al., 1999).

While the receptor genes were originally linked to NK killing of tumours and virus infected cells, recent studies have shown additional roles for Ly49 in other cell types. Ligands presented on the target cells, and Ly49 receptor expression patterns remain only partially deciphered. The likely requirements of these genes in combatting intracellular viral, bacterial, and parasitic pathogens remain to be explored.

Ly49 repertoire in inbred mouse strains

The methods for characterization of cells expressing Ly49 are limited. Only 11 of the Ly49 receptors have the recognising monoclonal antibodies. To make the issue harder, many of the antibodies are cross-reactive for more than one Ly49 gene. So, profiling NK cells precisely by the surface Ly49 receptors is hard. The survey of the Ly49 receptor expression is often done by searching the transcripts instead of detecting the protein or sorting the cells. Ly49 genes are polymorphic and polygenic: they have heterogeneity in the type, and the level of Ly49 molecules expressed in different mouse strains (Ortaldo et al., 1999). Besides the inhibiting and activating receptors, several Ly49 genomic sequences are non-functional pseudogenes. In total, there have been known approximately 20 – 30 members of Ly49 receptors including pseudogenes (Yokoyama and Seaman, 1993).

Different numbers of the Ly49 genes are encoded in the gene cluster depending on the mouse strain. BALB/c mouse strain possesses 9 genes (Anderson et al., 2005), Non-obese Diabetic Mouse (NOD/ShiLtJ or NOD) strain possesses 22 genes (Belanger et al., 2008), C57BL/6 possesses 16 genes (Makrigiannis et al., 2002, 2005) and the 129 strain has 20 genes (Anderson et al., 2005; Makrigiannis et al., 2005). Not all the Ly49

genes are protein coding genes. For example, Ly49 genes in CL57BL/6 strain have genes with a complete mRNA (Ly49a-j, and q) and pseudogenes (Ly49k, m, and n) (Brennan et al., 1994; Makrigiannis et al., 2000; McQueen et al., 1999; Silver et al., 2001; Smith et al., 1994). Ly49 haplotypes have a limited degree of conservation in the form of "framework" genes. The framework gene pairs are Ly49q-e, Ly49i-g, Ly49c-a. Strain-specific Ly49 genes are mediated by stop codons within the coding regions. Such diversity is driven by due to pathogenic challenges such as viral infections (Orange, 2002).

Stochastic expression of Ly49 receptors

Murine L49 genes are located as a cluster on chromosome 6, and human KIR genes are located on chromosome 19. The MHC-specific Ly49 receptors can discriminate among the various MHC-I alleles. The receptors are expressed in a variegated, overlapping fashion, such that each NK cell expresses several inhibitory and stimulatory receptors. The combinations of receptors resulted from the variegated expression enable NK cells to discriminate the self and non-self cells. The variegated pattern of receptor repertoire on NK cells results from the stochastic choice of which Ly49 genes are expressed.

Murine NK cells express up to six Ly49 receptors with an average of two to three Ly49 receptors on a cell in an overlapping fashion (Kubota et al., 1999a; Valiante et al., 1997). The maximum number of receptor combinations in the repertoire or its complexity can be approximated assuming that each NK cell can express from two to six Ly49 receptors. Provided that each mouse encodes 10 different inhibitory receptors, then approximately 1000 distinct types of NK cells can prevail expressing different receptor combinations. The sequence polymorphisms adding variations to Ly49 alleles at the same

locus in Ly49 heterozygous mice increase the possible NK cell population with the distinct repertoire. Ly49 gene expression is mainly monoallelic: most NK cells express either the maternal or the paternal allele of receptors, but not both (Held and Kunz, 1998; Held and Raulet, 1997; Held et al., 1995).

It has been proposed that the activation of Ly49 gene transcription is a stochastic process as the proportion of NK cells expressing two individual Ly49 receptors follows the chain rule, indicating the expression of the two Ly49 receptors are independent (Held and Kunz, 1998). The product rule provides a good estimate of the frequency of NK cell subpopulation expressing a given combination of receptors (Raulet et al., 1997; Valiante et al., 1997). Single-cell RT-PCR analysis of Ly49 expression has shown that in average one to four receptors are expressed on each NK cell with very low probability expressing more than five receptors at the same time on a cell, supporting the joint distribution of independent expression (Kubota et al., 1999a). That the stochastic expression follows the product rule implicates that different receptor genes are expressed independently.

The expression of the inhibitory receptors is overlapped so that a subset of NK cells share the receptors partially with other NK cells. Thus, each cell expresses multiple receptors and the combinatorial repertoire of the expressed receptors on NK cells generates population NK cells of unique specificities. This variegated pattern of receptor expression allows individual NK cells to discriminate between cells expressing different MHC-I molecules. The stochastic expression pattern applies to murine Ly49 receptors (Raulet et al., 1997) and KIR in humans (Valiante et al., 1997), despite their structural dissimilarities.

Even though a stochastic mechanism appears to govern the variegated expression of the inhibitory receptors following the product rule, deviations of the final NK cell

populations of the estimated proportion are expected since the education mechanisms censor the repertoire based on the expressed MHC-I molecules by the host. The education processes are still poorly understood.

LY49 GENE CLUSTER

Organisation of the gene cluster

The genes that encode the mouse Ly49 receptors are clustered in tandem in the natural killer gene complex (NKC) on mouse chromosome 6 with the exception of Ly49b (Yokoyama et al., 1990). Ly49b is located on chromosome 6, but outside the Ly49 cluster. The Ly49 gene cluster has been mapped in the C57BL/6 mouse genome (Brown et al., 1997; Depatie et al., 2000; McQueen et al., 1998). The highly related activators, Ly49h, k, and n are grouped, but in general, the Ly49 genes are not grouped with respect to activating and inhibitory functions or gene homology.

A study revealed the presence of an additional Ly49 promoter (Pro-1) and two non-coding exons (exon -1a and exon -1b) upstream of the previously defined promoter. The previously defined promoter was then renamed as Pro-2 (Saleh et al., 2002). The Pro-1 homologous regions were found at 4-10 kb upstream of Pro-2 in all examined Ly49 genes. Pro-1 is used to for transcription: Pro-1 transcripts were detected from the Ly49a, e, g, o, and v genes, and the activity was detected in bone marrow, fetal thymus, liver NK cells, and the murine LNK cell line. The inactivity of the Pro-1 in mature NK cells suggests its unique role as a promoter in the early stages of NK cells (Saleh et al., 2002). There is a recent report claiming that the Pro-1 is active in mature NK cell, but the Pro-1 in the mature NK cell is functioning as an enhancer instead of a promoter without the necessity of TATA (Gays et al., 2015a).

Regulation of expression

Ly49 gene families and the human analogues KIR gene families are composed of closely related genes with unidirectional transcription and mono-allelic yet independent and stochastic expression (Held and Kunz, 1998; Kubota et al., 1999a). The transcription level regulation is responsible for the unique pattern of gene expression (Kubota et al., 1999a). Transcriptional repression can generally be achieved either or both by limiting available transcription factors and by the inaccessibility of regulatory sequences such as the promoter and enhancer.

Chromatin remodelling and DNA methylation are one of the major mechanisms of transcriptional control for tissue-specific genes and the allele-specific gene expression (Attwood et al., 2014). Since this correlation of transcription and epigenetic states seems to be a common feature of immune system genes (Smale and Fisher, 2002), Ly49 expression may also be under the control of epigenetic states. The expression levels of various receptor subunits in NK cell subpopulations are evident at the mRNA level, suggesting differential transcriptional regulation of receptor gene expression (Held et al., 1995; Kubota et al., 1999a). The variegated expression of the Ly49 receptor presents interesting questions about the initiation and maintaining the expression in individual NK cells.

Methylation state was proposed on the basis of the relative stability of monoallelic expression of Ly49 receptors on cultured NK cells (Held and Raulet, 1997). Also, the selection of Ly49 gene maintained stable over the development. There must be a regulation mechanism like epigenetic control to maintain the expression over multiple cell divisions. Ly49a Pro-2 shows a strong link between DNA hypomethylation, histone acetylation, and transcriptional activity of the gene in mature NK cells. Methylation in

Pro-2 is high in non-expressing cells and low in expressing cells. Methylation level is high in fetal NK cells and is varying in nonlymphoid tissues.

Transcription factors in Ly49 transcription

Many transcription factors regulate the transcription of the Ly49 gene. During the conditional expression of a Ly49 gene by Pro-1 forward or reverse promoter, AML-1 is considered as a key transcription factor. Ly49a expression has been shown to be dependent on the transcription factor T cell factor-1 (TCF-1) during development (Held et al., 1999). However, the fact that TCF-1 is present in an equal amount in both Ly49A expressing and non-expressing primary NK cells makes TCF-1 hardly be the key regulator that controls which allele of Ly49A to express. Another transcription factor, activating transcription factor-2 (ATF-2) plays a role in transcription binding to a 13-bp regulatory element restricted to Ly49A-expression exhibiting promoter activity in EL-4 cells (Kubo et al., 1999).

Some transcription factors co-regulated certain Ly49 receptor genes resulting in co-expression of the receptors. For example, the transcription factor TCF-1 may control the expression of Ly49A and Ly49D but not the other Ly49 genes (Held et al., 1999). This co-regulation and the education during the maturation have the NK cell population with various combinations of Ly49 receptors deviate from the distribution calculated by the product rule.

Not only was the binding affinity of the forward and reverse promoters of Pro-1, but chromatin status also believed to play a role in the variegation of the Ly49 expression. The expression of the different combinations of Ly49 receptors – variegation, is complex and the regulation is difficult to be explained by transcription factors alone as the promoters may share common transcription factors binding sites.

Transcriptional regulation

Transcription is controlled in part by the assembly of relevant complexes of transcription factors in regulatory regions of genes. The abundance of transcription factors, localisation and higher order interactions among transcription factors, co-activators, or repressors are an array of mechanisms through which transcription is regulated. The second level of control is provided by epigenetic processes that affect the accessibility of transcription factors to the regulatory regions of their target genes within a highly ordered chromatin structure. By modulating the accessibility of transcription factors toward the regulatory regions of genes, epigenetic factors may silence, dampen or facilitate efficient transcription (Ansel et al., 2003; Egger et al., 2004; Felsenfeld and Groudine, 2003; Li, 2002; Raulet et al., 1997; Smale and Fisher, 2002).

A study presented that an upstream Ly49 promoter Pro-1 is bidirectional and acts as a probabilistic switch that determines the fate of the Ly49 gene and establishes the activity of the downstream promoter, Pro-2 (Saleh et al., 2002, 2004). Pro-1, the promoter at the upstream of exon -1a, is a bi-directional promoter. Depending on the binding to the forward or reverse direction, the Ly49 gene expression is activated or repressed. Binding to the reverse promoter represses the expression, while binding to the forward promoter starts transcription. The transcription process may change the chromatin state at the downstream which leads to constitutive expression of Ly49 genes in the mature NK cells (Pascal et al., 2006). The bidirectional promoter Pro-1 consists of an array of conserved transcription factor binding modules: TATA boxes at each 5' and 3' each flanked by C/EBP binding sites with AML-1 and NF-κB binding sites in the middle (Saleh et al., 2004). The forward and reverse directions for Pro-1 depend on the same transcription complex competing for each other. The transcription complex may assemble by chance and transcribe only in one direction. The forward transcription

process keeps proceeding past the downstream promoter, Pro-2, and in some cases, Pro-3 (McQueen et al., 2001; Saleh et al., 2004). The forward transcription process may dislodge an inhibitory complex, probably the closed chromatin state, at the upstream of Pro-2, which then maintains the expression of that Ly49 gene for the rest of the NK cell's life. Conversely, reverse transcription means the inhibitory complex stays at a key developmental moment, forever barring the NK cell from expressing that Ly49 (Saleh et al., 2004). The relative strength of the forward and reverse promoters is believed to be modulated by the strength of the C/EBP binding sites. Pro-1 is hypersensitive to DNase I in both Ly49A-expressing and -nonexpressing cells but not in nonlymphoid tissue (Tanamachi et al., 2004).

Maintenance of expression by epigenetic control

The mechanism that ultimately stabilises and maintains the final expression pattern is likely epigenetic. The usage of the Ly49 promoters shifts during the maturation of the NK cells. The bi-directional promoter Pro-1 is active only in immature NK cells. Once the NK cells mature, the promoter for the transcription is switched from Pro-1 to Pro-2, which leads to permanent expression of the Ly49 gene (Pascal et al., 2006). Once the Ly49 gene is successfully activated, its expression is stably maintained throughout multiple rounds of proliferation. In mice, the activated Ly49 receptors maintain the expression for at least 10 days *in vivo* (Dorfman and Raulet, 1998), and for at least 1-2 weeks of *in vitro* (Karlhofer et al., 1992).

Since the expression is maintained even after rounds of proliferation, the epigenetic control is implicated as the control mechanism. There is an inhibitory element immediately upstream of the Pro-2 of the Ly49c and j promoters (McQueen et al., 2001). The human KIR genes have conserved CpG islands among the promoters. Each KIR gene

is consistently methylated in silent KIR genes and demethylated in active KIR genes (Santourlidis et al., 2002). Unlike the human KIR genes, both Pro-1 and Pro-2 of the mouse Ly49 genes do not have CpG in conventional definition. However, the CpG dinucleotides in Ly49A$^-$ NK cells are heavily methylated, whereas the DNA from half of the Ly49A$^+$ cells are unmethylated and the other half are heavily methylated indicating that Ly49 methylation correlates with monoallelic expression in mice (Rouhi et al., 2006). Acetylation is also part of the epigenetic control of Ly49 genes. There is a significant difference in acetylation levels in Pro-2 regions between expressing and non-expressing NK cell lines (Chan et al., 2005). The acetylation level of KIR genes, however, is maintained high regardless of expression state (Uhrberg, 2005).

OBJECTIVES

While the probabilistic model of the bi-directional Pro-1 nicely describes the observed Ly49 expression patterns during the maturation of NK cell, it is not sufficient to explain the variegation of Ly49 receptors in NK cell population because the 'forward' and 'reverse' sequences in Pro-1 remain the same among the NK cells. There must be another layer of regulatory mechanism to differentiate the various Ly49 receptors in the NK cells. Furthermore, a recent report has shown that Pro-1 affinity does not always correlate with that Ly49's expression level (Gays et al., 2015b).

Variegated expression with possible epigenetic control of polymorphic Ly49 receptors suggests that some unknown factors may act in regulating Ly49 expression, giving forth an idea of the effects of nucleosome positioning as the unknown factor on the stochastic expression. I hypothesised that the accessibility to the promoters established by nucleosomes is an important mechanism to determine the use of forward or reverse promoters and the subsequent stochastic expression of the Ly49 receptors in

NK cells. The genomic sequences may work as a blueprint to organise the default layout of nucleosomes and transcription factors to work together: the organisation of the DNA sequences in the promoters differentiates the availability of the transcription factor binding sites and subsequent binding of the transcription factors by modulating the accessibility. The relationship between the transcription factor and the nucleosome determined by genomic sequences was examined by computational methods.

Certain transcription factor binding sites within the Ly49 promoters or the enhancer are sensitive to the nucleosome coverage. I hypothesized that these sensitive binding sites would be preferentially enriched within predicted nucleosome-bound regions of DNA. Additionally, our lab has previously shown that some transcription factors in nucleosome-covered regions are arranged 'in-phase' with nucleosomes indicated by a noticeable pattern of 10 bp periodicity of the distance distribution (Ioshikhes et al., 1999). Binding sites having such a periodic pattern would be able to orient themselves on the same side of DNA, since the 10 bp interval corresponds to one turn of a DNA double helix. The binding sites facing outward from the nucleosome when they are incorporated in a nucleosome would be available for target protein binding. TATA box binding sites require the 10 bp phasing of the binding sites to the proximal nucleosome as disrupting the phasing was shown to severely impact promoter function (Imbalzano et al., 1994).

In this research, I identified that the nucleosome positions in the Ly49 promoters mostly overlapped the forward Pro-1 site hindering the availability of the binding sites for the activation of the genes, while leaves the reverse Pro-1 freely available. The coverage of transcription factor binding sites by nucleosomes was systematically analysed identifying 'open' and 'covered' binding sites. Among the selected 17 transcription factors, I identified that AML-1 as the transcription factor of which binding

sites displayed both preferential nucleosome coverage and lack of 10 bp periodicity in distance distribution. The relationship between the factor binding sites and the nucleosome positions were studied by Association Rule Mining confirming the open and closed arrangement qualitatively. The enriched binding motifs on the DNA around nucleosomes revealed the genomic layout controlling both nucleosome positioning and transcription factor binding. I then confirmed, with the help of collaborators, the findings from the computational analysis using a cell line that naturally expresses Ly49A: AML-1 sites are preferentially depleted of nucleosomes throughout the promoter/enhancer regions of expressed Ly49 genes comparing to the sites of the unexpressed Ly49 genes within the same population, implying that nucleosome positioning is a possible mechanism regulating Ly49 expression *in vivo*.

PART 1: NUCLEOSOME POSITIONING SEQUENCE PATTERNS

Introduction

The nucleosome, which is the core subunit of chromatin structure, is made up of 147 bp of DNA wrapped around a histone octamer core. One of the important roles of the nucleosomes is stabilising and packing the long strand of DNA. The positively charged histones stabilise the negatively charged DNA backbone and make it possible to be packed in a cell. However, the role of nucleosomes is not limited to the packaging and stabilising DNA strands. Nucleosomes play regulatory roles such as gene activation (Bai and Morozov, 2010; Hu et al., 2013; Spitz and Furlong, 2012), transcription repression (Lickwar et al., 2009), or polymerase pausing (Levine, 2011). This regulatory activity is accomplished by limiting the access of the transcription factors to their binding sites on the DNA or bringing them close together by bending the DNA (Spitz and Furlong, 2012). Also, nucleosomes may contribute to the 3-dimensional structure of DNA and form transcription centre (Chakalova et al., 2005). By closing in or opening important regions of DNA, nucleosomes may block the binding of the transcription factors or facilitate the binding of transcription factors. It may even direct the binding to a specific region and increase the transcription efficiency. Therefore, knowing the precise position of a nucleosome, especially regarding the transcription start site, is important to understand the interaction between the nucleosome and other gene regulatory elements.

Nucleosomes are positioned either statistically or specifically (Kornberg and Stryer, 1988; Mavrich et al., 2008a). Statistic positioning of nucleosome represents nucleosomes are positioned freely and randomly on DNA as far as there is enough space: the nucleosome position is restricted by adjacent nucleosomes or DNA binding proteins. Otherwise, it can be placed in any available space on the DNA. On the other hand,

specific nucleosome positioning represents that the nucleosome position is closely regulated. The underlying DNA sequences (Collings et al., 2010; Ioshikhes et al., 2006; Peckham et al., 2007; Segal et al., 2006) or nucleosome remodelling factors (Chandy et al., 2006; Ito et al., 1997; Sadeh and Allis, 2011) are the major controllers of the specific positioning. Especially the DNA sequences are the fundamental factor in the specific nucleosome positioning.

In contrast to the well-defined, 4 to 10 bp long sequence motifs for transcription factor binding, the nucleosome sequence—the DNA region where histones bind to form a nucleosome—spreads over longer than 100 bp. Typically, the length of a nucleosome sequence is 146 to 147 bp for a canonical nucleosome (Kornberg, 1974). While the sequence motifs of transcription factor binding sites are relatively compact and well-defined, the nucleosome positioning sequences do not show specific sequence motifs. Instead, a dinucleotide periodicity, repeating appearances of dinucleotide in fixed intervals is the common characteristics of nucleosome positioning sequences (Ioshikhes et al., 1992). The periodic pattern of dinucleotides allows or restricts the bending of the DNA and the bending favours or disfavours the DNA wrapping the histone cores to form a nucleosome (Cui and Zhurkin, 2010; Gabdank et al., 2009; Zuccheri et al., 2001). Various NPS patterns have been proposed in eukaryotes (Ioshikhes et al., 2006, 2011; Salih et al., 2008; Segal et al., 2006). Among the many proposed patterns, the dinucleotide periodicity is commonly observed in the various NPS patterns, even though the proposed patterns differ from one another.

The dinucleotide NPS patterns are customarily classified into WW/SS (weak-weak/strong-strong) class and RR/YY (purine-purine/pyrimidine-pyrimidine) class as WW/SS and RR/YY are the most frequent dinucleotides showing periodicity in the NPS. The WW/SS patterns have alternating WW and SS dinucleotides at every 5 bp distances,

which leads to the periodic appearance of WW or SS dinucleotide at 10 bp apart of its own. The 10 bp periodicity locates the WW or SS dinucleotides on one side of the nucleosome DNA because the 10 bp periodicity is similar to the one turn of DNA helix. For the same reason, the WW and SS dinucleotides are located on the opposite side of the DNA as the distance between WW and SS dinucleotides is 5 bp. The location of WW dinucleotides on the nucleosomal DNA is where the major groove of the DNA faces away from the histone core. The SS dinucleotides, which are 5 bp apart from the WW dinucleotides, are thereby located where the major groove faces toward the histone.

The RR/YY NPS pattern is also characterised by the 10 bp periodic appearances of RR and YY dinucleotides. However, the positions of the dinucleotides are in counter-phase, which means they are at the same position but on the different strands of DNA. For example, AA (purine) and TT (pyrimidine) dinucleotides in the RR/YY pattern are in the counter-phase or on the different DNA strand. The locations of the AA and TT dinucleotides are 2 to 3 bp off from the WW and SS locations. So the locations are where the DNA backbone is close to the histones (Ioshikhes et al., 1996). Based on the locations of the dinucleotides in the NPS patterns, the WW/SS pattern may be more important regarding the binding of transcription factors or pioneer factors because the locations are related to the opening or squeezing of the major and minor grooves.

It was reported that some nucleosome sequences have anti-NPS patterns. The anti-NPS pattern is the opposite pattern of the corresponding NPS pattern (Ioshikhes et al., 2011). As the NPS pattern facilitates wrapping histone octamers by favouring the DNA bending, the opposite anti-NPS patterns need higher energy to bend the DNA to be wrapped around the histone octamers. Consequently, the nucleosome formed at the DNA with the anti-NPS pattern is less stable than those formed at the DNA with the NPS

pattern. The instability of the nucleosome can facilitate the binding of the transcription factors to DNA in place of the histones and initiate the transcriptional regulation.

Histone variant H2A.Z, which shares 60% identify with H2A, is characterised by its extended acidic patch on the surface and a unique C-terminal tail (Suto et al., 2000). The H2A.Z nucleosome, which contains the histone variant H2A.Z in place of H2A, is regarded an activator of the gene expression because the H2A.Z nucleosomes are often enriched at the promoters of active genes. However, the reported role of H2A.Z nucleosome in gene expression is not consistent but complex. Some of the phased H2A.Z nucleosomes at the 5' end of genes are related to the active transcription of the gene (Bargaje et al., 2012; Weber et al., 2010), while the enrichment of H2A.Z is observed of inactive genes in yeast (Guillemette et al., 2005). In some cases, the H2A.Z nucleosomes are enriched at the 5' ends of genes regardless of the gene activity and are lost following an increase in transcription (Raisner et al., 2005). H2A.Z affects the chromatin status: the H2A.Z enrichment is related to gene silencing (Swaminathan et al., 2005), or H2A.Z antagonises telomeric silencing and prevents the spread of heterochromatin by collaborating synergistically with a boundary element (Meneghini et al., 2003). Perplexing enough, H3.3 dependent recruitment of H2A.Z on promoters facilitates the compaction of chromatin to poise transcription, while H2A.Z mediates chromatin compaction and represses transcriptional activity (Chen et al., 2013). One of the speculated roles of the H2A.Z nucleosome is that it establishes transcription initiation instead of maintaining it (Santisteban et al., 2000).

I hypothesised that the NPS pattern in Drosophila determines the formation of nucleosomes at the regulatory regions as in the yeast. I analysed the nucleotide sequences of experimentally determined nucleosome positions to test if Drosophila nucleosomes have the NPS of the 10 bp periodicity comparable to the yeast patterns. I also

hypothesised that the NPS pattern of the canonical H2A nucleosome might differ from the patterns of the H2A.Z nucleosomes. The DNA sequences from the phased H2A and H2A.Z nucleosomes were analysed for periodicity, and the dinucleotide positions were compared. The positions of the nucleosomes on the genome were identified by aligning the sequence reads from ChIP-seq experiments. The nucleosome positions from the aligned sequence reads were determined, and the phased +1 nucleosomes were selected in the promoter regions. After retrieving the nucleotide sequences, I analysed the periodicity of the dinucleotides. The signal for the nucleosome positioning is subtle being stretched over a 146 bp long sequence. So the collective measures by summing up the dinucleotide frequencies of the selected sequences were necessary to increase the signal to noise ratio. To identify the periodicity in the nucleosome sequences, I used Fourier transform to convert the periodic patterns into a frequency domain with its spectral density. By examining the spectral density, the dominant frequency that is the inverse of the period could be identified. The detection of the NPS patterns required iterating processes to refine the patterns after removing noises: repeated selecting of the nucleosome sequences and running Fourier transform. The NPS patterns from the H2A and H2A.Z nucleosome sequences were compared with the yeast patterns for the detection of distinctions of the Drosophila NPS patterns and the effects of histone variants on the NPS patterns.

I also hypothesised that the genome sequences were the basic layout organising the H2A and H2A.Z nucleosome positions, and they may facilitate the binding of certain factors or interfere with the binding of other transcription factors. Hence, I scanned the promoters searching for the existence of the core promoter elements, such as BRE, Inr, CCAAT, and TATA followed by grouping the promoters according to the nucleosome type and the position. I then tested the co-occurrences of the core promoter elements with

the nucleosomes statistically. To investigate whether the genomic organisation has preferences for the type of nucleosomes such as H2A or H2A.Z and preferences for the position of the nucleosome, I analysed the enriched biological functions of the genes associated with H2A and H2A.Z nucleosomes and the NPS patterns.

Materials and Methods

IDENTIFICATION OF THE NUCLEOSOME POSITIONS

H2A nucleosome data were downloaded from Gene Omnibus (GSE30755), which are paired-end Illumina HiSeq sequencing data of 80 mL salt precipitated nucleosomal DNA from Drosophila S2 cell line (Teves and Henikoff, 2011). The quality of the short sequence reads was checked with FASTQ for the duplicates and non-specific repeats. The quality-checked sequence reads were aligned on the Drosophila R5.3 genome using Bowtie2 (Marygold et al., 2013). The paired sequence reads were identified with the sequence identification tag. Each position of the aligned reads of the paired sequence reads defines the beginning and the end of the sequenced DNA fragments.

After alignment, we selected the positions that were bound by phased nucleosomes. Because the short reads came from the paired sequencing reads, the size of the sequenced segments could be identified. The sequenced genomic segments with the size between 142 bp and 146 bp were selected for further analysis to restrict the selected sequences mainly bound by the phased nucleosomes. The middle positions of the selected sequence reads were found from the boundaries defined by the paired reads. From the genomic coordinates of the central positions, the read counts, the number of the aligned sequences residing at each position of the genome, were counted. The read counts on the genome were further analysed by Gene Track (Albert et al., 2008) to identify nucleosome peaks. Gene Track identified the nucleosome dyad or the centre of the nucleosome from the peaks with the highest read counts locally without being overlapped with other peaks within the 147 bp range. The identified nucleosome positions were filtered with the read count and the standard deviation of the dyad positions. The nucleosome positions with more than 2 read counts and the standard deviation less than the 75% quantile were selected as reliable nucleosome positions. The peak identification was made separately on

the Watson and Crick strands of the genome. Then, the identified nucleosome peak positions from one strand were compared to peaks in another strand, selecting peaks whose positions matched within 2 bp between the two strands.

The output of the Gene Track is the point coordinates of the nucleosome dyads on the genome. The genomic position of the nucleosome was determined by extending the dyad positions by 73 bp to the 5' and 3' directions to set the beginning and the end of the nucleosome boundaries. The nucleosome positions were saved in the 0-based BED file. The nucleotide sequences of the nucleosome-bound regions were fetched from the same Drosophila R5.3 genome using the genomic coordinate of the nucleosome positions with Bedtools (Quinlan and Hall, 2010a). The fetched genomic sequences were saved in FASTA format for further analysis. H2A.Z nucleosome positions were obtained from immunoprecipitated H2A.Z histone of *Drosophila melanogaster* embryo (Mavrich et al., 2008b). The H2A.Z nucleosome positions were identified in the same way as those of the H2A nucleosomes.

NUCLEOSOME DISTRIBUTION

Nucleosome occurrences were counted within the 2000 bp long sequence around transcription start sites. The identified nucleosome positions were counted if the dyad position resided within the 1000 bp upstream and 1000 bp downstream range of transcription start sites. The nucleosome counts were smoothed by taking the 3-bp window moving average: taking the average of the nucleosome counts at the current position, at the 1 bp previous position, and the 1 bp following position. The nucleosome counts were normalised by dividing the counts by the total number of the nucleosomes. The normalised occurrences at each position were plotted by aligning at transcription start sites.

IDENTIFYING THE +1 NUCLEOSOMES OF EACH GENE

The '+1 nucleosome' was defined as the first nucleosome located within 200 bp downstream from a transcription start site. The genomic coordinates of the transcription start sites of all known genes were obtained from the Drosophila melanogaster database using BioMart (Durinck et al., 2005) and R (R Development Core Team, 2012). We searched the presence of a nucleosome within the region which was 200 bp downstream from the transcription start site of each gene. The first nucleosome found in the 200 bp downstream region was marked as the +1 nucleosome of the corresponding gene. We searched for the +1 nucleosome from the H2A nucleosome data and repeated the same search with the H2A.Z nucleosome data. The +1 nucleosome was defined as the first downstream nucleosome within the 200 bp range of transcription start sites. Each gene has only one +1 nucleosome in our study. Once the +1 nucleosome was identified with each gene, the nucleosome and the corresponding gene pair were stored in one of two groups, an H2A +1 nucleosome-gene data set or an H2A.Z +1 nucleosome-gene data set, depending on the type of the +1 nucleosome. The +1 nucleosomes are located within the 200 bp range from transcription start sites. The *closest* command in the Bedtool suite was used to identify the nucleosomes closest to a transcription start site by comparing the genomic coordinates of all transcription start sites of known genes and the nucleosomes located within the 200 bp range from the transcription start sites.

GROUPING GENES BASED ON THE TYPE OF THE +1 NUCLEOSOME

Once the +1 nucleosome for each gene was identified, the nucleosome-gene pairs were further divided into three groups: H2A-only, H2A.Z-only, and H2A/H2A.Z. For each gene, its +1 nucleosome positions were compared between the H2A and the H2A.Z nucleosome. If only an H2A nucleosome was found as the +1 nucleosome for the gene, then the gene-nucleosome pair was stored in the H2A-only group; if only an H2A.Z

nucleosome was found as the +1 nucleosome for the gene, then the gene-nucleosome pair was stored in the H2A.Z-only group.; if both H2A and H2A.Z nucleosomes were found as the genes +1 nucleosome and the positions of the two +1 nucleosomes were overlapped at least by 73 bp, then the gene-nucleosome pair was stored in the H2A/H2A.Z group. The genomic coordinates of the transcription start sites, the +1 nucleosome position and the associated gene name were saved for further analysis. Comparison of the genomic coordinates of the nucleosome positions was performed with Bedtools *intersect* command.

DINUCLEOTIDE PATTERNS OF NUCLEOSOME SEQUENCES

Once the genomic coordinates of the nucleosomes were determined, the nucleotide sequences of the determined genomic coordinates were fetched from the Drosophila genome R5.3 using the *getfasta* command in the Bedtools suite. A unique number was assigned to the genomic coordinates for identification, and output was saved to a file in FASTA format.

The frequency patterns of the 20 dinucleotides were counted from the nucleotide sequences. The 20 dinucleotides consist of 16 primary patterns, which are the combinations of the four nucleotides, adenine, thymine, guanine, and cytosine (AA, AT, TA, TT, GG, GC, CG, CC, AG, GA, AC, CA, TG, GT, TC, CT) and four composite patterns of weak-weak (WW), strong-strong (SS), purine-purine (RR), and pyrimidine-pyrimidine (YY). Reverse complement strands were added for each sequence before counting the dinucleotide occurrences to consider the dinucleotide frequency in both strands reading from 5’ to 3’ direction.

Each nucleotide sequence was converted into a numerical format based on the presence of the dinucleotide. An example of the numeric representation of the AT and

CG dinucleotide patterns of a sequence is shown below. Note that the length of the dinucleotide pattern is 1 bp shorter than the length of the original sequence.

```
Nucleosome sequence:        5’-CGAGTATCGGAATCGTATGCC-3’
AT pattern:                 5’-0000010000010000 1000-3’
CG pattern:                 5’-1000000100000 1000000-3’
```

For each nucleotide sequence, the sequence was converted by custom Perl scripts to the numeric format for all 20 dinucleotide patterns.

The numeric patterns were analysed after being imported into R. The numbers at each position were summed through the complete set of the sequences to count the dinucleotide at each position and then divided by the total number of the sequence to calculate the occurrence frequency. The occurrence frequency for each dinucleotide pattern was then smoothed by 3-bp moving average and plotted along the nucleosome sequence aligned at the nucleosome dyad to look for the repeating appearance of the peaks. The smoothed pattern was used for periodicity analysis by Fourier transform.

DETECTING PERIODICITIES BY FOURIER TRANSFORM

Fourier transform converted the domain of the original data to the frequency domain. It detects the frequency from the repeating patterns and the contribution of each frequency to the plot. If the dinucleotide patterns had trends, then the trends were eliminated to make the patterns stationary. The detrending was done with LOESS (Locally Weighted Scatterplot Smoother) curve fitting. The pattern with a trend was fit with LOESS, and the predicted values were subtracted from the patterns. The resulting patterns were stationary without the trend. The occurrence of each dinucleotide pattern was converted to the frequency domain by Fourier transform, and then the spectral density of each frequency was compared to find the dominant frequency. The period can

be calculated as $1/frequency$ because our sampling rate is one base pair. The period and the spectrum were plotted using the *ggplot2* package in R.

For the composite patterns like WW, SS, RR, and YY, the dinucleotide counts of the corresponding dinucleotides were summed. For example, for the WW dinucleotide patterns, AA, AT, TT, and TA dinucleotide counts were summed. The patterns went through the same smoothing and Fourier transform for the detection of the periodicity.

BUILDING A MODEL FOR THE PERIODIC DINUCLEOTIDE PATTERN

Linear models were constructed using the identified period for the periodic dinucleotide pattern. The following equation was fitted using a linear model to estimate the parameters.

$$P = a\sin(2\pi\omega x) + \mathrm{b}\cos(2\pi\omega x) + e$$

Where ω is the identified frequency (1 / period), x is the nucleosome sequence position, and P is the dinucleotide occurrence.

The harmonic period was added to the model:

$$P = \mathrm{a}\sin(2\pi\omega x) + b\ \cos(2\pi\omega x) + \mathrm{c}\sin(4\pi\omega x) + d\cos(4\pi\omega x) + e$$

The equation was fitted using a linear model and the estimated parameters

REFINING OF THE PATTERNS BY CORRELATION

Correlation of individual nucleosome sequences to the particular dinucleotide patterns was calculated as in (Ioshikhes et al., 2011). For each pattern, the correlation was calculated between the initial dinucleotide pattern and each dinucleotide pattern. For example, the relationship between the initial WW dinucleotide pattern from the data set and the WW dinucleotide pattern of each nucleosome sequence. Then the nucleosome sequences were grouped based on the correlation coefficient. If the correlation coefficient is greater than 0, then the sequence was marked as positively correlated sequences and

stored together. If the correlation coefficient is less than 0, then the sequence was designated as negatively-correlated sequences and stored for further analysis. The calculation of the occurrence frequency and the detection of the significant periods by Fourier transform were carried out with the selected nucleosome sequences in the same way as previously described.

COMPARISON OF THE HISTONE STRUCTURES

Histone H3 sequences of yeast (P61830), Drosophila (P02299) and H3.1 sequences of mouse (P68433) and human (P68431) were retrieved from UniProt. The amino acid sequences were aligned using Clustalw2 (McWilliam et al., 2013). H2A histone sequences of yeast (P04911), Drosophila (P84051) and human (P0C0S8), and H2A.Z sequences of yeast (Q12692), Drosophila (P08985) and human (P0C0S5) were retrieved from UniProt. Those sequences were compared by multiple alignments using Clustalw2 (McWilliam et al., 2013) to find the conserved regions and the changes in the hydrophobicity. The tertiary structures of nucleosome core with H2A histone (2NQB) and H2A.Z histone (1F66) nucleosomes were retrieved from Protein Data Bank. The tertiary structures were superimposed on each other with PyMol to visualise the differences.

IDENTIFYING CORE PROMOTER ELEMENTS

The core promoter elements were identified from promoter sequences extracted from transcription start sites. The genomic coordinates of the transcription start sites of all known genes were expanded to 500 bp toward upstream (5’ direction) and 100 bp toward downstream (3’ direction). The 600 bp long nucleotide sequence was retrieved from the Drosophila R5.3 sequence with Bedtools. Each promoter was uniquely labelled with the associated gene name and the position. The promoter sequences were scanned

for the presence of core promoter elements, BRE, DPE, Inr, CCAAT, and TATA with Promoter Classifier (Gershenzon and Ioshikhes, 2005).

STATISTICAL TESTING OF THE CO-OCCURRENCES

Co-occurrence of the core promoter elements and the H2A or H2A.Z +1 nucleosome was searched to assess the enrichment of the core promoter elements. The transcription start sites of the promoter sequences were associated with known genes. The H2A and H2A.Z +1 nucleosomes were also related to known genes from the previous analysis. Those gene lists were compared to group the promoter sequences. Each core promoter elements were counted in grouped promoter sequences based on the H2A and H2A.Z +1 nucleosome. The enrichment of the core promoter elements was tested by Chi-square test. The total number of each core promoter element was counted and used as the expected value, and the count of each core promoter element in each gene set was used as the observed value. The processing and plotting of the data, not specified above, was done by custom scripts written in R using the ggplot2 package (R Development Core Team, 2012; Wickham, 2009)

ENRICHED BIOLOGICAL FUNCTIONS OF THE H2A AND H2A.Z +1 NUCLEOSOME

The selected genes in the H2A-only, H2A.Z-only, and H2A/H2A.Z subset were analysed for the enriched biological functions with Functional Annotation Clustering of DAVID web server (Huang et al., 2009). The standard gene symbols of the selected genes were prepared and used as the input of the DAVID. Enriched biological processes were chosen based on the enrichment score (greater than 2) and Benjamini-Hochberg false discovery rate (less than 0.05).

ENRICHED BIOLOGICAL FUNCTIONS OF GENES HAVING NPS PATTERNS

The correlation of nucleosome sequences to the WW/SS and RR/YY NPS patterns were calculated, and the positively correlated sequences were selected for both WW/SS and RR/YY patterns. The associated genes to the selected sequences were identified by the unique identification number given for the nucleosome sequences in the previous steps. The standard gene symbols were prepared and used as an input to the Functional Annotation Clustering in DAVID. Enriched biological processes were selected by the enrichment score (greater than 2) and Benjamini-Hochberg false discovery rate (less than 0.05).

Results

NUCLEOSOME DISTRIBUTION AROUND TRANSCRIPTION START SITE

Nucleosomes cover a large area of a chromosome. Besides the structural role of nucleosomes, the sequences covered by nucleosomes may not be readily available to other proteins to bind. If the covered chromosome regions are important in transcriptional regulation, the nucleosome coverage itself is part of the regulation. Nucleosomes are located either statistically or specifically, and the ones near promoters are believed to be positioned specifically. Selecting specifically positioned nucleosomes is important in searching for the nucleosome positioning sequence. The nucleosome distribution along the promoters gives information not only where nucleosomes are located but also gives a clue whether the nucleosomes are located statistically or specifically. If the nucleosomes are located specifically, which means they are positioned on designated positions, then the nucleosome distribution will show a definite profile at the nucleosome position. Otherwise, the nucleosome positions will spread out and the nucleosome distribution will produce a broad and low profile.

The H2A and H2A.Z nucleosome distributions around the Drosophila transcription start sites were analysed from the experimental data. Nucleosomes within the 1000 bp upstream and 1000 bp downstream region of the transcription start sites, were collected, and summed at each position across all known genes of Drosophila (**Figure 1**). The closer the nucleosome is to the transcription start sites, the sharper the profile is. More specifically, the first nucleosome at the downstream of the transcription start site has the strongest signal, and the subsequent nucleosomes away from the transcription start site are showing less defined profiles. The stronger signal means nucleosomes at the position were located specifically. The strong and sharp profile is the

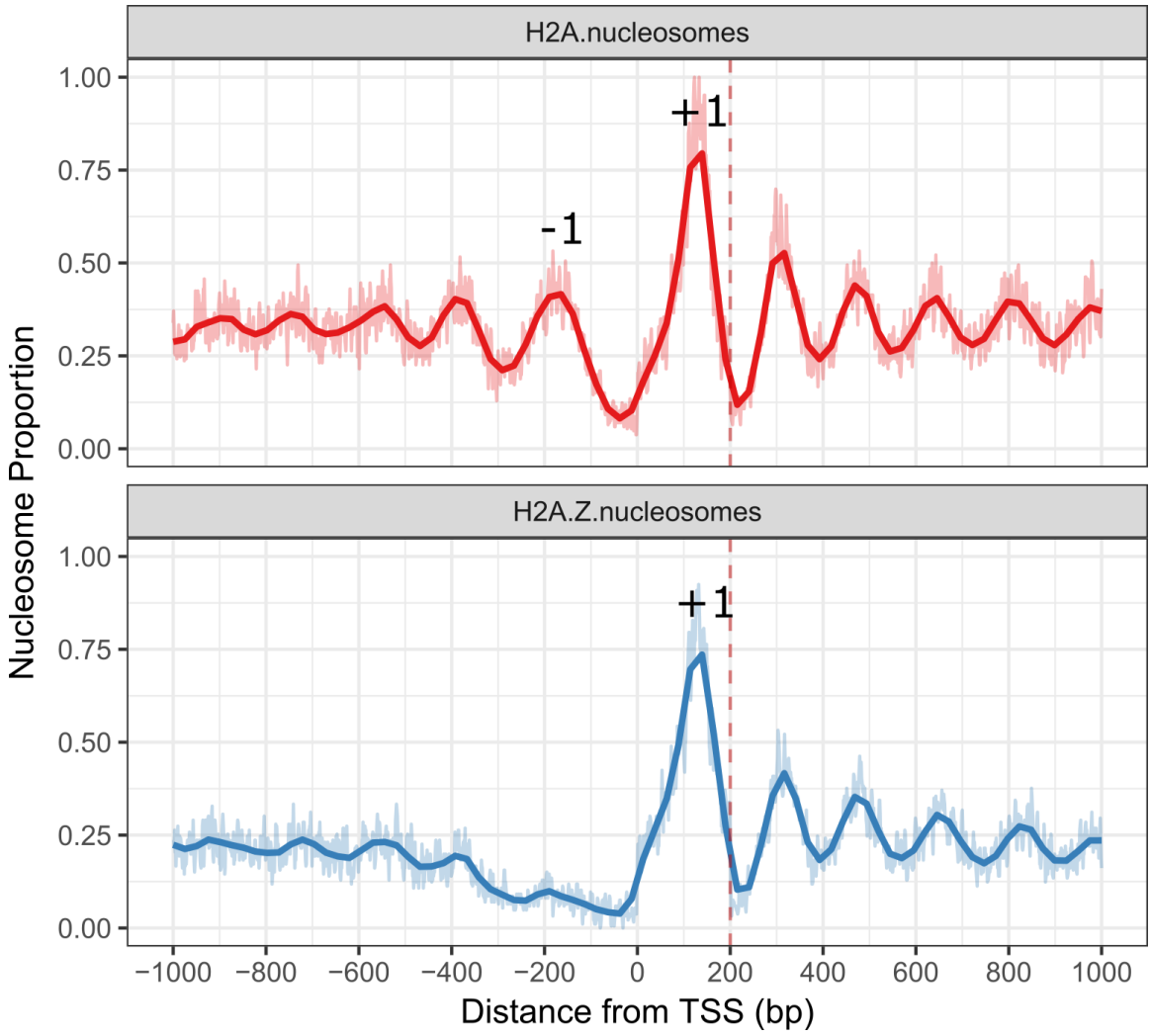

H2A.nucleosomes
H2A.Z.nucleosomes
+1
-1
+1
Nucleosome Proportion
1.00
0.75
0.50
0.25
0.00
−1000
−800
−600
−400
−200
0
200
400
600
800
1000
Distance from TSS (bp)

Figure 1. Nucleosome distributions on promoters. The distributions of H2A and H2A.Z nucleosomes were shown at 5' of genes by plotting normalised nucleosome occurrences. The genomic regions were aligned at the transcription start site (TSS). The upstream region was represented with a (-) notation, and the downstream region was represented with a (+) notation. The smoothed nucleosome occurrences (bold line) were calculated using the 3-bp moving average of the raw occurrences (thin line). The +1 nucleosome was defined as the first nucleosome within the 200 bp downstream of a TSS. The 200 bp boundary was marked by a dashed line. (A) H2A nucleosomes, including the +1 nucleosome and adjacent nucleosomes in both upstream and downstream regions are strongly positioned. (B) The +1 nucleosome and the downstream nucleosomes of H2A.Z were strongly positioned as the H2A nucleosomes. However, the lack the -1 nucleosomes, the nucleosome located at the immediate upstream of TSS, and the fuzzy distribution of the upstream nucleosomes differentiate the H2A.Z distribution from the H2A distribution.

sign of the phased nucleosome. The phased nucleosome was located within 200 bp from the transcription start site regardless of the H2A or H2A.Z nucleosome.

Conventionally, nucleosomes are numbered starting from the closest one to the transcription start site. The relative position of the transcription start site was denoted by giving (+) or (-) signs. The ones located at the upstream of the transcription start site are denoted with a minus sign (-), and the ones at the downstream of the transcription start site are denoted with a plus sign (+). So, the +1 nucleosome is the first nucleosome downstream of the transcription start site, and the +1 nucleosome is the most strongly positioned nucleosome. The positioning of the +1 nucleosome is not determined randomly by the space availability but directed by specific factors especially the nucleotide sequences. The nucleotide sequences of the strongly positioned +1 nucleosomes were selected for the analysis of the nucleosome positioning sequence.

Note that the H2A.Z nucleosome distribution lacks the -1 nucleosome, unlike the H2A nucleosomes distribution. The lack of -1 nucleosome forms a nucleosome free region (NFR) at the upstream of transcription start sites. The dynamic and less stable positioning of nucleosomes instead of non-binding accounts for the forming of NFR (Weiner et al., 2010). The fact that NFR is observed only in the H2A.Z nucleosome distribution but not in the H2A nucleosome distribution supports that nucleosome positioning in the promoter is rather specific than random.

SELECTION OF GENES BASED ON THE +1 NUCLEOSOME

We defined the +1 nucleosome as the first nucleosome appearing within 200 bp downstream of a transcription start site. The +1 nucleosome is the most phased one. The transcription starts sites of the all known genes were grouped depending on the type of

the +1 nucleosome as described in the Methods. The selection process of the genes and the selected number of genes based on the +1 nucleosome are shown in **Figure 2**. We started from the all identified nucleosome positions, and then selected one +1 nucleosome per transcription start site of each gene. The underlying nucleotide sequences were extracted from the Drosophila genome using the genomic coordinates of the +1 nucleosome positions.

The nucleosome distributions after the selection are shown in **Figure 3.** The selection was verified by the lack of H2A +1 nucleosome from the H2A.Z-only data set and H2A.Z +1 nucleosome from the H2A-only dataset. The distribution of the downstream nucleosomes from transcription start sites were affected by the presence of the +1 nucleosome. The H2A nucleosomes after the +1 nucleosome are well positioned in the H2A-only gene set, but the distribution of the H2A.Z nucleosomes in the same gene set, which lacks the H2A.Z +1 nucleosome was ill-defined (**Figure 3A**). In the same way, the H2A.Z distribution in the H2A.Z-only data set was well defined than the H2A distribution of the same gene set (**Figure 3C**). As expected, the distribution of both H2A and H2A.Z nucleosomes were well positioned in the H2A/H2A.Z gene set. The fuzziness of the distribution conforms to the previous findings that the downstream nucleosomes after the +1 nucleosomes appeared to be positioned statistically, while the +1 nucleosome was positioned specifically (Mavrich et al., 2008a).

The +1 nucleosome distribution was examined more in detail in 200 bp range from the transcription start site. The genes in the three groups, H2A-only, H2A.Z-only, and H2A/H2A.Z, made it possible to compare the nucleosome positions more precisely. Even though both H2A and H2A.Z nucleosomes were phased within the 200 bp regions, the precise positions were slightly different. The H2A.Z nucleosomes were apt to be

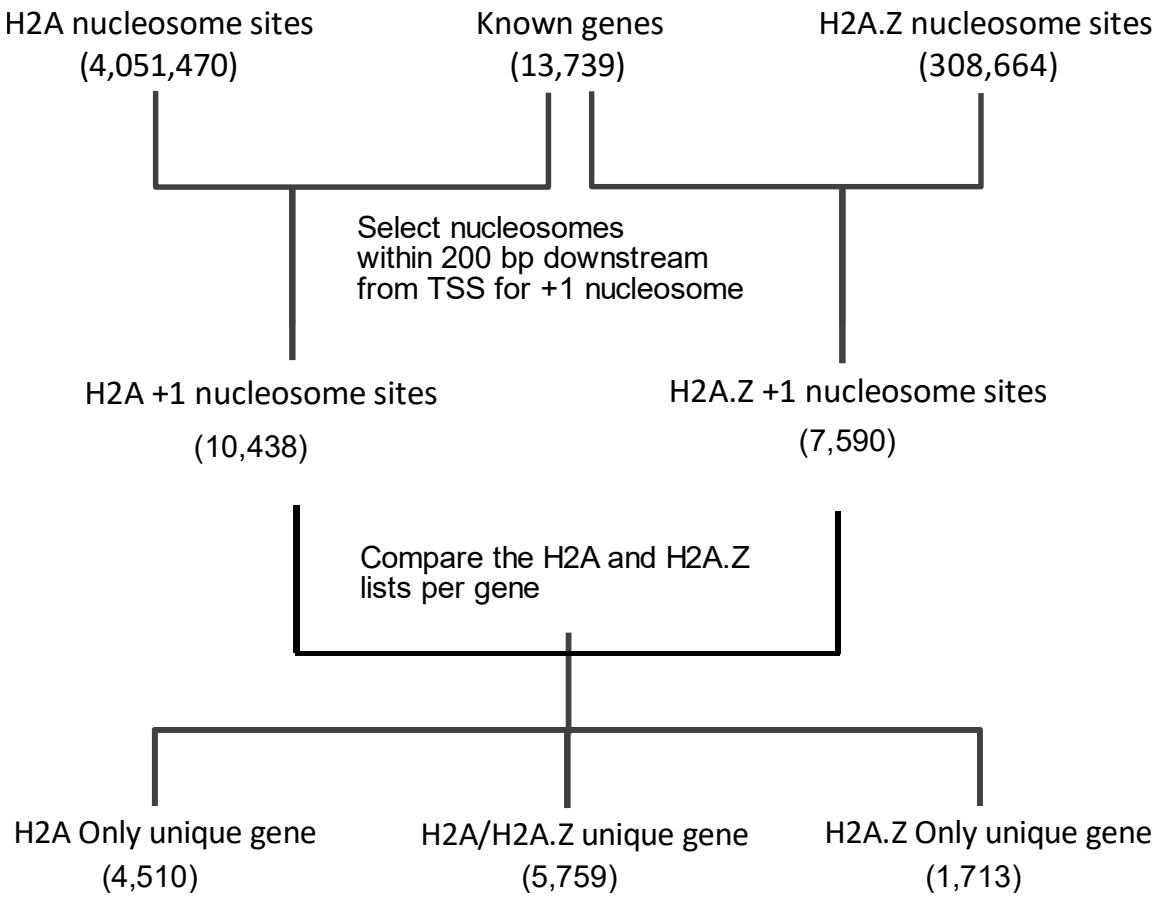

H2A nucleosome sites
(4,051,470)
Known genes
(13,739)
H2A.Z nucleosome sites
(308,664)
Select nucleosomes
within 200 bp downstream
from TSS for +1 nucleosome
H2A +1 nucleosome sites
(10,438)
H2A.Z +1 nucleosome sites
(7,590)
Compare the H2A and H2A.Z
lists per gene
H2A Only unique gene
(4,510)
H2A/H2A.Z unique gene
(5,759)
H2A.Z Only unique gene
(1,713)

Figure 2. Selection scheme of +1 nucleosomes and the associated genes. The genomic coordinates of the identified nucleosome positions from experimental data were compared with the TSS positions to find the +1 nucleosome of each gene. Because the +1 nucleosome was defined as the first nucleosome within 200 bp downstream of a TSS, each of the +1 nucleosomes could be associated with a gene as well as a TSS. The H2A and H2A.Z +1 nucleosome were identified separately. Then the +1 nucleosome was found both in H2A and H2A.Z nucleosomes; then the gene was marked as H2A/H2A.Z. If the +1 nucleosome was found either in the H2A or the H2A.Z list, then the gene was marked as H2A-only or H2A.Z-only, respectively. The marked pairs of the gene and the +1 nucleosome were saved for further analysis. The number of the genes and the nucleosome sequences throughout the selection processes were presented in the parenthesis.

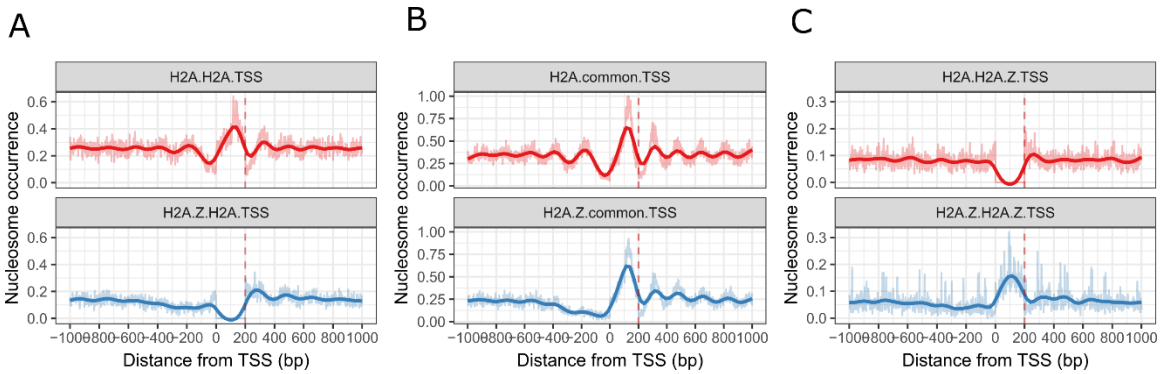

A
B
C
H2A.H2A.TSS
H2A.Z.H2A.TSS
H2A.common.TSS
H2A.Z.common.TSS
H2A.H2A.Z.TSS
H2A.Z.H2A.Z.TSS
Nucleosome occurrence
Distance from TSS (bp)

Figure 3. Nucleosome landscape around transcription start sites. The raw distribution (thin line) and the 3-bp smoothed nucleosome occurrences (bold line) of H2A (red) and H2A.Z (blue) nucleosomes on the 2000 bp around the transcription start sites are presented by aligning the transcription start site (TSS) at the centre. The upstream of a TSS was represented as the (-) distance, while the downstream was represented as the (+) distance. The H2A (red) and H2A.Z (blue) nucleosome distributions were calculated from the experimentally determined positions of H2A and H2A.Z nucleosomes, respectively. The nucleosome distributions from the three groups indicate the presence or the lack of the +1 nucleosome depending on the group, which verifies the selection process. (A) The nucleosome distributions in the H2A-only group. The H2A distribution (red) shows the presence of the H2A +1 nucleosome. The H2A.Z distribution (blue) lacks the +1 nucleosome. (B) The nucleosome distributions of the H2A/H2A.Z group. Both H2A (red) and H2A.Z (blue) distributions show the well-positioned +1 nucleosomes. The downstream nucleosomes were well-positioned also. (C) The nucleosome distributions of the H2A.Z-only group. The H2A distribution (red) lacks the +1 nucleosome, while the H2A.Z distribution has the +1 nucleosome.

phased at 100 bp, while H2A nucleosomes were phased at 80 bp and 120 bp positions downstream of TSS (Figure 4**A**). The H2A and H2A.Z +1 nucleosome distribution from each H2A-only and H2A.Z-only gene set were compared (Figure 4**A**). The differences suggest that H2A and H2A.Z may have different preference to the nucleotide sequences for positioning and the sequences differences of the promoters may affect the positions of the phased nucleosome. On the other hand, the H2A and H2A.Z nucleosomes from the H2A/H2A.Z gene set showed similar distribution suggesting the sequences may be preferred by both nucleosomes. (Figure 4**B**). We chose the phased +1 nucleosome sequences to identify nucleosome positioning sequences. We hypothesise that the +1 nucleosomes were positioned by the DNA sequences so that the +1 nucleosome-bound sequences were reliable sources to detect the nucleosome positioning sequences.

SEQUENCE ANALYSIS OF NPS PATTERNS OF H2A NUCLEOSOME SEQUENCES.

From the extracted nucleosome-bound sequences, we looked for nucleosome positioning sequence patterns, especially the 10 bp periodic occurrence of dinucleotides, which is the most well-known nucleosome positioning pattern in other organisms. We calculated the dinucleotide frequencies of 20 dinucleotides. The 20 dinucleotides consisted of 16 dinucleotides, which are the combinations of the four nucleotides, adenine (A), thymine (T), guanine (G), and cytosine (C). In addition to that, the patterns of four more composite dinucleotides were examined: weak (A and T), strong (C and G), purine (A and G), and pyrimidine (T and C). The four composite patterns were denoted as weak-weak (WW), strong-strong (SS), purine-purine (RR), and pyrimidine-pyrimidine (YY). First, each 146 bp long nucleosome sequence of a given gene set was converted to the corresponding dinucleotide code as described in the Methods. We added the reverse complementary sequence for each sequence because the NPS patterns have no

directionality. The occurrences of each dinucleotide were counted for all nucleosome sequences in the H2A-only data subsets using the converted dinucleotide codes.

Initial Dinucleotide patterns

From the initial dinucleotide patterns, 10 bp periodic occurrences were observed in WW and SS patterns **(Figure 5A**). The WW peaks were located at 15, 25, 35, 45, 55 bp positions from the nucleosome dyad, or centre. The positions of the WW peaks are where the major groove of the nucleosomal DNA is exposed to the outside. The SS patterns also showed repeating peaks which were 10 bp apart lying at 20, 30, 40, 50, 60 bp from the nucleosome dyad. The position of the SS peaks was 5 bp offset from the WW peaks, which

resulted in the SS peaks located where the minor grooves are exposed to the outside. The other dinucleotide patterns were also analysed. The AT and TA dinucleotides, which are part of the WW pattern, showed the 10 bp repeating peaks. In the same way, GC pattern showed the repeating peaks matching the SS pattern (**Figure 6**).

The periodicity of the WW and SS patterns were further analysed by Fourier transform. Fourier transform decomposes the pattern into periods and the spectral density. The spectral density indicates the contribution of the period in the repeating pattern. The analysis confirmed the visually recognized10 bp periodicity in the WW and SS patterns **(Figure 7**). The 10 bp period of the WW pattern has the strongest spectral density. The AA, TT, and AT patterns, which belong to the WW pattern, showed 10 bp period. They have other periods than the 10 bp according to the Fourier transform. The SS pattern showed a longer as well as the 10 bp period. Interestingly, the 10 bp period is the dominant period in the GG, CC, and CG dinucleotide patterns, even though the SS

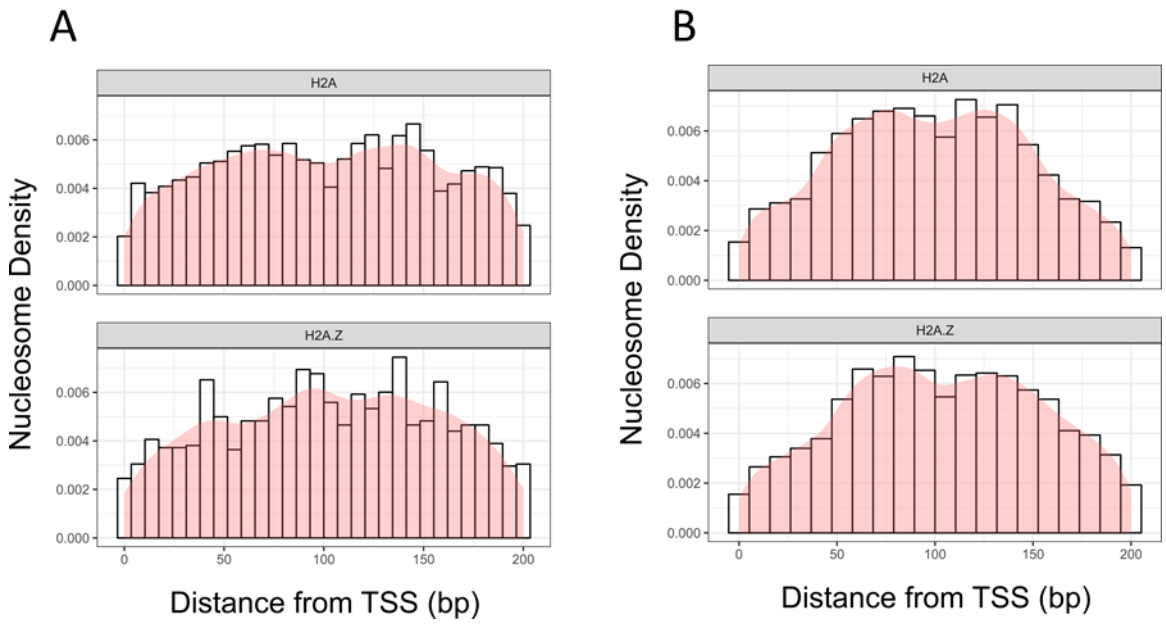

A
H2A
H2A.Z
Nucleosome Density
0.000
0.002
0.004
0.006
0
50
100
150
200
Distance from TSS (bp)
B
H2A
H2A.Z
Nucleosome Density
0.000
0.002
0.004
0.006
0
50
100
150
200
Distance from TSS (bp)

Figure 4. Distributions of the H2A and H2A.Z nucleosomes. The distributions of the +1 nucleosomes showed the precise positions of the +1 nucleosomes. (A) Comparison of the +1 nucleosome distributions between the H2A-only and the H2A.Z-only groups. The H2A +1 nucleosomes were positioned at 80 bp and 120 bp downstream from the TSS, while the H2A.Z +1 nucleosomes were positioned at 90 bp from the TSS. (B) Comparison of the +1 nucleosome distributions in the H2A/H2A.Z group. The H2A and the H2A.Z distributions on the same gene set were compared. The two nucleosomes showed similar distributions in the 200 bp region.

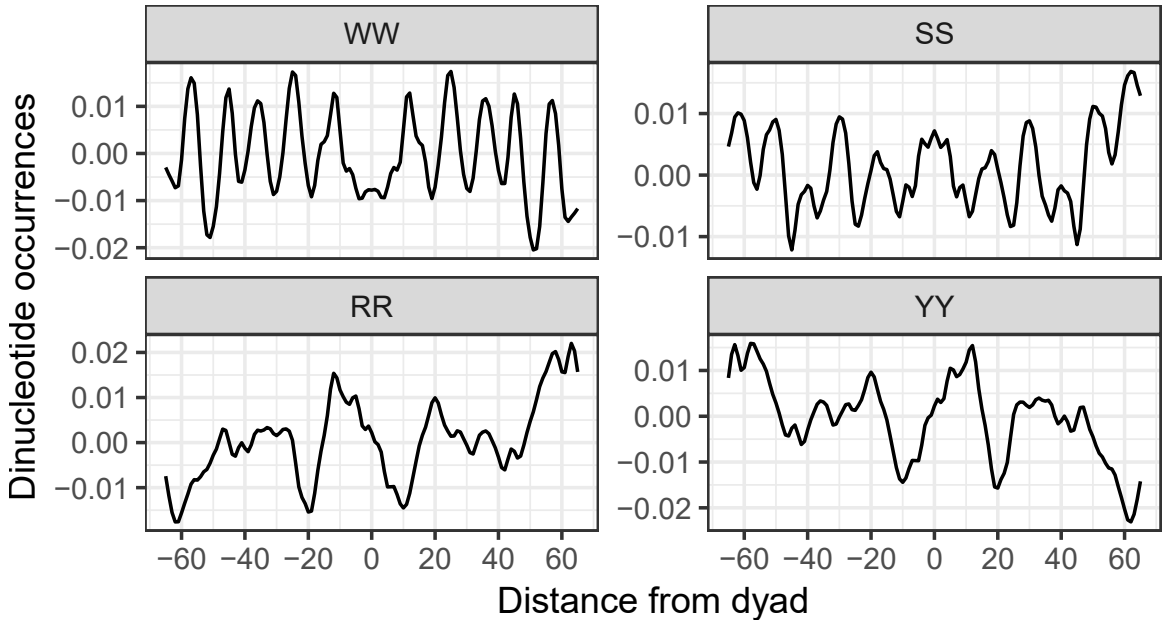

WW
SS
RR
YY
Dinucleotide occurrences
0.01
0.00
−0.01
−0.02
0.02
−60 −40 −20 0 20 40 60
Distance from dyad

Figure 5. Composite dinucleotide patterns of the H2A nucleosome sequences before refining. The average dinucleotide occurrences from the nucleosome sequences are presented as the patterns aligned at the nucleosome dyad. The WW dinucleotide pattern has its peaks at ±15, 25, 35, and 45 bp from the dyad (major groove facing outward). The SS dinucleotide has peaks ±20, 30, 40, 50, and 60 bp from the dyad (major groove facing toward histone). The interval between the peaks is larger around the nucleosome dyad. The interval of the RR and YY dinucleotide peaks are greater than 10 bp and less periodic than the WW and SS patterns.

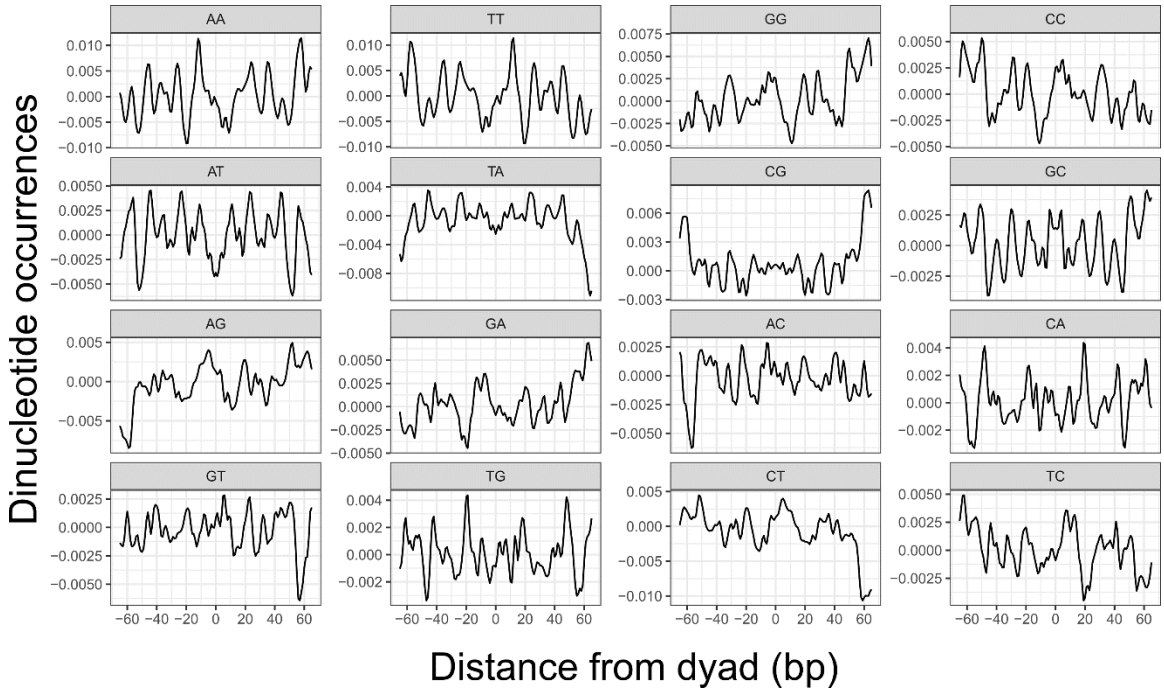

AA
TT
GG
CC
AT
TA
CG
GC
AG
GA
AC
CA
GT
TG
CT
TC
Dinucleotide occurrences
Distance from dyad (bp)

Figure 6. Dinucleotide patterns of the H2A nucleosome sequences before refining. The rest of the dinucleotide patterns from the H2A +1 nucleosome sequences are shown. The average dinucleotide occurrences from the nucleosome sequences are presented as the patterns aligned at the nucleosome dyad. The AT pattern, which is a part of the WW pattern, and the GC pattern, a part of the SS pattern, show periodic peaks with 10 bp interval. Even though the WW and the SS patterns showed periodic peaks with 10 bp interval (**Figure 5**), not all dinucleotide patterns belonging to the composite patterns show the 10 bp periodic patterns. The peak positions of the AA, TT, AT, and TA patterns are at the same position from the dyad (±25, 35, 45 bp) as the WW pattern peaks. The GC dinucleotide peaks remain in the same positions (±20, 30, 40, 50 bp) as the SS pattern even though some peaks are small.

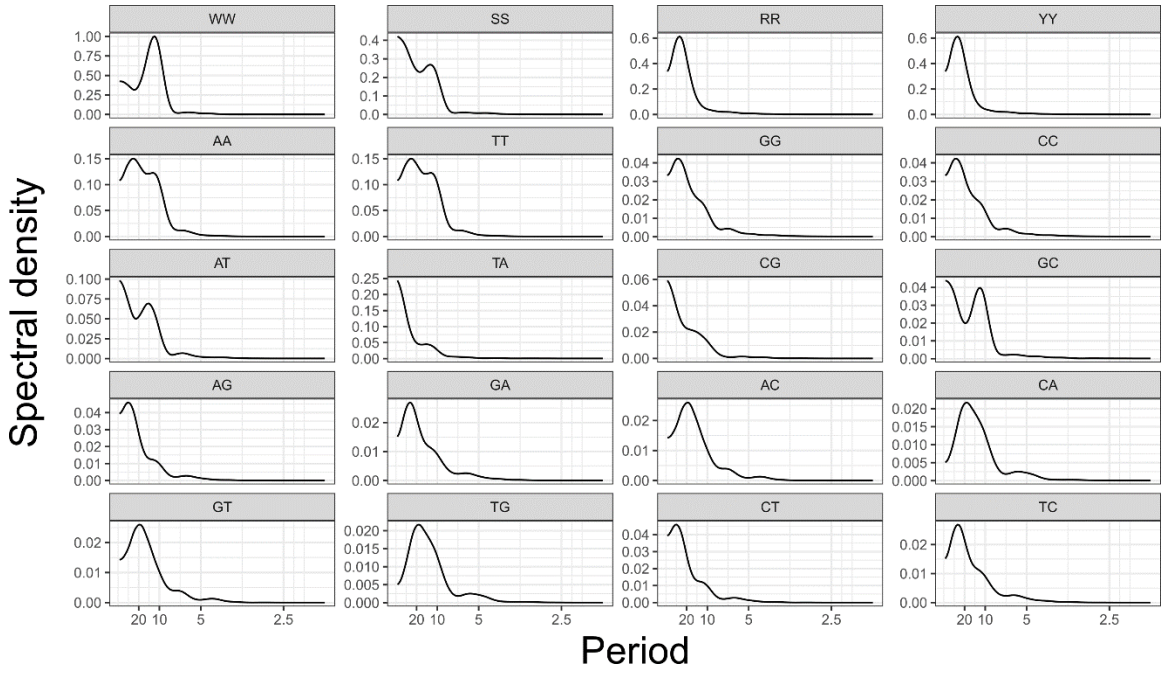

WW
SS
RR
YY
AA
TT
GG
CC
AT
TA
CG
GC
AG
GA
AC
CA
GT
TG
CT
TC
Spectral density
Period
20 10 5 2.5

Figure 7. Periodicities of the H2A dinucleotide patterns before refining. Fourier transform was used to detect the periodicity from the H2A dinucleotide patterns. The dominant period of the WW pattern is 10 bp. The SS pattern has the 10 bp period. Unlike the WW pattern, the SS pattern showed a longer period than 10 bp. The RR and the YY patterns have 20 bp periods. The AT, TT, and AT patterns, which are part of the WW pattern, show 10 bp periods, but the TA pattern does not. Out of the dinucleotide patterns comprising the SS pattern, the GC pattern has the 10 bp period. Some dinucleotide patterns have more than one period.

pattern has the 10 bp period. Only the GC pattern showed the 10 bp period, which may be the strong contributor to the 10 bp period of the SS pattern. The RR and YY patterns and the related AG, GA, TC, and TC patterns have longer periods than 10 bp. Fourier transform showed that the periods of the RR and YY patterns were 20 bp, which was expected from the peak width of the patterns. The initial analysis showed that WW and SS patterns have the expected 10 bp periodicity, but they also have other periods in the pattern, and RR and YY patterns have 20 bp periods.

Pattern model from the identified period.

A prediction model was built with the periods to verify the identified periods for the patterns. The model for patterns was constructed like the following equation with the frequency terms in it, and the model was fit to a general linear model. The frequency was calculated from the period identified in the Fourier transform. The equation was fit to each dinucleotide pattern to estimate the parameters:

$$P = a\sin(2\pi\omega x) + \mathrm{b}\cos(2\pi\omega x) + e$$

where ω is the identified frequency calculated by (1 / period), x is the nucleosome sequence position, and P is the dinucleotide occurrence.

Another model was built with by adding harmonic periods. The modified model is this:

$$P = \mathrm{a}\sin(2\pi\omega x) + b\ \cos(2\pi\omega x) + \mathrm{c}\sin(4\pi\omega x) + d\cos(4\pi\omega x) + e$$

By adding the harmonic frequency, which is the frequency multiplied by an integer, which is 2 in this case, the model may fit better. The raw patterns and the predicted patterns from the two models were plotted (**Figure 8**). The predicted patterns fit well in some part of the patterns, but the deviation accumulated going farther from the middle. The deviation was observed in all four patterns, and it was not reduced by adding the

harmonic terms. The deviation is telling that the estimated periods is not the representative periods. Indeed, by examining the patterns more closely, the -20 to 20 bp regions around the nucleosome dyad appear to have different periods. The period or the frequency of the term used in the model, from the Fourier transform is the average value of the overall pattern, which may not reflect the periods precisely. We checked local periods along the pattern. The local period along the pattern was checked with Fourier transform by moving a 50 bp wide window. The resulting periods were presented in the heat map (**Figure 9**). As expected, the period changes along the dinucleotide pattern generating various local periods. The WW pattern has 10 bp periods in the outer regions of the pattern, but the middle has a longer period of 20 bp. The SS pattern showed a distribution of local periods like the WW pattern: 10 bp periods in the outer region and longer periods in the middle. The RR and YY patterns showed the period of 20 bp near the middle as well as the outer regions. The varying local periods probably caused the deviation of the predicted pattern from the raw patterns.

The periodic model was fit again using the linear model, yet this time the model incorporated different periods for the middle and the outer regions. The results of the prediction improved significantly **(Figure 10**). The prediction for SS is almost perfect in the outer regions as well as in the middle. The prediction for WW was improved even though some deviation in the middle still existed. The RR and YY patterns, which are not periodic apparently, also were predicted satisfactorily by the periodic model. The prediction confirms that the dinucleotide patterns are periodic even though the initial patterns looked non-periodic apparently. The model also verified the estimated periods. Even though the model with local periods of 10 bp fit well the raw patterns, the Fourier showed the patterns have other periods than 10 bp. The extra periods may have resulted

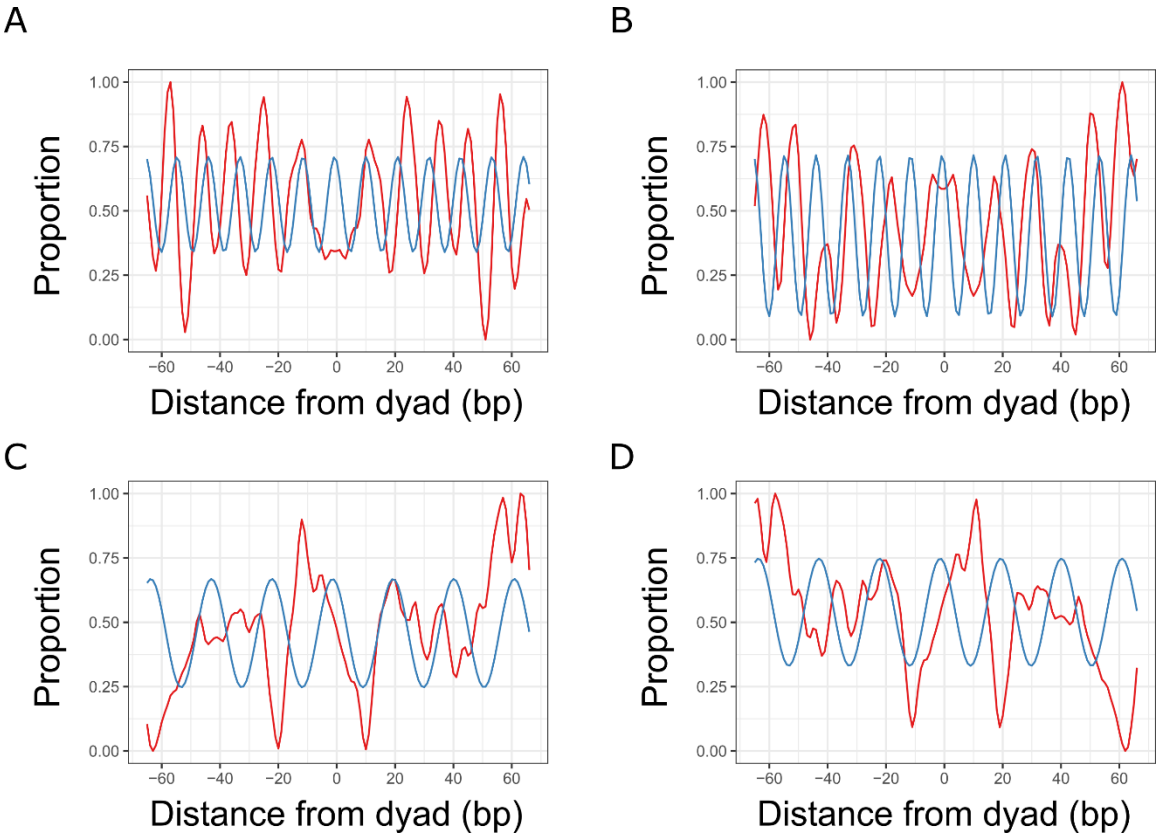

A
B
C
D
Proportion
Distance from dyad (bp)
1.00
0.75
0.50
0.25
0.00
−60
−40
−20
0
20
40
60

Figure 8. Fitting models to the dinucleotide patterns using the initially identified periods. A periodic pattern model was built using the overall periods identified for each dinucleotide pattern. (A – D) The WW (A), SS (B), RR (C), and YY (D) patterns were fitted using models generated with the estimated periods. The raw patterns (red) and the patterns from the model (blue) are shown in the plots aligned at the nucleosome dyad. The errors accumulate as the patterns moving away from the dyad. Visually the middle range of the patterns has different periods, which causes the deviation along the patterns.

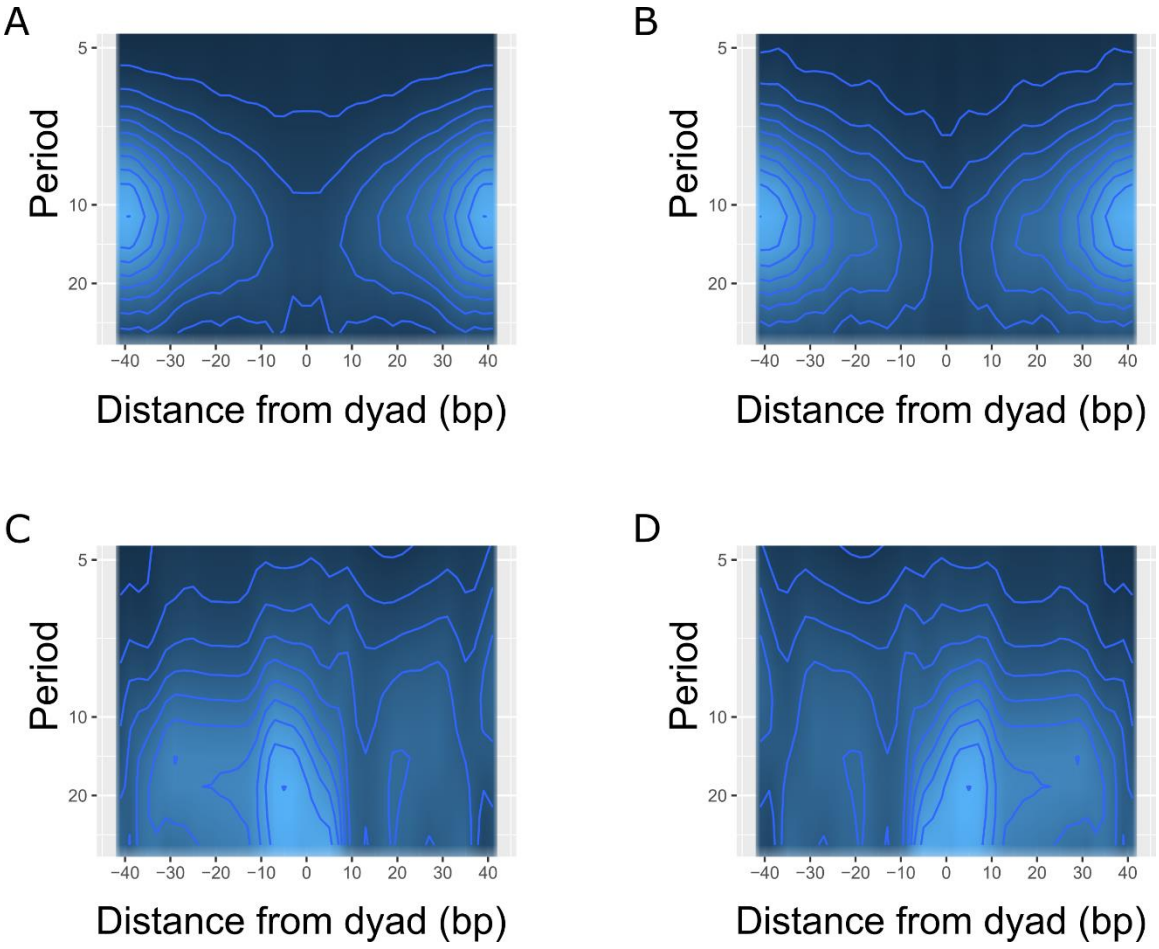

A
B
C
D
Period
5
10
20
−40 −30 −20 −10 0 10 20 30 40
Distance from dyad (bp)

Figure 9. Local periodicities of the refined H2A patterns. The Fourier transform detected the overall periods. The dinucleotide patterns appeared to have different local periods from the overall periods. Fourier transform was run on the 50 bp long segments of the patterns along the sequence. The result is the local periodicity moving along the dinucleotide pattern. The patterns were aligned at the nucleosome dyad at the centre. The middle of the x-axis is the nucleosome dyad, and the y-axis is the identified periods. The colour intensity shows the spectral density of the period at the given pattern position. (A – D) The spectral density of periods moving along the dinucleotide patterns, WW (A), SS (B), RR (C), and YY (D), respectively. The WW and SS dinucleotide patterns show a 10 bp period at the outer sides of the patterns. The 10 bp period changed near the nucleosome dyad: the period increases up to 20 bp. The periodicity of the RR and the YY dinucleotide patterns are stronger at the dyad. The dominant periods of 20 bp periods for the patterns, which are the same as the overall period, are seen in the middle of the patterns.

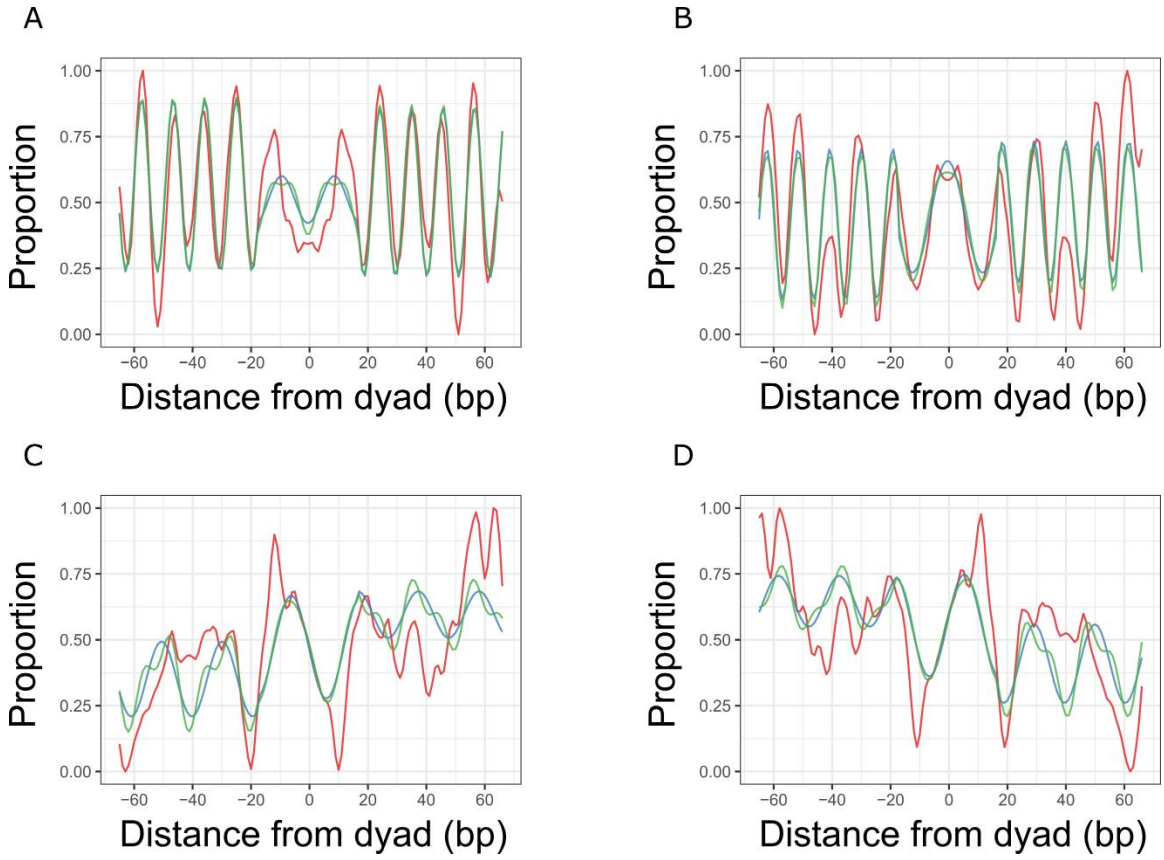

A
B
C
D
Proportion
Distance from dyad (bp)
1.00
0.75
0.50
0.25
0.00
−60
−40
−20
0
20
40
60

Figure 10. Fitting models with the estimated local periods. The periodic model was modified with the local periods as well as the overall periods. The patterns were aligned at the nucleosome dyad. The raw dinucleotide patterns (red) were shown with the predicted patterns generated from the models with the primary period's terms (blue), and by another model with additional harmonic period terms (green). (A - D) The WW (A), SS (B), RR (C) and YY (D) dinucleotide patterns are shown together with the fitted patterns, respectively. The model fitting improved in the four dinucleotide patterns compared to the fitting with the overall periods. Adding the harmonic period term improved the fitting models marginally. The models with the primary periods followed the major peaks and ignored some minor peaks. Adding the harmonic period terms made the model follow the minor peaks also.

from the mixture of the nucleosome sequences by non-phased nucleosome or nucleosomes positioned with anti-NPS sequences. The patterns improved following the refinement of the nucleosome sequences by correlation.

Identification of the NPS patterns by refining the patterns

Even though the model of the periodic patterns performed well in the prediction of the dinucleotide patterns, there were still rooms to improve. Also, the dinucleotide patterns contained multiple periods other than 10 bp alone. To improve the model and to check that the multiple periods are the nature of the patterns or from the noisy input data, the NPS patterns were refined using carefully selected nucleosome sequences. The nucleosome sequences may be the mixture of the NPS patterns and the anti-NPS patterns. We separated the nucleosomes sequences to the positively correlated sequences and the negatively correlated sequences by the correlation to the initially identified NPS patterns as described in the Methods. The dinucleotide patterns were explored from the positively and the negatively correlated sequences separately.

The four compound dinucleotide patterns were plotted along the sequences (**Figure 11**). The WW and SS patterns show sharp periodic patterns. The apparent period is almost 10 bp with the WW peaks at the ±15, 25, 35, 45 bp positions and the SS peaks at the ±20, 30, 40, 50 bp positions. The RR and YY patterns show more defined peaks compared with the initial patterns. The positively correlated patterns (**Figure 11A**) and the negatively correlated patterns (**Figure 11B**) are in the inverted phase. The interval of peaks of the negatively correlated patterns is the same 10 bp, but the peak positions are different. For example, the negatively correlated WW pattern has the peaks at ±20, 30, 40, 50 bp positions. The 5 bp shifted patterns put the WW dinucleotides where the major groove faces inside toward the histone. The changes of the orientation affect the

bendability of the DNA and the stability of the nucleosome on the anti-NPS. The negatively correlated RR and YY patterns show the same inverted phase to the positively correlated RR and YY patterns. The rest of the dinucleotide patterns were shown in **Figure 12**. The AA, TT, AT, and TA patterns, and the CC, GG, CG, and GC patterns, which comprise the WW and SS patterns, respectively, show clear periodic patterns. The positively correlated and the negatively correlated patterns also are inverse phased each other with similar periodicities.

The periodicity of the refined patterns was examined after decomposing with Fourier transform. The positively correlated and the negatively correlated are shown in **Figure 13**. The WW and SS patterns together with their comprising dinucleotide patterns show the well-defined 10 bp period. The longer extra periods, which were shown in the initial patterns disappeared. The negatively correlated WW and SS patterns also show the same period of 10 bp. The anti-NPS pattern did not affect the periodicity but the phase, or the peak position of the pattern, was opposite. The negatively correlated patterns did not have the longer extra period shown in the initial patterns, either. That the longer period did not remain in either positively or negatively correlated patterns but disappeared completely suggests that the extra period did not originate from random noise but resulted from the mixture of the sequences of NPS and anti-NPS patterns.

Interestingly, the periods of the RR, YY and the corresponding subpatterns were shifted from the 20 bp of the initial patterns to the 10 bp period. Both positively and negatively correlated patterns show the 10 bp period. Again, the longer periods resulted from the sequences of NPS and anti-NPS patterns. The separation of the sequences improved the period identification.

The separated patterns were fit with the pattern models incorporating the newly identified periods. The periodic models were fitted for the positively correlated patterns (**Figure 14**) and the negatively correlated patterns (**Figure 15**). Not only the WW and SS patterns, but the RR and YY patterns also fit well with the model of 10 bp period. Even though the periods of the RR and YY are hard to recognise apparently, the Fourier transform, and the model proves the 10 bp periodicity in the patterns.

The models were built with the dominant period only, or the dominant period and the harmonic period. The fitting of the periodic model was evaluated by cross-correlation with the sequence patterns (**Figure 16**). The correlation coefficients of WW and SS patterns were over 0.85 and higher than that of RR and YY patterns. The repeating peaks of the correlation coefficients are due to the periodicity of the patterns.

The improvement measured by the correlation is summarised in **Table 1**. The relationship between the model and the raw patterns are presented. The prediction for the refined patterns improved compared to the initial patterns. The improvement by adding the harmonic term is minimal in the refined patterns.

NPS patterns of H2A nucleosome sequences

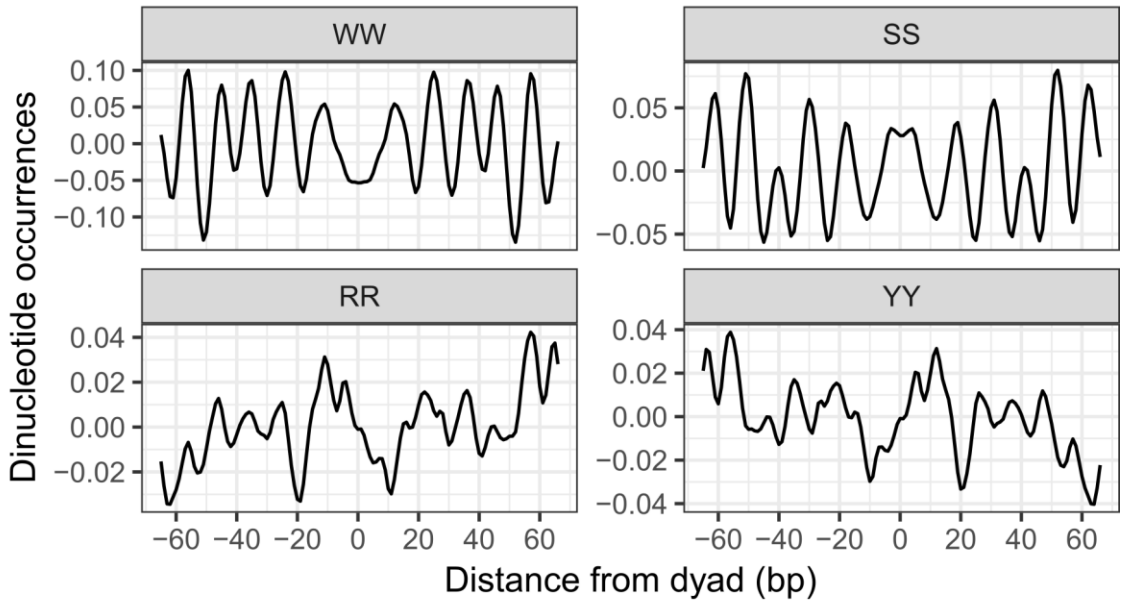
Dinucleotide sequence patterns of the H2A +1 nucleosome were analysed from the sequences of the H2A-only nucleosomes. The WW and SS dinucleotides initially show periodic patterns of 10 bp periods (**Figure 7**). To refine the NPS patterns, we selected the nucleosome sequences based on the correlation to the WW/SS or to the RR/YY NPS pattern as described in the Methods. The refined NPS patterns from the WW/SS positively-correlated nucleosome sequences in Drosophila demonstrate improved periodic patterns of WW/SS and RR/YY dinucleotides like the ones reported in yeast (**Figure 11**).

The WW peaks are located where the major groove faces *away from* the histone core: the SS peaks are located where the major groove faces *toward* the histone core. The periodicities are disrupted between -15 bp to +15 bp region around the dyad, unlike the yeast patterns. The RR and YY dinucleotides have fewer characteristic peaks than the WW and SS patterns. **Figure 12** shows the other dinucleotide patterns. Among them, weak-weak and strong-strong dinucleotide patterns (AT, TA, CG, and GC) show the same 10 bp periodicities as the WW/SS pattern. The RR/YY positively-correlated nucleosome sequences also show enforced RR/YY NPS patterns in the refined data set (**Figure 11B**).

Comparison between Drosophila and Yeast NPS patterns

The WW/SS NPS patterns of Drosophila are similar to the yeast pattern regarding the periodicity (**Figure 17**). The both yeast and Drosophila NPS patterns have a 10 bp period overall, while both patterns have longer than the 10 bp period around the dyad. At the dyad, the major groove of the nucleosomal DNA faces inward. The yeast pattern prefers the WW dinucleotide, while the Drosophila pattern prefers the SS dinucleotide at the position. Because the nucleosome sequence patterns between the yeast and Drosophila are different at the dyad, the amino acid sequences of the histone H3, which interacts with DNA at the dyad forming a nucleosome core, were compared. The multiple sequence alignments of the histone H3 sequences shows the distinct yeast H3 from the Drosophila H3 and other multicellular organisms (**Figure 18**). The differences include the three amino acids of the yeast protein (Q_{120} K_{121} K_{125}), which are located at the dyad. They are known as important in interaction with DNA in a nucleosome core (McBurney et al., 2016). The NPS patterns are correlated with the interacting histone sequences.

A


WW
SS
RR
YY
Dinucleotide occurrences
Distance from dyad (bp)

B

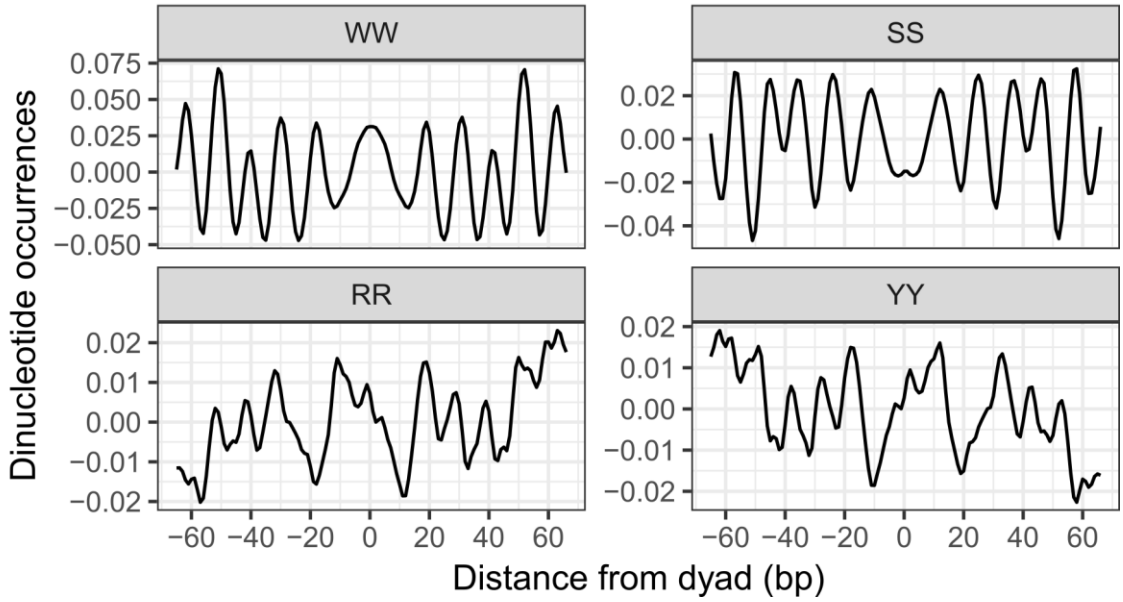
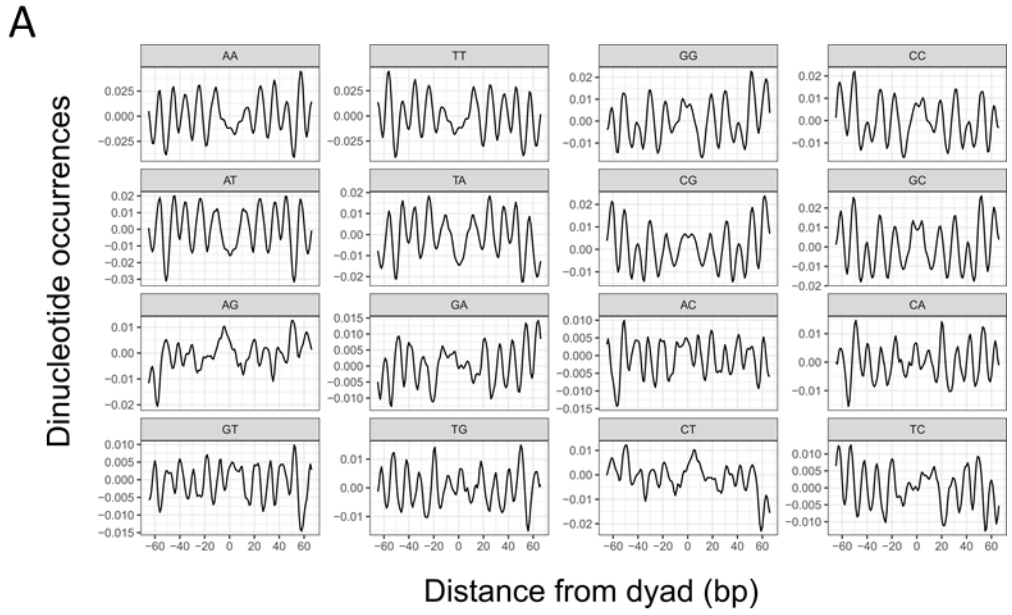

WW
SS
RR
YY
Dinucleotide occurrences
Distance from dyad (bp)

Figure 11. Refined composite dinucleotide patterns of the H2A nucleosome sequences. The dinucleotide occurrences from the selected nucleosome sequences are presented by aligning the nucleosome sequences at the nucleosome dyad. The nucleosome sequences were separated into positively and negatively correlated subsets based on the correlation to the initial WW/SS patterns. (A) Positively correlated patterns. Dinucleotide patterns were obtained from the subset of nucleosome sequences, which were positively correlated with the initial WW/SS patterns. The WW and SS patterns show repeating peaks at 10 bp apart. The peak positions are at ±25, 35, 45, and 55 bp from the dyad for the WW, and at ±20, 30, 40, 50 bp from the dyad for the SS patterns. The periodicity was disrupted in the middle area at the initial sequence set. The RR and YY peaks show roughly 10 bp apart peaks at the regions farther than 20 bp from the dyad. (B) Negatively correlated patterns. The dinucleotide patterns were obtained from the subset of nucleosome sequences, which were negatively correlated with the initial WW/SS patterns. The WW and SS patterns show periodic patterns of 10 bp apart peaks. However, the phases of the patterns are inverse to the positively correlated patterns: the WW peaks at ±20, 30, 40, 50 bp from the dyad and the SS peaks at ±25, 35, 45, and 55 bp from the dyad. The RR and YY patterns are inverted phase to the positively correlated patterns.


A
Dinucleotide occurrences
AA
TT
GG
CC
AT
TA
CG
GC
AG
GA
AC
CA
GT
TG
CT
TC
Distance from dyad (bp)

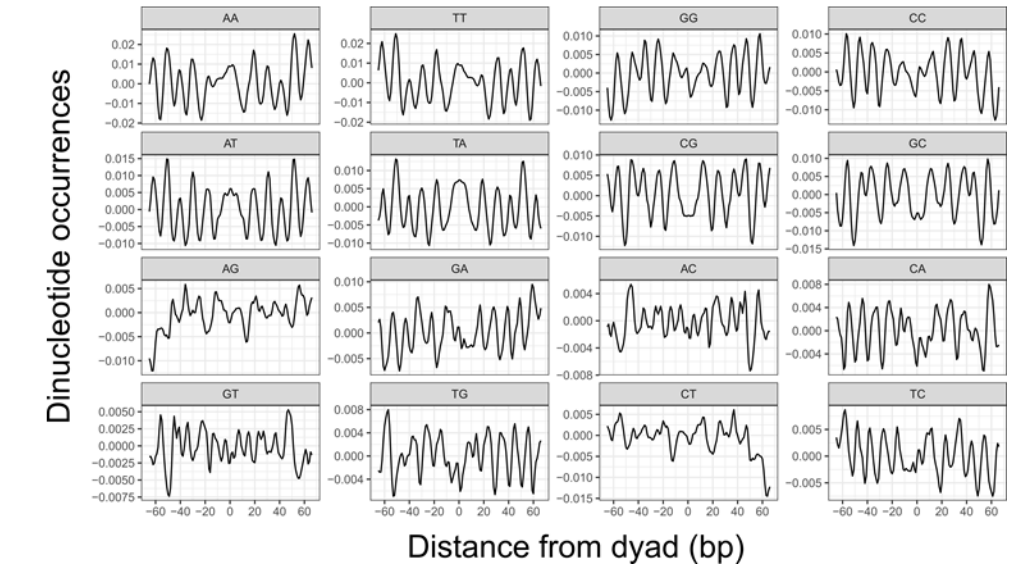
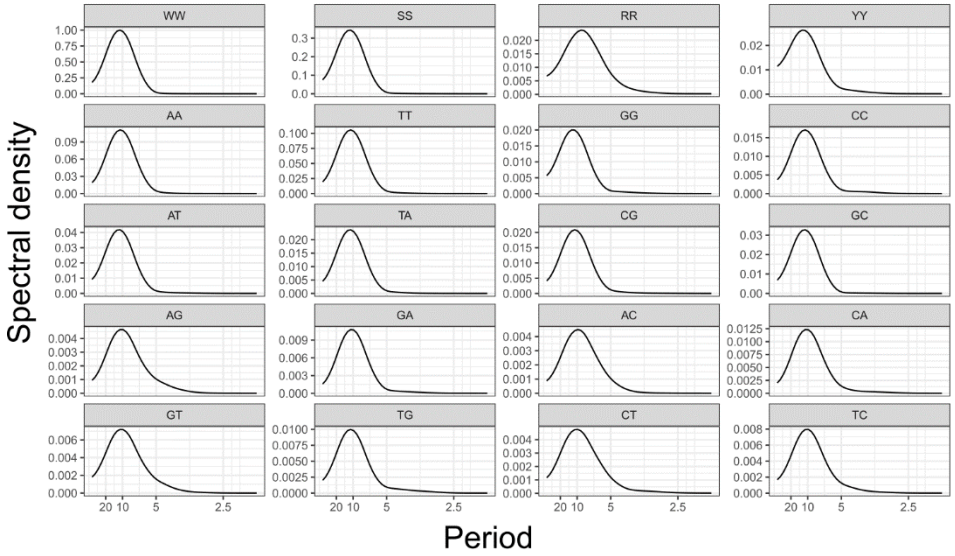

B
Dinucleotide occurrences
AA
TT
GG
CC
AT
TA
CG
GC
AG
GA
AC
CA
GT
TG
CT
TC
Distance from dyad (bp)

Figure 12. Refined dinucleotide patterns of the H2A nucleosome sequences. The dinucleotide patterns are presented. The dinucleotide patterns were obtained from the positively or negatively correlated nucleosome sequences to the initial WW/SS patterns. (A) Positively correlated patterns. Dinucleotide patterns were obtained from the subset of nucleosome sequences, which were positively correlated with the initial WW/SS patterns. The AA, AT, AT, and TA dinucleotides, which belong to the WW, showed less periodic patterns in the initial patterns (**Figure 6**), but the periodicity improved in these selected sequences. They show repeating peaks at 10 bp apart except for the middle region between -20 and +20 bp from the dyad. The peak positions are at ±25, 35, 45, and 55 bp from the dyad like the WW patterns. The CC, GG, CG, and GC dinucleotides, which belong to the SS, also show the same repeating peaks as the SS patterns with the peaks at ±20, 30, 40, 50 bp from the dyad. (B) Negatively correlated patterns. The dinucleotide patterns were obtained from the subset of nucleosome sequences, which were negatively correlated with the initial WW/SS patterns. The dinucleotide patterns are in the inverse phase of the same dinucleotide patterns of the positively correlated patterns.

A


WW
SS
RR
YY
AA
TT
GG
CC
AT
TA
CG
GC
AG
GA
AC
CA
GT
TG
CT
TC
Spectral density
20 10 5 2.5
Period

B

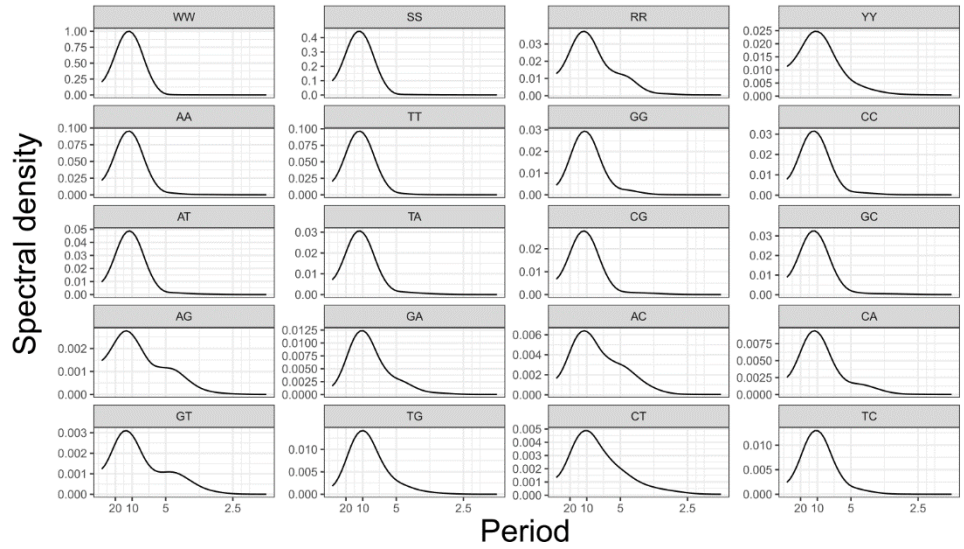

WW
SS
RR
YY
AA
TT
GG
CC
AT
TA
CG
GC
AG
GA
AC
CA
GT
TG
CT
TC
Spectral density
20 10 5 2.5
Period

Figure 13. Periodicities of the refined H2A dinucleotide patterns. Fourier transform was run on the refined dinucleotide patterns from the positively and negatively selected H2A +1 nucleosome sequences. (A) Positively correlated patterns. The 10 bp periodicity is dominant in many dinucleotide patterns: WW, SS, and their sub-patterns (AA, TT, AT, and TA for the WW pattern; GG, CC, GC, and CG for the SS pattern). The RR and YY patterns also show the dominant 10 bp period. (B) Negatively correlated patterns. All the dinucleotide patterns show the 10 bp periodicity. The inverse phase of the patterns does not affect the periodicity.

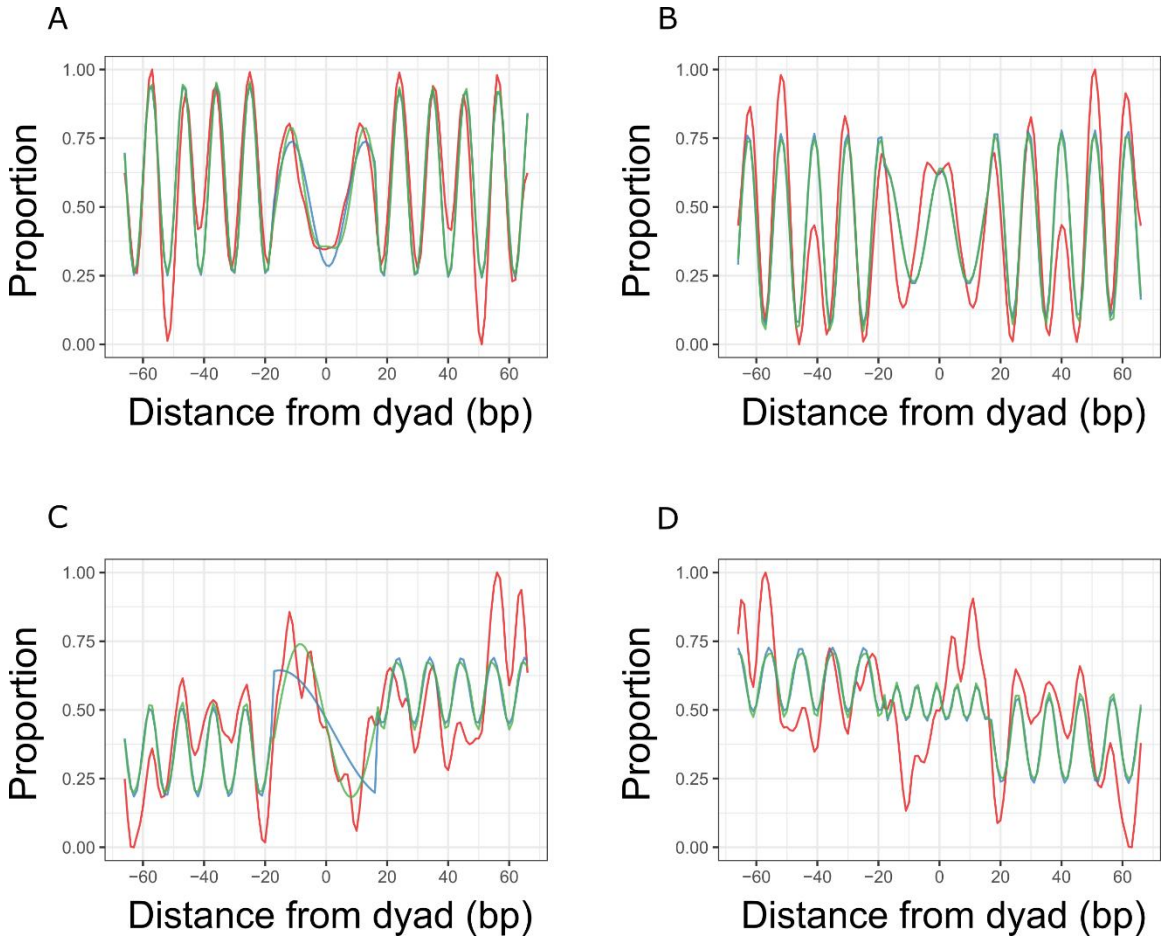

A
B
C
D
Proportion
Distance from dyad (bp)
1.00
0.75
0.50
0.25
0.00
−60
−40
−20
0
20
40
60

Figure 14. Building NPS models for the positively correlated H2A dinucleotide patterns. The local and overall periods were calculated from the refined dinucleotide patterns of the WW/SS positive subset. The parameters of the periodic model were estimated using a general linear model (GLM) with the estimated local and overall periods. The predicted patterns (green, blue) were plotted with the raw patterns (red) aligned at the nucleosome dyad. The fitting of the linear model was done with either a model containing the dominant periods only (blue) or with the model containing the dominant periods and the harmonics of the dominant periods (green). (A) WW pattern. The models with the 10 bp period fit well the raw pattern. Adding the harmonic period improved the model minimally. The fitting between the refined raw pattern and the model in this subset improved compared to the initial model. The prediction with the 10 bp period model proves that the WW pattern has a clear 10 bp periodicity. There are deviations from the model at ±50 bp position from the dyad. (B) SS pattern. The model for the SS pattern also fits well the refined raw pattern. The SS pattern also clearly has a 10 bp periodicity. At the nucleosome dyad, SS dinucleotide is preferred to the WW dinucleotide. The SS peaks at the 50 bp away from the dyad deviated from the model. (C). RR pattern. The model for the RR fits the raw pattern well despite the slight gradient. The 10 bp period of the RR pattern detected by Fourier transform was confirmed by the model built on the 10 bp period. There are deviations from the model at 55 bp position from the dyad. (D) YY pattern. The model for YY pattern shows a good fitting, which confirms the 10 bp period identified by Fourier transform. There are deviations of the pattern at ±10 and -55 bp from the dyad.

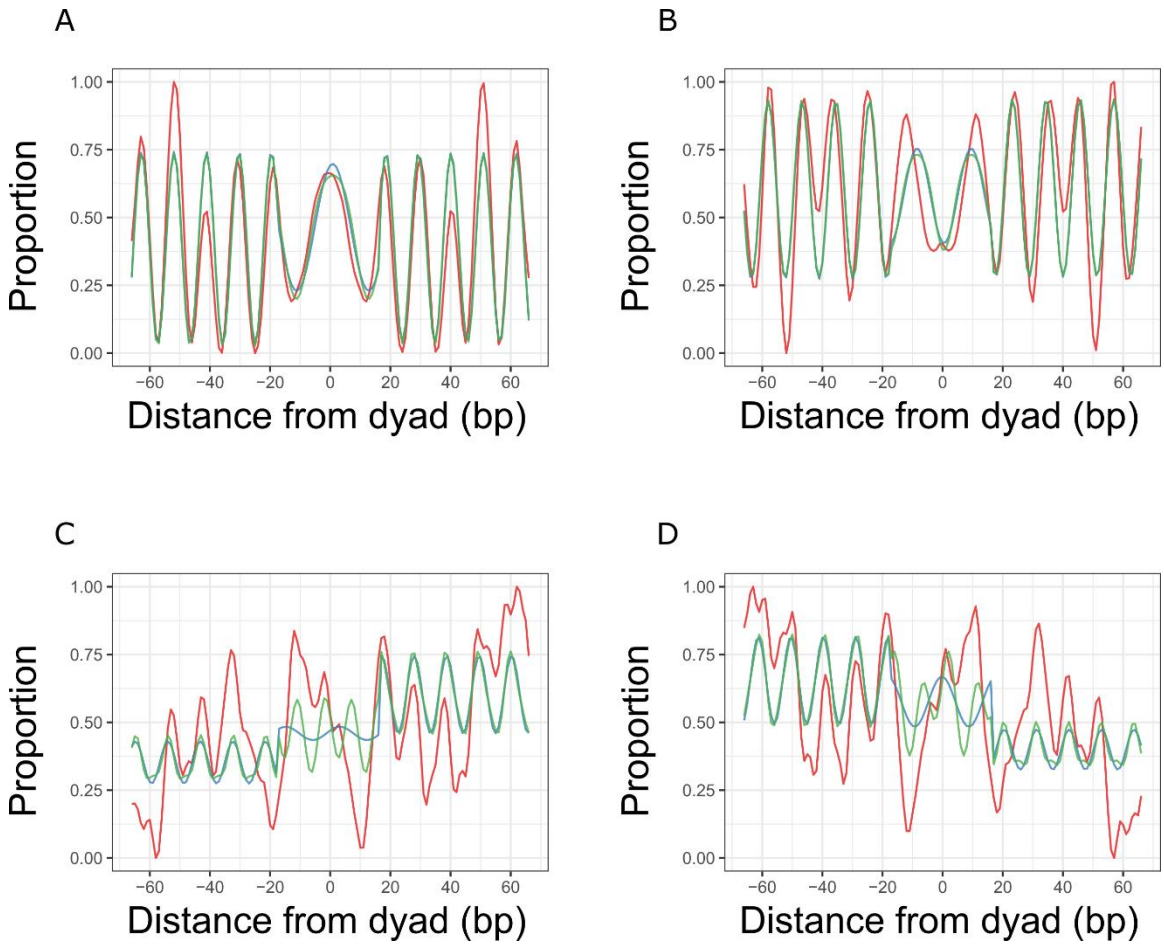

A
B
C
D
Proportion
1.00
0.75
0.50
0.25
0.00
−60
−40
−20
0
20
40
60
Distance from dyad (bp)

Figure 15. Building NPS models for the negatively correlated H2A dinucleotide patterns. The periodic model with the dominant period (blue) and the model with the dominant and the harmonic period (green) were built in the same way as for the **Figure 14**. They were overlaid on the refined raw patterns (red). (A) WW pattern. The model fits well the WW pattern from the subset of negatively correlated nucleosome sequences despite the inverse phase of the patterns. There is deviation at ±50 bp from the dyad. (B) SS pattern. The model only with the dominant period fits well the raw pattern. The peaks at ±50 bp apart from the nucleosome dyad deviated from the predicted pattern as in the WW/SS positively correlated subset. (C) RR pattern. The model with the 10 bp period fits the peak positions of the raw pattern with deviations. (D) YY pattern. The model predicts the peaks of the raw pattern. Adding the harmonic periods improved the model near the dyad area.

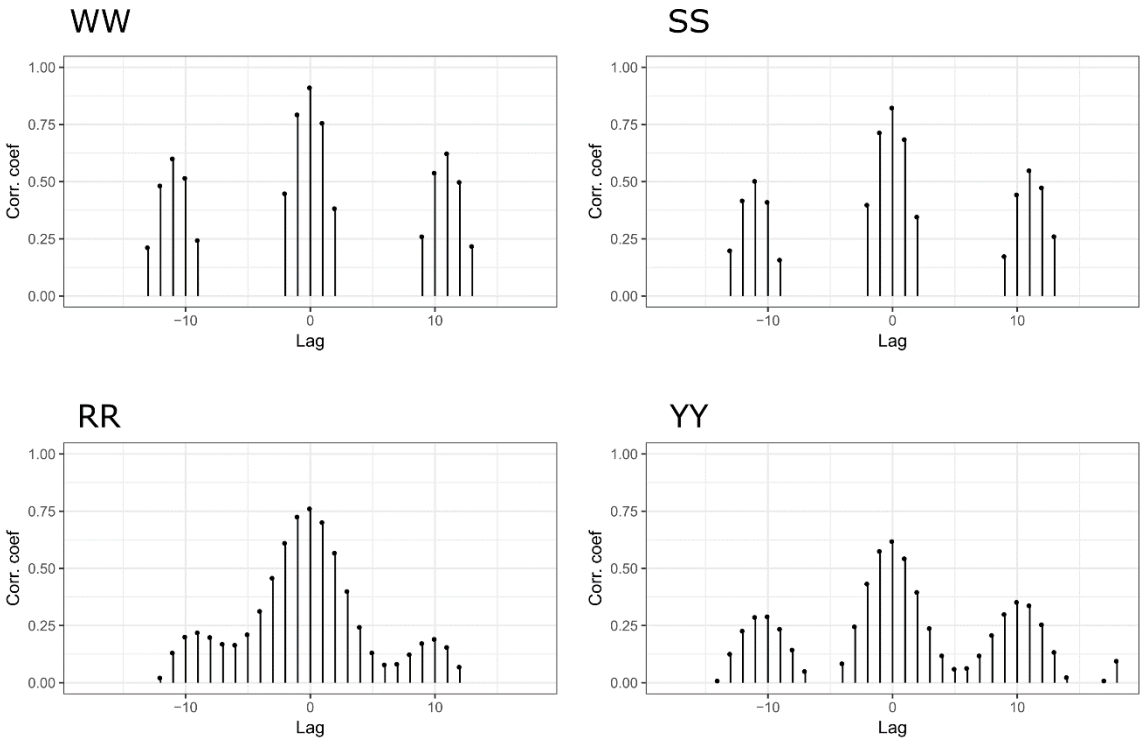
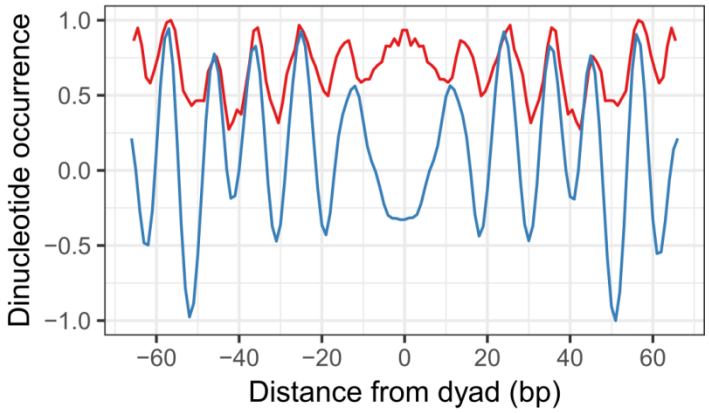

WW
SS
RR
YY
Corr. coef
Lag
1.00
0.75
0.50
0.25
0.00
−10
0
10

Figure 16. Cross-correlation of the predicted patterns and the refined sequence patterns. The correlation coefficients show good fittings of the predicted patterns. The WW and SS pattern predictions are better than the RR and YY patterns. The recurring peaks of the correlation coefficients indicate the periodic patterns of the sequence patterns.

	Initial		Positive		Negative	
	Dominant	Dominant + harmonic	Dominant	Dominant + harmonic	Dominant	Dominant + harmonic
WW	0.87	0.87	0.91	0.92	0.89	0.89
SS	0.76	0.76	0.82	0.82	0.82	0.82
RR	0.66	0.68	0.72	0.76	0.52	0.56
YY	0.67	0.69	0.62	0.62	0.54	0.57

Table 1. Performance summary of the H2A pattern models. The correlation coefficients between the predicted patterns and the sequence patterns are shown. The refined patterns improved in prediction compared to the initial patterns. The improvement by adding the harmonic terms in the refined patterns is minimal.

A


Dinucleotide occurrence
1.0
0.5
0.0
−0.5
−1.0
−60
−40
−20
0
20
40
60
Distance from dyad (bp)

B

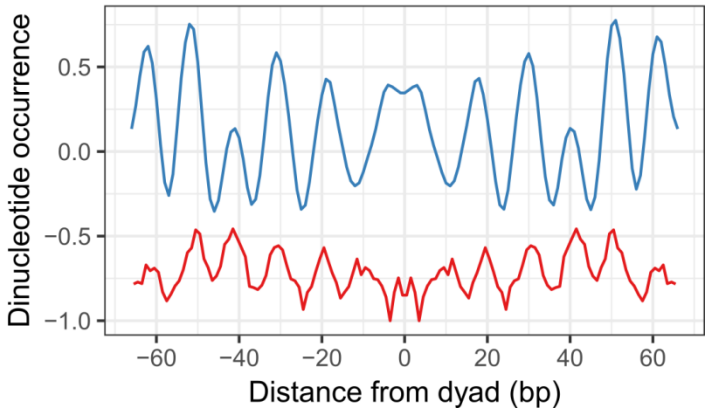

Dinucleotide occurrence
0.5
0.0
−0.5
−1.0
−60
−40
−20
0
20
40
60
Distance from dyad (bp)

Figure 17. NPS pattern comparison between yeast and Drosophila. The WW (A) and SS (B) patterns from the +1 nucleosome sequences of yeast (red) and Drosophila (blue) are shown. The WW and SS patterns from both yeast and Drosophila show a 10 bp periodicity with disruption in the middle between -20 and +20 bp from the dyad. The periodic peak positions of the WW and SS patterns are matching between yeast and Drosophila. However, the preferred dinucleotide at the dyad distinguishes the yeast patterns from the Drosophila patterns. The WW dinucleotide is preferred in yeast while the SS dinucleotide is preferable in Drosophila at the nucleosome dyad.

Histone H3

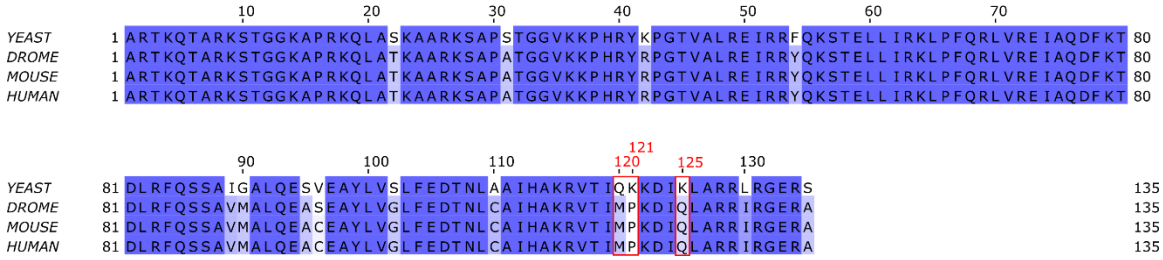

10
20
30
40
50
60
70
YEAST 1 ARTKQTARKSTGGKAPRKQLASKAARKSAPSTGGVKKPHRYKPGTVALREIRRFQKSTELLIRKLPFQRLVREIAQDFKT 80
DROME 1 ARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREIRRYQKSTELLIRKLPFQRLVREIAQDFKT 80
MOUSE 1 ARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREIRRYQKSTELLIRKLPFQRLVREIAQDFKT 80
HUMAN 1 ARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREIRRYQKSTELLIRKLPFQRLVREIAQDFKT 80
90
100
110
121
120
125
130
YEAST 81 DLRFQSSAIGALQESVEAYLVSLFEDTNLAAIHAKRVTIQKKDIKLARRLRGERS 135
DROME 81 DLRFQSSAVMALQEASEAYLVGLFEDTNLCAIHAKRVTIMPKDIQLARRIRGERA 135
MOUSE 81 DLRFQSSAVMALQEACEAYLVGLFEDTNLCAIHAKRVTIMPKDIQLARRIRGERA 135
HUMAN 81 DLRFQSSAVMALQEACEAYLVGLFEDTNLCAIHAKRVTIMPKDIQLARRIRGERA 135

Figure 18. Sequence alignment of histone H3 of eukaryotes. The yeast histone H3 sequence differs from those of other eukaryotes. The multiple sequence alignment of the histone H3 of *S. cerevisiae* (YEAST) and *D. melanogaster* (DROME), and H3.1 of *M. musculus* (MOUSE) and *H. sapiens* (HUMAN) show almost identical sequences except for the yeast H3. The red boxes mark three amino acids (Q_{120} K_{121} K_{125}), which are located at the nucleosome dyad and responsible for the histone binding to DNA in a nucleosome.

SEQUENCE ANALYSIS OF NPS PATTERNS OF H2A.Z NUCLEOSOME SEQUENCES

The H2A nucleosome patterns of Drosophila has been shown to have the 10 bp period. The period is evident in the WW and SS patterns and still valid in the RR and YY patterns even with small variations. The H2A.Z nucleosome has a different role *in vivo*. It is often found in the active genes replacing the H2A nucleosomes. The position of the H2A.Z +1 nucleosome is 20 – 30 bp shifted from the H2A nucleosomes. The changed positions with the histone variant may have resulted from the different NPS. The dinucleotide patterns from the strongly phased H2A.Z +1 nucleosome sequences were analysed in the same way as the H2A nucleosome sequences to investigate the NPS patterns to the H2A.Z nucleosome positioning.

Selection of the phased nucleosome (H2A.Z) for NPS analysis

The nucleosomes were filtered from the initially aligned sequences to select the phased nucleosomes. The sequence alignment was done in both sense and antisense strands of DNA. The nucleosome positions were identified separately in the two strands; then the identified nucleosome positions were compared. If only the positions of an identified nucleosome between the two strands were less than 2 bp apart, the nucleosome was selected for further analysis. The selection of the H2A.Z nucleosomes was started with the filtered nucleosomes and went through the selection process described in the Methods for the +1 nucleosomes and the H2A.Z-only nucleosomes. The +1 nucleosome is the most strongly positioned nucleosome in H2A.Z nucleosomes. The +1 nucleosome sequences were fetched from the Drosophila genome sequence in the same way as in the selecting H2A +1 nucleosome sequences.

Initial dinucleotide patterns

The selected nucleosome sequences were analysed for the dinucleotide patterns. The patterns show less distinctive than the H2A patterns, but still some periodic picks (**Figure 19**). The WW pattern is periodic in the outer regions. The peak positions are the same ±25, 35, 45, and 55 bp from the dyad as the H2A patterns. The strength of each peak varies among the positions. The variations of the strength are small for the SS pattern at the ±20, 30, 40, 50 bp positions with minor peaks in between. The periodicities of the RR and YY are obscure in the initial patterns.

Fourier transform was run on the pattern to identify the dominant and other periods. All the WW, SS, RR, and YY patterns show the 10 bp period as the dominant period (**Figure 20**). The subpatterns of the four patterns also have 10 bp periods as their dominant periods. In some patterns, AC, CA, GT, TG, show extra periods than 10. In general, the H2A.Z patterns also have 10 bp periods even though the periods were less recognisable in the initial raw patterns.

The local periods were searched for the H2A.Z patterns as well by moving the 50 bp window along the nucleosome sequence. The 2D plots show the changes of periodicity along the nucleosome sequence (**Figure 21**). The WW and SS patterns have strong 10 bp periods at the outer regions, while the RR and YY periods are stronger in the middle. As the H2A patterns, the local periods vary along the nucleosome sequence.

Pattern models were built for the H2A.Z patterns in a similar way as in the H2A patterns. The raw dinucleotide patterns, the prediction with the dominant period, and the prediction with the dominant and the harmonic periods are shown in **Figure 22**. The peaks at the outer regions of the WW and SS patterns have the constant interval of 10 bp positioned on the ±25, 35, 45, and 55 bp positions and ±20, 30, 40, 50 bp positions from the dyad, respectively. While the peak strengths predicted from the models are stable, the

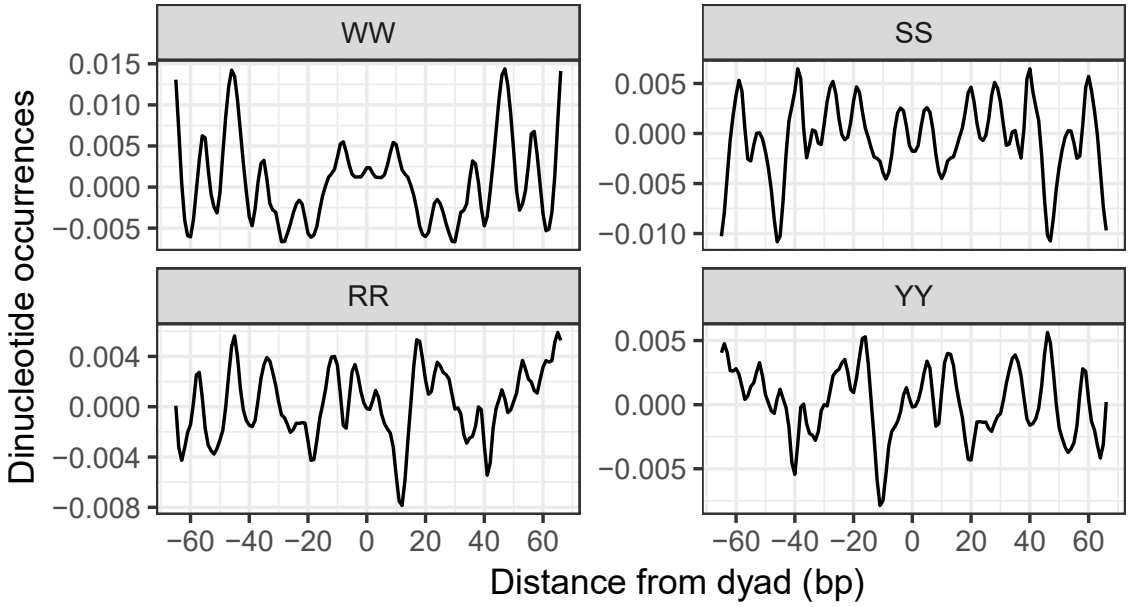

WW
SS
RR
YY
Dinucleotide occurrences
Distance from dyad (bp)
0.015
0.010
0.005
0.000
−0.005
−0.010
0.004
−0.004
−0.008
−60 −40 −20 0 20 40 60

Figure 19. Dinucleotide patterns of the H2A.Z nucleosome sequences. The normalised dinucleotide occurrences from the +1 nucleosome sequences are presented aligned at the nucleosome dyad. The WW patterns have repeating peaks at ±25, 35, 45, and 55 bp from the dyad. The SS patterns have the repeating peaks at ±20, 30, 40, 50, and 60 bp from the dyad. The periodicity of the RR and YY patterns are not as clear as the WW and SS patterns.

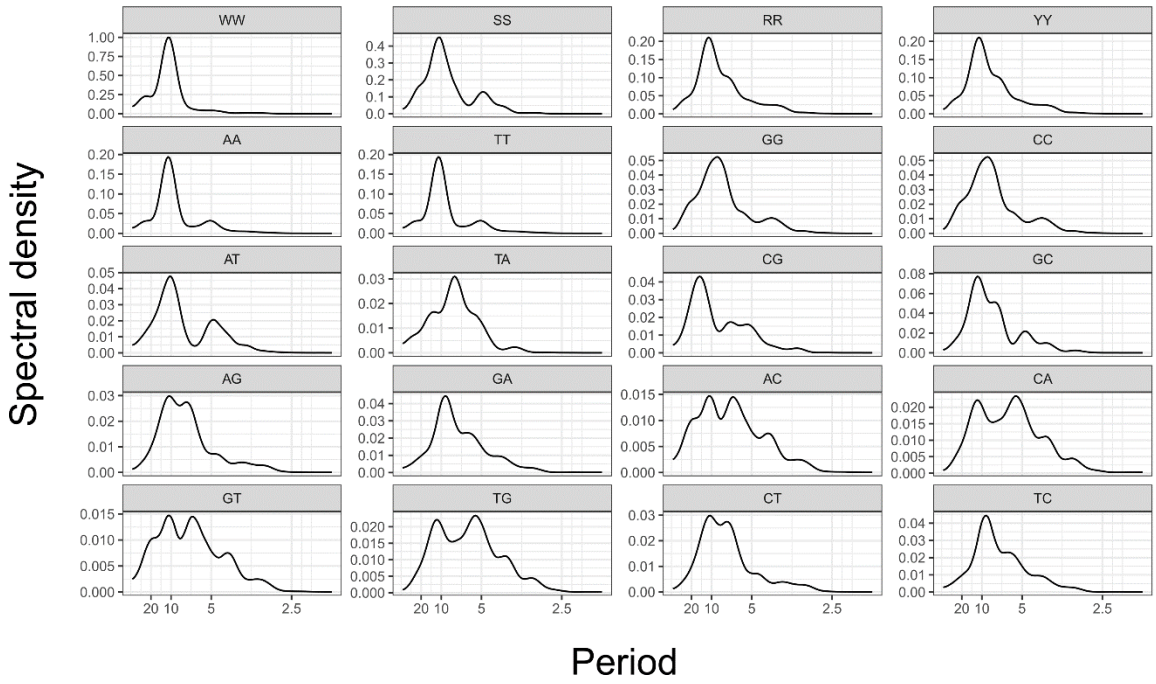

WW
SS
RR
YY
AA
TT
GG
CC
AT
TA
CG
GC
AG
GA
AC
CA
GT
TG
CT
TC
Spectral density
Period
20 10 5 2.5

Figure 20. Periodicities of the H2A.Z dinucleotide patterns before refining. The contribution of periods to the overall periodicity was represented as spectral density. The composite patterns, WW, SS, RR, and YY, have 10 bp periods. Some dinucleotide patterns have additional periods than 10 bp.

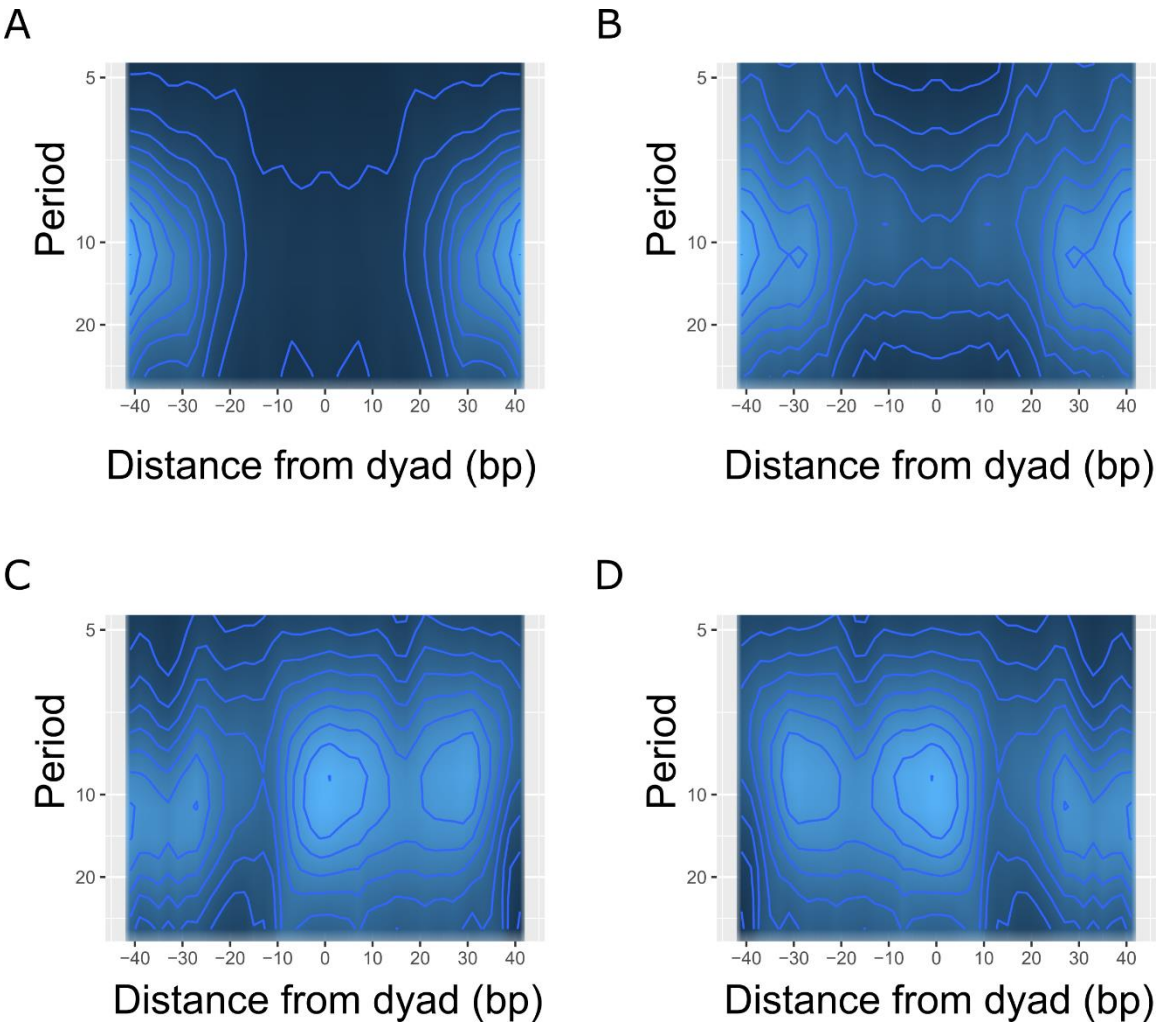

A
B
C
D
Period
5
10
20
−40 −30 −20 −10 0 10 20 30 40
Distance from dyad (bp)

Figure 21. Local periodicities of the H2A.Z dinucleotide patterns before refining. Fourier transform was run on the 50 bp segment moving along the dinucleotide patterns. The patterns (x-axis) were aligned at the nucleosome dyad and the spectral density of the periods (y-axis) was represented by the colour intensity. (A – D) The moving periodicities of the WW, SS, RR, and YY patterns, respectively. The WW and SS patterns show the 10 bp period at the outer regions of the pattern. The 10 bp periods of RR and YY are observed in the middle region.

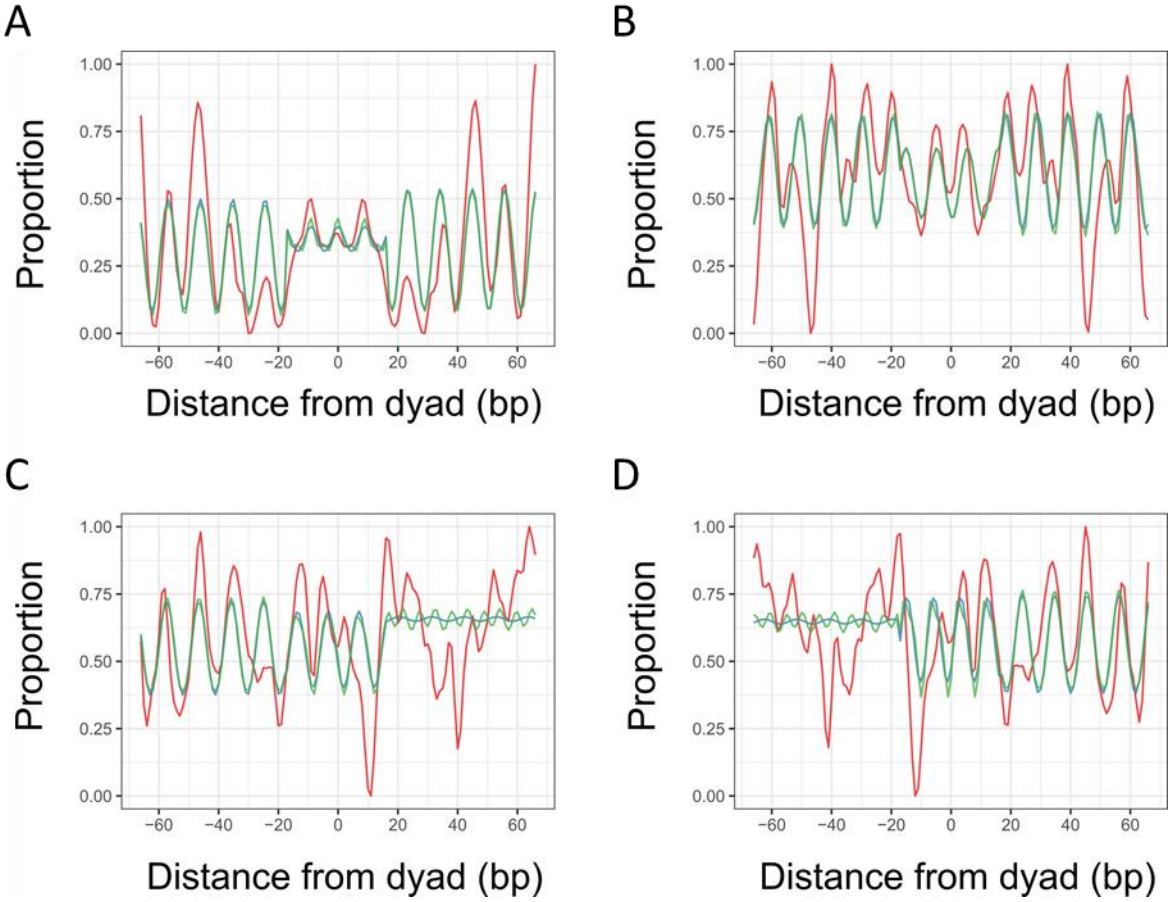

A
B
C
D
Proportion
1.00
0.75
0.50
0.25
0.00
−60
−40
−20
0
20
40
60
Distance from dyad (bp)

Figure 22. Fitting NPS models to H2A.Z dinucleotide patterns using the initially identified periods. The periodic models were built using the overall and the local periods identified by Fourier transform. The model with a term of the dominant periods or additional terms with the harmonic periods were estimated by the general linear model in the same way as the H2A models. The raw pattern (red), the predicted pattern the dominant period (blue), and the predicted pattern with the dominant period and the harmonic period (green) were presented. (A - D) WW, SS, RR, and RR patterns, respectively. The predicted pattern found the peak positions precisely, even though there were some deviations. In the WW and SS patterns, there are deviations at ±45 bp position from the dyad. The RR and YY patterns have periodic regions and less periodic regions. Adding the harmonic periods improved the fitting of the model minimally.

actual peak strengths vary. The WW peaks at ±45 bp are stronger than the predicted peaks. The RR and YY patterns have stronger periodicities in one-half of the pattern than the other half. Adding the harmonic terms improved minimally in all patterns.

The H2A.Z nucleosome sequences were separated according to the correlation to the NPS patterns as described in the Methods. The patterns from the positively and the negatively correlated sequences show well-defined periodic peaks with fewer peaks between the periodic peaks (**Figure 23**). The WW and SS patterns, regardless of the positive or negative correlation, show periodic peaks in inverted phase. The RR and YY also showed periodic peaks better defined than the initial patterns. The peak strength varies big for WW and SS patterns. Most of the H2A.Z dinucleotide patterns show the 10 bp periods identified by Fourier transform (**Figure 24**). Not only the overall periodicity, but the local periods are also 10 bp in WW, SS, RR, and YY dinucleotide patterns (**Figure 25**). The periods of the WW and SS patterns are 10 bp in the outer regions similar to the H2A patterns. However, the RR and YY pattern show different local periods from the H2A patterns. H2A patterns have the 10 bp in the middle (**Figure 9**), H2A.Z patterns have the 10 bp period asymmetrically.

The pattern model was modified according to the identified periods in the positively and the negatively correlated sequences. The modified models fit the WW, SS, RR, and YY patterns nicely regardless of the positive or negative correlation (**Figure 26**). The positively and the negatively correlated patterns are inverse phased each other. It is notable that some of the peaks deviate largely from the predictions. Especially the peaks at the ±45 bp positions show the large deviation in both WW and SS patterns. The peaks may explain the differences between the H2A and the H2A.Z nucleosome's preference toward the nucleotide sequences. The prediction with the modified models was carried out on the RR and YY patterns (**Figure 27**). The modified models improved from the

initial model before refining the patterns. The performance improvements of the models are summarised in **Table 2**. The modified models improved after refining the patterns. Adding the harmonic terms helped the model for the initial models but after refining the improvement is barely noticeable.

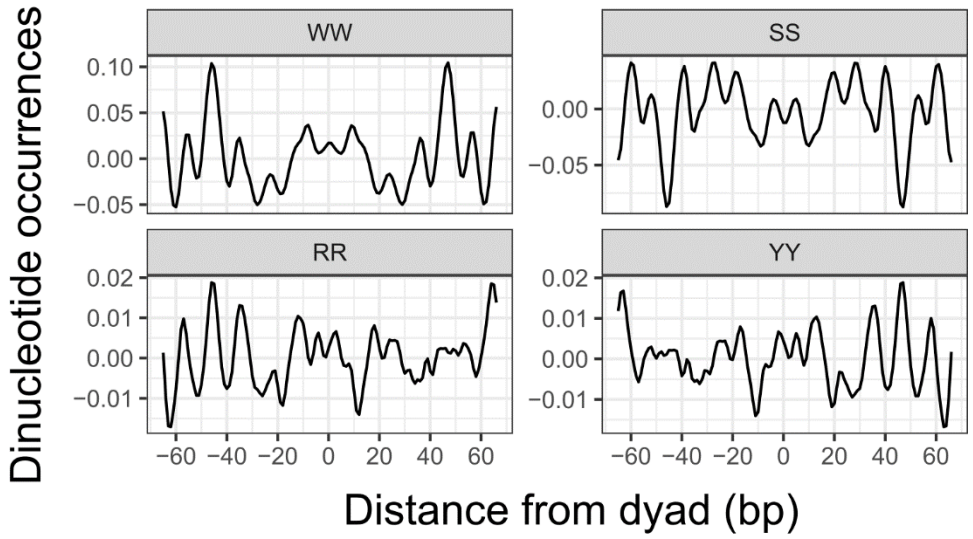
COMPARISON BETWEEN H2A AND H2A.Z NPS PATTERNS

The WW/SS and RR/YY patterns of the H2A were compared to the H2A.Z nucleosome sequences. Both H2A and H2A.Z patterns showed the peaks at the same positions as they have the 10 bp period (**Figure 28**). The WW peak positions are located where the major groove is facing away from the histone (±25, 35, 45, and 55 bp from the dyad), and the SS peak positions are located where the major groove is facing toward the histone core (±20, 30, 40, 50 bp from the dyad).

Even though the patterns are similar in terms of the periodicity, the variation in the peak strength is bigger in the H2A.Z patterns. The deviation of the H2A.Z peaks at the ±45 bp peak is more than three times than that that in the H2A. The deviation at the ±45 bp position is observed in all four dinucleotide patterns of WW, SS, RR, and YY. The RR and YY patterns of H2A.Z also have stronger peaks at the outer regions.

Based on the analysis of the dinucleotide patterns, I proposed the canonical NPS patterns of Drosophila nucleosome sequences (**Figure 29**). In the canonical patterns, the WW and SS dinucleotide peaks appear at the constant interval leading to the 10 bp period. The periodicity is lax in the middle of the patterns. The longer periodicity in the middle and the preference of SS dinucleotide, unlike the WW nucleotide of the yeast patterns, are the characteristics of the Drosophila patterns. The position of the WW and SS dinucleotides are important. The WW dinucleotides lie where the major groove faces outside, and the SS dinucleotides lie where the major groove faces inside.

A


WW
SS
RR
YY
Dinucleotide occurrences
Distance from dyad (bp)

B

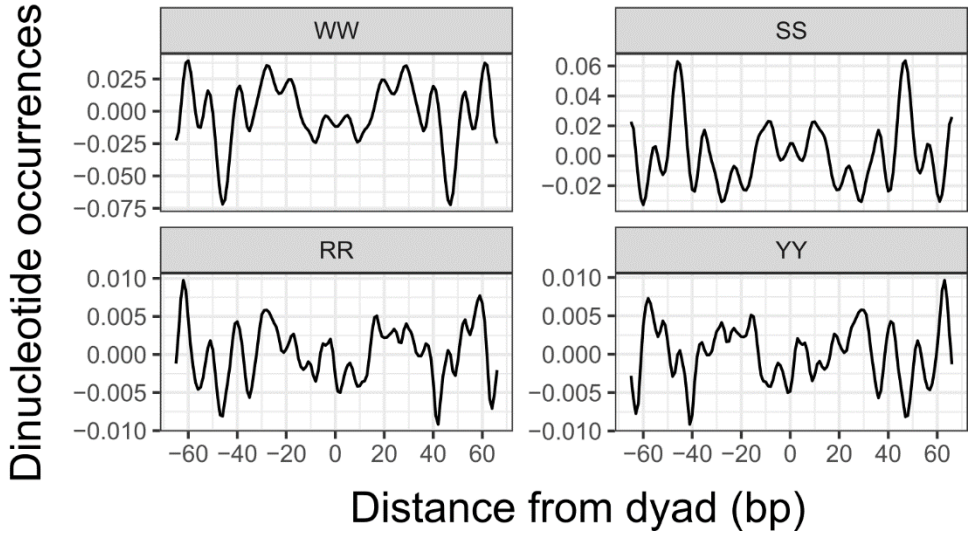
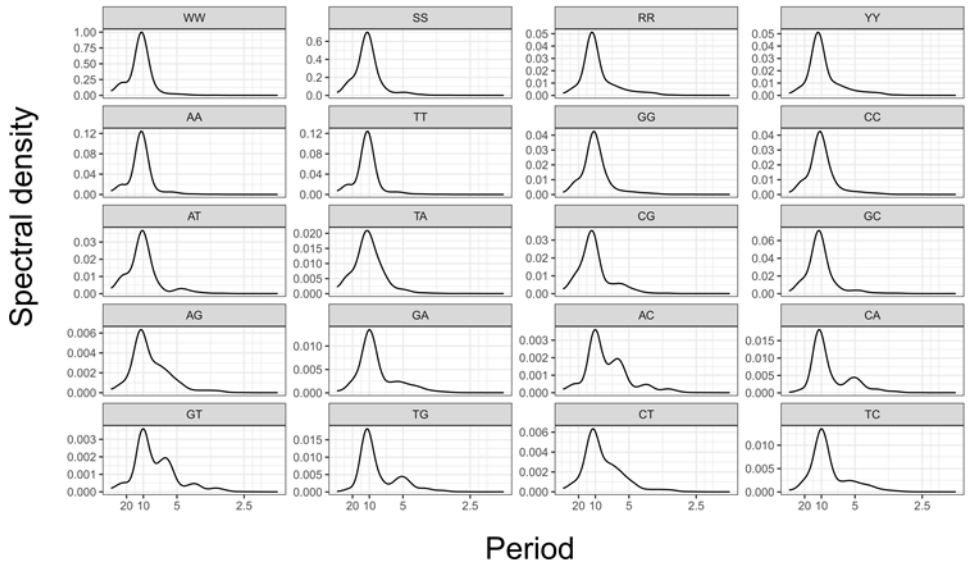

WW
SS
RR
YY
Dinucleotide occurrences
Distance from dyad (bp)

Figure 23. Refined composite dinucleotide patterns of H2A.Z nucleosome sequences. The dinucleotide patterns were obtained from the selected nucleosome sequences based on the correlation to the WW/SS patterns. (A) Positively correlated patterns. The WW and SS patterns have repeating peaks 10 bp apart. The WW peaks are at ±25, 35, and 45 bp from the dyad. The SS peaks are at 30, 40, and 50 bp from the dyad. The peak positions are the same as the corresponding H2A patterns. The WW pattern has major peaks at ±45 bp positions. (B) Negatively correlated patterns. The negatively correlated patterns show the same repeating peaks as the positively correlated patterns despite the inverse phase.

A


WW
SS
RR
YY
AA
TT
GG
CC
AT
TA
CG
GC
AG
GA
AC
CA
GT
TG
CT
TC
Spectral density
20 10 5 2.5
Period

B

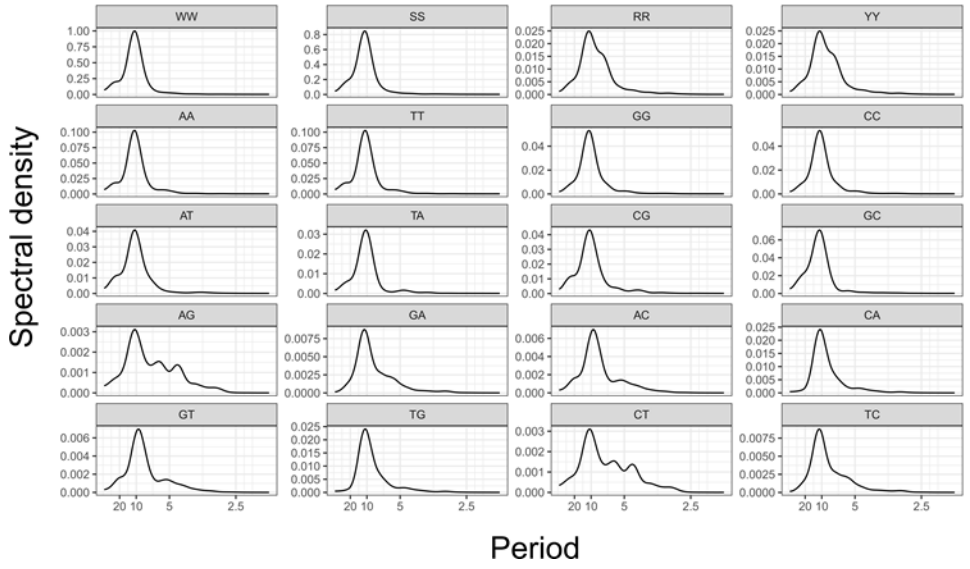

WW
SS
RR
YY
AA
TT
GG
CC
AT
TA
CG
GC
AG
GA
AC
CA
GT
TG
CT
TC
Spectral density
20 10 5 2.5
Period

Figure 24. Periodicities of the refined H2A.Z dinucleotide patterns. Most of the dinucleotide patterns, including the WW, SS, RR, and YY composite patterns, show 10 bp periodicity after refining. Some dinucleotide patterns have additional periods than the 10 bp period but the 10 bp period is still dominant. (A) Positively correlated patterns. (B) Negatively correlated patterns.

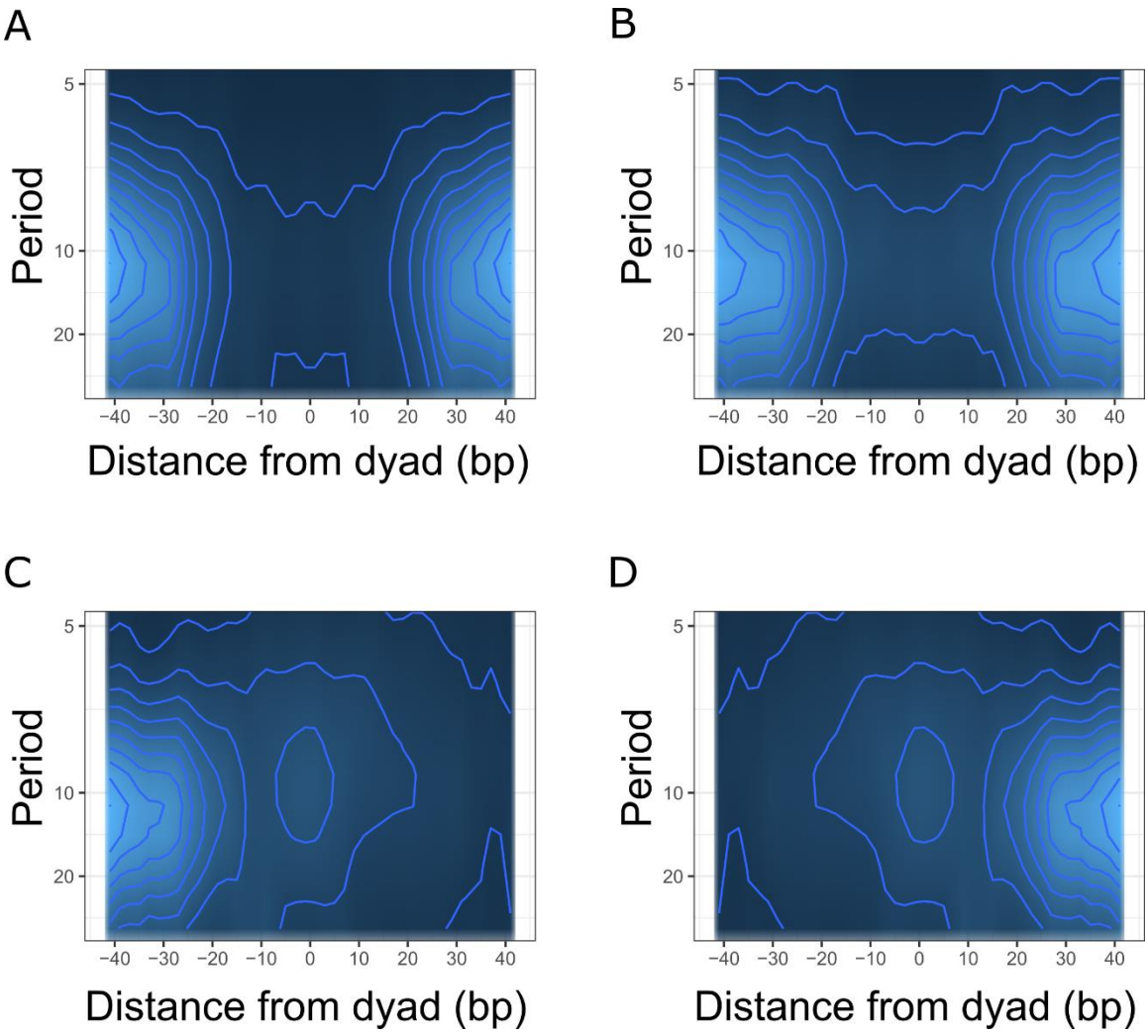

A
B
C
D
Period
5
10
20
−40 −30 −20 −10 0 10 20 30 40
Distance from dyad (bp)

Figure 25. Local periodicities of the refined H2A.Z patterns. The local periods were identified by running Fourier transform with the 50 bp segments of the dinucleotide patterns of the positively correlated sequences. The patterns were aligned at the nucleosome dyad (x-axis), and the spectral density was shown in the colour intensity of the periods (y-axis). (A – D). The periods of WW, SS, RR, and YY patterns, respectively. The WW and the SS patterns show the 10 bp periods at the outer regions of the patterns. The 10 bp periods of the RR and the YY patterns are on either side of the patterns complementing each other.

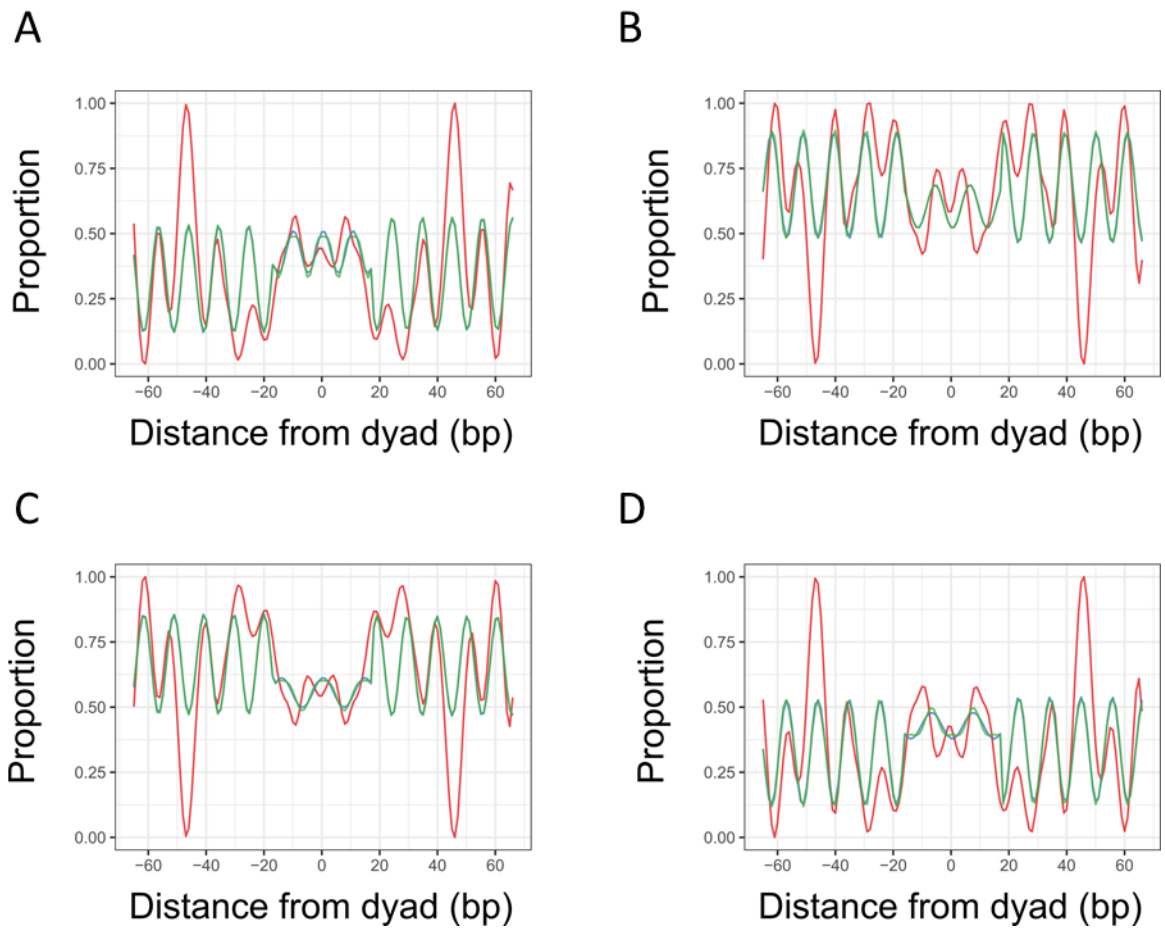

A
B
C
D
Proportion
Distance from dyad (bp)
1.00
0.75
0.50
0.25
0.00
−60
−40
−20
0
20
40
60

Figure 26. Fitting NPS models to the refined H2A.Z WW and SS dinucleotide patterns. The model was built with the periods identified from the WW/SS positively correlated sequences. The predicted patterns using the models with the dominant periods (blue) and the modified models the dominant and the harmonic periods (green) were shown together with the raw patterns (red). (A – B) Positively correlated WW and SS patterns, respectively. The predicted patterns fit the raw patterns. Adding the harmonic periods barely improved the models. There are deviations from the models at ±45 bp positions from the dyad in all four dinucleotide patterns. (C – D). Negatively correlated WW and SS patterns, respectively. The model fits well the negatively correlated patterns. There are deviations from the model at ±45 bp positions from the dyad like in the positively correlated patterns.

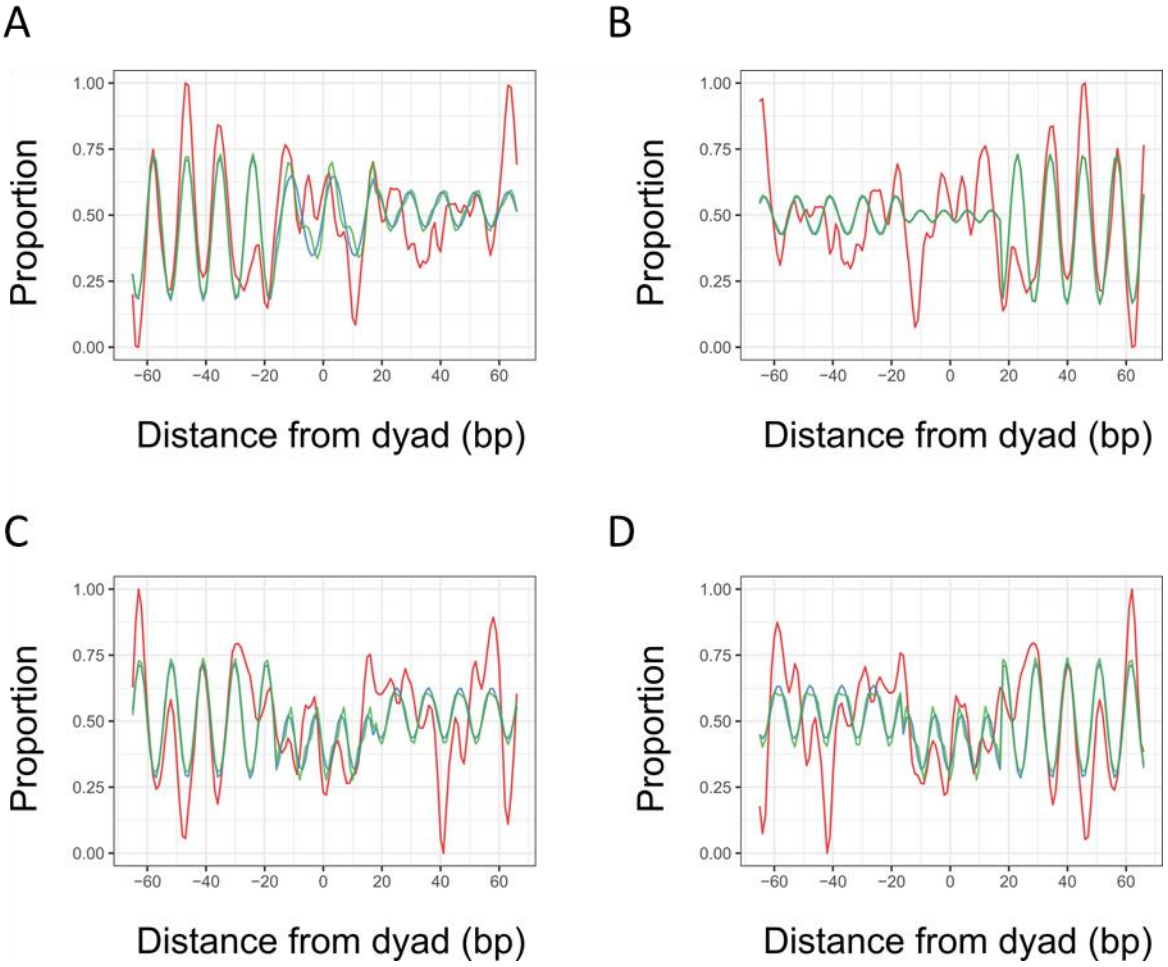

A
B
C
D
Proportion
1.00
0.75
0.50
0.25
0.00
−60
−40
−20
0
20
40
60
Distance from dyad (bp)

Figure 27. Fitting NPS models to the refined H2A.Z RR and YY dinucleotide patterns. The model was built with the periods identified from the WW/SS negatively correlated sequences. The predicted patterns using the models with the dominant periods (blue) and the modified models the dominant and the harmonic periods (green) were shown together with the raw patterns (red). (A – B) Positively correlated RR and YY patterns, respectively. The model fits the periodic part of the patterns. Adding the harmonic periods barely improved the models. (C – D). Negatively correlated RR and YY patterns, respectively. The model fits well the negatively correlated patterns. The less periodic parts of the patterns still have periodic peaks identified by the model.

	Initial		Positive		Negative	
	Dominant	Dominant + harmonic	Dominant	Dominant + harmonic	Dominant	Dominant + harmonic
WW	0.62	0.63	0.57	0.57	0.54	0.54
SS	0.61	0.61	0.54	0.54	0.54	0.54
RR	0.49	0.50	0.68	0.68	0.55	0.56
YY	0.50	0.51	0.63	0.63	0.56	0.56

Table 2. Performance summary of the H2A.Z pattern models. The correlation coefficients between the predicted patterns and the raw patterns are shown. The refined RR and YY patterns improved in prediction compared to the initial patterns. The improvement by adding the harmonic terms in the refined patterns is minimal.

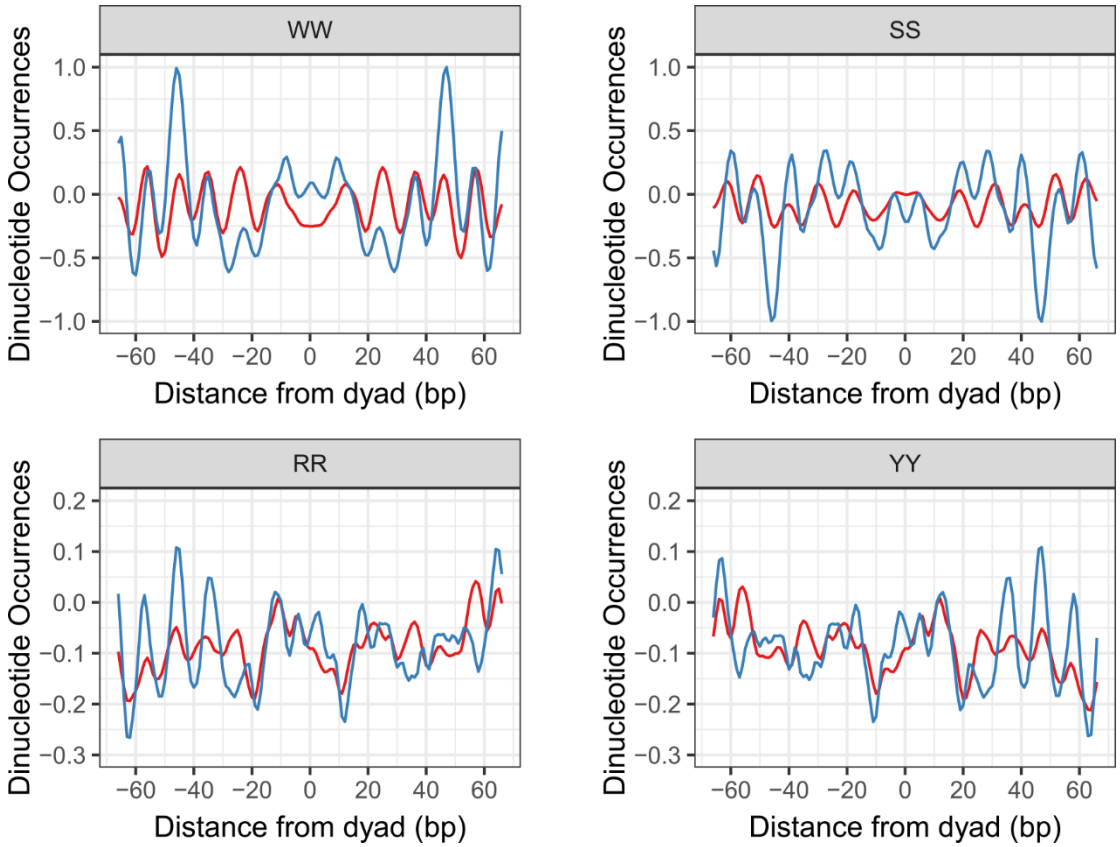

WW
SS
RR
YY
Dinucleotide Occurrences
Distance from dyad (bp)
1.0
0.5
0.0
−0.5
−1.0
0.2
0.1
−0.1
−0.2
−0.3
−60
−40
−20
0
20
40
60

Figure 28. Comparison of the NPS patterns of H2A and H2A.Z. The dinucleotide patterns of H2A (red) and the H2A.Z (blue) nucleosome sequences are shown. Both H2A and H2A.Z have similar WW patterns: highly periodic patterns with 10 bp periods. The WW patterns match the peak positions of ±25, 35, 45, and 55 bp from the dyad. The SS patterns between the H2A and the H2A.Z nucleosomes have the same period of 10 bp with matching peak positions, even with different magnitudes of peaks. The differences between the H2A and the H2A.Z patterns are observed in the middle of the patterns (between -20 and 20 bp from the dyad) and the ±45 bp positions from the dyad. Both WW and SS patterns show the differences at the ±45 bp position from the dyad. The RR and YY patterns of the H2A.Z nucleosome hold stronger peaks than the corresponding H2A patterns. The strongest peak of the RR and YY of the H2A.Z patterns also at the ±45 bp position.

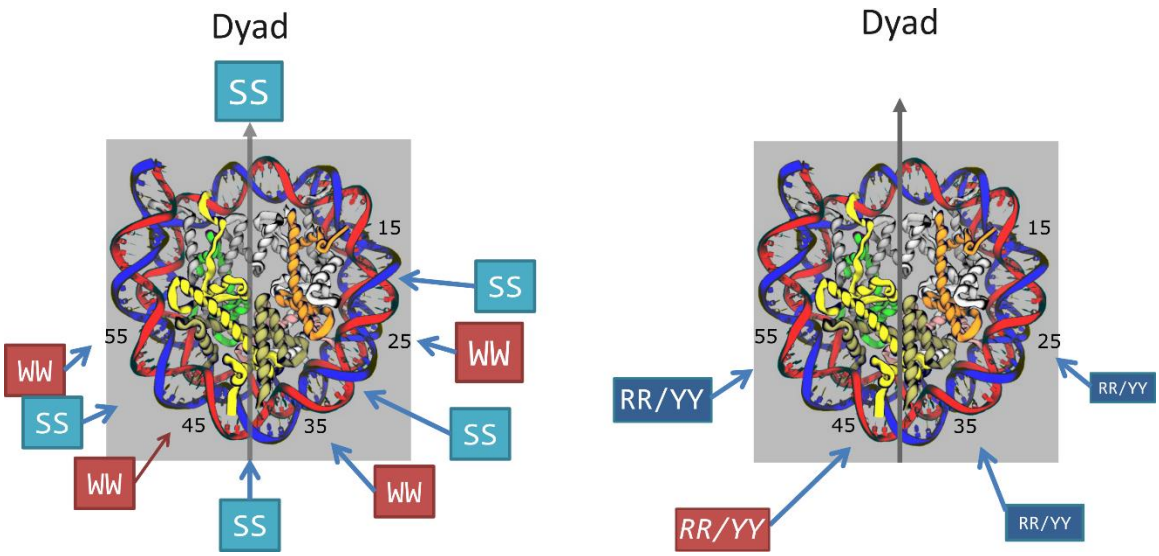

Dyad
SS
WW
SS
WW
SS
WW
SS
WW
SS
15
25
35
45
55
Dyad
RR/YY
RR/YY
RR/YY
RR/YY
15
25
35
45
55

Figure 29. Canonical nucleosome positioning sequences. Two models of nucleosome positioning sequences of Drosophila, WW/SS and RR/YY patterns, are presented. The labels show the preferred dinucleotides at the position, and the size of the labels represents the relative preference of the dinucleotides at the position. (A) WW/SS pattern model shows that the WW dinucleotides have a 10 bp period and the dinucleotide positions are located where the major groove faces away from the histone (±25, 35, 45, 55, 65 bp). Simultaneously, the SS dinucleotide pattern, with the same 10 bp period, has preferred dinucleotides located where the major groove faces toward the histone core (±30, 40, 50, 60 bp). The SS dinucleotides are the preferred dinucleotide at the dyad. (B) The RR/YY NPS pattern model shows that the RR and YY patterns of the 10 bp periods. The preference of the RR/YY dinucleotides is strong near the outer regions of the nucleosomal DNA, especially at ±45 bp position. The RR/YY pattern is weaker and less strict than the WW/SS patterns.

Structural differences of H2A and H2A.Z corresponding the pattern differences

The different peak positions of the H2A and H2A.Z NPS patterns direct the exposure of the major and minor groove toward or outside of the nucleosome core. The differences may be related to the different interactions between DNA and the histones. So, the peak positions were checked on the 3-dimensional structure of the nucleosome.

The 3-dimensional structures of the H2A and H2A.Z nucleosomes from yeast and human were compared. Because Drosophila H2A.Z nucleosome structure was unavailable at the time of the analysis, human H2A.Z nucleosome structure, whose protein sequences are conserved with the Drosophila H2A.Z histone (**Figure 30A**), was used instead. The loop 1 and the loop 2 of the H2A histone have been known as the DNA interacting motifs in a nucleosome (Thakar et al., 2009). The protein structures of them are different between H2A and H2A.Z histones according to the crystallography structure. The interacting DNA regions with the loop 1 and the loop 2 are the ±45 and ±55 bp positions from the dyad, which coincides with the peaks of H2A.Z differentiating the NPS patterns from the H2A patterns (**Figure 30B**).

ENRICHED BIOLOGICAL FUNCTIONS OF H2A-ONLY, H2A.Z-ONLY, AND H2A/H2A.Z BOUND GENES.

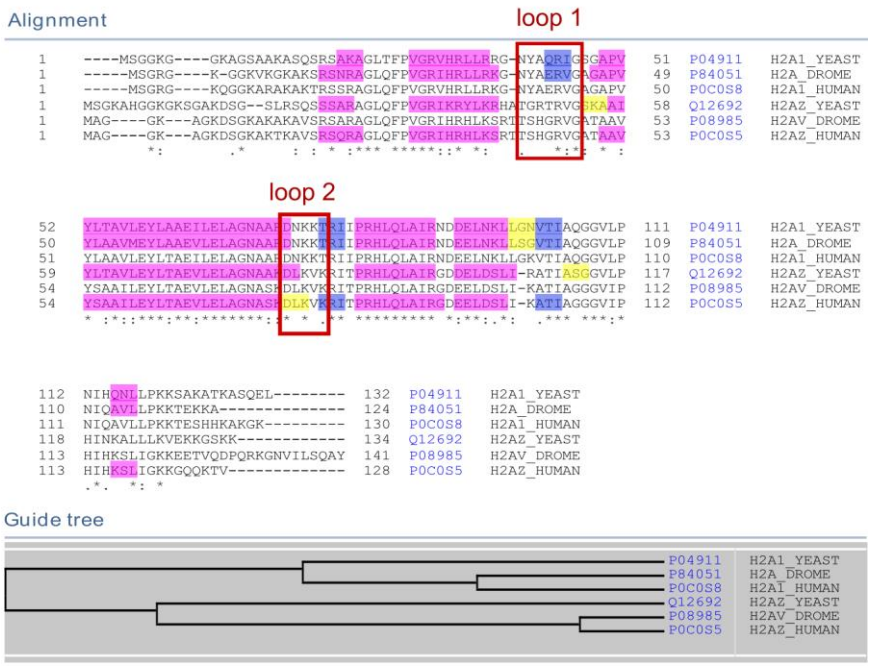
The enrichment tests for the biological functions were carried out with the three gene groups selected by the type of the +1 nucleosome: H2A-only, H2A.Z-only, and H2A/H2A.Z (**Figure 31**). The enriched biological functions are different between the three groups. Only the regulation of transcription is common between the H2A-only and the H2A/H2A.Z groups. The H2A/H2A.Z group contains house-keeping functions including biosynthesis and metabolic pathways functions. On the other hand, the enriched functions in the H2A-only group and the H2A.Z-only group were related to particular gene functions. The functions in the H2A-only genes are related to the cell

differentiation and development. The enriched functions in the H2A.Z-only group are related to the histone expression and nucleosome assembly.

The WW/SS and RR/YY NPS patterns were correlated with the biological functions (**Figure 32A**). The genes having the NPS patterns at their promoters were searched for the enriched functions. The Gene Ontology terms of biological processes were searched among the given genes, and the related terms were clustered for a better review. The enriched functions were selected based on the enrichment score (greater than 2) and Benjamini-Hochberg false discovery rate (less than 0.05). Regulation of transcription is the most enriched function in the genes associated with the nucleosomes having a strong WW/SS NPS pattern. The tightly regulated genes with the strong WW/SS NPS pattern implicates that the nucleosome may work together with transcription factors in the transcriptional regulation. On the other hand, genes with the negatively correlated WW/SS, the anti-NPS patterns, did not show enriched functions (**Figure 32B**). Some functions showed increased enrichment scores, but the FDR was not significant enough. The failure to identify the enriched functions suggests the wide-spread anti-NPS pattern over various genes.

The genes with the RR/YY NPS patterns show that the sensory perception, signal transmission or neuronal activities related functions are enriched (**Figure 33**). The localisation of the genes in the enriched function, sensory perception, revealed that most of the proteins expressed from the genes are secreted outside of the cell or membrane-bound proteins. The interesting genes in the group include Cadherins, Ecdysone-induces genes, His deacetylase, and NF-AT. The differences of enriched gene functions between the genes with different NPS patterns imply that the +1 nucleosome sequence and the positioning of the +1 nucleosome are related to the other promoter elements contribute to various gene regulation.

A


Alignment
loop 1
loop 2
Guide tree
P04911 H2A1_YEAST
P84051 H2A_DROME
P0C0S8 H2A1_HUMAN
Q12692 H2AZ_YEAST
P08985 H2AV_DROME
P0C0S5 H2AZ_HUMAN

B

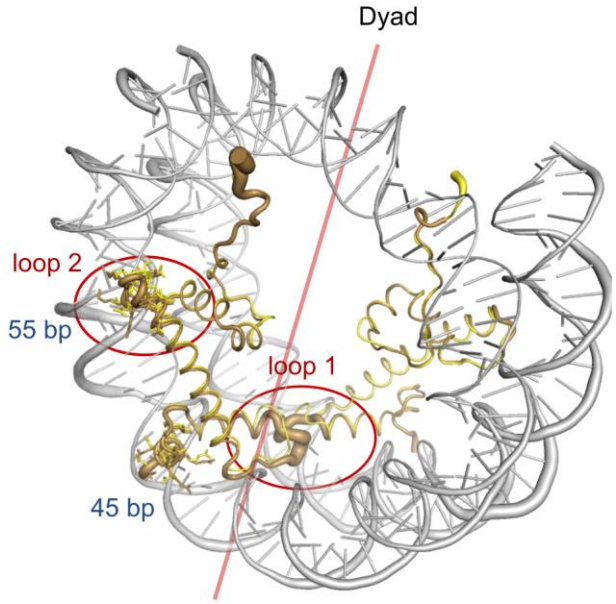

Dyad
loop 2
55 bp
loop 1
45 bp

Figure 30. Structural differences of the H2A and H2A.Z histones. (A) Protein sequences of H2A and H2A.Z histones from yeast, Drosophila, and human were aligned using ClustalW2. The red rectangles mark the interacting domains with DNA in a nucleosome, loop 1 and loop 2. The coloured blocks on the sequence view represent secondary structures: blue, beta-strand; pink, helix; yellow, turn. The tree view shows that Drosophila and human H2A.Z sequences are similar. (B) Comparison of the protein structure of H2A and H2A.Z nucleosome. The 3-dimensional structures of the H2A (light yellow) and H2A.Z (dark yellow) nucleosomes were superimposed. The flexibility of the protein is depicted as the thickness of the tube. The red boxes mark the different structures. The loop 1 and loop 2, the interacting regions between histones and DNA (Thakar et al., 2009), are different in 3-D structure as well as in protein sequences. The loop 1 and loop 2 regions are close to the ±45 and ±55 bp positions of nucleosome DNA sequence.

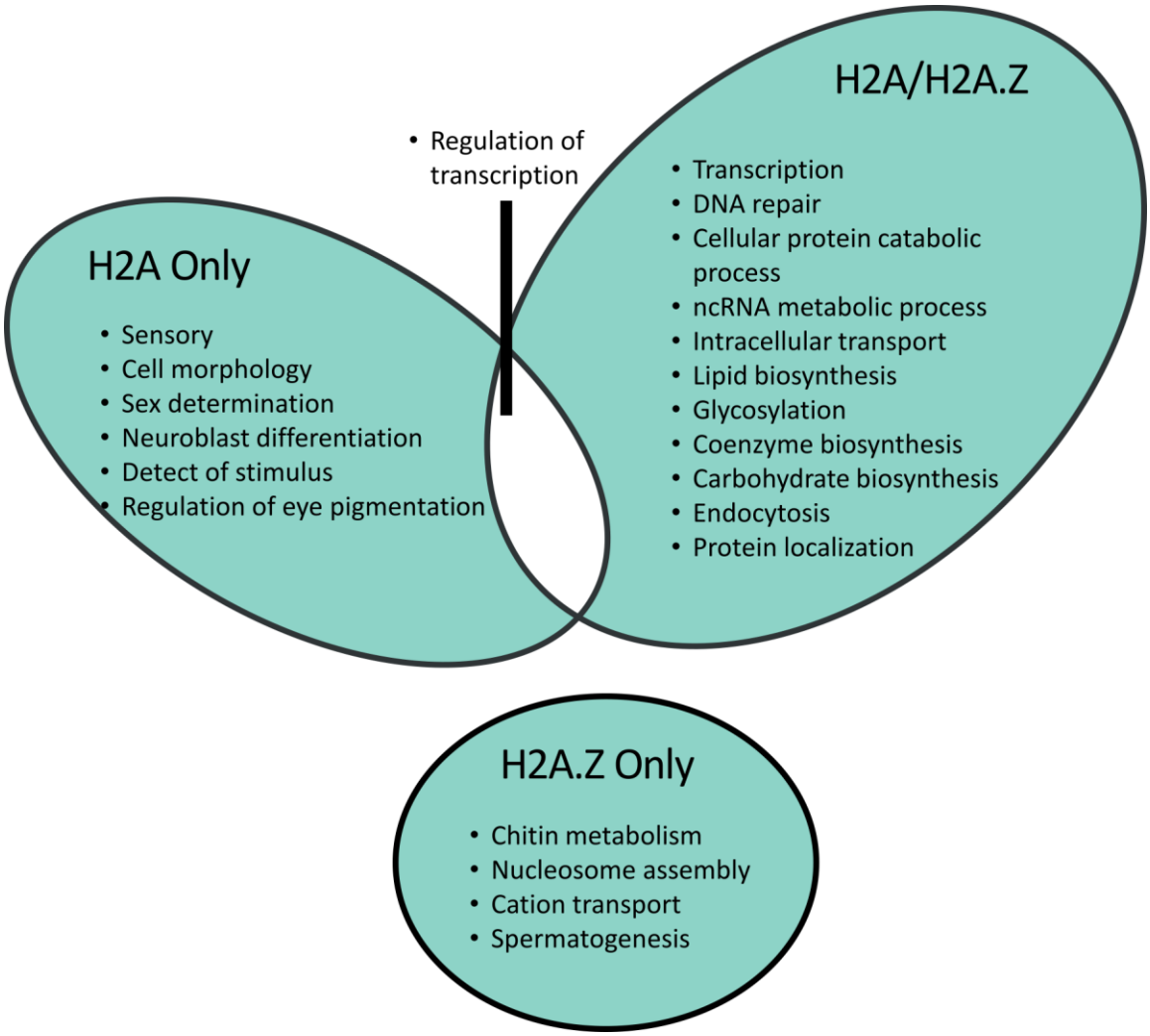
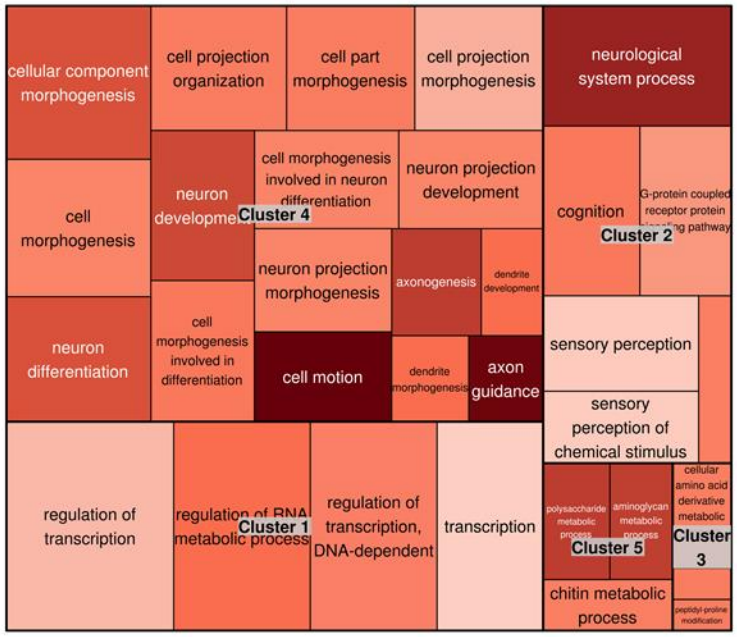

H2A/H2A.Z
• Regulation of transcription
• Transcription
• DNA repair
• Cellular protein catabolic process
• ncRNA metabolic process
• Intracellular transport
• Lipid biosynthesis
• Glycosylation
• Coenzyme biosynthesis
• Carbohydrate biosynthesis
• Endocytosis
• Protein localization
H2A Only
• Sensory
• Cell morphology
• Sex determination
• Neuroblast differentiation
• Detect of stimulus
• Regulation of eye pigmentation
H2A.Z Only
• Chitin metabolism
• Nucleosome assembly
• Cation transport
• Spermatogenesis

Figure 31. Enriched functions of genes grouped by the type of +1 nucleosomes. The enriched gene functions are exclusive to the three groups. Regulation of transcription is the only shared function between H2A-only and H2A/H2A.Z group. The functions of the H2A-only genes are mostly related to development/differentiation. Metabolic pathway related functions are enriched in the H2A/H2A.Z group genes. One of the enriched functions of the H2A.Z-only group is nucleosome assembly, the possible self-regulatory role of nucleosome on the nucleosome assembly proteins.

A


cellular component morphogenesis
cell projection organization
cell part morphogenesis
cell projection morphogenesis
neurological system process
cell morphogenesis
neuron developme
Cluster 4
cell morphogenesis involved in neuron differentiation
neuron projection development
cognition
Cluster 2
G-protein coupled receptor protein signaling pathway
neuron projection morphogenesis
axonogenesis
dendrite development
neuron differentiation
cell morphogenesis involved in differentiation
cell motion
dendrite morphogenesis
axon guidance
sensory perception
sensory perception of chemical stimulus
regulation of transcription
regulation of RNA metabolic process
Cluster 1
regulation of transcription, DNA-dependent
transcription
polysaccharide metabolic process
aminoglycan metabolic process
Cluster 5
cellular amino acid derivative metabolic
Cluster 3
chitin metabolic process
peptidyl-proline modification
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
Benjamini

B

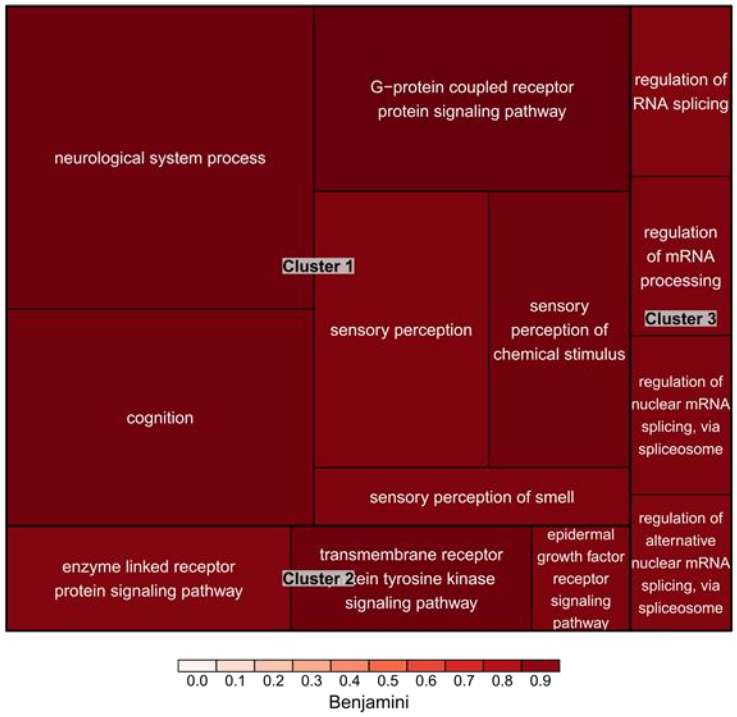

neurological system process
G−protein coupled receptor protein signaling pathway
regulation of RNA splicing
Cluster 1
sensory perception
sensory perception of chemical stimulus
regulation of mRNA processing
Cluster 3
cognition
regulation of nuclear mRNA splicing, via spliceosome
sensory perception of smell
enzyme linked receptor protein signaling pathway
transmembrane receptor
Cluster 2
ein tyrosine kinase signaling pathway
epidermal growth factor receptor signaling pathway
regulation of alternative nuclear mRNA splicing, via spliceosome
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
Benjamini

Figure 32. Enriched functions of genes containing WW/SS NPS patterns. Enriched gene functions were presented in a box plot. The colour represents the adjusted *p*-value, and the size of the rectangle represents the number of the genes with the functions. The genes related to the transcriptional regulation is the most enriched functions. (A) Genes containing the positive WW/SS NPS pattern in their promoters. Transcription regulated genes are enriched in cluster 1. (B) Genes containing the negative WW/SS NPS pattern in their promoters. No significantly enriched functions were detected.

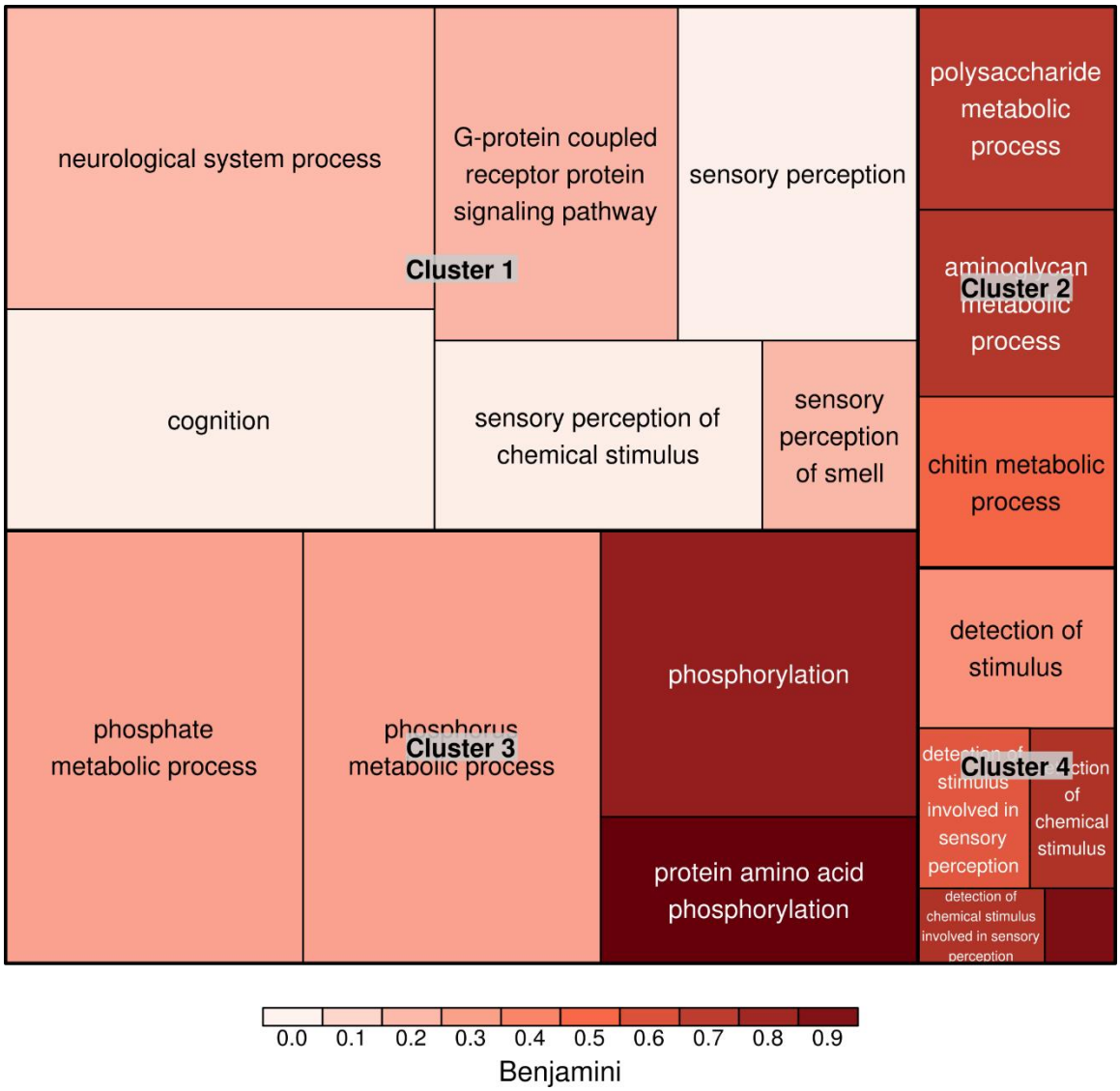

neurological system process
G-protein coupled receptor protein signaling pathway
sensory perception
polysaccharide metabolic process
Cluster 1
aminoglycan
Cluster 2
metabolic process
cognition
sensory perception of chemical stimulus
sensory perception of smell
chitin metabolic process
phosphorylation
detection of stimulus
phosphate metabolic process
phosphorus
Cluster 3
metabolic process
Cluster 4
stimulus involved in sensory perception
of chemical stimulus
protein amino acid phosphorylation
detection of chemical stimulus involved in sensory perception
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
Benjamini

Figure 33. Enriched functions of genes containing the RR/YY NPS pattern sequences. Enriched gene functions were presented in the box plot. The colour represents the adjusted *p*-value. The size of the rectangle represents the number of the genes with the functions. The enriched functions include a variety of roles. Cognition or sensory perception related genes were enriched.

CO-OCCURRENCES OF NUCLEOSOMES AND CORE PROMOTER ELEMENTS

Because the +1 nucleosome is located on the promoters, it may affect the binding of the other proteins such as transcription factors or the functions of the promoters. The possible effect of the nucleosomes in the promoters was examined by searching for the co-existence of the +1 nucleosomes and the core promoter elements.

The presence of the five core promoter elements, a TFIIB recognition element (BRE), an initiator element (Inr), a downstream promoter element (DPE), a CCAAT box, and a TATA box was analysed from the genomic sequences of the promoters. The core promoter elements usually reside within 60 bp upstream and 30 bp downstream of a transcription start site. More specific locations of the core promoter elements relative to the TSS are as follows: CCAAT box is located at -70 to -21 bp, BRE at -37 to -32 bp, TATA at -31 to -26 bp, Inr at -2 to +4 bp, and DPR +28 to 32 bp. Thanks to the close location of the core promoter elements to the location of +1 nucleosomes, the nucleosome can affect the binding of the proteins to the elements either by interfering with the binding by blocking the binding sites or facilitating the binding by presenting the binding sites. The number of the core promoter elements (observed) were counted and compared to the average number of the promoters (expected) in pooled promoters.

The presence of the specific core promoter elements was correlated with the existence of the H2A nucleosomes in the promoter (**Figure 34**). In the promoters with the +1 H2A nucleosomes, CCAAT box and TATA were observed more than expected, while BRE and Inr were more observed in the promoters without the +1 H2A nucleosomes. The number of observed DPE was not significantly different from the expected number of it.

CCAAT is known to be deficient in the promoters of housekeeping genes and is related to the tightly regulated genes, which are turned off when they are necessary and expressed in large quantity once needed. The presence of the +1 H2A nucleosome may

help the tight control of transcription. On the other hand, BRE and Inr are independent of the presence of the +1 H2A nucleosome. As Inr functions similar to TATA, it may not be surprising that the promoters having more Inr have less TATA, and *vice versa*. The A/T rich TATA box sequences, which disfavours the binding of histones, may be related to the dynamic nature of the -1 nucleosome. The BRE sequences are G/C rich.

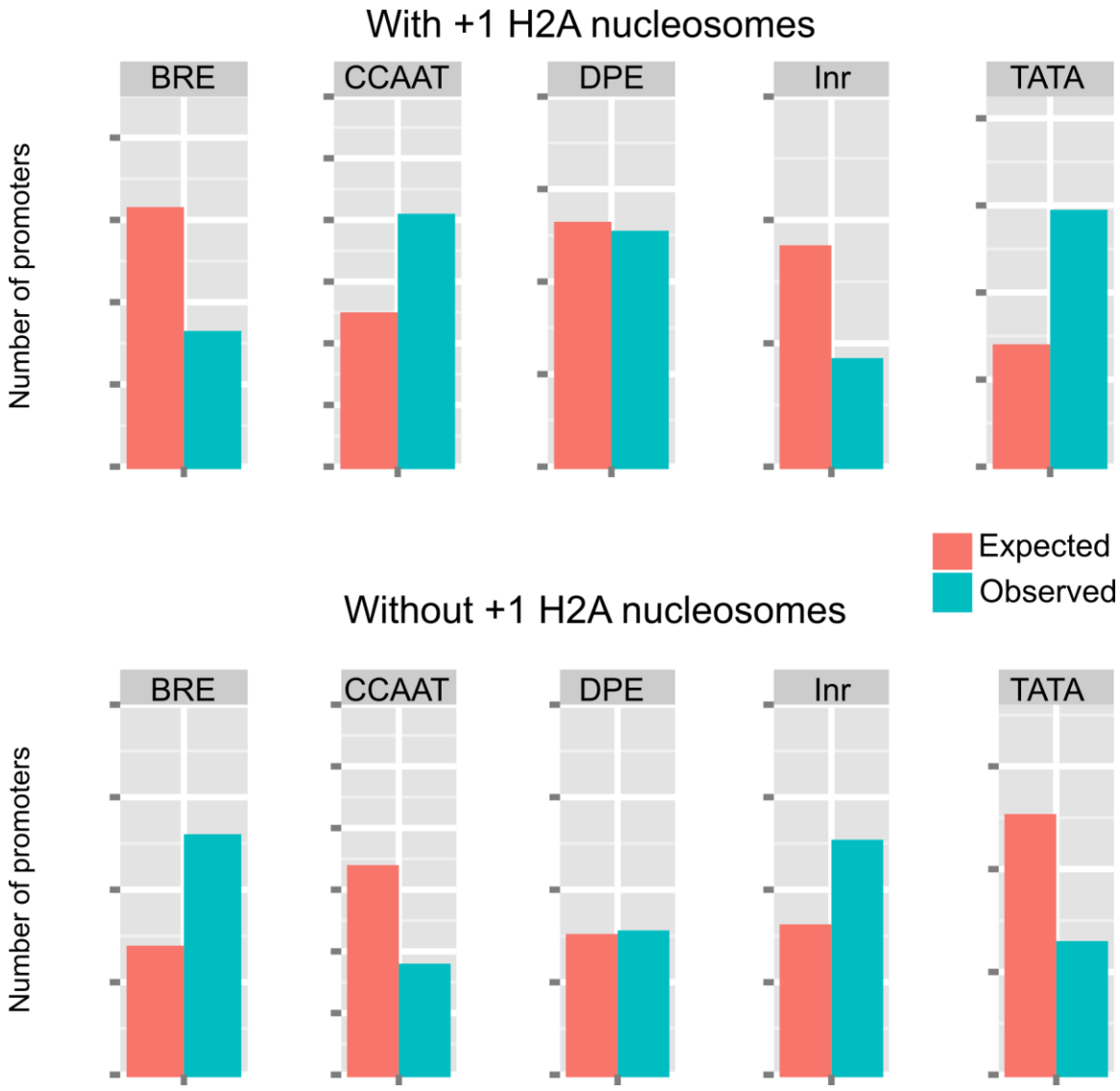

With +1 H2A nucleosomes
BRE
CCAAT
DPE
Inr
TATA
Number of promoters
Expected
Observed
Without +1 H2A nucleosomes
BRE
CCAAT
DPE
Inr
TATA
Number of promoters

Figure 34. Core promoter elements on genes selected by the +1 nucleosomes. The presence of the core promoter elements across the gene subsets grouped based on the type of the +1 nucleosome. The genes and the associated promoters were divided based on the presence of the +1 nucleosome, the H2A nucleosome within 200 bp range from the transcription start sites. The promoter sequences were scanned for core promoter elements. The observed (red) occurrences and the expected (blue) occurrences in the promoters are presented. The expected values were calculated from the entire promoters regardless of the H2A nucleosome. The promoters have a different organisation of the core promoter elements depending on the +1 nucleosome sequences. The TATA box and the CCAAT box were found more than expected in the promoters with H2A +1 nucleosome. On the other hand, BRE was found more in the promoters without the H2A +1 nucleosome (BRE, B recognition element; CCAAT, CCAAT box; DPE, downstream promoter element; Inr, initiator element; TATA, TATA box).

Discussion

Nucleosomes are a nuclear structure with important roles in the positive and negative regulation of gene expression, exerted by altering the binding of transcription factors. Thus, knowing nucleosome positioning is important in understanding the processes and the effects of nucleosomes in gene regulation.

Some nucleosomes are placed at specific positions, and others placed statistically (Mavrich et al., 2008a; Valouev et al., 2008). The specific positioning is observed near the 5' end of genes. The +1 nucleosomes are positioned specifically and affect the positioning of the downstream nucleosomes. Because the +1 nucleosomes are important in the positioning of the nucleosome array, I analysed the DNA sequences of the +1 nucleosomes to determine their NPS patterns. The dinucleotide patterns from the +1 H2A nucleosomes showed particular features. The WW and SS patterns had 10 bp periodicity, and RR and YY patterns also had 10 bp periodicity with extra periods of minor contribution. The periodic patterns were disrupted near the dyad. The pattern analysis revealed the presence of periodic patterns of dinucleotides. The patterns were refined by selecting the nucleosome sequences according to the correlation between the reference patterns and initial WW, SS, RR, and YY patterns. As a result, two canonical NPS patterns were proposed for Drosophila H2A nucleosomes.

The Drosophila WW/SS NPS pattern showed a periodic occurrence of WW and SS dinucleotides along the nucleosome sequence. The periodic pattern started at $\pm$15 bp from the dyad and reached up to $\pm$65 bp on each side of the nucleosome sequence. The periodic occurrences of the WW and SS have them located where the DNA major groove faces outside and inside, respectively. The rotational position of the dinucleotides on the DNA complied with the previously reported dinucleotides and DNA structures (Cui and

Zhurkin, 2010). The periodic positioning of WW and SS facilitates the nucleosome formation due to the anisotropic bending, a property of DNA bending towards a particular direction of nucleosomal DNA (Drew and Travers, 1985). WW dinucleotides tend to have a narrower minor groove, which facilitates the anisotropic bending towards the minor groove. Dinucleotides also have different physical properties such as the degree of sliding and bending (Tolstorukov et al., 2007). The differences in dinucleotide bendability form a specific curvature of the DNA based on the dinucleotide patterns, which influence the nucleosome formation on the DNA (Travers et al., 2009). In a recent study of nucleosome reconstitution using purified proteins showed that the involvement of the remodelling factors in positioning the -1/+1 nucleosome (Krietenstein et al., 2016). RSC recognises the poly(dA:dT) sequences and creates directional nucleosome free regions, which corresponds to the dynamic -1 nucleosome. INO80 alone positions +1 nucleosomes by recognising the dinucleotide DNA sequences and the helix twist by turn.

The periodic patterns were disrupted around the dyad in both yeast and Drosophila according to the local periodicities measured by moving windows. The nucleosome sequence between -15 to +15 bp surrounding the dyad interacts with H3 and H4 histones. Especially, H4 histones strongly interact at the $\pm$15 bp making the positions reactive (Pruss and Wolffe, 1993). The lack of the periodicity in the middle is an important feature of the nucleosomal DNA to interact with other proteins. While the dinucleotide periodicity is a shared feature between Drosophila and yeast NPS patterns, the higher occurrence of SS at the dyad and the disruption of the periodicity in the middle (-15 to +15 bp) distinguished the Drosophila WW/SS NPS pattern from the yeast patterns.

The differences in the NPS patterns between yeast and Drosophila gave interesting insights of the sequence preference according to the histone structures. Yeast nucleosomes are known to be highly euchromatic because they are intrinsically less stable than those of higher eukaryotes (Leung et al., 2016). The *in vitro* reconstitution using recombinant human, yeast, and human-yeast chimeric histones revealed that three amino acids near the C-terminus of histone H3 (Q_{120} K_{121} K_{125}) are responsible for the instability of the yeast nucleosomes. The three amino acids are located near the nucleosome dyad, and the mutations alter the nucleosome positioning *in vivo* and *in vitro* (McBurney et al., 2016). The *in vivo* effect of the nucleosome instability by H3 was demonstrated in yeast with synthetic genome where histone H3 was replaced with the human counterpart (Truong and Boeke, 2017). When the yeast was "reset" with human histones, the yeast showed a slow response to the environmental changes accompanied by the higher DNA occupancy by the human nucleosomes and reduced RNA contents. Converting five histone residues of H3 including the three amino acids back to their yeast sequences restored the robust growth rate. The pattern differences at the dyad are inferred as the effects of the histone structure. The preference of SS dinucleotides at the nucleosome dyad may contribute to more stable nucleosome positioning in Drosophila than in yeast because of the DNA bending property. The major groove faces towards the histone at the dyad, and the anisotropic bending of the SS dinucleotide tends to bend DNA toward the major groove (Richmond and Davey, 2003). Therefore, the preference of SS at the dyad may help form a proper curvature, which favours nucleosome formation.

Another NPS pattern of 10 bp periodicity, in addition to the WW/SS pattern, was identified. This other pattern had AA and TT dinucleotides in the counter phase, just as with the RR/YY NPS pattern of yeast (Ioshikhes et al., 2011). The RR and YY pattern in

Drosophila had a 10 bp periodicity with asymmetry. The nucleosome sequences with the RR and YY NPS patterns are flexible and not intrinsically curved, while sequences with the WW/SS NPS pattern are stiffer and intrinsically curved (Ioshikhes et al., 2011). The RR/YY pattern, with its periodicity mainly on one side, can accommodate the nucleosome formation on less periodic sequences, but this nucleosome positioning is expected to be less stable because the sequence is less favouring.

The enriched locations of the H2A.Z +1 nucleosomes around promoters were different from that of H2A +1 nucleosomes. H2A.Z is deposited by the 13 protein complex SWR1-C in a replication-independent way in yeast (Krogan et al., 2003; Wang et al., 2011) and by Tip60 in Drosophila (Clapier and Cairns, 2009). The truncated H2A.Z lacking the last 20 amino acids can still bind to the SWR1-C but did not provide the restriction with the spread of heterochromatin or chromatin anchoring. The structure of homotypic H2A.Z nucleosome is different from that of an H2A nucleosome (Weber et al., 2010). The differences in the locations and the structures implicated the differences of the sequence patterns of H2A.Z nucleosomes from the canonical patterns.

Indeed, H2A.Z nucleosome sequences had different patterns from the H2A NPS patterns. Both H2A and H2A.Z NPS patterns shared the same dinucleotide periodicities of 10 bp. However, the dinucleotide occurrences were not the same in all peak positions. The most significant differences between the patterns were the dinucleotide peaks at $\pm$45 bp from the dyad. The H2A.Z peaks largely deviated from the prediction model as well as from the H2A canonical patterns. Even though the differences in the physical properties of the H2A and H2A.Z histones are minuscule, the differences in H2A.Z histone affect the nucleosome positioning (Thakar et al., 2009). To understand the contributing factors to the different NPS patterns, I searched for the structural differences between H2A and H2A.Z. The different peak positions coincided to be in the proximity

of the loop 1 and loop 2 interaction domains where the H2A and H2A.Z protein sequences showing differences. Furthermore, the ±45 and ±55 bp positions of the nucleosomal DNA are where the H2A histones interact with DNA (Davey et al., 2002). Therefore, the differences of the DNA interaction domains of H2A and H2A.Z histones are correlated with the different NPS patterns.

H2A.Z is found to be essential for viability in many organisms such as Drosophila (Clarkson et al., 1999; van Daal and Elgin, 1992) and mouse (Faast et al., 2001). In yeast, the role of H2A.Z is known to be related to regulating transcription (Adam et al., 2001; Farris et al., 2005; Larochelle and Gaudreau, 2003; Santisteban et al., 2000). However, not all reports agree about the exact role of H2A.Z nucleosome in gene expression: the H2A.Z nucleosomes seem to be related to both active and inactive genes. The phased H2A.Z nucleosomes at the 5' end of genes are known to be related to the active transcription of the gene (Bargaje et al., 2012; Weber et al., 2010). However, enrichment of H2A.Z was also observed in inactive yeast genes (Guillemette et al., 2005), and was even related to repressing stress-induced genes under normal conditions (Lindstrom et al., 2006), or gene silencing by forming heterochromatin (Swaminathan et al., 2005). In other cases, the enriched H2A.Z nucleosomes were positioned at the 5' ends of genes regardless of the gene activity and were lost following an increase in transcription (Raisner et al., 2005).

Therefore, I propose the role of the H2A.Z nucleosomes as marking the genes to be transcribed, which is accomplished by the unique sequence patterns of H2A.Z nucleosomes. The H2A.Z nucleosome regulates the position of other nucleosomes (Guillemette et al., 2005) and maintains a nucleosome free region by being positioned at the 5' end of genes (Mavrich et al., 2008b). Additionally, H2A.Z nucleosomes are relatively immobile in the DNA, unlike H2A nucleosomes (Li et al., 2005). The sequence

pattern may contribute to the mobility change and enable H2A.Z nucleosomes to work as a barrier or an anchor.

GO enrichment analyses were done on the genes grouped by the +1 nucleosomes and the NPS patterns. The results showed that the +1 nucleosomes and the NPS patterns were related to the genes with distinct functions. The gene groups selected by the type of the +1 nucleosomes (H2A-only, H2A.Z-only, or H2A/H2A.Z) showed exclusive enriched gene functions among them. The functions of the H2A positioned genes were related to neuronal functions such as sensory, detection of a stimulus, and neuroblastoma differentiation; the enriched functions of the H2A/H2A.Z positioned genes were biosynthesis and metabolism-related processes. Interestingly, the H2A.Z positioned genes had nucleosome assembly functions as enriched gene functions, suggesting that H2A.Z may regulate the nucleosome assembly seemingly in self-regulated ways.

The analysis of promoter sequences showed that H2A nucleosome positions have tended to be coupled with TATA and CCAAT boxes. TATA and CCAAT are usually linked with tightly regulated inducible genes, which needs a lot of proteins quickly (Spitz and Furlong, 2012). H2A nucleosomes are related to the transcriptional regulation of complete repression and activation of genes.

The enriched functions of the genes with the canonical NPS patterns showed that the genes with the WW and SS NPS patterns were tightly regulated ones. The H2A nucleosome positioning was also correlated with the core promoter elements regulating gene expression tightly. The strong 10 bp periodicity of the WW/SS NPS patterns may maintain stable binding of histones repressing the expression until it is needed. Interestingly, the genes with RR and YY NPS patterns had the enriched function of neuronal signal transmission. The proteins expressed from the genes were localised at extracellular space or membrane. Cadherin and Ecdysone-induced genes were among the

enriched proteins. Cadherin is one of the essential genes in the development of the neuronal network. In mouse, its expression is controlled by nucleosome remodelling factors: the mutant lacking the remodelling factor failed to develop the healthy neuronal networks (Alvarez-Saavedra et al., 2014). The relationship between the NPS patterns and the remodelling factors would be an interesting subject to study further.

Choi and Kim (2008) showed that the gene expression is related to the promoter sequences affecting the organisation of the nucleosomes, and the periodicity of dinucleotides and DNA bending is related to the DNA-nucleosome interaction leading to variable gene expression. The results presented here also support the relationship between gene functions and the intrinsic nucleosome sequences. The unique biological roles of the grouped genes by the NPS patterns and the core promoter elements suggest the active role of the nucleosome in the transcriptional regulation. The genomic sequences of the promoters were organised in a way that the core promoter elements may collaborate with nucleosomes. Further investigation of the collaboration between the core promoter elements and nucleosomes will be worthy.

In this research, unique NPS patterns were identified in Drosophila. The canonical patterns for H2A nucleosomes were different from the yeast patterns, suggesting species-specific adaptations perhaps resulted from the histone structures. In addition, nucleosomes with histone variants, H2A.Z, had different NPS patterns, which may be related to the different role of the nucleosomes. The +1 nucleosomes specifically positioned by the NPS patterns are closely related to the promoter sequences for transcription factors serving as a part of a complex regulatory network.

PART 2: EFFECT OF NUCLEOSOME POSITIONING ON LY49 TRANSCRIPTION FACTOR BINDING AVAILABILITY

Introduction

Beyond this impressive organisational role, however, the nucleosome—a bulky and ubiquitous protein structure tightly bound to DNA (Lowary and Widom, 1998)—is believed to occupy a central role in the epigenetic regulation of gene transcription (Bai and Morozov, 2010; Wyrick et al., 1999). The role of histone in regulating gene expression has fallen out of interests once but has regained considerable attention since the discovery of the epigenetic elements. And considerably more attention is still required to understand this incredibly complex regulatory network. Not only these histone modifications and the chromatin, but nucleosomes are also believed to regulate gene expression by limiting the accessibility of a specific region of DNA (Lickwar et al., 2009). Originally, nucleosomes were thought to be evenly spaced every ~200 bp of DNA following being randomly positioned (Ioshikhes et al., 2006; Kornberg, 1974). However, many pieces of evidence display specifically positioned nucleosomes in the genome: either by the action of chromatin remodelling proteins (Ito et al., 1997) or by the basal affinity of a given stretch of DNA for a nucleosome (Peckham et al., 2007; Segal et al., 2006; Tillo et al., 2010). In Part 1 of this research, I identified the unique NPS patterns of Drosophila and the collaborative organisation of the nucleosome positions and the promoter elements. The computational analysis suggested that closely connected works of nucleosomes and transcription factors are organised by the genomic sequences. The DNA sequence patterns allowed *in silico* mapping of the 'default' nucleosome landscape of the genome (Ioshikhes et al., 2006; Kaplan et al., 2009)—that is, the nucleosome landscape before remodelling enzymes change it.

The goal of this part of the search is to examine how this 'default' sequence-determined nucleosome landscape interplay with transcription factors and correlate with gene expression. To that end, I have investigated the effects of nucleosome positioning on the expression of a family of immune genes, the Ly49 receptors.

Ly49 receptors—and their functional human analogues, the Killer-cell immunoglobulin-like receptors (KIR)—are expressed on natural killer (NK) cells and other lymphocytes. These receptors can recognise the class-I major histocompatibility complex (MHC-I) expressed on other cells. The recognition of MHC-I allows NK cells to distinguish healthy cells from cancers or virus-infected cells, which is critical to the innate immunosurveillance (Hanke et al., 1999; Kim et al., 2005a). I chose this system as a model to investigate the interplay of nucleosome positions with transcription factors. Aside from its importance in innate immunology, the Ly49 gene family has several attributes enough to become an ideal model for the investigation of the interplay of nucleosomes and transcription factors in transcriptional regulation.

The Ly49 genes all reside in a 650,000 bp region of chromosome 6 as cluster except for one gene and have very similar transcription factor requirements (**Figure 35**). However, expression of an individual Ly49 gene is stochastic. An NK cell can express a set of randomly chosen Ly49 receptors. This variegated expression is the result of stochastic expression of the individual Ly49 gene. As a result, each NK cell acquires a unique repertoire of Ly49 receptors during development and then maintains this repertoire throughout its life (Kubota et al., 1999b; Ortaldo et al., 1999; Pascal et al., 2006). The clustered organisation but stochastic expression of Ly49 genes, therefore, presents a good model system. The differences in nucleosome occupation between expressed and their silent neighbours can be comparable as both of them require the same transcription factors and would be equally influenced by other factors such as

chromosome looping, the actions of locus control regions, presence or absence of transcription factors, and any possible technical variations. Additionally, RMA, a common NK-T cell line expressing only Ly49A and none of the other Ly49 receptors, provides a convenient model for verifying this line of study. Despite the advantages, the stochastic expression of Ly49 in primary NK cells provokes a number of technical challenges for analysis.

The coordinated activity of promoters controls in part this stochastic expression for each gene (Saleh et al., 2004). Pro-1, which is used by immature NK cells, is the first promoter at the upstream of exon -1 and exon -2. Pro-1 is a bi-directional promoter consisting of a conserved transcription factor binding sites (**Figure 36**). There are central AML-1 and NF-κB binding sites, and TATA boxes at 5' and 3' of the promoter, each of which is flanked by C/EBP binding sites (Saleh et al., 2004). The forward and the reverse directions for Pro-1 is chosen randomly by the transcription complex. The transcription complex once choosing the direction, may transcribe in only one of the two directions; backward transcription represses the expression of the Ly49 genes, while forward transcription is to dislodge an inhibitory complex formed by methylation around the downstream second promoter and in some cases third promoter as well, which then maintains the expression of that Ly49 gene for the rest of the NK cell's life (McQueen et al., 2001; Saleh et al., 2004). Conversely, reverse transcription means the antisense transcript represses the transcription from the forward promoter. The forward and reverse transcription may be dependent on the ralative strengh of the promoters modulated by the C/EBP binding sites. In highly expressed Ly49 receptors, their Pro-1 has a "forward" sequence with higher affinity for these transcription factors than their "reverse" sequence, while receptors with low expression rates have a high-affinity "reverse" and low-affinity "forward".

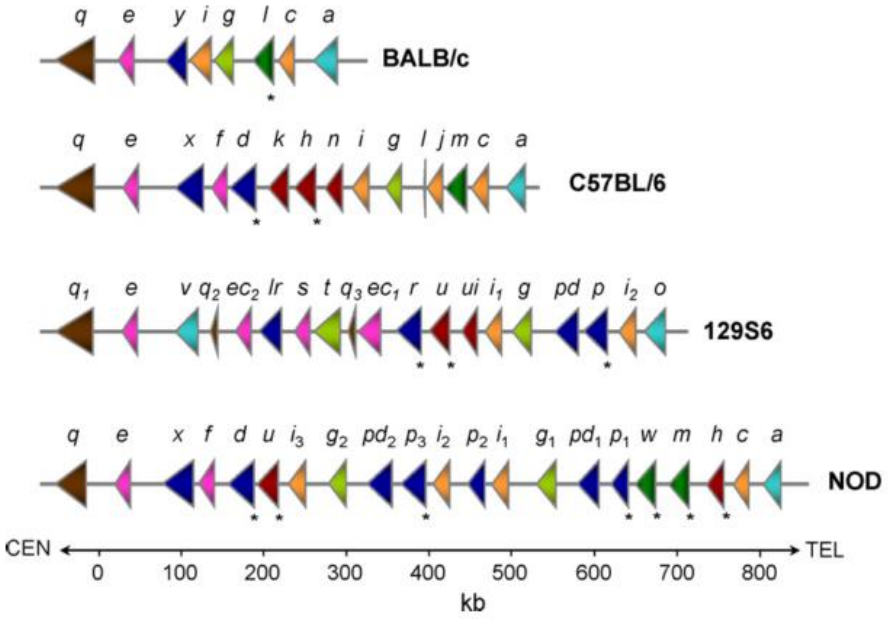

q e y i g l c a
BALB/c
q e x f d k h n i g l j m c a
C57BL/6
q1 e v q2 ec2 lr s t q3 ec1 r u ui i1 g pd p i2 o
129S6
q e x f d u i3 g2 pd2 p3 i2 p2 i1 g1 pd1 p1 w m h c a
NOD
CEN
TEL
0 100 200 300 400 500 600 700 800
kb

Figure 35. Ly49 gene clusters in mouse strains. The schematic organisation of the Ly49 gene cluster is depicted for four inbred mouse strains (BALB/c, C57BL/6, 129S6, and NOD). The orientation and relative span of the genes are indicated by the direction and size of the triangles (scale bar below in kilo bases). Different gene colours represent distinct Ly49 subfamily relationships. Genes encoding stimulatory Ly49 receptors are denoted with an asterisk (*). Ly49b is not included because it is not located in this cluster (Figure adapted from Carlyle et al. 2008).

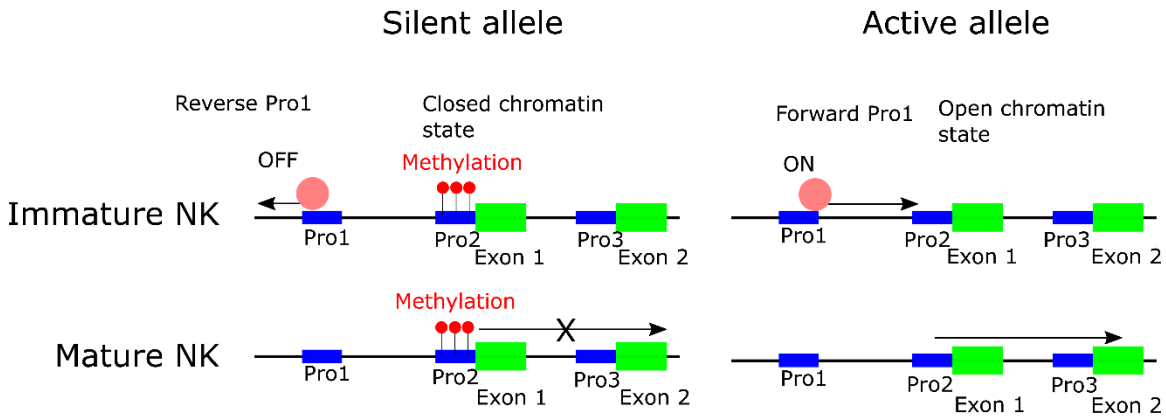

Silent allele
Active allele
Reverse Pro1
Closed chromatin state
Forward Pro1
Open chromatin state
OFF
Methylation
ON
Immature NK
Pro1
Pro2
Pro3
Exon 1
Exon 2
Methylation
X
Mature NK
Pro1
Pro2
Pro3
Exon 1
Exon 2

Figure 36. Organisation of the Ly49 promoter. The antisense promoter (reverse Pro-1) competes with the sense promoter (forward Pro-1). The binding to the reverse or forward Pro-1 is believed probabilistic based on the affinity of the promoter. The transcription from the forward Pro-1 detaches the inhibitory complex at the Pro-2, yet the mechanism has not been defined (Figure modified and redrawn from Pascal et al. 2006).

While the observed stochastic expression pattern of Ly49 receptors during the maturation of NK cells is accountable by the above model to some extent, the variegation of Ly49 receptors in NK cells could not be explained fully because the "forward" and "reverse" sequences in Pro-1 remains the same among the NK cells. There must be another layer of regulatory mechanism to differentiate the various Ly49 receptors in the NK cells. Furthermore, a recent report has shown that Pro-11 affinity does not always correlate with that Ly49's expression level (Gays et al., 2015b).

This finding that there may be another, as-yet-unknown layer of the regulatory elements at play in regulating Ly49 expression gave for the idea of the effects of nucleosome positioning on Ly49 expression. I proposed that the binding of nucleosome on the regulatory regions within the Ly49 cluster may alter the availability of the binding sites to certain necessary transcription factors, which are sensitive to the steric effects of nucleosome coverage. I expected that these sensitive binding sites would reside preferentially within predicted nucleosome-bound regions of DNA. Additionally, I examined the arrangement of the transcription factor binding sites regarding the nucleosome positions to test "in-phase" with nucleosomes by presenting a noticeable pattern of 10 bp periodicity in nucleosome-covered regions. Thanks to the helical structure of DNA, the 10 bp periodicity corresponds to one turn of a DNA double helix (Ioshikhes et al., 1999; Levitt, 1978). The binding sites with this periodicity would be able to put themselves in the same orientation in terms of the histone core of the nucleosome: facing "outward" from the histone core. Such positioning may serve to prevent or to potentiate the binding to the sites by target proteins (Lu et al., 1995).

In this study, I identified that nucleosome positions in the Ly49 promoter overlapped the forward Pro-1 promoter hindering the availability of the binding sites, while keeping the Pro-1 reverse promoter to remain available to the target proteins. The

coverage of transcription factor binding sites by nucleosomes was analysed visually and systematically identifying “open” and “covered” binding sites. Lymphocyte-specific transcription factors were selected to analyse the interplay with nucleosomes, and TATA was included as a control. Among the selected 17 transcription factors, I identified that AML-1 was a transcription factor sensitive to the nucleosome coverage as the binding sites displayed the preferential nucleosome coverage and lack of 10 bp periodicity. The relationship between the factor binding sites and the nucleosome positions were searched by Association Rule Mining and verified by logistic regression confirming the patterns qualitatively. The enriched binding motifs around nucleosomes revealed the genomic layout for both nucleosome positioning and transcription factor binding. With the help of collaborators, we confirmed that AML-1 sites were preferentially depleted of nucleosomes throughout the promoter/enhancer regions of the expressed Ly49 genes. The AML-1 sites were preferentially covered by nucleosome in the unexpressed genes within the same population, implicating nucleosome positioning as a probable mechanism to affect Ly49 expression *in vivo*.

Materials and Methods

PREDICTION OF THE NUCLEOSOME POSITIONING

The genomic DNA sequence of C57BL/6 mice containing the Ly49 gene cluster was available from GenBank (Wilhelm et al., 2002); sequences and annotations of the promoters and exons for 129/S6, BALB/c, and NOD mice were provided by Dr Makrigiannis's lab (Anderson et al., 2005; Belanger et al., 2008; Makrigiannis et al., 2005). The genomic DNA sequences of Ly49 gene cluster were used to predict the nucleosome affinity and the probability of nucleosome positioning by NuPoP (Xi et al., 2010). NuPoP calculates the nucleosome occupancy score and the affinity score on the given nucleotide sequence using a Hidden Markov model. The model oscillates between two states: nucleosome and linker. The probability of observing a nucleotide in a given sequence as a nucleosome was calculated under the 1^{st} or 4^{th} order Hidden Markov model depending on the four previous base composition of the sequence. The probability was calculated by Viterbi algorithm based on the occupancy score. The prediction model has several parameters for the optimised prediction. The species was set as a mouse and the model was set as the 4th order-Hidden Markov model. The occupancy and the affinity score outputs consist of the nucleotide coordinates and the calculated scores at the position. The Viterbi prediction generates the start and the end positions of the predicted nucleosome on the input sequences. The nucleosome dyad position was calculated as the middle of the start and end of a nucleosome from the Viterbi prediction. The nucleosome positions were saved in BED file format, and the affinity scores were stored in WIG file format by default. BED file format is a 0-start coordinate system, in which the coordinate of the first base in a sequence is set as 0. On the other hand, the WIG format is a 1-start coordinate system, in which the first base is set at 1. This different coordinate system was

carefully handled during the data import or data conversion not to make a shift in the coordinates.

ACCURACY OF THE PREDICTION

The nucleosome prediction accuracy was evaluated by comparing the predicted nucleosome positions and the publicly available MNase-Seq data set from Gene Expression Omnibus project (GEO) accession number 58005. The nucleosome positions from the predicted and GEO58005 were compared using BED tools. Nucleosomes of which predicted and MNase-Seq determined positions overlapped at least 50% were taken to represent a true prediction, while other nucleosomes were taken as false. The nucleosomes from the predicted and the MNase-Seq data were presented on the genome using UCSC genome browser.

The nucleotide sequences of the nucleosome-bound regions were retrieved from the given Ly49 gene family sequence using the predicted nucleosome positions. The nucleotide sequences of the nucleosome-bound regions were saved in FASTA format for further uses. The GC content and the ratio were calculated from the saved nucleosome-bound regions of Ly49 gene family, and the mouse chromosome 6 downloaded through the UCSC genome browser.

NUCLEOSOME LANDSCAPE AROUND LY49 GENE PROMOTERS

The promoter architecture was presented around Pro-1 region including Pro-1 and exon -1a. The landscape of the nucleosome positioning sites and the AML-1 binding sites in the promoters were examined to find the relationships. The predicted nucleosome positions and the AML-1 binding sites were displayed on the chromosome aligned at the position of exon -1a for each Ly49 gene in the C57BL/6 mouse. For genes without an exon -1a annotation, the average distance between the Pro-1 and exon -1a from other

annotated genes was used to approximate the location of exon -1a. The nucleosome coverage on the Pro-1 and Pro-2 promoters and the TF binding sites were calculated using BED tools. Each position of the nucleosome was compared with the Pro-1 and Pro-2 promoters, and the length of the overlapped nucleotides was recorded. If more than half the nucleotides of a promoter were covered, then it was marked as closed. Otherwise, the promoter was marked as open.

The nucleosome coverage on the TF binding sites was also calculated using BED tools. The predicted nucleosome positions and the TF binding sites were taken as input, and then the overlapped nucleotides were searched. If a transcription factor binding site and a nucleosome position were overlapped at least by one base pair, the transcription factor binding site was marked as covered. Otherwise, it was marked as open. The gene features (promoters and exons), nucleosome positions, and TF binding sites were visualised using Integrated Genome Viewer (IGV) for visual inspection. The same comparison and alignment were repeated with the nucleosome and TF binding sites data from the mouse strain BALB/c, 129S6, and NOD.

SEARCH FOR HYPERSENSITIVITY REGION

The hypersensitivity region, or hotspot, was searched from the publicly available data on UCSC genome browser. The Ly49 gene family annotation was imported to the UCSC genome browser's custom track. The custom track of the Ly49 gene annotations on chromosome 6 was displayed on the genome browser. For comparison, the hypersensitivity tracks from available tissues like B-cell, T-cell, brain, and heart were added to the genome browser. The Ly49 annotations were used as a guide and checked the hypersensitivity from the selected tissues at the Ly49 promoters and exons.

PREDICTION OF THE TRANSCRIPTION FACTOR BINDING SITES

The same genomic DNA sequences of the Ly49 gene clusters of the four mouse strains were used to predict transcription factor binding sites. The Position Weight Matrix (PWM) of the 17 transcription factors (AML-1, AP-1, C/EBP-beta, Egr-1, Egr-2, Egr-3, GATA-3, IRF-1, Ik-3, Lyf-1, MZF1, NF-AT, NF-κB, Oct-1, STAT3, Sp1, Tal-1alpha/E47, c-Ets-1(p54)) were retrieved from TRANSFAC or JASPAR public databases (Mathelier et al., 2016; Matys et al., 2006). TF binding sites were predicted taking the DNA sequences of the Ly49 gene clusters using either the MATCH or FIMO software (Grant et al., 2011; Kel et al., 2003). MATCH calculates the matrix similarity score for the given sequence. Finding putative TF binding sites depends on the cut-off values, the lowest value to decide the position of the sequence as a TF binding site, for core and matrix similarity. The cut-off values were set to minimise false positives to include only the sites a good similarity even though it may generate less number of sites. The TATA-binding sites were also predicted as a control. The genomic coordinates of the transcription factor binding sites were saved in BED file format.

DISTANCE DISTRIBUTION OF THE TRANSCRIPTION FACTOR BINDING SITES TO THE NEAREST NUCLEOSOME

The distance distribution of transcription factor binding sites around the nearest nucleosome was explored for each transcription factor. First, the predicted positions of the transcription factor binding sites and the nucleosome positions from the Viterbi prediction were converted to BED file format by a custom script. The predicted transcription factor binding sites and the nucleosome positions were then sorted by their positions on the chromosome. The nearest nucleosome for each transcription factor binding site was identified with the *closest* command of the BED tools suite (Quinlan and Hall, 2010b). Once the nearest nucleosome was identified per transcription factor binding

site, the distance between the pair was calculated from the middle position of the transcription factor binding site and the dyad position of the nearest nucleosome. The distance distribution was generated by counting the distances from all pairs of the transcription factor binding site and the nearest nucleosome and used to plot the histogram and the density estimation. The density of the distance distribution was estimated using Kernel Density Estimation by R software (R Development Core Team, 2012). The distribution histogram and the density were plotted aligned at the nucleosome dyad. The nucleosome boundaries were marked by dotted lines at 73 bp apart on both sides from the nucleosome dyad.

TEST THE MULTIMODALITY OF THE DISTRIBUTION

The multimodality of the distance distribution was measured by Kurtosis and tested by Hartigan's *dip test* (Hartigan and Hartigan, 1985). Kurtosis, the 4th measurement of moment assesses the shape of the distribution whether the distribution is heavy-tailed or light-tailed with respect to a normal distribution. Kurtosis for univariate data, $Y_1, Y_2, \ldots, Y_N$, was defined as

$$K[Y] = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4 \,/\, N}{s^4} - 3$$

where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points. Note that the standard deviation is computed with N in the denominator instead of $N - 1$. According to this definition, the kurtosis for a standard normal distribution is 0. The kurtosis was calculated for the distance distributions of the TF binding sites and the nucleosomes to assess whether the TF binding sites are within the nucleosome boundaries or out of the boundaries.

The Hartigan's *dip test* was performed using an R package, *diptest*. The test calculates the maximum differences between the sample distribution and the uniform

distribution. The calculation was repeated with a uniform distribution which minimises the differences. The null hypothesis of the dip test is that the sample distribution is unimodal; the alternative hypothesis is that the sample distribution is multimodal. A sample distribution with a p-value less than 0.05 is considered to be bimodal or multimodal.

ASSOCIATIONS BETWEEN VARIOUS VARIABLES

The visually identified associations between the nucleosome and TF binding sites, promoters and exons, and their nucleosome coverage were examined quantitatively. The association of the variables were evaluated with *support* and *confidence*. The *support* is defined as the percentage of groups that contain all the items listed in that association rule. The support is calculated as:

$$s[x] = \frac{N_x}{N}$$

where N_x is the number of groups containing the items x, N is the total number of all the groups.

The *confidence* is a percentage value that shows how frequently the rule head occurs among all the groups containing the rule body. The confidence c is calculated as:

$$C = P(Ex|Ey)$$
$$= P(E_x \cap E_y)$$

Where E_x and E_y are the items in the association rule. $P(E_x)$ and $P(E_Y)$ are the probability of the item E_x and E_y, respectively. The support and the confidence were calculated by counting the occurrences of the possible combination of variables. The following information at each Ly49 promoter was used as the variables for the association: the presence of nucleosomes and transcription binding sites at Ly49 promoters, the nucleosome coverage statuses (open or covered), and the forward and

reverse promoter (For or Rev). The visually identified patterns, for example, open AML-1 binding sites in reverse promoters, were assessed for the association by calculating the *support* and the *confidence*.

ASSOCIATION RULE MINING

Hidden associations of the factors and the promoter status were analysed by Association rule mining. Association rule mining is a machine learning method to discover the relations, or association, between various qualitative variables. Instead of calculating the association from a given set of conditions, the association rule mining searches for all possible combination of the given variables to find highly associated sets of the variables. The input variables were selected based on various criteria about the genomic features of the Ly49 gene family: promoters and exon regions, TF binding sites and nucleosome positions, open or covered status of the promoters and the TF binding sites, and Ly49 gene names. The associations among the variables were analysed by Apriori algorithm using *arule* package written for R (Hahsler et al., 2005).

Association Rule: If items x and y are present in a group, then item z is also present in the group. The rule is often presented like this:

$$[x, y] \Rightarrow [z]$$

The left-hand side (LHS) of the rule, $[x, y]$, is called *antecedent* or a rule head. The right-hand side (RHS) of the rule, $[z]$, is called *consequent*, or a rule body. The *support* and the *confidence* are the values that measure the association between the items in the rule. The support mainly shows how many cases of the items appeared in the total cases. So, support(x, y) is calculated in this way:

$$Support(x, y) = \frac{the\ number\ of\ cases, x\ and\ y}{the\ number\ of\ total\ cases}$$

The confidence indicates that the item z, where x and y are found together, is present in a certain percentage.

$$Confidence([x, y] => [z]) = \frac{Support(x, y, z)}{Support(x, y)}$$

For example, the rule is presented as follows for an association that forward Pro-1 and nucleosome-bound are associated with AML-1 binding sites:

$$[forward\ Pro1, nucleosome\ bound\ Pro1] => [AML - 1\ binding\ site]$$

And the support and the confidence were calculated like this:

$$Support(forward\ Pro1, nucleosome\ bound\ Pro1)$$

$$= \frac{the\ number\ of\ cases,\ nucleosome\ bound\ forward\ Pro1}{the\ number\ of\ Pro1}$$

ENRICHED MOTIF SEARCH

Enriched motifs are the sequence fragments which are commonly found on the set of the sequences. The nucleotide sequences of promoters or nucleosome bound regions in were searched for enriched sequence motifs. The nucleotide sequences of Pro-1 regions of Ly49 genes were extracted from the chromosome. The selected nucleotide sequences were given as an input to the Peak-Motif (Thomas-Chollier et al., 2012). Peak-Motif identified enriched motifs. The discovered enriched binding motifs were searched against JASPAR database to find matching or similar motifs of known transcription factors. The enriched motifs search was repeated with the sequences between exon 1 and exon 2 of the Ly49 gene family, which includes Pro-2/Pro-3 regions. The enriched motif search was also run with the sequences around the nucleosome positioning sites. The dyad positions of all nucleosomes or the nucleosomes in the Pro-1 promoters of Ly49 genes

were selected. The nucleosome dyad positions were extended by 150 bp toward 5' and 3' directions of the chromosome and the nucleotide sequences were extracted from the 300 bp long regions. The 300 bp long nucleotide sequences were divided into six groups of 50 bp long fragments from the 5' to the 3' direction. The 50 bp sequences were searched for enriched motifs with Peak-Motif.

IDENTIFYING THE 10 BP PERIODICITY OF TRANSCRIPTION FACTORS

The distance periodicity between the transcription factor binding sites and the proximal nucleosome was examined by Fourier transform. The transcription factor binding sites were selected from various regions: Pro-1 and Pro-2 promoter regions, nucleosome covered regions, and nucleosome free regions. For each selected transcription factor binding site, the nearest nucleosome was found by comparing the positions using Bedtools. The distances between the middle position of selected transcription factor binding sites and the dyad position of the nearest nucleosomes were calculated. The distances were then counted to get distribution. The distance distribution of all the selected pairs was analysed by Fourier transform for the periodicity, or they were grouped for each Ly49 gene before Fourier transform. Fourier transform calculated the spectral density of the period to generate a periodogram. The spectral density tells the contribution of each period to the repeating pattern. The period with maximum spectral density was considered as the dominant period of the distance distribution.

The leave-one-out analysis was performed to find the distance periodicity of each factor. The spectral density of the pooled distance distributions from the 17 transcription factors was used as a baseline diagram. The Fourier transform was repeated with the leave-one-out sample sets of 16 transcription factors after removing one transcription factor at a time from the whole sample set. The periodicities from the leave-one-out

samples were compared with the baseline periodicities. Once the deviation, the decrease or increase of the spectral density at the 10 bp period from the baseline, was observed in the periodicity generated from the leave-one-out sample, then the left-out factor was considered contributing to the 10 bp periodicity of the baseline. The transcription factors were grouped based on genomic regions (promoters or exons) or nucleosome-bound statuses to investigate the relationship between the periodicity and the role of the transcription factor.

NUCLEOSOME MAP OF RMA CELLS BY CHIP-SEQ

The predicted nucleosome binding positions were compared with the experimentally determined positions with MNase-Seq from the mouse NK-T cell line, RMA, as described in the reference (Henagan et al., 2015). The library was prepared by Génome Quebec at McGill University, and DNA samples were prepared and sequenced on an Illumina MiSeq (Illumina, San Diego, CA). FastQC checked the quality of the sequence reads, and the low-complex redundant sequence reads were removed before alignment. The sequence reads aligned on the mouse genome mm10 with Bowtie 2.0. The aligned short reads were used to find nucleosome peaks with MACS. The sequencing results have been deposited in the GEO database (accession number GSE71863).

DEVIANCY OF NUCLEOSOME POSITIONING IN RMA FROM THE PREDICTION

The predicted nucleosome-bound-regions and the MNase-Seq determined nucleosome-bound-regions were compared using the UCSC table browser. True predictions, nucleosome-bound-region in both predicted positions and MNase-Seq determined positions, and false predictions, nucleosome-bound-region in predicted positions and nucleosome-depleted-region in MNase-Seq determined positions were intersected using the UCSC table browser. The same process was repeated for each

transcription factor binding sites to identify the number of true-predictions and the false-predictions. The accuracy of the prediction was calculated for nucleosomes in general and in each TF binding site, and then summarised for each Ly49 gene. The results were presented as a heat map between the nucleosome predictions against the transcription factors.

STATISTICAL TESTS OF THE DEVIANCY

A chi-square test was constructed to detect the nucleosome deviancy from the prediction in Ly49 genes. The overall true-positive rate of each Ly49 gene across the transcription factors was set as the expected value. The true-positive of each Ly49 gene per transcription factor was set to the observed value. Individual chi-square statistics, as well as the overall statistic, were presented in a heat-map.

LY49A EXPRESSION ON RMA CELLS

RMA cells or isolated mouse splenocytes as a negative control were counted by flow cytometry after staining with antibodies against NK1.1, TCRβ, and the Ly49 receptors such as Ly49A/D (clone 4E5), Ly49D, Ly49C/I (5E6), Ly49E/F, Ly49G (4D11), and Ly49H (3D10). Antibodies were purchased from Becton Dickinson (San Jose, CA) or eBioscience (San Diego, CA), and samples were sorted using a Beckman Coulter CyAN-ADP and analysed using Kaluza (Beckman Coulter, Mississauga, ON).

CHROMATIN IMMUNOPRECIPITATION OF RMA CELLS FOR AML-1 BINDING SITES

Fragmented chromatin was prepared from RMA cells by MNase digest as described previously (Henagan et al., 2015), except that 150 bp fragments were selected by gel extraction. Chromatin was then incubated with a rabbit polyclonal antibody raised against AML-1 or immunoglobulin M (as an isotype control) overnight at 4 °C (Abcam). Antibody complexes were collected using protein A agarose beads, and DNA was

purified using a high pH chelating solution, as previously described (Flanagin et al., 2008; Nelson et al., 2006). The following primers were used:

Ly49A promoter:	AGGCCAGGGAAACCTGGTGTA
	AAGAGGTGGGGCACTGGACTG
Ly49A 6 kb up:	ACAGAACTCAGAGGGCAAAGGAAA
	TGGGCCACTTGGCCATTTATCT

Real-time PCR was performed using an Eppendorf thermal cycler.

PROTEIN INTERACTION

The protein interaction was analysed with GeneMania (Warde-Farley et al., 2010). The selected gene names from association rule mining and the enriched motifs were used to search for interacting proteins. The search organism was a mouse, and the network parameters were set as physical interactions and coexpression. GeneMania analysed the enriched functions of the interacting proteins. The enriched functions were ranked based on the False Discovery Rate (FDR).

Results

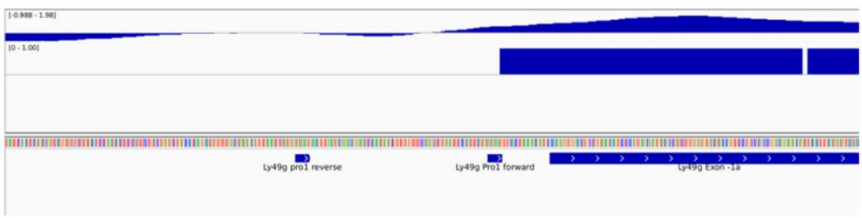
PREDICTION OF NUCLEOSOME POSITIONING SITES

Using the C57BL/6 mouse genome sequences of Chromosome 6, the nucleosome positioning sites were predicted on the Ly49 gene cluster. For the nucleosome prediction on the mouse, NuPoP (Wang and Xi, 2012) package was used. NuPoP produces nucleosome occupancy score, histone affinity score, and Viterbi prediction of the nucleosome position. The prediction calculates the affinity score and occupancy score from the genomic sequence through Fourth-order Hidden Markov Model. The Fourth-order Hidden Markov Model is the most sophisticated model available in the area, and the only predictor optimised for mouse genome sequence up to our knowledge. The species option was set as a mouse so that the parameters, which were optimised for the mouse genome sequence used for the model. The affinity score and the occupancy score calculates the binding affinity of the histone and the probability of the occupancy at the position, respectively. The Viterbi output takes the binding probability and produces dichotomy outputs: nucleosome-free region and nucleosome-occupied region on the input genome sequence. The Viterbi output was used to determine the nucleosome positions. An example view of the prediction around Ly49G promoter was shown in **Figure 37A**. The affinity score, the nucleosome position deduced from the Viterbi output, and the annotations of the genes, such as promoter and exons were presented. The predicted nucleosome positions were compared with the published nucleosome occupancy in mouse hepatocytes. The overall accuracy was 49% (**Figure 37B**). The prediction was made on the whole sequence of the Ly49 gene family. The nucleosome bound regions had slightly higher GC content than the average GC content of the Chromosome 6 (**Table 3**).

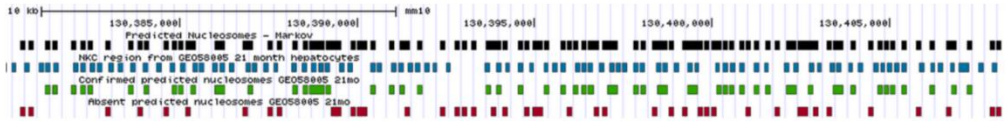
Once the probable nucleosome map on the Ly49 genes was calculated, we asked whether the nucleosome positions may contribute to the transcriptional regulation. The relative positions of the nucleosome positioning sites and the promoters and exons were tested to determine the possible interaction of the nucleosomes with the promoters and transcription factors. For each Ly49 gene in the C57BL/6, a specific pattern of nucleosome occupancy on the promoters and the exons was looked for. In most cases, the Pro-1 reverse promoter favours the open configuration, while the forward Pro-1 promoter favours a bound configuration (**Figure 38**). The same analysis was repeated for Ly49 gene families from BALB/c, 129/S6, and NOD mouse strains. In most Ly49 genes from the three mouse strains, I observed similar nucleosome positioning configuration around the Pro-1 promoter: an open configuration of the Pro-1 reverse promoter and a bound configuration of the forward Pro-1 promoter (**Figure 39** - **Figure 41**).

The presence of the AML-1 binding sites between the forward and the reverse promoters was noticeable, which agreed with the previous results from the C57BL/6 Ly49G promoter region (Saleh et al., 2004). The AML-1 binding sites were found in almost all Ly49 genes regardless of the mouse strain. The ubiquitous presence of the AML-1 binding sites at the promoter regions leads to an investigation of the identification of the binding sites of the immune specific transcription factors and the relations with the nucleosomes in promoters. The open and bound configurations of the reverse and forward Pro-1 can contribute to the accessibility of the chromosome to the proteins. The hypersensitivity regions, or hot spots, were examined in the published tissue samples. The hypersensitivity regions of Ly49 gene family on chromosome 6 were retrieved from lymphocytes such as B-cell and T-cell as well as brain and heart tissues (**Table 4**).

A



B



Confirmed Predictions: 48.8%

Figure 37. Hidden Markov prediction of nucleosome positions. (A) An example of an alignment showing the raw genomic affinity (1st track) and the probable occupancy (2nd track) of the Ly49G Pro-1 region (3rd track). (B) Prediction (black) was validated by comparing to a published nucleosome positions (blue) of MNase-Seq dataset from the GEO (accession number GEO58005). Predictions having > 50% overlap were marked as true predictions (green). Other predictions were taken as false (red).

Region	G+C	Total Bases	%GC
Chromosome 6	60606617	146336545	41.42%
Nucleosome-bound regions	43344070	94657763	45.79%

Table 3. GC content of the nucleosome covered regions. Nucleosome covered regions have slightly higher GC content than the average GC content of the chromosome 6. Repeating A and T or poly(dA:dT) tract is known as a determinant of nucleosome position, which is less favouring nucleosome occupation. Ly49 gene family: DNA covered by nucleosome: 305,466 bp Total DNA size: 572,712 bp.

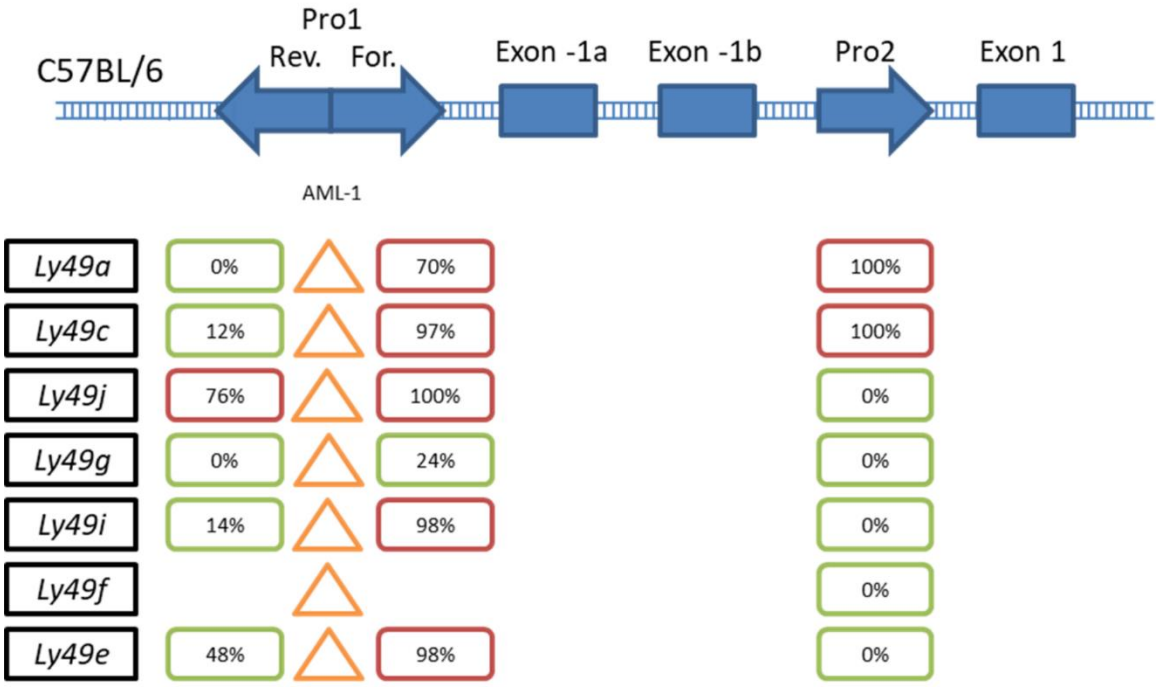

C57BL/6
Pro1
Rev.
For.
Exon -1a
Exon -1b
Pro2
Exon 1
AML-1
Ly49a
0%
70%
100%
Ly49c
12%
97%
100%
Ly49j
76%
100%
0%
Ly49g
0%
24%
0%
Ly49i
14%
98%
0%
Ly49f
0%
Ly49e
48%
98%
0%

Figure 38. Predicted nucleosome hindrance at extended promoter regions in C57BL/6. Schematic view of the predicted nucleosome positions and the Ly49 gene elements. Nucleosome positions were determined the 4th order Hidden Markov model of NuPoP package. The Ly49 gene elements were presented on top: promoters (blue arrow), exons (blue square). The relative nucleosome positions and AML-1 binding sites were presented below the Ly49 gene elements. Nucleosome (round square), AML-1 binding site (triangle). The number inside the square represents the proportion of the promoter length covered by the nucleosome.

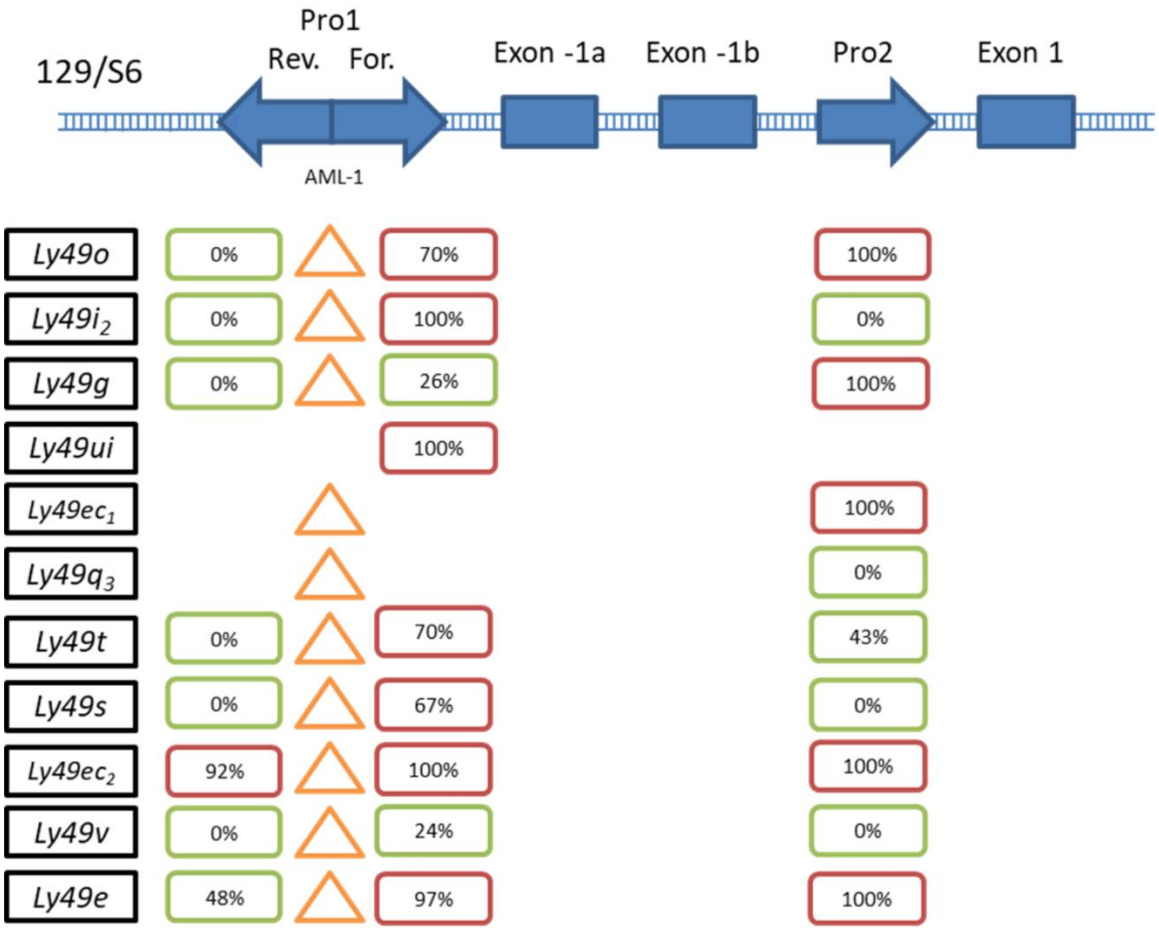

129/S6
Pro1
Rev.
For.
Exon -1a
Exon -1b
Pro2
Exon 1
AML-1
Ly49o
0%
70%
100%
Ly49i2
0%
100%
0%
Ly49g
0%
26%
100%
Ly49ui
100%
Ly49ec1
100%
Ly49q3
0%
Ly49t
0%
70%
43%
Ly49s
0%
67%
0%
Ly49ec2
92%
100%
100%
Ly49v
0%
24%
0%
Ly49e
48%
97%
100%

Figure 39. Predicted nucleosome hindrance at extended promoter regions in 129/S6. Schematic view of the predicted nucleosome positions and the Ly49 gene elements. Nucleosome positions were determined the 4th order Hidden Markov model of NuPoP package. The Ly49 gene elements were presented on top: promoters (blue arrow), exons (blue square). The relative nucleosome positions and AML-1 binding sites were presented below the Ly49 gene elements. Nucleosome (round square), AML-1 binding site (triangle). The number inside the square represents the proportion of the promoter length covered by the nucleosome.

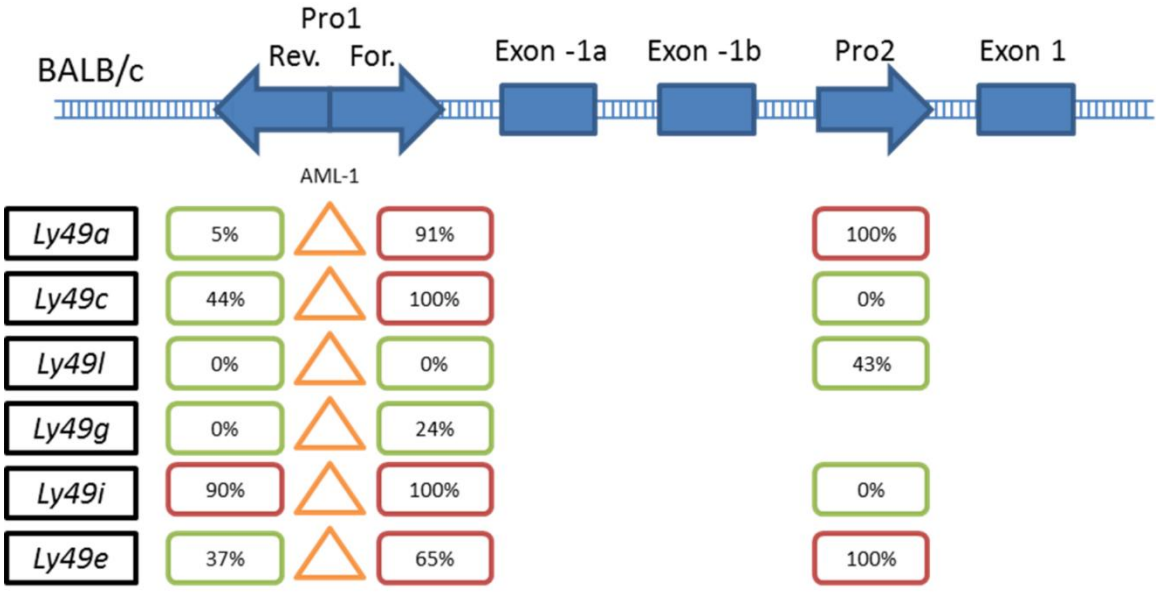

Pro1
Rev.
For.
BALB/c
Exon -1a
Exon -1b
Pro2
Exon 1
AML-1
Ly49a
5%
91%
100%
Ly49c
44%
100%
0%
Ly49l
0%
0%
43%
Ly49g
0%
24%
Ly49i
90%
100%
0%
Ly49e
37%
65%
100%

Figure 40. Predicted nucleosome hindrance at extended promoter regions in BALB/c. Schematic view of the predicted nucleosome positions and the Ly49 gene elements. Nucleosome positions were determined the 4th order Hidden Markov model of NuPoP package. The Ly49 gene elements were presented on top: promoters (blue arrow), exons (blue square). The relative nucleosome positions and AML-1 binding sites were presented below the Ly49 gene elements. Nucleosome (round square), AML-1 binding site (triangle). The number inside the square represents the proportion of the promoter length covered by the nucleosome.

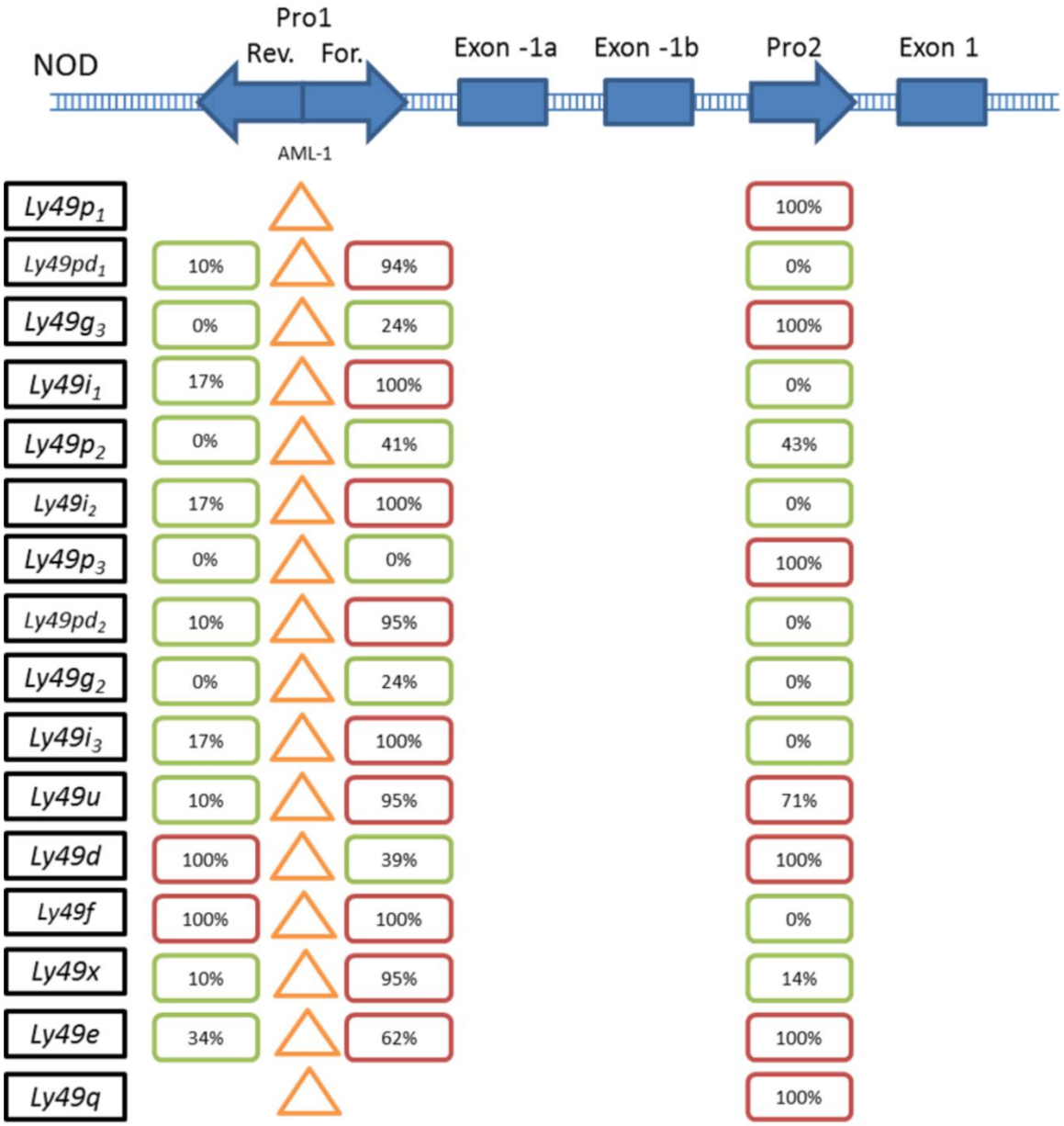

NOD
Pro1
Rev.
For.
Exon -1a
Exon -1b
Pro2
Exon 1
AML-1
Ly49p1
Ly49pd1
Ly49g3
Ly49i1
Ly49p2
Ly49i2
Ly49p3
Ly49pd2
Ly49g2
Ly49i3
Ly49u
Ly49d
Ly49f
Ly49x
Ly49e
Ly49q
10%
94%
0%
24%
17%
100%
0%
41%
17%
100%
0%
0%
10%
95%
0%
24%
17%
100%
10%
95%
100%
39%
100%
100%
10%
95%
34%
62%
100%
0%
100%
43%
0%
100%
0%
0%
71%
100%
0%
14%
100%
100%

Figure 41. Predicted nucleosome hindrance at Ly49 extended promoter regions in NOD. Schematic view of the predicted nucleosome positions and the Ly49 gene elements. Nucleosome positions were determined the 4th order Hidden Markov model of NuPoP package. The Ly49 gene elements were presented on top: promoters (blue arrow), exons (blue square). The relative nucleosome positions and AML-1 binding sites were presented below the Ly49 gene elements. Nucleosome (round square), AML-1 binding site (triangle). The number inside the square represents the proportion of the promoter length covered by the nucleosome.

Genes		Major hot spot			
Ly49	Klra	B-cells	T-cells	Brain	Heart
a	1	exon +1	exon -1	ND	ND
c	3	exon +1, -1	exon -1	ND	ND
d	4	ND	ND	ND	ND
e	5	ND	exon -1	ND	ND
f	6	Upstream	Upstream	ND	ND
g	7	exon +1	exon -1	ND	ND
h	8	ND	ND	ND	ND
i	9	exon -1	exon -1	ND	ND
j	10	ND	exon -1	ND	ND
k	11	exon 3	exon 3	ND	ND
m	13	ND	Upstream	ND	ND
n	14	ND	ND	ND	ND
q	17	ND	ND	ND	ND
x	22	ND	ND	ND	ND

Table 4. Hypersensitivity regions of the Ly49 gene cluster. The hypersensitivity region, or the hot spot, were searched on the chromosome of the Ly49 gene cluster using the publicly available data sets. The hypersensitivity regions, which are accessible by restriction enzymes, considered to be accessible by transcription factors. The hypersensitivity between lymphocytes and other tissues were compared. The Pro-1 and Pro-2 regions of Ly49 gene family in B-cells and T-cells, showed hypersensitivity, while Ly49 gene family in brain and heart were not accessible by the restriction enzymes. The differences in the DNA conformation may be maintained by nucleosomes.

In more than half of the Ly49 genes in B-cell or T-cell, hypersensitivity regions were found in the Ly49 promoters and the exons. In contrast, no promoters and exons of Ly49 genes in the brain and heart tissues showed the hypersensitivity. The presence of the hypersensitivity region is correlated with the expression level of the gene in NK cells. The Ly49J and Ly49E, whose reported expression levels are less than 1% in the NK cells, did not show the hypersensitivity in their promoters. Instead, the nucleosome affinity predicted from the Pro-1 sequences was high suggesting that the promoters probably bound by nucleosomes. So, the hypersensitive regions in the Ly49 promoters, important in the transcriptional regulation, may be maintained by nucleosomes.

IDENTIFICATION OF NUCLEOSOME COVERAGE ON TRANSCRIPTION FACTOR BINDING SITES

As the AML-1 binding sites were common in the Ly49 Pro-1, I next sought to determine whether other TF binding sites have a unique relationship with nucleosome positions: sensitive to the nucleosome coverage. Since Ly49 receptors are known only to be expressed in immune cells, I selected 17 transcription factors, whose functions are well-known in regulating lymphocyte gene expression (**Table 5**). The selected factors include factors with known Ly49 interactions such as AML-1 and C/EBP-beta (Saleh et al., 2004) and other common lymphocyte transcription factors. TATA was included as a control, which is expected to be positioned away from nucleosome dyads. The position weight matrices of those transcription factors were retrieved from JASPAR and TRANSFAC databases (**Figure 42**). The binding sites of all the selected factors were identified in the Ly49 gene family by calculating the score with MATCH. The cut-off values for the score for identification were set to minimise false-positives.

The identified binding sites were presented together with the nucleosome positions using IGV genome browser. The number of the open and nucleosome-bound TF binding sites was counted. If the identified TF binding sites were overlapped with the predicted nucleosome positions, then the TF binding site was deemed as a nucleosome-bound or covered TF binding site. Otherwise, the TF binding sites were marked as open. The ratio of open TF binding sites to all TF binding sites, which were counted in the Ly49 gene cluster, identified open TF binding sites and nucleosome-bound TF binding sites (**Figure 43**). More than 80% of TATA-binding sites are predominantly open: the binding sites avoid the possible nucleosome positioning. NF-AT, Oct-1, IRF-1, and GATA-3 followed the TATA in the openness of the binding sites. More than 50% of the binding sites of those factors are not overlapped with a nucleosome positioning site. They are supposed to be open as their nucleotide sequences avoid nucleosome positioning as a default genomic configuration.

On the other hand, other transcription factors had their binding sites overlapped with nucleosomes. Transcription factors like Sp-1, Egr-1, NF-κB, Egr-2, Ik-3, Lyf-1, MZF1, AML-1, Tal-α/E47, and AP-1 showed that more than 60% of their binding sites were nucleosome-bound or overlapped with nucleosome positions. Because these open and covered statuses are based on the genomic sequences, the statuses may be considered as a default configuration. Even though the actual statuses may vary depending on the tissues and cell statuses, the possibility that the TF binding sites are bound by nucleosomes *in vivo* is higher by these transcription factors than the open transcription factors unless other factors would be involved.

Symbol	Factor name	Class
AML-1	runt-factor AML-1	Runt
AP-1	activator protein 1	bZIP
C/EBP-beta	CCAAT/enhancer binding protein beta	bZIP
c-Ets-1(p54)	c-Ets-1(p54)	ETS
Egr-1	Egr-1/Krox-24/NGFI-A immediate-early gene product	C2H2 Zinc finger
Egr-2	Egr-2/Krox-20 early growth response gene product	C2H2 Zinc finger
GATA-3	GATA-binding factor 3	Zinc finger
IRF-1	interferon regulatory factor 1	Trp cluster
IK-3	Ikaros 3	C2H2 Zinc finger
Lyf-1	LyF-1	C2H2 Zinc finger
MZF1	MZF1	C2H2 Zinc finger
NF-AT	Nuclear factor of activated T-cells	Rel-related factor
NF-κB	NF-kappaB binding site	Rel-related factor
Oct-1	octamer factor 1	Pou domain
STAT3	signal transducer and activator of transcription 3	STAT
Sp1	stimulating protein 1	C2H2 Zinc finger
Tal-1 α /E47	Tal-1alpha:E47 heterodimer	bHLH
TATA	Retroviral TATA box	TATA (TBP)

Table 5. Selected immune specific transcription factors used in this study. These 17 transcription factors were selected for their roles in immunity. In addition to the AML-1, which is considered to play a role in Ly49 transcriptional regulation, other transcription factors are related to the transcriptional regulation in lymphocytes. TATA was selected as a control.

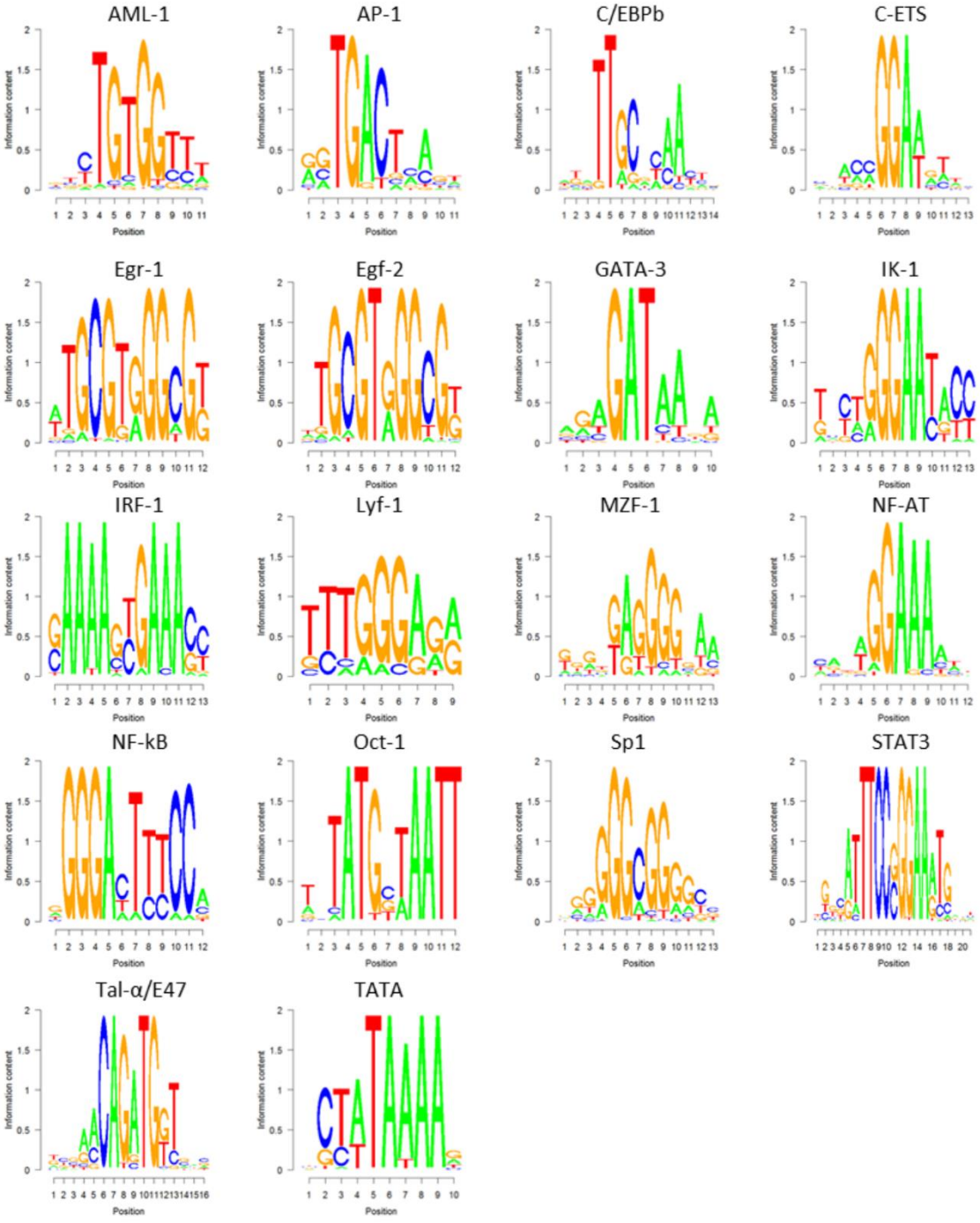

AML-1
AP-1
C/EBPb
C-ETS
Egr-1
Egf-2
GATA-3
IK-1
IRF-1
Lyf-1
MZF-1
NF-AT
NF-kB
Oct-1
Sp1
STAT3
Tal-α/E47
TATA
Information content
Position

Figure 42. Sequence motifs of the selected transcription factors. Position Weight Matrices (PWM) obtained from the TRANSFAC or JASPAR public databases are shown as sequence logos. The DNA sequences of the Ly49 gene family were scanned using the Position Weight Matrices to calculate the score for the transcription factor binding sites.

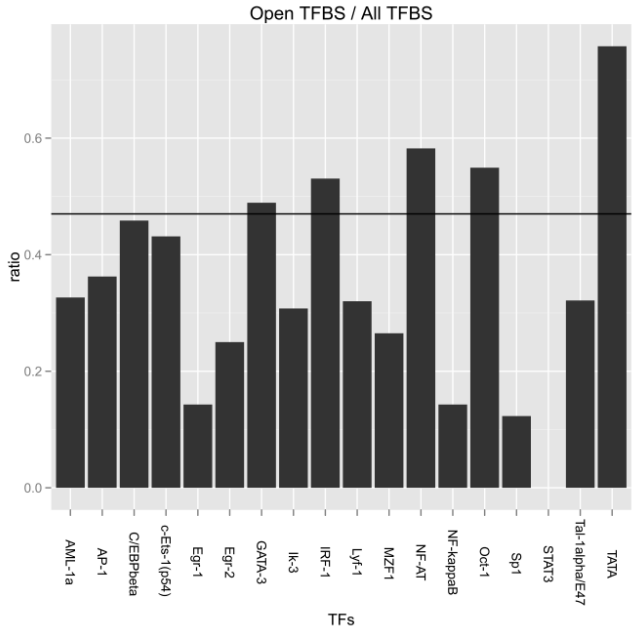

Open TFBS / All TFBS
ratio
0.0
0.2
0.4
0.6
AML-1a
AP-1
C/EBPbeta
c-Ets-1(p54)
Egr-1
Egr-2
GATA-3
Ik-3
IRF-1
Lyf-1
MZF1
NF-AT
NF-kappaB
Oct-1
Sp1
STAT3
Tal-1alpha/E47
TATA
TFs

Figure 43. The ratio of the open and covered transcription factor binding sites. The 50% line (black horizontal line) represents that even number of the TF binding sites are open or covered. TATA-binding sites are mainly open: not covered by nucleosomes. NF-AT, Oct-1, IRF-1, and GATA-3 follow TATA in the open TF binding sites. On the other hand, Egr-1, Sp1, NF-κB, Egf-2, Ik-3, Lyf-1, MZF1, AML-1, Tal-1-α, and AP-1 are more likely covered by nucleosomes. The AML-1, Lyf-1, MZF1 binding sites were also identified as covered by nucleosomes in the distance distribution analysis.

QUANTITATIVE ASSESSMENT OF THE NUCLEOSOME POSITIONING IN THE LY49 GENE PROMOTERS

To assess the open and nucleosome-bound status in Ly49 promoters and the TF binding sites quantitatively, the relationship was analysed by establishing association rules. In the association rule analysis, typically *items* and *groups* are specified. The *items* are the units in the identified rule of an association. The *groups* are the units that contain the items. For example, for the association of the presence of the AML-1 binding site and openness of the binding site, each of the AML-1 binding site and the openness are items, and the combination of the AML-1 binding site and the openness are the groups. In general, association rules reveal a correlation between qualitative variables.

According to the associations between the AML-1 binding sites and promoters in the C57BL/6 Ly49 gene cluster, the found associations supported our identification of the unique patterns between TF binding sites and the nucleosome sites in the promoter region: open Pro-1 reverse promoter and closed forward Pro-1 promoter, (**Table 6**). The confidence of the rule $[Forward\ Pro1] \Rightarrow [Closed]$ is 1.0, while the confidence of the rule $[Forward\ Pro1] \Rightarrow [Open]$ is 0. The support and the confidence indicate the association of open status for the Pro-1 reverse promoters. The confidence of the rule $[Reverse\ Pro1] \Rightarrow [Open]$ is 0.8, while $[Reverse\ Pro1] \Rightarrow [Closed]$ is 0.2. The rules found other associations among the binding sites, nucleosome positions and the promoters. The rule $[Reverse\ Pro1, Open] \Rightarrow [AML-1]$ has the confidence of 1.0 and the support of 0.5, which is interpreted that out of all cases counted, Reverse Pro-1, Open, AML-1 conditions found together in half of cases (support = 0.5), and whenever there are Reverse Pro-1 and Open, there is AML-1 binding site also (confidence = 1.0).

The association was repeatedly calculated in the Ly49 gene clusters were from other mouse strains of BALB/c (**Table 7**) and 129/S6 (**Table 8**). The analyses of BALB/c

and 129/S6 strains also support the same associations between the AML-1 binding sites and the nucleosome positioning in the Ly49 promoters: AML-1 binding sites are likely nucleosome-bound.

NUCLEOSOME COVERAGE OF TRANSCRIPTION FACTOR BINDING SITES

To identify the patterns of nucleosome coverage of the 17 transcription factors, I computed the distance distribution between the transcription factor binding sites and the nearby nucleosome dyad. Because a typical nucleosome covers 146 nucleotides, the TF binding sites within 73 bp from the nucleosome dyad is supposed to be overlapped or bound by a nucleosome, while the TF binding sites farther than 73 bp from the nearest nucleosome dyad is open.

First, the distance distribution was computed between the TF binding sites and the nearby nucleosomes in the Pro-1 reverse region of the Ly49 genes from the B57BL/6 mouse strain (**Figure 44**). The distance distribution shows that more NF-AT binding sites are at the boundaries and outside of the nucleosome than in the nucleosome covered region. TATA also show the similar distribution as the NF-AT: the binding sites locate at the nucleosome boundaries. On the other hand, AP-1, GATA-3, and Ik-3 demonstrate that many binding sites are within the nucleosome boundaries, which indicates that the TF binding sites of the factors are likely to be covered by nucleosomes. AML-1, C/EBP-beta, c-ETS-1, Lyf-1, MZF-1, and Sp-1 show that many binding sites are within the nucleosome boundaries. However, not all the binding sites are covered by nucleosomes because some binding sites are at the nucleosome boundaries as the peak of the density lies at the nucleosome boundaries. The promoter is open out of all cases, and where the Pro-1 reverse promoter is open, AML-1 binding sites are always found.

Rule	Support	Confidence
Closed -> AML	0.500	1.000
Closed -> No AML	0.000	0.000
Open -> AML	0.500	1.000
Open -> No AML	0.000	0.000
AML -> Open	0.500	0.500
AML -> Closed	0.500	0.500
For, Closed -> AML	**0.375**	**1.000**
Rev, Closed -> AML	**0.125**	**1.000**
For, Open -> AML	**0.000**	**N/A**
Rev, Open -> AML	**0.500**	**1.000**
AML -> For, Closed	0.375	0.375
AML -> Rev, Closed	0.125	0.125
AML -> For, Open	0.000	0.000
AML -> Rev, Open	0.500	0.500
For -> Closed	**0.375**	**1.000**
For -> Open	**0.000**	**0.000**
Rev -> Closed	**0.125**	**0.200**
Rev -> Open	**0.500**	**0.800**
Closed -> For	0.375	0.750
Open -> For	0.000	0.000
Closed -> Rev	0.125	0.250
Open -> Rev	0.500	1.000

Table 6. Associations of the AML-1 and the nucleosome coverage in C57BL/6 Ly49 cluster. The association was measured by support and confidence of the given combination of promoter elements. These notations are used in the table: For, forward Pro-1; Rev, reverse Pro-1, AML, AML-1 binding site; Open, the open status of the binding site; Closed, the nucleosome-bound status of the binding site.

Rule	Support	Confidence
Closed -> AML	0.75	1.00
Closed -> No AML	0.00	0.00
Open -> AML	0.25	1.00
Open -> No AML	0.00	N/A
AML -> Open	**0.25**	**0.25**
AML -> Closed	**0.75**	**0.75**
No AML -> Open	N/A	N/A
No AML -> Closed	N/A	N/A

Table 7. Associations of the AML-1 and the promoter elements in the BALB/c Ly49 cluster. The association was measured by support and confidence of the given combination of promoter elements. The AML-1 binding sites tend to be covered by a nucleosome or at least overlapped by a nucleosome positioning site. These notations are used in the table: For, forward Pro-1; Rev, reverse Pro-1, AML, AML-1 binding site; Open, the open status of the binding site; Closed, the nucleosome-bound status of the binding site.

Rule	Support	Confidence
Closed -> AML	0.500	0.857
Closed -> No AML	0.083	0.143
Open -> AML	0.417	1.000
Open -> No AML	0.000	0.000
AML -> Open	0.417	0.455
AML -> Closed	0.500	0.545
For, Closed -> AML	0.333	0.800
Rev, Closed -> AML	0.167	1.000
For, open -> AML	0.000	N/A
Rev, open -> AML	0.417	1.000
AML -> For, Closed	**0.333**	**0.364**
AML -> Rev, Closed	**0.167**	**0.182**
AML -> For, open	**0.000**	**0.000**
AML -> Rev, open	**0.417**	**0.455**
For -> Closed	**0.417**	**1.000**
For -> Open	0.000	0.000
Rev -> Closed	0.167	0.286
Rev -> Open	**0.417**	**0.714**
Closed -> For	0.417	0.714
Open -> For	0.000	0.000
Closed -> Rev	0.167	0.286
Open -> Rev	0.417	1.000

Table 8. Associations of the AML-1 and the promoter elements in the 129/S6 Ly49 cluster. AML-1 binding sites are more likely to open at the reverse Pro-1 promoter and less likely covered in the forward Pro-1. These notations are used in the table: For, forward Pro-1; Rev, reverse Pro-1, AML, AML-1 binding site; Open, the open status of the binding site; Closed, the nucleosome-bound status of the binding site.

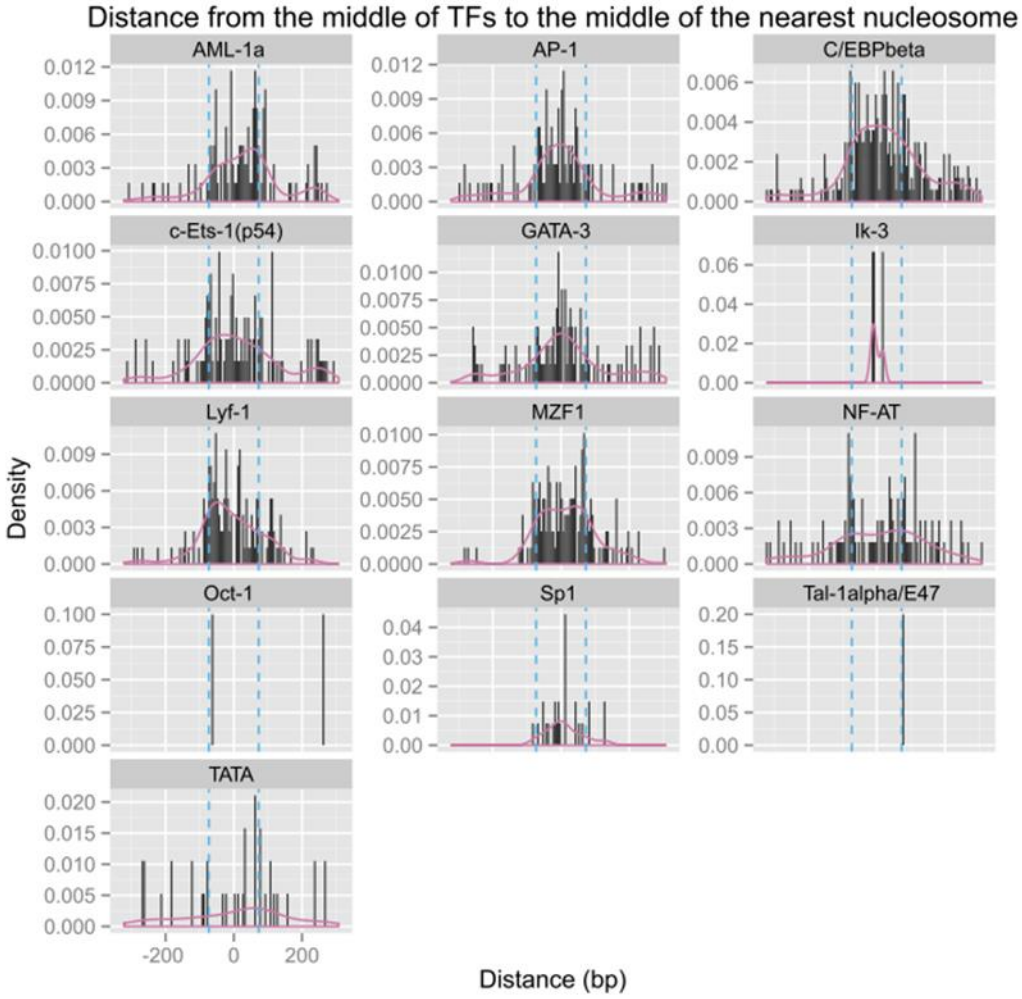

Distance from the middle of TFs to the middle of the nearest nucleosome
AML-1a
AP-1
C/EBPbeta
c-Ets-1(p54)
GATA-3
Ik-3
Lyf-1
MZF1
NF-AT
Oct-1
Sp1
Tal-1alpha/E47
TATA
Density
Distance (bp)
-200
0
200

Figure 44. Proximity of TF binding sites to nucleosomes in Pro-1 reverse promoters. The C57BL/6 nucleosome map as generated in **Figure 38** was compared individually to the transcription factor binding sites in the reverse Pro-1 promoters drawn from the TRANSFAC or JASPAR databases. TATA was included as a control. A histogram (black bar) and density (red line) are shown displaying each factor binding site's distance to the nearest nucleosome. Dashed vertical lines indicate the nucleosome boundary. The more binding sites are within the nucleosome boundaries, the factor's binding sites tend to be covered. Similarly, the more binding sites are outside of the boundaries, the factor's binding sites tend to be open.

To be more specific, the distance distribution to nucleosomes was computed from the TF binding sites in the forward Pro-1 promoters (**Figure 45**). The distance distributions of the factor bindings sites in the forward Pro-1 promoters had a similarity to the distance distribution of factor binding sites in the Pro-1 reverse promoters. NF-AT binding sites, as well as TATA-binding sites, are open in Pro-1 forward promoters as many binding sites are at the nucleosome boundaries. The estimated density distributions of the factors have multiple peaks lying on the nucleosome boundaries. AML-1, Lyf-1, MZF1 binding sites are within the nucleosome boundaries like the ones in the Pro-1 reverse promoters. However, unlike the Pro-1 reverse promoters, where the density of the distance skewed toward the nucleosome boundaries, the distribution in the forward Pro-1 promoters has the peak almost in the middle of the nucleosome boundaries. The density estimates of the factors show that majority of the binding sites are overlapped with the nucleosome positioning sites. Many C/EBP and c-ETS binding sites are found at the nucleosome boundaries as well as within the nucleosome boundaries. The shape of the density plots is skewed toward the boundaries and almost flat. These binding sites may be covered by nucleosomes but still available to the factors because many binding sites are found at the nucleosome boundaries.

The distributions of the binding site in the Pro-2 regions were different from the distributions of the Pro-1 regions (**Figure 46**). NF-AT and TATA-binding sites are consistently outside of the nucleosome boundaries as in the Pro-1 promoters. However, many other TF factors, whose binding sites were overlapped with nucleosomes in Pro-1 regions, are out of the nucleosome-bound sites. Previously nucleosome-bound binding sites, AML-1, AP-1, C/EBPbeta, c-Ets-1, and GATA-3 are outside of the nucleosome boundaries in the Pro-2 regions. The changes of C/EBP-beta and c-Ets-1 are remarkable: the density of the distribution in Pro-2 promoters show twin peaks at the nucleosome

boundaries instead of a single peak near the nucleosome dyad as for the density in the Pro-1 promoters. The shapes of the distance distribution of the Lyf-1 and MZF-1, which were within the nucleosome boundaries in Pro-1 promoters, were shifted toward the nucleosome boundaries. More binding sites are open compared to the binding sites in the Pro-1, even though many binding sites are still located inside the nucleosome boundaries.

To be more objective, the shape of the distribution was assessed quantitatively by measuring kurtosis, and the modality of the distribution was tested statistically with Hartigan's dip test. The kurtosis is the fourth moment of data, measuring the shape of the distribution along with the skewness, which is the third moment of data. The kurtosis, depending on the value, measures the "tailedness" of the distribution. The kurtosis of any univariate normal distribution varies either 3 or zero varying depending on the definition of the kurtosis. The differences between the two definitions are the matter of a constant. In our study, the definition of kurtosis was to set the kurtosis of the normal distribution as zero. The positive kurtosis, which is called leptokurtic, represents a narrower and tighter but longer tailed distribution than a normal distribution, while the negative kurtosis does a wider and often flatter-looking but shorter tailed distribution (**Figure 47**). In our cases, the negative kurtosis indicates that the binding sites are outside of the nucleosome boundaries, while the positive kurtosis indicates that the binding sites are denser within the nucleosome boundaries more likely to be bound by nucleosomes. The modality of the distribution, whether a distribution has one peak (unimodal) or two peaks (bimodal) was tested statistically by Hartigan's dip test. The null hypothesis to test is that the given distribution is unimodal. As the alternative hypothesis is that the distribution is multimodal, the statistic tests the unimodality of the distribution. The p-value less than 0.05 rejects the null hypothesis signifying that the distribution is not unimodal.

The distance distributions were generated from the 17 transcription factors and TATA binding sites with the nearby nucleosomes for the Ly49 cluster. The spatial relationship between TF and nucleosome positioning are examined (**Figure 48**). The *p*-value of the dip test and the measured kurtosis were presented with the plot. Three distinct patterns of TF-nucleosome spacing have been observed based on visual assessment and the calculated kurtosis by measuring the modality of the distribution quantitatively. In our results, distributions with kurtosis less than 0 indicate that the TF binding sites are located farther from the nucleosome dyad. Some factors – namely, NF-AT and TATA – showed a bimodal distribution with p-values of both less than 0.05 and a kurtosis less than -0.5. The bimodal peaks exist at or near the nucleosome boundary. The bimodal peaks suggest that these binding sites preferentially avoid the nucleosomes, and are unlikely to be hindered by neither nucleosomes nor sensitive to nucleosome interference. Other factors, AML-1, AP-1, Lyf-1, Sp-1, and MZF-1, show a tightened, unimodal distribution with positive kurtosis (> 0.5). The p-values of them after the dip test are less than 0.05 except for Lyf-1, proving the tighter than normal distribution of the TF binding sites around nucleosomes. The tightened distribution suggests that most factor binding sites are preferentially covered by the nucleosome. The rest of the factors analysed were found to be normally distributed around the nucleosome dyad with p-values greater than 0.05 or the kurtosis between -0.5 and 0.5, showing no preference or sensitivity to nucleosome coverage.

IDENTIFICATION OF THE TRANSCRIPTION FACTOR BINDING SITES DISPLAYING THE 10 BP PERIODICITY

Five transcription factor binding sites were identified, which were predicted to be covered by nucleosomes in Ly49 gene family. I hypothesised that those factors were the most sensitive to nucleosome interference due to the coverage. However, as mentioned

previously, the binding sites that display a strong 10 bp periodicity in the distance to the nucleosome dyad are able to exist on the same orientation on the nucleosome, and therefore be accessible even when they are covered by nucleosome (Ioshikhes et al., 1999). Those nucleosome-covered binding sites specifically lacking 10 bp periodicity will be disrupted by the turning of the DNA in the nucleosome. Due to the helical nature of DNA, it is possible for TF binding sites to be oriented to face outward even when wrapped around a histone core in a nucleosome, and therefore be accessible (**Figure 49**). This effect was measured in the Ly49 gene family by analysing the distance periodicity of each TF binding site from the nearby nucleosome positions following the selection of the TF binding sites from Pro-1 regions, nucleosome-bound regions, or nucleosome-bound regions in Pro-1. The distance distribution was analysed by Fourier transform to find the period, the interval of the distances in terms of base pair. The distance periodicities of the TF binding sites located in the Pro-1 regions have the 10 bp periodicity except for Ly49J and D (**Table 9**). Interestingly, the expression of the Ly49 genes is relatively low in NK cells. The proportion of NK cells expressing Ly49J and D is 5-8% (McQueen et al., 2001) and 50% (Ortaldo et al., 1999), respectively. It suggests that the 10 bp periodicity of TF binding sites may play a role in transcriptional regulation.

The leave-one-out analysis was used to take a quantitative and unbiased approach. In this analysis, the distribution of all 17 factors from the Pro-1 regions pooled together and then the periodicity was determined by Fourier transform, which forms the baseline periodicity. And then each factor was individually removed from the group of 17 factors to measure the effect of its removal on the overall periodicity, i.e., a factor displaying 10 bp periodicity will reduce the 10 bp periodicity in the periodogram when removed from the pooled sample set, while a factor lacking the 10 bp periodicity will increase the overall 10 bp periodicity when removed.

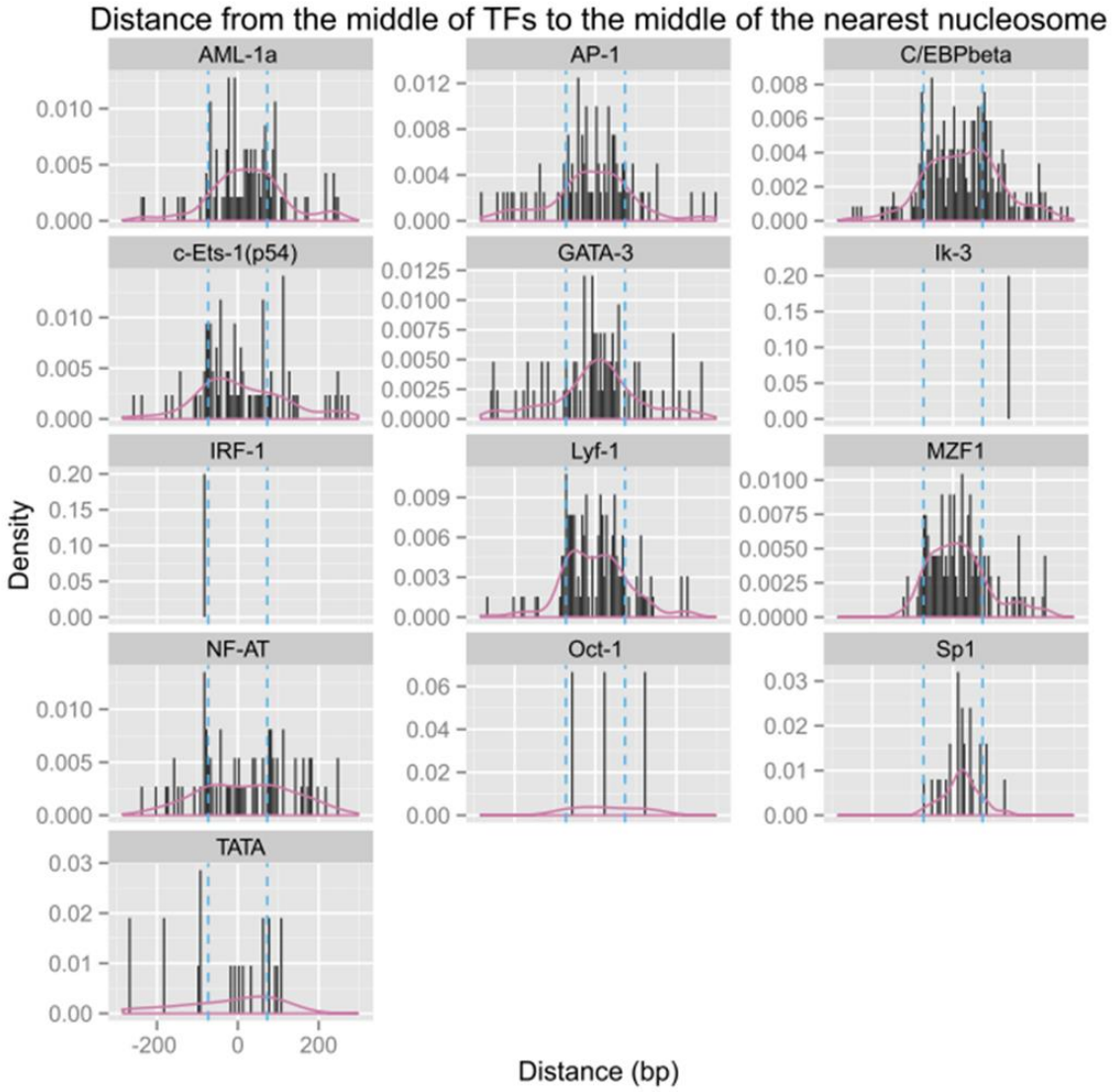

Distance from the middle of TFs to the middle of the nearest nucleosome
AML-1a
AP-1
C/EBPbeta
c-Ets-1(p54)
GATA-3
Ik-3
IRF-1
Lyf-1
MZF1
NF-AT
Oct-1
Sp1
TATA
Density
Distance (bp)
-200
0
200

Figure 45. Proximity of TF binding sites to nucleosomes in Pro-1 forward promoters. The C57BL/6 nucleosome map as generated in **Figure 38** was compared individually to the transcription factor binding sites in the forward Pro-1 promoters drawn from the TRANSFAC or JASPAR databases. TATA was included as a control. A histogram (black bar) and density (red line) are shown displaying each factor binding site's distance to the nearest nucleosome. Dashed vertical lines indicate the nucleosome boundary. The more binding sites are within the nucleosome boundaries, the factor's binding sites tend to be covered. Similarly, the more binding sites are outside of the boundaries, the factor's binding sites tend to be open.

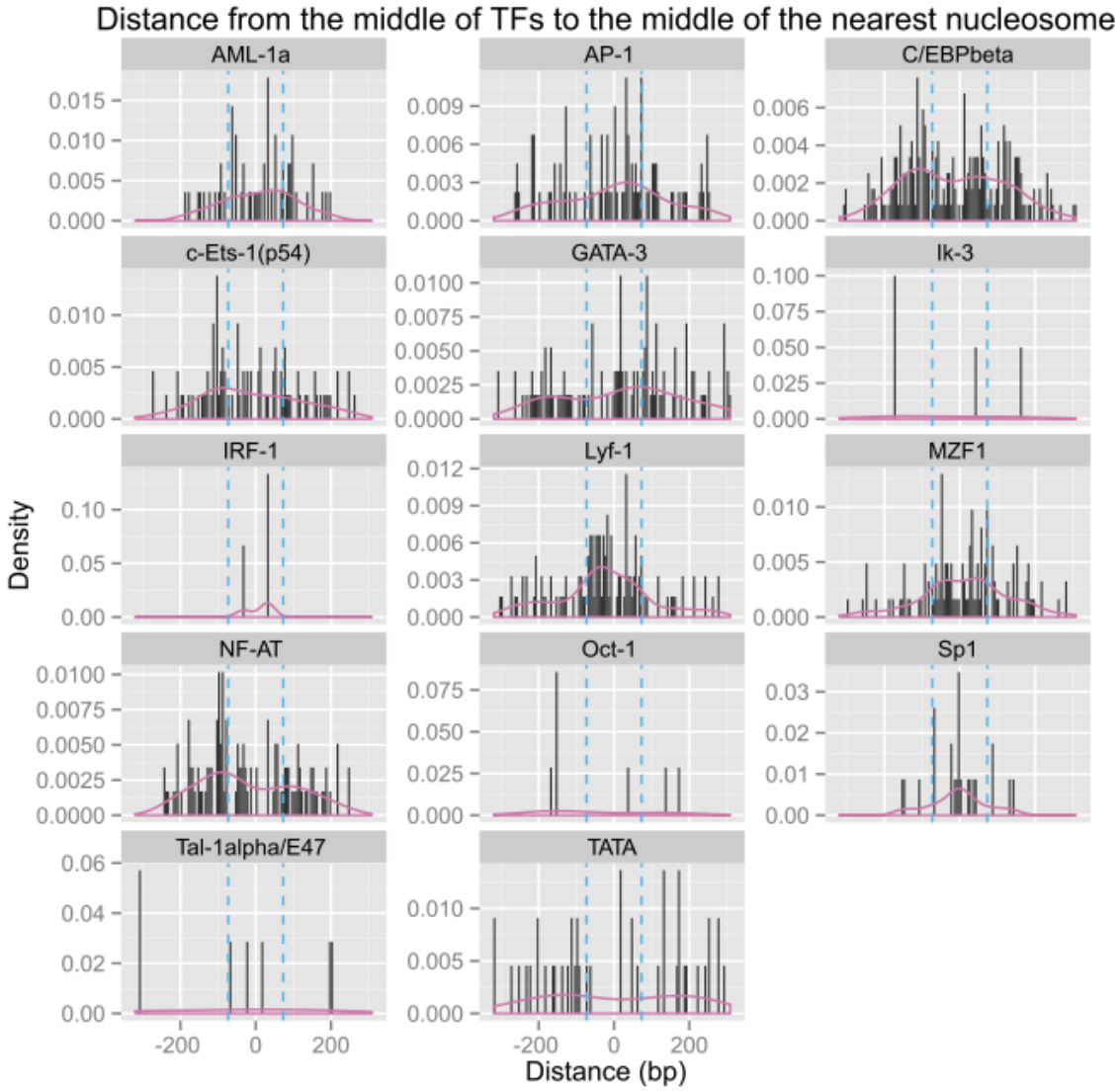

Distance from the middle of TFs to the middle of the nearest nucleosome
AML-1a
AP-1
C/EBPbeta
c-Ets-1(p54)
GATA-3
Ik-3
IRF-1
Lyf-1
MZF1
NF-AT
Oct-1
Sp1
Tal-1alpha/E47
TATA
Density
Distance (bp)
-200
0
200

Figure 46. Proximity of TF binding sites to nucleosomes in Pro-2 promoters. The C57BL/6 nucleosome map as generated in **Figure 38** was compared individually to the transcription factor binding sites in the Pro-2 promoters drawn from the TRANSFAC or JASPAR databases. TATA was included as a control. A histogram (black bar) and density (red line) are shown displaying each factor binding site's distance to the nearest nucleosome. Dashed vertical lines indicate the nucleosome boundary. The more binding sites are within the nucleosome boundaries, the factor's binding sites tend to be covered. Similarly, the more binding sites are outside of the boundaries, the factor's binding sites tend to be open.

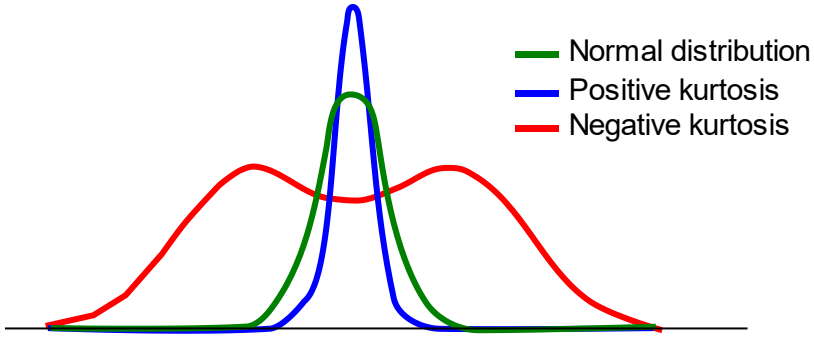

Normal distribution
Positive kurtosis
Negative kurtosis

Figure 47. Kurtosis and the shape of the distribution. The kurtosis, the fourth moment of a distribution measures the "tailedness" of the distribution. The positive kurtosis represents the tighter shape of the distribution with longer tails. On the other hand, the negative kurtosis indicates the wider shape of distribution with shorter tails. The kurtosis measures the tightness of the distribution of the distance between the TF binding sites and the nucleosome dyad.

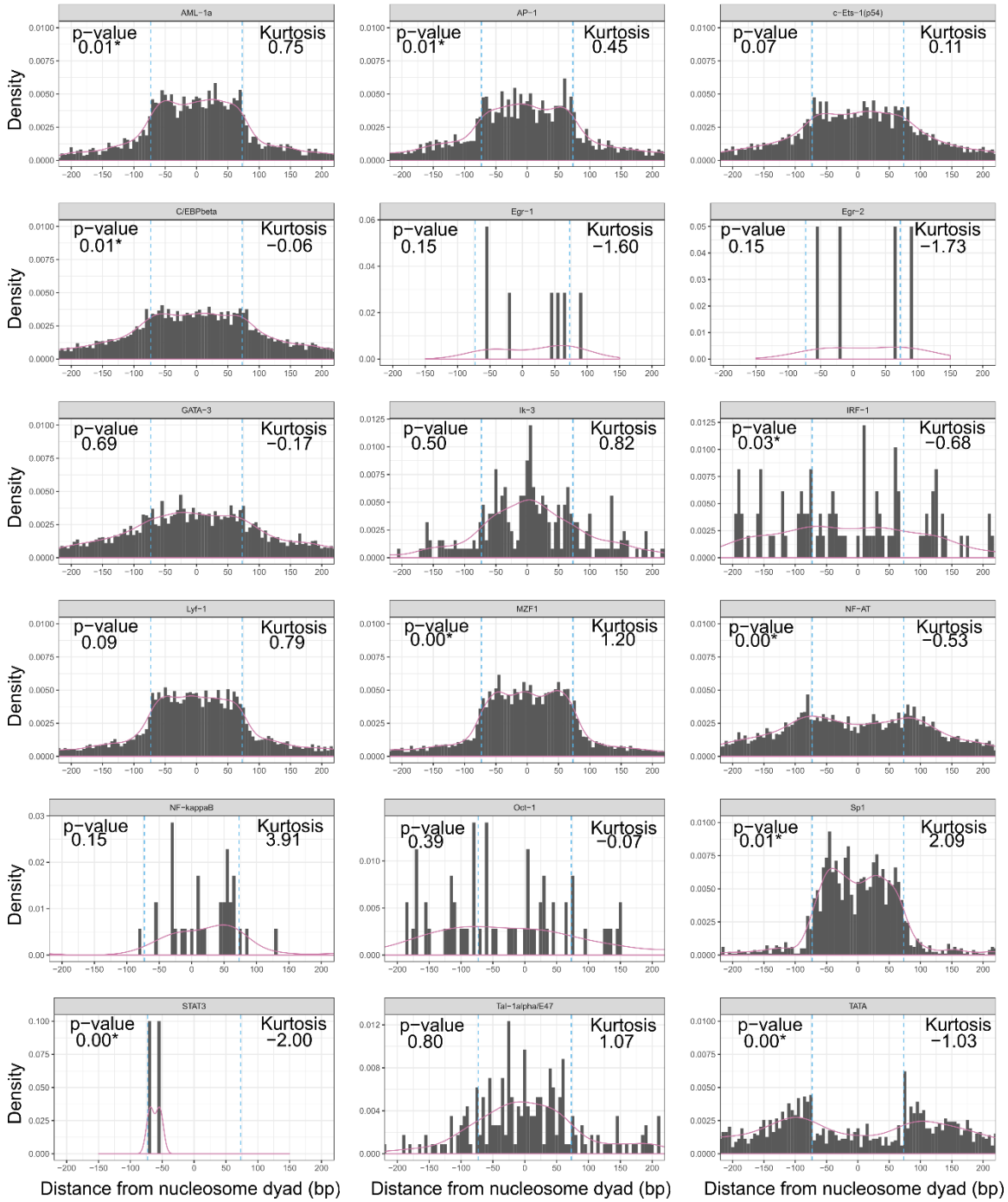

AML−1a
p−value 0.01*
Kurtosis 0.75
AP−1
p−value 0.01*
Kurtosis 0.45
c−Ets−1(p54)
p−value 0.07
Kurtosis 0.11
C/EBPbeta
p−value 0.01*
Kurtosis −0.06
Egr−1
p−value 0.15
Kurtosis −1.60
Egr−2
p−value 0.15
Kurtosis −1.73
GATA−3
p−value 0.69
Kurtosis −0.17
Ik−3
p−value 0.50
Kurtosis 0.82
IRF−1
p−value 0.03*
Kurtosis −0.68
Lyf−1
p−value 0.09
Kurtosis 0.79
MZF1
p−value 0.00*
Kurtosis 1.20
NF−AT
p−value 0.00*
Kurtosis −0.53
NF−kappaB
p−value 0.15
Kurtosis 3.91
Oct−1
p−value 0.39
Kurtosis −0.07
Sp1
p−value 0.01*
Kurtosis 2.09
STAT3
p−value 0.00*
Kurtosis −2.00
Tal−1alpha/E47
p−value 0.80
Kurtosis 1.07
TATA
p−value 0.00*
Kurtosis −1.03
Density
Distance from nucleosome dyad (bp)

Figure 48. Statistical tests of the proximity of TF binding sites to nucleosomes. The transcription factor binding sites plus TATA box sites were compared with the nucleosome positions. The distances between the centre of the binding site and the nearest nucleosome centre were presented in the histogram as density. The nucleosome coverage over the binding sites was quantitatively assessed by the kurtosis and diptest. Blue dotted lines denote the nucleosome boundaries.

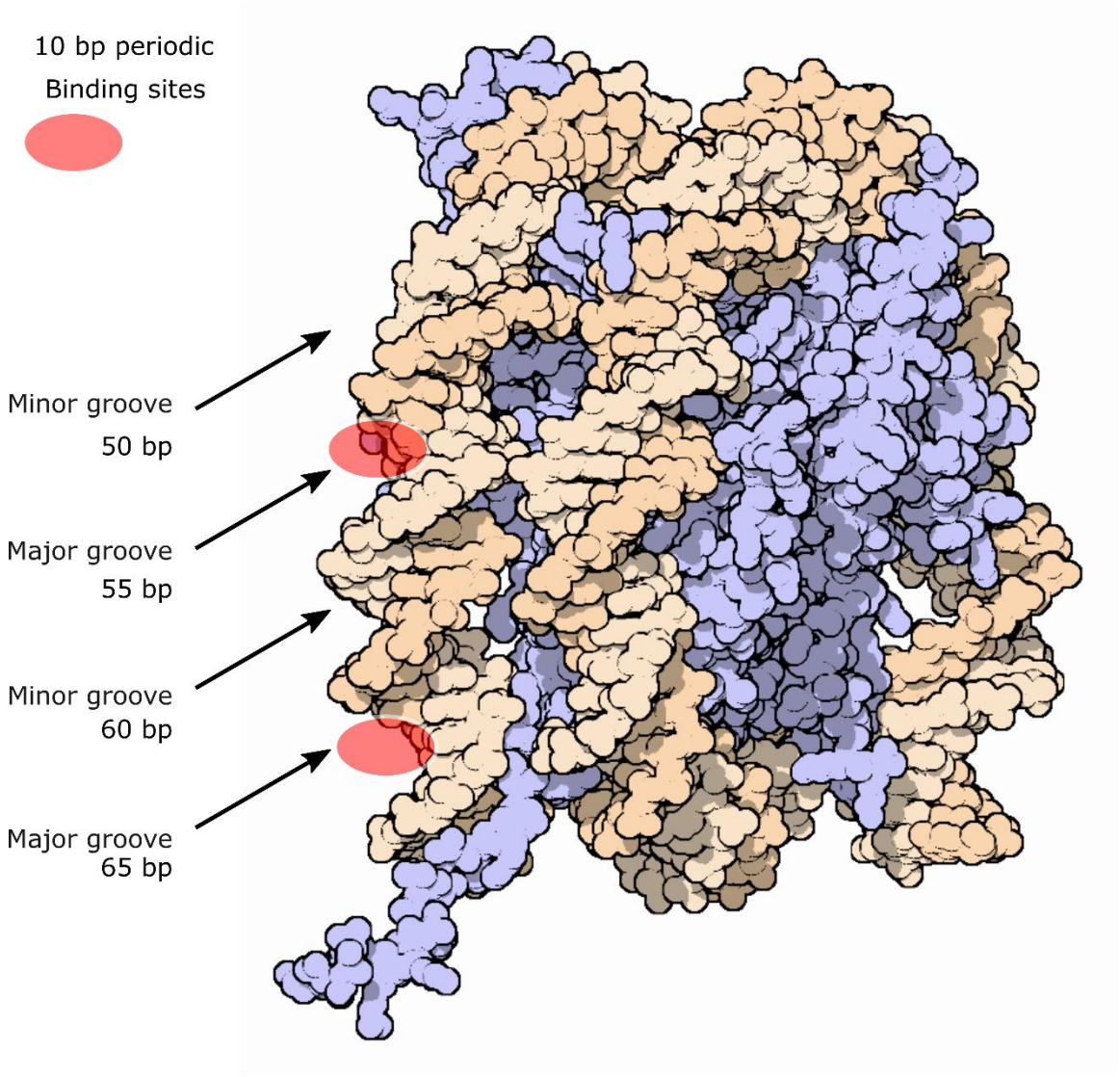

10 bp periodic
Binding sites
Minor groove
50 bp
Major groove
55 bp
Minor groove
60 bp
Major groove
65 bp

Figure 49. Schematic diagram the 10 bp periodic binding sites on a nucleosome. The transcription factor binding sites that are placed at 10 bp periodicity are oriented in the same direction, i.e., facing outward from the nucleosome surface. They can be still accessible by transcription factors even though packed into a nucleosome.

Gene	Dominant period	Other periods	Expression level	
A	10 bp	15, 20 bp	High	(Ortaldo et al., 1999)
C	10 bp	8, 20 bp	High	(Brennan et al., 1996)
J	8 bp	20 bp	Low (5 – 8%)	(McQueen et al., 2001)
G	12 bp	20 bp	High	(Ortaldo et al., 1999)
I	10 bp	8, 20 bp	High	(Brennan et al., 1996)
E	8 bp	20 bp	Low	(Gays et al., 2005)

Table 9. Distance periodicities of binding sites and the Ly49 expression. The periodicity of the TF binding sites to nearby nucleosomes in the Pro-1 region of each Ly49 gene and the expression level of the Ly49 genes are presented. The high expressed Ly49 genes (A, C, G, and I) have TF binding sites with the 10 bp periodicity to nearby nucleosomes. Low expressed Ly49 genes (J and E) do not have 10 bp periodicity. The distances were calculated between the middle of the TF binding sites and the nearest nucleosome dyad located in the Pro-1 promoters of the Ly49 gene family. The periodicity was analysed by Fourier transform.

AML-1 is the one which lacks the 10 bp periodicity in three Ly49 genes, Ly49 C, I, and J, as its removal from the pooled factors increases the signal at 10 bp. Removing Lyf-1 and MZF-1, on the other hand, had the reverse effects on the pooled periodicity: the periodicity of 10 bp decreased in Ly49A, C, G, and I (**Table 10**). Based on these results, AML-1 is the sensitive transcription factor to the nucleosome coverage thanks to the lack of 10 bp periodicity, as its removal from the pooled factors dramatically increases the signal at 10 bp in three Ly49 genes among the tested genes.

The same analysis was repeated with the TF factors within the nucleosome bound regions. The increase of the 10 bp periodicity after removal of the AML-1 from the pools samples, thus the lack of 10 bp periodicity of AML-1 binding sites, were observed again (**Figure 50**). The removal of Lyf-1 and MZF-1 decreased the 10 bp periodicity, implicating the 10 bp periodicity for the two factors. Collectively, these results indicate that, of the factors analysed, AML-1 is the one most sensitive to the surrounding nucleosomes, as it is the only factor that displays both a tendency to nucleosome coverage (**Figure 48**) and a lack of the specific 10 bp periodicity (**Figure 50**).

To determine whether AML-1's pattern of nucleosome co-occupancy and a lack of 10 bp periodicity are unique disposition only to the Ly49 gene family, these two analyses were extended to the entirety of mouse chromosome 6 (**Figure 51**). Surprisingly, many of the other transcription factors displayed much more pronounced coverage and an extreme loss of 10 bp periodicity, while AML-1 retained its nucleosome preference and displayed a very marginal lack of 10 bp periodicity. This marginal lack of 10 bp periodicity suggests that the arrangement of the nucleosomes with regard to AML-1 in Ly49 is an example of gene regulation mechanism via nucleosomal interference with one or more key transcription factor binding sites.

LY49 EXPRESSION STATE AND *IN VITRO* NUCLEOSOME MAPS

The Ly49 gene family exists as a cluster in the same region of chromosome 6. The Ly49 gene family have closely related promoter sequences and similar transcriptional regulation in terms of transcription factor binding sites. Nevertheless, individual Ly49 genes display remarkably different expression patterns. Indeed, RMA, a mouse NK-T cell line with a C57BL/6-derived Ly49 gene cluster, expresses the inhibitory Ly49A receptor, but no other Ly49 receptors with corresponding antibodies (**Figure 52**). The cell line was used for verification to determine whether the promoter of the actively expressing Ly49A would have a nucleosome landscape more divergent from the sequence-based predictions than other inactive Ly49 promoters. The nucleosome position map for RMA was generated using MNase-Seq data. The list of nucleosomes predicted to be present and confirmed by the MNase-Seq map ('true-positive') and a list of nucleosome predicted to be present but found absent on the MNase-Seq map ('false-positive') were presented (**Figure 53**). A heat map displays the prediction accuracy of each Ly49 promoter region at all the sites for each indicated transcription factor and across the whole region of interest as the background (**Figure 54A**). In this analysis, the selected promoter region includes the corresponding promoters 1, 2, and 3 of Ly49 gene family as well as any distal enhancer elements by extending the analysis approximately 8000 bp upstream of Pro-1. The relatively high degree of convergence with the prediction was represented by blue and the low degree of convergence by red.

Next, we asked whether specific transcription factor binding sites, for each promoter individually, diverge significantly in nucleosome occupancy, compared to the rest of the transcription factors at the promoter. A chi-square analysis was performed on the divergence of the transcription factor binding sites' nucleosome status. The divergence of the nucleosome status of each transcription factor from the prediction was

	A	C	E	G	I	J
AML-1	D	I	-	-	I	I
AP-1	-	D	-	I	-	I
C-EBPβ	-	-	-	-	-	-
c-ETS-1	-	-	-	-	-	-
Egr-1	-	-	-	-	-	-
Egr-2	-	-	-	-	-	-
GATA-3	-	I	-	-	-	-
IK-3	-	-	-	-	-	-
IRF-1	-	-	-	-	-	-
Lyf-1	D	D	-	D	D	-
MZF-1	D	D	-	D	D	-
NF-AT	-	-	-	D	D	-
NF-κB	-	-	-	-	-	-
Oct-1	-	-	-	-	-	-
Sp1	D	-	-	-	-	-
STAT3	-	-	-	-	-	-
Tal-1α	-	-	-	-	-	-

Table 10. TF binding sites with 10 bp periodicities in Pro-1. The periodicity was examined in the distances between the TF binding sites and the nucleosome dyads in the Pro-1 promoters. One factor was taken out of the calculation to find the effects of the factor on the periodicity. The table summarises the changes of the 10 bp periodicity after taking one factor out: no changes (-), increase of the 10 bp periodicity (I), a decrease of the 10 bp periodicity (D). Note that the reduction of the 10 bp periodicity indicates that the removed factor has the 10 bp periodicity. Removing Lyf-1 and MZF-1 binding sites reduced the 10 bp periodicity, which suggests the binding sites have 10 bp periodic distances from the nucleosomes. Removing AML-1 and AP-1 binding sites increased the 10 bp periodicity suggesting the organisation of the AML-1 and AP-1 binding sites differ from Lyf-1 and MZF-1. NF-AT and c-ETS-1 show the 10 bp period in Ly49G and Ly49 I. The binding sites have no 10 bp period in Ly49E and J, which have low expression level.

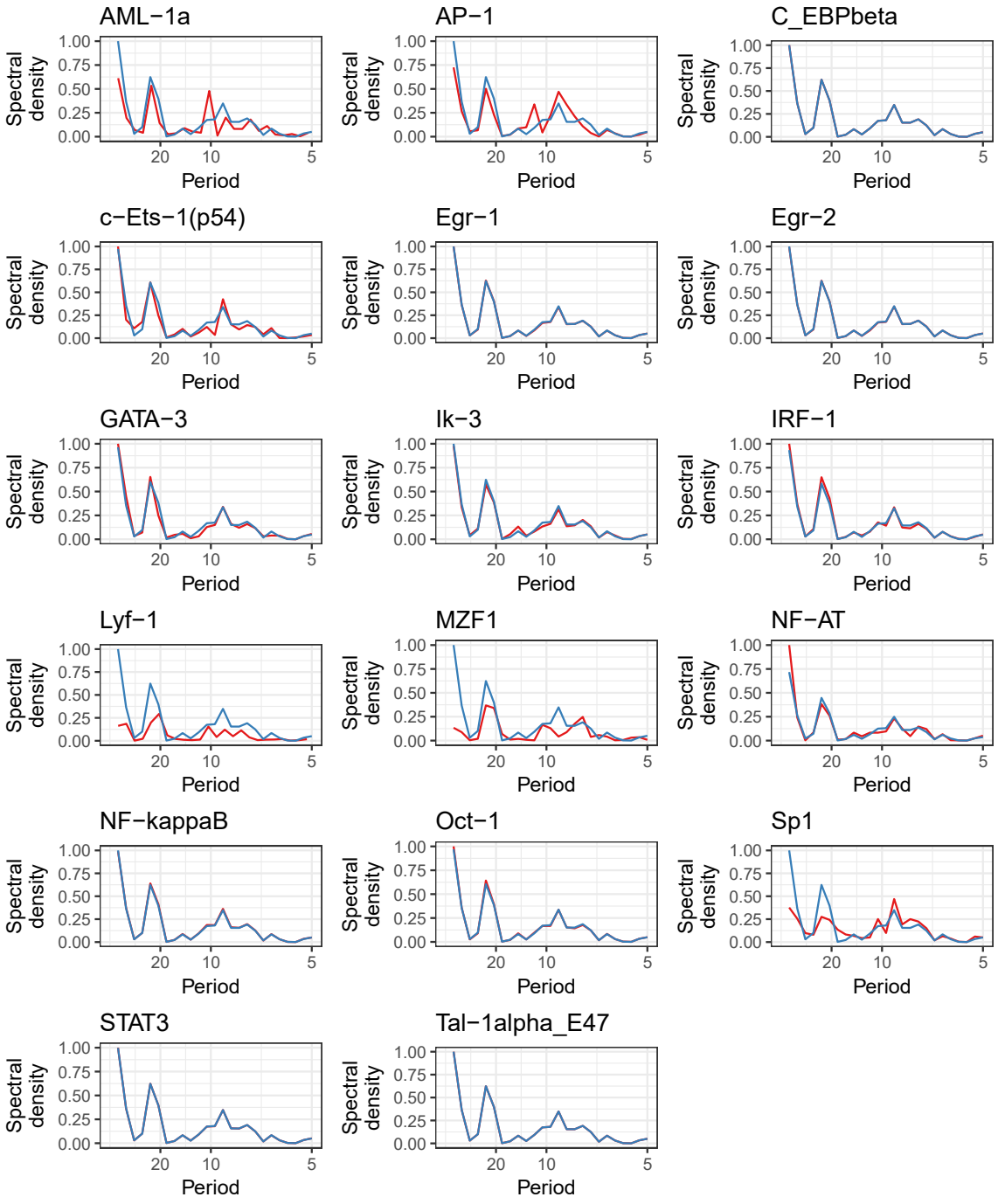
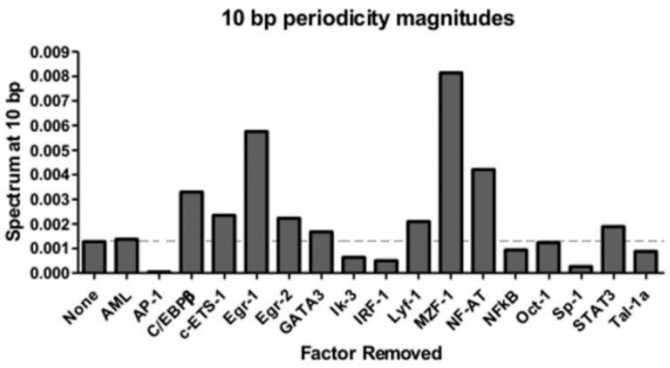

AML−1a
AP−1
C_EBPbeta
c−Ets−1(p54)
Egr−1
Egr−2
GATA−3
Ik−3
IRF−1
Lyf−1
MZF1
NF−AT
NF−kappaB
Oct−1
Sp1
STAT3
Tal−1alpha_E47
Spectral density
1.00
0.75
0.50
0.25
0.00
20
10
5
Period

Figure 50. Distance periodicities of the nucleosome-covered TF binding sites in the Ly49 cluster. AML-1 lacks the 10 bp periodicity within the nucleosome-bound Ly49 regions. The distance periodicity from all 17 transcription factors (blue line) in the nucleosome covered Ly49 cluster was compared with distance periodicity from the leave-out-out distribution (red line). It should be noted that the increase of 10 bp periodicity in the leave-one-out distribution upon the removal of a factor indicates lack of 10 bp periodicity for the factor.

A


10 bp periodicity magnitudes
Spectrum at 10 bp
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
None
AML
AP-1
C/EBPβ
c-ETS-1
Egr-1
Egr-2
GATA3
Ik-3
IRF-1
Lyf-1
MZF-1
NF-AT
NFkB
Oct-1
Sp-1
STAT3
Tal-1a
Factor Removed

B

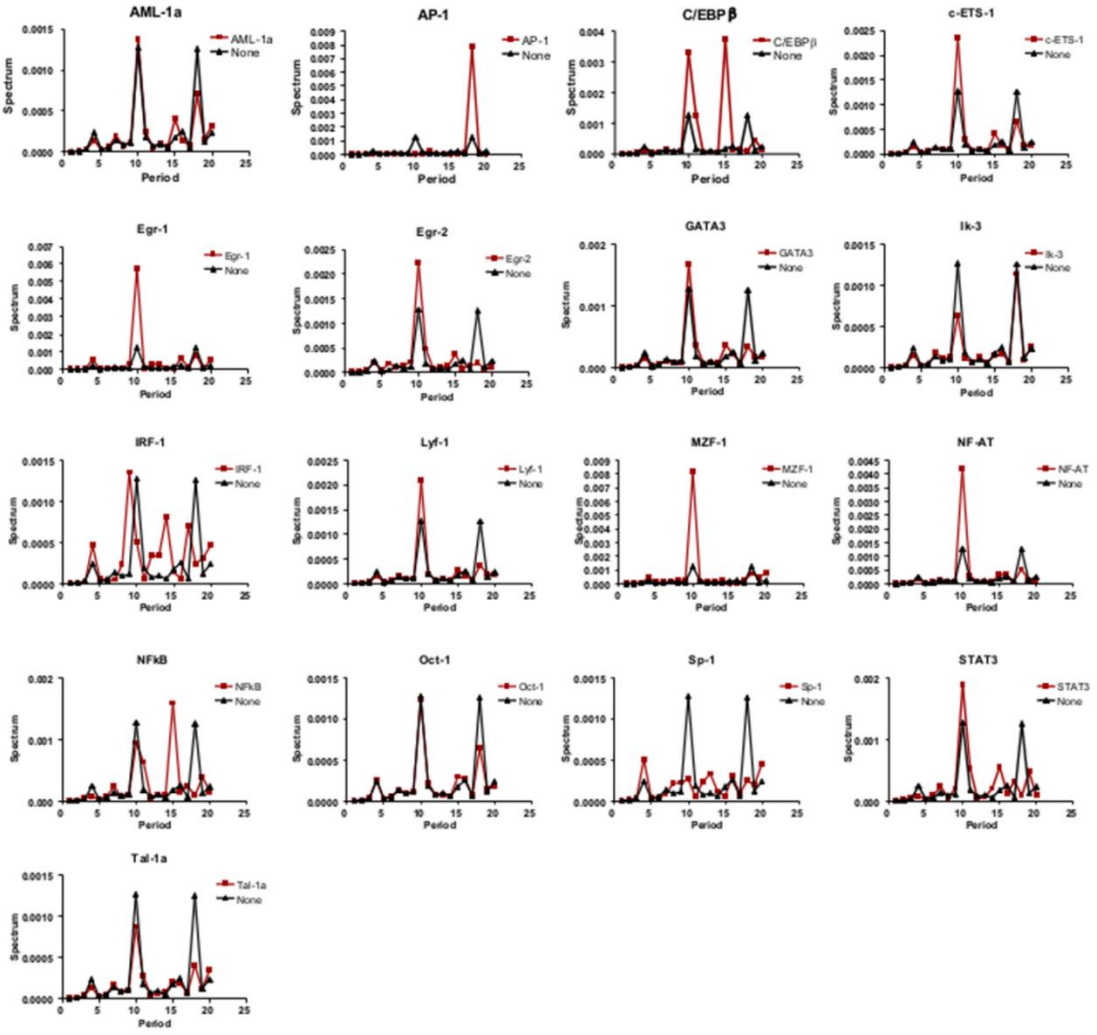

AML-1a
AML-1a
None
AP-1
AP-1
None
C/EBPβ
C/EBPβ
None
c-ETS-1
c-ETS-1
None
Egr-1
Egr-1
None
Egr-2
Egr-2
None
GATA3
GATA3
None
Ik-3
Ik-3
None
IRF-1
IRF-1
None
Lyf-1
Lyf-1
None
MZF-1
MZF-1
None
NF-AT
NF-AT
None
NFkB
NFkB
None
Oct-1
Oct-1
None
Sp-1
Sp-1
None
STAT3
STAT3
None
Tal-1a
Tal-1a
None
Spectrum
Period

Figure 51. Distance periodicities of all nucleosome-bound TF binding sites on chromosome 6. The 10 bp distance periodicity was lost in transcription factors in chromosome 6 beyond the Ly49 gene cluster. It should be noted that the increase of 10 bp periodicity in the leave-one-out distribution upon the removal of a factor indicates lack of 10 bp periodicity for the factor. (A) The bar graph summarises the effect of removing each factor individual on the 10 bp periodicity. (B) Graphs show the periodogram of the pooled data (black) and the effect of removing the indicated factor from the pooled data (red).

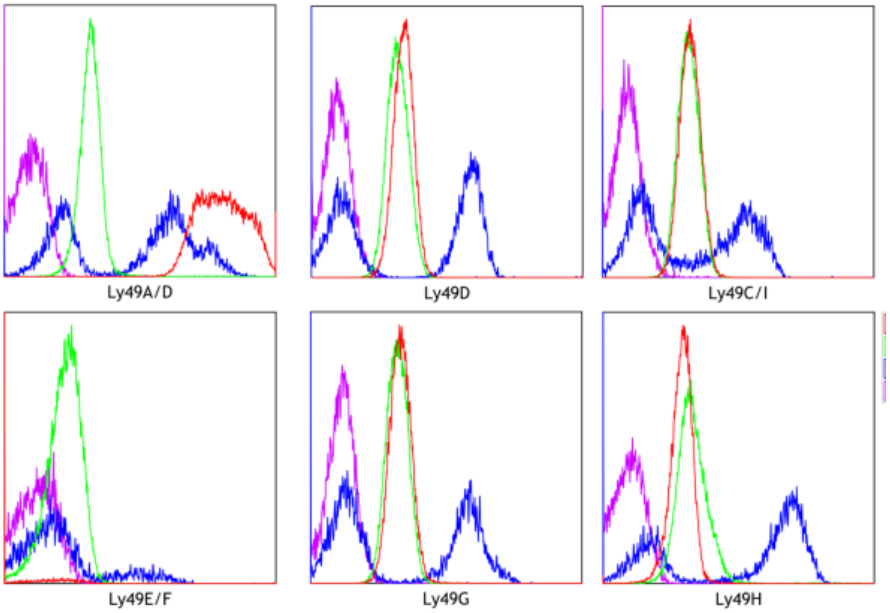

Ly49A/D
Ly49D
Ly49C/I
Ly49E/F
Ly49G
Ly49H

Figure 52. Ly49A expression on RMA cells. RMA cells or isolated mouse splenocytes were analysed by flow cytometry after staining with antibodies against NK1.1, TCRb, and the indicated Ly49 receptors. RMA (red line) stains positive only for Ly49A/D when compared to the isotype (green). It should be noted that positive for Ly49A/D and negative for Ly49D indicates the single expression of Ly49A in RMA cells. The normal NK cells isolated from C57BL/6 splenocytes as a positive (blue) and negative (isotype, violet) control.

mm10
5 kb
130,387,000
130,387,500
130,388,000
130,388,500
130,389,000
130,389,500
130,390,000
130,390,500
130,391,000
130,391,500
130,392,000
130,392,500
130,393,000
Predicted Nucleosomes - Markov
Nucleosomes from RNA genome by MNase-Seq
Confirmed predicted nucleosomes RNA
Absent predicted nucleosomes RNA
AML-1

Figure 53. The deviation of the nucleosome positioning from the prediction in RMA. Example analysis of predicted versus actual nucleosome positioning sites and their association with predicted TF binding sites. Taking a region of interest 15,000 bp upstream and 1,000 bp downstream of the Ly49A transcription start site, predicted nucleosomes (black) were compared to the results of the MNase-Seq experiment (blue) using the UCSC table browser to find intersections of predictions with nucleosome-bound regions (true predictions, green) or nucleosome-free regions (false predictions, red).

	AML-1	AP-1	C/EBPβ	c-ETS-1	GATA3	Lyf-1	MZF-1	NF-AT	Sp1	Whole
Ly49A	26%	31%	33%	24%	43%	27%	27%	35%	22%	39%
Ly49C	50%	56%	42%	41%	58%	37%	36%	35%	38%	47%
Ly49J	49%	37%	45%	43%	38%	44%	50%	55%	53%	46%
Ly49G	39%	39%	36%	27%	28%	43%	40%	38%	42%	44%
Ly49I	37%	42%	41%	48%	39%	38%	44%	48%	36%	42%
Ly49H	30%	41%	41%	44%	45%	39%	39%	38%	42%	42%
Ly49K	38%	47%	38%	25%	37%	38%	30%	34%	27%	44%
Ly49D	34%	25%	32%	36%	35%	31%	32%	34%	27%	40%
Ly49F	54%	49%	45%	43%	38%	44%	46%	45%	36%	48%
Ly49E	34%	34%	34%	40%	41%	34%	34%	39%	43%	42%
Ly49Q	32%	29%	41%	46%	36%	33%	33%	46%	41%	42%

Figure 54. Nucleosome depletion from the predicted positions *in vivo*. Nucleosome positions on Ly49A expressing cell line RMA was determined by MNase-Seq. The degree of true positive nucleosomes was calculated by comparing the MNase-Seq determined positions to the predicted positions per Ly49 gene across the transcription factors as the background value (Whole). The relative degree of accuracy was calculated by repeating the comparison for each transcription factor.

tested against the overall predictive convergence across the transcription factor for each Ly49 promoter (the 'Whole' column in **Figure 54**) as the expected divergence, and the predictive convergence for each transcription factor binding as the observed divergence. Two Ly49 promoters were significantly divergent from the predicted nucleosome occupancy at the tested transcription factor binding sites: the solely expressed Ly49A and the pseudogene Ly49K (**Figure 55**). All other Ly49 genes were not expressed in RMA and not found to differ significantly from their predicted nucleosome landscapes. The findings suggest that the divergence from the predicted nucleosome configuration is related to the suppressed expression. As indicated by the heat map, AML-1, c-ETS-1, Lyf-1, and MZF-1 were the factors most responsible for the significant divergence in Ly49A (**Figure 55**).

These divergent factors, except for c-ETS-1 that are marginal in nucleosome coverage with the dip test p-value of 0.07, were all preferentially covered by nucleosomes according to the previous distance analysis (**Figure 48**). The nucleosome divergence among the factors was related possibly for one factor driving the depletion of the nucleosome in another factor. Logistic regression analysis on all pairs between the four factors determined whether this could be the case. The logic of this analysis is that if condition A is good to predict condition B, but the opposite is not true, then the condition A is likely to be the cause of the condition B but not vice versa. The nucleosome state, true or false positive, was used as a predictor to determine the nucleosome state of the nearest binding site or the reporter variable. The nucleosome state of the reporter variable was determined whether the two sites were within 147 bp of each other or not after controlling for the distance between the two sites.

Predicting the nucleosome state of the nearest c-ETS-1 site with the nucleosome state of an AML-1 site was neither better nor worse than predicting the nucleosome state

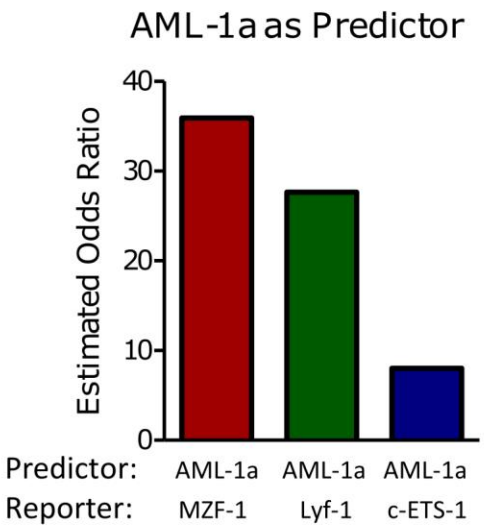
of the nearest AML-1 site with the nucleosome state of a c-ETS-1 site, suggesting that neither of these sites impacts the other unequally (**Figure 56A-B**). The knowledge of the nucleosome coverage state of an AML-1 site, however, was a stronger predictor of the nucleosome coverage state of the nearest MZF-1 or Lyf-1 binding site than either of these sites was for predicting the state of AML-1, suggesting that AML-1-based divergence may be preceding the divergence of Lyf-1 and MZF-1 influencing the nucleosome state of Lyf-1 and MZF-1 (**Figure 56A-B**). AML-1, Lyf-1, and MZF-1 were preferentially covered factors in the nucleosome-bound-regions except for c-ETS-1.

As AML-1 is well-known as a required factor for Ly49 expression in mouse NK cells (Saleh et al., 2004), a chromatin immunoprecipitation was performed to validate the presence of AML-1 at the Ly49A promoter in RMA cells (**Figure 56C**). Immuno precipitation of the DNA sequences with anti-AML-1 resulted in a 10-fold enrichment of the Ly49A promoter sequences against precipitation with an isotype. Conversely, the AML-1 binding confined to the promoter and did not bind to the enhancer region of the Ly49A promoter—a 6 kb upstream of the promoter—was not enriched following anti-AML-1 enrichment. The immediately upstream of exon 1 (corresponding to Pro-2) and upstream of exon 2 (corresponding to Pro-3) of Ly49 genes were analysed. Due to the smaller number of identified genetic elements of nucleosomes and AML-1 binding sites, a statistical analysis was not feasible. However, for each of Pro-2 and Pro-3, the AML-1 binding site was identified, and the probable coverage at the binding sites was determined (**Table 11**). Surprisingly, Ly49A displays similar or even more convergence at AML-1 sites in Pro-2 and Pro-3 than many of the non-expressed Ly49 genes, indicating that the nucleosome divergence for AML-1 in expressed genes is restricted to Pro-1 and the enhancer elements. In summary, the nucleosome divergence was significant on Pro-1.

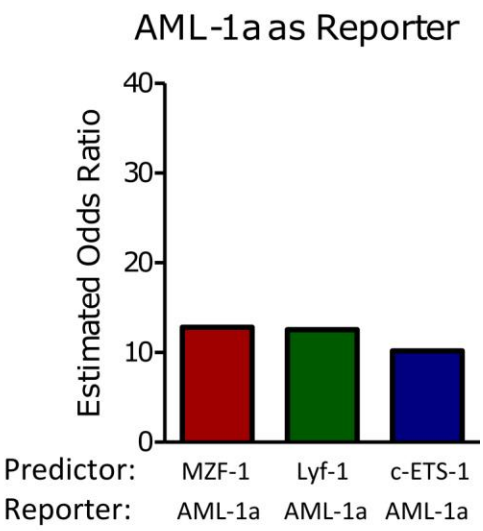
	AML-1	AP-1	C/EBPβ	c-ETS-1	GATA3	Lyf-1	MZF-1	NF-AT	Sp1	Total	P-value
Ly49A	3.3	0.7	1.4	3.3	0.3	4	3.6	0.2	1.278	18	*
Ly49C	0.1	0.8	0.8	0.3	1.3	2.4	2.3	1.2	0.003	9.4	ns
Ly49J	0.2	0.9	0	0.1	0.6	0.1	0.3	0.9	0.813	4	ns
Ly49G	0.4	0.4	1.9	3.4	3.6	0	0.2	0.4	0.05	10.4	ns
Ly49I	0.4	0	0	0.5	0.1	0.3	0.2	0.4	0.034	2	ns
Ly49H	3.6	0	0.1	0	0.1	0.3	0.3	0.3	0.077	4.8	ns
Ly49K	0.8	0.2	1.2	4.8	0.6	1.2	6	1.2	1.134	17	*
Ly49D	1	3.3	3.2	0.3	0.4	3.1	2	0.6	1.512	15.3	ns
Ly49F	0.5	0	0.3	0.2	1.1	0.3	0.1	0.1	0.035	2.6	ns
Ly49E	1.3	1.5	2.8	0.1	0	2.4	2	0.1	0.147	10.3	ns
Ly49Q	1.9	2.4	0	0.3	0.6	2.7	2.2	0.3	0.049	10.4	ns

Figure 55. Chi-square tests of the nucleosome depletion from the predicted positions. The statistical significance of the nucleosome depletion was tested by chi-square tests. The overall nucleosome accuracy—the 'Whole' column of **Figure 54** was set up as the expected value, and the relative accuracy of each Ly49 was set up as the observed value. The nucleosome depletion is significant in Ly49A and Ly49K (A p-value < 0.05 was considered significant).

A


AML-1a as Predictor
Estimated Odds Ratio
40
30
20
10
0
Predictor: AML-1a AML-1a AML-1a
Reporter: MZF-1 Lyf-1 c-ETS-1

B


AML-1a as Reporter
Estimated Odds Ratio
40
30
20
10
0
Predictor: MZF-1 Lyf-1 c-ETS-1
Reporter: AML-1a AML-1a AML-1a

C

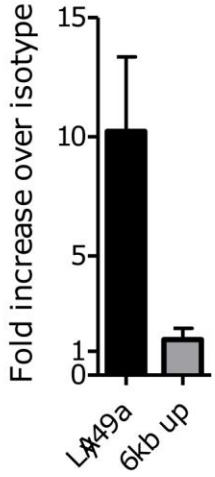

Fold increase over isotype
15
10
5
1
0
Lyf49a
6kb up

Figure 56. Confounding factor AML-1 in nucleosome depletion. (A - B) Logistic regression analysis predicted the nucleosome depletion of one factor binding site upon the depletion of another factor binding site of the selected factors (AML-1, c-ETS-1, Lyf-1, and MZF-1), using (A) AML-1 as the predictor, or (B) AML-1 as the reporter. In each case, AML-1 was a better predictor of the other two nucleosome-preferring factors, Lyf-1 and MZF-1, than any of the others was of AML-1. AML-1 performed equally well as predictor or reporter against c-ETS-1, a factor that had no preference for nucleosome coverage, and so unlikely to be involved as a confounding variable. (C) ChIP against the RMA chromatin using anti-AML-1 enriched the Ly49a region over 10-fold. The 6 kbp upstream region displayed no enrichment. An irrelevant anti-isotype antibody was used as a background control in both cases.

Gene	Pro-2	Pro-3
Ly49A	80%	29%
Ly49C	67%	57%
Ly49J	100%	0%
Ly49G	20%	33%
Ly49I	0%	43%
Ly49H	33%	43%
Ly49D	22%	33%
Ly49F	100%	50%
Ly49E	50%	13%
Ly49Q	100%	20%

Table 11. Nucleosome deviation at the AML-1 sites in Ly49A expressing RMA. Convergence analysis of the nucleosome positions in the genomic regions corresponding to Pro-2 and Pro-3 was performed by noting how many AML-1 sites were found to be covered by a nucleosome compared to how many AML-1 sites were present in total. Note that Pro-3 has only been detected as transcriptionally active in Ly49G and Ly49J.

The enhancer elements for exon 1 of the expressed Ly49A in RMA showed high nucleosome divergence at AML-1 binding sites, which was not observed in Pro-2 and Pro-3.

ASSOCIATION RULE MINING

To verify the sensitivity between the factor binding sites and find more relations between them, I searched for the underlying associations between the TF binding sites and their nucleosome coverage state quantitatively. The Association Rule Mining algorithm was applied to identify the relationship, or the rule, between qualitative variables. The rule is defined as an association between the qualitative variables such as the presence of the binding sites and the nucleosome coverage. Various information was used as items, the variables to be searched for association in the association rule mining: the name of the TF binding sites, nucleosome coverage state of each TF binding site, the location of the TF binding sites and the nucleosome bound regions such as a forward Pro-1 promoter, Pro-1 reverse promoter, Pro-2, and all Ly49 genes. The unsupervised learning algorithm searched for rules between all possible combinations of the given variables and identified associations. The association rule mining identified rules between the AML-1 binding site and the open state in Pro-1 regions (**Figure 57**). The AML-1 binding sites in the Pro-1 reverse promoter tend to be open (nucleosome-free), while the AML-1 binding sites in the forward Pro-1 regions tend to be closed (nucleosome-bound). The algorithm also identified several other rules. Another identified rule is the association of AML-1 nucleosome state with the Lyf-1 and MZF-1 nucleosome state. The open AML-1 binding sites are likely to be associated with open Lyf-1 and MZF-1 binding sites. The reverse rule was weaker than this rule.

As a conclusion, the AML-1 site is open and does not have 10 bp periodicity in the distance to the nearest nucleosome dyad, which suggests that AML-1 needs to remain open to be accessible to the transcription factor. AML-1 also predicts Lyf-1, MZF-1 divergence, which suggests AML-1 triggers the removal of the nucleosome on the Lyf-1 and MZF-1 binding sites.

ENRICHED MOTIFS ON PROMOTER REGIONS.

The sequences of the Ly49 gene were analysed for enriched motifs to identify possible cis-elements interacting with nucleosomes. The sequences of the promoter regions (Pro-1 and Pro-2/Pro-3) and around the nucleosome positions were selected for the analysis. First, enriched motifs were searched in the 150 bp long sequences including Pro-1. The AML-1 motif is among the enriched motifs from Pro-1 sequences (**Table 12**). The sequence motif for AML-1, whose binding sites were found at the Pro-1 promoter, was highly ranked in the enriched motifs. The enriched motif search verifies the finding of the AML-1 binding sites at the Pro-1 regions.

The motif search was expanded to the region between exon 1 and exon 2 containing Pro-2/Pro-3 promoters, which are important in mature NK cells. The profile of the enriched motifs is different from the profile from the Pro-1 regions (**Table 13**). The motifs found in Pro-1 region including AML-1 were not found in Pro-2/Pro-3 regions. Instead, GATA, Foxhead (FOX), and Pax family motifs were enriched in the Pro-2/Pro-3 regions. The enriched motifs were grouped by each Ly49 gene (**Table 14**). GATA1 and GATA3 were found in all Ly49 genes. However, Foxhead family motifs (Foxa2, Foxo3, Foxl1, Foxd1) showed distinct presences in certain Ly49 genes even though they were highly ranked enriched motifs when all Ly49 genes considered together. Ly49M, N, K, and Q, where the FOX motifs were not found, are pseudogenes. The lack of motifs for

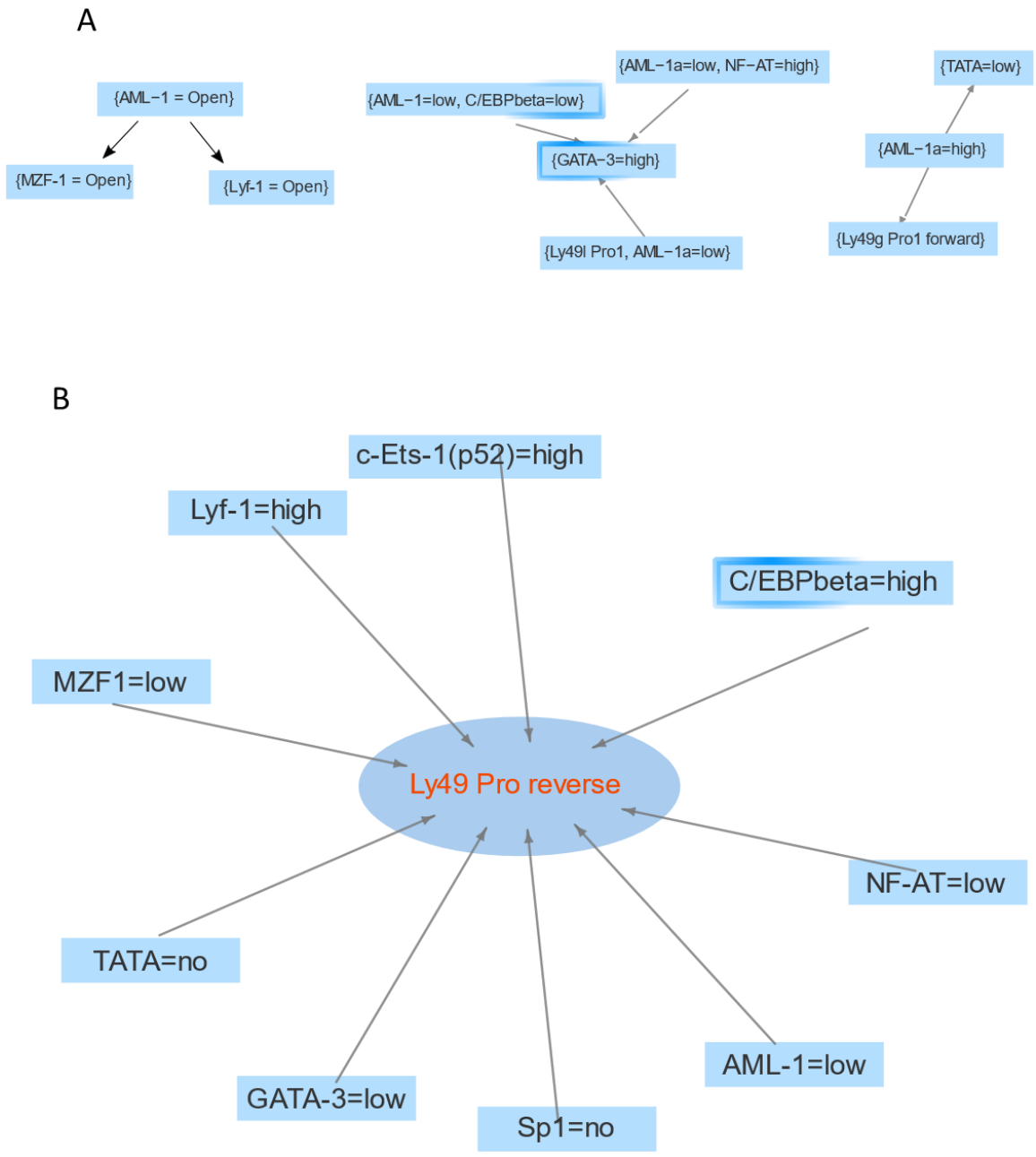

A
{AML−1 = Open}
{MZF-1 = Open}
{Lyf-1 = Open}
{AML−1a=low, NF−AT=high}
{AML−1=low, C/EBPbeta=low}
{GATA−3=high}
{Ly49l Pro1, AML−1a=low}
{TATA=low}
{AML−1a=high}
{Ly49g Pro1 forward}
B
c-Ets-1(p52)=high
Lyf-1=high
C/EBPbeta=high
MZF1=low
Ly49 Pro reverse
NF-AT=low
TATA=no
AML-1=low
GATA-3=low
Sp1=no

Figure 57. Association rules found between Ly49 promoter elements. (A) AML-1 related rules were shown out of the discovered rules by the association rule mining, AML-1 nucleosome bound state, and the abundance of the binding sites in the promoters are associated with the state of other factors. The open AML-1 is associated with open MZF-1 and Lyf-1, but the reverse association is less significant. Low abundance of AML-1 binding sites tends to be associated with a high number of binding sites of GATA-3. The two factors may play distinct roles or work in different promoter regions. On the other hand, a high number of binding sites of AML-1 is associated with the forward Pro-1 promoter. AML-1 may play a role in the forward Pro-1 promoter. (B) Rules associated with open promoters. The various states of the promoter elements such as open-close state, abundance in a promoter, were searched for rules with associated items. These items are most associated with the open promoters. The rule found an association of the abundance of the factors in the open promoters of Ly49 genes. For example, AML-1, MZF-1, GATA-3, NF-AT binding sites were low in the open promoters, and those conditions were associated with Ly49 reverse Pro-1. The rule confirms the pattern found in the Pro-1 regions from the visual inspection of the patterns: Pro-1 reverse promoters tend to be open, while the Pro-1 forward promoters tend to be closed.

TF	Consensus sequence	JASPAR motif	Match Rank
FEV	.....CAGGAAAT.	CAGGAArT	1
ZNF354C	.GTGGAT.....	GTGGAK	2
SPIB	....ACAGGAA...	wsmGGAA	3
AML-1	.AAAACCACAAACAGC........	MAAACCACARAMMMM	4
SPI1	......AGGAAAT.	AGGAAGT	5
AML-1	.....AAAACCACAAACAGC....	MAAACCACARAMMMM	6
AML-1	.AAAACCACAAACAGC........	MAAACCACARAMMMM	7

Table 12. Enriched motifs in Pro-1 regions of Ly49 gene family in C57BL/6 mouse strain. Enriched motifs were searched in the 150 bp long DNA sequences containing Pro-1 regions, which are the Pro-1 promoter regions of Ly49. The enriched motifs identify the most common factors binding to the Pro-1 region in the Ly49 genes. AML-1 is one of the enriched motifs. Enriched motifs were analysed from the set of DNA sequences, and the similarity was searched with the known DNA motifs of transcription factors.

JASPAR Motif	Sequence Consensus	TF Motif	Match Rank
NE2L1::MafG	....CATGAC...	sATGAC	1
SPI1	.....AGGAGGy.	AGGAAGT	2
Foxa2	..TATTTACAWr	TRTTTACwyw..	3
Gata1	......GCAGATAry	rsWGATAAg..	4
GATA3	TryTATCT........	TYTTATCT	5
Pdx1	....CATTAG..	MATTAG	6
FOXO3	..TATTTACA..	TGTTTACM	7
FOXL1	.wATAAATA...	wwdayATA	8
Gata1	TryTATCTGC......	.YCTTATCWSY	9
SPIB	...AGAGGAG...	wsmGGAA	10
FOXD1	.YTATTTAC...	WTGTTTAC	11
SPIB	wkCCTCT......	TTCCKSW	12
EHF	....GAGGAGGy.	SAGGAAGK	13
Pax2	...AGTCATGT...	wgTCAykb	14
Pax2	...ACATGACT....	VMRTGACW	15
Gata1	......GCAGATAry	RSAGATAAG..	16
Pax2	...ACATGACT..	VMRTGACW	17

Table 13. Enriched motifs between exon 1 and exon 2 regions of Ly49 gene family. The enriched sequence motifs were analysed in the same way as the **Table 12**. The enriched motifs in this region are different from the ones in Pro-1 region. FOXL1 is the most common transcription factor, seven out of thirteen genes. FOX gene family transcription factors are among the common factors enriched in the region. GATA and PAX are also commonly found factors.

Genes	Ly49A	Ly49C	Ly49M	Ly49J	Ly49G	Ly49I	Ly49N	Ly49K	Ly49D	Ly49F	Ly49X	Ly49E	Ly49Q
NFE2L+::MafG	+	-	+	-	+	-	+	+	+	-	+	+	-
SPI1	+	+	+	+	+	+	+	+	+	+	+	+	-
Foxa2	+	+	-	+	+	-	-	-	-	+	-	+	+
Gata+	+	+	+	+	+	+	+	+	+	+	+	+	+
GATA3	+	+	+	+	+	+	+	+	+	+	+	+	+
Pdx1	+	+	-	+	+	+	-	-	-	+	-	+	-
FOXO3	+	+	-	+	+	-	-	-	-	+	-	+	+
FOXL1	+	+	-	+	+	+	-	-	-	+	-	+	-
Gata1	+	+	+	+	+	+	+	+	+	+	+	+	+
SPIB	+	+	+	+	+	+	+	+	+	+	+	+	-
FOXD1	+	+	-	+	+	-	-	-	-	+	-	+	+
SPIB	-	+	+	+	+	+	+	+	+	+	+	+	+
EHF	+	+	+	+	+	+	+	+	+	+	+	+	-
Pax2	+	-	+	-	+	-	+	+	+	-	+	+	-
Pax2	-	-	+	-	-	-	+	+	+	-	+	+	+
Gata1	+	+	+	+	+	+	+	+	+	+	+	+	+
Pax2	+	-	+	-	+	-	+	+	+	-	+	+	-

Table 14. Enriched motifs in the Ly49 genes. The enriched motifs were in each Ly49 gene. GATA1 and GATA3 were the most commonly found motifs: they appear in seven out of 13 genes. The presence of the FOXL1 motif is not correlated with the expression of the genes as the motif was found in some low expression genes (Ly49E and J) but not another low expression gene (Ly49Q). The pseudogenes (Ly49M, N, K, and X), however, do not have the FOXL1 binding motif.

Foxhead family in the pseudogenes suggests the vital role of Foxhead family in Ly49 expression, especially in mature NK cells. The protein interaction was investigated with the factors whose binding motifs were enriched in the Ly49 promoters. Interestingly, the genes with the enriched motifs at this region interact with another DNA binding proteins. Many of the interacting proteins are transcription regulators or erythrocyte differentiation related genes (**Figure 58**).

The chromosome sequences around nucleosome bound regions for the motif search was selected to investigate the relation of the factors to the nucleosome. The same enriched motifs found in the Pro-1 promoters, FEV, SPI1, and AML-1, were also enriched in the nucleosome bound regions (**Figure 59**). Interestingly, the enriched motifs were only found at the nucleosome boundaries, but no sequences motifs were found enriched in the nucleosome core region, 50 bp from the nucleosome dyad in either side. The enriched motifs, Klf4, SPI1, MEF2C, FEV, Gfi1b, Erg, and GABPA motifs are located at least 50 bp apart from the nucleosome dyad. The AML-1 binding motif is found at region 65 to 75 bp apart from the nucleosome dyad, which corresponds to the boundary of the nucleosome. The distance may put the AML-1 binding sites bound by nucleosomes as previously identified, but the location is at the nucleosome boundaries, which makes the AML-1 binding sites are available to the factors more easily than being in the middle of the nucleosome bound region. On the other hand, GABPA binding motif, which was found in the 50 to 100 bp from the nucleosome dyad, was clearly out of the nucleosome boundaries lying 115 to 135 bp from the nucleosome dyad.

The motif search around nucleosomes was expanded to include all nucleosome positions and surrounding regions in the Ly49 gene cluster (**Figure 60**). The 300 bp long nucleotide sequences of nucleosome positions and the surrounding sequences were searched for enriched motifs after dividing them into 50 bp long fragments. The profile of

the enriched motifs is rather similar to the enriched motifs from Pro-2/Pro-3 regions than that of Pro-1 regions. Foxhead family and GATA family, the enriched motifs in Pro-2/Pro-3 regions, were also found in the nucleosome-bound sequences. AML-1 was enriched at the nucleosome boundaries as in the other promoter regions. GATA family (GATA1, GATA2, GATA3) motifs were found within the nucleosome boundaries. The enrichment of GATA family motifs in the nucleosome bound regions matches that the binding sites of those factors are bound by nucleosome in the proximity analysis (**Figure 48**). Besides the enriched GATA family, the enrichment of the SP1 and MZF1 motifs near the nucleosome boundaries suggests that the factor binding sites are likely to be bound by nucleosomes as identified in the distance distribution to the nearby nucleosomes. The interwoven organisation of the TF binding sites with the nucleosome positions, within and outside of nucleosome-bound regions, indicate the sophisticated genomic organisation of the *cis*-regulatory elements.

The enriched motifs at the boundaries of nucleosomes had poly(dA:dT) tracts (**Figure 61**). The poly(dA:dT) tract is known as one of the determinants of nucleosome organisation (Segal and Widom, 2009). The poly(dA:dT) stretch limits the nucleosome positioning due to the stiff bending property of the DNA. The presence of similar sequences as the binding motifs for Foxhead family is an indication of genomic sequences organisation controlling the nucleosome and TF binding.

In contrast, the enriched sequences between exon 1 and exon 2 lacked the poly(dA:dT) sequences and Fox gene family motifs (**Table 13**). Considering that the divergence is greater in Pro-1 regions but not in the Pro-2/Pro-3 regions, the Pro-1 regions are more sensitive to the nucleosome positioning and under the direct influence of the nucleosome.

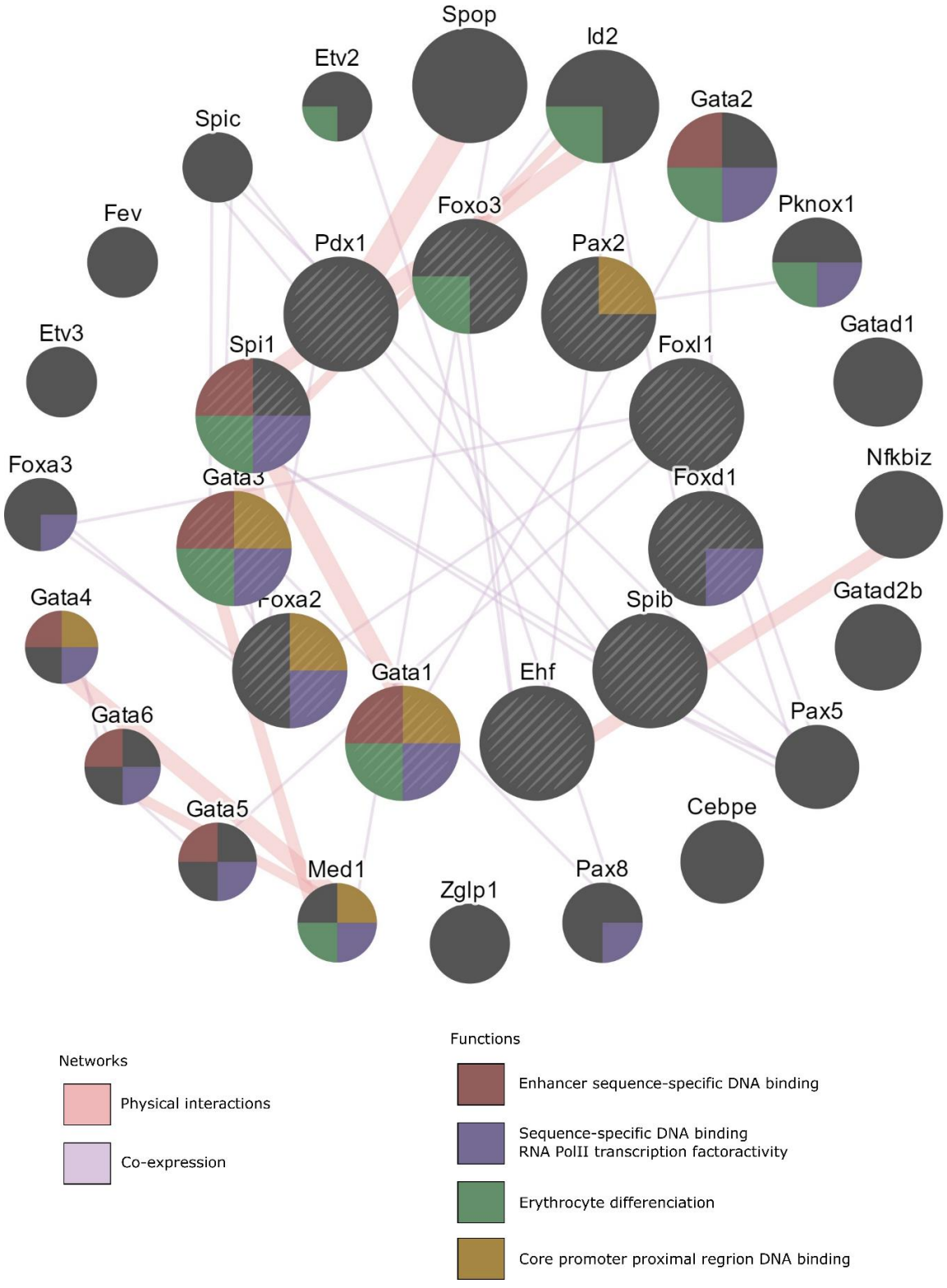

Spop
Id2
Etv2
Gata2
Spic
Pknox1
Fev
Foxo3
Pdx1
Pax2
Etv3
Gatad1
Spi1
Foxl1
Foxa3
Nfkbiz
Gata3
Foxd1
Gata4
Gatad2b
Foxa2
Spib
Gata1
Ehf
Gata6
Pax5
Gata5
Cebpe
Med1
Pax8
Zglp1
Networks
Physical interactions
Co-expression
Functions
Enhancer sequence-specific DNA binding
Sequence-specific DNA binding
RNA PolII transcription factoractivity
Erythrocyte differenciation
Core promoter proximal regrion DNA binding

Figure 58. Protein interactions of the transcription factors enriched between Exon 1 and Exon 2. The transcription factors, whose binding motifs are enriched in the Ly49 gene family, interact with other proteins physically. The interaction and the closely positioned binding motifs may form them as a *cis*-regulatory module (CRM) in the regulation of Ly49 genes in mature NK cells. The colour of the genes represents the biological functions of the genes. The diagonal mark denotes genes used for the search.

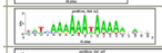
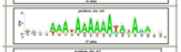
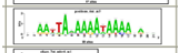
-150 ~ -100	-100 ~ -50	-50 ~ 0	0 ~ 50	50 ~ 100	100 ~ 150
Klf4, SPI1, MEF2C	FEV, Gfi1b, AML-1	No match	No match	No match	Erg, GABPA

Figure 59. Enriched motifs around nucleosomes in the Pro-1 region. DNA sequences of 300 bp long (150 bp from either side of the nucleosome dyad in Pro-1 regions of Ly49 gene family) were analysed for the enriched motifs. The 300 bp long sequences were divided into six 50 bp long sequences depending on the distance from the nearby nucleosome dyad. AML-1 binding motifs are enriched between -100 to -50 bp region from the nucleosome dyad. More specifically, the AML-1 binding motif is found at -75 to -65 bp from the nucleosome dyad (data not shown), which is just about the boundary of the nucleosome. On the other hand, GABPA binding motif was found at 115 to 135 bp from the nucleosome dyad, which is farther from the nucleosome dyad than AML-1. The nucleosome-covered regions (-50 ~ 50 bp) did not show enriched TF binding motifs.

Distance from the nucleosome dyad (bp)

-150~ -100	-100 ~ -50	-50 ~ 0	0 ~ 50	50 ~ 100	100 ~ 150
BRCA1, FOXD1, FOXD3, FOXI1, FOXP1, NFATC2, NFIC SOX2, SOX6, SOX9, SPI1, SPZ1, SRY, YY1	ARID3A, CDX2, ETS1, FOXD3, FOXH1, FOXI1, FOXO3, FOXP1, FOXQ1, GATA1, NR2E3, NRF1, **AML-1**, SOX10, YY1, ZNF354C	BHLHE40, GATA1, MAX, MYCN, MZF1, PDX1, PRDM1, SOX5, SOX9, SP1, SPI1, SPIB, TFAP2A, THAP1, YY1	ARID3A, ARNT::AHR IRF1, JUN::FOS, NR4A2, SOX2, SOX6, SOX9, TCF7L2	BRCA1, CDX2 CDX2, FOXA1, FOXD1, FOXL1, FOXO3, GATA1, GATA2, GATA3, LHX3, NFIC, NKX301, SOX5, YY1	ELK1, ETS1, FOXD1, FOXF2, FOXO3, MZF1, NKX2-5 NKX3-2, REL, SOX2, SOX5, SOX9, SP1, ZNF354C

Figure 60. Enriched motifs around all nucleosomes in the Ly49 cluster. Enriched motifs were searched in genomic sequences around all nucleosomes in the Ly49 gene cluster of C57BL/6. The 300 bp long sequences, 150 bp upstream and downstream from a nucleosome dyad, was divided by 50 bp long sequence fragments. Enriched motifs were searched from the 50 bp long sequences. Binding motifs for FOX proteins were the most commonly enriched protein binding motifs around nucleosomes. FOXO3 binding motif was found on either side of the nucleosome boundary.

Logo	
	YY1, Gata1, ETS1
	CDX2, ARID3A
	FOXP1, FOXI1, FOXO3
	No match
	SOX10, NR2e3
	CDX2, Foxq1
	No match
	FOXI1
	No match
	No match
	No match
	No match
	Gata1, YY1
	No match
	FoxH1, ZNF354C, AML-1

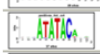
Logo	
	YY1, Gata1
	No match
	No match
	NFIC
	No match
	CDX2
	No match
	FOXL1
	No match
	No match
	No match
	NKX3-1, GATA3
	GATA2
	No match
	No match

Figure 61. Enriched motifs at the nucleosome boundaries. The enriched motifs were searched with the sequences of 50 bp to 100 bp from the nucleosome dyad in upstream and downstream of the Ly49 gene family. Many enriched motifs contain poly(dA:dT) sequences. The poly(dA:dT) tract is one of the determinants of nucleosome organisation, which hinders the nucleosome positioning (Segal and Widom, 2009). The presence of the poly(dA:dT) stretch at the nucleosome boundaries may define or at least help to position a nucleosome at the specific position.

PROTEIN INTERACTION OF THE FACTORS

The previous results of the sensitivity and association rule mining strongly suggest that accessibility of the Lyf-1 and MZF-1 binding sites are highly regulated by nucleosome positioning, and AML-1 binding may trigger the removal of the nucleosome resulting in exposure of those binding sites. However, Lyf-1 and MZF-1 do not have nucleosome remodelling activity, which may be required to access the binding sites of the nucleosome-bound region. The physically interacting proteins and co-expressed proteins with these factors were searched from GeneMania database (**Figure 62**). Interestingly, Lyf-1 interacts with nucleosome remodelling factors (Chd3, Smarchcc, Smarcd3). It also interacts with many factors known to regulate lymphocyte (Ik23, Ikzfl, Foxp3). The nucleosome remodelling factors recruited by Lyf-1 may remove the nucleosome and make the binding sites accessible to the factors.

Discussion

In this study, I have shown that nucleosome positioning was involved in regulating Ly49 expression, not just by condensing the DNA, but by specifically interfering with transcription factor binding sites, especially AML-1 transcription factor binding sites. First, I have shown that the AML-1 binding site within the Ly49 gene cluster was predicted to be sensitive to the coverage by a nucleosome by computational analysis. The AML-1 binding sites in the promoter regions tended to reside within a nucleosome boundary, meaning that AML-1 binding sites have higher chances to be part of the nucleosomal DNA. Moreover, the lack of 10 bp periodicity of the AML-1 binding sites, unlike other covered TF binding sites, put the bindings sites "out-of-phase" with the nucleosomes and could not adopt a conformation allowing the binding sites to be in the same orientation on the surface of the nucleosome.

The predicted nucleosome coverage was verified using a cell line, RMA. The expressed Ly49 genes showed preferential nucleosome depletion in the promoters, especially at AML-1 binding sites compared to the inactive Ly49 genes. The nucleosome depletion at AML-1 sites was not likely to be the result of other confounding TF binding sites, but AML-1 was the primary target site for the nucleosome coverage. The knowledge of nucleosome depletion at AML-1 binding sites was shown to be as good or better at predicting the depletion of other nucleosome depleted TF binding sites, but not *vice versa*. Altogether, these results suggest that Ly49 expression is regulated, at least partially, by nucleosome occupancy at AML-1 binding sites.

I chose the Ly49 gene cluster to examine the interplay of nucleosomes and transcription factors for both immunological importances and practical reasons in bioinformatics. Immunologically, Ly49 receptors on NK cells are of great importance

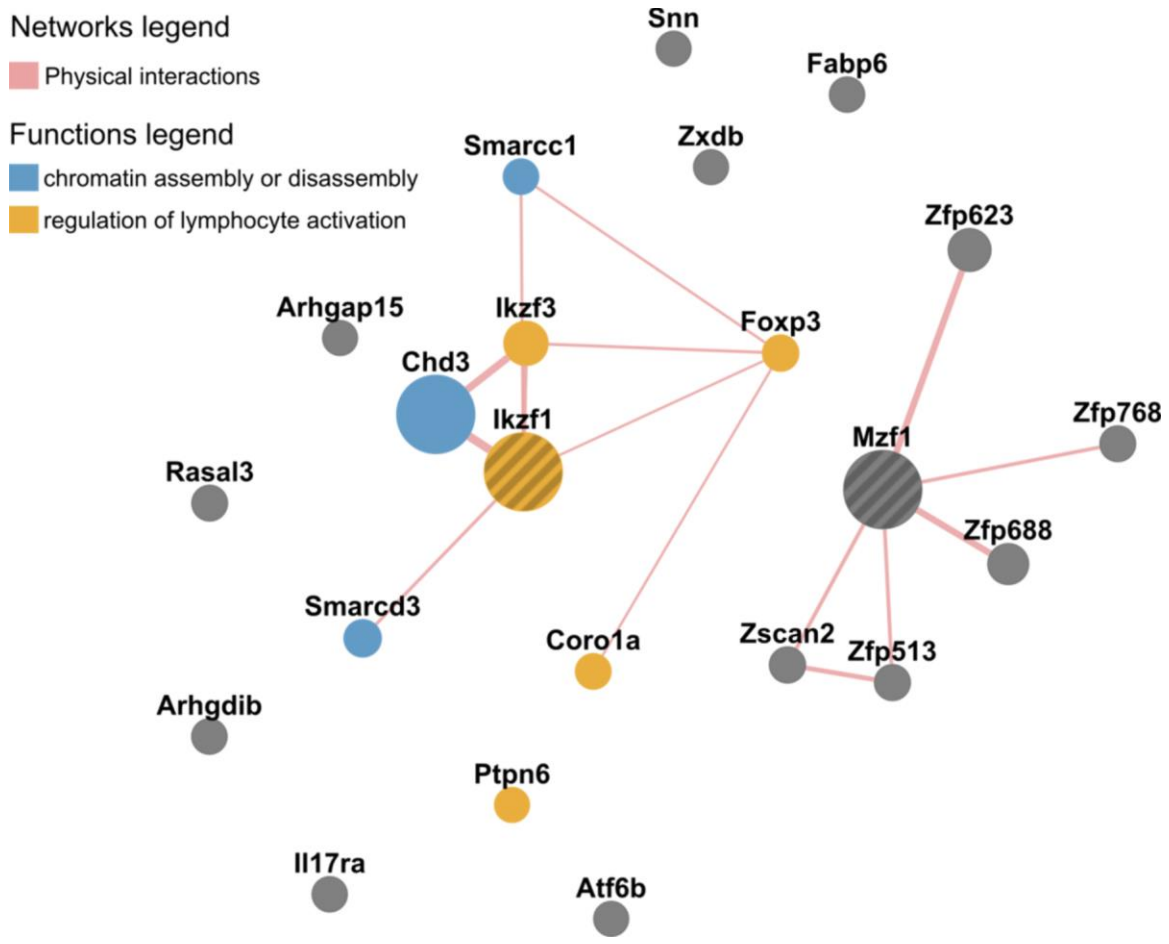

Networks legend
Physical interactions
Functions legend
chromatin assembly or disassembly
regulation of lymphocyte activation
Snn
Fabp6
Smarcc1
Zxdb
Zfp623
Arhgap15
Ikzf3
Foxp3
Chd3
Ikzf1
Zfp768
Mzf1
Rasal3
Zfp688
Smarcd3
Coro1a
Zscan2
Zfp513
Arhgdib
Ptpn6
Il17ra
Atf6b

Figure 62. Proteins interacting with Lyf-1 and MZF1. The interacting proteins with Lyf-1 and Mzf1 were presented. Transcription factors related to the chromatin assembly (blue circles, Chad3, Smarcd3, Smarcc1) and lymphocyte activation (yellow circles) were among the physically interacting proteins.

playing a primary role in innate immunity. Moreover, Ly49 gene expression on NK cells is a complex phenomenon: the genes are polymorphic, and the expression is stochastic leading to heterogeneous subpopulations of NK cells. This heterogeneity allows the population of NK cells to be able to respond to virtually any alteration to homoeostasis, which discriminates damaged or foreign cells from the normal and self-cells (Ljunggren and Kärre, 1990). The variegated expression of Ly49 receptors is the key to the NK cell function. Understanding the Ly49 transcriptional regulation is essential to comprehend the roles of NK cells more thoroughly. Even though there are studies on transcription factors and *cis*-elements involved in the complicated Ly49 expression, there have been few studies about the role of nucleosome positioning.

For the practical reasons of bioinformatics and computational analysis, the Ly49 genes residing in a cluster provides attractive conditions. The gene cluster of the Ly49 family provides a good model system where the analysis is possible from visual and manual to statistical and systematic. It is possible to make comparisons between related expressed and non-expressed genes, using one group as a control for the others, while keeping any trans-acting factors and large-scale genetic factors, such as chromosome looping or the action of locus control regions, are even out on the genes under investigation, regardless of the expression state. This Ly49 gene family, or KIR in human datasets, are a good model system for other bioinformatics or systematic analyses.

Considering the known role of AML-1 in Ly49 transcriptional regulation, it is intriguing that the computational analysis identified and confirmed AML-1 binding sites—especially within the Pro-1/enhancer region—as the most significant player in terms of nucleosomal regulation. One of the notable features of the Ly49 Pro-1 region is a central AML-1 binding site, occurring in almost all Ly49 genes, identified as being required for the bidirectional promoter Pro-1 to function in either direction (Saleh et al.,

2004). Thus, targeting AML-1 binding sites by nucleosomes and disrupting the availability represents a potent mechanism of activating and repressing Ly49 expression. That was also confirmed and explained by the comparison of expressed and silent genes in the cell line. Interestingly, the presence of the AML-1 site in the middle of both forward and reverse promoters may need further study. One may argue that the presence of the AML-1 site for both forward and reverse promoters suggest that nucleosome regulation is unlikely to impact the stochastic expression of Ly49 gene directly, but rather affects any gene activity at all regardless of forward or reverse. However, without knowing the absolute requirement of the AML-1 site for both forward and reverse transcription, it is still open that nucleosome may impact the stochastic expression of Ly49 genes because the positioning of nucleosome itself is stochastic. If either the forward or the reverse transcription is allowed, then the stochastic or random positioning of nucleosome may impact the stochastic expression of the Ly49 genes. The impact of random positioning of nucleosomes on the stochastic expression of other genes is an interesting area to investigate.

Pro-1 is known to be active only in developing NK cells. That the nucleosome regulation is targeting the Pro-1/enhancer region may also indicate that the regulatory effects are exerted during the development and maturation. As the other epigenetic controls, the transcriptional regulation by nucleosome positions may operate in the larger scope of regulations such as tissue development.

The analysis of the enriched motifs showed that winged helix transcription factor FOX binding motifs are positioned at the nucleosome boundaries. Some FOX proteins, e.g. FOXA, are known as a pioneer factor. The binding motifs enriched at the nucleosome boundaries imply the possible genomic organisation of the nucleosome positioning for pioneer factors. Moreover, it will be interesting to examine in the future

that the nucleosomes are H2A.Z because H2A.Z and the winged helix transcription factor Foxa2 act together to remodel nucleosomes and to regulate gene activation, thus promoting embryonic stem cell differentiation into endoderm (Li et al., 2012).

Finally, I suggest as the next step to study such nucleosome sensitivity of AML-1 binding sites at other, non-Ly49 genes, to determine whether the sensitivity to nucleosome occupancy is a unique trait for Ly49 regulation and AML-1 is ubiquitously sensitive to nucleosome occupancy, or even the other transcription factors are sensitive to the nucleosome occupancy as AML-1.

Our analysis of nucleosome-transcription factor interplay in the whole-chromosome range revealed the surprising result. On that scale, many other transcription factors displayed the nucleosome coverage and lacked the 10 bp periodicity, even the ones showing the 10 bp periodicity in regulatory regions. The nucleosome coverage and lack of 10 bp periodicity, the characteristics that identified AML-1 as a nucleosome sensitive factor in the Ly9 cluster were more widely spread on the binding sites in extended chromosome than the regulatory regions. The fact that many nucleosome-covered transcription factor binding sites in the regulatory regions have displayed the 10 bp periodicity suggests that the binding sites still remain available to the protein factors to bind to. However, on a genome-wide scale, many of the transcription factor binding sites became or had the potential to be nucleosome-sensitive: covered by nucleosome and lack of 10 bp periodicity. Within a given gene or gene family, however, only a small number of factors, AML-1 in the Ly49 family, retain this nucleosome-sensitivity and may be under the control of nucleosome coverage, while the other factors are insensitive to the nucleosome coverage by avoiding nucleosomes or by getting in-phase with them.

The 10 bp periodicity of the TF binding site's distance to the nearest nucleosome may be related to not only the orientation on the nucleosome but also the chromatin

folding. The spatial organization of chromatin domains and compartments are getting more attention in gene regulation. Hi-C or chromosome conformation capturing methods revealed spatial organizations of chromosomes such as spatially separated parental chromosomes (Du et al., 2017) and separate folding compartments with different GC content and CTCF (CCCTC-binding factor) binding sites (Xie et al., 2017). The different folding and chromatin domains adopted by active and inactive X chromosomes suggest the changes of the spatial organisation of chromatin domains in response to regulation (Wang et al., 2016). Nucleosome repeating length (NRL), which is the combined length of nucleosome core and linker and is typically 155 – 240 bp long, have rotational settings by the linker of 10 n + 5 bp length (n = 1, 2, and so on) (Lohr, 1981; Wang et al., 2008). The fixed rotational settings are negatively correlated with chromatin folding that marks active genomes (Correll et al., 2012).

The 10 bp periodicity may be an important trait determining the sensitivity to a nucleosome coverage; then a factor can be either sensitive or insensitive to nucleosome coverage depending on the configuration of the nearby nucleosome positioning sites, which broaden the role of a transcription factor beyond the existence of TF binding sites during the transcriptional regulation. A genome-wide characterization and identification of these nucleosome-sensitive factors and their associated genes could provide a novel method of functional clustering of genes. This analysis of the periodicity from the distance distribution of binding sites relative to the nucleosome positions will contribute to the understanding of how chromatin compaction and nucleosome positioning shape the gene expression landscape.

Not only this complex expression of Ly49 provides a good model but it also introduces hindrances into the present study especially for the experimental verification. Most of the randomly expressed Ly49 genes in an NK cell population—usually the

inhibitory receptors—display a mono-allelic expression pattern (Rouhi et al., 2006, 2009). Therefore, NK cells expressing the Ly49 in questions are required to be enriched first from the primary NK cell population, and even after that only half of the genomic material of the enriched cells contains a Ly49 expressing components. For this reason, the availability of the RMA cell line is valuable, which naturally expresses Ly49A. The expression of Ly49A in the entire population presents the cell line as a convenient model to verify our predictions in a natural setting without further introducing mutations or other inhibitory agents. Future work may expand performing this analysis to the heterogeneous pool of natural killer cells, which would allow the comparison between inhibitory and activating Ly49 receptors, of which the RMA cell line lacks.

Also in the future, the follow-up study of the protein interaction of the transcription factors of interest. Lyf-1 and MZF-1 were identified for their interaction with many proteins, and AML-1 could predict the nucleosome removal of those factors during the Ly49 gene expression. The functions or biological roles of the interacting proteins are not yet clearly defined. Further studies on the interactions and propagation of the signals during the gene expression would find important networks or connections in the transcriptional regulation of Ly49 genes and NK cell maturation.

GENERAL DISCUSSION

The attention to the histones and nucleosomes has swung between two opposite sides since its discovery. Initially, the histones were considered as the molecule conveying the genetic information thanks to the diverse proteins (Kornberg and Lorch, 1999). DNA, which consists of only four bases was not considered complex enough to carry genetic information. DNA was merely considered to play a simple role of bridging the histones. However, later it was found that histones were one of the most conserved proteins across species. The most conserved histone H4 sequences have >95% identity across all known sequences (Baxevanis and Landsman, 1996). The apparent diversity was an artefact during the acid extraction. Far from being diverse, histones could hardly carry the diverse information. And yet, the possibility of histone being regulatory molecules did not discount. Based on the findings of variations in the ratio of the histones in the vertebrates, the idea of combinatorial control of histones, the several different types of repeating arrangements of histones, held as an alternative (Huberman, 1973). The combinatorial control was finally dismissed after the unique repeating structure of the nucleosome was recognised (Kornberg and Lorch, 1999). The glory of the regulatory molecule shifted to DNA, and the histones were regarded as a complement of the DNA.

Then, the discovery of the histone acetyltransferase (HAT) activity of transcription factors brought back the regulatory role to the nucleosome (Struhl, 1999). Through the activities of HAT and histone deacetylase (HDAC), the chromatin state of compactness varies between euchromatin and heterochromatin, and facilitates or silences the gene expression. Packaging promoters in nucleosome have an inhibitory effect on transcription *in vivo* (Han and Grunstein, 1988). Hence, knowing nucleosome positioning is important to understand transcription.

There has been a discussion about whether nucleosomes are placed randomly or in sequence-specific manners on the DNA (Kornberg and Stryer, 1988). The paradox is that almost all DNA sequences are packaged in chromatin, so the sequence specificity is inconceivable. However, there have been reports of non-random locations of nucleosomes. The idea of statistical distributions, non-random locations by a stochastic mechanism was introduced (Kornberg and Stryer, 1988). In the statistical positioning, the statistical distribution of nucleosomes was calculated with minimum assumptions that agree with experiments. The result was the evenly distributed nucleosomes adjacent to a boundary. The boundary could by a DNA bound protein or a specifically placed nucleosome. This barrier model also reported later (Mavrich et al., 2008a). The barrier position is crucial to know the nucleosome positions; The +1 nucleosome is enriched at downstream of transcription start site in yeast (Albert et al., 2007; Mavrich et al., 2008a; Montgomery et al., 2001; Yuan et al., 2005). This study also found the enriched H2A and H2A.Z +1 nucleosomes in Drosophila at downstream of the transcription start sites. The +1 nucleosome sequences were selected to search for the underlying DNA sequences that specifically place nucleosomes.

Periodic appearances of dinucleotides is a prominent sequence-specific determinant for nucleosome positioning (Johnson et al., 2006; Mavrich et al., 2008b; Valouev et al., 2008). The dinucleotide patterns for nucleosome positioning, or nucleosome positioning sequences were well-known in eukaryotes such as yeast (Kaplan et al., 2009; Segal et al., 2006), *C. elegans* (Salih et al., 2008), chicken (Satchwell et al., 1986), and Human (Kogan et al., 2006), but not fully reported in Drosophila. The H2A and H2A.Z nucleosome positioning patterns were identified using *in vivo* nucleosome positions from experimental data. Because the periodic appearance of a dinucleotide of NPS is subtle stretching over a 146 bp long sequence, the nucleosome sequences had to

be analysed collectively and iteratively to reduce the background noise. The DNA fragments selected from nucleosomes were sequenced by high-throughput sequencing technology using a paired-sequencing tag. Thanks to the paired-sequencing tag, the boundaries of the sequenced DNA fragments could be determined without the need of estimation. Otherwise, the boundaries need to be estimated after statistical approximation of the aligned data increasing the noise. Choosing the well-positioned nucleosome is important as the fuzzy nucleosomes add more noise to the already weak signal. The paired-end sequencing improved to determine the nucleosome positions with more confidence.

As previously mentioned, the sequence-specific positioning of the nucleosome is prominent in setting the boundary of nucleosome landscape. The nucleosome enrichment around the transcription start site was reported in other organisms (Barski et al., 2007; Kristell et al., 2010; Schones et al., 2008). The nucleosome landscape in the promoters was explored so to find that the +1 nucleosome was enriched at downstream of transcription start sites as in the other organisms. The difference is, however, yeast has enriched H2A.Z -1 nucleosome, while Drosophila lacks the -1 nucleosome.

Fourier transform calculated the period from the dinucleotide patterns. The dominant period identified from Fourier transform was verified by building a linear periodic model. The initial analysis to identify the Drosophila NPS showed overall 10 bp periodicity in WW and SS dinucleotide patterns, while RR and YY had different periodicity. A periodic model was built to verify the 10 bp period by predicting the patterns. The initial model was marginally useful to predict the patterns: the deviation increased as the peaks were located farther from the middle of the pattern. It indicates that even though the overall period of the pattern is 10 bp, the local period may be different locally. The changes of the periodicity along the sequence were detected by

moving analysis by Fourier transform. The flanking regions of the nucleosome sequences had the 10 bp period, while the period was approximately 20 bp in the middle of the nucleosome sequence. There has been a report that the DNA conformation in the midst of a nucleosome is flatter than the surrounding region (Luger et al., 1997; Tolstorukov et al., 2007). The structural differences are reflected in the periods of the NPS.

There are anti-NPS-patterns, which also govern the nucleosome positioning Ioshikhes et al. (2011). The selected sequences for the NPS may have been the mixture of the sequences with the NPS patterns and the anti-NPS patterns. The initial dinucleotide patterns were used to enrich sequences with the NPS patterns by separating the positively correlated sequences for NPS patterns and the negatively correlated sequences for anti-NPS patterns based on the correlation. The positively correlated sequences generated more evident NPS patterns. The linear model verified the overall 10 bp period of the dinucleotide patterns with variations near the dyad. Canonical NPS patterns for Drosophila nucleosomes were proposed as 10 bp periodic WW/SS patterns with SS preference at the dyad.

Besides the H2A nucleosome sequences, the nucleosome sequences of H2A.Z nucleosome were also analysed. H2A.Z nucleosomes have different biological roles from the H2A nucleosomes (Jackson and Gorovsky, 2000). H2A.Z nucleosome is often found in the active genes, and it is believed to be related to the activation of a gene, even though there are conflicting reports. The different landscape of H2A.Z nucleosomes in promoters foretold that the NPS pattern may not be the same. In addition to the lack of H2A.Z -1 nucleosomes, closer examination showed that the positions of the H2A and H2A.Z +1 nucleosome were slightly different.

With the similar analysis to identify the H2A nucleosome positioning patterns, the H2A.Z nucleosome positioning patterns were identified. Despite the same 10 bp

periodicity, the identified H2A.Z NPS patterns differ from the H2A NPS patterns with strong deviations at ±45 bp positions from the nucleosome dyad. The X-ray structure has revealed an extended acidic patch on the nucleosome surface of H2A.Z/H2B dimers (Suto et al., 2000). Loop 1, one of the structurally different residues between H2A and its variant, is where the DNA interacts with. The structural differences may induce the subtle differences in the NPS patterns. There are conflicting reports about the stability of H2A.Z as well as of its perplexing roles. Both reduced ionic-strength dependent stability (Abbott et al., 2001) and (Zhang et al., 2005) and increased stability (Park et al., 2004) were reported. The differences between H2A and H2A.Z in the protein structures and the accompanying NPS patterns probably impact on the stability or the mobility of the H2A.Z nucleosome. The exact changes will be the further subject of study.

The promoter sequences were organised to accommodate different core promoter elements and the nucleosome sequences. The promoters containing the H2A positioning sequences were more likely to have CCAAT and TATA elements. The enriched function of the associated genes was transcription regulation. As the CCAAT box is usually found in the tightly regulated genes (Spitz and Furlong, 2012), the positioning of nucleosomes in proximity may function as a regulator together with transcription factors to repress or to activate the gene expression. Various roles of nucleosome as part of the transcriptional regulation have been reported: poising the RNA polymerase (Levine, 2011; Teves and Henikoff, 2011), elongation (Adam et al., 2001). The genomic sequence itself may as well be organised as a layout accommodating both transcription factors and nucleosomes.

To elucidate the interplay between the nucleosome and the transcription factors, I chose the Ly49 gene cluster as a model system. Ly49 proteins are the surface receptor expressed in mouse NK cells. The genes are located on chromosome 6 as a cluster. They

are important in innate immunity, and their stochastic expression and the clustered genes have pragmatic advantages in this computational analysis.

The transcript from the reverse promoter of the bidirectional promoter, Pro-1, repressed the expression, while the transcript from the forward promoter activates the transcription. The transcription is believed to change the closed chromatin state induced by methylation at Pro-2. The changes of the chromatin status remain the same during the maturation leading to the permanent activation or repression in mature NK cells (Saleh et al., 2002). The pivotal element for the stochastic expression of the Ly49 genes is how to select the forward or reverse promoter. Because of the similar promoters among the Ly49 genes clusters, transcriptional regulation only by transcription factors is not sufficient to explain the stochastic expression of the Ly49 genes.

I hypothesised that nucleosome positioning at the Pro-1 may be entailed in the selection of the reverse and the forward promoter as an added layer of regulation. The nucleosome positions and the transcription factor binding sites were examined on the Ly49 gene cluster of C57BL/6 mouse. Visual inspection of the map revealed the unique arrangement between them. The Pro-1 reverse promoters, which repress the expression in the immature NK cells, were preferentially open by avoiding the nucleosome positions so that they are readily available for transcription. The forward Pro-1 promoters, however, were more likely to be covered, or overlapped by the nucleosome position. Also, the binding sites for AML-1, which is one of the known regulators of Ly49 expression, were almost always found in the midst of the reverse and the forward Pro-1 promoter. The same arrangement was also observed in other mouse strains such as BALB/c, 129/S6, and NOD.

The arrangement was further verified quantitatively by the calculated association between the nucleosome positions, open Pro-1 reverse promoters, and the covered

forward Pro-1 promoters. Because the analysis was done only with the genomic sequences, the results can be considered as a default setting of the promoters encoded in the genome without effects of the other factors. Upon the finding that forward Pro-1 promoter is covered, the relation between the nucleosome positions and the transcription factor binding sites were extended to see whether other transcription factors were affected by the nucleosome positions. In other words, the nucleosome coverage on the binding sites was systematically analysed with more immune cell specific transcription factors.

First, transcription factors were identified, which were possibly covered by nucleosome: the TF binding sites within 73 bp range from the nearest nucleosome dyad. The distance distribution showed that some factors were more likely to be covered by nucleosomes while others were avoiding nucleosomes. TATA, which used as a control, has clearly shown that majority of TATA boxes were out of the nucleosome reach. The binding sites of the NF-AT and C/EBPb were out of the nucleosome positions. On the other hand, there were other factors whose binding sites were located within the boundaries of the nucleosome, i.e., around 73 bp apart from the nucleosome dyad, or completely covered by the nucleosome sites.

One of the nucleosome-covered factors is AML-1. AML-1 is a known factor in the Ly49 gene regulation, and its binding sites were likely to be covered by nucleosomes. For AML-1 to be working, its binding sites should be exposed to be accessible by the factor. The AML-1 does not have nucleosome remodelling activity according to the sequence profile (Lo Coco et al., 1997; Tanaka et al., 1997). AML-1 and other factors, whose binding sites are covered by a nucleosome, need help from another protein that removes the nucleosome, or another mechanism to be able to bind the nucleosome-covered DNA. It has been reported that the 10 bp periodic distance of the binding sites to the nearest nucleosome dyad might make the factor available even though the binding

sites included in the nucleosome thanks to the DNA wrapping the histones putting the sequences on every 10 bp apart on the same side of the DNA (Ioshikhes et al., 1999).

The distances between the nucleosome covered factor binding sites and the nucleosomes dyad were analysed for the periodicity in every factor using a leave-one-out test. It recognised the lack of the periodicity in the nucleosome covered AML-1 binding sites, while another nucleosome covered Lyf-1 and MZF-1 binding sites showed 10 bp periodicity. The 10 bp periodicity arranged the factors' binding sites available even when they were overlapped by a nucleosome. AML-1, unlike the factors with 10 bp periodicity, could be completely shut down by nucleosomes or sensitive to nucleosome

The regulation by nucleosomes coverage was verified in RMA cell line, which expresses Ly49A only (Kärre et al., 1986). At least half of the predicted nucleosome sites were verified with the experimental data. Furthermore, the deviations of the nucleosome positioning from the predicted ones were compared between the expressed and non-expressed Ly49 genes. The deviation was statistically significant in Ly49a and Ly49k. Because the Ly49k is a pseudogene, the deviation from the prediction is practically happening only in the active gene. The predicted sites of nucleosomes and factors can be considered as the default settings of the genome. The changes of nucleosome positioning *in vivo* from the default settings tell the activity of cellular factors working.

In addition to the nucleosome depletion, AML-1 binding was verified by immunoprecipitation with the anti-AML-1 antibody. The Ly49A promoters were enriched more than 10-fold in the precipitated sequences. The immunoprecipitation proved that the depletion of nucleosomes and the binding of AML-1 in the Ly49A promoter. The nucleosome coverage status of AML-1 could predict the status of the Lyf-1 and MZF-1 in the promoters, but the opposite did not hold leading to the conclusion that AML-1 depletion precedes the depletion of Lyf-1 and MZF-1. Together, the

nucleosome is regulating the transcription of Ly49 gene expression by covering one of the key factors, AML-1 binding sites.

The 10 bp periodicity is necessary for the nucleosome-covered factor binding sites to be accessible without the help of remodellers. The binding sites in the promoters tend to have the 10 bp periodicity, which makes the binding sites available even though it was covered by nucleosomes. Interestingly, the periodicity was not observed when the search incorporated all the binding sites in entire chromosome 6 instead of the promoter regions. The 10 bp periodicity is the unique feature of the promoters. Because the lack of the periodicity in the nucleosome-covered binding sites, many of the binding sites may not be the true binding sites *in vivo*. The relatively short sequence motifs for transcription factors can exist on the genome by chance. Predictions of transcription binding sites often generate too many false positives. There have been significant efforts in to improve the computational identification of the functional binding sites from the physical binding sites (Hannenhalli, 2008; Wasserman and Sandelin, 2004). I propose that incorporating the nearby nucleosome positions into the prediction will improve the accuracy of the prediction. By expanding the searching for the periodicity in the promoters of other genes than Ly49, it may possible to see the periodicity in more general cases. Having identified the periodicity between the binding sites and the proximal nucleosome, a better model could be built minimising false prediction.

This study of the interplay between nucleosome positions and transcription factor bindings in Ly49 gene cluster can be expanded toward human cases, KIR genes. KIR genes are involved in the MHC-I expression and the maturation of NK cells in human. However, KIR genes are not the homologous of Ly49 genes, even though they share the same roles in NK cell and innate immunity. It would be worthwhile if the same principle can be applied to the genes of the same function but different origins.

These Part 1 and Part 2 studies explored how nucleosomes are positioned and the interaction with other factors in the cell. The periodic patterns of genome sequences are one of the factors determining nucleosome positions. The 10 bp periodic NPS pattern is a weak signal requiring sophisticated analysis of many nucleosome sequences to distinguish the pattern from the background noise. However, the weak and subtle signal is critical in organising the transcription factors and nucleosomes in the genome. The long and subtle NPS pattern makes it possible to position nucleosomes in the practically overall genome, and yet strong enough to position some of the nucleosomes at specific positions.

The nucleosome positioning can be viewed as the positioning of studs in construction. In general, the rule governing the placement of the studs is the equal spacing of 12 or 16-inch apart. However, some of the studs are placed strategically at specific positions. For example, the king studs are located to define the spaces for putative windows or doors supporting the opening; The other studs were located following the king stud. The +1 nucleosomes are like the king stud: it defines and supports the opening for the transcription so that the TF factors can work through it. The layout is coded in the DNA sequences beforehand. Moreover, the genomic sequence is organised in a fascinating way. It incorporates the two layers of regulation. The long loosely defined nucleosome positioning patterns could accommodate the more specific motifs for the binding of transcription factors.

The genomic sequences are like a blueprint arranging all the activities in a cell. The predictions based on the genomic sequence can be inferred as the default setting without the effects of the ongoing dynamics of living cells. The arrangement will change, but the genomic settings set the limits of the actions.

The periodic appearance of dinucleotide over a long stretch of the sequences is less stringent than the specific transcription factor binding motifs. It is possible for the dinucleotide patterns and the TF binding sites to coexist on DNA. The dinucleotide signal is subtle, so it will not interfere with the TF binding motifs at the same time direct the positioning of the nucleosome. It is like merging of two layers of regulation in one-dimensional DNA sequence: TF binding sites over the background of nucleosome positioning sequences. The analysis of nucleosome positions in addition to the factor binding sites will be a computational tool to extract more information from the genomic sequences even before validating with costly experiments. Not only the practical advantages but also the new model of genetic networks of merged layers of the nucleosome positioning and the factor bindings on 1-dimensional genomic sequences may also elucidate the wonders of organising information in the biological systems.

References

Abbott, D.W., Ivanova, V.S., Wang, X., Bonner, W.M., and Ausió, J. (2001). Characterization of the stability and folding of H2A.Z chromatin particles: Implications for transcriptional activation. J. Biol. Chem.

Adam, M., Robert, F., Larochelle, M., and Gaudreau, L. (2001). H2A.Z Is Required for Global Chromatin Integrity and for Recruitment of RNA Polymerase II under Specific Conditions. Mol. Cell. Biol.

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

Albert, I., Wachi, S., Jiang, C., and Pugh, B.F. (2008). GeneTrack--a genomic data processing and visualization framework. Bioinformatics *24*, 1305–1306.

Alvarez-Saavedra, M., De Repentigny, Y., Lagali, P.S., Raghu Ram, E.V.S., Yan, K., Hashem, E., Ivanochko, D., Huh, M.S., Yang, D., Mears, A.J., et al. (2014). Snf2h-mediated chromatin organization and histone H1 dynamics govern cerebellar morphogenesis and neural maturation. Nat. Commun. *5*, 4181.

Anderson, S.K., Dewar, K., Goulet, M.-L., Leveque, G., and Makrigiannis, a P. (2005). Complete elucidation of a minimal class I MHC natural killer cell receptor haplotype. Genes Immun. *6*, 481–492.

Ansel, K.M., Lee, D.U., and Rao, A. (2003). An epigenetic view of helper T cell differentiation. Nat. Immunol. *4*, 616–623.

Arase, H. (2002). Direct Recognition of Cytomegalovirus by Activating and Inhibitory NK Cell Receptors. Science (80-.).

Arents, G., and Moudrianakis, E.N. (1993). Topography of the histone octamer surface: repeating structural motifs utilized in the docking of nucleosomal DNA. PNAS.

Arents, G., Burlingame, R.W., Wang, B.C., Love, W.E., and Moudrianakis, E.N. (1991). The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix. Proc. Natl. Acad. Sci. *88*, 10148–10152.

Attwood, J.T.T., Yung, R.L.L., and Richardson, B.C.C. (2014). DNA methylation and the regulation of gene transcription. Cell. Mol. Life Sci. C. *59*, 241–257.

Bai, L., and Morozov, A. V (2010). Gene regulation by nucleosome positioning. Trends Genet. *26*, 476–483.

Bargaje, R., Alam, M.P., Patowary, A., Sarkar, M., Ali, T., Gupta, S., Garg, M., Singh, M., Purkanti, R., Scaria, V., et al. (2012). Proximity of H2A.Z containing nucleosome to the transcription start site influences gene expression levels in the mammalian liver and brain. Nucleic Acids Res. *40*, 8965–8978.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Baxevanis, A.D., and Landsman, D. (1996). Histone sequence database: A compilation of highly-conserved nucleoprotein sequences. Nucleic Acids Res. *24*, 245–247.

Belanger, S., Tai, L.-H., Anderson, S.K., and Makrigiannis, a P. (2008). Ly49 cluster sequence analysis in a mouse model of diabetes: an expanded repertoire of activating receptors in the NOD genome. Genes Immun. *9*, 509–521.

Beĺanger, S., Tu, M.M., Rahim, M.M.A., Mahmoud, A.B., Patel, R., Tai, L.-H.H., Troke, A.D., Wilhelm, B.T., Landry, J.-R.R., Zhu, Q., et al. (2012). Impaired natural killer cell self-education and "missing-self" responses in Ly49-deficient mice. Blood *120*, 592–602.

Van Beneden, K., Stevenaert, F., De Creus, A., Debacker, V., De Boever, J., Plum, J., and Leclercq, G. (2001). Expression of Ly49E and CD94/NKG2 on Fetal and Adult NK Cells. J. Immunol. *166*, 4302–4311.

Benson, D.M., Hofmeister, C.C., Padmanabhan, S., Suvannasankha, A., Jagannath, S., Abonour, R., Bakan, C., Andre, P., Efebera, Y., Tiollier, J., et al. (2012). A phase 1 trial of the anti-KIR antibody IPH2101 in patients with relapsed / refractory multiple myeloma. Blood *120*, 4324–4333.

Binstadt, B.A., Brumbaugh, K.M., Dick, C.J., Scharenberg, A.M., Williams, B.L., Colonna, M., Lanier, L.L., Kinet, J.-P., Abraham, R.T., and Leibson, P.J. (1996). Sequential Involvement of Lck and SHP-1 with MHC-Recognizing Receptors on NK Cells Inhibits FcR-Initiated Tyrosine Kinase Activation. Immunity *5*, 629–638.

Biron, C.A., Byron, K.S., and Sullivan, J.L. (1989). Severe Herpesvirus Infections in an Adolescent without Natural Killer Cells. N. Engl. J. Med. *320*, 1731–1735.

Biron, C.A., Nguyen, K.B., Pien, G.C., Cousens, L.P., and Salazar-Mather, T.P. (1999). NATURAL KILLER CELLS IN ANTIVIRAL DEFENSE: Function and Regulation by Innate Cytokines. Annu. Rev. Immunol. *17*, 189–220.

Boyington, J.C., Motyka, S. a, Schuck, P., Brooks, a G., and Sun, P.D. (2000). Crystal structure of an NK cell immunoglobulin-like receptor in complex with its class I MHC ligand. Nature *405*, 537–543.

Brennan, J., Mager, D., Jefferies, W., and Takei, F. (1994). Expression of Different Members of the Ly-49 Gene Family Defines Distinct Nautral Killer Cell Subsets and Cell Adhesion Properties. J. Exp. Med. *180*, 2287–2295.

Brennan, J., Lemieux, S., Freeman, J.D., Mager, D.L., and Takei, F. (1996). Heterogeneity among Ly-49C natural killer (NK) cells: characterization of highly related receptors with differing functions and expression patterns. J. Exp. Med. *184*, 2085–2090.

Brown, M.G., Fulmek, S., Matsumoto, K., Cho, R., Lyons, P. a, Levy, E.R., Scalzo, A. a, and Yokoyama, W.M. (1997). A 2-Mb YAC Contig and Physical Map of the Natural

Killer Gene Complex on Mouse Chromosome 6. Genomics *42*, 16–25.

Burshtyn, D.N., Scharenberg, A.M., Wagtmann, N., Rajagopalan, S., Berrada, K., Yi, T., Kinet, J.-P., and Long, E.O. (1996). Recruitment of Tyrosine Phosphatase HCP by the Killer Cell Inhibitory Receptor. Immunity *4*, 77–85.

Burshtyn, D.N., Yang, W., Yi, T., and Long, E.O. (1997). A Novel Phosphotyrosine Motif with a Critical Amino Acid at Position -2 for the SH2 Domain-mediated Activation of the Tyrosine Phosphatase SHP-1. J. Biol. Chem. *272*, 13066–13072.

Campbell, K.S., Dessing, M., Lopez-Botet, M., Cella, M., and Colonna, M. (1996). Tyrosine phosphorylation of a human killer inhibitory receptor recruits protein tyrosine phosphatase 1C. J. Exp. Med. *184*, 93–100.

Carlyle, J.R., Mesci, A., Fine, J.H., Chen, P., Bélanger, S., Tai, L.-H., and Makrigiannis, A.P. (2008). Evolution of the Ly49 and Nkrp1 recognition systems. Semin. Immunol. *20*, 321–330.

Chakalova, L., Debrand, E., Mitchell, J. a, Osborne, C.S., and Fraser, P. (2005). Replication and transcription: shaping the landscape of the genome. Nat. Rev. Genet. *6*, 669–677.

Chan, P.Y., and Takei, F. (1989). Molecular cloning and characterization of a novel murine T cell surface antigen, YE1/48. J. Immunol. *142*, 1727 LP-1736.

Chan, H.-W., Miller, J.S., Moore, M.B., and Lutz, C.T. (2005). Epigenetic Control of Highly Homologous Killer Ig-Like Receptor Gene Alleles. J. Immunol. *175*, 5966–5974.

Chandy, M., Gutiérrez, J.L., Prochasson, P., and Workman, J.L. (2006). SWI/SNF displaces SAGA-acetylated nucleosomes. Eukaryot. Cell *5*, 1738–1747.

Chen, P., Zhao, J., Wang, Y., Wang, M., Long, H., Liang, D., Huang, L., Wen, Z., Li, W., Li, X., et al. (2013). H3.3 actively marks enhancers and primes gene transcription via opening higher-ordered chromatin. Genes Dev. *27*, 2109–2124.

Choi, J.K., and Kim, Y.-J. (2008). Epigenetic regulation and the variability of gene expression. Nat. Genet. *40*, 141–147.

Chung, H.-R., and Vingron, M. (2009). Sequence-dependent nucleosome positioning. J. Mol. Biol. *386*, 1411–1422.

Clapier, C.R., and Cairns, B.R. (2009). The biology of chromatin remodeling complexes. Annu. Rev. Biochem. *78*, 273–304.

Clarkson, M.J., Wells, J.R., Gibson, F., Saint, R., and Tremethick, D.J. (1999). Regions of variant histone His2AvD required for Drosophila development. Nature.

Lo Coco, F., Pisegna, S., and Diverio, D. (1997). The AML1 gene: A transcription factor involved in the pathogenesis of myeloid and lymphoid leukemias. Haematologica *82*, 364–370.

Collings, C.K., Fernandez, A.G., Pitschka, C.G., Hawkins, T.B., and Anderson, J.N.

(2010). Oligonucleotide sequence motifs as nucleosome positioning signals. PLoS One *5*, e10933.

Correa, I., and Raulet, D.H. (1995). Binding of diverse peptides to MHC class I molecules inhibits target cell lysis by activated natural killer cells. Immunity *2*, 61–71.

Correll, S.J., Schubert, M.H., and Grigoryev, S.A. (2012). Short nucleosome repeats impose rotational modulations on chromatin fibre folding. EMBO J. *31*, 2416–2426.

Cui, F., and Zhurkin, V.B.V.B. (2010). Structure-based analysis of DNA sequence patterns guiding nucleosome positioning in vitro. J. Biomol. Struct. Dyn. *27*, 821.

van Daal, A., and Elgin, S.C. (1992). A histone variant, H2AvD, is essential in Drosophila melanogaster. Mol. Biol. Cell.

Davey, C.A., Sargent, D.F., Luger, K., Maeder, A.W., and Richmond, T.J. (2002). Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. J. Mol. Biol. *319*, 1097–1113.

Denning, T.L., Granger, S., Mucida, D., Graddy, R., Leclercq, G., Zhang, W., Honey, K., Rasmussen, J.P., Cheroutre, H., Rudensky, A.Y., et al. (2007). Mouse TCR +CD8 Intraepithelial Lymphocytes Express Genes That Down-Regulate Their Antigen Reactivity and Suppress Immune Responses. J. Immunol. *178*, 4230–4239.

Depatie, C., Lee, S.-H., Stafford, A., Avner, P., Belouchi, A., Gros, P., and Vidal, S.M. (2000). Sequence-Ready BAC Contig, Physical, and Transcriptional Map of a 2-Mb Region Overlapping the Mouse Chromosome 6 Host-Resistance Locus Cmv1. Genomics *66*, 161–174.

Dimasi, N., and Biassoni, R. (2005). Structural and functional aspects of the Ly49 natural killer cell receptors. Immunol Cell Biol *83*, 1–8.

Dorfman, J.R., and Raulet, D.H. (1998). Acquisition of Ly49 receptor expression by developing natural killer cells. J Exp Med *187*, 609–618.

Drew, H.R., and Travers, A.A. (1985). DNA bending and its relation to nucleosome positioning. J. Mol. Biol. *186*, 773–790.

Du, Z., Zheng, H., Huang, B., Ma, R., Wu, J., Zhang, X., He, J., Xiang, Y., Wang, Q., Li, Y., et al. (2017). Allelic reprogramming of 3D chromatin architecture during early mammalian development. Nature *547*, 232–235.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics *21*, 3439–3440.

Egger, G., Liang, G., Aparicio, A., and Jones, P.A. (2004). Epigenetics in human disease and prospects for epigenetic therapy. Nature *429*, 457–463.

Elliott, J.M., Wahle, J.A., and Yokoyama, W.M. (2010). MHC class I–deficient natural killer cells acquire a licensed phenotype after transfer into an MHC class I–sufficient

environment. J. Exp. Med. *207*, 2073–2079.

Faast, R., Thonglairoam, V., Schulz, T.C., Beall, J., Wells, J.R.E., Taylor, H., Matthaei, K., Rathjen, P.D., Tremethick, D.J., and Lyons, I. (2001). Histone variant H2A.Z is required for early mammalian development. Curr. Biol.

Farris, S.D., Rubio, E.D., Moon, J.J., Gombert, W.M., Nelson, B.H., and Krumm, A. (2005). Transcription-induced chromatin remodeling at the c-myc gene involves the local exchange of histone H2A.Z. J. Biol. Chem.

Felsenfeld, G., and Groudine, M. (2003). Controlling the double helix. Nature *421*, 448–453.

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

Flanagin, S., Nelson, J.D., Castner, D.G., Denisenko, O., and Bomsztyk, K. (2008). Microplate-based chromatin immunoprecipitation method, Matrix ChIP: A platform to study signaling of complex genomic events. Nucleic Acids Res. *36*.

Fraser, R.M., Keszenman-Pereyra, D., Simmen, M.W., and Allan, J. (2009). High-resolution mapping of sequence-directed nucleosome positioning on genomic DNA. J. Mol. Biol. *390*, 292–305.

Gabdank, I., Barash, D., and Trifonov, E.N. (2009). Nucleosome DNA bendability matrix (C. elegans). J. Biomol. Struct. Dyn. *26*, 403–411.

Garboczi, D.N., Utz, U., Ghosh, P., Seth, A., Kim, J., VanTienhoven, E.A., Biddison, W.E., and Wiley, D.C. (1996). Assembly, specific binding, and crystallization of a human TCR-alphabeta with an antigenic Tax peptide from human T lymphotropic virus type 1 and the class I MHC molecule HLA-A2. J. Immunol. *157*, 5403 LP-5410.

Gays, F., Martin, K., Kenefeck, R., Aust, J.G., and Brooks, C.G. (2005). Multiple Cytokines Regulate the NK Gene Complex-Encoded Receptor Repertoire of Mature NK Cells and T Cells. J. Immunol. *175*, 2938–2947.

Gays, F., Aust, J.G., Reid, D.M., Falconer, J., Toyama-Sorimachi, N., Taylor, P.R., and Brooks, C.G. (2006). Ly49B Is Expressed on Multiple Subpopulations of Myeloid Cells. J. Immunol. *177*, 5840–5851.

Gays, F., Taha, S., and Brooks, C.G. (2015a). The Distal Upstream Promoter in Ly49 Genes, Pro1, Is Active in Mature NK Cells and T Cells, Does Not Require TATA Boxes, and Displays Enhancer Activity. J. Immunol. *194*, 6068–6081.

Gays, F., Taha, S., and Brooks, C.G. (2015b). The Distal Upstream Promoter in Ly49 Genes, Pro1, Is Active in Mature NK Cells and T Cells, Does Not Require TATA Boxes, and Displays Enhancer Activity. J. Immunol. *194*, 6068–6081.

Gershenzon, N.I., and Ioshikhes, I.P. (2005). Promoter classifier: software package for

promoter database analysis. Appl. Bioinformatics *4*, 205–209.

Gosselin, P., Mason, L.H., Willette-Brown, J., Ortaldo, J.R., McVicar, D.W., and Anderson, S.K. (1999). Induction of DAP12 phosphorylation, calcium mobilization, and cytokine secretion by Ly49H. J. Leukoc. Biol. *66*, 165–171.

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

Guillemette, B., Bataille, A.R., Gévry, N., Adam, M., Blanchette, M., Robert, F., and Gaudreau, L. (2005). Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. PLoS Biol. *3*, e384.

Hahsler, M., Grün, B., and Hornik, K. (2005). arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. J. Stat. Softw. *14*.

Han, M., and Grunstein, M. (1988). Nucleosome loss activates yeast downstream promoters in vivo. Cell.

Hanke, T., Takizawa, H., McMahon, C.W., Busch, D.H., Pamer, E.G., Miller, J.D., Altman, J.D., Liu, Y., Cado, D., Lemonnier, F. a., et al. (1999). Direct assessment of MHC class I binding by seven Ly49 inhibitory NK cell receptors. Immunity *11*, 67–77.

Hannenhalli, S. (2008). Eukaryotic transcription factor binding sites--modeling and integrative search methods. Bioinformatics *24*, 1325–1331.

Hara, T., Nishimura, H., Hasegawa, Y., and Yoshikai, Y. (2001). Thymus-dependent modulation of Ly49 inhibitory receptor expression on NK1.1+gamma/delta T cells. Immunology *102*, 24–30.

Harp, J.M., Hanson, B.L., Timm, D.E., and Bunick, G.J. (2000). Asymmetries in the nucleosome core particle at 2.5 Å resolution. Acta Crystallogr. Sect. D Biol. Crystallogr.

Hartigan, J.A., and Hartigan, P.M. (1985). The Dip Test of Unimodality. Ann. Stat. *13*, 70–84.

Held, W., and Kunz, B. (1998). An allele-specific, stochastic gene expression process controls the expression of multipleLy49family genes and generates a diverse, MHC-specific NK cell receptor repertoire. Eur. J. Immunol. *28*, 2407–2416.

Held, W., and Raulet, D.H. (1997). Expression of theLy49A gene in murine natural killer cell clones is predominantly but not exclusively mono-allelic. Eur. J. Immunol. *27*, 2876–2884.

Held, W., Roland, J., Raulet, D.H., and H. (1995). Allelic exclusion of Ly49-family genes encoding class I MHC-specific receptors on NK cells.

Held, W., Kunz, B., Lowin-Kropf, B., Van de Wetering, M., and Clevers, H. (1999). Clonal acquisition of the Ly49A NK cell receptor is dependent on the trans-acting factor TCF-1. Immunity *11*, 433–442.

Henagan, T.M., Stefanska, B., Fang, Z., Navard, A.M., Ye, J., Lenard, N.R., and

Devarshi, P.P. (2015). Sodium butyrate epigenetically modulates high-fat diet-induced skeletal muscle mitochondrial adaptation, obesity and insulin resistance through nucleosome positioning. Br. J. Pharmacol. *172*, 2782–2798.

Hu, G., Cui, K., Northrup, D., Liu, C., Wang, C., Tang, Q., Ge, K., Levens, D., Crane-Robinson, C., and Zhao, K. (2013). H2A.Z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation. Cell Stem Cell *12*, 180–192.

Huang, D.W., Sherman, B.T., and Lempicki, R. a (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Huberman, J.A. (1973). Structure of Chromosome Fibers and Chromosomes. Annu. Rev. Biochem. *42*, 355–378.

Imbalzano, A.N., Kwon, H., Green, M.R., and Kingston, R.E. (1994). Facilitated binding of TATA-binding protein to nucleosomal DNA. Nature *370*, 481–485.

Ioshikhes, I., Bolshoy, A., and Trifonov, E.N. (1992). Preferred positions of AA and TT dinucleotides in aligned nucleosomal DNA sequences. J. Biomol. Struct. Dyn. *9*, 1111–1117.

Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M., and Trifonov, E.N. (1996). Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. J. Mol. Biol. *262*, 129–139.

Ioshikhes, I., Trifonov, E.N., and Zhang, M.Q. (1999). Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. Proc. Natl. Acad. Sci. U. S. A. *96*, 2891–2895.

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

Ioshikhes, I.P., Hosid, S., and Pugh, B.F. (2011). Variety of genomic DNA patterns for nucleosome positioning. Genome Res. *21*, 1863–1871.

Iouzalen, N., Moreau, J., and Méchali, M. (1996). H2A.ZI, a new variant histone expressed during Xenopus early development exhibits several distinct features from the core histone H2A. Nucleic Acids Res.

Ito, T., Bulger, M., Pazin, M.J., Kobayashi, R., and Kadonaga, J.T. (1997). ACF, an ISWI-containing and ATP-utilizing chromatin assembly and remodeling factor. Cell *90*, 145–155.

Jackson, J.D., and Gorovsky, M.A. (2000). Histone H2A.Z has a conserved function that is distinct from that of the major H2A sequence variants. Nucleic Acids Res.

Johnson, S.M., Tan, F.J., McCullough, H.L., Riordan, D.P., and Fire, A.Z. (2006). Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. Genome Res.

Joncker, N.T., Shifrin, N., Delebecque, F., and Raulet, D.H. (2010). Mature natural killer cells reset their responsiveness when exposed to an altered MHC environment. J. Exp. Med. *207*, 2065–2072.

Jonsson, A.H., Yang, L., Kim, S., Taffner, S.M., and Yokoyama, W.M. (2010). Effects of MHC Class I Alleles on Licensing of Ly49A+ NK Cells. J. Immunol.

Kamogawa-Schifter, Y., Ohkawa, J., Namiki, S., Arai, N., Arai, K.I., and Liu, Y. (2005). Ly49Q defines 2 pDC subsets in mice. Blood *105*, 2787–2792.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., et al. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362–366.

Karlhofer, F.M., Ribaudo, R.K., and Yokoyama, W.M. (1992). MHC class I alloantigen specificity of Ly-49+ IL-2-activated natural killer cells.

Kärre, K., Ljunggren, H.G., Piontek, G., and Kiessling, R. (1986). Selective rejection of H–2-deficient lymphoma variants suggests alternative immune defence strategy. Nature *319*, 675–678.

Kel, A.E., Gössling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V, and Wingender, E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. *31*, 3576–3579.

Kiessling, R., Klein, E., and Wigzell, H. (1975). „Natural" killer cells in the mouse. I. Cytotoxic cells with specificity for mouse Moloney leukemia cells. Specificity and distribution according to genotype. Eur. J. Immunol. *5*, 112–117.

Kim, S., Iizuka, K., Kang, H.-S.P., Dokun, A., French, A.R., Greco, S., and Yokoyama, W.M. (2002). In vivo developmental stages in murine natural killer cell maturation. Nat. Immunol. *3*, 523–528.

Kim, S., Poursine-Laurent, J., Truscott, S.M., Lybarger, L., Song, Y.-J., Yang, L., French, A.R., Sunwoo, J.B., Lemieux, S., Hansen, T.H., et al. (2005a). Licensing of natural killer cells by host major histocompatibility complex class I molecules. Nature *436*, 709–713.

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M. a, Richmond, T. a, Wu, Y., Green, R.D., and Ren, B. (2005b). A high-resolution map of active promoters in the human genome. Nature *436*, 876–880.

Kogan, S.B., Kato, M., Kiyama, R., and Trifonov, E.N. (2006). Sequence structure of human nucleosome DNA. J. Biomol. Struct. Dyn. *24*, 43–48.

Koh, C.Y., Blazar, B.R., George, T., Welniak, L. a, Capitini, C.M., Raziuddin, a, Murphy, W.J., and Bennett, M. (2001). Augmentation of antitumor effects by NK cell inhibitory receptor blockade in vitro and in vivo. Blood *97*, 3132–3137.

Kornberg, R.D. (1974). Chromatic structure: a repeating unit of histones and DNA.

Science (80-.). *184*, 868.

Kornberg, R.D., and Lorch, Y. (1999). Twenty-Five Years of the Nucleosome, Fundamental Particle of the Eukaryote Chromosome. Cell *98*, 285–294.

Kornberg, R.D., and Stryer, L. (1988). Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. Nucleic Acids Res. *16*, 6677–6690.

Kornberg, R.D., and Thomas, J.O. (1974). Chromatin Structure : Oligomers of the Histones. Science (80-.). *184*, 865–868.

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

Kristell, C., Westholm, J.O., Olsson, I., Ronne, H., Komorowski, J., and Bjerling, P. (2010). Nitrogen depletion in the fission yeast Schizosaccharomyces pombe causes nucleosome loss in both promoters and coding regions of activated genes. Genome Res.

Krogan, N.J., Keogh, M.C., Datta, N., Sawa, C., Ryan, O.W., Ding, H., Haw, R.A., Pootoolal, J., Tong, A., Canadien, V., et al. (2003). A Snf2 Family ATPase Complex Required for Recruitment of the Histone H2A Variant Htz1. Mol. Cell.

Kubo, S., Nagasawa, R., Nishimura, H., Shigemoto, K., and Maruyama, N. (1999). ATF-2-binding regulatory element is responsible for the Ly49A expression in murine T lymphoid line, EL-4. Biochim. Biophys. Acta - Gene Struct. Expr. *1444*, 191–200.

Kubota, A., Kubota, S., Lohwasser, S., Mager, D.L., and Takei, F. (1999a). Diversity of NK cell receptor repertoire in adult and neonatal mice. J. Immunol. *163*, 212–216.

Kubota, a, Kubota, S., Lohwasser, S., Mager, D.L., and Takei, F. (1999b). Diversity of NK cell receptor repertoire in adult and neonatal mice. J. Immunol. *163*, 212–216.

Kwon, C.S., and Wagner, D. (2007). Unwinding chromatin for development and growth: a few genes at a time. Trends Genet. *23*, 403–412.

Lanier, L.L., Corliss, B.C., Wu, J., Leong, C., and Phillips, J.H. (1998). Immunoreceptor DAP12 bearing a tyrosine-based activation motif is involved in activating NK cells. Nature *391*, 703–707.

Larochelle, M., and Gaudreau, L. (2003). H2A.Z has a function reminiscent of an activator required for preferential binding to intergenic DNA. EMBO J.

Leung, A., Cheema, M., González-Romero, R., Eirin-Lopez, J.M., Ausió, J., and Nelson, C.J. (2016). Unique yeast histone sequences influence octamer and nucleosome stability. FEBS Lett. *590*, 2629–2638.

Levine, M. (2011). Paused RNA Polymerase II as a Developmental Checkpoint. Cell *145*, 502–511.

Levitt, M. (1978). How many base-pairs per turn does DNA have in solution and in chromatin? Some theoretical calculations. Proc. Natl. Acad. Sci. U. S. A. *75*, 640–644.

Lewin, B. (1994). Chromatin and gene expression: Constant questions, but changing answers. Cell *79*, 397–406.

Li, E. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. Nat. Rev. Genet. *3*, 662–673.

Li, B., Pattenden, S.G., Lee, D., Gutiérrez, J., Chen, J., Seidel, C., Gerton, J., and Workman, J.L. (2005). Preferential occupancy of histone variant H2AZ at inactive promoters influences local histone modifications and chromatin remodeling. Proc. Natl. Acad. Sci. U. S. A. *102*, 18385–18390.

Li, Z., Gadue, P., Chen, K., Jiao, Y., Tuteja, G., Schug, J., Li, W., and Kaestner, K.H. (2012). Foxa2 and H2A.Z Mediate Nucleosome Depletion during Embryonic Stem Cell Differentiation. Cell *151*, 1608–1616.

Liao, N., Bix, M., Zijlstra, M., Jaenisch, R., and Raulet, D. (1991). MHC class I deficiency: susceptibility to natural killer (NK) cells and impaired NK activity. Science (80-.). *253*, 199–202.

Lickwar, C.R., Rao, B., Shabalin, A.A., Nobel, A.B., Strahl, B.D., and Lieb, J.D. (2009). The Set2/Rpd3S pathway suppresses cryptic transcription without regard to gene length or transcription frequency. PLoS One *4*, e4886.

Lindstrom, K.C., Vary, J.C., Parthun, M.R., Delrow, J., and Tsukiyama, T. (2006). Isw1 Functions in Parallel with the NuA4 and Swr1 Complexes in Stress-Induced Gene Repression. Mol. Cell. Biol.

Ljunggren, H.-G., and Kärre, K. (1990). In search of the "missing self": MHC molecules and NK cell recognition. Immunol. Today *11*, 237–244.

Loch, S., and Tampé, R. (2005). Viral evasion of the MHC class I antigen-processing machinery. Pflugers Arch. Eur. J. Physiol. *451*, 409–417.

Lohr, D.E. (1981). Detailed analysis of the nucleosomal organization of transcribed DNA in yeast chromatin. Biochemistry.

Lowary, P.T., and Widom, J. (1998). New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. J. Mol. Biol. *276*, 19–42.

Lu, Q., Wallrath, L.L., and Elgin, S.C. (1995). The role of a positioned nucleosome at the Drosophila melanogaster hsp26 promoter. EMBO J.

Lubliner, S., and Segal, E. (2009). Modeling interactions between adjacent nucleosomes improves genome-wide predictions of nucleosome occupancy. In Bioinformatics, pp. 348–355.

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

Maeda, M., Lohwasser, S., Yamamura, T., and Takei, F. (2001). Regulation of NKT Cells by Ly49: Analysis of Primary NKT Cells and Generation of NKT Cell Line. J. Immunol. *167*, 4180–4186.

Makrigiannis, A.P., and Anderson, S.K. (2000). Ly49 gene expression in different inbred mouse strains. Immunol. Res. *21*, 39–47.

Makrigiannis, A.P., Etzler, J., Winkler-Pickett, R., Mason, A., Ortaldo, J.R., and Anderson, S.K. (2000). Identification of the Ly49L protein: evidence for activating counterparts to inhibitory Ly49 proteins. J. Leukoc. Biol. *68*, 765–771.

Makrigiannis, A.P., Pau, A.T., Schwartzberg, P.L., McVicar, D.W., Beck, T.W., and Anderson, S.K. (2002). A BAC Contig Map of the Ly49 Gene Cluster in 129 Mice Reveals Extensive Differences in Gene Content Relative to C57BL/6 Mice. Genomics *79*, 437–444.

Makrigiannis, a P., Patel, D., Goulet, M.-L., Dewar, K., and Anderson, S.K. (2005). Direct sequence comparison of two divergent class I MHC natural killer cell receptor haplotypes. Genes Immun. *6*, 71–83.

Marygold, S.J., Leyland, P.C., Seal, R.L., Goodman, J.L., Thurmond, J., Strelets, V.B., and Wilson, R.J. (2013). FlyBase: improvements to the bibliography. Nucleic Acids Res. *41*, D751-7.

Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., et al. (2016). JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *44*, D110–D115.

Matys, V., Kel-Margoulis, O. V, Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. *34*, D108-10.

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

Mavrich, T.N., Jiang, C., Ioshikhes, I.P., Li, X., Venters, B.J., Zanton, S.J., Tomsho, L.P., Qi, J., Glaser, R.L., Schuster, S.C., et al. (2008b). Nucleosome organization in the Drosophila genome. Nature *453*, 358–362.

McBurney, K.L., Leung, A., Choi, J.K., Martin, B.J.E., Irwin, N.A.T., Bartke, T., Nelson, C.J., and Howe, L.A.J. (2016). Divergent residues within Histone H3 dictate a unique chromatin structure in Saccharomyces Cerevisiae. Genetics.

McQueen, K.L., Freeman, J.D., Takei, F., and Mager, D.L. (1998). Localization of five new Ly49 genes, including three closely related to Ly49c. Immunogenetics *48*, 174–183.

McQueen, K.L., Lohwasser, S., Takei, F., and Mager, D.L. (1999). Expression analysis

of new Ly49 genes: most transcripts of Ly49j lack the transmembrane domain. Immunogenetics *49*, 685–691.

McQueen, K.L., Wilhelm, B.T., Takei, F., and Mager, D.L. (2001). Functional analysis of 5' and 3' regions of the closely related Ly49c and j genes. Immunogenetics *52*, 212–223.

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y.M., Buso, N., Cowley, A.P., and Lopez, R. (2013). Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res. *41*, W597-600.

Meneghini, M.D., Wu, M., and Madhani, H.D. (2003). Conserved Histone Variant H2A.Z Protects Euchromatin from the Ectopic Spread of Silent Heterochromatin. Cell *112*, 725–736.

Middleton, D., and Gonzelez, F. (2010). The extensive polymorphism of KIR genes. Immunology *129*, 8–19.

Miele, V., Vaillant, C., D'Aubenton-Carafa, Y., Thermes, C., and Grange, T. (2008). DNA physical properties determine nucleosome occupancy from yeast to fly. Nucleic Acids Res. *36*, 3746–3756.

Montgomery, K.T., Lee, E., Miller, A., Lau, S., Shim, C., Decker, J., Chiu, D., Emerling, S., Sekhon, M., Kim, R., et al. (2001). A high-resolution map of human chromosome 12. Nature *409*, 945–946.

Morozov, A. V., Fortney, K., Gaykalova, D.A., Studitsky, V.M., Widom, J., and Siggia, E.D. (2009). Using DNA mechanics to predict in vitro nucleosome positions and formation energies. Nucleic Acids Res.

Nelson, J.D., Denisenko, O., and Bomsztyk, K. (2006). Protocol for the fast chromatin immunoprecipitation (ChIP) method. Nat. Protoc. *1*, 179–185.

Nikolova, E.N., Kim, E., Wise, A. a., O'Brien, P.J., Andricioaei, I., and Al-Hashimi, H.M. (2011). Transient Hoogsteen base pairs in canonical duplex DNA. Nature *470*, 498–502.

Nunn, M.E., Djeu, J.Y., Glaser, M., Lavrin, D.H., and Herberman, R.B. (1976). Natural Cytotoxic Reactivity of Rat Lymphocytes Against Syngeneic Gross Virus-Induced Lymphoma2. JNCI J. Natl. Cancer Inst. *56*, 393–399.

Olcese, L., Lang, P., Vély, F., Cambiaggi, A., Marguet, D., Bléry, M., Hippen, K.L., Biassoni, R., Moretta, A., Moretta, L., et al. (1996). Human and mouse killer-cell inhibitory receptors recruit PTP1C and PTP1D protein tyrosine phosphatases. J. Immunol. *156*, 4531–4534.

Olcese, L., Cambiaggi, A., Semenzato, G., Bottino, C., Moretta, A., and Vivier, E. (1997). Human killer cell activatory receptors for MHC class I molecules are included in a multimeric complex expressed by natural killer cells. J. Immunol. *158*, 5083–5086.

Olins, A.L., and Olins, D.E. (1974). Spheroid chromatin units (ν bodies). Science (80-.). *183*, 330.

Orange, J.S. (2002). Human natural killer cell deficiencies and susceptibility to infection. Microbes Infect. *4*, 1545–1558.

Ortaldo, J.R., Winkler-Pickett, R., Mason, A.T., and Mason, L.H. (1998). The Ly-49 Family: Regulation of Cytotoxicity and Cytokine Production in Murine CD3⁺ Cells. J. Immunol. *160*, 1158 LP-1165.

Ortaldo, J.R., Mason, a T., Winkler-Pickett, R., Raziuddin, a, Murphy, W.J., and Mason, L.H. (1999). Ly-49 receptor expression and functional analysis in multiple mouse strains. J. Leukoc. Biol. *66*, 512–520.

Ozsolak, F., Song, J.S., Liu, X.S., and Fisher, D.E. (2007). High-throughput mapping of the chromatin structure of human promoters. Nat. Biotechnol.

Park, Y.J., Dyer, P.N., Tremethick, D.J., and Luger, K. (2004). A new fluorescence resonance energy transfer approach demonstrates that the histone variant H2AZ stabilizes the histone octamer within the nucleosome. J. Biol. Chem.

Pascal, V.V., Stulberg, M.J., and Anderson, S.K. (2006). Regulation of class I major histocompatibility complex receptor expression in natural killer cells: one promoter is not enough! Immunol. Rev. *214*, 9–21.

Peckham, H.E., Thurman, R.E., Fu, Y., Stamatoyannopoulos, J.A., Noble, W.S., Struhl, K., and Weng, Z. (2007). Nucleosome positioning signals in genomic DNA. Genome Res. *17*, 1170–1177.

Pruss, D., and Wolffe, A.P. (1993). Histone-DNA contacts in a nucleosome core containing a Xenopus 5S rRNA gene. Biochemistry *32*, 6810–6814.

Quinlan, A.R., and Hall, I.M. (2010a). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Quinlan, A.R., and Hall, I.M. (2010b). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

R Development Core Team (2012). R: A language and environment for statistical computing. (R Foundation for Statistical Computing).

Raisner, R.M., Hartley, P.D., Meneghini, M.D., Bao, M.Z., Liu, C.L., Schreiber, S.L., Rando, O.J., and Madhani, H.D. (2005). Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin. Cell *123*, 233–248.

Rajagopalan, S., and Long, E.O. (2005). Viral evasion of NK-cell activation. Trends Immunol. *26*, 403–405.

Raulet, D.H. (2004). Interplay of natural killer cells and their receptors with the adaptive immune response. Nat. Immunol. *5*, 996–1002.

Raulet, D.H., and Vance, R.E. (2006). Self-tolerance of natural killer cells. Nat. Rev.

Immunol. *6*, 520–531.

Raulet, D.H., Held, W., Correa, I., Dorfman, J.R., Wu, M.F., and Corral, L. (1997). Specificity, tolerance and developmental regulation of natural killer cells defined by expression of class I-specific Ly49 receptors. Immunol. Rev. *155*, 41–52.

Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A.M. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc. Natl. Acad. Sci.

Richmond, T.J., and Davey, C.A. (2003). The structure of DNA in the nucleosome core. Nature *423*, 145–150.

Robinson, P.J.J., Fairall, L., Huynh, V.A.T., and Rhodes, D. (2006). EM measurements define the dimensions of the "30-nm" chromatin fiber: Evidence for a compact, interdigitated structure. Proc. Natl. Acad. Sci. *103*, 6506–6511.

Romagné, F., André, P., Spee, P., Zahn, S., Anfossi, N., Gauthier, L., Capanni, M., Ruggeri, L., Benson, D.M., Blaser, B.W., et al. (2011). therapeutic antibody that augments natural killer − mediated killing of Preclinical characterization of 1-7F9 , a novel human anti – KIR receptor therapeutic antibody that augments natural killer – mediated killing of tumor cells. Hematology *114*, 2667–2677.

Rouhi, A., Gagnier, L., Takei, F., and Mager, D.L. (2006). Evidence for Epigenetic Maintenance of Ly49a Monoallelic Gene Expression. J. Immunol. *176*, 2991–2999.

Rouhi, A., Lai, C.B., Cheng, T.P., Takei, F., Yokoyama, W.M., and Mager, D.L. (2009). Evidence for high bi-allelic expression of activating Ly49 receptors. Nucleic Acids Res. *37*, 5331–5342.

Roy, S., Ernst, J., Kharchenko, P. V, Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C. a, Ma, L., Lin, M.F., et al. (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science (80-.). *330*, 1787–1797.

Sadeh, R., and Allis, C.D. (2011). Genome-wide "re"-modeling of nucleosome positions. Cell *147*, 263–266.

Saleh, A., Makrigiannis, A.P., Hodge, D.L., and Anderson, S.K. (2002). Identification of a Novel Ly49 Promoter That Is Active in Bone Marrow and Fetal Thymus. J. Immunol. *168*, 5163–5169.

Saleh, A., Davies, G.E., Pascal, V., Wright, P.W., Hodge, D.L., Cho, E.H., Lockett, S.J., Abshari, M., and Anderson, S.K. (2004). Identification of probabilistic transcriptional switches in the Ly49 gene cluster: a eukaryotic mechanism for selective gene activation. Immunity *21*, 55–66.

Salih, F., Salih, B., and Trifonov, E.N. (2008). Sequence structure of hidden 10.4-base repeat in the nucleosomes of C. elegans. J. Biomol. Struct. Dyn. *26*, 273–282.

Santisteban, M.S., Kalashnikova, T., and Smith, M.M. (2000). Histone H2A.Z Regulates Transcription and Is Partially Redundant with Nucleosome Remodeling Complexes. Cell *103*, 411–422.

Santourlidis, S., Trompeter, H.-I., Weinhold, S., Eisermann, B., Meyer, K.L., Wernet, P., and Uhrberg, M. (2002). Crucial Role of DNA Methylation in Determination of Clonally Distributed Killer Cell Ig-like Receptor Expression Patterns in NK Cells. J. Immunol. *169*, 4253–4261.

Satchwell, S.C., Drew, H.R., and Travers, A.A. (1986). Sequence periodicities in chicken nucleosome core DNA. J. Mol. Biol. *191*, 659–675.

Scharton-Kersten, T.M., and Sher, A. (1997). Role of natural killer cells in innate resistance to protozoan infections. Curr. Opin. Immunol. *9*, 44–51.

Schenkel, A.R., Kingry, L.C., and Slayden, R.A. (2013). The Ly49 gene family. A brief guide to the nomenclature, genetics, and role in intracellular infection. Front. Immunol. *4*, 1–8.

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

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

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.-P.Z., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772–778.

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., and Gaul, U. (2008). Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature *451*, 535–540.

Siewiera, J., El Costa, H., Tabiasco, J., Berrebi, A., Cartron, G., Bouteiller, P., and Jabrane-Ferrat, N. (2013). Human Cytomegalovirus Infection Elicits New Decidual Natural Killer Cell Effector Functions. PLoS Pathog. *9*.

Silver, E.T., Gong, D., Hazes, B., and Kane, K.P. (2001). Ly-49W, an activating receptor of nonobese diabetic mice with close homology to the inhibitory receptor Ly-49G, recognizes H-2D(k) and H-2D(d). J. Immunol. *166*, 2333–2341.

Sivakumar, P. V, Gunturi, A., Salcedo, M., Schatzle, J.D., Lai, W.C., Kurepa, Z., Pitcher, L., Seaman, M.S., Lemonnier, F.A., Bennett, M., et al. (1999). Cutting edge: expression of functional CD94/NKG2A inhibitory receptors on fetal NK1.1+Ly-49- cells: a possible mechanism of tolerance during NK cell development. J. Immunol. *162*, 6976–6980.

Smale, S.T., and Fisher, A.G. (2002). Chromatin structure and gene regulation in the immune system. Annu. Rev. Immunol. *20*, 427–462.

Smith, H.R., Karlhofer, F.M., and Yokoyama, W.M. (1994). Ly-49 multigene family expressed by IL-2-activated NK cells. J. Immunol. *153*, 1068–1079.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

Struhl, K. (1999). Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. Cell *98*, 1–4.

Suto, R.K., Clarkson, M.J., Tremethick, D.J., and Luger, K. (2000). Crystal structure of a nucleosome core particle containing the variant histone H2A.Z. Nat. Struct. Biol. *7*, 1121–1124.

Swaminathan, J., Baxter, E.M., and Corces, V.G. (2005). The role of histone H2Av variant replacement and histone H4 acetylation in the establishment of Drosophila heterochromatin. Genes Dev. *19*, 65–76.

Tagami, H., Ray-Gallet, D., Almouzni, G., and Nakatani, Y. (2004). Histone H3.1 and H3.3 Complexes Mediate Nucleosome Assembly Pathways Dependent or Independent of DNA Synthesis. Cell *116*, 51–61.

Talbert, P.B., and Henikoff, S. (2010). Histone variants--ancient wrap artists of the epigenome. Nat. Rev. Mol. Cell Biol. *11*, 264–275.

Tanaka, T., Tanaka, K., Ogawa, S., Kurokawa, M., Mitani, K., Yazaki, Y., Shibata, Y., and Hirai, H. (1997). An acute myeloid leukemia gene, AML1, regulates transcriptional activation and hemopoietic myeloid cell differentiation antagonistically by two alternative spliced forms. Leukemia *11 Suppl 3*, 299–302.

Tanamachi, D.M., Moniot, D.C., Cado, D., Liu, S.D., Hsia, J.K., and Raulet, D.H. (2004). Genomic Ly49A Transgenes: Basis of Variegated Ly49A Gene Expression and Identification of a Critical Regulatory Element. J. Immunol. *172*, 1074–1082.

Tarek, N., Luduec, J.B. Le, Gallagher, M.M., Zheng, J., Venstrom, J.M., Chamberlain, E., Modak, S., Heller, G., Dupont, B., Cheung, N.K. V, et al. (2012). Unlicensed NK cells target neuroblastoma following anti-GD2 antibody treatment. J. Clin. Invest. *122*, 3260–3270.

Tassi, I., Le Friec, G., Gilfillan, S., Takai, T., Yokohama, W.M., and Colonna, M. (2009). DAP10 associates with Ly49 receptors but contributes minimally to their expression and function in vivo. Eur. J. Immunol. *39*, 1129–1135.

Teif, V.B., and Bohinc, K. (2011). Condensed DNA: Condensing the concepts. Prog. Biophys. Mol. Biol. *105*, 208–222.

Teves, S.S., and Henikoff, S. (2011). Heat shock reduces stalled RNA polymerase II and nucleosome turnover genome-wide. Genes Dev. *25*, 2387–2397.

Thakar, A., Gupta, P., Ishibashi, T., Finn, R., Silva-Moreno, B., Uchiyama, S., Fukui, K., Tomschik, M., Ausio, J., and Zlatanova, J. (2009). H2A.Z and H3.3 histone variants

affect nucleosome structure: biochemical and biophysical studies. Biochemistry *48*, 10852–10857.

Thoma, F., and Acta, B. (1992). Nucleosome positioning. Biochim. Biophys. Acta - Gene Struct. Expr. *1130*, 1–19.

Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D., and Van Helden, J. (2012). RSAT peak-motifs: Motif analysis in full-size ChIP-seq datasets. Nucleic Acids Res. *40*.

Tillo, D., and Hughes, T.R. (2009). G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics *10*, 442.

Tillo, D., Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Field, Y., Lieb, J.D., Widom, J., Segal, E., and Hughes, T.R. (2010). High nucleosome occupancy is encoded at human regulatory sequences. PLoS One *5*, e9129.

To, K., Agrotis, A., Besra, G., Bobik, A., and Toh, B.H. (2009). NKT cell subsets mediate differential proatherogenic effects in ApoE -/- mice. Arterioscler. Thromb. Vasc. Biol. *29*, 671–677.

Tolstorukov, M., Colasanti, A., and McCandlish, D. (2007). A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. J. Mol. Biol. *371*, 725–738.

Toyama-Sorimachi, N., Tsujimura, Y., Maruya, M., Onoda, A., Kubota, T., Koyasu, S., Inaba, K., and Karasuyama, H. (2004). Ly49Q, a member of the Ly49 family that is selectively expressed on myeloid lineage cells and involved in regulation of cytoskeletal architecture. Proc. Natl. Acad. Sci. *101*, 1016–1021.

Travers, A., Caserta, M., Churcher, M., Hiriart, E., and Di Mauro, E. (2009). Nucleosome positioning--what do we really know? Mol. Biosyst. *5*, 1582–1592.

Trinchieri, G. (1989). Biology of natural killer cells. Adv. Immunol. *47*, 187–376.

Truong, D.M., and Boeke, J.D. (2017). Resetting the yeast epigenome with human nucleosomes. bioRxiv.

Uhrberg, M. (2005). Shaping the human NK cell repertoire: An epigenetic glance at KIR gene regulation. Mol. Immunol. *42*, 471–475.

Unanue, E.R. (1997). Studies in listeriosis show the strong symbiosis between the innate cellular system and the T-cell response. Immunol. Rev. *158*, 11–25.

Valiante, N.M., Uhrberg, M., Shilling, H.G., Lienert-Weidenbach, K., Arnett, K.L., D'Andrea, A., Phillips, J.H., Lanier, L.L., and Parham, P. (1997). Functionally and Structurally Distinct NK Cell Receptor Repertoires in the Peripheral Blood of Two Human Donors. Immunity *7*, 739–751.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., et al. (2008). A high-resolution, nucleosome

position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res. *18*, 1051–1063.

Vély, F., Olivero, S., Olcese, L., Moretta, A., Damen, J.E., Liu, L., Krystal, G., Cambier, J.C., Daëron, M., and Vivier, E. (1997). Differential association of phosphatases with hematopoietic co-receptors bearing immunoreceptor tyrosine-based inhibition motifs. Eur. J. Immunol. *27*, 1994–2000.

Vey, N., Bourhis, J.-H., Boissel, N., Bordessoule, D., Prebet, T., Charbonnier, a., Etienne, a., Andre, P., Romagne, F., Benson, D., et al. (2012). A phase I trial of the anti-inhibitory KIR monoclonal antibody IPH2101 for acute myeloid leukemia (AML) in complete remission. Blood *120*, 1–3.

Wagtmann, N., Biassoni, R., Cantoni, C., Verdiani, S., Malnati, M.S., Vitale, M., Bottino, C., Moretta, L., Moretta, A., and Long, E.O. (1995). Molecular clones of the p58 nk cell-receptor reveal immunoglobulin- related molecules with diversity in both the extracellular and intracellular domains. Immunity *2*, 439–449.

Wang, J., and Xi, L. (2012). A introduction on NuPoP R package NuPoP functions. 7–9.

Wang, A.Y., Aristizabal, M.J., Ryan, C., Krogan, N.J., and Kobor, M.S. (2011). Key Functional Regions in the Histone Variant H2A.Z C-Terminal Docking Domain. Mol. Cell. Biol. *31*, 3871–3884.

Wang, J.P., Fondufe-Mittendorf, Y., Xi, L., Tsai, G.F., Segal, E., and Widom, J. (2008). Preferentially quantized linker DNA lengths in Saccharomyces cerevisiae. PLoS Comput. Biol. *4*.

Wang, S., Su, J., Beliveau, B.J., Bintu, B., Moffitt, J.R., Wu, C., and Zhuang, X. (2016). Spatial organization of chromatin domains and compartments in single chromosomes. Science (80-.). *353*, 598–602.

Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C.T., et al. (2010). The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. *38*, 214–220.

Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. *5*, 276–287.

Weber, C.M., Henikoff, J.G., and Henikoff, S. (2010). H2A.Z nucleosomes enriched over active genes are homotypic. Nat. Struct. Mol. Biol. *17*, 1500–1507.

Weiner, A., Hughes, A., Yassour, M., Rando, O.J., and Friedman, N. (2010). High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. Genome Res. 90–100.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis (Springer New York).

Wilhelm, B.T., Gagnier, L., and Mager, D.L. (2002). Sequence analysis of the ly49

cluster in C57BL/6 mice: a rapidly evolving multigene family in the immune system. Genomics *80*, 646–661.

Wong, H., Victor, J.-M., Mozziconacci, J., Liu, A., and Stevens, M. (2007). An All-Atom Model of the Chromatin Fiber Containing Linker Histones Reveals a Versatile Structure Tuned by the Nucleosomal Repeat Length. PLoS One *2*, e877.

Wyrick, J.J., Holstege, F.C., Jennings, E.G., Causton, H.C., Shore, D., Grunstein, M., Lander, E.S., and Young, R. a (1999). Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. Nature *402*, 418–421.

Xi, L., Fondufe-Mittendorf, Y., Xia, L., Flatow, J., Widom, J., and Wang, J. (2010). Predicting nucleosome positioning using a duration Hidden Markov Model. BMC Bioinformatics *11*, 346.

Xie, W.J., Meng, L., Liu, S., Zhang, L., Cai, X., and Gao, Y.Q. (2017). Structural Modeling of Chromatin Integrates Genome Features and Reveals Chromosome Folding Principle. Sci. Rep. *7*, 2818.

Yadi, H., Burke, S., Madeja, Z., Hemberger, M., Moffett, A., and Colucci, F. (2008). Unique Receptor Repertoire in Mouse Uterine NK cells. J. Immunol. *181*, 6140–6147.

Yankulov, K. (2015). Book review: Epigenetics (second edition, eds. Allis, Caparros, Jenuwein, Reinberg). Front. Genet. *6*.

Yokoyama, W.M., and Seaman, W.E. (1993). The Ly-49 and NKR-P1 gene families encoding lectin-like receptors on natural killer cells: the NK gene complex. Annu. Rev. Immunol. *11*, 613–635.

Yokoyama, W.M., Jacobs, L.B., Kanagawa, O., Shevach, E.M., and Cohen, D.I. (1989). A murine T lymphocyte antigen belongs to a supergene family of type II integral membrane proteins. J. Immunol. *143*, 1379–1386.

Yokoyama, W.M., Kehn, P.J., Cohen, D.I., and Shevach, E.M. (1990). Chromosomal location of the Ly-49 (A1, YE1/48) multigene family. Genetic association with the NK 1.1 antigen. J. Immunol. *145*, 2353 LP-2358.

Yuan, G., Liu, Y., Dion, M., Slack, M., Wu, L., and SJ (2005). Genome-scale identification of nucleosome positions in S. cerevisiae. Science (80-.). *309*, 626–630.

Zhang, H., Roberts, D.N., and Cairns, B.R. (2005). Genome-wide dynamics of Htz1, a histone H2A variant that poises repressed/basal promoters for activation through histone loss. Cell *123*, 219–231.

Zhang, Y., Moqtaderi, Z., Rattner, B.P., Euskirchen, G., Snyder, M., Kadonaga, J.T., Liu, X.S., and Struhl, K. (2009). Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. Nat. Struct. Mol. Biol. *16*, 847–852.

Zlatanova, J., and Thakar, A. (2008). H2A.Z: view from the top. Structure *16*, 166–179.

Zuccheri, G., Scipioni, A., Cavaliere, V., Gargiulo, G., De Santis, P., and Samorì, B.

(2001). Nucleosome Positioning Pattern Derived from Oligonucleotide Compositions of Genomic Sequences. Proc. Natl. Acad. Sci. U. S. A. *98*, 3074–3079.

Zucchi, I., Mento, E., Kuznetsov, V.A., Scotti, M., Valsecchi, V., Simionati, B., Vicinanza, E., Valle, G., Pilotti, S., Reinbold, R., et al. (2004). Gene expression profiles of epithelial cells microscopically isolated from a breast-invasive ductal carcinoma and a nodal metastasis. Proc. Natl. Acad. Sci. U. S. A. *101*, 18147–18152.

Contributions of Collaborators

Andrew Wight, a PhD student and one of our collaborators in the Makrigiannis' lab performed the experiments using RMA: preparation of samples for ChIP-Seq, measuring the expression of Ly49 genes in RMA, immunoprecipitation of RMA cells for AML-1 binding sites. He analysed the GC content of nucleosome positioning regions and the chromosome 6, computed the deviancy of nucleosome positioning, and performed the logistic regression and the chi-square test. He drafted the figures depicting these results.

Curriculum Vitae

Professional Experiences

Research Associate, Biologics Resources, LLC, Germantown, MD (2015 – 2017)

- Developing antibodies treating Aplastic Anemia using Phage library
- Process development of recombinant protein production
- Writing SOPs for the production
- Building database to keep the records to comply GMP regulations

Technology Consultant, Valley Crosser Computing, Niagara Falls, ON (2008-2010)

- Gave businesses technical consultation regarding document management.
- Built databases for storing data and managing workflows.

Senior Microbiologist, E3 Laboratories Inc., Niagara-On-The-Lake, ON (2007-2008)

- Ensure test methods and laboratory procedures to comply with the approved standards under Safe Water Act.
- Conducted biological tests for drinking water and wastewater.
- Kept documentation for the QA/QC of the test procedures, wrote SOPs.
- Built a digital document management system for the retention of the test documents according to the regulation.

Researcher, LG Biomedical Institute, Inc., La Jolla, CA (2000-2003)

- Analysed microarray gene expression data using various statistical methods
- Discovered novel cancer targets and diagnostic markers and filed patents.
- Developed a new cloning method of full-length genes from IMAGE library (4x faster than previous methods).
- Cloned novel cancer target genes for validation.
- Verified the cloned genes by DNA sequencing.
- Developed integrated databases of gene information and shortened the reporting time for the candidate genes.
- Developed reporting system to share analysis results and reports internal- and intra-team members.

Research associate, LG Chem, Ltd., Daejun, Korea (1998-2000)

- Purified proteins for NMR, X-ray crystallography, cell-based assay, and pre-clinical tests.
- Provided recombinant proteins for cancer therapy and HCV vaccines.

Research associate, Hanhyo Institute of Technology (1995-1998)

- Optimised high cell density culture reaching up to 40 g/l dry cell weight.
- Purified proteins for characterization, cell-based assays, and pre-clinical tests.
- Cultured influenza virus, identified strains, tested the virus activities.

Education

PhD Candidate, Microbiology and Immunology/Bioinformatics, University of Ottawa, Ottawa, ON (2011 – present)

Title of Thesis: Computational study of interplay between nucleosomes and transcription factors on gene expression
Expected graduation date: Jan 2015

Statistics, San Diego State University, San Diego, CA (2003-2006)
Finished course requirements in the graduate school.

MSc in Microbiology, Seoul National University, Seoul, Korea (1993-1995)
Title of thesis: Temperature effect on the production of beta-lactamase on recombinant *Streptomyces lividans* PD6
Modelled and optimised the culture condition to maximise the product and the plasmid stability.

BSc in Microbiology, Seoul National University, Seoul, Korea (1989-1993)

Work Related Training and certificates

- Foundations of Clinical Research, San Jose, CA 2006
- Visiting Scientist: Data Mining and Statistics for Microarray Data, GeneLogic, Gaithersburg, MD Oct. 1 - Nov. 21, 2001
- Techniques in Bioinformatics and Comparative Genomics, BTCI, Madison, Wisconsin June 18 - 23, 2000
- National Certificate in Information Technology (1st degree), Korea 1997

Awards

- Ontario Graduate Scholarship, 2012
- Admission Scholarship, University of Ottawa 2011
- Admission Scholarship, University of Ottawa 2010

Volunteering

- Let's Talk Science, Ottawa, ON
- Science fair judge.
- Science Fair judge, Canada-Wide Science Fair, Toronto, ON
- Elected council member of graduate student association, University of Ottawa

Publications and Patents

Patents

Gene families associated with cancers, PCT/KR03/02161

The advanced method for the preparation of HCV protease, Application no. 99-18294

Process for Preparing Recombinant Proteins Using Recombinant Saccharomyces cerevisiae, KP 0250804

Publications

Yang, D., and Ioshikhes, I. (2016). Drosophila H2A and H2A.Z Nucleosome Sequences Reveal Different Nucleosome Positioning Sequence Patterns. J. Comput. Biol. cmb.2016.0173.

Wight, A*., Yang, D*., Ioshikhes, I., and Makrigiannis, A.P. (2016). Nucleosome Presence at AML-1 Binding Sites Inversely Correlates with Ly49 Expression:

Revelations from an Informatics Analysis of Nucleosomes and Immune Cell Transcription Factors. PLoS Comput. Biol. 12, 1–19. (* Equal contribution)

Alvarez-Saavedra, M., De Repentigny, Y., Lagali, P.S., Raghu Ram, E.V.S., Yan, K., Hashem, E., Ivanochko, D., Huh, M.S., Yang, D., Mears, A.J., et al. (2014). Snf2h-mediated chromatin organisation and histone H1 dynamics govern cerebellar morphogenesis and neural maturation. Nat. Commun. 5, 4181.

Bae, C.S., Yang, D.S., Chang, K.R., Seong, B.L., and Lee, J. (1998). Enhanced secretion of human granulocyte colony-stimulating factor directed by a novel hybrid fusion peptide from recombinant Saccharomyces cerevisiae at high cell concentration. Biotechnol. Bioeng. 57, 600–609.

Bae, C.S., Yang, D.S., Lee, J., and Park, Y.H. (1999). Improved process for production of recombinant yeast-derived monomeric human G-CSF. Appl. Microbiol. Biotechnol. 52, 338–344.

Lee, J., Choi, S.I., Jang, J.S., Jang, K., Moon, J.W., Bae, C.S., Yang, D.S., and Seong, B.L. (1999). Novel secretion system of recombinant Saccharomyces cerevisiae using an N-terminus residue of human IL-1 beta as secretion enhancer. Biotechnol. Prog. 15, 884–890.

Yang, D.S., Bae, C.S., and Lee, J. (1997). Production of recombinant human granulocyte-colony-stimulating factor in high cell density yeast cultures. Biotechnol. Lett. 19, 655–659.

Temperature effect on the production of beta-lactamase on recombinant Streptomyces lividans PD6 Master Thesis, 1995

Publications in preparation

A. K. M. Firoj Mahmud, Doo Yang, Per Stenberg, Ilya Ioshikhes, Soumyadeep Nandi, Analysis of Drosophila Transcription Factor Interaction Network

Refereed poster presentations

Doo Seok Yang, Ilya Ioshikhes, 2012, Various nucleosome positioning patterns in Drosophila, Translational Bioinformatics Conference

Doo Seok Yang, Ilya Ioshikhes, 2012, Nucleosome positioning is defined by various nucleosome sequence patterns, CCSB and MPSA

Soumyadeep Nandi, Doo Yang, Ilya Ioshikhes, 2012, Identifying the cluster of regulatory elements in proximal promoters of Drosophila melanogaster, CCSB and MPSA

Yuchu Grace Hsiung, Doo Seok Yang, Chung-mi Kim, Jeong-lim Kim, Tae-saeng Choi, Hyun-ho Chung, Sang Seok Koh, 2001, Genomic analysis of tumor progression using transgenic mouse model system, AACR-NCI-EOTC International Conference on Molecular Targets and Cancer Therapeutics.

Yuchu Grace Hsiung, Doo Seok Yang, Chung-mi Kim, Jeong-lim Kim, Tae-saeng Choi, Hyun-ho Chung, Sang Seok Koh, 2002, Potential suppressor of tumorigenesis, AACR annual meeting.

From: "Ballen, KarenDate: Apr 28, 2017 3:18 PM

Subject: RE: LiebertPub Website Customer Question

Dear Doo Yang:

Copyright permission is granted for use of the figures from your article published in JOURNAL OF COMPUTATIONAL BIOLOGY, April 2017, 24/4, pp. 289-298, in your thesis.

Kind regards,

Karen Ballen

Manager, Reprints, Permissions, and Open Access

ELSEVIER LICENSE
TERMS AND CONDITIONS

May 12, 2017

This Agreement between doo yang ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4106541115658
License date	May 12, 2017
Licensed Content Publisher	Elsevier
Licensed Content Publication	Seminars in Immunology
Licensed Content Title	Evolution of the Ly49 and Nkrp1 recognition systems
Licensed Content Author	James R. Carlyle,Aruz Mesci,Jason H. Fine,Peter Chen,Simon Bélanger,Lee-Hwa Tai,Andrew P. Makrigiannis
Licensed Content Date	December 2008
Licensed Content Volume	20
Licensed Content Issue	6
Licensed Content Pages	10
Start Page	321
End Page	330
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Order reference number	
Original figure numbers	Figure 1
Title of your thesis/dissertation	Computational study of nucleosome positioning and the interplay with transcription factors
Expected completion date	May 2017
Estimated size (number of pages)	200
Elsevier VAT number	GB 494 6272 12
Requestor Location	doo yang 4208-451 Smyth Rd. Ottawa, ON K1H 8M5 Canada Attn: doo yang
Total	0.00 CAD
Terms and Conditions	

Creative Commons License Deed

Attribution 4.0 International (CC BY 4.0)

This is a human-readable summary of (and not a substitute for) the license.

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material

for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.