Nucleosome positioning in Arabidopsis

Sarah Louise Usher

Thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

The University of Warwick,
Warwick HRI/Rothamsted Research

September 2009

Dr Graham King (Rothamsted Research)
Dr Smita Kurup (Rothamsted Research)
Dr Guy Barker (Warwick HRI)

This project was carried out under a BBSRC funded strategic studentship.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank the following people, firstly my academic supervisors, Dr. Graham King, Dr. Smita Kurup and Dr. Guy Barker, for all of their support, help and encouragement throughout the PhD. I would also like to thank Paul Verrier and Stephen Powers for help with bioinformatics and statistics. Thanks to all who have helped in my journey who have not been specifically mentioned. Also, thanks to my fellow PhD students and members of the research group at Rothamsted who provided a strong sense of community, and much useful advice and motivation. Thanks to the BBSRC for funding this research.

Declaration

This thesis is the sole work of the author, and no part has previously been published or presented for another degree.

The DNA sequence datasets used throughout this thesis were processed by Graham King as described within Chapter 2.

Warwick HRI's departmental guidelines were used in the preparation of this thesis.

Summary

The aim of this project was to test hypotheses relating to nucleosome positioning in Arabidopsis to provide a basis for better understanding of epigenetic transcriptional regulation in plants. Prior to this study, virtually no information existed regarding nucleosome positioning in plants. Eukaryote chromosomes consist of chromatin, composed of nucleosomes separated by linker DNA of variable lengths. Nucleosomes consist of 147 bp of DNA wrapped 1.7 times around a histone octamer. Whilst no consensus nucleosome positioning DNA sequence exists, sequence preferences influence positioning, and contribute to the complex epigenetic processes which act to control transcriptional activity. These details of the underlying mechanisms are known to differ between the plant and animal kingdoms.

High-throughput sequencing technologies were utilised to generate large datasets of mono- and di-nucleosome sequences from wild-type Arabidopsis. These enabled genome-wide analysis and inference of plant-specific patterns of nucleosome positioning and sequence properties. Further data were generated from a methyltransferase antisense (*MET1*) which is depleted in methylated CG epigenetic marks.

The internal distributions of dinucleotides within Arabidopsis nucleosomes were similar to those observed in non-plant eukaryotes. A unique periodicity in the distribution of linker lengths was detected in Arabidopsis wild type chromatin. In contrast, the *MET1* antisense line displayed the expected periodicity, indicating systematic differences in chromatin organisation. There was a significant increase in nucleosome occupancy within exons compared with introns. However, this difference was less marked in the *MET1* antisense. Specific patterns of nucleosome phasing were observed around transcription start sites. Linker lengths within rRNA gene clusters associated with nucleolar organiser regions (NORs) differed depending on chromosome of origin, suggesting differences in higher order chromatin structure between the NORs. Comparison of the nucleosome position and DNA methylation within the rRNA gene cluster revealed interesting differences between the two regions, which may reflect interactions affecting chromatin structure and transcriptional regulation.

Abbreviations

^{o}C	degrees Celsius
^{5m}C	5-methylcytosine
ACF	Autocorrelation function
Ath	*Arabidopsis thaliana*
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
ChiP	chromatin immunopreciptitation
Chr	Chromosome
CL	Confidence limit
Col	Columbia
CSV	Comma separated variable
DNA	Deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
F	Forward
FISH	Fluorescence *in-situ* hybridisation
g	grams
IGR	Intergenic region
Inr	Initiator
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ITS	Internal transcribed spacer
IUPAC	The International Union of Pure and Applied Chemistry
KDE	Kernel Density Estimation
M	Molar
m	milli (10^{-3})
miRNA	Micro-RNA
Mbp	Mega basepairs
MNase	Micrococcal nuclease
MTE	motif ten element
n	Number
NASC	Nottingham *A. thaliana* Stock Centre
NCBI	National Centre for Biotechnology Information
NFR	Nucleosome free region

NOR	Nucleolar organiser region
NPRD	Nucleosome Position Region database
NRL	Nucleosome repeat length
nt	Nucleotide
PACF	Partial autocorrelation function
PcG	Polycomb group
Pol I	RNA polymerase I
Pol II	RNA polymerase II
R	Reverse
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
SNP	Single nucleotide polymorphism
snRNA	Small nuclear RNA
snoRNA	Small nucleolar RNA
tRNA	Transfer RNA
TAIR	The Arabidopsis Information Resource
TE (buffer)	Tris EDTA buffer
Tr.E	Transposable element
TSS	Transcriptional start site
µ	micro (10^{-6})
U	units
url	Uniform resource locator
UTR	Untranslated region
V	volts
vs.	versus
WT	Wild-type
W	Nucleotides
X-gal	5-bromo-4-chloro-3-indolyl-b-D-galactopyranoside
$\bar{y}$	Arithmetic mean

IUPAC nucleotide codes were adopted throughout this thesis.

IUPAC code	Base abbreviation
A	A
C	C
G	G
T	T
M	A or C
R	A or G
W	A or T
S	C or G
Y	C or T
K	G or T
V	A, C or G
H	A, C or T
D	A, G or T
B	C, G or T
N	A, C, G or T

Preface

The thesis is presented in eight chapters. The background literature and motivation for the project is outlined in Chapter 1. Methods used in more than one results chapter are described in the general materials and methods Chapter 2 Each results chapter follows the convention of presentation as a paper, reviewing relevant literature, describing specific materials and methods, and presenting and discussing findings.

Chapter 3 presents the characteristics of Arabidopsis nucleosome DNA sequences isolated in this project. The aim of this chapter was to determine any sequence preferences in Arabidopsis with respect to nucleosome positioning.

Chapter 4 presents linker length variation in Arabidopsis, with the aim of determining any patterns within the distribution of linker lengths.

Chapter 5 Presents the patterns of nucleosome occupancy within the Arabidopsis genome with respect to regulatory regions. The aim of this chapter is to determine any species-specific nucleosome occupancy.

Chapter 6 presents the position of DNA methylation and nucleosome positioning over the rRNA region, with the aim of determining any relationship between epigenetic marks.

Chapter 7 the results of the thesis are discussed in terms of questions raised, and potential future work.

References are given in Chapter 8.

Chapter 1

General introduction

1 General Introduction

1.1 Epigenetics

Epigenetics can be defined as heritable, reversible modifications to chromatin which affect phenotype and gene expression, but which do not affect the underlying DNA sequence. The term 'epigenetics' (epi meaning over or above), was coined by Conrad Waddington following his work on the development of *Drosophila* wing morphology and cellular differentiation in development (Waddington, 1953).

Since then, epigenetic phenomena have been widely observed throughout eukaryotes and generally are associated with regulatory mechanisms that affect phenotype, which relate to chromatin structure (Tchurikov, 2005). Chromatin can be thought of as existing in two states: i) heterochromatin which is compact and generally transcriptionally silent and ii) euchromatin which is more relaxed and generally associated with transcriptional activity. Specific epigenetic marks are associated with each class of chromatin (Fransz *et al.,* 2006).

1.2 Epigenetic marks

Chromatin structure is influenced through DNA methylation of cytosine residues, and modification of histone protein N-terminal tails. Such modifications are stable through mitotic and/or meiotic cell divisions giving rise to a heritable activation state (Zilberman and Henikoff, 2005).

1.2.1 DNA methylation

DNA methylation is one of the most abundant epigenetic marks, with 5.26 % of all genomic cytosines methylated in Arabidopsis (Lister *et al.,* 2008). DNA methylation occurs at CG, CNG and CNN in plants and mammals (Lister *et al.,* 2009). The pericentromeric regions, repeats and transposable elements have been found to be highly methylated in Arabidopsis. In addition methylation is found in the promoters of developmentally regulated genes and in the gene body of one third of all constitutively expressed genes in Arabidopsis (Cokus *et al.,* 2008; Zhang *et al.,* 2006). DNA methylation is initiated and maintained through *DOMAINS*

REARRANGED METHYLTRANSFERASE 2 (DRM2), METHYLTRANSFERASE1 (MET1), and *CHROMOMETHYLASE 3* (*CMT3). MET1* and *DRM2* share homology with the mammalian methyltransferases *DNA (cytosine-5-)-methyltransferase 1* (*DNMT1)* and *DNA (cytosine-5-)-methyltransferase 3 (DNMT3),* respectively while *CMT3* is plant-specific. Both *MET1* and *CMT3* are responsible for maintaining established DNA methylation marks and *DRM2* is responsible for establishing new methylation marks (*de novo* methylation).

The DNA-maintenance methyltransferase *MET1* methylates the 5' carbon of cytosine residues in the CG context, using s-adenosylmethionine as the donor for the methyl group (Cheng, 1995). Mutants in *met-1* have drastically reduced levels of CG methylation, and display a variety of developmental abnormalities (Finnegan *et al.,* 1996). While, up regulation of transposable elements was not significant in the *met-1* or *cmt3* single mutants it was significantly higher in the *met-1 cmt3* double mutant, suggesting that the two enzymes may act redundantly to silence transposable elements (Kato *et al.,* 2003b).

The plant-specific *CMT3* has been shown to interact, through the chromodomain, with the tails of histone H3 when it is methylated at both lysines at position 9 and 27 (Lindroth *et al.,* 2004). Therefore, *CMT3* may be involved in silencing of homeotic genes such as *FLOWERING LOCUS C (FLC)* (Bastow *et al.,* 2004).

De novo DNA methylation in Arabidopsis occurs through the DRM1 and 2 DNA methyltransferases. DNA methylation by DRM 1 and 2 is directed by siRNAs. Experiments involved the transformation of Arabidopsis plants with *FWA.* FWA is silenced in the wild-type, but remains active in the *drm1* and *drm2* mutants and in the siRNA pathway *rna-dependant rna polymerase 2*, *dicer-like3, sde4* and *argonaute4* mutants (Chan *et al.,* 2004). Thus the repression of FWA is through siRNA directed DNA methylation.

1.2.2 Histone modification

Histones are globular proteins with N-terminal tails that are available for modification. Post-translational modifications of histone N-terminal tails within nucleosome have an impact on transcriptional activity and chromatin structure. The pattern of histone modification is referred to as the histone code, since specific epigenetic marks are often associated with specific chromatin structures and transcriptional states (Figure 1.1) (Henikoff *et al.*, 2004). Histones are mostly modified at the N-terminal tails by classes of enzymes such as histone methyl-transferases (HMTs) and histone acetyl-transferases (HATs) (Grant, 2001). These enzymes are required both for *de novo* modification, and maintenance of epigenetic marks. Histone modification appears to be a more complicated mechanism than DNA methylation, the resultant action being dependent on the type and location of modification. Histone modification enzymes contain domains that recognise epigenetic marks; chromodomain containing proteins (such as Heterochromatin Protein 1, HP1) recognise methylation while bromodomain containing proteins (such as DDM1) recognise acetylated histones. These proteins seek out marks and target those areas for further modification resulting in a cascade of modification along a chromosomal region.

In *Arabidopsis*, flowering is dependent on down regulation of FLC following vernalisation. Levels of FLC remain low during development suggesting an epigenetic mechanism but studies have shown that there is no change in DNA methylation levels before and after vernalisation, there is however a change in the methylation state of histone H3. Before vernalisation, Histone H3 in di-methylated lysine 4, which has been recognised as a transcriptionally active mark. Following vernalisation methylation at this site is reduced in conjunction with di-methylation of Histone H3 at lysines 9 and 27. Methylation at these sites triggers binding of chromodomain containing HP1 and induce heterochromatin formation and hence silencing. Low levels of FLC are maintained throughout development by expression of VRN1 and 2. VRN1 encodes a DNA binding protein and VRN2 encodes a homologue of the *Drosophila* SU(Z)12 a member of the Polycomb group Proteins (PcG) family of proteins which have been implicated in maintaining the silence of Drosophila HOX genes through mitotic cell divisions (Bastow *et al.*, 2004).

Di- and tri- methylation of histone H3 at lysine 9 has been shown to be associated with the promoters of silenced genes in Arabidopsis. The floral repressor *FLC* is down-regulated following vernalisation by *VRN1* and *VRN2*, allowing floral development. Following cold treatment, *VRN2*, a homologue of the *Drosophila* Polycomb-group (PcG) E(Z), is recruited to the 5' region of FLC and causes di-methylation of H3K9 (Bastow *et al.,* 2004).

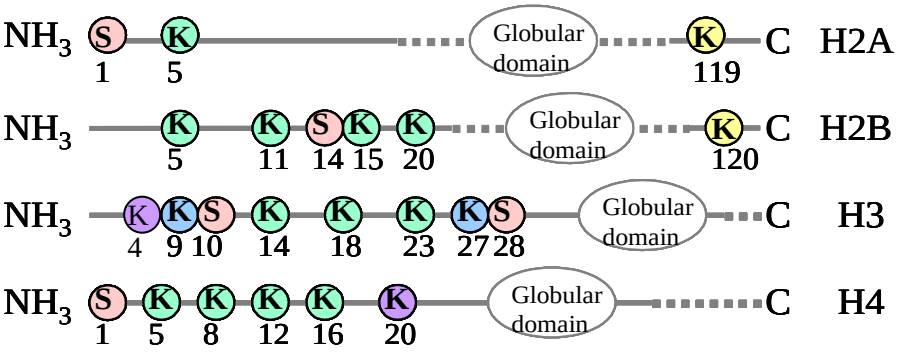

Figure 1.1 Diagram showing the sites of post-translational modification of the histone tails of the 4 core histones. Pink circles represent sites of phosphorylation, green: represents sites of acetylation, yellow: represents sites of ubiquitination, purple: sites of methylation and blue: sites of acetylation and methylation. The type of post-translation modification depends on the transcriptional activity of the DNA within the region of the nucleosome.

1.2.3 Chromatin structure at the Nucleolar organiser regions

The nucleolar organiser regions in Arabidopsis, are clustered into regions on chromosomes 2 and 4. They consist of the 18S, 5.8S and 25S rRNA genes in two clusters (nucleolar organiser regions (NORs)) of 700 ± 130 tandemly repeated copies on chromosomes 2 and 4, with a single repeat present on chromosome 3 (European Union Chromosome 3 Arabidopsis Genome Sequencing Consortium, 2000). The rRNA genes are separated by two internal transcribed spacer sequences and repeats are separated by an intergenic spacer The coding sequences are transcribed by RNA polymerase I (Pol I) and rRNA transcripts constitute 50-80 % of transcribed RNA in eukaryotic cells, although transcription of only a subset of repeats is needed to fulfil this requirement (Pruess and Pikaard, 2007). While the histone modification and DNA methylation status are being studied, there is little known about the chromatin structure of these regions.

1.3 The structure of the nucleosome

Nucleosomes are the building blocks of chromatin, and are generally positioned every 157- 240 bp in eukaryote chromatin (Englehardt, 2007). Nucleosomes consist of a core particle, a linker-associated histone H1 (or sometimes H5) and linker DNA (Figure 1.2). The core particle consists of 147bp DNA (Davey and Richmond, 2002) wrapped in a left-hand superhelical turn (of 10.17 bp per turn), 1.67 times around the histone octamer, with the central 129 of 147 bp organised over the histone (Richmond and Davey 2003).

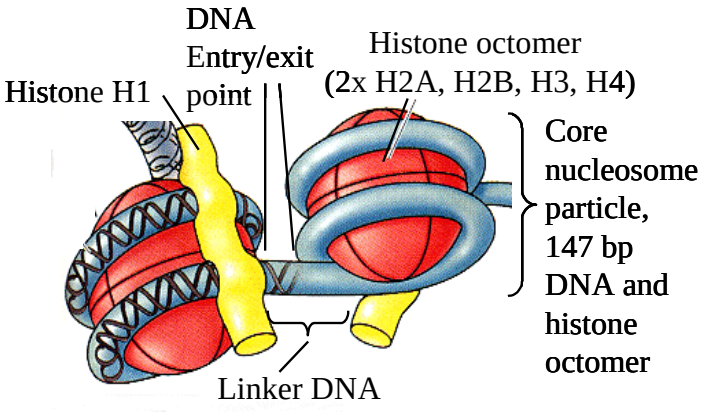

Figure 1.2. The nucleosomal core particle, linker DNA and linker-associated histone. Adapted from Purves *et al.*, (1997)

Evidence from x-ray crystallography shows that the 147 bp nucleosome DNA is not bent uniformly around the histone, but has sharp bends at specific points. Nucleosome DNA is 1.3 times more likely to be bent into the major groove, with the major groove facing in towards the histone octamer. Specific points of interaction exist between the nucleosomal DNA and the histone (Rhodes 1997, Richmond *et al.*, 1984) and the interaction between the histone octamer and the DNA is strongest around the dyad-axis region, with a weaker interaction up to 60 bp from the dyad and a 5 bp periodicity in contact points (Hall *et al.*, 2009).

The histone octamer consists of two molecules of each of histone H2A, H2B, H3 and H4 proteins, organised into pairs of H3/H4 and H2A/H2B dimers. Histone H1 is associated with linker DNA and protects the DNA at the entry/exit point to the

nucleosome. H1 also aids stability and prevents slippage of the nucleosomal position (Bustin *et al.*, 2005). Linker histone has also been shown to have a greater affinity for methylated DNA and contributes to chromatin compaction (McArthur and Thomas 1996).

Nucleosomes that form at the active centromere contain a centromere-specific variant of histone H3, CENP-A (also called CenH3) (Bernad *et al.*, 2009). Study of *Drosophila* nucleosomes has revealed that these variants, rather than being targeted towards the centromere, occur at other sites in the chromosome, but are later removed and replaced by canonical histone H3 (Dalal *et al*., 2007a). The structure of centromeric nucleosomes differs from the majority in that they consist of a histone tetramer rather than an octamer. The nucleosome core consists of one copy of each Histone H2A, H2B, CENP-A and H4 and is more susceptible to enzymatic digestion. Electron microscopy had provided further evidence for small particles, separated by long linkers that resist condensation under physiological conditions (Dalal *et al.*, 2007b). In fact the nucleosome repeat length of the small particles does not seem to differ from that of canonical nucleosomes.

1.4 Nucleosome Positioning

Almost any stretch of DNA has the ability to form nucleosomes. The use of prokaryote DNA in nucleosome reconstitution experiments demonstrates this effect, as prokaryote DNA is not packaged in the same way as eukaryotic DNA *in vivo* and is not associated with histones (Trifonov, 1984). Reports of preferential positioning, nucleosome-free regions and different nucleosome formation potentials of coding and non-coding DNA suggest that nucleosome formation is affected by specific nucleotide sequences or their associated properties (Vinogradov, 2005; Anselmi *et al.*, 1999). In addition, nucleosomes tend to be positioned on DNA with a higher GC content that the genomic average (Chung and Vingron, 2008).

Although nucleosome positioning is not dependant on specific base recognition by histone proteins, the location of a nucleosome is influenced by a number of factors. These include inherent mechanical properties of the sequence, such as the intrinsic curvature and flexibility, which are determined by base composition of local DNA sequence (Zuccheri *et al.*, 2001). Binding of non-histone proteins, such as those for recombination and transcription, compete with histone proteins for binding sites. Boundary effects and steric hindrance between adjacent nucleosomes are also implicated in positioning as there is a minimum length of linker DNA between nucleosomes that is required to enable formation of higher order structures (Baldi *et al.*, 1996).

Chromatin is a dynamic system where nucleosomes are formed at, and move toward sites of least resistance (low free energy) (Widlund *et al.*, 1999). Nucleosomes tend to be positioned at sites with strong nucleosome positioning signals and require energy (ATP) to be remodelled. These sites in the genome offer enhanced stability and have a higher probability of occupancy by a nucleosome (Fitzgerald and Anderson, 2004).

1.4.1 Influence of local DNA structure on nucleosome position

While specific contact points exist between the histone octamer and the nucleosomal DNA, there is no sequence recognition by the histone octamer to serve as a signal for nucleosome positioning. Nucleosome formation is influenced by local DNA structure determined by local DNA sequence, with particular contributions from constituent

dinucleotides (Fitzgerald and Anderson, 1999). DNA curvature and flexibility are thought to be major determinates of nucleosome positioning (Kiyama and Trifonov, 2002). The curvature and flexibility of a DNA molecule are determined by base sequence (Scipioni *et al.*, 2002; Pina *et al.*, 1990). Intrinsically curved DNA is curved in solution regardless of the presence or absence of protein (Cloutier and Widom, 2004). Flexible DNA, whilst intrinsically straight, is able to deviated from straight and follow different paths.

In order to establish the relationship between DNA structural properties and nucleosome position, DNA sequences have been synthesised and analysed for their ability to form nucleosomes *in vitro* (Shrader and Crothers, 1989). Sequences containing 5 repeats of a 10 bp sequence [TATAAACGCC] were shown to have the highest affinity for nucleosome formation by competitive reconstitution experiments with bulk DNA and natural nucleosome positioning sequences. However, this DNA sequence was later shown to be unable to position a nucleosome *in vivo*. This suggests optimal that DNA structure parameters considered in the original experiments are not sufficient alone to position nucleosomes *in vivo* (Tanaka *et al.*, 1992).

1.4.2 Nucleosome positioning sequence preferences

Sequence analysis of nucleosome DNA fragments has identified what appears to be nucleosome positioning sequence preferences. One of the most studied nucleosome positioning patterns is the [AA/TT] dinucleotide motif (Ioshikhes *et al*., 1996). The [AA] and [TT] dinucleotides tend to be placed counter-phase to each other, (with 5 bp intervals) and were discovered in experimentally mapped nucleosomal sites to be approximately 5 bp apart and occur with a 10.3 (+/-0.2) bp periodicity, (Figure 1.3). However this counter-phase pattern has since been demonstrated to decay over approximately 50 bp, whereas the in-phase [AA/TT] dinucleotide repeating pattern is longer-range, decaying over approximately 200 bp (Cohanim *et al.*, 2006). While the [AA/TT] repeating pattern has been well characterised, nucleotide repeats with different base compositions have also been identified. For example, similar dinucleotide patterns consisting of a more general [RR/YY] repeating motif have been discovered using sequences of clones from over 1000 human dinucleosome DNA fragments (Kato *et al.*, 2005, Kogan *et al.*, 2006). Further analyses revealed

that the dinucleotide repeat [CC/GG] appears to be the dominant nucleosome positioning signal in human chromatin, while the [AA/TT] repeat appears to have little influence (Kogan *et al.*, 2006; Kogan and Trifonov, 2005).

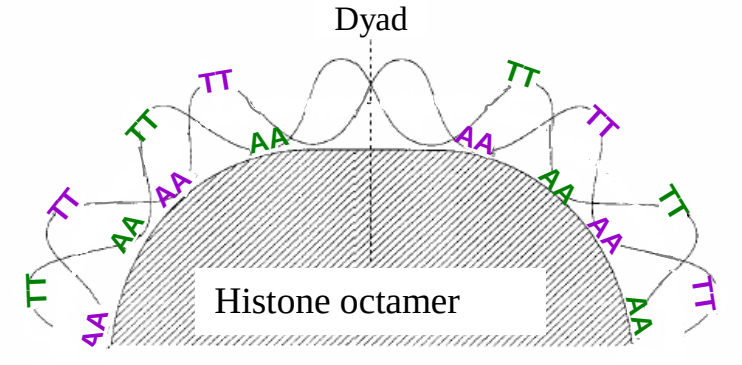

Figure 1.3 Representation of the organisation of periodic DNA sequence patterns around the histone octamer. Positions of dinucleotides on opposite strands are marked in purple and green. Adapted from Ioshikhes *et al.*, (1996)

A trinucleotide motif [VWR] with a 10.5 bp periodicity was identified whilst training a computer program to identify splice site junctions in human DNA (Baldi *et al.*, 1996). This was later confirmed as a nucleosome-positioning signal in yeast (Stein and Bina, 1999). The sequence occurs periodically throughout the human genome. In experiments to identify strong nucleosome positions in mouse chromatin, a 10 bp sequence [TATAA(A/C)CG(T/C)C] was isolated which has the strongest affinity for nucleosome positioning identified to date (Widlund *et al.*, 1997). This sequence was identified following multiple rounds of competitive reconstitution experiments.

A different class of nucleosome positioning motif, not previously described, was discovered in the DNA sequence of the unicellular organism *Trichomonas vaginalis*. This organism is known to have short 5' and 3' UTRs and a large number of duplicated genes. A 120.9 bp periodicity was discovered, which was more pronounced in gene sequences, and which started between the 5'UTR and the coding sequences with a preference for GC at the third codon, and a weaker preference at the first codon. The repeating 120.9 bp pattern decreased towards the 3' ends of the genes, suggesting a preference for nucleosome positioning within genic regions in this organism (Chen *et al.*, 2008).

Sequences containing [TGGA] repeats have been shown to impair nucleosome formation through successive nucleosome reconstitution experiments, where the un-reconstituted DNA is selected (Cao *et al.,* 1998). DNA fragments containing [TGGA] repeats (and fragments containing runs of the dinucleotide steps in [TGGA]) are thought to form secondary structures detrimental to nucleosome formation. These motifs occur in tandem repeats throughout eukaryotic genomes and are thought to be important as positioning signals for nucleosomal arrangement. For example, size polymorphisms of tandem [TGGA] repeats at the 5' end of the human myelin basic protein have been associated with subsets of multiple sclerosis cases (Tienari *et al.,* 1998). Other motifs thought to impair nucleosome formation include [CCG] repeats (ref) and poly [dA/dT] tracts (ref).

1.4.3 Nucleosome positioning and gene expression

During transcription, 5' ends of transcribed regions must be made available to the transcriptional machinery, whilst downstream sequences may remain unavailable. Nucleosomes positioned at promoter regions influence gene transcription, and relocation of nucleosomes has a dramatic effect on transcription rates (Parthasarthy and Gopinathan, 2006). Nucleosome occupancy is depleted at the promoter regions of many transcriptionally active genes: for example, positioned nucleosomes at promoter regions of polymerase II transcribed genes in *Salmonella* phage 6 inhibit transcription (Lorch *et al.,* 1987). However, once transcription has been initiated, RNA polymerase is able to 'read-through' nucleosomes positioned in the coding regions. This is thought to be because of the space required at the promoter regions for the transcriptional machinery.

The PHO5 promoter in yeast has been extensively characterised with respect to nucleosome position during transcriptional activation (Boeger *et al.,* 2003, Boeger *et al.,* 2004, Boeger *et al.,* 2005). During transcriptional activation, two of the four nucleosomes positioned at the promoter are lost. Incomplete removal of nucleosomes at the promoter suggests a mechanism of removal and re-formation following the passage of the transcriptional machinery (Boeger *et al.,* 2004). Microarray analysis of yeast chromatin following heat-shock revealed that 65 % of genes had their transcript level altered by a factor of two, and also showed changes in nucleosome

occupancy. Nucleosome occupancy was found to be inversely proportional to the level of transcriptional activity. Furthermore, genes which became active following heat shock showed a loss of nucleosome occupancy at the promoter, while genes which became silenced gained nucleosome occupancy at the promoter (Lee *et al.,* 2004). It was hypothesised that nucleosome depletion at promoters of actively transcribed genes is a characteristic of eukaryotic chromatin structure.

Histone variants H2A.Z have been identified in global studies of yeast chromatin. Nucleosomes containing H2A.Z typically flank a nucleosome-free region, which extends approximately 200 bp upstream of the transcriptional start and includes the promoter region (Raisner and Madhani, 2006). Histone variant H2A.Z is thought to be a mark of euchromatin in yeast as it is not present at promoters of inactive genes. A genome-wide map of H2A.Z-containing nucleosomes revealed the presence of H2A.Z at DNA methylation-depleted regions in Arabidopsis (Zilberman *et al.,* 2008). A peak in H2A.Z occupancy was observed in the position of the first nucleosome after the transcriptional start site in Arabidopsis chromatin, which is consistent with the position of H2A.Z peaks observed in yeast (Mavrich *et al.,* 2008b), *Drosophila* (Mavrich *et al.,* 2008a) and in transcriptionally active human chromatin (Schones *et al*, 2008). The genome-wide study in Arabidopsis revealed a lack of H2A.Z-containing nucleosomes from heavily methylated regions such as the pericentromeric regions and transposable elements, suggesting that DNA methylation excludes deposition, or perturbs formation of H2A.Z-containg nucleosomes. Conversely, a mutation in the *PIE1* domain of the *Swr1* complex, responsible for deposition of H2A.Z-containing nucleosomes, resulted in genome-wide hypermethylation within gene bodies, suggesting that H2A.Z also protect the DNA from methylation (Zilberman *et al.,* 2008). These results suggest that H2A.Z-containing nucleosomes are a feature of transcriptional activity at Pol II promoters across kingdoms.

While nucleosome positioning at promoter regions provides a barrier to transcription for some genes, transcription rates of RNA III polymerase transcribed genes (tRNA's) are enhanced by positioning of nucleosomes at the promoter regions (Shivaswamy *et al.,* 2006). Positioning of nucleosomes at the promoter of these

regions brings distally placed promoter elements and transcriptional start sites within close proximity

1.5 Chromatin remodelling

Chromatin structure has a requirement to be dynamic, as an unalterable chromatin state would not allow for differential transcriptional activity. This dynamic nature is achieved through chromatin remodelling. Remodelling of chromatin refers to the movement of nucleosomes to bring about a change in chromatin structure. This may occur through many mechanisms, for example, changes in ionic concentration, salt concentration and thermal energy required to induce nucleosome movement both short range and over longer distances (Marky and Manning, 1995). Remodelling complexes possess nucleosome re-positioning abilities and their action can result in de/condensation of chromatin.

1.5.1 Translational and rotational nucleosome positioning

There are two types of nucleosome positioning in relation to DNA sequence: translational and rotational. Translational nucleosome positioning refers to the linear region of DNA that the nucleosome occupies. There does not appear to be a consensus sequence for strong translational nucleosome positioning. However, DNA sequence patterns, such as the [AA/TT] dinucleotide repeat, confer a higher statistical probability with which a histone octamer selects a contiguous stretch of 147 base pairs DNA (Thalstrom *et al.,* 2004). These sequence patterns could constitute likely signals which represent preferred local sequences involved in direct DNA protein interaction. Translational positioning is more likely to occur as a result of ATP-dependant chromatin remodelling of a rotationally positioned nucleosome away from the energetically favourable DNA structure.

Rotational positioning refers to the orientation in which the DNA molecule is positioned as it wraps around the histone octamer with respect to the parts of the molecule that are facing towards the octamer. This has implications for binding of regulatory elements, such as promoter elements, to the DNA molecule as part of the molecule is ‘protected’ by the histone proteins and part is exposed. Rotational

positioning signals give permanent directional curvature to the DNA and have been shown to occur over TATA – containing promoters. Therefore, they also act as translational signals, as a position with a strong rotational preference (or curvature) is more likely to have a nucleosome occupying it. The β-phaseolin gene in tobacco is active during embryogenesis, but transcriptionally silent at all other developmental stages. The silenced state is consistent with a rotationally-positioned nucleosome over three phased TATA boxes within the promoter (Li *et al.*, 1998). The promoter nucleosome is removed by chromatin remodelling during seed development, when the gene becomes transcriptionally active.

Two classifications of nucleosome movement exist, sliding and translocation. Sliding refers to the movement of a nucleosome along a DNA molecule, while translocation refers to the removal of the histone octamer to another DNA molecule (Lorch *et al.*, 2004).

1.5.2 ATP dependant remodelling complexes

The ATP-dependent remodelling complex SWI/SNF was discovered in *Sachharomyces cerevisiae*. This discovery arose from a series of genetic studies, which were intending to identify functions of specific genes (Winston and Carlson, 1992, Hsieh and Fischer, 2005). These included (among others) the *HO* endonuclease gene (involved in mate-type switching) and the *SUC* invertase gene (which is required for hydrolysis of glucose to raffinose). The analyses resulted in the discovery of a set of overlapping regulatory genes and were named SWI/SNF (SWI- switch and SNF- sucrose non-fermenter). These have since been shown to act as transcriptional regulators for many genes and are evolutionarily conserved across eukaryotes (Vignali *et al.*, 2000).

The SWI/SNF remodelling complexes are divided into several classes. These are characterised by the proteins, other than the ATPase, found in the complex. The different classes exhibit different mechanisms and vary in their targets and actions. They are abundant in the genome, with over 40 present in Arabidopsis.

The SWI2/SNF2 class has a conserved ATPase domain, a bromodomain in C-terminal, as well as two other conserved domains of unknown function (Vignali e*t al.*, 2000). These complexes are mostly involved in alteration of the winding of the DNA molecule around a histone octamer, and therefore are mostly involved in transcriptional activation and repression. The SWI2/SNF2 group include:

- SWI/SNF in yeast (ySWI/SNF)
- Brahma complexes isolated from *Drosophila*
- Human BRM (hBRM)
- BRG1 complexes (mouse homologue mBRG1)

ISWI (Imitation SWItch) also, has a conserved ATPase domain, but is smaller than the SWI2/SNF2 class, having fewer subunits and shows homology to the SWI2/SNF2 proteins only over the ATPase domain. They require the presence of Histone H4 N-terminal tail for full activation (they are activated by DNA in nucleosomes). ISWI action includes relocation of nucleosomes by sliding along DNA and is mostly involved with maintaining higher order structure and chromatin assembly (Längst and Becker, 2001). The ISWI group include:

- ATP-utilising chromatin assembly and remodelling factor (ACF)
- Nucleosome remodelling factor (NURF)
- Chromatin assembly complex (CHRAC)

1.5.3 Mode of action, twist defect and loop recapture

Two models for the mechanism of the movement of a nucleosome along the DNA molecule have been suggested. These are the twist defect model and the loop recapture model (Fig 1.4) (Längst and Becker, 2004). The twist defect model relies on local alterations of DNA torsion around the nucleosome, brought about by the

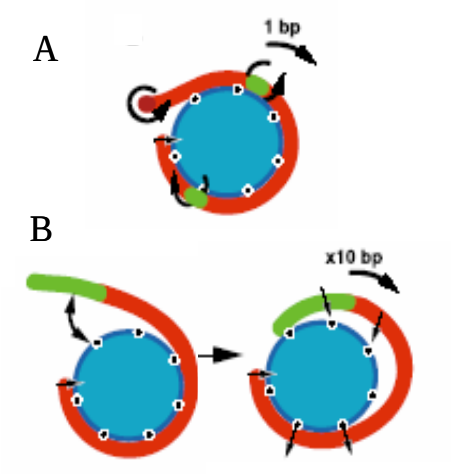

Figure 1.4 Schematic diagram to show the possible remodelling mechanisms. **A**: twist defect and **B:** loop recapture models
Image from (Langst *et al*,. 1997)

action of remodelling complexes. These changes in direction twist the DNA away from the surface of the nucleosome, and therefore change the rotational position. The loop recapture model suggests that the DNA molecule detaches at a specific point. More of the DNA molecule is incorporated into the nucleosome while still in contact with the histone octamer, creating a loop or bulge. This loop is propagated along the length of the nucleosome DNA by subsequent detachment in front of the loop, and reattachment of the DNA molecule and the octamer behind the bulge. This results in a stepwise movement of the nucleosome along the DNA molecule to an alternate position (Flaus and Owen-Hughes, 2004). There is evidence to show that the loop recapture model is likely to be the prevalent mechanism *in vivo* (Strohner *et al.*, 2005, Lorch *et al.*, 2004).

1.6 Variation in linker length and higher-order chromatin structure

Linker DNA refers to the DNA between nucleosome core particles and can exist as either naked or linker histone-bound. It has been described as being between 10 and 90 nt in length (Robinson *et al.,* 2006a) and thought to comprise of sequences which favour a straight or rigid structure, such as poly [A] tracts (Marx *et al.,* 2006). The linker histone 'anchors' nucleosomes in place by attachment of one end of the protein to the linker DNA and the other end to the DNA within the nucleosome (Bustin *et al.,* 2005). Variations in linker length have been observed between and within species. Estimations of linker length were collected from the literature and published as a species comparative list (van Holde, 1979). Fourier analyses of this collection of linker lengths revealed a 10 bp periodicity for linker lengths, which is thought to reflect the rotational positioning of nucleosomes on 10 bp sequence repeats (Widom, 1992). A similar periodicity was reported in Human DNA (Kato *et al.,* 2003). However these linker lengths were derived from nucleosome positions predicted by the DNA sequence patterns, while the periodicity was not found in the distribution of lengths of >1000 dinucleosome fragments isolated from K625 cells.

Variation in linker length is thought to be related to the level of transcriptional activity of the cell from where the chromatin originated, with shorter linker lengths and more tightly packed nucleosomes indicated areas of lower transcriptional activity (Olins *et al.,* 1976). An association between linker length and chromatin fibre diameter was established by electron microscopy measurements of chromatin from three tissues with different nucleosome repeat lengths (NRL, the length from the start of one nucleosome to the start of the next): chicken erythrocyte (NRL = 208 bp), mouse thymus (NRL = 195 bp) and sea cucumber spermatozoa (NRL = 277 bp). While the level of compaction, and therefore the chromatin fibre diameters varied with salt concentrations, the size of the fibres consistently reflected the NRL s of the chromatin from which they were formed (Alegre and Subirana, 1989).

To understand the higher-order folding of chromatin, experiments involving measurements under different conditions have been performed. This revealed that compaction is dependent on salt concentration. Nucleosomes containing linker histones in low salt adopt a zig-zag conformation, while in the absence of H1, the

'beads on a string' structure is observed (Sun *et al.*, 2005). In high salt concentrations (10 mM monovalent) the trajectory of the DNA as it exists the nucleosome is altered and forms a 'stem structure'(Woodcock and Horowitz 1997).

Two models for chromatin compaction beyond the nucleosome level exist: the solenoid and the zig-zag. The solenoid structure (also called one-start helix structure) was originally suggested by Finch and Klug (1976). In this model, the nucleosomes are packed adjacent to one another, with the linker DNA bent between nucleosomes to accommodate placement of the neighbouring nucleosome continuing the coil established by the nucleosome. Therefore, in the solenoid model the 30 nm fibre diameter is not affected by the length of the linker DNA. In the zig-zag model (also known as crossed linker and 2-start helix structure models) the linker crosses the middle of the 30 nm fibre, and neighbouring nucleosomes sit opposite each other in the 30 nm fibre (Wong *et al.*, 2007). A zig-zag model arrangement of nucleosomes would favour either straight or flexible DNA but not likely to favour curved DNA. The resulting chromatin fibre would be affected by changes in linker length, however the ability of DNA to stretch around the histone octamer allows for discrepancies in linker length of up to 4 bp without affecting higher order structure (Davey *et al.*, 2002).

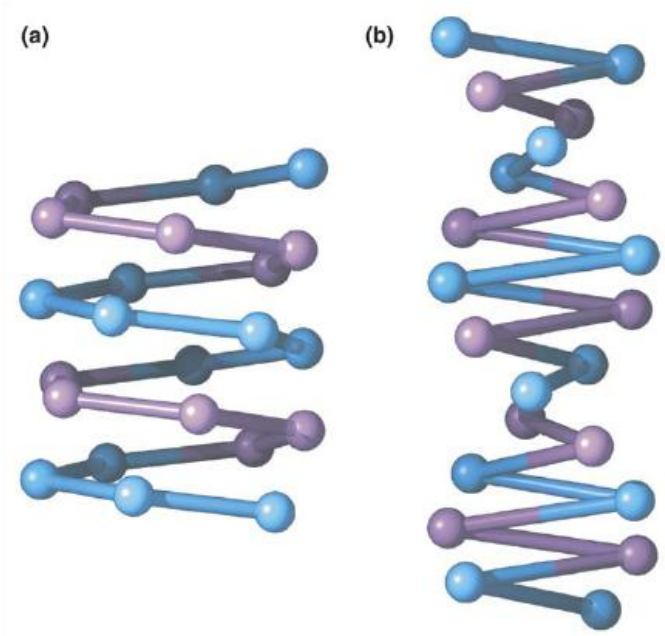

Figure 1.5 Schematic representations of the different models of chromatin fibre folding, A: the solenoid (or 1-start) model and B: the zig-zag (or 2-start) model. Image from Robinson *et al.*, (2006b).

Recent work in this area suggests that a non-linear relationship exists between linker length and chromatin fibre diameter (Robinson *et al.*, 2006a). The strong nucleosome-positioning DNA sequence '601' was used to reconstitute nucleosome arrays of up to 72 nucleosomes, with evenly spaced nucleosomes and linkers ranging from 10 to 70 bp. Measurements of chromatin fibres showed two classes (1) arrays with NRL from177-207 had an average diameter of 33 nm and 10.8 nucleosomes/11 nm and (2) arrays with NRLs of 217-237 bp had a diameter of 43-ish nm and 14-15 nucleosomes/11nm. They suggested a model where the preferred structure of the chromatin fibre is a left-handed one-start helix (solenoid) and that the linker histone contributes to bending intrinsically straight DNA into the correct trajectory.

This work has been further enhanced by computer modelling of linker trajectory and the organisation of nucleosomes within chromatin fibre, for increasing linker lengths. These models suggest a different structure of fibre for each 10 bp 'step' in linker length, and also suggests that both solenoid-type and zig-zag type structure may be present (Wong *et al.*, 2007).

1.7 Recent advances and tools for chromatin studies.

Epigenetic studies are benefiting from recent advances in microarray and sequencing technologies. Previous studies in Arabidopsis had identified methylated regions of heterochromatin on chromosome 4 (Lippman *et al.*, 2004) at 1 kb resolution. However, global profiling of Arabidopsis DNA methylation sites using the Affymetrix GeneChip® *Arabidopsis*1.0R tiling array, has revealed patterns of DNA methylation specific to particular gene classes (Zhang *et al.*, 2006). Of all expressed genes, 61.5 % were found to be un-methylated and this group was enriched for transcription factors; 33.3 % were methylated in the body of genes, which tended to be constitutively expressed genes; and 5.2 % genes were methylated in the promoter. Promoter methylated genes were more likely to be expressed in a tissue-specific manner. The Arabidopsis methylome has been even more accurately mapped to base pair resolution by Solexa sequencing of bisulphite-treated DNA (Lister *et al.*, 2008).

The tiling microarray approach was used to identify the translational positions of 2278 nucleosomes in *Saccharomyces cerevisiae* (Yuan *et al.*, 2005). Nucleosome DNA fragments were liberated by micrococcal nuclease digestion and hybridised to microarrays containing 50 bp probes tiled every 20 bp. Probes covered most of chromosome III and a further 223 genes on other chromosomes. These experiments revealed that 65-69% nucleosomes in yeast are well positioned. De-localised (or 'fuzzy') nucleosomes tend to be found at highly expressed genes, which suggests that nucleosomes are transiently disassembled and re-assembled, thus allowing for the passage of RNA polymerase. This was the first time that high-resolution microarrays had been used to study nucleosome positioning. Analysis of the yeast nucleosome landscape by tiling arrays revealed that 81 % of the genome is occupied by nucleosomes (Lee *et al*., 2007). This experiment revealed that the intergenic regions are nucleosome-depleted compared with the coding regions. A feature which has now been shown in human chromatin (Schones *et al.*, 2008). This experiment also revealed a pattern in nucleosome occupancy at the transcriptional start sites (TSS), where there appears to be higher occupancy to the right of the TSS (upstream).

Chromatin structure at human promoters has also been investigated using a tiling microarray. Mononucleosomal DNA was hybridised to microarrays containing

probes 50 nucleotides in length and staggered every 10 bp (Ozsolak *et al.*, 2007). Microarray probes covered 1.5 kb promoter regions of 3692 genes. Seven cell lines were compared and a high degree (90 %) of reproducibility between experiments was observed. Approximately 88% of the promoters investigated were found to contain at least one positioned nucleosome, and expressed genes were more likely to have a nucleosome-depleted region of approximately 100 bp spanning the transcriptional start site. It was suggested that this approach could be used as a tool for the analysis of chromatin structure in disease, as the microarray contained promoter regions of all the genes represented on the Affymetrix Human Cancer Array.

The increase in available nucleosome positioning data will aid the refinement of nucleosome prediction models. A novel prediction algorithm has been developed, based on nucleosome positioning sequence preferences identified from 199 experimentally determined yeast nucleosome DNA fragments (Segal *et al.*, 2006). The algorithm was designed to recognise [AA], [TT] and [TA] dinucleotides with a 10 bp periodicity, and was able to predict approximately 50% of nucleosome positions experimentally determined using a tiling microarray (Yuan *et al.*, 2005). The nucleosome positions identified using the microarray were also compared to computationally identified nucleosome positioning signals for gene regions (-1,000 bp to +800 bp relative to the transcriptional start site) (Ioshikhes *et al.*, 2006). These analyses revealed high degree of accuracy in predicting nucleosome positions at yeast promoters. These models have since been superseded by the more recent probabilistic model developed by the Segal lab (Kaplan *et al.*, 2008). This was derived from yeast *in vitro* positioning data and scores the nucleosome formation potential. This model therefore represents only nucleosome sequence preferences.

Massively parallel sequencing technologies have been utilised to map the positions of nucleosomes containing the histone variant H2A.Z in *Saccharomyces cerevisiae* cells (Albert *et al.*, 2007). While some nucleosomes are always strongly positioned, many exhibit some cell to cell translational variation. This technology allows a more accurate analysis of nucleosome position than tiling microarrays, as multiple sequence reads are obtained for translational nucleosome positions. A map of H2A.Z has also been constructed for Arabidopsis, which revealed an interesting

relationship between this histone variant and DNA methylation (Zilberman *et al.*, 2008). H2A.Z-containing nucleosomes appear to protect the DNA from DNA methylation within euchromatic regions, conversely, the H2A.Z –containing nucleosomes are excluded from heavily methylated areas such as the pericentromeric repeats. High-throughput sequencing of the yeast and *Drosophila* nucleosomes has revealed some interesting patterns of occupancy around the TSS (Mavrich *et al,* 2008a; Mavrich *et al.,*2008b). Both genomes show a peak in nucleosome occupancy after the TSS, and ordered nucleosome phasing thereafter (within the gene region). However the distance of the peak was different for the two organisms resulting in the TSS being nucleosome-free in Drosophila, but the TSS was buried within the nucleosome in yeast. This suggests that there may be different mechanisms regulating transcriptional control in these two species.

Recent advances in high-throughput technologies have allowed the study of genome-wide locations of histone modifications and variants in Arabidopsis. The tri-methylation of histone H3 at lysine 27 is established and maintained by PcG proteins and is involved in silencing more than 4,000 Arabidopsis genes (Zhang *et al.*, 2007). Conversely, the histone modifications H3K4me^3 and H3K9ac are associated with the transcriptionally active genes involved in drought stress in Arabidopsis (Kim *et al.*, 2008).

1.8 Nucleosome positioning databases

A publicly available set of nucleosomal sites and sequences has been available in a nucleosome position region database (NPRD) (Levitsky *et al.,* 2005). The database contains a compilation of experimentally determined nucleosome positions. Each entry relates to an individual nucleosome site, and details associated with that site, for example, the relationship between nucleosome position and type of gene activity, type of nucleosome position (rotational or translational) and factors influencing nucleosome position. The NPRD contains sequences representing a variety of eukaryotic and viral nucleosome sites over a variety of genomic regions (Figure 1.6). However, only 2 plant-derived data are represented (*Zea mays*). The current version of the database contains 112 entries (February 2007) and is still in development (112 entries, last checked 01/09/09).

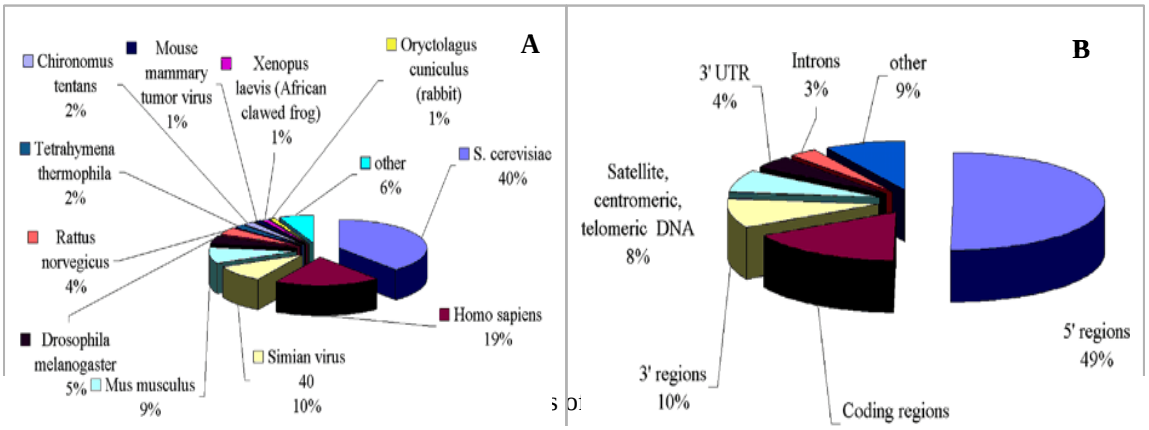

Figure 1.6 A Chart showing the relative numbers of organisms represented in the NRPD (2005 version). B Shows a representation of genomic regions represented in the database. Images from Levitsky *et al.,* (2005)

1.9 Project aims

The motivation for this project arose from a lack of plant-derived nucleosome positioning data, and the need to understand the regulation of plant gene expression at the chromatin level. At the beginning of this project, there were plant-derived nucleosome positioning data for only a few individual genes, but no global studies had been carried out.

1.9.1Scientific Hypotheses to be tested

1. Arabidopsis nucleosome sequence composition is different from the genome average and there are differences in nucleosome and linker DNA.

2. Arabidopsis nucleosome sequences display specific distributions in their dinucleotide complement which are known to contribute to the bending of DNA around the nucleosome and that these preferences are similar to those discovered in other taxa

3. The distribution of linker length in Arabidopsis is comparable to that determined for other taxa and Periodicity exists within the Arabidopsis linker length distribution.

4. Features of the observed distribution of Arabidopsis linker length show similarities to the distribution of Arabidopsis intron length, which may indicate a relationship between introns in the DNA and chromatin structure.

5. Nucleosomes are evenly distributed throughout the Arabidopsis genome, and biases in nucleosome occupancy differ between coding and non-coding regions in Arabidopsis.

6. Ordered positioning of nucleosomes exists around the TSSs in Arabidopsis, comparable to the organisation found within other taxa.

7. The Arabidopsis Nucleolar Organiser Regions display different nucleosome patterns of positioning dependent on chromosomal origin, due to the transcriptionally heterogeneous nature of these regions.

8. Overall differences in nucleosome positioning exist between wild-type plants and plants with depleted ^{5m}C levels.

1.9.2 Experimental approaches

To address the above hypotheses, the following approaches were taken:

Dinucleosome libraries have been constructed from *A. thaliana* tissue. A subset of clones has been sampled to produce a set of ~500 dinucleosomal sequences. These sequences were used to determine linker length distribution and analysed for sequence patterns.

High-throughput sequencing technologies (Roche 454 FLX and Solexa) were used to determine the sequence and genomic positions of mononucleosomes and dinucleosomes throughout the Arabidopsis genome. Specific-dinucleotide and linker length distributions were calculated from DNA fragment data and nucleosome occupancy within specific gene classes and at the transcriptional start sites were investigated.

Mononucleosome DNA was isolated from the Arabidopsis *MET1*-antisense to determine any differences in the sequences composition, nucleosome spacing and genomic occupancy of nucleosomes in a DNA methylation-deficient environment.

Chapter 2

General materials and methods

2.1 General laboratory protocols

All reagents from Sigma unless otherwise stated.

2.1.i Plant material

Arabidopsis thaliana Columbia-0 (Col-0) and antisense *MET1* line 10.1, T4 generation (C24 ecotype) (Finnegan *et al.,* 1996) were sterilised in 25 % Parazone for five minutes, washed with 70% ethanol for two minutes, rinsed twice and re-suspended in sterile water. Seeds were cold-treated at 4 ^{o}C for five days and sown onto sterile half-strength Murashige and Skoog media (M0404) with Gamborgs vitamins, containing 1% sucrose and 1% agar (Bacto), pH 5.7. Plates were incubated at 18 - 22^{o}C in constant light for 7 days. Seedlings were transferred to soil (Levington F2's and intercept 5GR 280g/m^{3}) and grown in long day conditions (16 hours light 300 μmol m−2 sec−1 from Osram HQI-BT 400W/D metal halide lamps at 23^{o}C day65% relative humidity / 8 hours dark, 18^{o}C 70% relative humidity). Leaf material (rosette leaves 6, 7 and 8) was harvested 18-21 days post germination and cross-linked immediately following harvest (Section 2.1.ii).

2.1.ii Formaldehyde cross-linking of protein and DNA

The cross-linking procedure was carried out in a fume hood. Formaldehyde cross-linking of protein and DNA was adapted from the method of Gendrel *et al.,* (2005). The fresh Arabidopsis leaf material was added to 37 ml of 1% formaldehyde. The leaf/formaldehyde mixture was vacuum-infiltrated at room temperature for 15 minutes (Heto, Sue-300). To stop the cross-linking reaction from progressing any further, 2.5 ml of 2 M glycine (to final concentration of 0.125 M) was added to the leaf tissue and was vacuum-infiltrated for an additional 5 minutes. The tissue was rinsed with sterile water to remove the formaldehyde and blotted dry between paper towels. Nuclei were isolated from cross-linked tissue immediately following the cross-linking step.

2.1.iii Nuclei isolation

The nuclei isolation protocol was adapted from the method of Bowler *et al.*, (2004). Fresh leaf tissue (20 g) was ground in 30 ml pre-cooled Nuclei extraction buffer, filtered through two layers of miracloth (Calbiochem, cat.# 475855) into cold 50 ml sterile centrifuge tubes (TPP). The extract was centrifuged at 2500 x g for 10 minutes at 4^{o}C using a bench top centrifuge (Sigma 3K10). The supernatant was discarded and the extract washed three times by re-suspension in 50 ml ice-cold nuclei wash buffer and centrifuged at 3000 x g for six minutes. Following the final wash, the supernatant was discarded and the pellet was re-suspended in 1 ml nuclei wash buffer. Nuclei were layered over an ice-cold 1.5 M sucrose (Fisher) solution in a sterile centrifuge tube and centrifuged for three minutes, 3000 x g at 4^{o}C to remove remaining chloroplasts. The supernatant was removed and the nuclei pellet was re-suspended in 4 x pellet volume in nuclei storage buffer and an equal volume of glycerol was added. The suspension was divided into two aliquots, the first containing 25% of the total volume of the nuclei suspension, the other containing 75%. Nuclei were frozen in liquid nitrogen and stored at -80 ^{o}C.

Nuclei isolation buffer:

1 M Hexalene glycol (Fisher Scientific)

10 mM PIPES/KOH pH 7.0 (Fluka)

10 mM $MgCl_2$ (Fluka)

0.2 % Triton X-100 (Promega)

5 mM 2-mercaptoethanol

1 mM AEBSF (4-(2-Aminoethyl)benzenesulfonyl fluoride hydrochloride (A8456))

Nuclei wash buffer

as isolation buffer, except 0.5 M hexylene glycol

Nuclei storage buffer

50 mM Tris-HCl, pH 7.5

5 mM $MgCl_2$ (Fluka)

0.1 mM EDTA (Duchefa)

2.1.iv Preparation of nucleosomal DNA

In order to determine the optimal concentration of micrococcal nuclease (MNase, Worthington Biochemical Corporation, USA) enzyme to produce nucleosome DNA fragments of the desired length, a small analytical digest was performed on a quarter of the nuclei preparation. This was performed for each new batch of chromatin in nuclei prepared.

2.1.v Preparation of the enzyme

MNase was re-suspended in MNase buffer at a concentration of 15 U/μl and 50 μl aliquots were stored at -20^{o}C prior to use.

MNase buffer
10 mM Tris-HCl pH 7.5
0.5 mM EDTA
0.5 mM DTT

2.1.vi Quantitative Micrococcal Nuclease digests

The optimal concentration of MNase required to digest partially chromatin in nuclei was determined by digestion of 25% of the total nuclei preparation. Nuclei were thawed on ice, pelleted by centrifugation at 3000 x g for three minutes and re-suspended in 450 μl of buffer M. Aliquots of 150 μl were transferred to 3, 1.5 ml microfuge tubes. Chromatin in nuclei was digested with 1, 10 and 50 μl aliquots of a 1/500 dilution MNase (1/500 = 0.03 U/μl) for 4 minutes at 37^{o}C. The reaction was stopped by the addition of 15 μl 0.5 M EDTA and DNA extracted.

Buffer M
50 mM Tris-HCl pH 7.4
60 mM KCl (Riedol-de Hean)
3 mM $CaCl_2$ (Fisher Scientific)
50 % glycerol (Fisher Scientific)
0.34 M sucrose (Fisher Scientific)

To isolate nucleosomal DNA fragments, nuclei were transferred to a 2 ml tube and DNA extraction buffer was added to a final volume of 1 ml. The nuclei were incubated at 55^{o}C for 30 minutes. DNA was extracted using 24:1 chloroform:isoamyl alcohol, precipitated with ice-cold 100% ethanol, washed with ice-cold 70% ethanol and dried in a vacuum drier for 3-5 minutes. The DNA pellet was re-suspended in 20 μl 1 x TE buffer (pH 8) and treated with 10 μg RNase A/T mix (Fermentas) at 37^{o}C for 30 minutes. Formaldehyde cross-links were reversed by incubation at 65^{o}C for 2 hours. DNA fragments were separated by electrophoresis on a 1.5% agarose (Bioline)/TBE gel containing Ethidium Bromide (Bio-Rad) (1 % vol), at 100 V for 40 minutes. Fragments were viewed using Bio-Rad Quantity One UV geldoc system.

DNA extraction buffer

100 mM Tris-HCl pH 8

50 mM EDTA pH 8

150 mM NaCl (Fisher Scientific)

1 % SDS

100 μg/ml proteinase K

2.1.vii Preparative Micrococcal nuclease digests

The remaining nuclei preparation was gently thawed on ice, pelleted and re-suspended in M buffer, to twice the pellet volume. The suspension was digested with the optimal concentration of enzyme determined in the serial digest. Following extraction, DNA was re-suspended in 40 μl sterile water, quantified (UV Nanodrop) and separated on an agarose gel (as above).

2.1.viii Cloning

DNA bands of approximately 300-400 bp were excised and purified using the Promega 'Wizard DNA gel purification and Clean-Up kit' according to the manufacturer's instructions. Fragments were cloned using the *Eco*RV site of

pSTBlue cloning vector ('Perfectly Blunt Cloning Kit '(Novagen)) following the manufacture's instructions, with the ligation reaction time increased to 25 minutes. Colonies were grown statically overnight at 37^{o}C on 2 xYT containing 1% Agar (Bacto ™), 50 µg/ml carbenicillin (Melford Laboratories), 12.5 µg/ml tetracycline (Duchefa), 70 µg/ml X-gal (Melford Laboratories), and 80 µM IPTG (Melford Laboratories),.

Bacterial strain

Novablue Singles ^{TM} high-efficiency competent cells (Merck)

Genotype:

endA1 hsdR17 (rK12– mK12+) supE44 thi-1 recA1 gyrA96 relA1 lac F'[proA+B+ lacIqZDM15::Tn10] (TetR)

Colonies were screened for the presence of inserts by colony PCR using vector-specific primers:

T7 promoter (F) 5'-TCTAATACGACTCACTATAGGG-3',

U19-mer (R) 5'-GTTTTCCCAGTCACGACGT-3'.

Amplification conditions were: 94^{o}C/ 4 minutes, 30 x (94^{o}C/ 30 seconds, 52^{o}C/ 30 seconds, 72^{o}C/ 1 minute), 72^{o}C/ 5 minutes and hold at 10^{o}C).

Colonies containing inserts were transferred to 96 deep-well (1.2 ml) plates containing 200 µl YT, 10 % glycerol and 50 µl/ml carbenicillin. Cultures were grown statically at 37^{o}C for 12 hours and stored at -80^{o}C. Plasmid DNA was amplified by from bacterial colonies by rolling circle amplification using the Templiphi™ amplification kit (GE Healthcare). Bacterial cells were transferred into a 0.2 ml PCR tube using a cocktail stick and a 1.25 µl sample buffer was added. Cells were lysed for 5 minutes at 95^{o}C. A mixture of 1.25 µl enzyme buffer and 0.05 µl enzyme mix (containing ϕ29 polymerase and random hexamer primers) for each sample was prepared, and 1.25 µl was added to each tube of cooled lysed cells. The reaction continued overnight at 30^{o}C, and was stopped by incubation for 10 minutes at 65^{o}C.

2.1.ix Determination of clone library quality

To check quality of library, inserts were sequenced directly from plasmids using Big Dye terminator kit V.31 (Roche) according to the manufacturer's instructions, using one-eighth of the Big Dye reaction mix recommended. Amplification conditions were: 96°C/ 1.5 minutes, 35x (96°C/ 10 seconds, 50°C/ 5 seconds, 60°C/ 4 minutes) and 22°C hold). Test reactions were sequenced at Oxford sequencing services.

2.1.x Sequencing of clone library

Plasmids containing inserts were amplified by the Templiphi™ rolling circle amplification (as above) in 96-well PCR plates. Plasmid DNA from the clone library was quantified and sequenced at the QIAGEN Genomic and Sequencing Services facility.

2.1.xi Large scale sequencing of Arabidopsis nucleosome fragments

Dinucleosome DNA fragments were prepared as described above (2.1.vii.) for sequencing by Roche 454 FLX. Library preparation of Arabidopsis Col-0 dinucleosome fragments by Roche 454 FLX was carried at the John Innes Genome Centre, Norwich, and sequencing was carried out at Cogenics (Hope end, Takeley, Essex, CM22 6TA) via the John Innes Genome Centre, Norwich.

Dinucleosome and mononucleosome fragments were prepared as described above (2.1.vii.) from wild-type and *MET1* antisense plants. In order for the samples to be mixed and sequenced using a single solexa channel, a different 4 bp 'barcode/tag' was ligated to the ends of the fragments from each sample. This enabled the separation of sequence data into the two samples following sequencing. The barcode 'tags' were removed from the sequences prior to the data being released from GATC. Library preparation (including barcode and adapter ligations) and sequencing were carried out at GATC Biotech AG (Jakob-Stadler-Platz 7, D-78467 Konstanz, Germany).

2.1.xii Hybridisation to the Affymetrix GeneChip® *Arabidopsis* Tiling 1.0R Array

The Affymetrix GeneChip® *Arabidopsis* Tiling 1.0R Array was used to investigate genomic patterns of nucleosome occupancy. The array consists of ~3.2 million probe pairs (perfect-match and mis-match) covering the whole non-repetitive Arabidopsis genome, on a single chip. The probe sequences correspond to the reverse DNA strand and are 25 bp in length. They are tiled with an average spacing of 10 bp apart, resulting in an average resolution of 35 bp.

Mononucleosome DNA fragments were labelled with biotin and hybridised to the array. Labelling and hybridisations were carried out at the Nottingham *A. thaliana* Stock Centre (NASC) Affymetrix service.

2.2 General data analysis protocols

2.2.i Sequence BLAST alignment

Nucleosome DNA fragments were aligned to the Arabidopsis Col-0 reference genome (TAIR 8) using the Time-Logic Tera-probe BLAST (Smith-Waterman algorithm) alignment on a DeCypher server. Matches with 97% or greater similarity (for the clone fragments), or 100% similarity (for the 454 FLX and Solexa DNA fragments) to the reference genome were used in the final datasets. The Tera-probe alignments were carried out by Graham King, Rothamsted Research.

2.2.ii Paired-end matching

The Tera-probe BLAST alignments output was parsed using the programme [Blastparser2.exe] (written in Delphi 7 by Graham King, Rothamsted Research); the output for each end was then imported into separate 'hits' tables in a MySQL database. Assignment of genomic co-ordinates for paired end sequences was obtained via database queries, using the following criteria: where the alignments in each pair were within 620 bp and where the alignments in each pair were on opposite genomic DNA strands. The 5' co-ordinates of the two ends on opposite strands were output to a .CSV file, along with the chromosome and fragment ID. The sequence alignments were sorted into two groups for each dataset: all alignments (which include sequences that align to more than one genomic position with 100% similarity) and unique alignments (which exclude sequences which align to the genome more than once). All sequences were used for analysis of sequence characteristics and linker length distributions, while unique alignments only were used in analyses of nucleosome position. Paired-end matching and sorting into unique and non-unique datasets was performed by Graham King, Rothamsted Research. The Delphi 7 scripts were executed, and the MySQL database was built and accessed using a PC.

2.2.iii Extraction of nucleosome sequences

Nucleosome sequences were extracted from the Col-0 reference Arabidopsis genome using a PERL script called [find_seq.pl] (Table 2.1). All PERL scripts were

executed using a desktop PC and the program Perl Express 2.5 (freeware, url: http://www.softdll.com/phpaspperl/perlexpress.html). Sequences were trimmed into subsets of mononucleosome and linker, using the *LEFT* and *RIGHT* functions in Excel.

Table 2.1 A list of PERL scripts generated and used in this study.

Perl script	Description
find_seq.pl	Extracts sequences from large sequence files using lists of co-ordinates.
end_dinuc.pl	Counts the end dinucleotide (terminal dinucleotide in position 1) and generates a frequency table.
count_dinuc.pl	Counts each dinucleotide in a given sequence, over a moving window of two (so each base is considered twice, the preceding and following one).
rand_seq.pl	Uses the rand function of PERL to generate a series of random numbers within a given size.
count_gc.pl	Counts the numbers of A,C,G and T for each sequences in a sequence file
pos_dinuc.pl	Uses PERLs pos function to generate a list of the first position of a given dinucleotide for each sequence in a sequences file
count_freqs.pl	Uses a list of numbers (the output from pos_dinuc.pl) and generates a frequency table (CSV format)

2.2.iv Detection of periodicity

Periodicities in dinucleotide and linker length distribution were calculated using Fourier analysis, which is a standard method for detecting sequence periodicities (Satchwell *et al.*, 1986). Distributions were once or twice-differenced to eliminated the overall trend in the data (twice-differencing was required in most cases). The periodicity was calculated using a programme written by Steve Powers (Biomathematics and Bioinformatics Dept.) Rothamsted Research, utilising the periodtest procedure in GenStat® (2008, 11th Edition, VSN International, UK). Non-parametric 95% confidence limits for important peaks were calculated by the equation:

$$CI = 0.1026 \times \sqrt{\frac{\text{Variance}_{(\text{periodogram})}}{2}}$$

Periodicities were statistically significant where the peak was larger than the neighbouring peaks + the CL.

2.2.v Kernel Density Estimation

Kernel density estimation was used to smooth the nucleosome distributions in order to reveal long-term trends in the distribution. Kernel density estimates were calculated in Genstat v11, using the method of Sheather and Jones (1991), and 2048 grid points.

2.2.vi Construction of nucleosome occupancy files

To investigate nucleosome occupancy within different annotation regions, occupancy files were constructed. A gene annotation code was designed (Table 2.2) and assigned to each nucleotide position in the Arabidopsis genome in successive rounds, so that each new pass overwrites the previous annotation. Annotation codes were written in the following order: transcriptional units, translational units, 5' and 3' UTRs, exons and introns, 5' UTR exons and intron and 3' UTR exons and intron (Figure 2.1). For alternatively-spliced annotation, introns are given preference over exons since the introns are written after exons. The annotation code appears as a column in the nucleosome occupancy files alongside the nucleotide position.

Table 2.2 Annotation codes assigned to each bp position in Arabidopsis in construction of nucleosome occupancy files.

	miRNA	Misc RNA	Protein coding	Pseudo-gene	rRNA	snoRNA	snRNA	tRNA
	mi	sc	pc	pg	rr	no	sn	tr
NS	NS_mi	NS_sc	NS_pc	NS_pg	NS_rr	NS_no	NS_sn	NS_tr
TU	TU_mi	TU_sc	TU_pc	TU_pg	TU_rr	TU_no	TU_sn	TU_tr
TR	TR_mi	TR_sc	TR_pc	TR_pg	TR_rr	TR_no	TR_sn	TR_tr
U5	U5_mi	U5_sc	U5_pc	U5_pg	U5_rr	U5_no	U5_sn	U5_tr
U3	U3_mi	U3_sc	U3_pc	U3_pg	U3_rr	U3_no	U3_sn	U3_tr
EX	EX_mi	EX_sc	EX_pc	EX_pg	EX_rr	EX_no	EX_sn	EX_tr
IN	IN_mi	IN_sc	IN_pc	IN_pg	IN_rr	IN_no	IN_sn	IN_tr
E5	E5_mi	E5_sc	E5_pc	E5_pg	E5_rr	E5_no	E5_sn	E5_tr
I5	I5_mi	I5_sc	I5_pc	I5_pg	I5_rr	I5_no	I5_sn	I5_tr
E3	E3_mi	E3_sc	E3_pc	E3_pg	E3_rr	E3_no	E3_sn	E3_tr
I3	I3_mi	I3_sc	I3_pc	I3_pg	I3_rr	I3_no	I3_sn	I3_tr

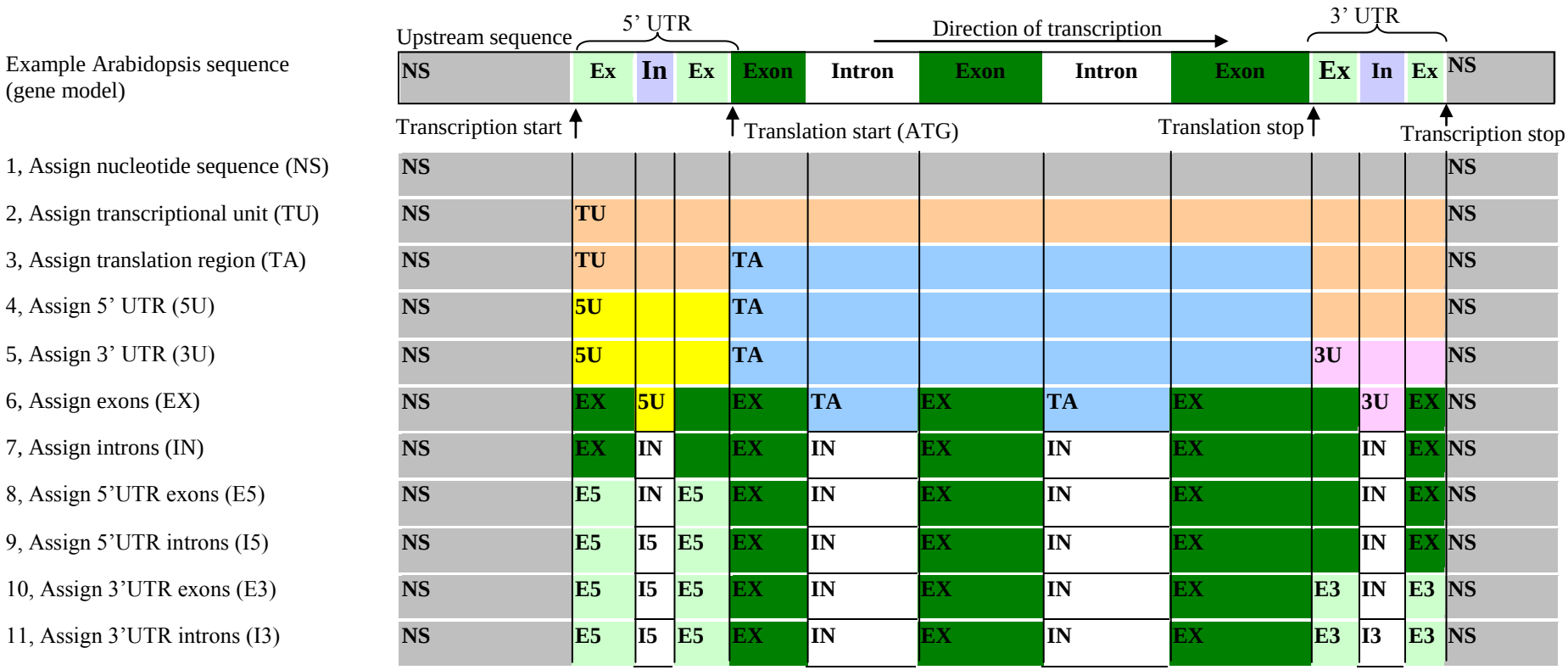

Figure 2.1 Schematic showing the order of assignment of gene annotation to each Arabidopsis nucleotide. The annotation is re-written in successive rounds until all annotations are added.

Nucleosome occupancy was calculated by counting the number times each nucleotide position is represented within a nucleosome DNA fragment. Measures of nucleosome occupancy were added to a column in the nucleosome occupancy files.

To investigate the effect of DNA methylation on nucleosome positioning, previously published DNA methylation data (Lister *et al.*, 2008) were downloaded from supplementary material from the science direct website (**url**:http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4S9G216-).

DNA methylation data was added to the nucleosome occupancy files in the same manner as the nucleosome occupancy. The nucleosome annotation files were constructed by Graham King, Rothamsted Research using a PC.

2.2.vii Description of external datasets

A description of the external (publicly-available) datasets used in this thesis are shown in Table 2.3.

Table 2.3 A description with references of the publicly available datasets used in this thesis.

Dataset	Species	Nucleosome class	Technology	Number of entries	Reference
List_wt_meth	Arabidopsis	^{5m}C	Bs-solexa		Lister *et al*, 2008
List_met-1_meth	Arabidopsis	^{5m}C	Bs-solexa		Lister *et al*, 2008
Johns_Celegans	*C. elegans*	mono	454	327,294	Johnson *et al.*, 2006
Trav_chick_dinuct	Gallus	di (trim)	Traditional Sanger sequencing	49	Satchwell and Travers 1989
Trav_chick_dinucu	Gallus	di (untrim)	Traditional Sanger sequencing	103	Satchwell and Travers 1989
Trav_chick_monoc	Gallus	mono	Traditional Sanger sequencing	152	Satchwell and Travers 1989
Kato_03_hum	*H. sapiens*	di	Traditional Sanger sequencing	1,002	Kato *et al.*, 2003
Kato_05_hum	*H. sapiens*	di	Traditional Sanger sequencing	100	Kato et al, 2005
yeast_mono	*S. cereviseae*	mono	Traditional Sanger sequencing	199	Segal *et al.*, 2006

2.2.viii Analysis of tiling array

Preliminary analyses of tiling data were carried out using the Affymetrix Tiling Array Software (TAS)®.

Chapter 3

The characteristics of Arabidopsis nucleosome sequences

3.1 Introduction

3.1.1 Introduction to nucleosome positioning signals

Early evidence for the non-random positioning of nucleosomes arose from the visualisation of chromatin isolated from chicken erythrocyte, chicken liver and cultured calf cells by electron microscopy (Oudet *et al.*, 1975). It was demonstrated that chromatin consists of a flexible chain of repeating spherical structures (nucleosomes), separated by thin filaments of varying length (DNA linkers). Since then, there has been great interest in discovering what are the determinants of nucleosome positioning.

One of the first mononucleosome sequence datasets produced was from chicken erythrocytes (Drew and Travers, 1986). DNaseI digestion of reconstituted nucleosomes revealed non-random positioning of dinucleotides around the histone octomer, with [AA/TT] dinucleotides tending to be positioned with minor grooves facing in toward the histone octomer. The [CC/GG] dinucleotides were found to be positioned with the minor grooves facing out. The dinucleotides were positioned with a periodicity of 10.17 bp, and thought to be related to the helical twist.

The chicken erythrocyte sequences were further analysed to determine sequence periodicities (Satchwell *et al.*, 1986). Fourier analysis of the position of the dinucleotides along the nucleosome sequences confirmed a periodicity of 10.2 bp for the dinucleotides [AA/TT] along the nucleosome core sequence at positions where the minor groove faces inward towards the histone octomer. In addition, periodicities were also seen for [CC/GG] and [TG/CA] dinucleotides, both with an opposite phasing to the [AA/TT]. However, the number of sequences was too small to reliably quantify periodicity for the other dinucleotides.

The periodic dinucleotide sequence patterns of preferred nucleosome formation sites do not occur along the length of the nucleosomal DNA at positions 10 bp apart, but the characteristic distribution observed is the result of the cumulative addition of many distributions. This can make nucleosome positioning signals difficult to detect in small samples. To overcome this, one approach was adopted whereby sequences

were aligned by the position of the [AA/TT] dinucleotides along nucleosome sequences collected from the literature (Ioshikhes *et al.,* 1996). A total of five different algorithms were used to detect the [AA/TT] repeat, which was demonstrated to have a periodicity of 10.3 (±0.2) bp. The alignments also demonstrated a symmetrical pattern along the length of the DNA molecule. The occurrence of a dinucleotide resulted in the opposite dinucleotide in the complementary position along the same strand of DNA, for example, [AA] repeats at positions 18 and 131-132 had [TT] counterparts at positions 127-8 and 14. The alignments also demonstrated counter-phase behaviour of the [AA/TT] dinucleotides at roughly half-phase (or ~5.5 bp) from each other ($[AANNNTTNNN]_n$).

It is thought that the counter-phase dinucleotide pattern discriminates flexible DNA from curved DNA, since the wedge angles for alternate [AA] and [TT] dinucleotides change the direction of the DNA molecule every 5-6 bases (Ioshikhes *et al.,* 1996). Furthermore, the in-phase dinucleotide periodical repeat is thought to correspond to the curvature of DNA (Bolshoy *et al.,* 1991, Cohanim *et al.,* 2006).

The [AA/TT] nucleosome positioning signal has since been found in nucleosome sequences from many species including *S. cerevisiae* (Albert *et al.,* 2007; Peckham *et al.,* 2007; Segal *et al.,* 2006), *C. elegans* (Johnson *et al.,* 2006) and *D. melanogaster* (Mavrich *et al.,* 2008a). However, it should be noted that while [AA/TT] dinucleotides were found at nucleosome positions in the *D. melanogaster* genome, the [CC/GG] dinucleotide distribution patterns appeared to be a more efficient predictor of nucleosome occupancy.

Whilst the [AA/TT] periodically repeating pattern has been most extensively studied, other dinucleotide distribution patterns contribute to nucleosome positioning. Kato *et al.* (2003) isolated 1,002 dinucleosome sequences from human K562 cells. Following alignment and analyses of these sequences, a dinucleotide repeating pattern consisting of [RR/YY] dinucleotides with a periodicity of 10.38 bp was observed. The [AA/TT] dinucleotide signal, whilst present, was not the largest contributor to the pattern. Further analysis of this set of human dinucleosome sequences revealed a stronger preference for [CC/GG] dinucleotide repeats than [AA/TT] (Kogan *et al.,* 2006). A recent study of nucleosome DNA fragments from

the sheep β-lactoglobulin gene also showed a preference for [CC/GG] periodic repeats (Fraser *et al.,* 2009).

The differences between species for nucleotide position motifs suggest that there may be a species bias toward which particular dinucleotide has a stronger effect on nucleosome positioning. This is important if the third base in a codon does have a role in ensuring the maintenance of the dinucleotide pattern, as the third base of codons are species determined (Cohanim *et al.,* 2006). Further evidence for a species bias in nucleosome positioning signals arose from alignments of sequences around splice-site junctions (Kogan and Trifonov, 2005). These sequences, when alignments were restricted by species, revealed a preference for [CC/GG] dinucleotide patterns in human and mouse sequences and [AA/TT] preferences in *Arabidopsis* and *C. elegans*.

Successive competitive reconstitution experiments used mouse nucleosomal DNA, isolated nucleosome sequences from the most strongly-positioned nucleosome in mouse chromatin (Widlund *et al.,* 1997). The nucleosome DNA sequences revealed five classes of nucleosome positioning preferences including repeating [CA] dinucleotides, repeating [CAG] trinucleotides, and [TATA] tetrads [TATAA(A/C)CG(T/C)C] which represent the strongest positioning motif isolated *in vivo*. Fluorescent *in-situ* hybridisation of the nucleosomal sequences from mouse metaphase chromosomes revealed these sequences occur around the centromeric region (Widlund *et al.*, 1997).

3.1.2 Aims of study in this chapter

To date, the studies of nucleosome positioning signals have involved mammalian, insect and yeast genomes. Plant genomes have yet to be studied. With this in mind, the following hypotheses were formulated:

1. Arabidopsis nucleosome sequence composition is different from the genome average and there are differences in nucleosome and linker DNA.

2. Arabidopsis nucleosome sequences display specific distributions in their dinucleotide complement which correspond to those known to contribute to the bending of DNA around the nucleosome.

3. Arabidopsis nucleosome sequence preferences are similar to those discovered in other taxa.

To test the hypotheses above, the approach taken here involved sequencing, employing different technologies, of mono and dinucleosome fragments isolated from Arabidopsis chromatin. The different technologies applied allowed increasing resolution of the nucleosome positions.

3.2 Methods

Sequence processing and PERL scripts are described in Chapter 2, Table 2.1.

3.2.1 Determination of end-dinucleotides at the site of MNase activity

The frequency of the occurrence of each dinucleotide at the active site of MNase was determined using a PERL script: [end_dinuc.pl]. Dinucleotides from both ends of each sequence were counted, the nucleotide at the edge being considered the first base regardless of which end of the sequence it had originated from.

3.2.2 Calculation of observed/expected ratios for dinucleotides at sites of MNase activity

The dinucleotide occurrence for each dinucleotide in the genome was counted using a PERL script: [count_dinuc.pl].

The expected dinucleotide values for each dinucleotide were calculated by :

$$\frac{\sum X_1 X_2}{n}$$ Equation 3.1

where X_1 = base 1, X_2 = base 2, so that $\sum X_1 X_2$ is the sum of occurrence at base 1 multiplied by occurrence at base 2, moving along the forward strand of genome, for consecutive pairs of bases X_1 and X_2; and n is the total number of dinucleotides in the genome.

The ratio of the observed occurrence for a sequence to that expected for the whole genome was calculated as:

$$R = \left(\frac{\sum X_1 X_{2(seq)}}{n_{(seq)}}\right) \Big/ \left(\frac{\sum X_1 X_{2(genome)}}{n_{(genome)}}\right)$$

Equation 3.2

where the numerator is the observed proportion of occurrence in a sequence.

The statistical significance was determined using a Chi-squared test.

The Arabidopsis Col-0 reference genome was downloaded from the TAIR (version 8) website (url at: www.arabidopsis.org). Random Arabidopsis sequences of length 147 bp were generated using a PERL script: [rand_seq.pl].

The human genome (v36.3), chicken genome (v2.1), *D. melanogaster* genome and *S. cerevisiae* genome were downloaded from the National Centre for Biotechnology Information (NCBI) ftp site (url at: ftp://ftp.ncbi.nih.gov/genomes/).

3.2.3 Calculation of GC content

GC content of DNA fragments within each dataset was calculating using a PERL script: [count_gc.pl]. Mann-Whitney U tests of the medians were performed using GenStat® (2008, 11th Edition, VSN International, UK).

Arabidopsis BLAST datasets (TAIR version 8) were downloaded from the TAIR website (url at: www.arabidopsis.org).

3.2.4 Determination of dinucleotide frequencies

Dinucleotide frequencies of nucleosome sequences were determined as described in section 3.2.2. Observed values were calculated from the frequencies of each individual base within that sequence. Observed/expected ratios for each sequence were calculated as:

$$R = \left(\frac{\sum X_1 X_{2(seq)}}{n_{(seq)}}\right) \Bigg/ \left(\left(\frac{\sum X_{1(seq)}}{n_{(seq)}}\right) \times \left(\frac{\sum X_{2(seq)}}{n_{(seq)}}\right)\right)$$

Equation 3.3

where X_1 = base 1, X_2 = base 2 moving along the sequence of the forward strand for pairs of bases X_1 and X_2; and n = DNA fragment length.

3.2.5 Distribution of dinucleotides along the nucleosome DNA fragment

The positions of each dinucleotide within nucleosome sequences were determined using a PERL script: [pos_dinuc.pl]. Frequency tables were constructed using a PERL script: [count_freqs.pl].

3.2.6 Periodicity of dinucleotide distribution.

Fourier analysis and autocorrelation tests of the frequency tables (calculated in 3.2.5) used the 'periodtest' and time series procedures in Genstat v11. Fourier analysis is described in Chapter two, Section 2.2.iv.

3.2.7 Construction of a nucleosome sequence database

Nucleosome sequence data were collected from various publicly-available sources and used to populate a relational nucleosome positioning database implemented in Microsoft Access. Data sources are summarised in Table 3.1 and an entity-relationship diagram of the database is shown in Figure 3.1.

Table 3.1 Sources of publicly-available nucleosome sequence data collected and used to populate a nucleosome sequence database.

Data source	Format	No. of sequences	Method of isolation	Year entered into database	Reference
Andrew Travers (personal communication)	MS Excel	177	MNase digestion of chromatin	2006	Satchwell and Travers (1989)
Andrew Travers (personal communication)	MS Excel	157	MNase/S1 digestion of chromatin	2006	Drew and Travers (1986)
I. Ioshikhes (via Graham King)	FASTA	207	Mixture of MNase/DNase1 isolated nucleosomes	2006	Ioshikhes *et al.* (1996)
NPRD (Download)	FASTA	100	Mixture of MNase/DNase1 isolated nucleosomes	2006	Levitsky *et al.* (2005)

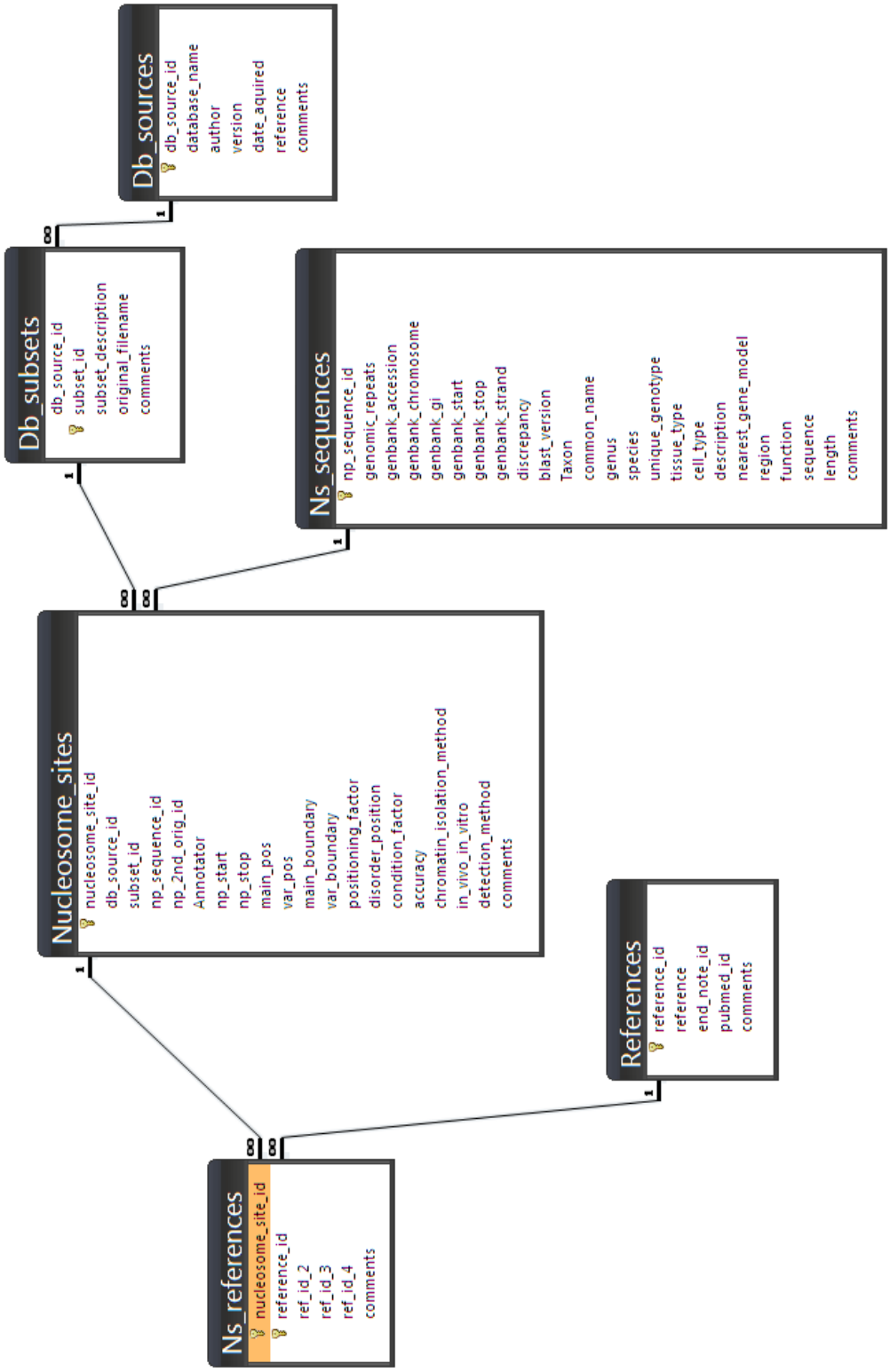

Figure 3.1 Entity-relationship diagram showing tables and fields of the relational nucleosome sequence database.

3.3 Results

Much effort was expended in attempting to acquire nucleosome positioning data from the Arabidopsis tiling arrays, with four (nucleosome) datasets and one control dataset being generated. Following preliminary analysis of the tiling array datasets, a nucleosome signal which was uniformly distinct from the control could not be detected, and therefore the results are not discussed.

3.3.1.i Generation of Arabidopsis nucleosome DNA fragments

Sequencing of nucleosome DNA fragments isolated from Arabidopsis leaf tissue chromatin resulted in four datasets. An overview of this is shown in Table 3.2.

Table 3.2 Arabidopsis datasets generated in this study

Dataset	Origin	Raw sequencing reads	Reads aligned	Unique hits	Mono-nucleosome	Di-nucleosome
Ath_01_wt	WT_clone	576	550	476	N/A	476
Ath_02_wt	WT 454	43,420	25,227	19,548	19,548	N/A
Ath_03_wt	WT (F)	1,536,234	1,097,364	908,572	16,589*	924,381*
	WT (R)	1,536,234				
Ath_04_MET1	*MET1* (F)	2,543,279	1,702,915	1,280,099	1,054,734*	98,365*
	MET1 (R)	2,543,279				
Ath_05_wt	WT (F)				3,557,287	
	WT (R)					

*including repeats, F: forward DNA strand, R: reverse DNA strand

3.3.1.ii Dinucleosome DNA fragments from a clone library

A dinucleosome library, containing over 3,000 clones, was constructed from Arabidopsis leaf tissue. A total of 576 cloned fragments were sequenced. Dinucleosome DNA fragments were aligned to the Arabidopsis Col-0 reference genome (TAIR 8) using the Time-Logic Tera-probe BLAST (Smith-Waterman algorithm) alignment on a DeCypher server. Matches with 97% or greater similarity to the reference genome and over 294 bp in length were included in the final set of 476 fragments. This set was nominated as *Ath_01_wt*.

Fragments which were less than 294 nucleotides in length were removed from the final set to give the set for analysis. The theoretical minimum nucleotide length of a dinucleosome fragment is assumed to be 294 nucleotides, which is the length of two nucleosomes (1 nucleosome = 147 nucleotides) (Richmond and Davey, 1993). However, this does not take into consideration minimum linker length or steric hindrance. In addition, fragments which aligned to organelle genomes were also removed from all datasets, since organelles do not form true nucleosomes *in vivo* (Salganik *et al.*, 1991).

3.3.1.iii Dinucleosome DNA fragments sequenced by 454 FLX technology

Arabidopsis dinucleosome DNA was sequenced using a 1/8th 454 pico-titre plate. This generated a collection of 43,420 sequence reads, varying from 41 to 368 nucleotides in length. The DNA sequences were aligned to the Arabidopsis Col-0 reference genome (TAIR8), as in the previous section (3.3.1.ii). Fragments which aligned with 99% similarity or more were included in the final set. Fragments which aligned to more than one genomic region (genomic repeats) were omitted, leaving a final set of 19,548 fragments. This set was nominated as *Ath_02_wt*.

Since 454 FLX sequence reads do not extend to dinucleosome length, fragments were trimmed to 147 nucleotides from the 5' end and treated as mononucleosome fragments in further analyses, as shown in the schematic Figure 3.2. Length distributions of datasets *Ath_01_wt* and *Ath_02_wt* are shown in Figure 3.3.

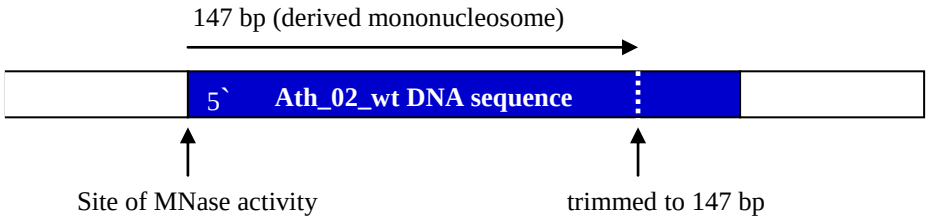

Figure 3.2 Schematic showing the processing of Arabidopsis dinucleosome fragments from dataset *Ath_02_wt.* Sequences were trimmed to 147 bp from the sites of MNase activity and treated as mononucleosomes.

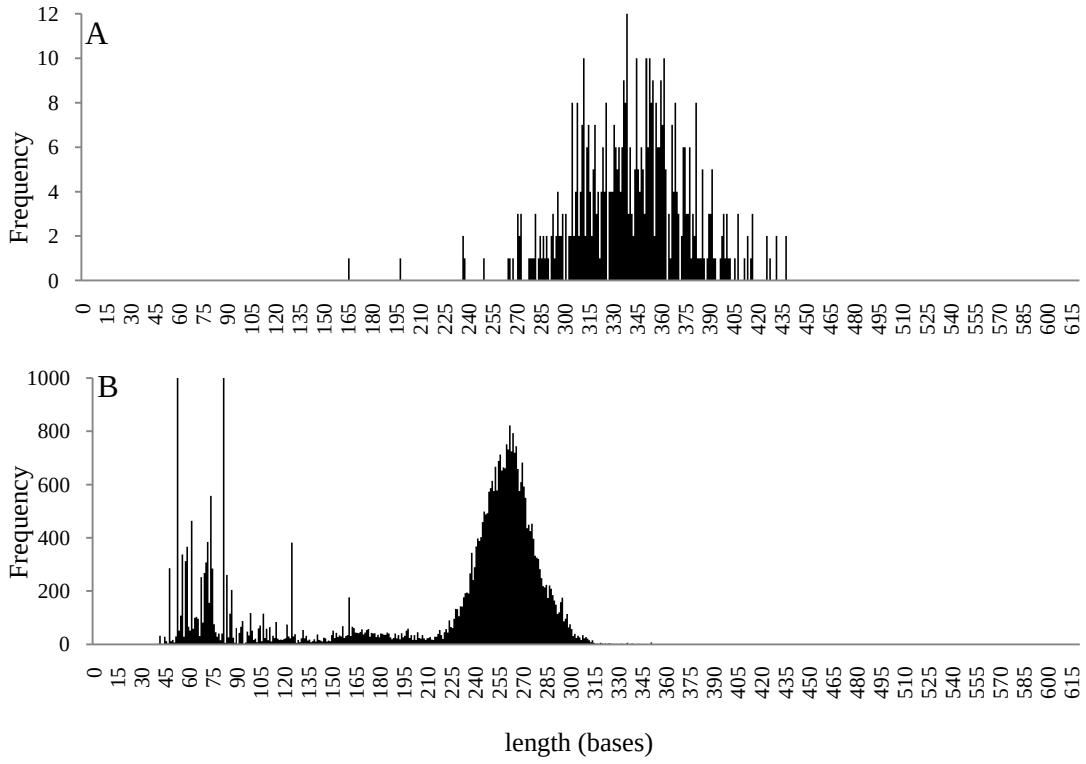

Figure 3.3 Distributions of dinucleosome DNA fragment lengths, **A:** length distribution of the nucleosome fragments of dataset *Ath_01_wt*. **B:** shows the length distribution of sequences from dataset *Ath_02_wt*, prior to trimming to 147 bp.

3.3.1.iv Nucleosome DNA fragments sequenced using Illumina (Solexa) technology

A total of 2 µg each of Arabidopsis wild-type (Col-0) and anti-sense *MET1* (C24 ecotype) mono (*MET1*) and dinucleosome (wild-type) fragments were sequenced using the Illumina Genome Analyzer sequencing technology, on 1/8 th of a channel. For an experiment where samples are mixed and sequenced in the same channel, fragments are tagged with a 4 bp 'barcode' which allows the samples to be identified and sorted following the sequencing reactions. The barcode and primer account for 18 bp of the sequence. The nucleosome fragment sequences obtained were thus 32 bp from each end of the fragment (see schematic Figure 3.4.). A total of 1,536,234 pairs of reads (paired having read from each end of the fragment) were generated for wild-type and 2,5432,79 pairs of reads were generated for anti-sense *MET1*.

Both mononucleosome and dinucleosome DNA fragments were isolated from the wild-type and anti-sense *MET1* tissues and pooled. For sequencing, the Solexa project was outsourced to and external company. Unfortunately, only the mononucleosome

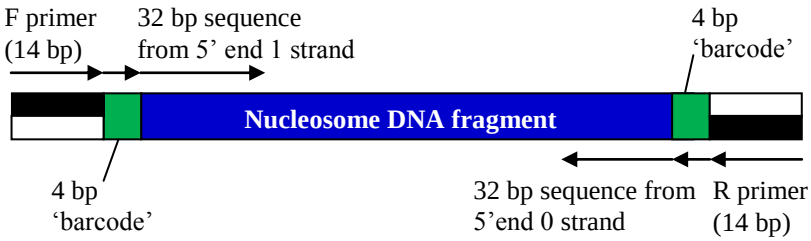

Figure 3.4 Schematic showing the process of paired-end Illumina sequencing of Arabidopsis nucleosome DNA fragments. Different 4 bp barcodes are applied to each sample when multiple samples are mixed prior to sequencing to allow for sorting later.

fragments from the anti-sense *MET1* dataset and dinucleosome fragments from the wild-type following excision of fragments from agarose gels. It was decided to complete the analysis of nucleosome positions with these fragments. However it became clear later that a further dataset was required to enable comparisons between the wild-type and anti-sense *MET1* datasets. Therefore a further wild-type mononucleosome dataset was generated, but due to the late arrival of this dataset, it will be presented here, but not discussed until Chapter 5.

The Solexa DNA sequence fragments were aligned to the Arabidopsis Col-0 reference genome, (TAIR8), as in 3.3.1.ii. Positions were assigned to pairs of sequences where they were aligned with 100% sequence similarity to the Arabidopsis reference genome, the pairs being aligned within 620 nucleotides from each other and on opposite DNA strands. This resulted in a total of 1,097,364 wild-type and 1,702,915 anti-sense *MET1* (di) nucleosome positions. These sets were nominated *Ath_03_wt* and *Ath_04_MET1*, respectively. A further wild-type mononucleosome dataset was nominated *Ath_05_wt* (n = 3,557,287). The length distributions of *Ath_03_wt*, *Ath_04_MET1* and *Ath_05_wt* are shown in Figure 3.5.

Datasets *Ath_03_wt* and *Ath_04_MET1* contain all aligned nucleosome sequences. Subsets of each dataset were constructed, where sequences that aligned to a single genomic region were defined as unique hits as shown in Table 3.2. For sequence analysis, all sequences which aligned to nuclear chromosomes 1-5 were used for each dataset. For the purposes of analysis, nucleosome and linker regions were separated by splitting the dinucleosome sequences 147 bp from the 5' and 3' ends as

shown in the schematic (Figure 3.6). The resulting fragments were nominated as *(WT/MET1)_left*, (*WT/MET1)_di_linker* and *(WT/MET1)_right.*

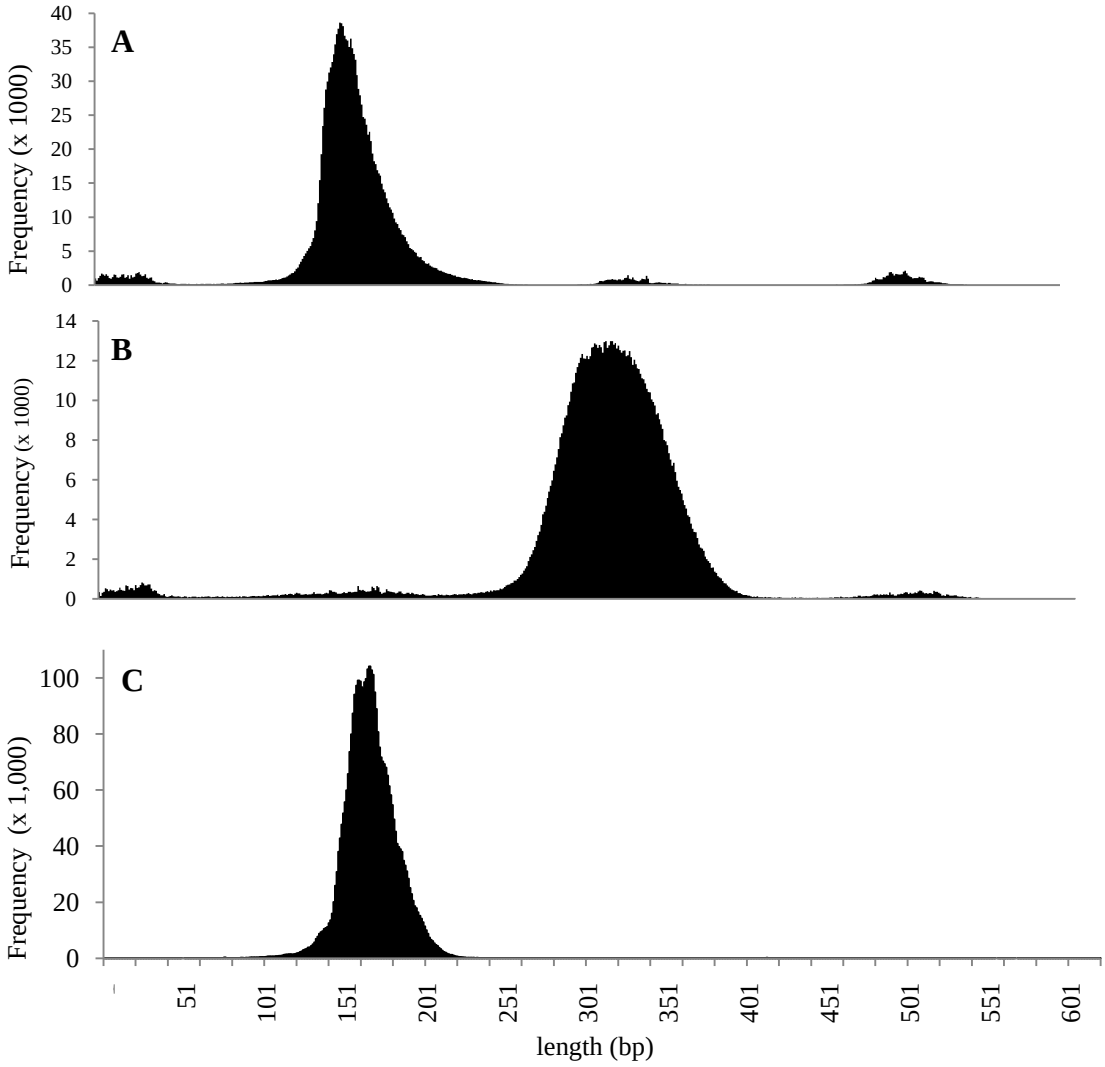

Figure 3.5 Length distribution of mono and di nucleosome sequences of datasets A: *Ath_04_MET1*, **B**: *Ath_03_wt* and **C:** *Ath_05_wt*

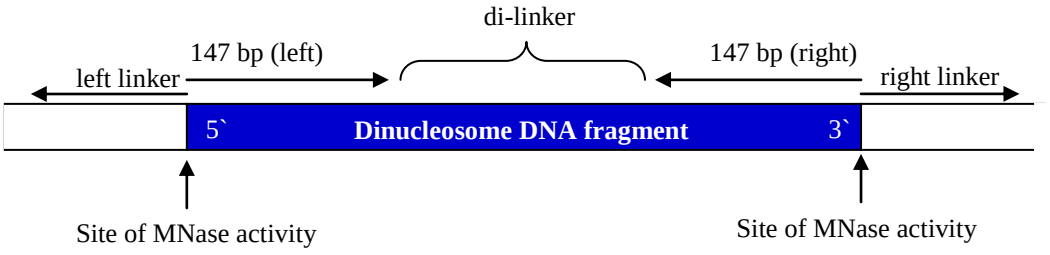

Figure 3.6 Schematic showing the division of dinucleosome sequences from datasets *Ath_03_wt* and *Ath_04_MET1* into nucleosome and linker, and the generation of derived linker datasets.

Mononucleosome fragments were identified as fragments between 130 - 177 bp in length, and named the '*WT_mono'* and '*MET1_mono'* datasets for wild-type and anti-sense *MET1*, respectively. In addition, the sequences 52 bp to the left and right of the nucleosome DNA fragments were collected, and allocated into '*left_linker'* and '*right_linker'* datasets, respectively.

For the purposes of sequence analysis, the *wt_left* (Table 3.3) and *MET1_mono* datasets were used, in order for the wild-type and anti-sense *MET1* datasets to be of a comparable size. For comparison to genomic DNA sequences, a random set of 584,032 sequences between 147 and 150 bp in length were generated from the Arabidopsis Col-0 reference genome (TAIR8) using PERL scripts: rand_seq.pl and find_seq.pl (section 3.2.2). A summary of the derived datasets is shown in Table 3.3.

Table 3.3 A summary of datasets derived from *Ath_03_wt* and *Ath_04_MET1*

Ath derived	Description	Length (bp)	No of sequences
wt_left	147 bp from 5' edge dinucleosome fragment (+ strand)	147	924,381
wt_right	-147 bp from 3' edge dinucleosome fragment (- strand)	147	924,381
wt_mono	DNA fragments >130 bp <177 bp	130-177	16,589
wt_di-linker	DNA fragment remaining from dinucleosome after subtraction of left and right nucleosome	10-326	924,381
MET1_left	147 bp from 5' edge dinucleosome fragment (+ strand)	147	98,365
MET1_right	-147 bp from 3' edge dinucleosome fragment (- strand)	147	98,365
MET1_mono	DNA fragments >130 bp<177 bp	130-177	1,054,734
MET1_di-linker	DNA fragment remaining from dinucleosome after subtraction of left and right nucleosome	10-326	98,365
wt_left_linker	-52 bp from 5' edge nucleosome fragment	52	924,381
wt_right_linker	+52 bp from 3' edge nucleosome fragment	52	924,381
MET1_left_linker	-52 bp from 5' edge nucleosome fragment	52	16,589
MET1_right_linker	+52 bp from 3' edge nucleosome fragment	52	16,589
Ath_rand	Randomly generated from the Arabidopsis Col-0 reference genome	147-150	584,032

3.3.2.i Specificity of micrococcal nuclease

In order to determine whether the digestion of chromatin had been affected by any bias in MNase base preference, the incidence of each base was recorded for the first and last position from the forward strand of each sequence in each Arabidopsis dataset (with the exception of *Ath_03_wt*, where the nucleotide at the MNase site was recorded). Nucleotides [W] accounted for between 82 and 97% of the nucleotides in the first and last positions in the Arabidopsis nucleosome fragments, with [WG] being the most favoured dinucleotide, this occurring between 40 and 49% of the Arabidopsis sequences. Interestingly, the dinucleotide [WC] occurred at a lower frequency, indicating the MNase preference in for [W] followed by [G]. However, one set (*Ath_03_wt*) did show an increased preference for [GT] (9%) at the MNase cut-site when compared to the other Arabidopsis data. The percent of occurrence of [W], [WG] and [WC] are shown in Table 3.4.

Table 3.4 Frequency of [W], [WG] and [WC] at the site of MNase activity for each nucleosome dataset

Arabidopsis dataset	Number of ends of nucleosome DNA fragments	Nucleotide/dinucleotide (% nucleosome fragments) [W]	[WG]	[WC]
Ath_01_wt	952	94	49	13
Ath_02_wt	19,548	97	44	22
Ath_03_wt	2,554,122	82	40	15
Ath_04_MET1	2,306,198	84	40	15

To test the hypothesis that the occurrence of each dinucleotide at the site of MNase activity is not random, the observed/expected ratios were calculated (equation 3.2), comparing the occurrence of dinucleotides at the MNase cut-site to the occurrence of each dinucleotide throughout the Arabidopsis Col-0 reference (TAIR8) genome (equation 3.1). The ratios are shown in Figure 3.7.

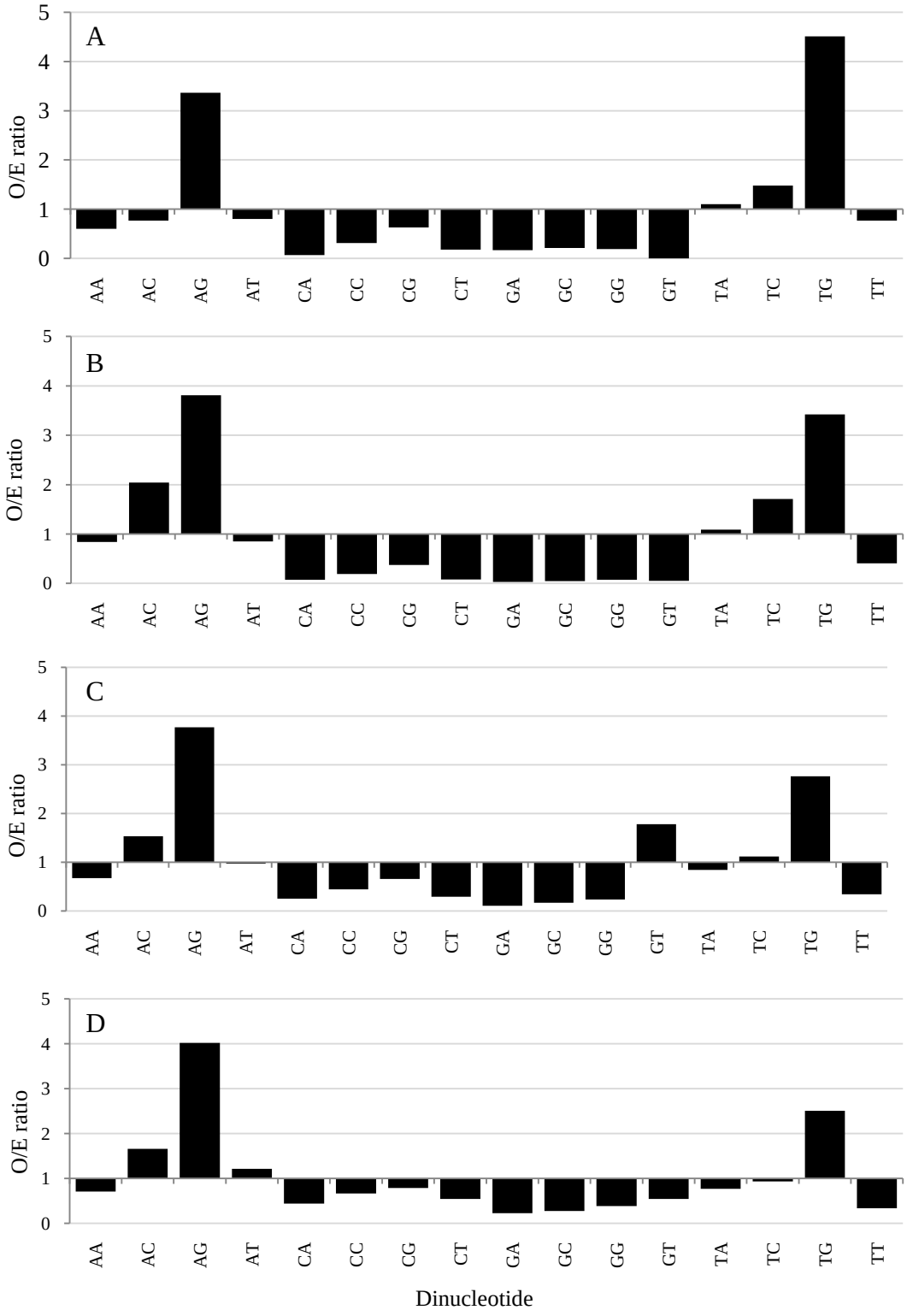

Figure 3.7 Observed/expected ratios for the occurrence of each dinucleotide at the site of MNase activity for Arabidopsis datasets, A: *Ath_01_wt*, **B**: *Ath_02_wt*, **C**: *Ath_03_wt* and **D**: *Ath_04_MET1*

The ratios show significantly greater than expected occurrences (p=0.05 (Chi-Square)) of [AG] in all Arabidopsis datasets, and [TG] in the *Ath_01_wt* and *Ath_02_wt* datasets.

The data obtained here suggests that the Arabidopsis chromatin has a bias towards a greater than expected occurrence of [WG] at the MNase cut-site, when compared to dinucleotide occurrence throughout the whole genome. To determine if this is a unique feature of this genus, the observed/expected ratios of dinucleotide occurrence at the site of MNase activity were calculated for other taxa. Data were taken from publicly available sources (chicken (Drew and Travers, 1986), yeast (Segal *et al.*, 2006), *C. elegans* (Johnson *et al.*, 2006) and human (Kato *et al.*, 2005, Kato *et al.*, 2003), where the nucleosome DNA fragments were prepared by MNase digestion of chromatin. The significance of the occurrence of each dinucleotide compared to the genome was determined using a Chi-squared test. The ratios are shown in Table 3.5, significantly higher than expected occurrences are shown in bold; while significantly lower than expected occurrences are indicated with an asterisk.

There appears to be considerable variation in the observed/expected ratios between each species tested, and within the three chicken sets. The higher than expected occurrence of [WG] observed at Arabidopsis sites on MNase activity were also observed in the *C. elegans* dataset. The chicken and yeast datasets show a bias toward either [AG] or [TG], with the exception of the trimmed chicken dataset. The two human datasets show a bias towards [CC], [CG] and [GG] dinucleotides at the site of MNase activity, when compared to the whole human genome. As the previously reported sequence preference for MNase activity has been determined to be [WG], these results suggest that there may be more factors involved when chromatin is digested with this enzyme such as chromatin conformation and DNA accessibility.

Table 3.5 Observed/expected ratios of dinucleotide occurrence at the site of MNase activity of nucleosome sequences. The Arabidopsis datasets are shown for comparison to the publicly-available data for other taxa. Chi-squared calculated using percent occurrence of dinucleotides at MNase/whole genome. **Bold** = significantly higher than expected; * = significantly lower than expected(p=0.05).

	Ath_01_wt *n*=950	*Ath_02_wt* *n*=25,227	*Ath_03_wt* *n*=2,194,728	*Ath_04_ MET1* *n*=3,405,830	chick_di_ trim *n* = 98	chick_di_ untrim *n* = 206	chick_mono _untrim *n* = 304	yeast_ mono *n* = 398	*C.elegans*_ mono *n*=327,294	kato_03_ human *n* = 2004	kato_05_ human *n* = 200
AA	0.6*	0.8*	0.7*	0.7*	1.5*	1.0*	1.2*	1.5*	0.5*	0.2	0.1
AC	0.8*	2.0*	1.5*	1.7*	**4.3**	1.4*	0.7*	0.6*	2.7	2.8	2.0*
AG	**3.4**	**3.8**	**3.8**	**4.0**	0.8*	**4.1**	2.1	2.0*	**3.7**	2.4	2.1
AT	0.8*	0.9*	1.0*	1.2*	**3.7**	2.6	0.8*	**4.0**	0.6*	0.5*	0.3*
CA	0.1*	0.1*	0.3*	0.4*	0.3*	**4.5**	**3.3**	**5.0**	0.1*	2.3	2.2
CC	0.3*	0.2*	0.4*	0.7*	1.8*	0.0*	1.6*	0.0*	0.4*	**5.2**	**5.7**
CG	0.6*	0.4*	0.7*	0.8*	2.0*	0.9*	**8.1**	0.2*	0.7*	**11.0**	**15.0**
CT	0.2*	0.1*	0.3*	0.5*	1.1*	**3.2**	**3.5**	2.2	0.2*	1.9*	1.0*
GA	0.2*	0.0*	0.1*	0.2*	3.1	0.9*	2.9	1.5*	0.1*	2.1	1.2*
GC	0.2*	0.0*	0.2*	0.3*	2.5	0.4*	0.9*	0.0*	0.1*	2.3	2.6
GG	0.2*	0.1*	0.2*	0.4*	2.6	0.4*	1.6*	0.0*	0.1*	**4.8**	**5.7**
GT	0.0*	0.1*	1.8*	0.5*	**6.3**	1.3*	1.2*	1.2*	0.1*	2.9	2.6
TA	1.1*	1.1*	0.8*	0.8*	0.0*	1.8*	1.1*	**3.3**	0.8*	0.4*	1.1*
TC	1.5*	1.7*	1.1*	0.9*	**3.4**	0.9*	2.0*	1.2*	2.8	2.6	**3.7**
TG	**4.5**	**3.4**	2.8	2.5	0.3*	**4.1**	**3.8**	**5.0**	**3.5**	1.9*	1.9*
TT	0.8*	0.4*	0.3*	0.3*	1.3*	1.2*	1.5*	0.8*	0.4*	0.2	0.2*

n = number of ends of nucleosome DNA fragments

3.3.2.ii The GC content of nucleosomal DNA fragments

To test the hypothesis that nucleosomal DNA has a higher GC content than the average for the genome, the GC content of each sequence was calculated for the *wt_left*, *MET1_mono* and random genomic sequences subsets. The distribution of GC content (expressed as a percentage) for both subsets is shown in Figure 3.8.

The mean GC contents of the *wt-left*, *MET1_mono* and *Ath_rand* subsets were 41.3%, 40.1% and 36.0%, respectively. The GC contents of both nucleosome DNA subsets were higher than the mean for the random genomic sequences and the genome average of 36%.

As genic regions of the Arabidopsis genome are known to have higher GC content than the intergenic regions, GC contents were calculated for Arabidopsis exons, introns, 5'-UTR's, 3'-UTR's, transposable elements and intergenic sequences. The nucleosome sequences have a GC content most similar to the exon subset (41.3% for the *wt_left* compared to 42.5% for the exon), and all other subsets have lower GC contents. This suggests that the nucleosome datasets are likely to be enriched for exon sequences.

In order to test whether there were statistically significant differences between the GC contents of the nucleosome DNA datasets and the Arabidopsis genomic subsets, two-sample Mann-Whitney-U tests were performed. Each Arabidopsis genomic subset was tested against the GC content of *wt_left* and *MET1_mono*.

All tests showed significant differences between medians, although the extreme size of the datasets tested may have influenced the outcome of the test. The results of the Mann-Whitney-U tests are shown in Table 3.6.

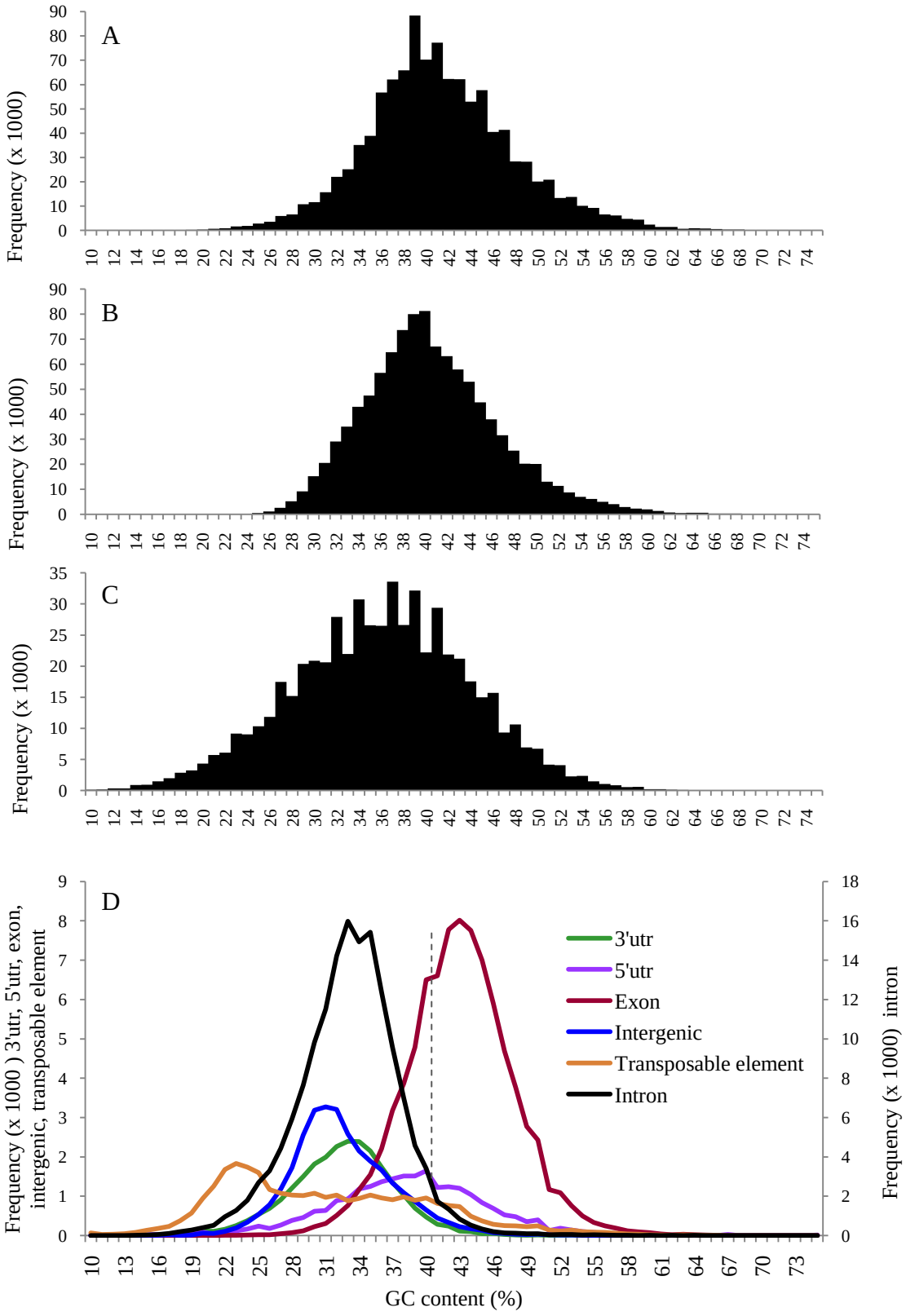

Figure 3.8 Distribution of GC content (%) for each sequence in the Arabidopsis datasets A: *wt_left* nucleosome DNA, **B:** *MET1_mono* nucleosome DNA, **C:** *Ath_rand* and **D:** Arabidopsis TAIR 8 blast sets (3'UTR's, 5'UTR's, exons, introns, transposable elements and intergenic sequences). The dashed line indicates the mean for *wt_left* nucleosomes. The distribution of intron sequences is on the right-hand *y*-axis for clarity.

Table 3.6 Mean, standard deviation and median of the distribution of GC content (%) of *wt_left* and *MET1_mono* nucleosome sequences (highlighted in yellow) and the Arabidopsis (TAIR 8) genomic subsets (exon, intron, 3'UTR, 5'UTR, transposable elements and intergenic DNA sequences). In addition, the results of Mann-Whitney tests (p-values) of each genomic subset vs. *wt_left* and *MET1_mono* are reported, testing the null hypothesis that the medians are equal.

Fragment subset	mean ($\bar{y}$)	stdev	median	*wt_left* test statistic	*wt_left* p-value	*MET1*_mon test statistic	*MET1*_mon p-value
Exon (*n*=154,240)	42.5	4.9	42.5	73.71	p<0.001	137.41	p<0.001
wt_left	41.3	6.6	41.0				
MET1_mono	40.1	6.2	39.6				
5'UTR (*n*=24,267)	38.0	7.9	38.0	71.10	p<0.001	44.45	p<0.001
Ath_rand	36.0	8.0	36.3	405.81	p<0.001	316.67	p<0.001
Intron (*n*=86,850)	32.6	4.6	32.8	477.06	p<0.001	441.89	p<0.001
3'UTR(*n*=25,273)	32.1	4.9	32.2	209.44	p<0.001	196.47	p<0.001
Intergenic (*n*=31,089)	32.0	4.6	31.5	231.41	p<0.001	218.93	p<0.001
Transposable element (*n*=31,060)	30.7	9.3	29.4				

3.3.2.iii Dinucleotide occurrence in nucleosome DNA fragments

Specific dinucleotides have previously been shown to be important for nucleosome positioning in many studies (Satchwell *et al.,* 1986, Ioshikhes *et al.*, 1996) in terms of dinucleotide frequency and position. The dinucleotides [AA], [TT], [AT] and [TA] have been shown to be important, with many nucleosome prediction algorithms based on patterns of distribution of these dinucleotides (Segal *et al.,* 2006, Ioshikhes *et al.,* 2006).

To test the hypothesis that Arabidopsis nucleosome positions are influenced by local sequence composition, dinucleotide observed/expected ratios were calculated for the Arabidopsis nucleosome DNA sequences for datasets *wt_left* and *MET1_left*. The expected values were calculated from the local nucleotide count and observed values were calculated from the local dinucleotide count (equation 3.3). The observed/expected dinucleotide ratios were also calculated for the whole Arabidopsis genome.

The mean observed/expected ratios of dinucleotide occurrence for each dataset are shown in Figure 3.9. There is little variation between the mean observed/expected ratios of the *wt_left* and *MET1_mono* nucleosome datasets. Both [AA] and [TT] dinucleotide frequencies occurred more often than expected (1.07 and 1.08 respectively) when compared to the local nucleotide count. However, when the nucleosome sequence distributions are compared to the whole genome, there appears to be very little variation. The dinucleotides [CG] and [TA] frequencies were both lower than expected (0.90 and 0.67, respectively for *wt_left*, 0.91 and 0.68, respectively for *MET1_mono*) when compared to the local nucleotide count, which is also observed in the whole genome.

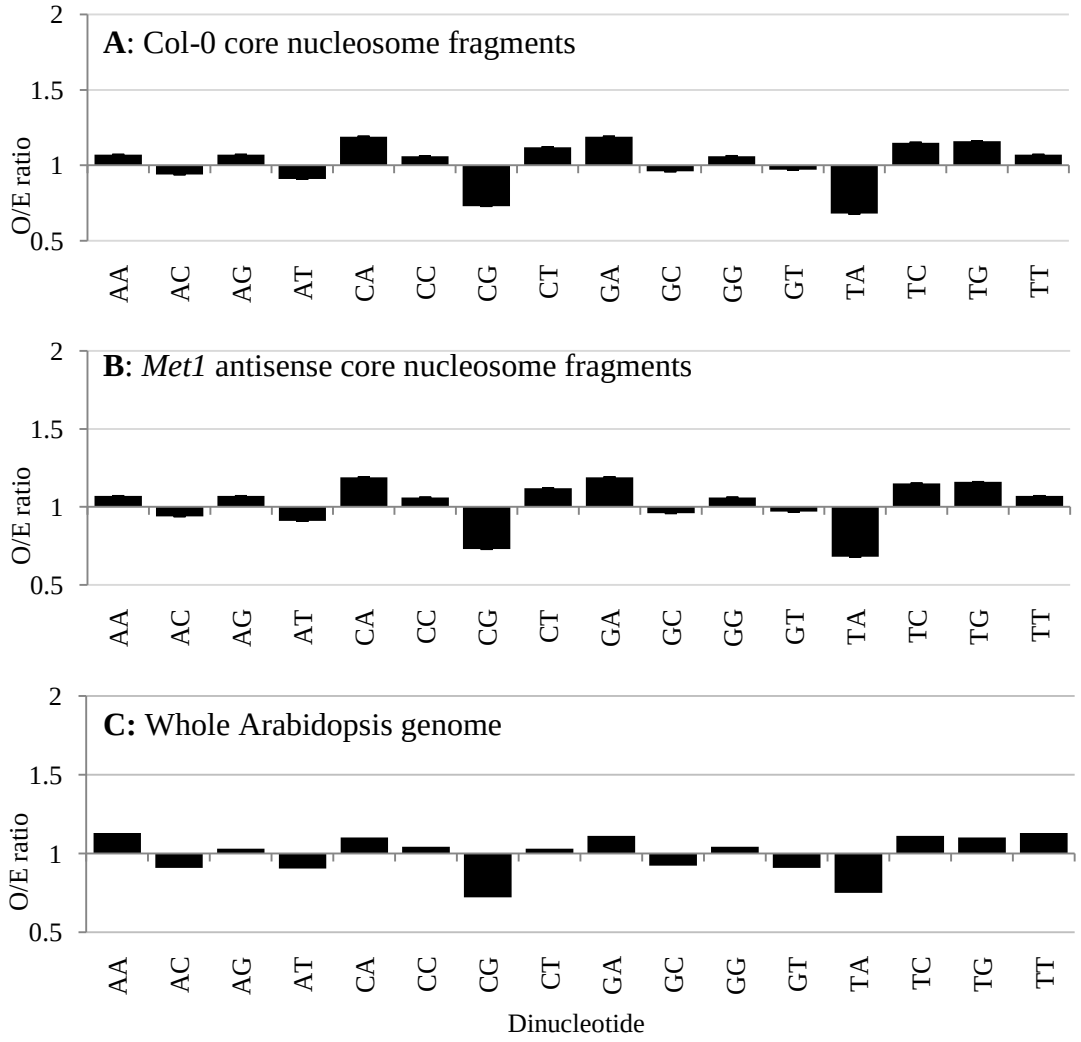

Figure 3.9 Mean observed/expected ratios of the occurrence of each dinucleotide for each DNA sequence in the nucleosome Arabidopsis datasets A: *wt_left* nucleosome DNA, **B**: *MET1_mono* nucleosome DNA, **C**: the whole genome. Expected values were calculated from the local DNA sequence composition. Error bars show standard error of the mean.

For comparison, the dinucleotide observed/expected ratios were also calculated for the linker DNA sequences from datasets *wt_di-linker* and *MET1_di-linker*, Figure 3.10. There appears to be slightly more variation in the occurrence of dinucleotides between the linker datasets than there is between the nucleosome datasets. Neither nucleosome nor linker datasets appear to show a strong bias towards a specific dinucleotide.

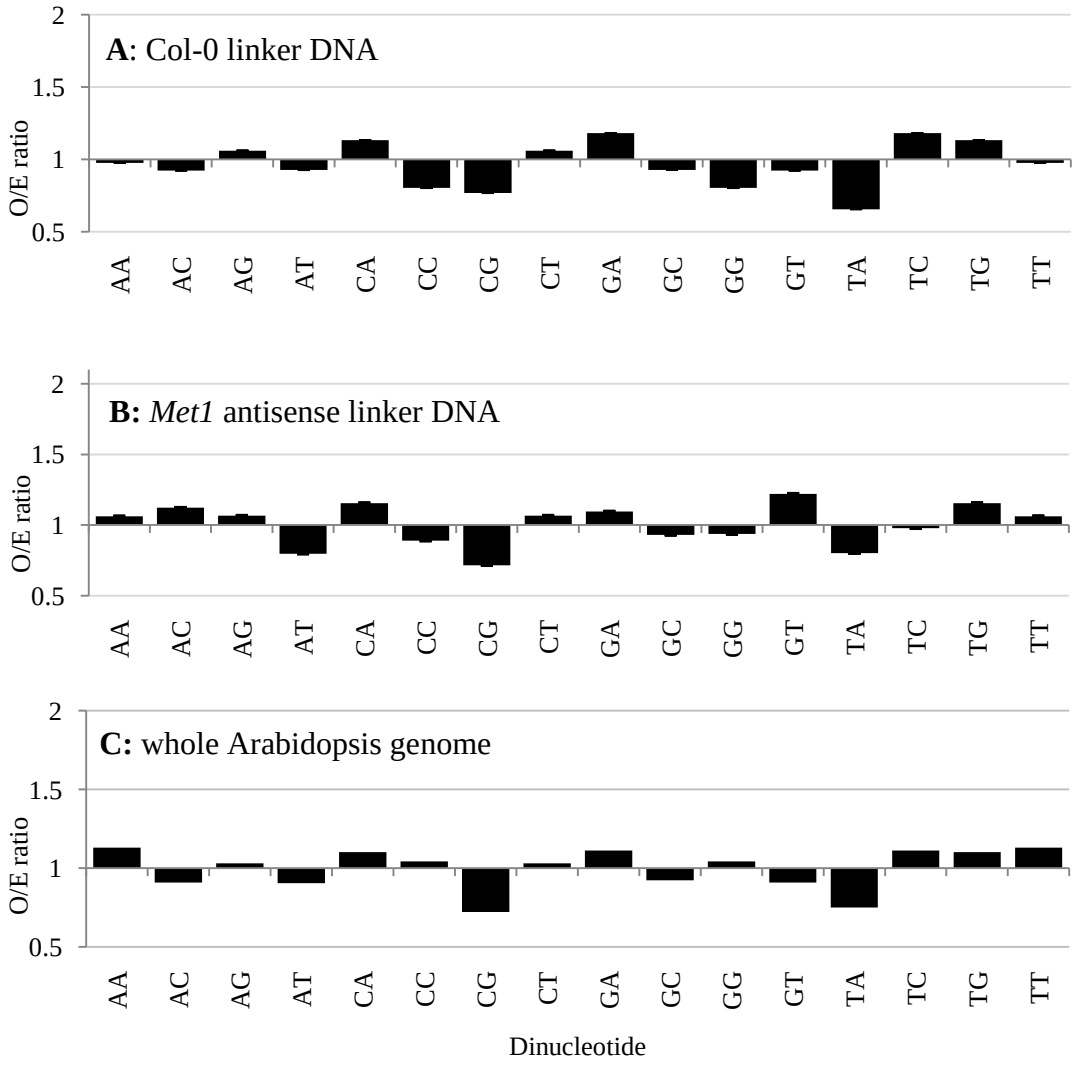

Figure 3.10 Mean observed/expected ratios of the occurrence of each dinucleotide for each DNA sequence in the linker Arabidopsis datasets A: *wt_di-linker* dataset, **B**: the *MET1_di-linker* dataset and **C** the whole genome. Expected values were calculated from the local DNA sequence composition. Error bars show standard error of the mean

3.3.2.iv Dinucleotide distribution within Arabidopsis nucleosome DNA fragments

In many studies, the [AA/TT] dinucleotides have been found to occur at a higher frequency at preferred sites within the 147 bp core nucleosome sequence, with a 10.3-10.4 bp periodic repeat, and reduced amplitude at the dyad axis (Ioshikhes *et al.*, 1996).

In order to test the hypothesis that dinucleotides occur at preferred sites within Arabidopsis nucleosome sequences, the position of each [AA] and [TT] dinucleotide along the nucleosome core DNA fragment was determined for the *wt_left* and *MET1_mono* datasets. The frequency at each position within the nucleosome core sequence was plotted and is shown in Figure 3.10.

A position for the dyad axis could be estimated by using the apparent symmetry of the [AA/TT] distribution. This was placed at the mid-point between the first two peaks which are located at the edge of the region where the signal appears to be reduced in amplitude (Figure 3.10). This region is thought to indicate reduced bending of DNA at the nucleosome dyad axis (Ioshikhes *et al.*, 1996). In these datasets, this position was detected ~83 bp in from the 5' edge of the nucleosome fragment, suggesting that during the digestion of chromatin, the MNase had not digested up to the nucleosome edge, since the mid-point of the nucleosome is thought to be around 73 bp (Richmond and Davey, 2003). This could either be due to the base-preference this enzyme exhibits, or because, for the present datasets, the enzyme had reduced access to the DNA, possibly because of the presence of histone H1 (An *et al.*, 1998).

.

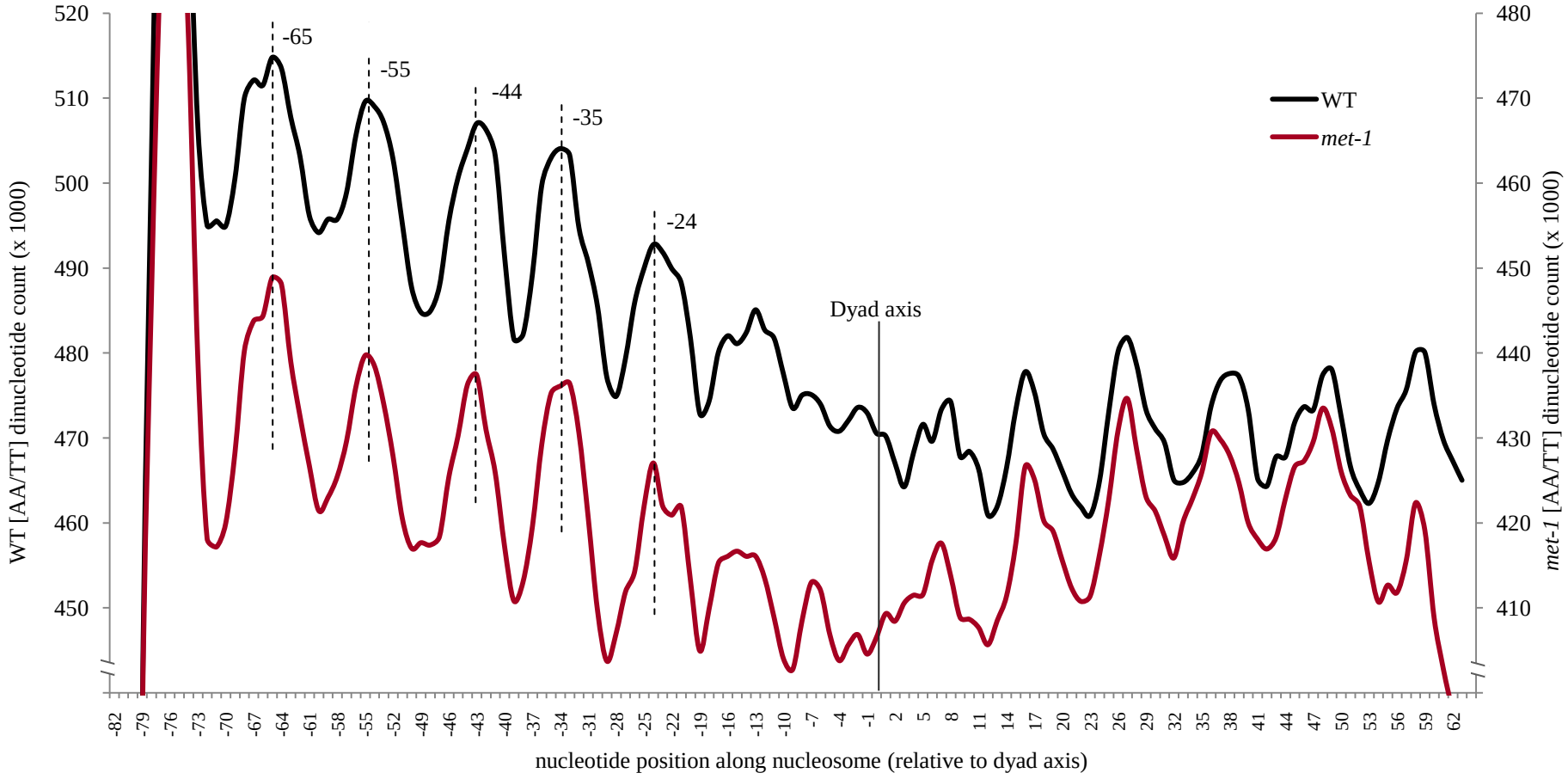

Figure 3.11 Distribution of [AA/TT] dinucleotides within nucleosome DNA fragments. Dataset *wt_left* is indicated with a black line (*y*-axis on left) and dataset *MET1_mono* is indicated with a red line (*y*-axis on right). The estimated dyad axis is indicated by a solid line, dashed lines indicate peaks in both datasets.

The [AA/TT] dinucleotides displayed a non random distribution along the nucleosome DNA fragment. Peaks were observed at positions -65, -55, -44, -35, -24, -13, -7, +7, +16, +27, +38, +48, +57 relative to the dyad axis, with peaks displaying greater amplitude towards the 5' end of the nucleosome DNA fragment.

The positions of all [CC] and [GG] dinucleotides were determined for the nucleosome DNA fragments in datasets *wt_left* and *MET1_mono*, as shown in Figure 3.11. The distribution for [CC/GG] did not show the same periodicity as the [AA/TT] distribution. However, peaks were observed in both the *wt_left* and *MET1_mono* datasets at -73, -39 and -5, indicating that this dinucleotide may have a different role in positioning the DNA around the histone octomer. In particular, the peak at -73 bp suggests a bias for [CC/GG] at the nucleosome/linker boundary, which has not previously been reported.

In order to test that the dinucleotide distributions observed could be attributed to nucleosome sequences, the distributions of [AA/TT] and [CC/GG] dinucleotides were calculated for the *Ath_rand* dataset (Figure 3.12). While peaks and troughs were observed in the distributions of both [AA/TT] and [CC/GG] dinucleotides, they appeared to be random. The distributions of dinucleotides within the random DNA sequences are dissimilar to the distributions of [AA/TT] dinucleotides in the nucleosome DNA sequences. Also, the distribution of [AA/TT] dinucleotides for the random DNA sequences do not display the apparent periodic distribution with an increase amplitude toward the 5' end observed in the distribution within the nucleosome DNA sequences. The peaks at specific nucleotide positions observed in the distribution of [CC/GG] dinucleotides for nucleosome sequences are also absent from the distribution for random DNA sequences.

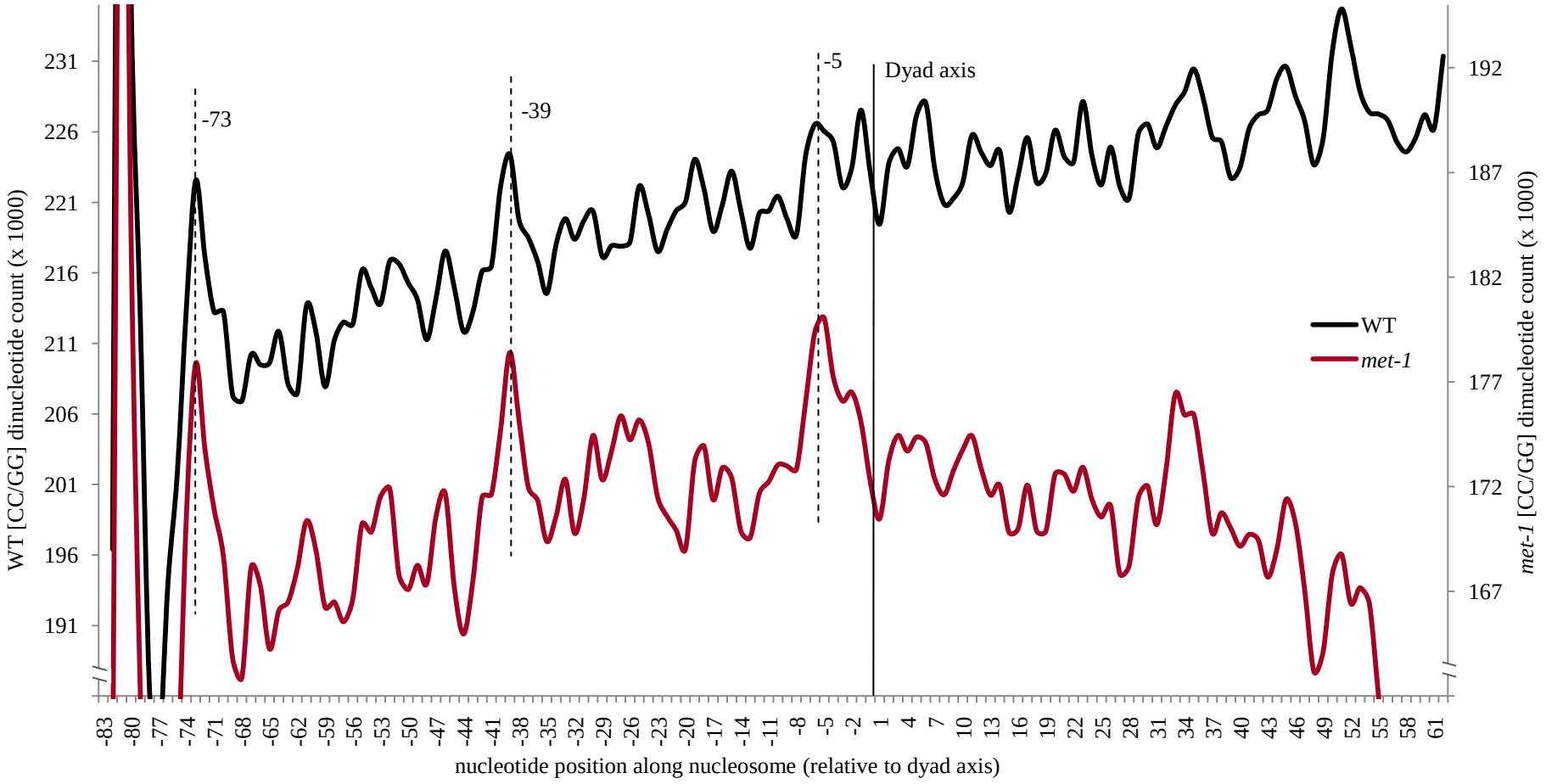

Figure 3.12 Distribution of [CC/GG] dinucleotides along the nucleosome DNA fragments. Dataset *wt_left* is indicated with a black line (*y*-axis on left) and dataset *MET1_mono* is indicated with a red line (y-axis on right). The estimated dyad axis is indicated with a solid line, dashed lines indicate peaks in both datasets.

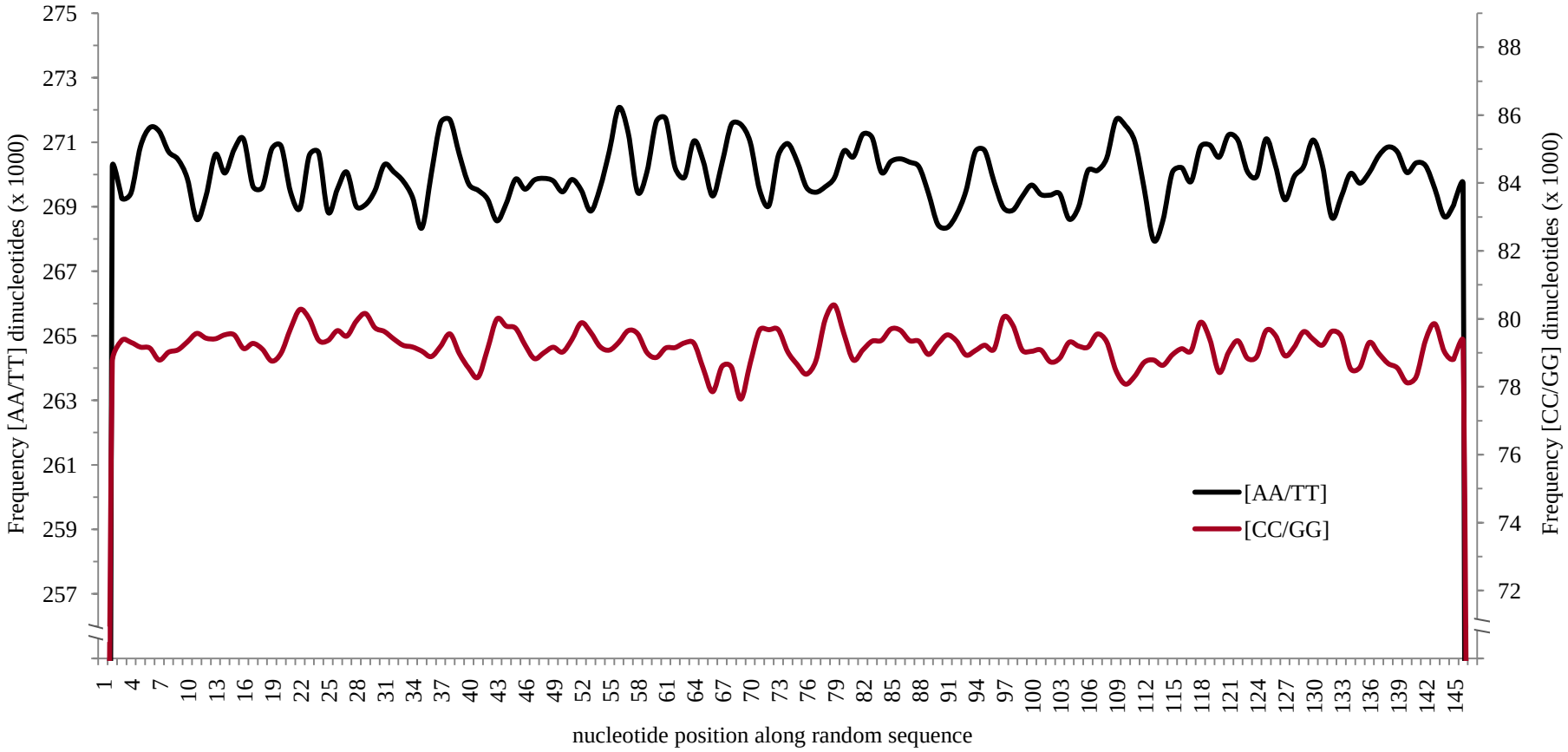

Figure 3.13 Distribution of [AA/TT] dinucleotides (black line) and [CC/GG] dinucleotides (red line) within the *Ath_rand* dataset.

To test the hypothesis that nucleosomes form on DNA sequences which have patterns of dinucleotide distribution distinct from linker DNA sequences, the distributions of [AA/TT] and [CC/GG] were calculated for *wt_left_linker* and *MET1_left_linker*. The distributions, shown in Figure 3.13, indicate that differences are observed in the distributions of [AA/TT] dinucleotides between the nucleosome and linker DNA sequences. The apparently smooth, regular peaks of the nucleosome distributions are not observed in the linker distributions. This would be expected if the role of the linker DNA in chromatin is different from that of nucleosome DNA. In addition, some differences are observed in the distributions of [AA/TT] dinucleotides between wild-type and anti-sense *MET1* linker DNA. In particular, the distributions within the wild-type and anti-sense *MET1* from 0 to ~ -35 bp appear to be a mirror-image of each other (horizontally inverted at the *x*-axis). To test whether the apparent mirror-image in the distributions of [AA/TT] dinucleotides within the *wt_left_linker* and *MET1_left_linker* datasets was statistically significant, a test for autocorrelation was used to identify important peaks and troughs in the data. However, this failed to show any significant ($p < 0.05$) peaks in either dataset. For example, the peak at -10 bp in the wild-type corresponds to a trough at this position in the anti-sense *MET1* distribution. This could be due to an overall shift of nucleosome position in either dataset.

The distributions of the [CC/GG] dinucleotides appear to be less organised than the [AA/TT] distribution in the *wt_left_linker* dataset. However, the distribution for *MET1_left_linker* appears to have a cyclical pattern with peaks at -2, -8, -15, -22, -30, -35, -42 and -50 bp.

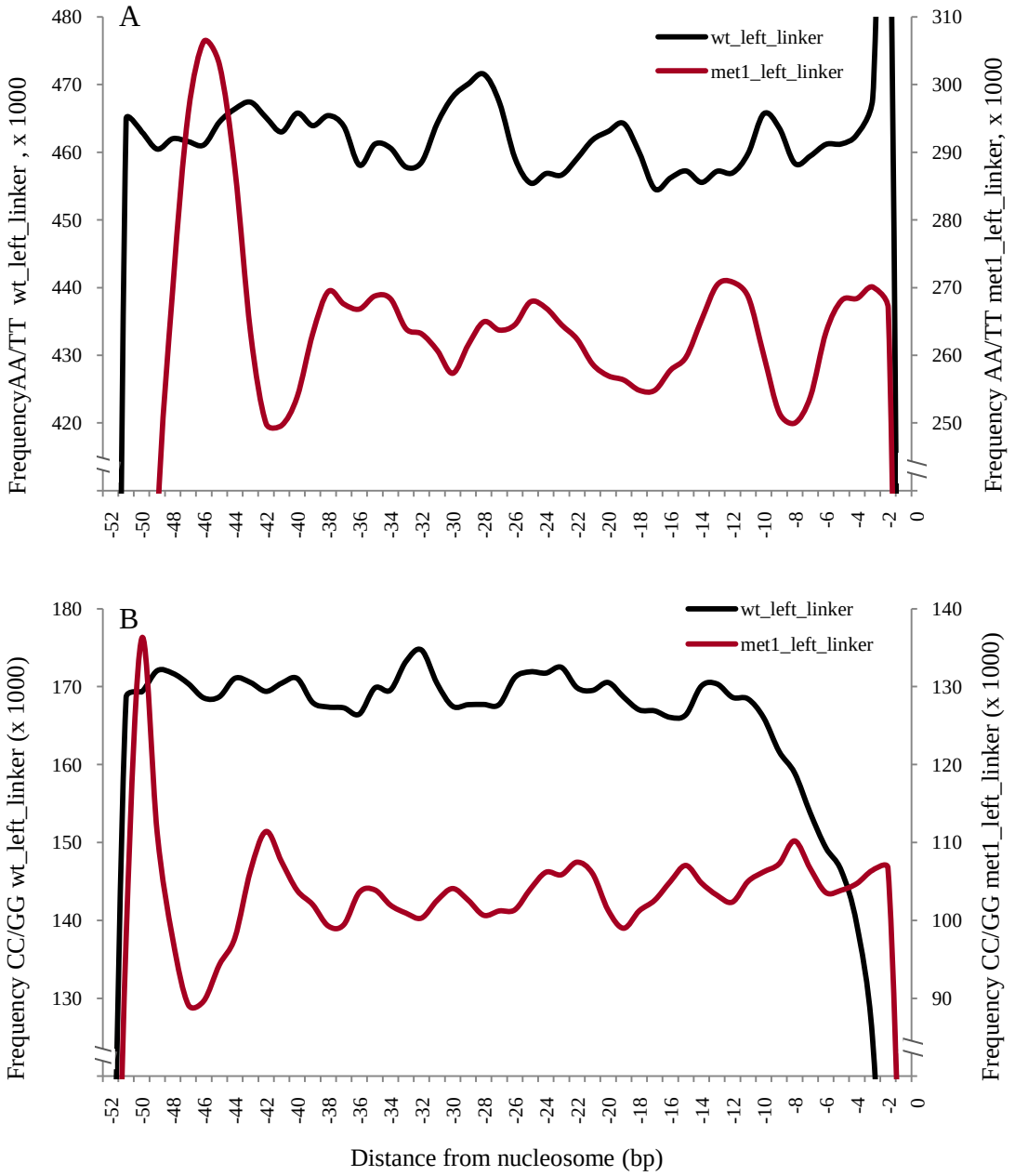

Figure 3.14 Distribution of dinucleotides along the linker DNA fragments of *wt_left_linker* and *MET1_left_linker* of **A:** [AA/TT] dinucleotides and **B:** [CC/GG] dinucleotides. The 5' edge of the nucleosome starts at position 0 on the right-hand side of each linker DNA fragment.

3.3.2.v Periodicity of dinucleotides in Arabidopsis nucleosome DNA fragments

In order to test for periodicity in the distributions of dinucleotides along the nucleosome, Fourier analysis was performed on the frequencies of each complimentary pair of dinucleotides for the datasets *wt_left*, *MET1_mono* and *Ath_02_wt*. Prior to assessing periodicity, the distribution was twice-differenced to remove the overall trend in the data. The twice differenced [AA/TT] dinucleotide of the *wt_left* dataset is shown in Figure 3.14.

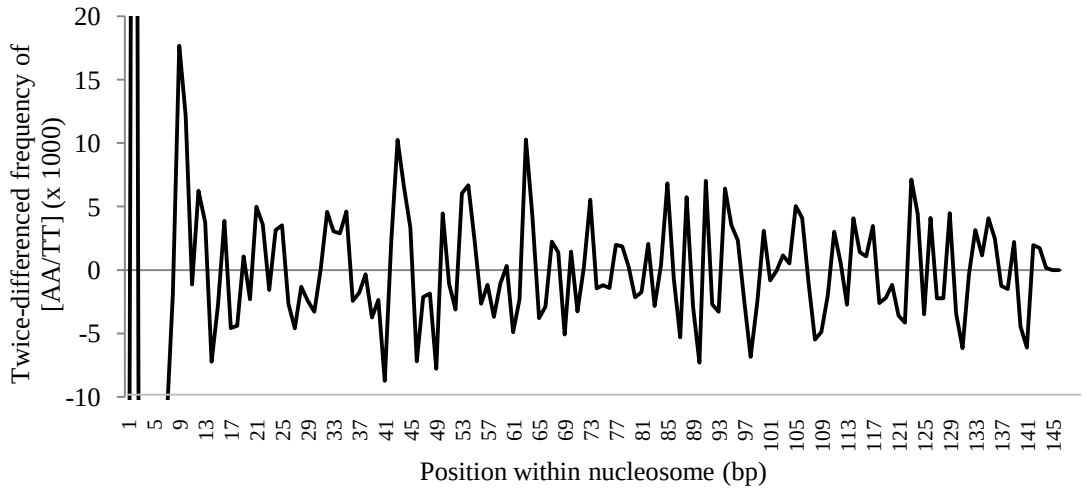

Figure 3.15 The twice-difference distribution of the [AA/TT] dinucleotide frequencies along nucleosome DNA fragments for the dataset *wt_left*.

To determine important periodicities in the distributions, Fourier-analysis was applied to the twice-differenced data and a periodogram was constructed. The periodogram is shown in Figure 3.15. The periodogram was inspected for statistically significant peaks, which indicate harmonics with important contributions to the variance in the data. The first statistically significant ($p < 0.05$) peak at amplitude of 0.0966 corresponds to a periodicity of 10.35 bp in the distribution of [AA/TT] dinucleotides. Table 3.7 shows the periodicities calculated by Fourier analysis for datasets *wt_left*, *MET1_mono*, *Ath_02_wt* and the periodicity of [AA/TT] and [CC/GG] in the *Ath_rand* dataset.

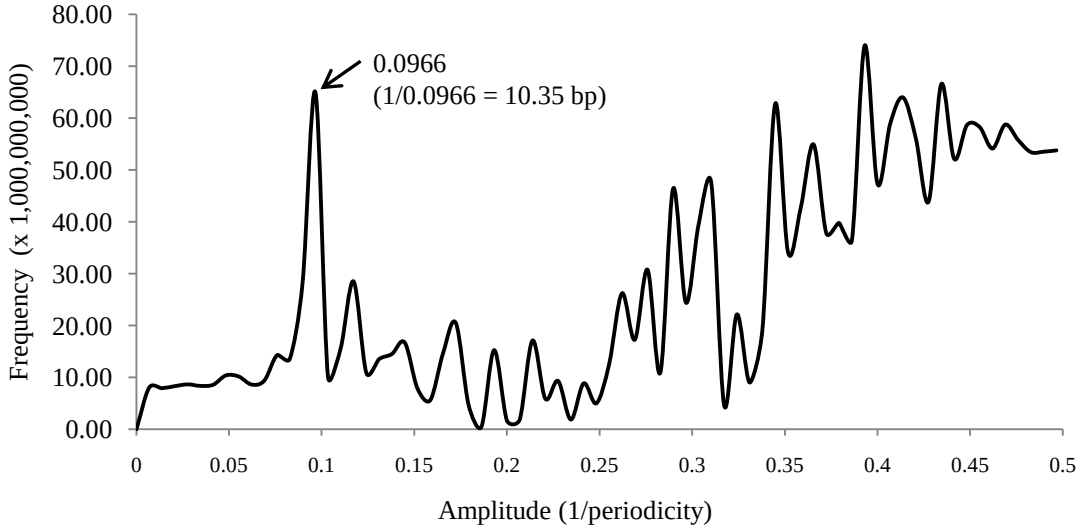

Figure 3.16 Periodogram constructed by Fourier analysis of the [AA/TT] dinucleotide distribution for the *wt_left* dataset. The arrow indicates the first important harmonic in the periodogram, indicating a periodicity of 10.35 bp for [AA/TT] dinucleotides.

Table 3.7 Statistically significant periodicities calculated by Fourier analysis for each complementary pair of dinucleotides for the nucleosome datasets *wt_left, MET1_mono, Ath_02_wt* and *Ath_rand*. The largest statistically significant ($p < 0.05$) periodicity is shown in each case.

Dinucleotide pair	Periodicity (bp)			
	wt_left	*MET1_mono*	*Ath_02_wt*	*Ath_rand*
AA/TT	10.35	20.71*	10.35	12.08
AC/GT	10.35	13.18*	20.71*	
AG/CT	6.31	10.35	10.35	
AT/TA	10.35	7.63	10.35	
CA/TG	6.19	13.18	10.35	
CG/GC	28.99*	28.99*	18.11	
CC/GG	13.18	13.18	8.53	24.15
GA/TC	13.18*	13.18*	6.19	

* These analyses also show evidence of periodicity of 10.35 bp

The dinucleotide [AA/TT] has a periodicity of 10.35 bp for both the *wt_left* and *Ath_02_wt* datasets, while the *MET1_mono* shows an important periodicity at 20.71 bp. However, for this dataset, there was also a significant harmonic relating to 10.35, suggesting that there may be preferred positions occurring more often in the *MET1_mono* dataset than in the others. The [CC/GG] dinucleotide periodicity was calculated to be 13.18 bp in the *wt_left* and *MET1_mono* datasets and 8.53 bp for the *Ath_02_wt* dataset.

There appears to be a difference in the periodicity of the [AC/GT] dinucleotide between anti-sense *MET1* and both (*wt_left* and *Ath_02_wt*) wild-type datasets. For the [AT/TA] dinucleotide, both wild-type datasets show a periodicity of 10.35 bp, while the anti-sense *MET1* has a periodicity of 7.63 bp. This suggests differences in the contribution of dinucleotides to the nucleosome positioning signal between anti-sense *MET1* and wild-type.

The [CA/TG] dinucleotide shows a different periodicity for each dataset. It is possible that this dinucleotide does not contribute to the overall nucleosome positioning signal. However, it could be the case (as with all dinucleotides) that this dinucleotide contributes to a specific set of nucleosome positions, for example, alternative positions around TSSs. Several differences are observed between *Ath_02_wt* and the other two sets. The random DNA sequences show periodicity in the distribution of [AA/TT] of 12.08 bp and in the distribution of [CC/GG] dinucleotides of 24.15 bp. The periodicities of the [AA/TT] and [CC/GG] distributions were also calculated for the *wt_left_linker* and *MET1_left_linker* as for the nucleosome datasets. These are shown in Table 3.8. The [CC/GG] dinucleotide occurred with a 10.20 bp periodicity in both the *wt_left_linker* and the *MET1_left_linker* datasets, although there was also evidence of a 7.28 bp periodicity for [CC/GG] in the anti-sense *MET1* linker.

Table 3.8 The periodicities of the linker datasets: *wt_left_linker* and *MET1_left_linker*

Dinucleotide pair	Periodicity (bp)	
	wt_left_linker	*MET1_left_linker*
AA/TT	4.64	7.28
CC/GG	10.20	10.20

Both the wild-type and anti-sense *MET1* datasets show differences between the nucleosome DNA and the linker DNA in terms of the periodicity of [AA/TT] dinucleotides.

3.3.3.i Nucleosome Sequence database

The construction of the Nucleosome Sequence database was initiated in 2006, when the amount of available nucleosome positioning data was small and not collated into a single resource. The Nucleosome Sequence database was populated with data from a number of sources with the aim of producing a single source for all the publicly available datasets at the time. The database contains nucleosome position information on species shown in Table 3.9.

Table 3.9 A list of the species represented in the Nucleosome Sequence database, collected from a variety of sources, and the number of entries for each species.

Species	Number of entries	Data source
Bovine spp	7	Ioshikhes *et al.* (1996)
Chlorocebus spp	1	Ioshikhes *et al.* (1996)
Cricetulus spp	1	Ioshikhes *et al.* (1996)
Crithidia fasciculata	1	Ioshikhes *et al.* (1996)
Drosophila melanogaster	32	Ioshikhes *et al.* (1996), NPRD: Levitsky *et al.* (2005)
Euplotes eurystomus	4	Ioshikhes *et al.* (1996)
*Gallus spp**	338	A. Travers. (Drew and Travers (1986), Satchwell and Travers (1989)), Ioshikhes *et al.* (1996)
Homo sapiens	18	Ioshikhes *et al.* (1996), NPRD: Levitsky *et al.* (2005)
Human immunodeficiency virus type 1	6	Ioshikhes *et al.* (1996)
Human papillomavirus type 16	2	NPRD: Levitsky *et al.* (2005)
Lytechinus variegatus	1	Ioshikhes *et al.* (1996)
Mouse mammary tumour virus	6	Ioshikhes *et al.* (1996)
Mus musculus	8	Ioshikhes *et al.* (1996), NPRD: Levitsky *et al.* (2005)
Oxytricha nova	3	Ioshikhes *et al.* (1996)
Psammechinus miliaris	1	Ioshikhes *et al.* (1996)
Rattus norvegicus	12	Ioshikhes *et al.* (1996), NPRD: Levitsky *et al.* (2005)
Saccharomyces cerevisiae	275	Ioshikhes *et al.* (1996), NPRD: Levitsky *et al.* (2005)
Schizosaccharomyces pombe	12	Ioshikhes *et al.* (1996)
Simian virus 40	53	Ioshikhes *et al.* (1996), NPRD: Levitsky *et al.* (2005)
Tetrahymena thermophila	11	Ioshikhes *et al.* (1996)
Xenopus laevis	5	Ioshikhes *et al.* (1996)
Zea mays	2	Ioshikhes *et al.* (1996)

Datasets generated since 2005 utilising new sequencing technologies and high-resolution tiling arrays has vastly increased the amount of nucleosome positioning data available. From this time, many datasets contribute to give insight into the distribution of nucleosome positions across whole genomes (Mavrich *et al.*, 2008a, Zhang *et al.*, 2008, Albert *et al.*, 2007, Johnson *et al.*, 2006, Lee *et al.*, 2007, Yuan *et al.*, 2005). At the time of writing, these sets had not been added to the original database, but those that were available are included here as a collection of nucleosome positions for comparisons and analysis. These are summarised in Table 3.10.

Table 3.10 A summary of some of the large nucleosome position datasets generated by high-throughput, next generation technologies, which were available, at the time of writing, for data comparisons and analysis.

Species	Nucleosome class	Technology	Number of entries	Reference
H. sapiens	dinucleosome	Traditional Sanger sequencing	1,002	Kato *et al.*, 2003
H. sapiens	dinucleosome	Traditional Sanger sequencing	100	Kato *et al.*, 2005
C. elegans	mononucleosome	454 (pyrosequencing)	187,863 (aligned)	Johnson *et al.*, 2006
D. melanogaster	mononucleosome (H2AZ-containing)	GS20FLX and tiling array	207,025 (aligned)	Mavrich *et al.*, 2008a
S. cerevisiae	mononucleosome	Traditional Sanger sequencing	199	Segal *et al.*, 2006
S. cerevisiae	mononucleosome (H2AZ-containing)	Pyrosequencing	322,000	Albert *et al.*, 2007
S. cerevisiae	mononucleosome	Tiling array	2,278	Yuan *et al.*, 2005

3.4 Discussion

3.4.1 Micrococcal nuclease sequence preferences

Sequence preferences at the site of MNase activity have been reported previously. A study to determine sequence preferences undertaken by Dingwall and colleagues (Dingwall *et al.,* 1981) showed a preference for [WG] in bovine DNA. In addition, a preference for [WG] has also been demonstrated at the site of MNase activity in yeast (Segal *et al.,* 2006). In contrast, human nucleosome DNA did not show a [WG] bias, although a preference for [GG] dinucleotides was observed. This led to speculation that the MNase preference might form part of a nucleosome positioning signal outside the nucleosome core boundary (Kogan *et al.,* 2006). However, in these studies, nucleotide preferences at the site of MNase activity were based on counts of occurrences, and did not take into consideration the frequency of dinucleotides in the genome from which the chromatin was prepared. Comparisons of observed/expected ratios calculated in this study give a more accurate picture of the effect of MNase preference on digested chromatin.

The Arabidopsis observed/expected ratios for MNase nucleotide preference show the bias towards [WG] similar to what may be expected given those identified by Dingwall *et al.,* (1981). Observed/expected ratios calculated for other taxa showed that different nucleotides occurred at frequencies greater than expected when compared to the respective genomic ratios. In the human datasets, the preferences for [CG] were 11-15 times higher than expected when compared to the genome occurrence of [CG]. A preference for [CG] was also observed in untrimmed chicken nucleosomes, eight times more often than expected when compared to the whole genome. The reported MNase preference for [WG] is observed in all datasets analysed with the exception of the trimmed chicken dataset. However, the observed occurrence of other dinucleotides at the sites of MNase action suggests that the MNase preference may be dependent on other factors such as the local chromatin structure. This could either be due to the presence of histone H1, or steric hindrance preventing MNase access to closely-spaced nucleosomes.

It is unlikely that any sequence-specificity which MNase possesses is adversely affecting the analysis of sequence composition in this study. The distribution of dinucleotides within the nucleosome sequence (section 3.3.2.iv) shows a possible 10 bp gap between the edge of the DNA fragment and the start of the nucleosome, according to the estimate of the dyad axis. The presence of histone H1 is thought to protect 5-15 bp linker DNA from MNase digestion (An *et al.*, 1998) and could be the cause of the gap, alternatively it could be incomplete digestion of the linker to the nucleosome edge. Arabidopsis chromatin was formaldehyde cross-linked prior to MNase digestion. This should ensure that nucleosomes do not alter their translational position during the preparation to provide fragments and so should preserve the *in vivo* nucleosome positions as well as the positioning signals associated with them.

Another apoptotic enzyme, DNA fragmentation factor 40/caspase-activated deoxyribonuclease (DFF40/CAD) has been developed for the purpose of studying chromatin structure and is reportedly superior to MNase in its specificity for linker DNA as a substrate (Widlak and Garrard, 2006). However, this enzyme is also reported to have a stronger sequence preference than MNase and shows a bias towards sequences which contain [RRRYRYYY], cleaving between the 4^{th} and 5^{th} position.

3.4.2 Nucleosome DNA fragment GC content

The sequence composition of nucleosome DNA fragments was compared to the genome average with respect to GC content. The nucleosome DNA fragments had a considerably higher GC content than the average, and when compared to subsets of the Arabidopsis genome, the mean GC of nucleosome sequences (41% GC) was between that associated with exons (42%) and the 5'UTRs (38%). This may suggest that the nucleosome datasets are enriched for exonic sequences. A previous computationally-based study places nucleosome positioning signals at intron/exon boundaries (Kogan and Trifonov, 2005). If this is the case, the GC content of the nucleosome sequences would suggest that the nucleosome occupies the 'exonic side' of the boundary, as the GC content for nucleosome sequences is well above that for introns (33% GC). It could be the case that exonic nucleosomes have more defined

positions compared to intronic sequences, which may be less structured. This may result in exonic sequences being represented more often in nucleosome datasets.

3.4.3 Nucleosome DNA fragment dinucleotide occurrence

The observed/expected ratios of dinucleotide occurrence within nucleosome DNA fragments showed that there was little deviation from the expected values when compared to the local nucleotide count. Occurrences of [AA], [TT] and [CA] were higher than expected. However, the occurrence of the [AA] and [TT] dinucleotides appear to be higher in the Arabidopsis genome when compared to the nucleosome sequences. These dinucleotides have all been implicated in positioning nucleosomes in previous studies (Widlund *et al.,* 1997, Ioshikhes *et al.,* 1996). The wild-type linker dataset (*wt_di-linker*) shows slightly more variation between the datasets when compared to the nucleosome dataset. It is likely that the positioning of specific dinucleotides is more important for nucleosome positioning than the overall abundance, and therefore, those dinucleotides are not likely to occur in nucleosome sequences more often than expected.

3.4.4 Distributions of dinucleotides within nucleosome DNA

The characteristic distribution of the dinucleotides [AA/TT] within the nucleosome DNA fragments observed in this study has been observed previously (Ioshikhes *et al.,* 1996). The in-phase 10 bp periodic repeat of [AA/TT] suggests a curved structure in the DNA (Satchwell *et al.,* 1986). The distribution of [AA/TT] dinucleotides within Arabidopsis nucleosome DNA fragments has a periodicity of 10.35 bp, so this is further evidence for sequence-dependent DNA bending around the nucleosome.

The dinucleotides [CC] and [GG] are thought to have a similar distribution along the nucleosome DNA as [AA] and [TT], but occur a half-turn (5 bp) out of phase with the position of [AA/TT] (Ioshikhes *et al.,* 1996). The [CC/GG] dinucleotide distribution did not show the expected counter-phase distribution to that of the [AA/TT] in Arabidopsis. The [CC/GG] distribution shows peaks at specific points

along the nucleosome in both the wild-type and anti-sense *MET1* datasets. This suggests a different role for [CC/GG] from that of [AA/TT] in nucleosome positioning. A periodicity of 13.18 bp in the distribution of [CC/GG] dinucleotides further supports this hypothesis.

Differences in the periodicity of [AC/GT] dinucleotides were observed between anti-sense *MET1* and both wild-type datasets. This may suggest differences in the contribution of dinucleotides to the nucleosome positioning signal between anti-sense *MET1* and wild-type. This is of interest since this dinucleotide has been shown to contribute to the nucleosome positioning signal (Segal *et al.,* 2006, Widlund *et al.,* 1997). Several differences were observed between *Ath_02_wt* and the *Ath_03_wt* datasets. This could be due to subtle differences in environmental growing conditions experienced by the plants on the different occasions the nucleosome fragments were isolated.

Of particular interest is the distribution of dinucleotides within the derived linker sequences. The wild-type and anti-sense *MET1* datasets show an almost mirror-image in [AA/TT] distribution, but appear to have different periodicities within the distribution. This adds further evidence to the hypothesis that nucleosome occupancy in anti-sense *MET1* is different from wild-type. Investigation of autocorrelation did not identify important peaks and troughs in the wild-type and anti-sense *MET1* dinucleotide distributions. However, inspection of the difference between peaks in the distribution of [AA/TT] when comparing the wild-type and anti-sense *MET1* linkers suggests that there may be a difference between the two datasets of approximately five to eight bp (or multiples of five to eight bp).

The periodicities in the distribution of [AA/TT] dinucleotides along the linkers of wild-type and anti-sense *MET1* are different, with the wild type linker distribution appearing to have many small peaks which are absent from the *MET1_left_linker* distribution. This may account for the difference in periodicities. The periodic distribution of [AA/TT] dinucleotides within nucleosome DNA is thought to extend to approximately 200 bp (~50 bp beyond the edge of the nucleosome) (Ioshikhes *et al.,* 1996) and represents alternative nucleosome positions. This was not observed in Arabidopsis. However, the distributions of [CC/GG] dinucleotides in the left linkers

of the wild-type and anti-sense *MET1* datasets did show a periodicity of 10.20 bp. This periodicity was not observed in the nucleosome DNA sequences, suggesting that [CC/GG] dinucleotides have different roles for nucleosome positioning in nucleosome and linker DNA, for example, [CC/GG] dinucleotides may exclude nucleosomes in some contexts. This phenomenon has not been previously reported. In general, there appears to be little information regarding the sequence characteristics of linker DNA compared to nucleosome DNA in the literature.

The differences observed for dinucleotide occurrence and distribution between wild-type and anti-sense *MET1* datasets may reflect differences in chromatin structure brought about by DNA methylation (or the absence of DNA methylation). Previous studies have shown that DNA methylation can alter the position of a nucleosome *in vitro*. Reconstitution of nucleosomes onto methylated and unmethylated versions of the chicken β-globin gene promoter demonstrated that the presence of DNA methylation altered nucleosome occupancy at this site (Davey *et al.*, 1997). When DNA methylation occurred near to the dyad axis of a nucleosome (1.5 helical turns from the dyad axis), nucleosome occupancy was dramatically reduced. DNA methylation at other positions had little or no effect. Further studies demonstrated that DNA methylation of a prokaryotic DNA sequence (*E. coli tyr*T promoter) altered the preferred rotational positioning of nucleosomes (Buttinelli *et al.*, 1998). Addition of 5-methylcytosine groups to the major grooves shifted the nucleosome position three bases towards the 3' end of DNA fragments. This suggests that presence of 5-methylcytosine groups on nucleosome DNA may affect the bending of DNA around the histone core by altering the base-stacking parameters of the dinucleotides.

3.4.5 Summary

The characteristics of nucleosome sequences in Arabidopsis show some similarities to those isolated in some other genera, such as higher GC content compared to the genomic average (Ozsolak *et al.*, 2007), the occurrence of [A] and [T] at the site of MNase activity, and the distribution of [AA/TT] dinucleotides with a periodicity of 10.35 bp (Segal *et al.*, 2006; Ioshikhes *et al.*, 1996; Satchwell *et al.*, 1986). In contrast, Arabidopsis nucleosome sequences also exhibit distinct patterns of

[CC/GG] dinucleotide distribution within nucleosome sequences and differences in the distribution of dinucleotides between the nucleosome and linker DNA. Particularly interesting are the differences revealed in sequence preferences for nucleosome and linker DNA between wild-type and anti-sense *MET1*. These differences suggest a role for DNA methylation in nucleosome positioning (see Chapter 5).

Chapter 4

Nucleosome spacing and linker length variation

4.1 Introduction

4.1 Introduction to linker DNA and linker length variation

Linker DNA is defined as the DNA connecting two nucleosomes. Linker histones (H1/H5) anchor the nucleosome in place and contribute to the compaction of the chromatin fibre by interacting with DNA, both in the nucleosome core and in the linker region (Fan *et al.,* 2003).

Early studies of the structure of chromatin involved the visualisation of chromatin by electron microscopy (Olins and Olins, 1974, Finch *et al.,* 1975, Oudet *et al.,* 1975). These studies revealed chromatin to be a heterogeneous molecule composed of repeating units of 'spherical bodies' (or nucleosomes) connected by 'DNA filaments' (linker DNA). These experiments provided the first evidence that chromatin structure is conserved between species, but that the linker length and level of chromatin compaction is variable.

Differences in DNA linker length were also noticed by Lohr and colleagues (Lohr *et al.,* 1977). Staphylococcal nuclease digestion of chicken erythrocyte, yeast and HeLa chromatin revealed differences in the nucleosome repeat length (the distance between the centres of two adjacent nucleosomes) but not in the DNA wrapped around the core particle. This suggested that the variability was due to the differences in linker length between these species. The differences in linker lengths between the three species included in the study were attributed to the transcriptional activity of the cell-type from which the chromatin originated. The linkers were found to be more variable in the HeLa and yeast chromatin, which is transcriptionally active, whereas the relatively inactive chicken erythrocyte chromatin was found to have regularly ordered nucleosome spacing. Further study of chicken erythrocyte, yeast and HeLa chromatin led to the observation of a linker length phasing in chromatin of 10_n (or multiples of the helical repeat of DNA). In yeast chromatin the linker length phasing was 10_n + 5 bp which suggested that nucleosomes in yeast chromatin are positioned in an anti-parallel manner with respect to the adjacent nucleosome (Lohr *et al.,* 1979).

Further evidence for the relationship between transcriptional activity and linker length was provided by experiments determining the linker length of the developing chicken from embryo to adult (Weintraub, 1978). The nucleosome repeat length of a 4-day old chick embryo was shown to be 190 bp (equivalent to a linker length of 43 bp). This increased to a nucleosome repeat length of 212 bp (equivalent to a linker length of 65 bp) in the adult chicken. The increase in nucleosome repeat length correlated with an increase in the number of molecules of erythrocyte-specific histone H5 from 0.2 molecules H5/nucleosome in the embryo to 1 molecule H5/nucleosome in the adult.

The nucleosome repeat lengths from many studies were collated and published in a taxa-organised list containing 185 entries (van-Holde, 1989). These data were used to determine if nucleosome repeat lengths do indeed show the multiples of preferential lengths that had been suggested previously (Widom, 1992). Firstly, probability distribution functions were constructed for average nucleosome repeat lengths. These were reported with error values, and the probability distributions were investigated using Fourier analysis. This revealed a periodicity of 9 to 10 bp within these data which is consistent with the periodicity of the helical repeat of DNA. This result disagreed with an earlier study where chromatin was reconstituted onto salmon sperm DNA and the psf 2124 plasmid without the presence of a linker histone. Micrococcal nuclease digestion and investigation by electron microscopy revealed that these nucleosomes, when closely spaced, appeared to have a preferred linker length of multiples of 7 bp (Noll *et al.*, 1980). However, since then, an approximate 10 bp periodicity has been demonstrated in the distribution of linker lengths in yeast chromatin (Wang *et al.*, 2008; Cohanim *et al.*, 2005) and in the predicted linkers of human chromatin (Kato *et al.*, 2003).

Linker length is thought to be a major determinant of higher-order chromatin structure, and specifically contributes to the diameter of the '30 nm' fibre (Robinson, *et al.*, 2007, Wong *et al.*, 2007). Measurements of chromatin fibres reconstituted from ordered nucleosome arrays revealed two classes. The first has a diameter ~35 nm for nucleosome repeat lengths of 177 to 207 bp (equivalent to linker length of 30 to 60 bp). The second class of fibre constructed from nucleosome arrays with repeat

lengths of 217 to 237 bp (equivalent to linker length of 70 to 90 bp) had a chromatin fibre diameter of ~45 nm (Robinson *et al.,* 2007). These measurements were used to model the trajectory of the linker DNA within the chromatin fibre (Wong *et al.,* 2007). These authors show that the most energetically favourable chromatin fibre conformation for each class of linker length is likely to result in a different chromatin fibre structure.

A relationship between intron length and nucleosome repeat length was suggested following examination of human and mouse exon and intron lengths (Beckmann and Trifonov, 1991). Periodicities between 200 and 210 bp were detected within the exon/intron and intron/exon length distributions for both human and mouse datasets. The periodicity detected was found to be sustained over long distances ~2000 bp. This led to the suggestion that introns may be contributing to the overall chromatin higher-order structure, by maintaining nucleosome positioning signals within genic DNA. The [AA/TT] nucleosome positioning signal was used in a computational study to investigate nucleosome positioning around exon/intron and intron/exon boundary sequences (Dennisov, *et al.*, 1997). This study used a set of ~4000 sequences mainly from the human and mouse genomes, and revealed that the splice junction is positioned within the mid 15 bases of the nucleosome (dyad axis). It was suggested that the placement of nucleosomes over the splice junction may provide protection of the splice junction against mutation. A computational study used experimentally determined and predicted splice-site sequence data to investigate dinucleotide phasing around splice sites for a number of species (Kogan and Trifonov, 2005). This study was able to position both the exon/intron and intron/exon boundaries to 1-2 bp from the nucleosome mid-point (dyad axis), providing further evidence of protective nucleosome positioning at gene splice-sites.

In Arabidopsis, the 18S, 5.8S and 25S rRNA genes are organised into two clusters (nucleolar organiser regions (NORs)) of 700 ± 130 tandemly repeated copies on chromosomes 2 and 4, with a single repeat present on chromosome 3 (European Union Chromosome 3 Arabidopsis Genome Sequencing Consortium, 2000). The rRNA genes are separated by two internal transcribed spacer (ITS) sequences and repeats are separated by an intergenic spacer (IGR) (Figure 4.1). The coding sequences are transcribed by RNA polymerase I (Pol I). The rRNA transcripts

constitute 50-80 % of transcribed RNA in eukaryotic cells, although transcription of only a subset of repeats is needed to fulfil this requirement (Pruess and Pikaard, 2007). The NORs can be visualised by silver staining of metaphase chromosomes, and are referred to as secondary constrictions. These are regions of less condensed chromatin, as previously transcriptionally active rRNA genes remain associated with RNA Pol I transcription factors in order to allow rapid transcription of rRNA in the following cell cycle (McStay, 2006). Silenced rRNA regions are condensed at metaphase and do not form secondary constrictions. Silver staining and FISH of rye NORs on metaphase chromosomes revealed domains of differing chromatin structure within a NOR, with condensed rDNA adjacent to a decondensed secondary constriction, suggesting different regulation of rRNA genes within a NOR (Caperta *et al.,* 2002). Psoralen cross-linking and restriction digestion of murine Friend cell chromatin has demonstrated that two different chromatin states exist in rDNA regions: psoralen accessible (nucleosome depleted) and psoralen inaccessible (organised nucleosomes) (Conconi *et al.,* 1989).

The phenomenon of nucleolar dominance has been used as a model to study the mechanisms of rRNA gene silencing in Arabidopsis. In *A. suecica*, a naturally occurring allotetraploid of *A. thaliana* and *A. arenosa*, the NORs originating from the *A. thaliana* parent are silenced, while the *A. arenosa* rRNA genes are transcribed (or dominant) (Chen *et al.,* 1998). Treatment of this Arabidopsis hybrid with 5'aza-2-deoxycytosine (a methyltransferase inhibitor) or trichostatin A (a histone deacytalase inhibitor) de-represses the *A. thaliana* rRNA genes. Some *A. arenosa* rRNA genes are also upregulated following 5'aza-2-deoxycytosine treatment, suggesting that a subset of these genes is also silenced in the Arabidopsis hybrid, and it is likely that the same mechanism controls gene silencing and rRNA gene dosage. Further work revealed that the promoters of the *A. thaliana* rRNA genes in *A. suecica* are heavily methylated and the genes are associated with H3K9me2, a known heterochromatic mark. The transcribed *A. arenosa* rRNA gene promoters are hypomethylated and associated with the euchromatic mark H3K4me3, whilst the silenced subset of *A. Arenosa* rRNA genes are associated with H3K9me2 and show hypermethylation at the promoter (Lawrence *et al.,* 2004).

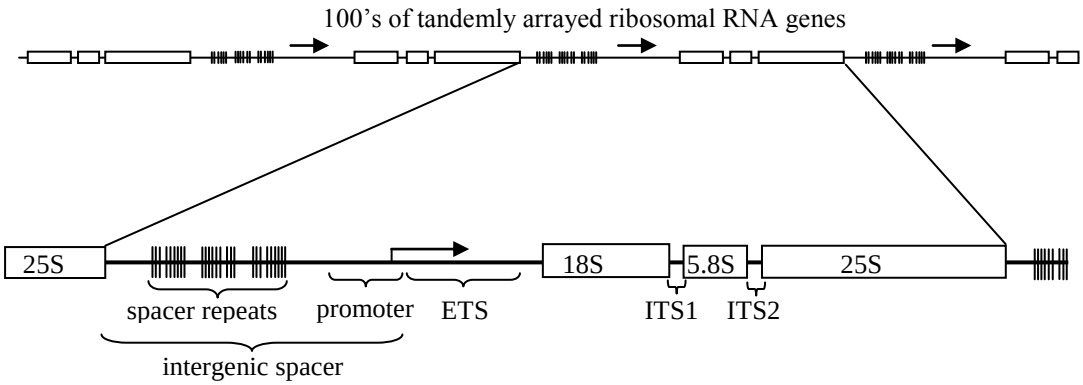

Figure 4.1 Organisation of the 18S, 5.8S and 25S rRNA genes, the internal transcribed spacers and intergenic spacer. Image reproduced from Pruess and Pikaard (2007).

Considering the important contribution of linker length variation to chromatin higher-order structure, the following hypotheses were constructed and tested:

1. The distribution of linker length in Arabidopsis is comparable to that determined for other taxa.

2. Periodicity exists within the Arabidopsis linker length distribution similar to that identified in other taxa.

3. Features of the observed distribution of Arabidopsis linker length show similarities to the distribution of Arabidopsis intron length, which may indicate a relationship between introns in the DNA and chromatin structure.

4. A distinct linker length distribution exists within the nucleolar organiser regions of the Arabidopsis genome, which reflects region-specific higher-order chromatin structure.

4.2. Methods

4.2.1 Calculation of linker length

Linker DNA lengths were calculated by subtraction of 294 bp (2 x nucleosome length (147 bp)) from the dinucleosome lengths of fragments for each of the data sets *Ath_01_wt*, *Ath_03_wt* and *Ath_04_MET1*.

Linker lengths were estimated from mononucleosome positions for each of datasets *Ath_02_wt* and *Ath_04_MET1* by calculating the distance between the end of a nucleosome position and the start of the adjacent nucleosome position from the start position of nucleosome, using the formula [n+1$_{start}$ - n_{end}] as shown in Figure 4.2A.

Where n_{end} - n+1$_{start}$ < 0, the lengths were omitted as they represent overlapping nucleosome positions (nucleosome in alternative positions in different cells) rather than linker lengths (nucleosome spacing) as shown in Figure 4.2B.

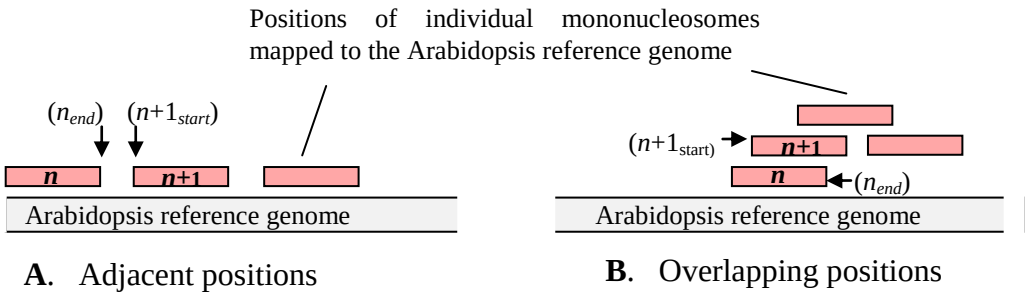

Figure 4.2 Schematic showing the method for estimation of linker length from mononucleosome DNA datasets, A: shows positions of adjacent nucleosome and B: shows overlapping nucleosomes, which are likely to represent alternative positions for the same nucleosomes from different cells.

4.2.2 Calculation of length periodicities

Fourier analyses and partial autocorrelations were performed on frequencies calculated in section 4.2.1 in order to determine any periodicities in linker length distribution. The procedures are described in Chapter 2, Section 2.2.iv.

4.2.3 Determination of Arabidopsis exon and intron length and periodicity

Intron sequences and annotations were downloaded from the TAIR (v8) blast dataset ftp site (url: ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets). Intron lengths were measured directly from the sequence lengths. Exon lengths were inferred by subtracting the start position of intron 2 from the end position of intron 1, for introns *i*, (i+1$_{start}$ - i_{end}) as shown in Figure 4.3. The distances from the start of an intron to the start of the next intron (i+1$_{start}$ - i_{start}), and from the start on an intron to the end of the next intron were also calculated.

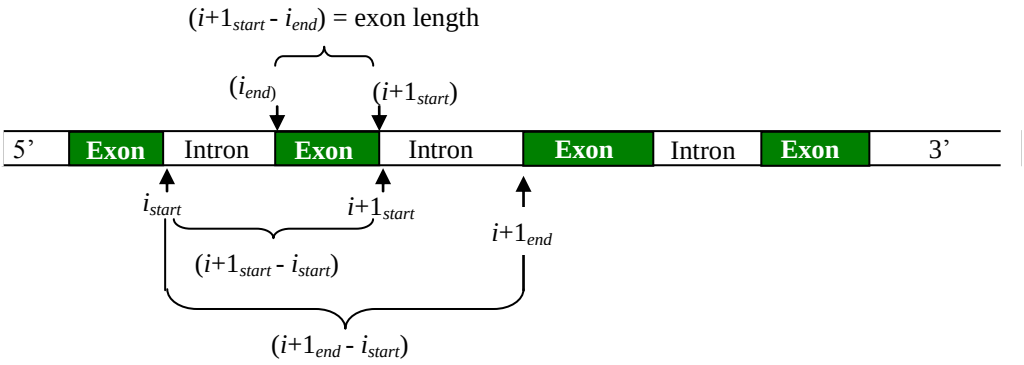

Figure 4.3 Schematic showing the method for calculating exon and intron lengths from intron position data and for calculating intron/exon combination datasets.

Any periodicities in the distributions of exon and intron lengths were determined by Fourier analysis as previously described.

4.2.4.i Determination of linker length within an Arabidopsis nucleolar organiser region

Dinucleosome DNA fragments from dataset *Ath_03_wt* which mapped within 1000 bp* in the 5' region of a miscellaneous RNA gene and within 1000 bp in the 3' region of the 18S rRNA genes, repeated on chromosomes 2 and 3, were extracted. The data sub-sets and regions represented from each chromosome are shown in Table 4.1. The two regions (one from each chromosome) contain the 5S and 18S rRNA genes, of which 730 ±100 copies are thought to be present on Chr 2, and correspond to the nucleolar organiser region (NOR) on that chromosome. The repeats are not accounted for in the annotated reference genome.

*Only 871 bp upstream of the AT2G01008 gene as this is the edge of the sequence data available (TAIR v8).

The region on Chr 3 is thought to contain a single copy of the repeat. Only fragments which mapped to unique sequences on either Chr 2 or Chr 3 were used for the analyses.

Table 4.1 The two subsets extracted from the *Ath_03_wt* dataset which represent the annotated rRNA regions on Chromosomes 2 and 3. The gene models, descriptions and positions relative to the annotated reference genome are shown.

Dataset	Dataset region (bp)	Gene model (TAIR)	Gene description	Chr	Gene start position (bp)	Gene stop position (bp)
Ath_03_rna2	1000 to 6,782	AT2G01008	other RNA	2	1,871	2,111
		AT2G01010	18S rRNA	2	3,706	5,513
		AT2G01020	5S rRNA	2	5,782	5,945
Ath_03_rna3	14,205,903 to 14,211,902	AT3G41761	other RNA	3	14,206,903	14,207,064
		AT3G41768	18S rRNA	3	14,208,663	14,210,470
		AT3G41979	5S rRNA	3	14,210,739	14,210,902

Linker lengths were calculated for each dataset as described in section 4.2.1, and plotted together with a smoothed curve calculated from an 8 bp sliding window. The latter removed small-scale periodicity and revealed overall trends in the distribution.

4.2.4.ii Kernel Density Estimation

In order to inspect the linker length distributions of datasets *Ath_03_rna2* and *Ath_03_rna3* for multi-modal components within the distribution, Kernel Density Estimation (KDE) was used to smooth the linker length frequency distribution and reveal important longer-range peaks in the distributions. The density estimations were performed in GenStat® (2008, 11th Edition, VSN International, UK) with the bandwidth calculated using the method of Sheather & Jones (1991), and using a set of 1,024 grid points.

4.2.4.iii Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test is a non-parametric test used to determine whether two distributions are significantly different, comparing the cumulative distribution function for each. The cumulative distributions of the linker lengths were calculated in GenStat® (2008, 11th Edition, VSN International, UK) and the Kolmogorov-Smirnov *D*-statistic was calculated as the largest difference between the cumulative distributions at any value of *x*, where *x* represents increasing linker length.

The critical D-value was calculated as:

$$D\alpha = c(\alpha)\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

where $c(\alpha)$ = 1.36 when α = 0.05, and where n_1 and n_2 are the numbers of observations in the two sets of linker lengths (distributions) being compared.

The critical *D*-value is compared to the calculated Kolmogorov-Smirnov *D*-statistic. A calculated *D*-value larger than the critical value indicates a statistically significant difference between distributions at the α (or p) = 0.05 level of significance.

Results

4.3.1 Linker length distribution

To test the hypothesis that linker DNA in Arabidopsis is variable in length and has a non-random in distribution, the linker length distribution was investigated using the datasets *Ath_01_wt*, *Ath_02_wt*, *Ath_03_wt* and *Ath_04_MET1*.

4.3.1.i Calculation of linker length from dinucleosome DNA fragments

For dinucleosome sequences (DNA fragments with a length > 294 bp) from datasets *Ath_01_wt*, *Ath_03_wt* and *Ath_04_MET1*, linker lengths were calculated by subtracting twice the nucleosome length (147 bp, Richmond and Davey, 1993) from the dinucleosome DNA fragment length (see Figure 4.4). However, this does not take into consideration any minimum linker length, steric hindrance, nor the presence of histone H1. Fragment lengths less than 294 were omitted from the analysis as this is the minimum possible dinucleosome length (2x 147 bp). Using these constraints, it is possible that dinucleosome fragments length over 442 bp (linker length = 147 bp) may actually originate from a tri-nucleosome fragment with no linker DNA. For this reason, fragments over 442 bp in length were also omitted from the analysis. The distributions of linker length for datasets *Ath_01_wt*, *Ath_03_wt* and *Ath_04_MET1* are shown in Figure 4.5

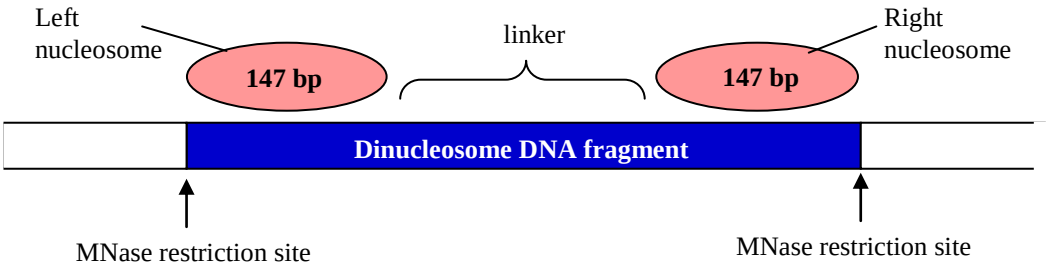

Figure 4.4 Schematic showing the calculation of linker length by subtracting 2x 147 bp (two nucleosomes) from the length of a dinucleosome DNA fragment.

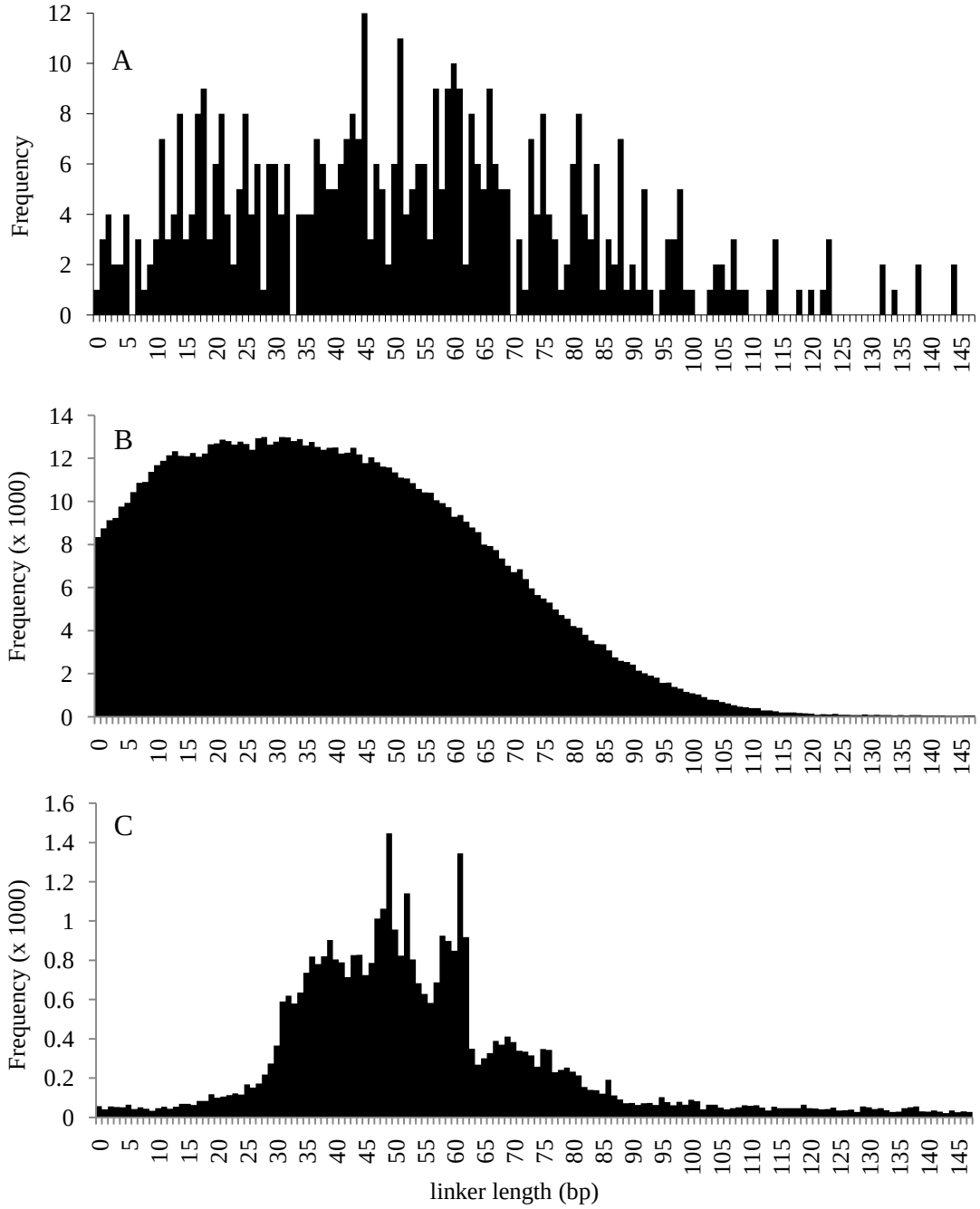

Figure 4.5 Distributions of calculated linker lengths from dinucleosome-length sequences from the datasets A: *Ath_01_wt* B: *Ath_03_wt* and *Ath_04_MET1*.

The *Ath_01_wt* (n=485) and *Ath_04_MET1* (*n* = 39,608) datasets show non-random distributions of linker length with multiple peaks within the distributions. The peaks are likely to represent the distributions around preferred linker lengths for Arabidopsis dinucleosomes. The *Ath_03_wt* dataset (*n* = 903,305) does not appear to show such peaks within the distribution, which could be due to the larger sample size

(or under-sampling of the other datasets). The distributions of *Ath_01_wt* and *Ath_04_MET1* suggest that there is a multi-modal distribution for linker length.

The mean linker length ($\bar{y}$) and standard deviation were calculated for each dataset and are shown in Table 4.2. This shows that there appears to be more similarity between the mean linker lengths of the *Ath _01_wt* and *Ath_04_MET1* datasets, with the medians being identical. The standard deviations are large for each data set, and indicate a large range of linker lengths in Arabidopsis.

Table 4.2 Linker lengths calculated from Arabidopsis dinucleosomes from the datasets *Ath_01_wt, Ath_03_wt* and *Ath_04_MET1*

Data set	*n*	*Mean* ($\bar{y}$) (bp)	Median (bp)	Standard deviation
Ath_01_wt	485	52.39	51	29.46
Ath_03_wt	903,305	40.01	38	24.89
Ath_04_MET1	39,608	54.67	51	23.10

4.3.1.ii Derived linker lengths from mononucleosome positions

In order to estimate the linker length variation in the mononucleosome datasets, the distances between the positions of mononucleosomes were calculated for the *Ath_02_wt* and *Ath_04_MET1* datasets. Linker lengths were calculated as the distance between individual mononucleosome DNA fragments mapped onto the Arabidopsis genome, as shown in the schematic (Figure 4.6). Distances less than 147 bp (or linker length < 0) were omitted from the analysis as these lengths represent alternative nucleosome positions. In addition, distances larger than 294 bp (linker lengths > 147) were omitted from the analysis as beyond this length, there is no certainty that the distance represents dinucleosome linkers (see section 4.3.1.i.).

The distributions of the derived linker lengths for datasets *Ath_02_wt* (*n* = 1,545) and *Ath_04_MET1* (*n* = 154,084) are shown in Figure 4.7.

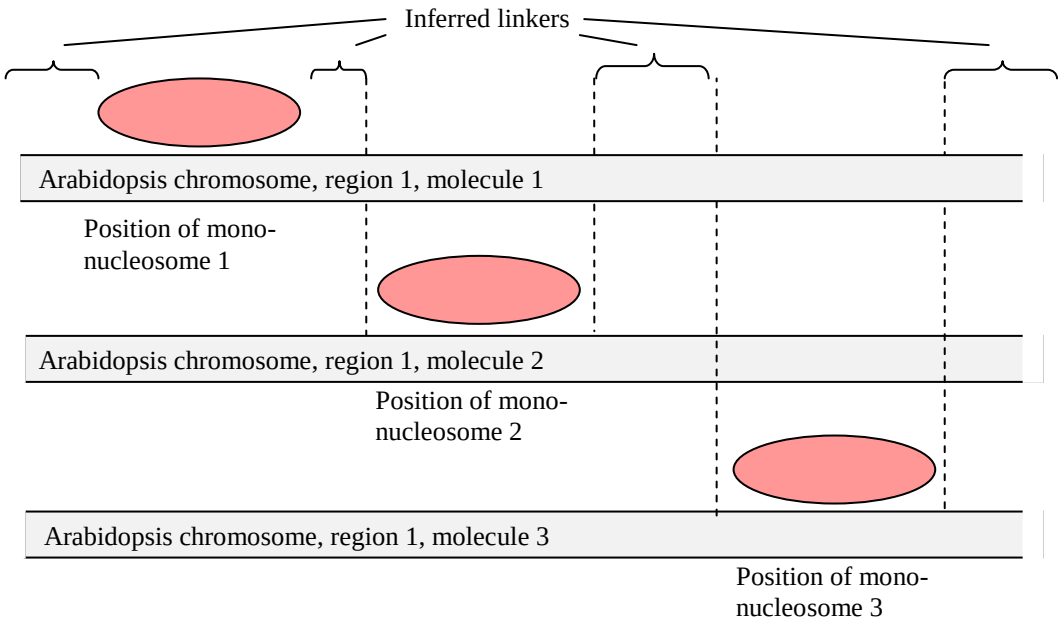

Figure 4.6 Schematic showing the method for inferring linker lengths from a set of mononucleosome DNA fragments. The diagram demonstrates different nucleosome positions in different molecules over the same region of an Arabidopsis chromosome.

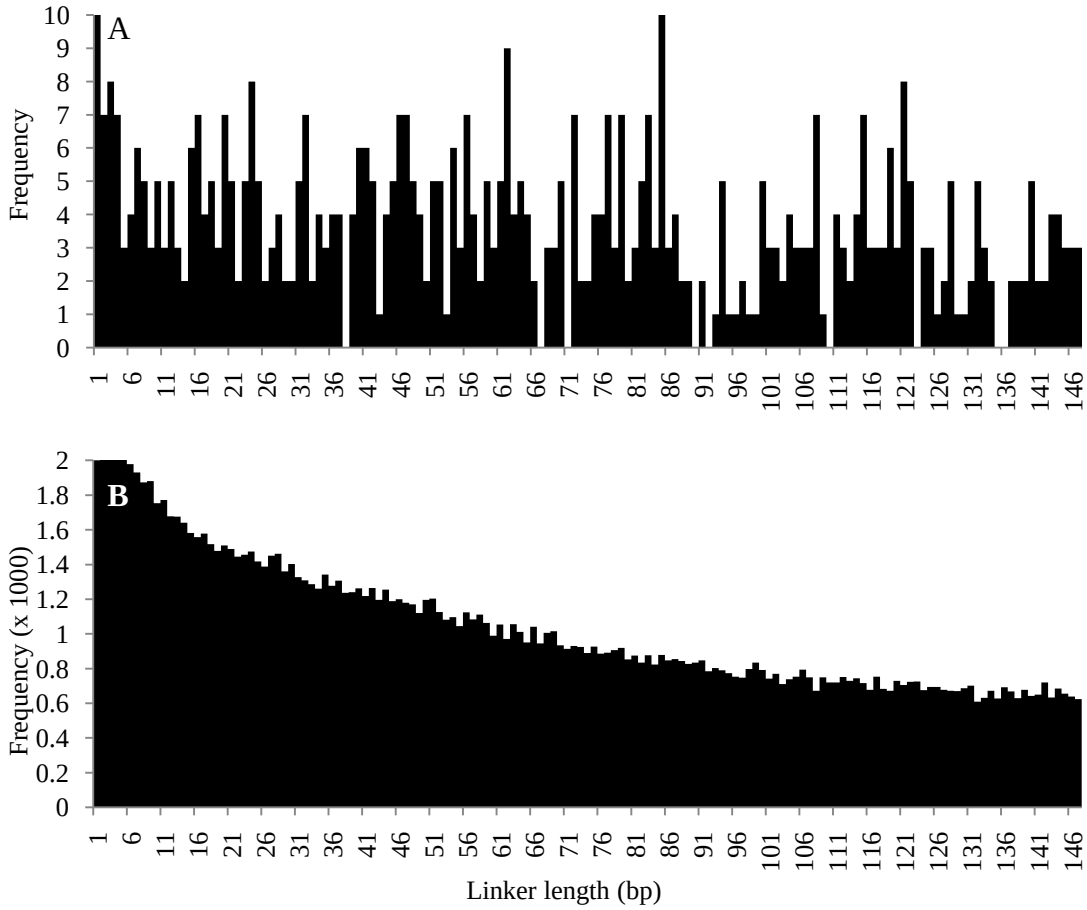

Figure 4.7 Distribution of the inferred linker lengths from datasets, A: *Ath_02_wt* and B: *Ath_04_MET1*.

The linker lengths derived from the *Ath_02_wt* data show a distribution similar to that of the *Ath_01_wt* linker lengths with peaks throughout the distribution, possibly representing preferred length for linker DNA in Arabidopsis. The distribution of the *Ath_04_MET1* derived linker lengths is more continuous, with less well defined maxima within the distribution. It may be that the smaller datasets are under-sampled, showing only a subset of the dinucleosome lengths present in the genome.

The mean linker length ($\bar{y}$) and standard deviation were calculated for each dataset and are shown in Table 4.3. The mean and median values for the derived linker lengths are both higher than those from the linker lengths calculated from dinucleosome DNA fragments.

Table 4.3 Mean, median and standard deviation of linker lengths derived from mononucleosome positions using the datasets *Ath_02_wt* and *Ath_04_MET1*.

Dataset	Number of values	Mean ($\bar{y}$) (bp)	Median (bp)	Standard deviation (bp)
Ath_02_wt	549	64.58	61	42.64
Ath_04_MET1	155,566	58.95	52	42.43

4.3.2 Periodicity in linker length distribution

Previous studies have presented evidence suggesting that a periodicity of approximately 10 bp exists in linker length distributions (Kato *et al.,* 2003, Widom *et al.,* 1992). The periodicity is thought to be a consequence of nucleosome positioning influenced by the DNA helical twist. In order to test the hypothesis that Arabidopsis chromatin exhibits preferred nucleosome linker lengths, periodicity was investigated using Fourier analysis and partial autocorrelation. Linker lengths calculated in the previous sections 4.3.1.i. and 4.3.1.ii. for datasets *Ath_01_wt*, *Ath_02_wt*, *Ath_03_wt* and *Ath_04_MET1* were used for the analyses.

The results of Fourier analyses of the distribution of linker lengths in Arabidopsis are shown in Table 4.4, and suggest a periodicity of 7.68 bp for the *Ath_01_wt*, 7.45 bp for *Ath_02_wt* and 7.34 bp for *Ath_03_wt* dinucleosome datasets. A periodicity between 9.2 bp and 9.8 bp was calculated for the linker length from the dinucleosome fragments of *Ath_04_MET1* and a periodicity of 12.25 bp was calculated for the distribution of distances between nucleosome positions from the *Ath_04_MET1* dataset. The partial autocorrelation function (PACF) is described as the correlation between observations n_l and $n_{l+k,}$ having accounted for all other correlations from n_l to n_{l+k-1}, and gives an indication of periodicity in the data. PACFs were calculated from the linker length distributions for Arabidopsis datasets *Ath_01_wt*, *Ath_02_wt*, *Ath_03_wt* and *Ath_04_MET1* and are shown in Table 4.4. In general, the PACF tends to support the periodicity calculated by Fourier analysis.

The calculated linker length periodicity for Arabidopsis is less than the ~10 bp previously reported for other taxa (Kato *et al.,* 2003, Widom *et al.,* 1992). In order to provide a direct comparison, the Fourier and partial autocorrelation analyses were performed using dinucleosome DNA fragments from the publicly available datasets: *Kato_03_hum, Kato_05_hum, and Trav_chick_dinuc*. In addition, distances between mononucleosome positions from the *Johns_Celegans* dataset were also used (Table 4.4). The periodicities in the linker length distributions derived from these datasets have not previously been calculated. Descriptions of these datasets can be found in the general materials and methods (Chapter 2, section 2.2.vii). The periodicities calculated by Fourier analysis and PACF are shown in Table 4.4. A periodicity of 8

bp was detected within the distribution of the *Kato_03_hum* dataset while a periodicity of 6.34 bp was calculated for the *Kato_05_hum* dataset. The discrepancy between these two datasets could be due to the small sampling size of the *Kato_05_hum* set (n = 100) compared to the *Kato_03_hum* dataset (n = 1002). The partial autocorrelation suggests a periodicity of 8-9 bp for the linker length distribution in both datasets, which is the closest to the ~7 bp detected here for Arabidopsis. The *Trav_chick_dinuc* dataset was the only one which exhibited the reported 10 bp periodicity within the linker length distribution, which is also seen in the PACF. The *Johns_Celegans* dataset exhibits a periodicity of ~ 5 bp within the linker length distribution, although there was also some evidence for a periodicity of 9.75 bp for this dataset. A periodicity of 4 bp was detected in the *Johns_Celegans* linker length distribution using PACF, which is closer to the ~5 bp Fourier analysis result, suggesting a periodicity of 4-5 bp for this dataset. Overall, these results suggest that linker length preferences may be taxa-specific.

Table 4.4 Statistically significant (p < 0.05) periodicities calculated by Fourier analyses and PACF for linker length distributions derived from Arabidopsis datasets *Ath_01_wt, Ath_02_wt, Ath_03_wt* and *Ath_04_MET1*. In addition, the periodicities of *Kato_03_hum, Kato_05_hum, Trav_chick_dinuc* and *Johns_Celegans* are shown.

Dataset	Linker length derived from:	Periodicity Fourier (bp)	Confidence limit (Fourier) (bp)	Periodicity (PACF) (bp)
Ath_01_wt	dinucleosome fragment	7.68	7.30 - 8.30	7
Ath_03_wt	dinucleosome fragment	7.34	6.68 - 7.73	6
Ath_04_MET1	dinucleosome fragment	9.2-9.8	8.7 - 10.50	10
Ath_02_wt	mononucleosome position	7.45	7.10 - 7.84	8
Ath_04_MET1	mononucleosome position	12.25	11.31 - 13.37	14
Kato_03_hum	dinucleosome fragment	*8.00	7.62 - 8.89	8-9
Kato_05_hum	dinucleosome fragment	6.34	6.00 - 6.53	8
Trav_chick_dinuc	dinucleosome fragment	10.66	10.00 - 11.43	9-10
Johns_Celegans	mononucleosome position	**5.20	5.12 - 5.29	4

* There was also evidence of periodicity of ~6 bp
** There was also evidence of periodicity of 9.75 bp

4.3.3 Relationship between linker length and intron length

Previous studies using the nucleosome positioning signal [AA/TT] dinucleotide have shown that nucleosomes are likely to be positioned at 5'intron/exon3' boundaries (Denisov *et al.*, 1997). Further modelling based on sequence periodicities around intron/exon boundaries suggested that the nucleosome dyad axis is likely to be positioned one or two bases upstream of the intron/exon boundary (Kogan and Trifonov, 2005). This would place half of the nucleosome in the exon and half in the intron. If nucleosomes are positioned at both the 5'intron/exon3' and 5'exon/intron3' boundaries, it may be expected that average intron length is related to linker length. In order to test the hypothesis that such a relationship exists in Arabidopsis, the lengths of *exons*, *introns* and the distance from the start of an intron to the end of the following exon (*5'intron-exon3'*) and the distance from the start of an intron to the end of the following intron (*5'intron-exon-intron3'*) were calculated as shown in the schematic (Figure 4.8).

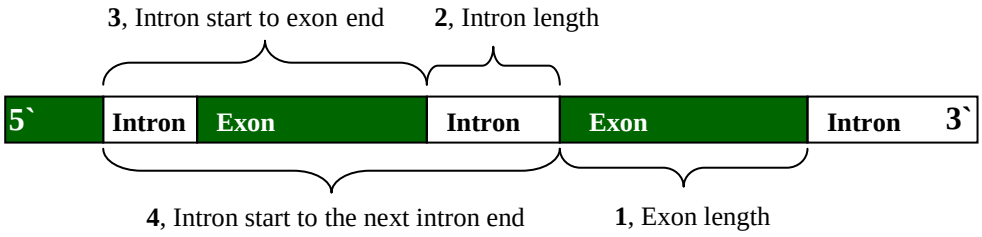

Figure 4.8 Schematic showing how lengths of exons (1), introns (2), 5'intron-exon3' (3) and 5'intron-exon-intron3' (4) were determined.

The length distributions were calculated and are shown in Figure 4.9. The mean ($\bar{y}$), median and standard deviations of the lengths for the *exon* ('1' in Figure 4.8), *intron* ('2' in Figure 4.8), *5'intron-exon3'* ('3' in Figure 4.8) and *5'intron-exon-intron3*' ('4' in Figure 4.8) datasets were calculated. The *exon* DNA fragments have a maximum in the distribution (modal group) at 73 bp, coinciding with a half-nucleosome distance (73.5 bp). The maxima for the *intron*, *5'intron-exon3'* and *5'intron-exon-intron3'* DNA fragments were 85, 182 and 275 bp respectively.

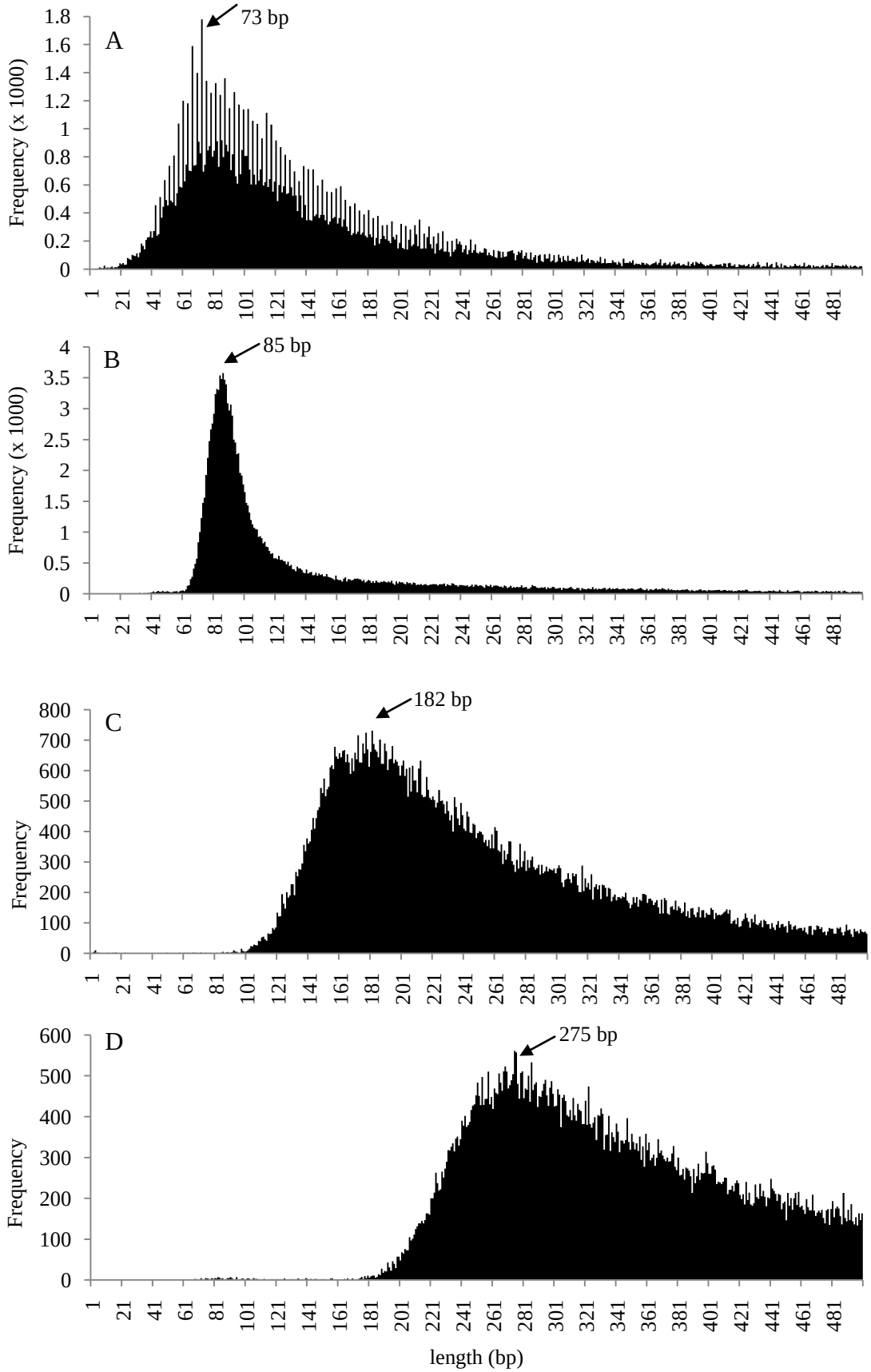

Figure 4.9 Distributions of the length of Arabidopsis genomic components, A: exons, B: introns, C: the distances from intron start to exon end, and D: the distance from intron start to the next intron end.

To further investigate the nucleosome positioning at 5'intron/exon3' boundaries, periodicities within the distributions of the *exon, intron, 5'intron-exon3'* and *5'intron-exon-intron3*' datasets were determined using Fourier analysis. The calculated periodicities are shown in Table 4.5. A periodicity of 3 bp was detected within the exon length distribution. This is most likely the signal associated with the 3 bp codon length reported previously (Choon and Yan, 2008). A periodicity of ~7 bp was detected in the distribution of Arabidopsis intron length which is close to the periodicity detected within the distribution of linker lengths. The periodicities of ~ 8.3 bp detected within the distributions of the *5'intron-exon-3'*and the *5'intron-exon-*intron3' datasets are different to the ~7 bp detected in the distribution of linker lengths.

Table 4.5 The mean, standard deviation and median lengths for Arabidopsis exons, introns, intron start to exon end and from intron start to the next intron end. In addition, periodicities in the distribution, obtained by Fourier analysis, for each feature are shown.

Dataset	($\bar{y}$) (bp)	Median (bp)	Standard Deviation (bp)	Kurtosis	Skewness	Periodicity (Fourier) (bp)	Confidence limit (bp)
exon	170	116	202	98.30	7.07	3.00	2.99 - 3.01
intron	165	99	179	174.07	7.87	7.19	7.14 - 7.24
5'intron-exon3'	322	242	265	74.73	5.70	8.33	8.26 - 8.40
5'intron-exon-intron3'	474	375	321	52.79	4.64	8.36	8.26 - 8.40

The above analyses were repeated for the first introns relative to the 5' end of transcriptional units, as a subset, since first introns have been shown to possess distinct properties from subsequent introns (Jeong, *et al.*, 2006). First introns have been shown to both enhance and inhibit transcription (Rose *et al.*, 2008) and are on average longer in length than subsequent introns (Bradnam and Korf, 2008). The distribution of first intron lengths are shown in Figure 4.10. The mode of the distribution (indicated by an arrow on the histogram) is the same for the first introns as it is for all introns. The mean length is 253 bp, with a standard deviation of 268 bp, and the median is 155 bp. When tested using Fourier analysis, the distribution of first intron lengths showed a 7.57 bp (CL: 7.51 – 7.63 bp) period. This value is slightly higher than the periodicity in the distribution of all introns, and falls within the confidence limits for complete Arabidopsis wild-type datasets.

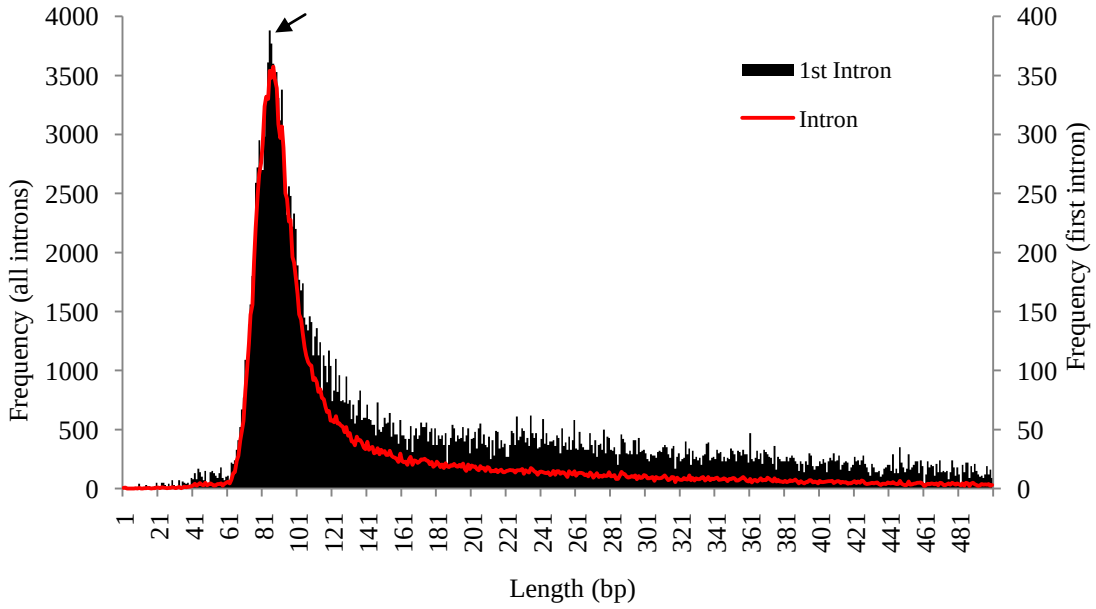

Figure 4.10 Distribution of Arabidopsis first intron lengths, the arrow indicates the mode with a value of 85 bp. The distribution of all introns (from Figure 4.6) is also shown by the red line for comparison. For clarity, the frequencies for the two distributions are shown on different *y*-axes; the frequencies for all introns on the left-hand axis and those for first introns on the right-hand axis.

4.3.4 Linker length variation in an Arabidopsis nucleolar organiser region

In order to test the hypothesis that linker length, and therefore chromatin structure varies between different regions in the Arabidopsis genome, two regions containing similar rRNA gene sequences were compared, these being from 1,000 bp to 6,945 bp on Chromosome 2 and from 14,205,903 bp to 14,211,902 bp on Chromosome 3, as described in the methods (section 4.2.4). Only linker lengths which were calculated from dinucleosome sequences which uniquely map to either Chr 2 or Chr 3 were included in the analyses. The distributions are shown in Figure 4.11. As the Arabidopsis linker length distribution has been shown to have a periodicity of ~ 7-8 bp in this study, a smoothed curve over a sliding window of 8 bp was calculated to reveal overall trends in both distributions. This shows a peak in the distribution of linker lengths at 67 bp within the dataset *Ath_03_rna2*, and two peaks in the distribution of linker lengths, at 30 bp and 42 bp, for the *Ath_03_rna3* dataset.

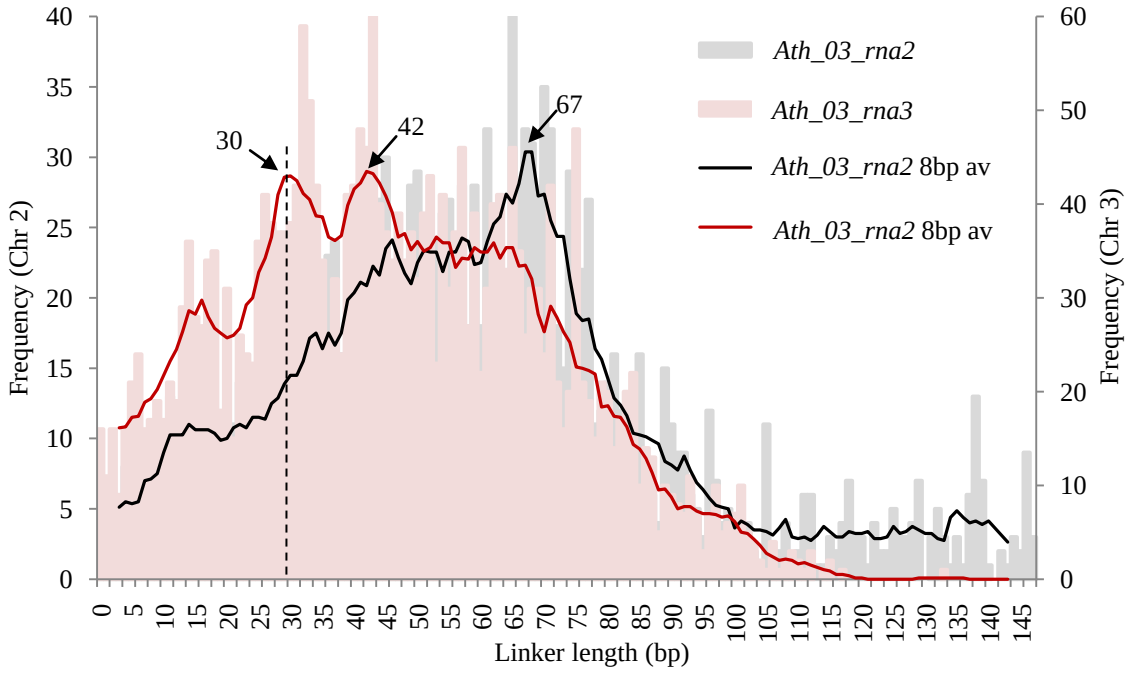

Figure 4.11 Distribution of Linker lengths for datasets *Ath_03_rna2* and *Ath_03_rna3*. An 8 bp moving average is indicated for the *Ath_03_rna2* distribution by a black line and for the *Ath_03_rna3* distribution by a red line. Peaks at 67 bp in the *Ath_03_rna2* distribution and at 30 bp and 42 bp in the *Ath_03_rna3* distributions are indicated. For clarity, distributions for datasets *Ath_03_rna2* and *Ath_03_rna3* are plotted on different *y*-axis scales (Chr2 on the left and Chr3 on the right).

Basic statistics of the distributions and are shown in Table 4.6. Linker lengths of *Ath_03_rna2* have a median of 58.6 bp, while the linker lengths from *Ath_03_rna3* have a median of 46.5 bp. Fourier analysis was used to identify any periodicity in the distribution of linker lengths for each dataset *Ath_03_rna2* and *Ath_03_rna3* (Table 4.6). The periodicity detected in the distribution of linker lengths from the *Ath_03_rna2* is very similar to the periodicity calculated for the whole *Ath_03_wt* dataset (Table 4.4). However, the periodicity detected in the distribution of linker lengths for the dataset *Ath_03_rna3* is higher than that for the whole *Ath_03_wt* dataset. This suggests that different linker length preferences may exists for different regions in Arabidopsis chromatin.

Table 4.6 Mean, median, standard deviation and periodicity (Fourier analysis) of datasets *Ath_03_rna2* and *Ath_03_rna3.*

Dataset	*n*	*Mean* ($\bar{y}$)	Median	Standard deviation	Periodicity (Fourier)	CL for periodicity
Ath_03_rna2	1719	59.63	58	30.59	7.73	7.35-8.17
Ath_03_rna3	2752	46.49	45	24.42	8.65	8.17-9.19

In order to test for multimodality in the linker length distributions, Kernel Density Estimation was used to smooth the distribution. This allows the identification of important peaks within a distribution. The probability density plots for the datasets *Ath_03_rna2* and *Ath_03_rna3* are shown in Figure 4.12. The probability density plots show non-normal distributions for both datasets. The *Ath_03_rna2* dataset shows a main peak at 67 bp and further peaks at 51bp and 123 bp. The *Ath_03_rna3* dataset shows peaks at 43 bp and 51 bp, with a further peak observed at 68 bp.

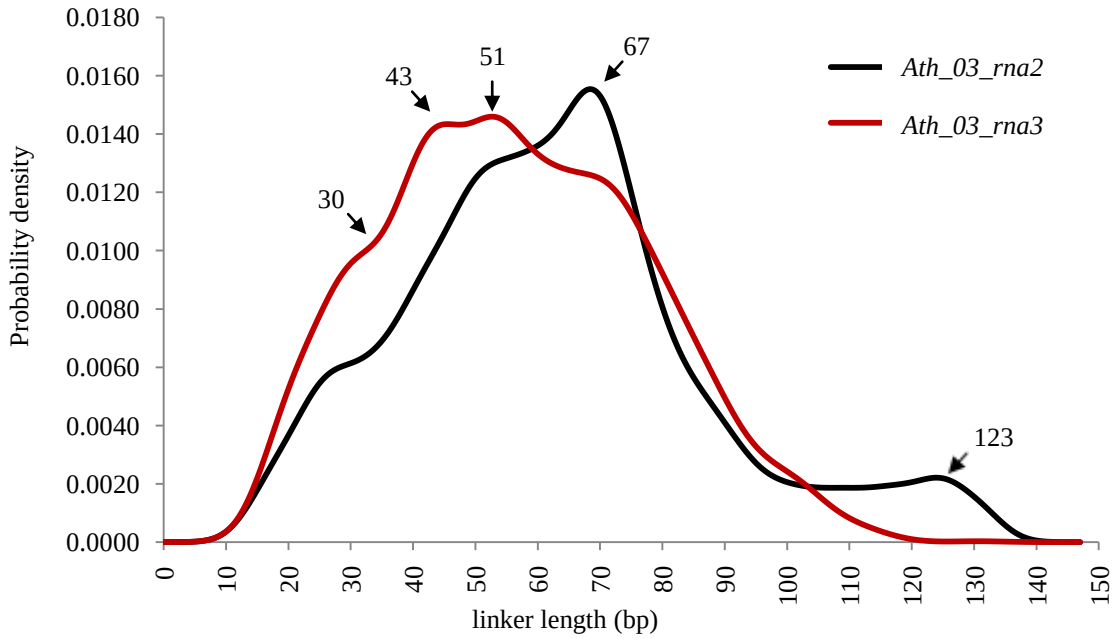

Figure 4.12 Probability density plots for datasets *Ath_03_rna2* and *Ath_03_rna3*. The dataset *Ath-03_rna2* is represented with a black line and *Ath_03_rna3* is represented by a red line. Arrows indicate important peaks (linker lengths) in the smoothed distributions detected by Kernel Density Estimation.

Kurtosis and skewness measurements were calculated for each probability density plot and are shown in Table 4.7. Kurtosis gives an indication of the relative 'peakedness' of a distribution compared to the normal distribution, with positive values indicating a peaked distribution and negative values indicate a relatively flat (non-peaked) distribution. The kurtosis value for each dataset is negative, with the plot for the *Ath_03_rna2* having a kurtosis of -0.91, which is higher than the value for the *Ath_03_rna3* plot (-1.53). This suggests that the *Ath_03_rna2* smoothed distribution has more peaks, or a more important peak than the smoothed distribution of the *Ath_03_rna3* dataset. The skewness measurement gives an indication of the symmetry of a distribution around the mean. A positive value indicates a skewed distribution with a tail extending towards positive values (higher than the mean) and a negative value indicates a skewed distribution with a tail extending towards negative values (lower than the mean). The smoothed distributions of both datasets *Ath_03_rna2* and *Ath_03_rna3* have positive values indicating the distributions are skewed with more values higher than the mean. The skewness value for the *Ath_03_rna2* is higher than that for the *Ath_03_rna3* distribution, indicating that the distribution of the *Ath_03_rna2* has more data points with values higher than the mean (i.e. is more skewed towards higher values).

Table 4.7 Kurtosis and skewness values for the smoothed distributions of linker lengths from the *Ath_03_rna2* and *Ath_03_rna3* datasets.

Dataset	Kurtosis	Skewness
Ath_03_rna2	-0.91	0.70
Ath_03_rna3	-1.53	0.30

To determine whether there are significant differences in the distributions of linker lengths from datasets *Ath_03_rna2* and *Ath_03_rna3*, a Kolmogorov-Smirnov test was used. The cumulative distribution function for each distribution was calculated, and the difference between the cumulative distributions at each linker length was determined. The *D*-statistic, which is the largest difference between the cumulative distributions, was 0.1019. This is larger than the critical *D*-value, indicating that the distributions of linker length for datasets *Ath_03_rna2* and *Ath_03_rna3* are significantly ($p < 0.05$) different. The differences between the cumulative distributions were above the critical value for all linker lengths between 34 bp and 116 bp. The cumulative distributions are shown in Figure 4.13.

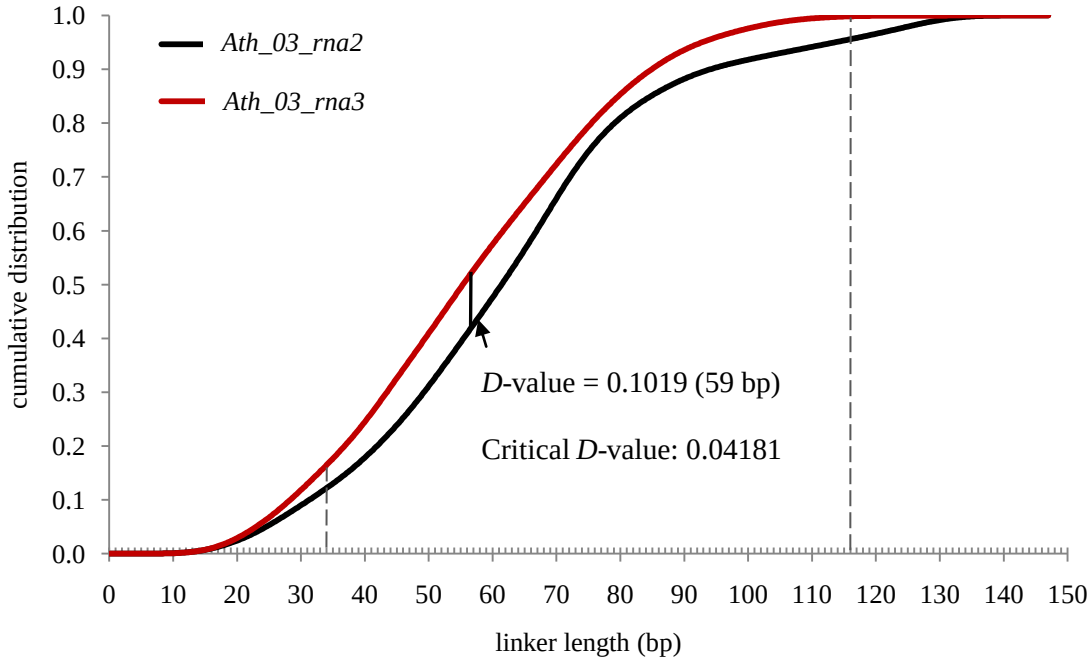

Figure 4.13 Cumulative distribution function plot for datasets *Ath_03_rna2* and *Ath_03_rna3*. The value of the Kolmogorov-Smirnov *D*-statistic is indicated by a solid vertical black line. The cumulative distributions are significantly different at all linker lengths between the dashed lines (between linker lengths from 34 bp and 116 bp).

4.4. Discussion

4.4.1 Arabidopsis linker length distribution

Linker length was estimated from Arabidopsis dinucleosome sequences by subtracting twice the length of the nucleosome core (147 bp). Steric hindrance between nucleosomes is thought to impose constraints on linker lengths at short distances from the nucleosome, with linker lengths of up to 5bp, and from 10 to 15 bp thought to be unfavourable (Mengeritsky and Trifonov, 1983). Other, more recent studies, consider 10 bp as the minimum linker length (Segal *et al.,* 2006, supplementary data). However, there appears to be no agreement regarding minimum linker length in the literature.

In this study, linker lengths were calculated from the positions of mononucleosomes for the *Ath_02_wt, Ath_04_MET1* and the *C. elegans* data sets. These linker lengths represent the relative distance between mononucleosomes from mapped positions of individual nucleosomes rather than the direct measurement of the physical distance between two nucleosomes, as derived from a dinucleosome dataset. Therefore, the mapped positions cannot be as reliable an indicator of linker length as linker length datasets derived from dinucleosome DNA fragment lengths. However, calculation of linker lengths from mapped mononucleosome positions has been successfully used to estimate linker lengths in yeast (Lubliner and Segal, 2009). Linker length was used as a measure of inter-nucleosome interactions, which were incorporated into a nucleosome prediction model. The distribution of derived yeast linker lengths showed a 10 bp periodicity, which had previously been demonstrated in yeast chromatin (Wang *et al.,* 2008). This suggests that linker lengths derived from mapped mononucleosome positions are likely to be representative of *in vivo* linker lengths.

Linker lengths displayed a non-random distribution, in accordance with previous observations (Lohr and Van-Holde, 1979, Kato *et al.,* 2003, Cohanim *et al.,* 2005). The distributions of linker lengths from the *Ath_03_wt* and *Ath_04_MET1* derived from mononucleosomes did not show the same ‘peaked’ distributions as the other datasets. This may be due to under-sampling for the *Ath_01_wt*, *Ath_04_MET1* (dinucleosome) and *Ath_02_wt* datasets, which are therefore relatively small. The

differences between distributions, in terms of periodicity, could be tested by using a large number (~1000) of random subsets of the data (linker lengths) from each dataset. The periodicity for each subset would then be calculated to provide a 'bootstrap' (Effron and Tibshirani, 1994) mean periodicity and 95% CL, for each distribution. Non-overlapping CLs would then indicate statistically significant differences between the distributions.

To compare Arabidopsis linker lengths with those from other taxa, linker lengths were collated from the literature (Table 4.8). The mean linker lengths observed in this study for Arabidopsis wild-type were found to be between 52 bp and 64 bp, with the exception of the *Ath_03_wt* dataset which had a mean of 40 bp. This suggests that Arabidopsis linker length falls closest to those calculated from human and chicken chromatin. The mean linker length for the *Ath_03_wt* dataset falls between those calculated from human and yeast chromatin. A general feature of the current and previous studies is considerable variation in the estimation of linker length from the same taxa. This may reflect variation associated with cell and tissue-specific expression patterns.

Table 4.8 A list of linker lengths for different species collected from literature.

Organism	Tissue type/condition	Mean linker length (bp)	Reference
C. elegans		68	Johnson *e al.,* 2006
B. taurus	calf brain, cortical neurone	21	Allan *et al.,* 1984
Gallus spp.	erythrocyte	58*	Satchwell and Travers, 1989
Gallus spp.	erythrocyte	68**	Satchwell and Travers, 1989
H. sapiens	K562 cells	28*	Kato *et al.,* 2003
H. sapiens	K562 cells	48**	Kato *et al.,* 2003
H. sapiens	-	38	Kahamann and Rake, 1993
H. sapiens	Down's syndrome	26	Kahamann and Rake, 1993
Kluyveromyces lactis	-	24	Heus *et al.,* 1993
M. musculus	bulk chromatin	48	Cioffi *et al.,* 2006
M. musculus	liver	25	Cioffi *et al.,* 2006
Rattus spp.	pre-natal brain	53	Jaeger and Kuenzle, 1982
Rattus spp.	post-natal brain	23	Jaeger and Kuenzle, 1982
Tetrahymenia spp.	telomeres	7	Cohen and Blackburn, 1998
S. cerevisiae	grown to lag phase	44***	Yuan *et al.,* 2005

*Calculated from all DNA fragments
**Calculated from DNA fragments greater than 294 bp
*** Calculated from nucleosome positions defined using tiling arrays

Previous studies have demonstrated differences in linker length between cells types exhibiting different transcriptional activity (Grigoryev *et al.,* 1990). As the data presented here are generated from many cell types (within an organ), it is likely that multiple chromatin conformations are represented.

4.4.2 Periodicity with linker length distribution in Arabidopsis

The periodicity observed in the distribution of Arabidopsis linker lengths is different from the reported 10 bp periodicity for yeast and human chromatin (Widom, 1992, Kato *et al.,* 2003, Wang *et al.,* 2008). However linker lengths of multiples of 7 bp have been reported previously for reconstituted nucleosomes with short linkers (Noll *et al.,* 1980). The Arabidopsis anti-sense *MET1* did exhibit the ~10 bp periodicity in the linker length distribution. This suggests that the overall chromatin structure of Arabidopsis anti-sense *MET1* may be different from wild-type (by chromatin re-modelling). DNA methylation has been shown to both inhibit and promote nucleosome formation in human and mouse imprinting control regions (Davey *et al.,* 2003). Studies of chromatin conformation in mouse embryonic stem cells which lack Dnmt3a and DNmt3b showed an altered binding of linker histone (Gilberts *et al.,* 2007). This revealed that linker histones are less mobile (or more-tightly bound) in methylation-deficient mouse cells. This is thought to be due to increased binding in the linker rather than the nucleosome. However, the reduced DNA methylation was not thought to alter chromatin compaction.

It has previously been suggested that the discrepancy between the periodicity of linkers obtained from dinucleosome sequences and the 10 bp repeat is an artefact of the method of dinucleosome DNA fragment preparation (Cohanim *et al.,* 2005). If the periodicity detected is indeed due to an experimentally induced artefact, and the true value is around 10 bp, then the discrepancy is likely to be seen in every case where the method has been implemented. This was not the case when investigating the linker length variation in Arabidopsis, human and *C. elegans*. It is therefore likely that the distributions of linker length and the periodicities they exhibit are a true reflection of the state of the chromatin at the time of nucleosome extraction. Any deviations from the proposed 10 bp periodicity are likely to arise from structural changes in chromatin conformational brought about by chromatin remodelling. It is particularly interesting that when the Arabidopsis genome is de-methylated (<20%

wild-type levels), the periodicity in linker length distribution returns to the previously reported 10 bp. This suggests that there may be a preferred 'default' setting for the positioning of nucleosomes, which is over-ridden by other epigenetic processes in wild-type (functional) chromatin.

4.4.3 Relationship between linker length and intron length

Previous *in vivo* studies and computational analysis of DNA sequences have shown that nucleosomes are likely to be positioned at 5'intron/exon3' boundaries (Denisov, *et al.,* 1997). The results of these studies suggest that nucleosomes are positioned with the nucleosome dyad axis 0 bp to 15 bp from the intron/exon boundary, positioning the one half of a nucleosome within the intron and the other half in the exon. If nucleosomes are indeed positioned at 5'intron/exon3' boundaries, it may be expected that the length of the 5'intron/exon3' unit would reflect the nucleosome repeat length (NRL = one nucleosome and one linker). In this study, the distribution of intron length was compared to the distribution of nucleosome linker length. The distribution of the 5'intron-exon3' has a mean length of 322 bp, which does support the expectation, although the exon and intron length distributions have mean lengths of 170 bp and 165 bp respectively, which is consistent with the length of a nucleosome core and one molecule of histone H1 (~165 -167 bp). This may suggest that nucleosomes are positioned more often within either the exon, or intron regions, rather than with the nucleosome dyad positioned at the 5'intron/exon3' boundary as previously suggested. However the mean GC content of Arabidopsis introns is lower than that observed in Arabidopsis nucleosome core sequences (Chapter 3, Section 3.3.2.ii), which suggests that nucleosomes are positioned more often within exonic DNA.

To investigate nucleosome positioning at 5'intron/exon3' boundaries further, the periodicity in the distributions of the exon, intron and 5'intron/exon3' lengths was determined. The distribution of intron length has a periodicity of 7.87 bp, which is close to the periodicity of Arabidopsis linker length determined in this study, whereas the *5'intron-exon3'* dataset has a periodicity of 8.33 bp. This suggests that nucleosomes may be positioned more often in exon regions, with introns occurring more often in linker regions.

The first introns relative to the 5' end of a transcriptional unit have been shown to have properties which distinguish them from subsequent introns in Arabidopsis. This includes them having longer length (Bradnam and Korf, 2008), and motifs which can promote and inhibit gene expression (Chung *et al.*, 2006). The first intron length in Arabidopsis has a mean of 253 bp and a median of 155 bp. The periodicity in the distribution of first intron length was 7.57, which is slightly lower than the periodicity in the distribution of all intron lengths. Although nucleosome positioning may be different at first introns compared to subsequent introns, this was not supported by the periodicity in the length distribution determined in this study. It would be interesting to compare specific histone modifications to nucleosome occupancy at first and subsequent introns in different length classes in order to investigate this further.

4.4.4 Effect of linker length on higher-order chromatin structure

Electron microscopy of chromatin fibres reconstituted from ordered nucleosome arrays possessing linkers of known lengths has shown that linker length variation dramatically influences the conformation of higher-order chromatin structure (Robinson *et al.*, 2007). From these experiments, it appears that the diameter of the 30 nm chromatin fibre is related to linker length in a non-linear fashion. Chromatin fibres consisting of nucleosomes with 30 bp to 60 bp of linker DNA tend to have a diameter of ~35 nm, whereas fibres consisting of nucleosome with 70 bp to 90 bp of linker DNA tend to have a diameter around 45 nm. These measurements were used to model the trajectory of the linker DNA within the 30 nm chromatin fibre for each linker length (Wong *et al.*, 2007). The most energetically favourable conformation was described for each linker length from 30 bp to 90 bp, resulting in seven different chromatin fibre conformations. A comparison of the modelled chromatin conformations is reproduced (Wong *et al.*, 2007), and displayed along with the linker length distribution of Arabidopsis anti-sense *MET1* in Figure 4.14. There appears to be a correspondence between the proposed chromatin fibre structures and the peaks in the linker length distribution of the *Ath_04_MET1* dataset.

The Wong *et al.* (2007) model suggests a shortest permissible linker length of 30 bp, with chromatin fibres having linker lengths between 30 bp and 60 bp having a fibre diameter of 35 nm. It also suggests that there is a change in chromatin fibre diameter between 60 bp and 70 bp from 35 nm to 45 nm. In the distribution of *Ath_04_MET1* linker length, there appear to be relatively few linkers between 62 bp and 63 bp, with 60.5 % of the distribution lying between 30 bp and 59 bp. This may reflect the relative distribution of chromatin fibre-class in Arabidopsis anti-sense *MET1*. A further 25.5 % of the *Ath_04_MET1* linker lengths fall between 60 bp and 90 bp, which correspond to the 45 nm chromatin fibre in the Wong *et al.* (2007) model (with the remaining 14.0% accounting for values 0-29 bp and 91-147 bp in length). This analysis suggests that the majority of chromatin in Arabidopsis anti-sense *MET1* is organised into chromatin fibres of ~35 nm in diameter.

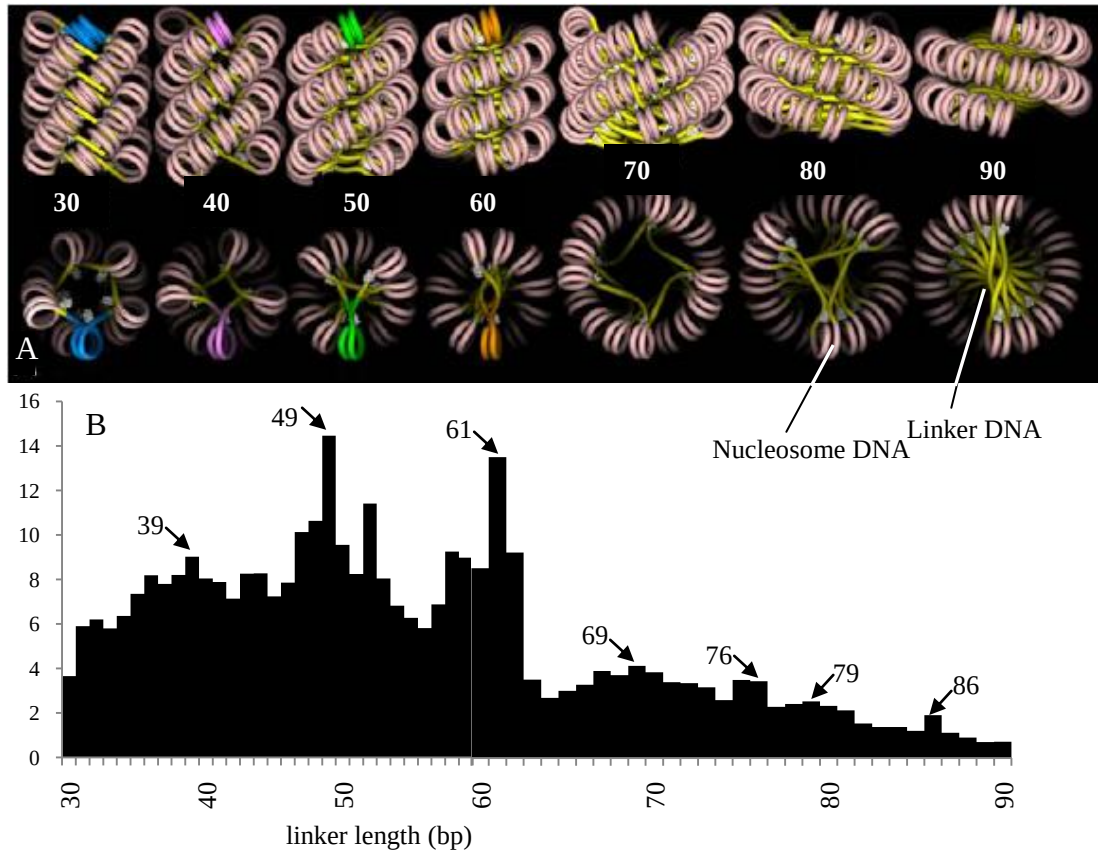

Figure 4.14 Demonstration of the possible relationship between A: models of the structure of the 30 nm fibre at increasing linker lengths (reproduced from Wong *et al.*, (2007)) and B: Arabidopsis linker length calculated from the dinucleosome sequences of the *Ath_04_MET1* dataset. The *x*-axis shows the frequency of each length (x 100). Each linker length is aligned to the chromatin proposed structure for that linker length. Peaks in the linker length distribution which approximately correspond to different chromatin conformations in the Wong *et al.* (2007) model are indicated.

The differences within the Arabidopsis distributions are not as apparent for any of the Arabidopsis wild-type datasets, which may be a feature of the different sample sizes of the datasets.

The Wong *et al.* (2007) model of the 30 nm fibre structure assumes a linker histone (H5 rather than H1 in this model) at each nucleosome, although other studies suggests that this is not the case for many higher eukaryotes. The number of molecules of linker histone per nucleosome was found to range from 0.03 (yeast) to 1.3 (chicken) (Woodcock *et al.,* 2006). The presence of linker histone (H1/H5) is thought to promote chromatin condensation, although the level of compaction is thought to be affected by linker length (Routh *et al.,* 2008).

4.4.5 Nucleosome positioning in the nucleolar organiser region in Arabidopsis

The rDNA regions, on Arabidopsis Chromosomes 2 and 3, were chosen to investigate potential differences in linker length in regions with similar DNA sequence, since they were well represented in all datasets. It is likely that the large number of samples acquired is due to the 730 ± 100 rRNA repeats (NOR, chromosomes 2 and 4) which are not accounted for explicitly in the annotation of the Col-0 (TAIR v8) reference genome.

The rRNA genes have been shown to be associated with both active (euchromatic) and inactive (heterochromatic) chromatin marks in Arabidopsis (Lawrence *et al.,* 2004). Bisulphite-treated rRNA promoters revealed that most are methylated, and chromatin immunopreciptitation (ChiP) revealed that methylated promoters were associated with H3K9me^2, a mark of heterochromatin. Conversely, hypomethylated promoters were associated with H3K4me^3 and Pol I, which indicated that those rRNA genes are active (Lawrence *et al.,* 2004). With this in mind, the linker length distributions of the two rRNA regions were investigated to determine if two different chromatin structures could be detected. The two regions, represented in datasets *Ath_03_rna2* and *Ath_03_rna3* have different linker lengths: the *Ath_03_rna2* distribution has a mean of 60 bp and a median of 58 bp while the *Ath_03_rna3* distribution has a mean of 46 bp and a median of 45 bp. This was supported by the Kolmogorov-Smirnov test for differences; the cumulative distributions of linker lengths from both datasets were statistically significantly different ($p < 0.05$) from

each other. This suggests that the two regions are likely to have different chromatin structures and could reflect differences in the regulation of rRNA genes in the two regions.

4.4.6 Structure of the rRNA regions on Chromosomes 2 and 3

The positions of the 18S and 5S rRNA genes on Chr 2 and 3 were determined in this study using the annotation from the TAIR v8 Col-0 reference genome. The annotated rRNA genes on Chr 2 are most likely a single-copy representation of the nucleolar organiser region at the beginning of Chr 2, which comprises of 730 ±100 copies of the rRNA region (Wambutt *et al.,* 2000). The 18S and 5S genes are also present in the centromeric region of Chr 3. However this is thought to be just a single copy (European Union Chromosome 3 Arabidopsis Sequencing Consortium, 2000) and therefore, it does not seem likely that the large sampling of the nucleosome data represents the rRNA genes present on Chr 3. It may be that the data corresponds to copies of the rRNA genes in the NOR on Chr 4, as this region is absent in the annotated Arabidopsis reference genome (TAIR v8) and the sequence is almost identical (Wambutt *et al.,* 2000). In addition, at the time of writing, the 25S rRNA gene is also surprisingly still absent from the annotation of the reference genome (TAIR v8), although the sequence has been published (NCBI accession X52320, Unfried and Gruendler, 1991).

During the time the analyses of linker length distribution in the rRNA regions were performed, a new version of the Arabidopsis reference genome became available (TAIR v9, June, 2009), but the updated annotation still lacks the 25S, ITS, IGR and NOR4 regions. In addition, the updated reference genome also translocates the relative position of the annotated rRNA genes present on Chr 3, but not the sequence, overall, so the new version is unlikely to affect the results of the linker length distribution analyses. Alignment of the published 25S rDNA sequence to the reference genome (TAIR v9) places the 25S rRNA gene alongside the 5S rRNA gene, and the ITS regions between the rRNA genes (Figure 4.15). Alignment of the intergenic region (IGR) alongside the rRNA genes was necessary as it also was not represented in the annotated reference genome. The Arabidopsis IGR sequence has

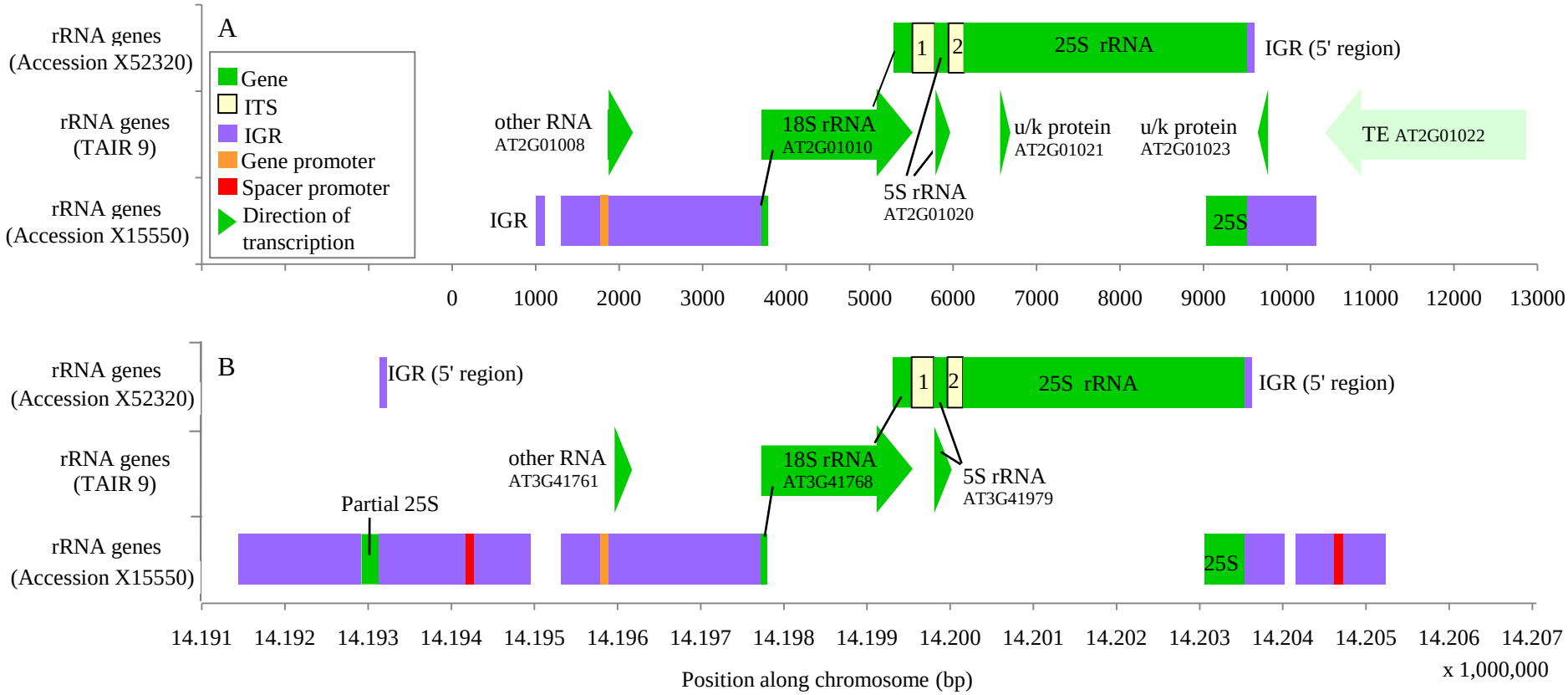

Figure 4.15 Structure of the annotated rRNA regions on A: Chromosomes 2, and B: Chromosome 3. Gene positions are shown in green, the ITS regions in yellow, IGR regions in purple, the gene promoters in orange and the spacer promoters in red. Accessions X15550 (Arabidopsis IGR) and X52320 (Arabidopsis ITS and 25S regions) were aligned to the Arabidopsis reference genome (TAIR v9) to position the non-annotated 25S, ITS1, ITS2 and IGR. In addition, the positions of the gene and spacer promoters are shown within the IGR. The transposable element on Chromosome 2 (AT2G01022) signals the end of the rDNA region in the annotated genome. The absence of the spacer promoter and some of the IGR from Chr 2 suggests that part of the sequence is missing from the annotation.

previously been published (NCBI accession X15550, Luschnig *et al.,* 1993), and aligns partly upstream of the annotated rRNA gene positions, and partly downstream, which is unsurprising as it is part of the repeating unit and may signal the start and end of the rRNA region in the annotation. The gene and spacer promoter DNA sequences (Doelling *et al.,* 1993) were also aligned (shown with the IGR positions, Figure 4.11), and this shows that part of the IGR containing the spacer promoter is missing from the annotated region on Chr 2.

4.4.7 Summary

The distribution of linker length in Arabidopsis wild-type and anti-sense *MET1* chromatin was investigated and found to be within a similar size range to that previously identified for other taxa. A periodicity of 7-8 bp was detected within the wild-type dinucleosome distributions which differed from the expected ~10 bp periodicity. Investigation of periodicity in the distributions of previously published nucleosome datasets revealed a period of ~8 bp in human chromatin and 5 bp in *C. elegans* chromatin, which were also different from the expected 10 bp. A period of ~9-10 bp was detected in the distribution of linker lengths from the Arabidopsis anti-sense *MET1*, suggesting that DNA methylation is affecting nucleosome positioning and higher-order chromatin structure. Similarity between the periodicity of linker length and intron length distributions were detected, which suggest there may be a relationship between nucleosome spacing and intron length. It is interesting that the average intron (165 bp) and exon (170 bp) lengths are both close to the size of a nucleosome and linker histone (~167 bp). Of particular interest is the distribution of linker length within the rRNA regions. Dinucleosome sequences which map to either Chr 2 or Chr 3 were used to investigate region-specific chromatin structure in wild-type chromatin. This revealed two subsets with different distributions in linker length, possibly reflecting differences in higher order chromatin structure and gene regulation over the NORs. Unfortunately this analysis was not possible to perform with data from the anti-sense *MET1*. The *Ath_04_MET1* dataset consists mostly of mononucleosome DNA fragments and the extensive sampling over the rRNA regions resulted in only overlapping mononucleosome sequences, making linker estimation extremely difficult.

Chapter 5

Nucleosome distribution and patterns of occupancy

5.1 Introduction

Recent advances in high-throughput, technologies have massively increased the size and number of nucleosome datasets, making possible, concurrent with this thesis, study of genomic organisation patterns of nucleosome occupancy. Large-scale sequencing of nucleosome DNA fragments, and hybridisation of DNA to tiling arrays, in yeast and *C. elegans*, have in the last 2 – 3 years revealed that nucleosomes have a non-random genomic distribution, and tend to be positioned more often within the coding regions, whilst intergenic regions are relatively nucleosome depleted (Lee *et al.,* 2007; Johnson *et al.,* 2007; Yuan *et al.,* 2005).

In recent studies, the role of nucleosome positioning in the promoter regions has attracted much attention. Hybridisation of yeast mononucleosome DNA fragments to a high-resolution, overlapping tiling array revealed a nucleosome-free region (NFR), approximately 150 bp in length, commencing 200 bp upstream from annotated genes (Yuan *et al.,* 2005). NFRs in the regions upstream of the transcriptional start site (TSS) have since been found in a variety of organisms including *Drosophila* (Mavrich *et al.,* 2008a), *C. elegans* (Johnson *et al.,* 2007) and human (Schones *et al.,* 2008; Ozsolak *et al.,* 2007). NFRs are thought to occur at all promoters in the yeast genome, regardless of transcriptional activity, and this is thought to reflect a 'poised for transcription' chromatin state (Yuan *et al.,* 2005). In contrast, the *Drosophila* genome does not appear to have NFRs throughout the genome. Thus genes with the promoter motifs: TATA, Inr or MTE, lack the NFR (Mavrich *et al.,* 2008a).

NFRs have been found to be flanked on either side by a well-positioned nucleosome within genes transcribed by Pol II (Yuan *et al.,* 2005). Furthermore, these flanking nucleosomes were found to contain the histone variant H2A.Z in yeast (Albert *et al.,* 2007), whereas H2A.Z was only found within nucleosomes of highly transcribed genes in *Drosophila* (Marvich *et al.,* 2008a). In addition, the H2A.Z-containing nucleosomes adopted a more ordered pattern of organisation than bulk nucleosomes, and are thought to confer a more open chromatin structure.

For transcription to occur, a pre-initiation complex, comprising of the transcription factor, Pol II and activator proteins is bound at the promoter prior to transcription. Current evidence suggests that to allow for binding of the pre-initiation complex, the positioned nucleosome upstream of the TSS is either re-positioned or removed (Workman, 2006).

5.1.2 Aims of study in this chapter

Whilst much effort has been put into study of nucleosome organisation around the promoter/transcriptional start site, and to some extent the 5' and 3' ends of genes and the mechanism of transcription, the occupancy around the exon, intron and intergenic regions have been neglected. As the positioning around the TSS varies between yeast and *Drosophila*, it is likely that species-specific patterns of nucleosome organisation may exist both within the transcriptional units and in their 5' and 3' regions.

With this in mind, the following hypotheses were constructed and tested:

1. Nucleosomes are evenly distributed throughout the Arabidopsis genome.

2. Biases in nucleosome occupancy differ between coding and non-coding regions in Arabidopsis.

3. Ordered positioning of nucleosomes exists around the TSSs in Arabidopsis, comparable to the organisation found within other taxa.

4. The Arabidopsis Nucleolar Organiser Regions display different nucleosome patterns of positioning dependent on chromosomal origin, due to the transcriptionally heterogeneous nature of these regions.

5. Overall differences in nucleosome positioning exist between wild-type and the Arabidopsis anti-sense *MET1*.

5.2 Materials and methods

5.2.1 Genomic representation in datasets and distribution of fragments

To determine the distribution between the chromosomes of fragments from each dataset, the percent of each chromosome represented in the nucleosome datasets is given by:

$$n=\left(\frac{F_{(Chr\ x)}}{F_{(total)}}\right)\times 100$$

Where:

n = % represented each chromosome represented within the dataset

$F_{(Chr\ x)}$ = fragments within the dataset which align to Chr x

$F_{(total)}$ = total number of fragments within the dataset

To determine the percent of nucleosome-occupied bases on each chromosome, the number of unoccupied bases was determined calculated by subtracting the start co-ordinate of *n_k+1* from the end co-ordinate of *n_k* (Figure 5.1) to give the distance between occupied bases: nuc_$k_{(end)}$ – nuc_$k+1_{(start}$.

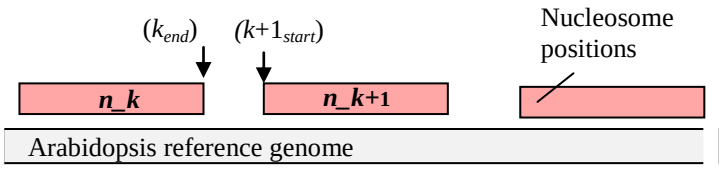

Figure 5.1 Schematic showing the method for calculating the distance between the positions of nucleosome fragments aligned to the Arabidopsis reference genome.

The percent occupancy was calculated using the following equation:

$$\%\ \text{occupancy} = 100 - \left(\left(\frac{\sum b_{(d>0)}}{b_{(total)}}\right)\times 100\right)$$

Where:

$b_{(d>0)}$ = distance between occupied bases, summed over all the gaps between fragments

$b_{(total)}$ = total number of bases in the chromosome.

To investigate the distribution of nucleosomes within chromosome 1, the frequency of the start co-ordinate of each nucleosome DNA fragment was calculated in bin sizes of 10,000 bp. A similar operation was performed to investigate a 2 Mbp region with a 1,000 bp bin size.

5.2.2 Genomic patterns of nucleosome occupancy

To begin to understand patterns of nucleosome occupancy in Arabidopsis, nucleosome occupancy scores were calculated (Chapter 2, section 2.2.vi). A count of nucleosome molecules occupying any given co-ordinate was determined, and an occupancy value assigned to each co-ordinate (Figure 5.2). The nucleosome occupancy scores were initially viewed using a prototype genome browser developed at Rothamsted by Graham King.

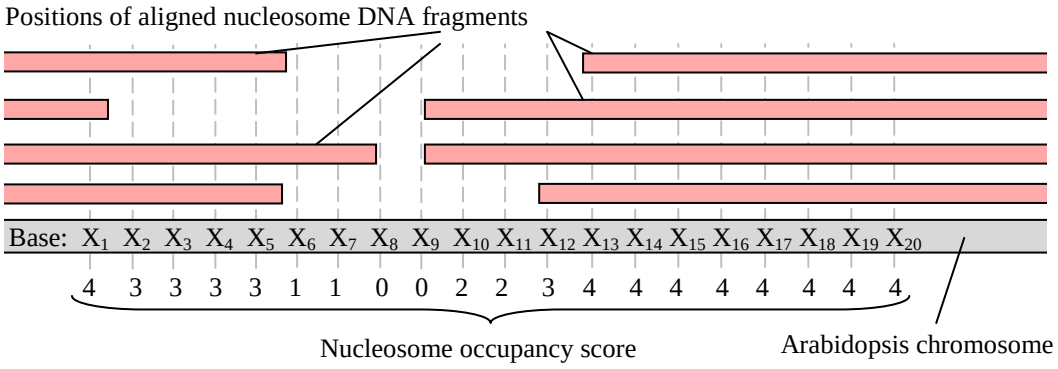

Figure 5.2 Schematic showing the assignment of nucleosome occupancy scores to each position within the Arabidopsis chromosome.

5.2.3 Distribution of occupancy within exons and introns

Annotation classes were assigned to each position, along with the nucleosome occupancy score as described in Chapter 2, section 2.2.vi.

For exons, the average occupancy (number of molecules) per base was calculated using the equation:

$$\text{occupancy/bp} = \frac{\sum\left(b_{(1..n)} \times o_{(1..n)}\right)_{(exon)}}{B_{(total\ exon)}}$$

Where:

b = number of bases (in exon category);

o = number of molecules (occupancy);

$B_{(total\ exon)}$ = number of bases in Arabidopsis nuclear genome DNA (in exon category).

The exon/intron ratios were calculated for nucleosome occupancy, using the formula:

$$ratio_{(exon)} = \frac{Occupancy/bp_{(exon)}}{(Occupancy/bp)_{(exon)} + (Occupancy/bp)_{(intron)}}$$

5.2.4 Nucleosome occupancy around transcriptional start sites

In order to determine the presence and occupancy of nucleosomes around the TSS in Arabidopsis, nucleosome DNA fragment co-ordinates were compared to a set of published, experimentally determined and predicted TSS (Tanaka *et al.*, 2009).

In order to determine what percent of TSS was occupied by at least one nucleosome, the distance between the TSS and nearest upstream nucleosome boundary was calculated by subtracting the co-ordinate of the TSS position from the nucleosome end-position (Figure 5.3), where distances between 0 and 147 (bp) indicate that the TSS is occupied.

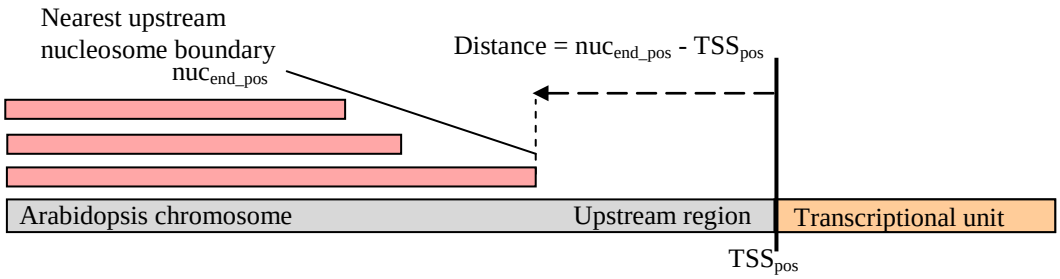

Figure 5.3 Schematic showing the method for determination of the distance between the TSS and the nearest upstream nucleosome boundary.

In order to determine the nucleosome occupancy within the vicinity of the TSS, the distance between the TSS and the nucleosome boundary was calculated by subtracting the midpoint of every nucleosome position from the TSS (where the nucleosome midpoints are positioned to within 1000 bp of the TSS) (Figure 5.4).

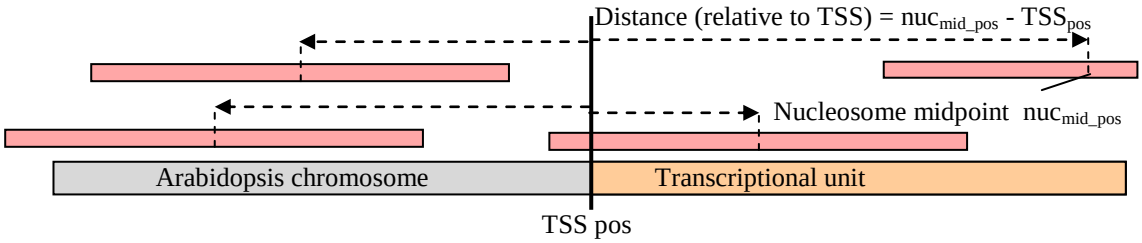

Figure 5.4 Schematic showing the method for determination of the distance between the TSS and every nucleosome midpoint within 1000 bp.

5.2.5 Kernel Density Estimation

In order to inspect the frequency of nucleosome occupancy in the vicinity of the TSS for evidence of ordered nucleosome positioning, Kernel Density Estimation was used. This smoothed the occupancy frequency distribution around the TSS and is able to reveal longer-range peaks in the distributions. The density estimations were performed in GenStat® (2008) with the bandwidth calculated using the method of Sheather & Jones (1991), and using a set of 2,048 grid points.

5.3 Results

5.3.1 Genomic distribution and coverage of nucleosome sequences

In order to determine if the nucleosome sequences of datasets *Ath_03_wt* *Ath_04_MET1* and *Ath_05_wt* are distributed evenly across the genome, the number of sequences which align to each chromosome was calculated and expressed as a percent of the total sequences. The percent of the genome contained within each chromosome was also calculated for comparison and is shown in Table 5.1. Both Chrs 2 and 3 appear to be over-represented in the *Ath_05_wt* dataset. Chr 3 is also over-represented in the *Ath_04_MET1* dataset. This is most likely because of the large number of nucleosome DNA fragments which align to the rRNA genes annotated on Chrs 2 and 3.

Table 5.1 Percent of sequences from datasets *Ath_03_wt, Ath_04_MET1* and *Ath_05_wt* which align to each chromosome. The percent of the genome on each chromosome is shown for comparison.

Chr	Genome	*Ath_03_wt*	*Ath_04_MET1*	*Ath_05_wt*
1	25.53	24.67	24.15	22.99
2	16.53	16.52	16.47	18.01
3	19.69	21.49	22.67	22.57
4	15.59	15.35	15.56	15.07
5	22.65	21.97	21.14	21.37

The number of bases occupied by nucleosomes was calculated for each chromosome and expressed as a percentage of the total number of bases for each chromosome, for the datasets *Ath_03_wt* *Ath_04_MET1* and *Ath_05_wt* (Table 5.2). Occupancy is lower for all chromosomes in the *Ath_05_wt* dataset than the other two. Presence of linker DNA in the *Ath_03_wt* may falsely increase the number of occupied bases due to the wider coverage of each molecule.

Table 5.2 Percent of occupied bases for each chromosome for the datasets *Ath_03_wt, Ath_04_MET1* and *Ath_05_wt.* Values for the total genome are highlighted in bold.

Occupied in each set	Chr 1	Chr 2	Chr 3	Chr 4	Chr 5	Total
Ath_03_wt	66.04	66.05	68.80	67.52	66.89	**67.01**
Ath_04_MET1	63.86	62.88	64.38	63.26	63.76	**63.68**
Ath_05_wt	48.96	50.55	52.71	51.34	50.06	**50.58**

In order to determine the distribution of nucleosome sequences within each chromosome, the 5' (start) co-ordinate of each nucleosome DNA fragment was plotted, using a bin size of 10,000 bp. The distribution of nucleosome sequences from datasets *Ath_02_wt*, *Ath_03_wt* and *Ath_04_MET1* is shown in Figure 5.5. There appear to be some differences between the wild-type plots and the plot of anti-sense *MET1* positions around the centromere region of chromosome 1 (region in the grey box, Figure 5.5). To investigate this further, the co-ordinates of the nucleosome start sites were plotted over a region from 14,000,000 bp to 16,000,000 bp using a 1,000 bp bin size, as shown in Figure 5.6. In the expanded view of the 2,000,000 bp region, there appears to be little difference between the wild-type and anti-sense *MET1* nucleosome start co-ordinates at this scale. Any differences observed appear to be associated with the relative abundance of nucleosome sequences in each bin.

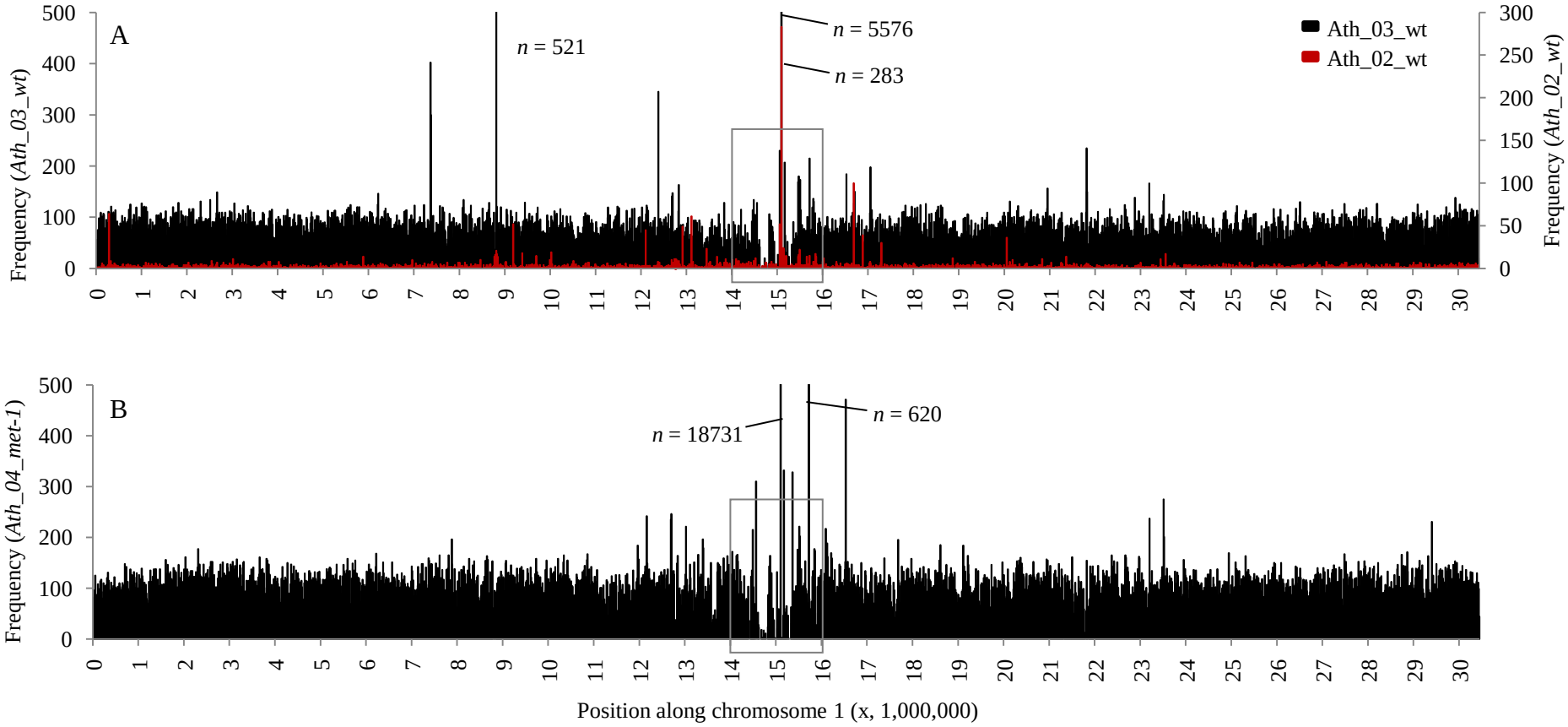

Figure 5.5 The density of nucleosome sequence fragment start co-ordinates within chromosome 1 for datasets A: *Ath_02_wt, Ath_03_wt* and B:*Ath_04_MET1,* plotted with a 10,000 bp bin size. For clarity, the *Ath_02_wt* dataset is plotted on a separate *y*-axis. The region in the grey box indicates some differences between the *Ath_03_wt* and *Ath_04_MET1* datasets, which are plotted at higher resolution. Figure 5.6.

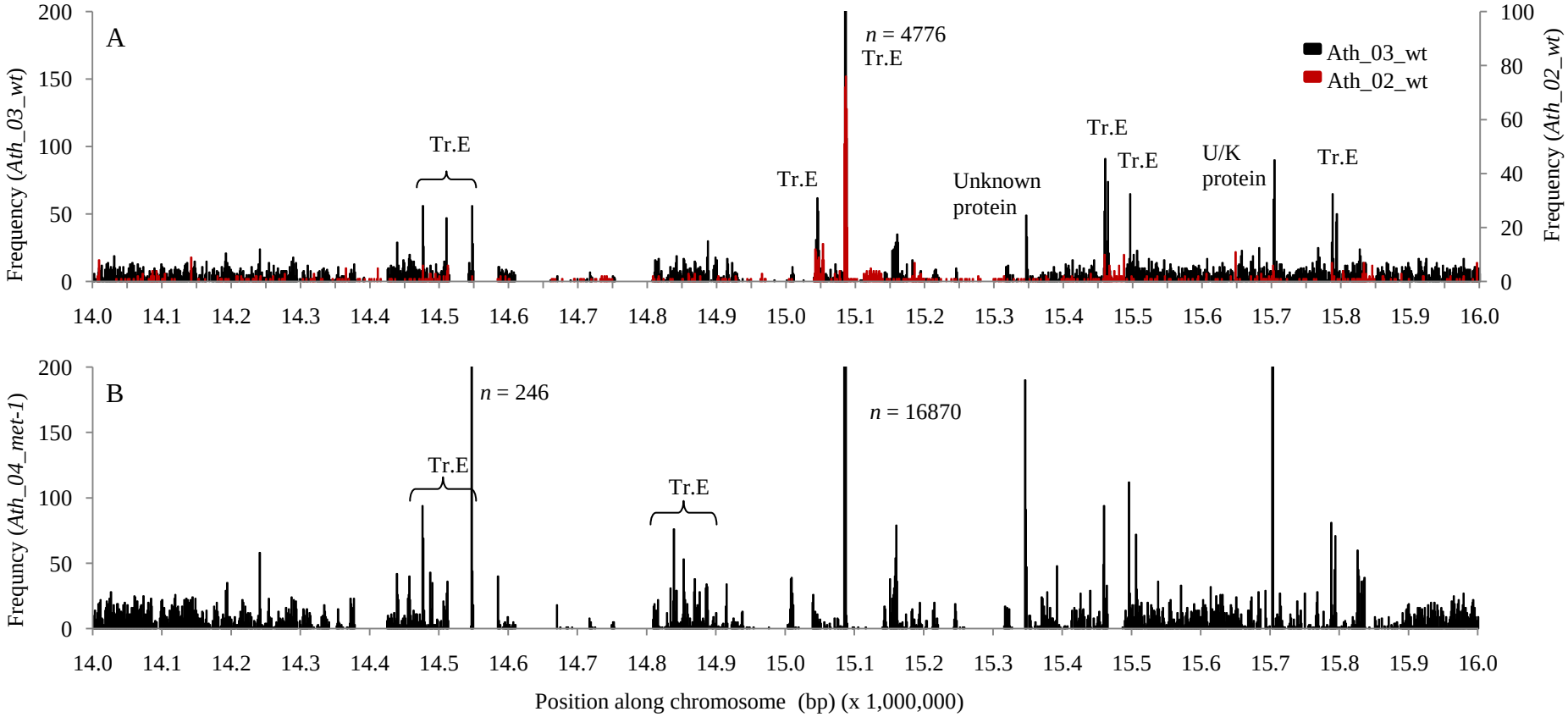

Figure 5.6 The density of nucleosome sequence fragment start co-ordinates within a close-up of a 2 Mbp region (boxed in Figure 5.5), within chromosome 1 for datasets A: *Ath_02_wt, Ath_03_wt* and B: *Ath_04_MET1*, plotted with a 1,000 bp bin size. For clarity, the *Ath_02_wt* dataset is shown on a separate *y*-axis. Where peaks are observed, the corresponding gene descriptions at those co-ordinates are shown.

5.3.2 Trends of nucleosome occupancy

Mapped nucleosome occupancies were viewed using a prototype genome browser, developed by Graham King, Rothamsted Research (data not shown). This revealed differences in nucleosome positioning between the *Ath_03_wt* and *Ath_04_MET1* datasets. Two variables affect these datasets: mononucleosome vs. dinucleosome and anti-sense *MET1 vs*. wild-type. Therefore, to understand and ascribe any observed differences to one or other variable, a further comparable wild-type mononucleosome sequence dataset was generated. However this was only available towards the very end of the project. The nucleosome DNA which was used to produce this dataset originated from the same sample (and same MNase digest) as the material from which the *Ath_03_wt* dataset was produced.

5.3.2.i Genomic trends of nucleosome occupancy

The nucleosome occupancy of datasets *Ath_03_wt*, *Ath_04_MET1* and *Ath_05_wt* was compared, and a few general trends in nucleosome positioning emerged. There are many cases for datasets *Ath_03_wt*, *Ath_04_MET1* and *Ath_05_wt* where the peak in nucleosome occupancy is high over a relatively long stretch of DNA (500 – 700 bp) (spread), indicating conserved, de-localised (or fuzzy) positioning. In contrast, high peaks in the nucleosome occupancy with a narrow spread occur, indicating conserved well-positioned nucleosomes. In addition, there are regions of lower occupancy throughout the genome and some regions which appear to be nucleosome-free. For the *Ath_03_wt* and *Ath_05_wt* datasets, the majority of nucleosome occupancy appears to be within the transcriptional units, with the intergenic regions generally having a lower occupancy, or being nucleosome-depleted. Overall, nucleosome positioning tends to be more regularly spaced in the *Ath_04_MET1* dataset, with increased occupancy within the intergenic region when compared to the wild-type.

5.3.2ii Nucleosome occupancy in specific chromosomal regions

To understand further some of the properties and potential functional consequences of differences in nucleosome occupancy observed between the datasets, two regions on chromosome 1 were chosen for further investigation: one, containing the imprinted gene *PHERES1* (co-ordinates 24,258,093 – 24,271,483 bp) and two, a

region containing a cluster of genes controlling floral development (co-ordinates 25,985,494 – 26,013,322 bp). The genes present in each region along with a brief description are shown in Tables 5.3 and 5.4.

Table 5.3 Gene models present in a selected region (1) on chromosome 1

Gene model	Gene name	Chr	TAIR 8 start co-ordinate	TAIR 8 stop co-ordinate	Strand	Description
AT1G65300	*PHERES 2* (*AGL38*)	1	24,258,592	24,259,428	F	MADS box TF
AT1G65310	*ATXTH17*	1	24,260,879	24,262,136	F	Xyloglucan endotransglucosylase/ hydrolase
AT1G65320	N/A	1	24,263,551	24,266,614	R	CBS-domain, unknown function
AT1G65330	*PHERES 1* (*AGL37*)	1	24,270,144	24,270,983	R	MADS box, imprinted, regulated by MEA and FIE

Table 5.4 Gene models present in a selected region (2) on chromosome 1

Gene model	Gene name	Chr	TAIR 8 start co-ordinate	TAIR 8 stop co-ordinate	Strand	Description
AT1G69120	*AP1*	1	25,985,993	25,989,976	R	MADS box, floral homeotic gene
AT1G69150	N/A	1	25,997,322	25,998,875	F	DC1 domain-containing protein
AT1G69160	N/A	1	26,003,877	26,005,251	F	Unknown protein
AT1G69170	*SPL6*	1	26,008,731	26,010,926	F	TF, petal differentiation and expansion
AT1G69180	*CRC*	1	26,011,128	26,012,722	R	Carpel and style development, floral meristem ID

The 5’ and 3’ ends of the genes within region 1 showed differences in the nucleosome occupancy between the three datasets (Figure 5.7). The red arrows indicate differences between the two wild-type datasets and the *Ath_04_MET1* dataset at the 5’ boundaries of *PHERES1* and *PHERES2*. Since nucleosome occupancy at the 5’ and 3’ regions is known to be important for transcriptional regulation, this suggests that these genes may be subject to different transcriptional

regulation in the *anti-sense MET1*. This region also demonstrates the presence of more regularly spaced nucleosomes within the *Ath_04_MET1* dataset. For example, in Figure 5.7 black arrows indicate nucleosome present at positions in the *Ath_04_MET1* dataset, which are absent from the other two datasets *Ath_03_wt* and *Ath_05_wt*. Regions of high occupancy with a wide spread, representing fuzzy positioning of nucleosomes (arrow 1) contrast with regions of high occupancy and narrow spread representing well-ordered positioning of nucleosomes (arrow 2). In addition, a nucleosome free region is observed between the *PHERES2* and AT1G65310 genes.

Differences in nucleosome occupancy within the 5' and 3' genic regions are also observed within region 2 (indicated by red arrows, Figure 5.8). However, there were also differences at the 5' and 3' ends between the *Ath_03_wt* and *Ath_05_wt* datasets in this example. The black arrows indicate occupancy in the *Ath_04_MET1* dataset which is not observed in the *Ath_03_wt* and *Ath_05_wt* datasets. The differences observed in this example all lie within the intergenic regions. In both the region 1 and 2 examples, the higher nucleosome occupancy peaks tend to be consistent with the positions of the exons. In addition, where defined, the 5' and 3' UTRs tend to have low nucleosome occupancy or are nucleosome-depleted.

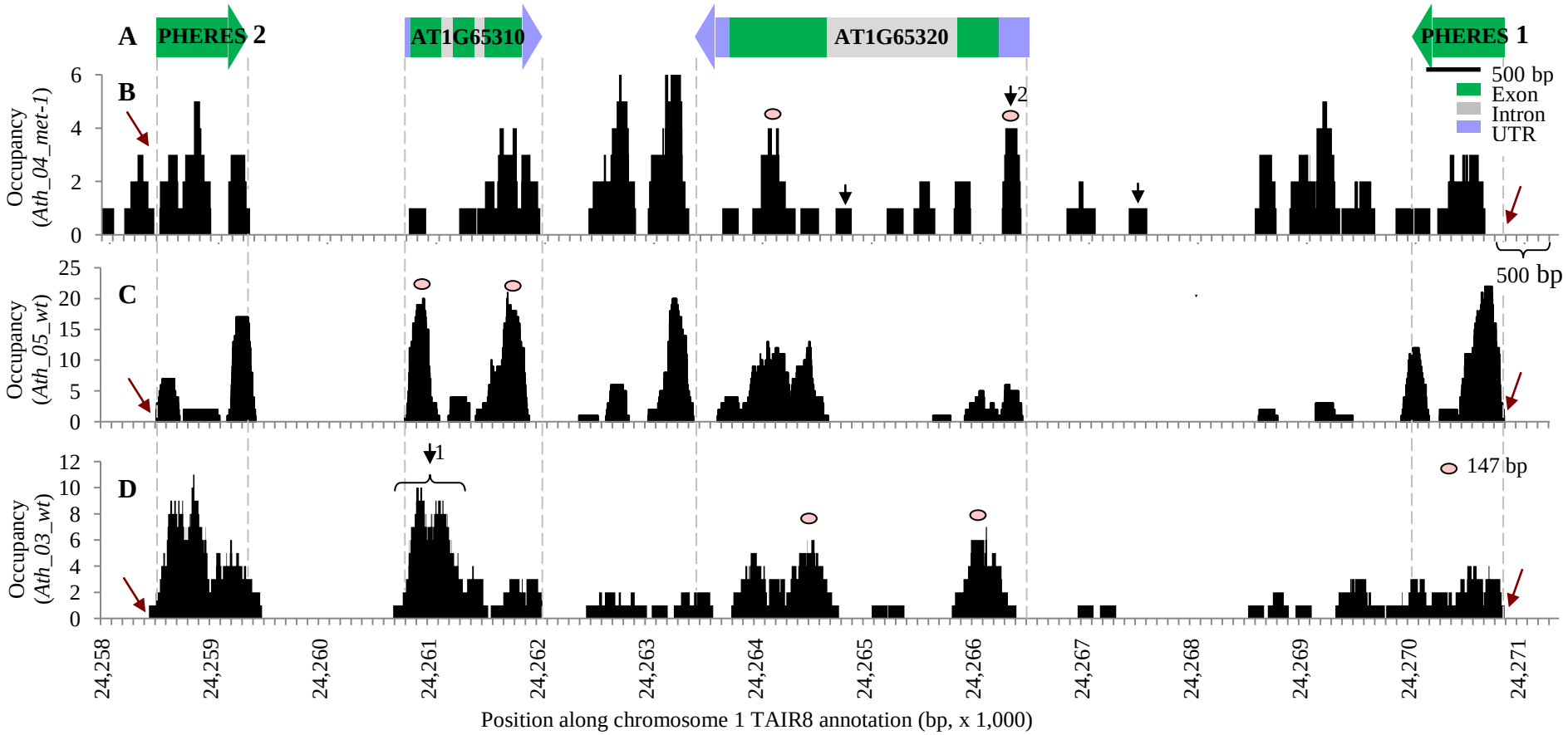

Figure 5.7 Nucleosome occupancy of datasets B: *Ath_04_MET1* C: *Ath_03_wt,* and D:*Ath_05_wt,* over a 14,500 bp region (region 1) within chromosome 1, containing the imprinted gene *PHERES1* (A). Red arrows indicate 5' and 3' regions where the nucleosome occupancy differs between datasets. Regions of occupancy observed in *Ath_04_MET1* which are not observed in the wild-type datasets are indicated with black arrows. For comparison, ellipses of nucleosome size are added.

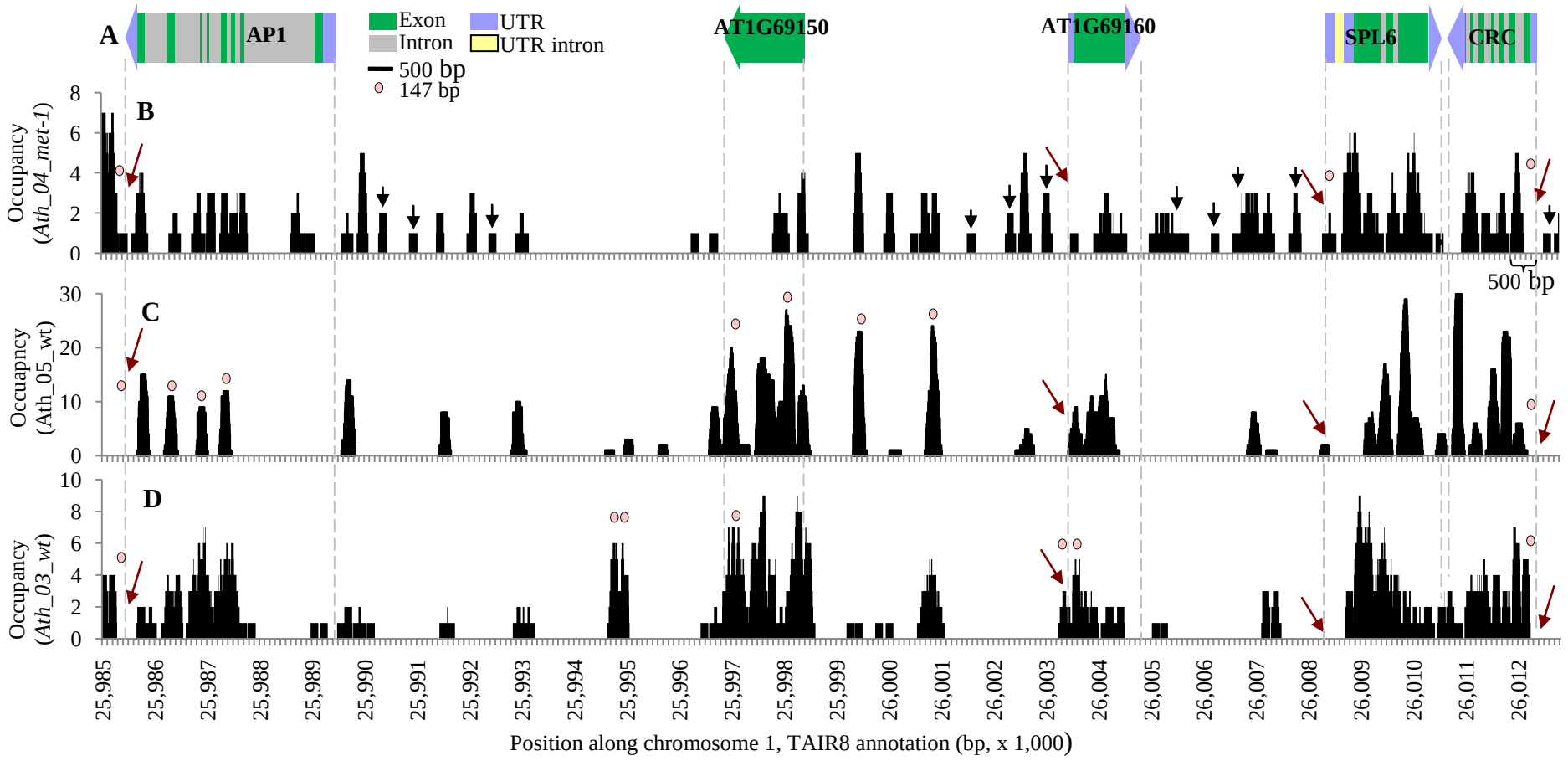

Figure 5.8 Nucleosome occupancy of datasets B: *Ath_04_MET1* C:*Ath_03_wt,* and D:*Ath_05_wt,* over a 28,000 bp region (region 2) within chromosome 1, containing floral homeotic genes (A). Red arrows indicate 5' and 3' regions where the nucleosome occupancy differs between datasets. Regions of occupancy observed in *Ath_04_MET1* which are not observed in the wild-type datasets are indicated with black arrows. For comparison, ellipses representing nucleosome size (147 bp) are added.

5.3.2.iii. Nucleosome occupancy within coding *vs.* non-coding regions

Based on these general and specific observations, a more systematic analysis was carried out. In order to determine whether there is a bias in nucleosome occupancy between introns and exons in Arabidopsis, the ratios of occupancy within genic:intergenic and exonic:intronic DNA were calculated for different gene annotation classes for the datasets *Ath_03_wt*, *Ath_04_MET1* and *Ath_05_wt* (Table 5.5). For comparison, the overall intron:exon ratios for each annotation class represented on Chr 1 were also calculated.

Table 5.5 Exon/intron ratios of nucleosome occupancy and average occupancy/base within different annotation classes on chromosome 1, for datasets the *Ath_03_wt, Ath_04_MET1* and *Ath_05_wt*. Overall exon/intron ratios and % of different gene annotation classes on Chr 1 are also shown. The average occupancy/bp and ratios of exon/intron are highlighted

Annotation class	*Ath_03_wt*		*Ath_04_MET1*		*Ath_05_wt*		Chr 1 %	Chr 1 ratio
	Ratio	*Occupancy*	*Ratio*	*Occupancy*	*Ratio*	*Occupancy*		
Genic	0.663	**3.88**	0.591	**4.61**	0.717	**7.02**	58.40	0.584
Intergenic	0.337	**1.97**	0.409	**3.20**	0.283	**2.77**	41.60	0.416
protein coding:								
Exon	0.718	**8.70**	0.685	**9.83**	0.710	**9.19**	32.4	0.673
Intron	0.282	**3.42**	0.315	**4.51**	0.290	**3.75**	15.8	0.327
5'UTR exon	0.551	3.56	0.543	4.41	0.673	5.22	3.87	0.847
5'UTR intron	0.449	2.89	0.457	3.72	0.327	2.54	0.700	0.153
3'UTR exon	0.561	3.57	0.526	4.22	0.699	5.29	3.43	0.835
3'UTR intron	0.439	2.79	0.474	3.81	0.301	2.28	0.676	0.165
Pseudogene:								
Exon	0.527	3.89	0.562	4.42	0.656	8.33	0.583	0.964
Intron	0.473	3.49	0.438	3.44	0.344	4.38	0.0310	0.036
misc RNA								
Exon	0.569	4.56	0.499	3.98	0.688	6.61	0.0839	0.667
Intron	0.431	3.46	0.501	3.99	0.312	3.00	0.0310	0.333
5'UTR exon	0.854	17.46	0.833	19.22	0.910	30.13	0.0855	0.305
5'UTR intron	0.146	2.98	0.167	3.86	0.090	3.00	0.0427	0.695
3'UTR exon	0.565	3.65	0.490	4.21	0.706	6.80	0.0652	0.661
3'UTR intron	0.435	2.81	0.510	4.38	0.294	2.83	0.0335	0.339
tRNA								
Exon	0.575	5.32	0.489	4.88	0.184	7.19	0.0006	0.337
Intron	0.425	3.93	0.511	5.10	0.816	31.81	0.0010	0.663

The genic:intergenic ratios for the *Ath_03_wt* and *Ath_05_wt* indicate that the genic regions of chromosome 1 are occupied by nucleosomes more often than the intergenic regions. The genic/intergenic ratio calculated for the *Ath_04_MET1*

dataset is lower than that for the *Ath_05_wt* for exons and is closer to the total genic/intergenic ratio, which suggests that the nucleosomes in the Arabidopsis *anti-sense MET1* are more evenly distributed than in the wild-type and that the average nucleosome occupancy/ bp in the *Ath_04_MET1* dataset is 3.2, compared to 2.7 in *Ath_05_wt*. In general, the exon:intron ratios indicate that exons are occupied by nucleosomes more often than introns in all datasets. The *Ath_04_MET1* exon:intron ratios, while indicating more occupancy of exons than introns, are nearer 0.5. This suggests that there is a more even distribution of nucleosomes between introns and exons. The exon/intron ratios for the dataset *Ath_03_wt* tend to be nearer those calculated for the *Ath_04_MET1* datasets.

5.3.3 Nucleosome positioning around transcriptional start sites

Previous studies in the small yeast genome have indicated a nucleosome free region approximately 150 bp in length, 200 bp upstream from the transcriptional unit (or within the promoter region) (Yuan *et al.*, 2005). In the absence of extensive promoter information for Arabidopsis, nucleosome presence at the transcriptional start sites was determined using a set of experimentally determined and predicted TSS (Tanaka *et al.*, 2008).

5.3.3.i Nucleosome presence at the transcriptional start site (TSS)

The distances between the TSS and nearest upstream nucleosome were calculated for both the forward and reverse strands of chromosome 1 for the datasets *Ath_03_wt* and *Ath_04_MET1* (Figure 5.9) and divided into three categories: those that are occupied, those which have the nearest nucleosome boundary within 200 bp upstream and those which have the nearest nucleosome boundary more than 200 bp upstream.

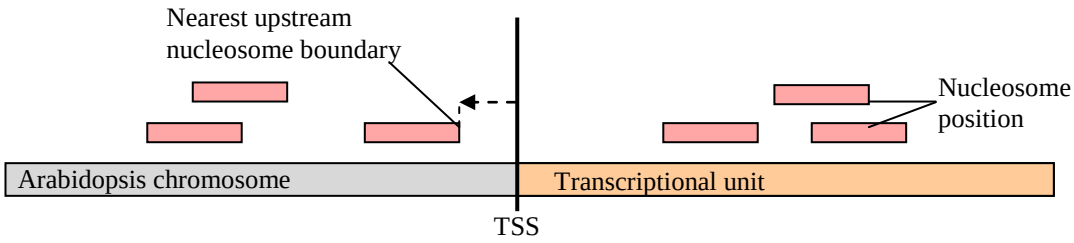

Figure 5.9 Schematic showing the method for determination of the distance between the TSS and the nearest upstream nucleosome boundary.

The percentage of TSS in each of the categories was calculated and is shown in Table 5.6. For both *Ath_03_wt* and *Ath_04_MET1*, 72 % of the TSS were occupied by at least one nucleosome DNA fragment. In addition, for ~17 % of the TSSs, the nearest nucleosome boundary was within the region 200 bp of upstream of the TSSs. For the remaining ~11 % of the TSSs, the distance between the nearest upstream nucleosome boundary and TSS was more than 200 bp. As many of the TSSs in this dataset are derive from a prediction algorithm, there are alternative TSSs for the same gene model within the dataset. The high occupancy of the TSSs may be due to alternative TSSs close to each other occupied by the same nucleosome fragment. In order to adjust for this, the categories were re-calculated using only one TSS for each

gene model (the furthest upstream in each case). This did not appear to alter the occupancy estimation, with 72 % TSS occupied over both the forward and reverse strands in the *Ath_03_wt* and *Ath_04_MET1* datasets. In addition, there was little difference in the relative numbers of occupied TSSs between the datasets *Ath_03_wt* and *Ath_04_MET1*.

Table 5.6 The presence of nucleosomes at, and upstream of TSSs within chromosome 1 for the datasets: *Ath_03_wt* and *Ath_04_MET1*.

	Ath_03_wt			*Ath_04_MET1*		
Category	Forward	Reverse	All	Forward	Reverse	All
all TSS						
Occupied	72.77	72.05	72.40	72.77	72.15	72.46
0 to -200 bp	15.85	17.92	16.71	15.77	17.37	16.58
-200 + bp	11.38	10.64	10.89	11.46	10.48	10.96
1 TSS only						
Occupied	70.15	73.08	71.62	70.15	73.14	71.65
0 to -200 bp	16.64	16.42	16.53	16.57	16.36	16.47
-200 + bp	13.21	10.50	11.85	13.27	10.50	11.89

5.3.3.ii Nucleosome occupancy around transcriptional start sites

In order to determine the nucleosome occupancy around the TSS, the midpoint of each nucleosome in the dataset *Ath_05_wt* was calculated. For all TSS within the first 6,000,000 bp on the forward strand of Chr 1, the distance between each nucleosome midpoint and each TSS was calculated (Figure 5.10).

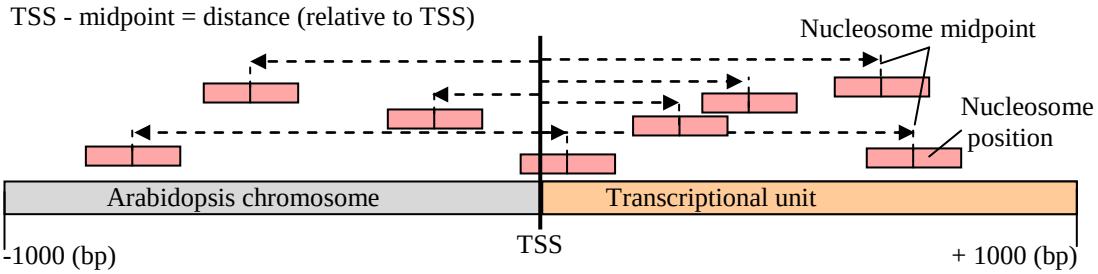

Figure 5.10 Schematic showing the method for calculating the frequency of nucleosome occupancy around the TSSs by subtracting the position of the TSS from the nucleosome midpoint.

For each TSS, the distance between the TSS and every nucleosome midpoint within 1000 bp of the TSS was plotted against the position on Chr 1, to produce a density map (Figure 5.11.A). There appear to be regions within Chr 1 where the nucleosome

occupancy is denser, specifically in the vicinity of co-ordinates ~2 Mbp, ~3 Mbp, 4-4.5 Mbp and 5-6 Mbp. The regions of greater density appear to occur in the region to the right (immediately downstream) of the TSS. In addition, there are less dense regions of nucleosome occupancy around 1.5 Mbp along the chromosome. The differences in nucleosome density could be due to differences in gene density. To test this, the density of the TSSs within 0-6 Mbp Chr 1 (10,000 bp bin size) is shown in Figure 5.11.B. The nucleosome depleted region around 1.5 Mbp does appear to coincide with a lack of TSS in this area. In comparison, the nucleosome-dense areas around 2 Mbp, 4-3 Mbp and 5-6 Mbp appear to co-incide with an increase in TSS frequency within these regions. In order to calculate the estimated occupancy of the TSS, a subset of the distances between each TSS and all nucleosomes within 73 bp was produced. Of all the TSS in the region, 69.71 % were occupied by nucleosomes. The frequency of nucleosome occupancy relative to the position of the TSS on Chr 1 is shown in Figure 5.11.C.

To test the hypothesis that specific patterns of occupancy exist around the TSS, the frequency of nucleosome midpoints for each position around the TSS was plotted (Figure 5.11.D). A peak in the distribution is observed 25-72 bp after (downstream) from the TSS. Nucleosome occupancy also appears to be greater downstream of the TSS compared with upstream, indicating that the region upstream may be nucleosome depleted or relatively nucleosome free, which is consistent with previous studies (Mavrich *et al.,* 2008a;b; Yuan *et al.,* 2005). In addition, there appears to be some periodicity in the distribution, which may indicate ordered positioning around the TSS. A periodicity of 10.05 bp (CI 9.99-10.10 bp) was detected within the distribution by Fourier analysis. However this is likely to be due to periodicity with the distribution of alternate TSS. Kernel Density Estimation was used to reveal longer-range peaks within the distribution (Figure 5.12).

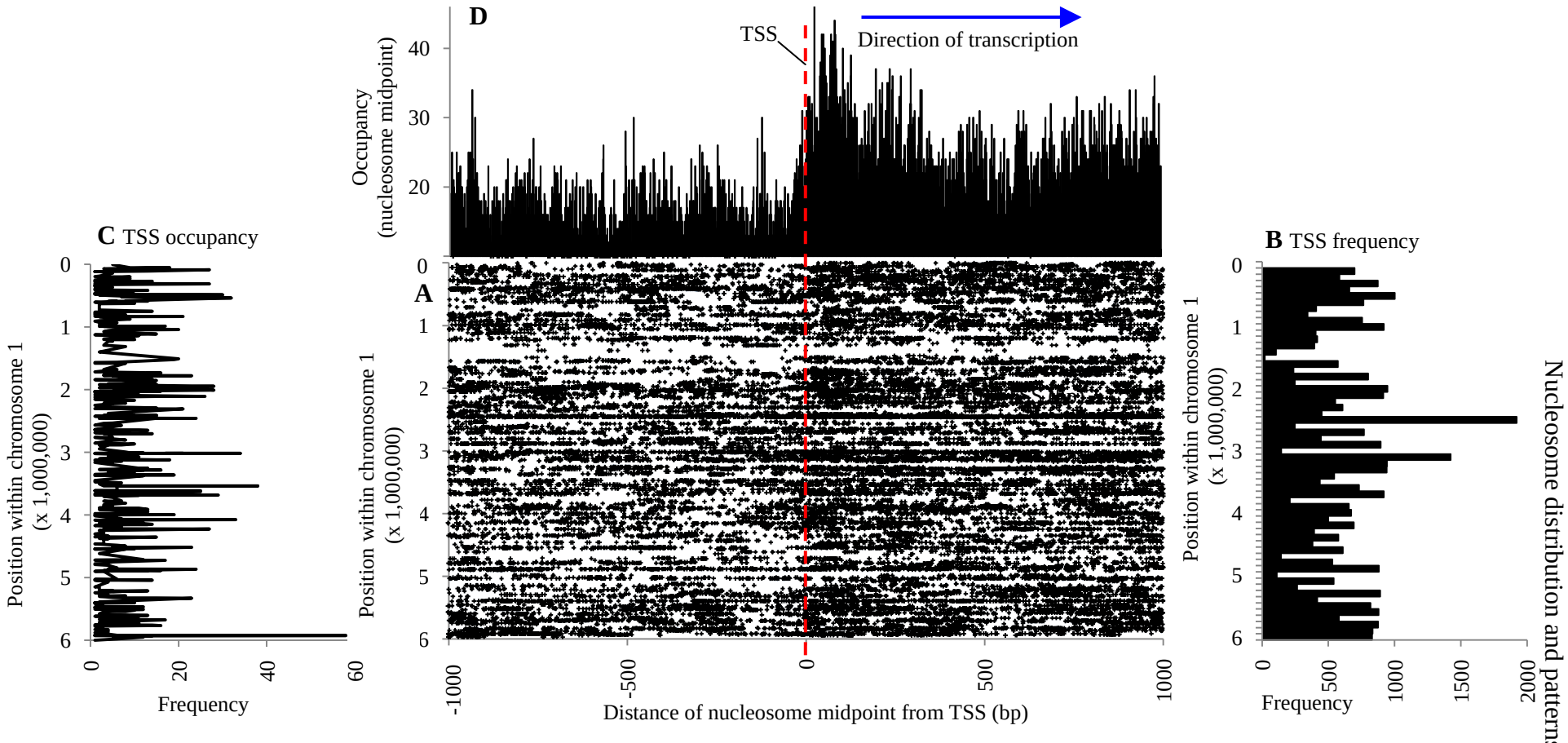

Figure 5.11 Nucleosome occupancy around the TSS of the first 6 Mbp of the forward strand of chromosome 1. **A**: The distances of all nucleosomes within 1000 bp either side of the TSS relative to the position on Chr 1, **B**: the nucleosome occupancy of the TSS relative to the position within Chr 1, **C**: TSS density (relative to the position on Chr 1, and **D**: frequency of nucleosome occupancy at each position within 1000 bp of the TSS.

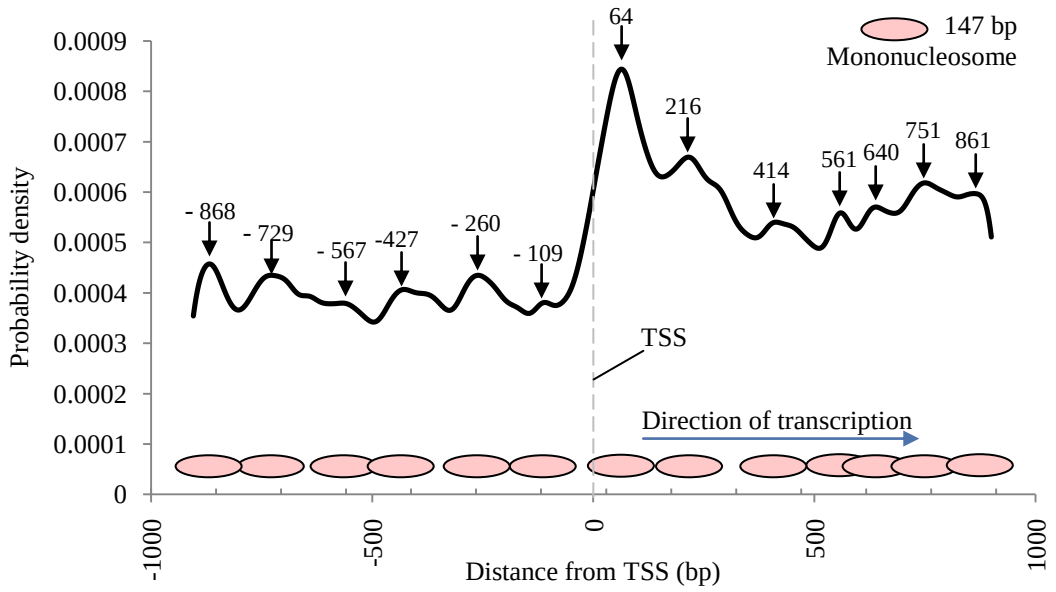

Figure 5.12 Kernel Density Estimation of nucleosome occupancy around the TSSs within the first 6 Kb of chromosome 1 (forward strand). The distance (bp) between the TSS and important peaks in the nucleosome occupancy are indicated. Internal *x*-axis markers are spaced at 147 bp and ellipses representing nucleosome size are added for comparison.

Smoothing of the distribution revealed a peak in the nucleosome occupancy at +64 bp with respect to the TSS. This is consistent with ~70 % of the TSS being occupied by nucleosomes. Further peaks were observed in the smoothed distribution of nucleosome occupancy, both upstream and downstream of the TSS, which may indicate ordered positioning around the TSS. To investigate this further, the distance between the apex of each peak was determined and the mean distance was calculated to be 144.1 bp ± 9.2 bp with a 95 % CI of 123.9 - 164.4 bp for a sample of 12 distances. The calculation was repeated, but omitting the distances between peaks which occur after +500 bp with respect to the TSS, since these peaks appear to correspond to overlapping nucleosome positions within this region as such they may correspond to the end of ordered positioning (or correspond to another, overlapping signal ~+ 500 bp from the TSS). The mean for this sub set is 158.7 bp ± 6.3 bp with a 95% CI of 144.3 – 173.2 bp, for a sample of 9 distances.

5.3.4 Nucleosome occupancy around the Arabidopsis nucleolar organiser region

The regions containing rRNA genes annotated on chromosomes 2 and 3 have different biases in linker length, indicating differences in chromatin structure between the two regions (Chapter 4, section 4.3.4). Again it should be noted that the sequence of Arabidopsis NORs on Chrs 2 and 4 are incomplete in the current reference genome. The sequence of a single repeat on Chr 3 probably represents the 700-900 copies known to exist in the Chr 4 NOR. In order to test the hypothesis that differences exist in the nucleosome positioning and occupancy, which may affect the chromatin structure of the Arabidopsis NOR regions, the nucleosome occupancy from datasets *Ath_03_wt*, *Ath_04_MET1* and *Ath_05_wt* around the homologous rRNA genes were aligned, and are shown in Figure 5.13.

There appear to be differences in the positioning and occupancy of the rRNA genes in all datasets. Most notably, a peak in the nucleosome occupancy around the 3' boundary of the 25S rRNA gene is observed within Chr 2, but is absent from the corresponding position on Chr 3 in all datasets, (indicated on Figure 5.13 by arrows 1,4 and 6). The nucleosome positioning over the 18S rRNA gene appears to be well-ordered and conserved between Chr 2 and 3 for the dataset *Ath_05_wt*. However, there are interesting differences between Chr 2 and 3 and the 5' and 3' boundaries of the 18S rRNA gene. A peak in the nucleosome occupancy within Chr 2 is observed approximately 150 bp upstream of the 5' boundary of the 18S rRNA gene, which appears to be absent within Chr 3 (indicated by arrows 2, Figure 5.13). In addition, a further peak in the nucleosome occupancy is observed which appears to span the 3' edge of the 18S rRNA gene on Chr 3, which is absent from Chr 2 (indicated by arrows 3, Figure 5.13). Differences are also observed in the nucleosome positioning between Chr 2 and 3 in the *Ath_03_wt* dataset, over a region which spans the 5' boundary of the 18S rRNA gene and approximately 350 bp upstream. The occupancy of dinucleosomes at the 5' boundary of the 18S rRNA gene on Chr 2 (n = ~400 molecules) appears to be approximately half of that on Chr 3. However, the positioning on Chr 2 appears to be well-ordered with two well-defined peaks representing the two nucleosomes (indicated by arrows 5, Figure 5.13). The well-defined peaks are not observed in the corresponding position on Chr 3.

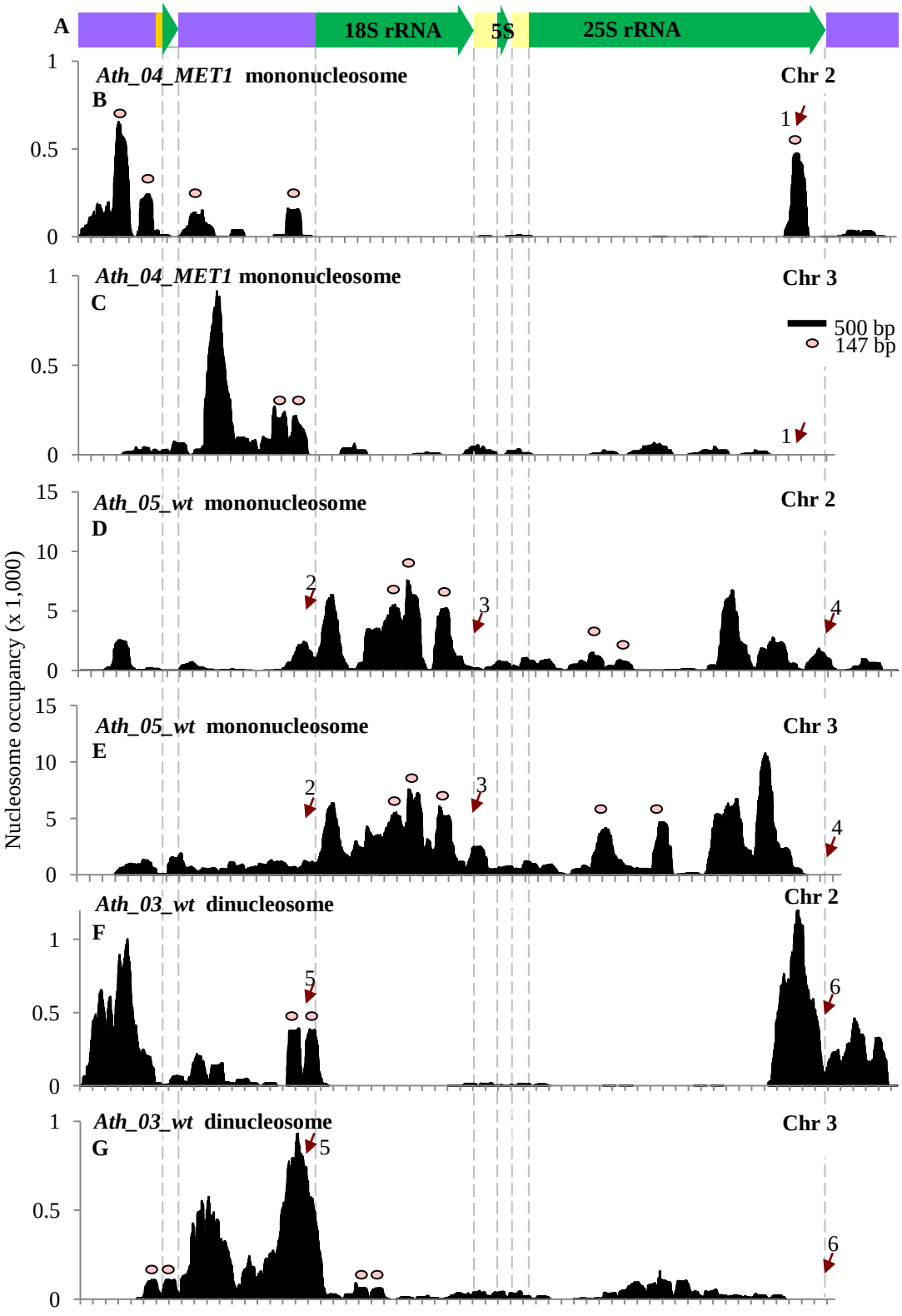

Figure 5.13 Alignment of (A) rRNA genes on chromosomes 2 and 3 and nucleosome occupancy of datasets *Ath_04_MET1* (B and C), *Ath_05_wt*,(D and E), and *Ath_03_wt* (F and G). The relative positions of the rRNA genes, ITS and IGE are shown in track A. For size comparisons, ellipses represent mononucleosome size (147 bp) and *x* axis tick marks are spaced at 147 intervals. Arrows 1, 4 and 6 indicate differences in occupancy at 3' end of 25S rRNA gene between chromosomes 2 and 3, and arrows 2, 3 and 5 indicate differences in occupancy around the 18S rRNA gene.

In addition to differences in nucleosome positioning and occupancy over rRNA regions between chromosomes 2 and 3, differences are also observed between the datasets *Ath_03_wt*, *Ath_04_MET1* and *Ath_05_wt* over the rRNA region within Chr 2 and Chr 3. To enable ready comparisons between datasets over the rRNA region on the same chromosome, the nucleosome occupancies of datasets *Ath_03_wt*, *Ath_04_MET1* and *Ath_05_wt* were re-plotted against either Chr 2 or Chr 3 (Figures 5.14 and 5.15, respectively).

The most apparent differences between the datasets on Chr 2 are within the occupancy of nucleosomes from the dataset *Ath_05_wt* compared to the other two datasets. The observed nucleosome occupancy around the 18S rRNA gene was much higher in the *Ath_05_wt* dataset, than the other two datasets. The occupancy of the 5S and 5' end of the 25S rRNA genes is also relatively higher in the *Ath_05_wt* dataset than the *Ath_03_wt* or *Ath_04_MET1* datasets. The occupancy over the three rRNA gene regions is either relatively low or absent in the *Ath_03_wt* and *Ath_04_MET1* datasets, giving the impression of nucleosome-free regions over these genes. In addition, higher peaks in nucleosome occupancy at the 3' end of the 25S rRNA gene in the *Ath_03_wt* and *Ath_04_MET1* datasets appear to correspond to a low occupancy region in the *Ath_05_wt* dataset.

A similar pattern of nucleosome occupancy within the 18S and 25S rRNA is also observed in Chr 3 (Figure 5.15). The *Ath_05_wt* dataset shows well-ordered high occupancy compared with the much lower occupancy for *Ath_03_wt* and *Ath_04_MET1* datasets over the same gene. The datasets *Ath_03_wt* and *Ath_04_MET1* appear to show more similarities than either the two mononucleosome datasets (*Ath_05_wt* and *Ath_04_MET1*) or the two wild-type datasets (*Ath_03_wt* and *Ath_05_wt*).

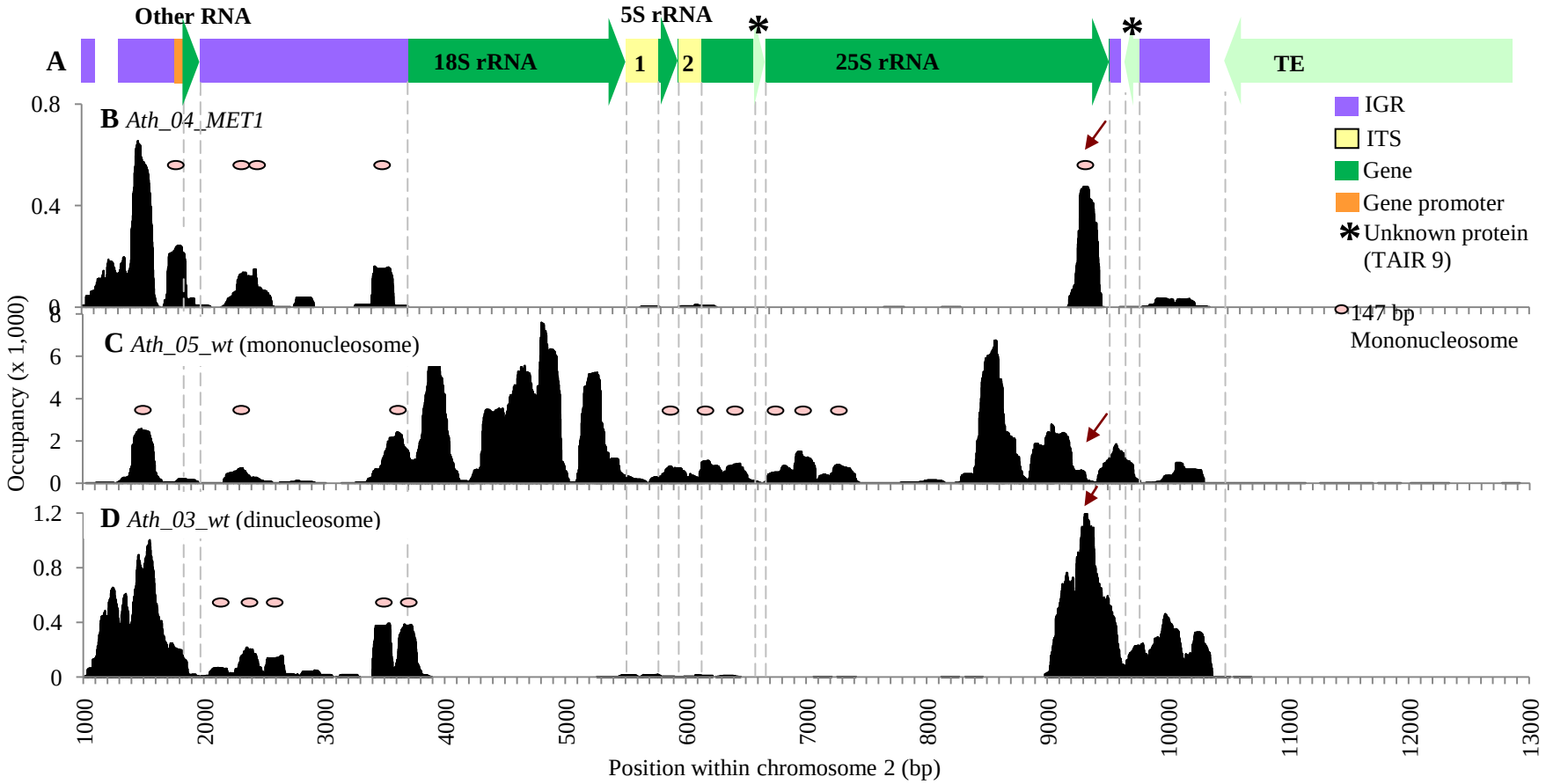

Figure 5.14 Nucleosome occupancy of the rRNA regions on chromosome 2 for datasets B: *Ath_04_MET1*, C: *Ath_05_wt* and D: *Ath_03_wt.* The positions of the rRNA genes, ITS and IGR are indicated in track A. For size comparisons, ellipses represent mononucleosome size (147 bp). The arrows indicate similarities in occupancy between *Ath_03_wt* and *Ath_04_MET1* dataset not observed in the *Ath_05_wt* dataset.

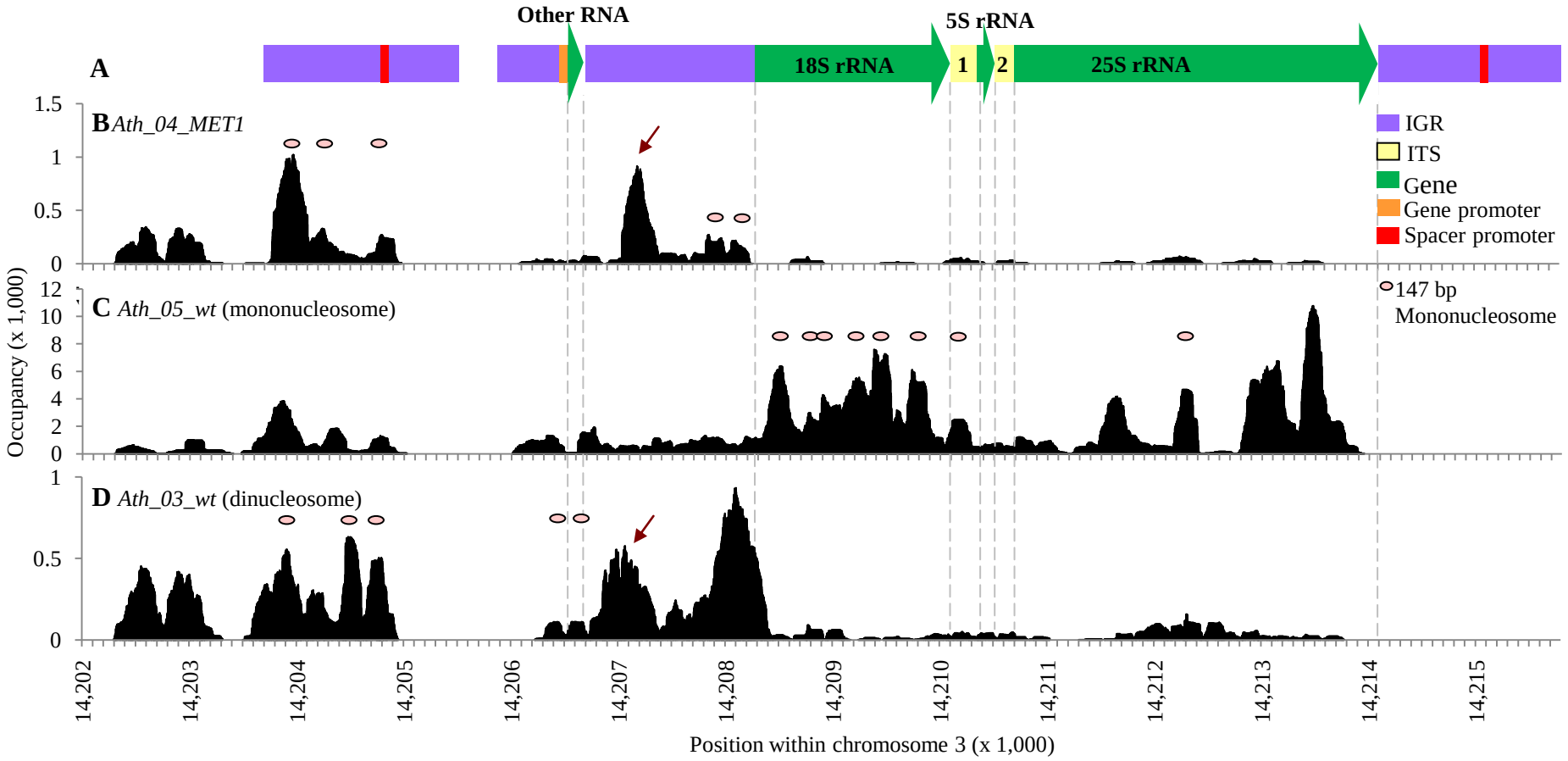

Figure 5.15 Nucleosome occupancy of the rRNA regions on chromosome 3 for datasets B: *Ath_04_MET1*, C: *Ath_05_wt* and D: *Ath_03_wt*. The positions of the rRNA genes, ITS and IGR are indicated in track A. For size comparisons, ellipses represent mononucleosome size (147 bp). The arrows indicate similarities in occupancy between *Ath_03_wt* and *Ath_04_MET1* dataset not observed in the *Ath_05_wt* dataset.

5.4 Discussion

5.4.1 Genomic trends of nucleosome distribution

Nucleosomes were found to be distribution throughout the Arabidopsis genome, in a non-uniform manner. and with particular regions of high occupancy. Of particular interest was the positioning within the rRNA regions annotated on chromosomes 2 and 3. These observations are consistent with previous observations in yeast (Yuan *et al.,* 2005) and *C. elegans* chromatin (Johnson *et al.*, 2007). It is not surprising that peaks are observed around rRNA, as these regions consist of highly repeated sequences, some of which may not be represented by the correct copy number within the current Arabidopsis annotation. In addition, transposable elements are known to be highly methylated (epigenetically regulated), and also contain repetitive sequences (Lippman *et al.,* 2004). It should be noted that the datasets *Ath_03_wt*, *Ath_04_MET1* and *Ath_05_wt* were filtered for multiple hits within the genome (for the 32 bp and 76 bp tags) prior to these analyses.

Genic regions were found to be occupied more often than intergenic regions in wild-type Arabidopsis chromatin as can be seen within the two examples of nucleosome occupancy from chromosome 1 (Figures 5.7 and 5.8). This observation of nucleosome occupancy is also consistent with previous observations in yeast (Lee *et al.,* 2007). Nucleosomes were found to be present in 87 % of the transcribed sequences in yeast, in contrast to 53 % of the intergenic sequences (Lee *et al.*, 2007). This could be due to the positioning of nucleosomes, with the intergenic regions being less well defined than positioning over specific transcribed sequences. This in turn may lead to conserved nucleosome positions within the transcribed regions being sampled more often than those within intergenic regions. Alternatively, the chromatin structure (influenced by epigenetic marks) may be different within the transcribed region compared to the intergenic, which may render the chromatin either more resistant or more susceptible to digestion by MNase. The differences observed within this study between the wild-type mononucleosome and dinucleosome datasets, particularly within the rRNA region, certainly seem to support the hypothesis that different classes of nucleosome exist, which also appear to have different susceptibilities to enzymatic digestion.

Previous studies in yeast have demonstrated both de-localised and well-defined nucleosome positions (Yuan *et al.,* 2005). The de-localised nucleosome positions were associated more often with highly transcribed genes, and as a result were attributed to the disassembling and re-assembling of nucleosomes during the passage of RNA polymerase II through the coding region. In the current project for Arabidopsis, regions of high occupancy with a wide spread, indicating de-localised (or 'fuzzy') positioning were observed here in Arabidopsis. In addition, nucleosome peaks of high occupancy with narrow spread, indicating well-ordered positioning, were also observed. Within the two examples of nucleosome occupancy on Arabidopsis chromosome 1, both the de-localised and well-ordered classes of nucleosomes are observed. However, the de-localised nucleosome positions appear mostly to be confined to the transcriptional units in both *Ath_03_wt* and *Ath_05_wt* datasets, which could be due to the passage of Pol II within these genes (Schwabish and Strul, 2004). Well-defined nucleosome positions are observed both within the transcriptional units and within the intergenic regions.

The *Ath_04_MET1* dataset appeared to have more instances of de-localised nucleosome positions within the intergenic regions, possibly associated with the loss of ^{5m}CG epigenetic marks. Furthermore, these tend to be located in a region with low occupancy or well-defined nucleosome positions in the corresponding locations in the wild-type datasets. This suggests a de-regulation of positioned nucleosomes in the *anti-sense MET1* within genic regions, resulting in higher occupancy in the intergenic regions.

5.4.2 Nucleosome occupancy in 5' region of *PHERES1* and *PHERES2*

Differences were observed in the positioning of nucleosomes at 5' and 3' ends of genes between the wild-type *Ath_03_wt* and *Ath_05_wt* datasets, and also in the *Ath_04_MET1* dataset within the examples shown for chromosome 1. The *PHERES1* and *PHERES2* genes display particularly interesting behaviour with regard to nucleosome occupancy. In wild-type plants, *PHERES1* is imprinted on the maternal allele in the developing embryo following fertilisation. Silencing of this allele is brought about by methylation of H3K27 by FIS, a polycomb-group complex of proteins comprising of *MEDEA, FIE* and *FIS2* (Köhler *et al,* 2006). The paternal

PHERES1 allele is silenced following fertilisation, by DNA methylation. In *anti-sense MET1*, the paternal *PHERES1* allele is expressed in the embryo and endosperm, whereas the maternal allele remains repressed (Köhler and Makarevich, 2006). The wild-type datasets *Ath_03_wt* and *Ath_05_wt* show nucleosome occupancy at the TSS and within the 5' region of *PHERES1*. The *Ath_04_MET1* dataset shows a loss of occupancy around the TSS of *PHERES1*. The *PHERES2* gene is thought to have a similar temporal pattern of transcription to *PHERES1* (Villar *et al*, 2009). The *Ath_04_MET1* dataset showed increased occupancy of *PHERES2* between the TSS and 300 bp upstream of the TSS, which was not observed in the *Ath_03_wt* and *Ath_05_wt* datasets. These observations suggest that differences may exist in transcriptional regulation within the leaves of Arabidopsis *anti-sense MET1* as nucleosome occupancy within the 5' region upstream of the TSS is thought to be important for transcriptional regulation (Yuan *et al.,* 2005). To test this hypothesis, the transcript levels of *PHERES1* and *PHERES2* would need to be measured in the same material from which the chromatin was extracted.

5.4.3 Nucleosome occupancy within exons and introns

The exon/intron ratio of nucleosome occupancy within chromosome 1 was found to be altered in the *Ath_04_MET1* datasets compared with the wild-type. Calculations of the ratio of exons and introns for various annotation classes revealed that for protein-coding regions, nucleosome occupancy is skewed towards exonic DNA in protein coding, 5' UTR and 3' UTR regions. Occupancy has also been shown to be skewed in the yeast genome (Lee *et al.,* 2007). The *Ath_04_MET1* dataset ratio was nearer to that of the exon/intron ratio of the whole chromosome. This suggests that nucleosomes are positioned more often in the exons than introns in wild-type Arabidopsis, and that exon-specific positioning may be compromised in the Arabidopsis *anti-sense MET1*. In wild-type Arabidopsis with normal 5'me-C levels, one third of the transcribed genes are methylated within the gene body (Zhang *et al.,* 2006). It may be that the gene body DNA methylation either directly or indirectly guides nucleosome positioning towards exons. Methylation within the gene body occurs almost exclusively in the [CG] context. Gene body methylation has been shown to be higher within in the transcribed regions when compared to the upstream and downstream regions and be skewed towards the 3' end of the transcribed region (Cokus *et al.,* 2008). In addition, the profiles of methylation within the transcribed

regions were reported for a number of single, double and triple methylation mutants. This demonstrated a complete loss of [CG] methylation in the *met-1* mutant. However, when *met-1, drm1* and *drm2* are all missing, [CG] methylation within the transcribed region appears to be compensated by *CMT3* with CHG methylation which occurs in a similar distribution profile to the wild-type [CG] profile (Cokus *et al.*, 2008).

5.4.4 Nucleosome occupancy at the transcriptional start site

Previous studies in yeast have demonstrated a nucleosome-free region of 150 bp in length, approximately 200 bp upstream from annotated genes (Yuan *et al.*, 2005). The nucleosome occupancy of the first 6 Mbp of Arabidopsis Chr 1 was investigated in the *Ath_05_wt* dataset, and a nucleosome-free region of this size (150 bp) was not observed in this subset. However, the nucleosome occupancy was lower within the upstream (-1000 to 0 bp, relative to the TSS) region than in the downstream region. This is consistent with a further study in yeast (Lee *et al.*, 2007) and more recently in *Drosophila* (Mavrich *et al.*, 2008a). A clear peak in nucleosome occupancy was observed in Arabidopsis around +25 to +70 bp relative to the TSS, with the corresponding peak in the smoothed distribution at +64 bp. As the nucleosome distances were measured from the midpoint, this would place the TSS within the first 10 bp of the nucleosome. This is in contrast to the positioning around the TSS in yeast and *Drosophila* (Mavrich *et al.*, 2008a). For comparison, the Arabidopsis nucleosome occupancy around the TSS with the smoothed distribution overlaid is aligned with the occupancy for yeast and *Drosophila* (Figure 5.16).

While the position of the first peak downstream of the TSS (the +1 nucleosome) differs between all three taxa, the position of +2 and +3 peaks in the Arabidopsis distribution appear to coincide with +2 in yeast and +2 in *Drosophila*, respectively. The position of the -1 nucleosome was found to be conserved in *Drosophila* and yeast, but does not appear to be conserved in Arabidopsis, with the first main peak being further upstream than in yeast or *Drosophila*. However, well ordered nucleosome positioning is observed both upstream and downstream of the TSS in all distributions, suggesting that while the position of the occupancy differs between taxa, well-ordered, long-range nucleosome positioning around the TSS is present in all three and is likely to contribute to gene regulation.

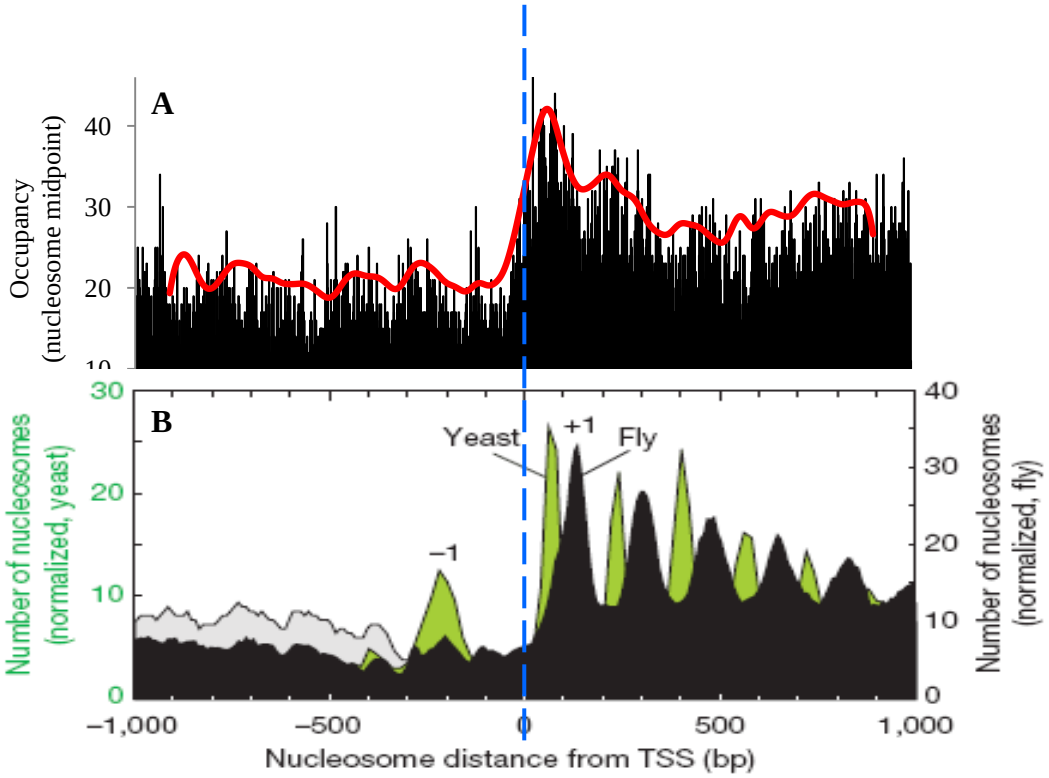

Figure 5.16 The nucleosome occupancy around the TSSs of Arabidopsis, yeast and Drosophila. **A**: the occupancy around the TSS of the first 6 Mbp in Arabidopsis chromosome 1, with the smoothed distribution overlaid. **B**: the H2A.Z-containing nucleosome positioning around the TSS of yeast and Drosophila. Image from Mavrich *et al.*, (2008). The blue dashed line indicates the position of the TSS.

A study of the H2A.Z histone variant positions in yeast position the TSS within the first helical turn of the nucleosome (Albert *et al.*, 2007). This is consistent with the peak in occupancy in Arabidopsis between +63 and +70 bp, relative to the TSS.

5.4.5 Nucleosome occupancy with the Arabidopsis nucleolar organiser regions

The rRNA genes in the NORs of Arabidopsis display well-conserved non-random positioning. The annotation of single copies of these genes account for ~700-900 copies within each NOR (assuming the single repeat on Chr 3 represents the NOR on Chr 4). High occupancy around the NOR was also observed in *C. elegans* chromatin (Johnson *et al.*, 2007). Differences between the two annotated NORs are observed in all datasets *Ath_03_wt*, *Ath_04_MET1* and *Ath_05_wt*. These differences may represent variation in regulation over the two NORs, particularly with the peaks observed in the *Ath_05_wt* dataset around the 18S rRNA gene.

A peak at 100-150 bp upstream of the 5' edge was observed on Chr 2 but not on Chr 3, while a peak at the 3' end was observed on Chr 3 but not on Chr 2. In addition, the Ath_03_wt dataset has a well-defined dinucleosome at the 5' end of the 18S rRNA gene on Chr 2, and a de-localised peak over the same region on Chr 3. The de-localised (or fuzzy) nucleosome positions are thought to correspond with highly expressed genes (Yuan *et al.*, 2005). Whilst a complete NOR is not thought to be silenced or active (Preuss and Pickaard, 2007), these results suggest that one NOR may be relatively more active and the other may be more silenced. The level of occupancy within the annotated rRNA regions is 1,000 larger than other regions of the genome, which is consistent with conserved positioning across the 700-900 copies of the rRNA region in each NOR.

The differences observed over the rRNA regions between the *Ath_03_wt* and *Ath_05_wt* datasets were unexpected. The region corresponding to the position of the genes on both chromosomes appear to be occupied by mononucleosomes, while dinucleosomes tended to occupy the IGRs. This could indicate differences in susceptibility to enzymatic digestion resulting in different classes of nucleosomes being present in the bands of a nucleosome ladder. The two wild-type datasets *Ath_03_wt* and *Ath_05_wt* were produced from the same material, in the same chromatin digest and cut from the same gel. The known difference between the samples was the size of the fragment (and that the populations of fragments were sequenced on separate occasions). Had the *Ath_05_wt* dataset not been sequenced, it might have been concluded that the 18S rRNA gene was a nucleosome-free region, as there was far less occupancy over the gene in this dataset than in the other two datasets. This is an important issue as many conclusions made regarding the positioning of nucleosomes relative to annotated positions are based on one type of molecule (many mononucleosomes). It may be the case that the fragment size resulting from chromatin digestion gives insight into the high-order chromatin structure.

5.4.6. Summary

The distribution of nucleosomes within chromosome 1 was relatively even, with occasional peaks observed around transposable elements, centromeres and rRNA regions. Genic regions were found to be occupied more often than intergenic

regions, which is consistent with other studies. However, for the *Ath_04_MET1* dataset, the ratio of genic to intergenic regions occupied by nucleosomes is closer to the genomic ratio for genic to intergenic regions. This indicates a more even distribution of nucleosome positions. Differences in the nucleosome occupancy around the 5' ends of the *PHERES* genes are observed between wild-type and *anti-sense MET1* chromatin, suggesting alternative regulation of these genes in the *anti-sense MET1*. Nucleosomes exhibit specific patterns of occupancy within introns and exons, and around the TSS. Differences in nucleosome occupancy around the TSS were observed between Arabidopsis and other taxa, with well-ordered positioning a feature in all taxa. Within the NORs, subtle differences are observed in the nucleosome occupancy between chromosomes in each dataset, indicating differences in regulation of the rRNA genes. However, more surprising were the differences observed between nucleosome classes, suggesting that chromatin regions have different susceptibilities to enzymatic digestion.

Chapter 6

Effect of DNA methylation on nucleosome position

6.1 Introduction

DNA methylation occurs at [CG] nucleotides in mammals, and [CG], [CNG] and [CNN] in plants. DNA methylation through *MET1* occurs in the context of [CG]. The DNA pericentromeric regions, DNA repeats and transposable elements are all highly methylated in Arabidopsis in all sequence contexts (Cokus *et al.,* 2008). In addition, relatively high levels of DNA methylation have been found within the gene body of over one third of expressed genes in Arabidopsis, with 5% methylated within the promoter (Cokus *et al.,* 2008). Furthermore, genes with methylation in the gene body tend to be constitutively expressed, while genes with promoter methylation tend to be developmentally controlled, indicating that transcriptional control by DNA methylation is position-dependent (Zhang *et al.,* 2006). Arabidopsis DNA methyltransferase anti-sense *MET1* plants have developmental abnormalities, including decreased size, decreased apical dominance, abnormal leaf morphology, abnormal flower morphology and decreased fertility (Finnegan *et al.,* 1996). However, the effect of loss of [CG] methylation from transposable elements is thought to be minimal as *MET1* and *CMT3* have been shown to act redundantly to repress the mobility of *CACTA* elements (Kato *et al.,* 2003).

Experiments involving whole-genome bisulphite sequencing have revealed patterns of DNA methylation in Arabidopsis in detail. There appear to be different types of methylation within the different functional and structural components of Arabidopsis, with complex interactions between methyltransferases acting to control levels of DNA methylation. The methylation of pericentromeric regions, DNA repeats and transposable elements occurs at [CG], [CHG] and [CHH] di/trinucleotides, while gene body methylation occurs at [CG] dinucleotides (Cokus *et al.,* 2008). In addition, 55 % of the methylation in DNA from floral tissue was found to be at [CG] dinucleotides (Lister *et al.,* 2008). In the *met-1* mutant (*met1-3*) the level of [CG] methylation is >1 % of that in the wild-type, although [CHG] methylation is increased, and is found within the gene bodies of more than 2,000 genes (Lister *et al.,* 2008). This study also compared the sites of DNA methylation with the snRNAome and found snRNAs associated with one third of the sites of DNA methylation and down-regulation of snRNA transcription in the *met-1* mutant.

There have been few studies of the effect of DNA methylation on nucleosome positioning (Pennings *et al.*, 2005). Nucleosomes containing H1 have been shown to form on methylated DNA in mouse (Ball *et al.*, 1983). However, DNA methylation has been shown to alter the position of a nucleosome in the chicken β^A-globin promoter (Davey *et al.*, 2004). The β^A-globin promoter contains a $[CpG]_3$ element which occupies a position approximately 1.5 helical turns from the dyad axis *in vitro*, which was disrupted when the DNA was methylated. However, DNA methylation was not shown to alter the nucleosome positioning over the imprinting control region of the mouse *Igf2*r gene (Davey and Allan, 2003). Whole-genome analysis of the histone variant H2A.Z revealed an antagonistic relationship between DNA methylation and H2A.Z-containing nucleosomes (Zilberman *et al.*, 2008). H2A.Z – containing nucleosomes were excluded from highly methylated regions such as the pericentromeric region. In addition, mutation of the complex responsible for deposition of H2A.Z resulted in genome-wide hypermethylation. This suggests that wild-type levels of H2A.Z-containing nucleosomes protect euchromatic DNA from DNA methylation.

With this in mind, the aim of this chapter was to determine if patterns of nucleosome occupancy over the Arabidopsis rRNA regions correlate with the pattern of DNA methylation in the same region, using publicly-available DNA methylation data (Lister *et al.*, 2008).

6.2 Methods

To enable comparisons between the nucleosome occupancy and DNA methylation data, the positions of DNA methylation (Lister *et al.*, 2008) were added to the annotation files as previously described (Chapter 2, section 2.2.vi), by assignment of the methylation score to each position along the Arabidopsis genome. These datasets were generated by Graham King, Rothamsted Research.

6.3 Results

In order to investigate the effect of DNA methylation, nucleosome occupancy was compared to previously published ^{5m}C in Arabidopsis wild type the *met-1* mutant (Lister *et al.*, 2008). Due to time limitations and tissue differences, comparisons were made over sequences representing NOR regions only.

6.3.1 Nucleosome occupancy and DNA methylation within the Arabidopsis nucleolar organiser region (NOR).

The Arabidopsis rRNA regions, which are annotated on chromosomes 2 and 3, have been shown in this project to vary in nucleosome linker length distribution and in nucleosome occupancy (Chapter 4, Section 4.3.4. and Chapter 5, Section 5.3.4). In order to test the hypothesis that DNA methylation and nucleosome positioning, the nucleosome occupancy from datasets *Ath_03_wt*, *Ath_05_wt* and *Ath_04_MET1* over the rRNA regions annotated on Chrs 2 and 3 were plotted (Figures 6.1 and 6.2). The DNA methylation data for wild-type were plotted along with the *Ath_03_wt* and *Ath_05_wt* datasets.

There appears to be some distinct patterns of DNA methylation with respect to nucleosome position between Chrs 2 and 3. For the *Ath_05_wt* dataset, the peaks in DNA methylation appear to coincide with the boundaries of the nucleosome occupancy specifically over the 18S and 25S rRNA genes on Chr 2 (indicated by black arrows Figure 6.1). In addition, the region immediately upstream of the 5S rRNA gene appears to be highly methylated (compared to the 18S and 25S rRNA genes), indicated by red arrow 1 (Figure 6.1). The nucleosome positioning relative

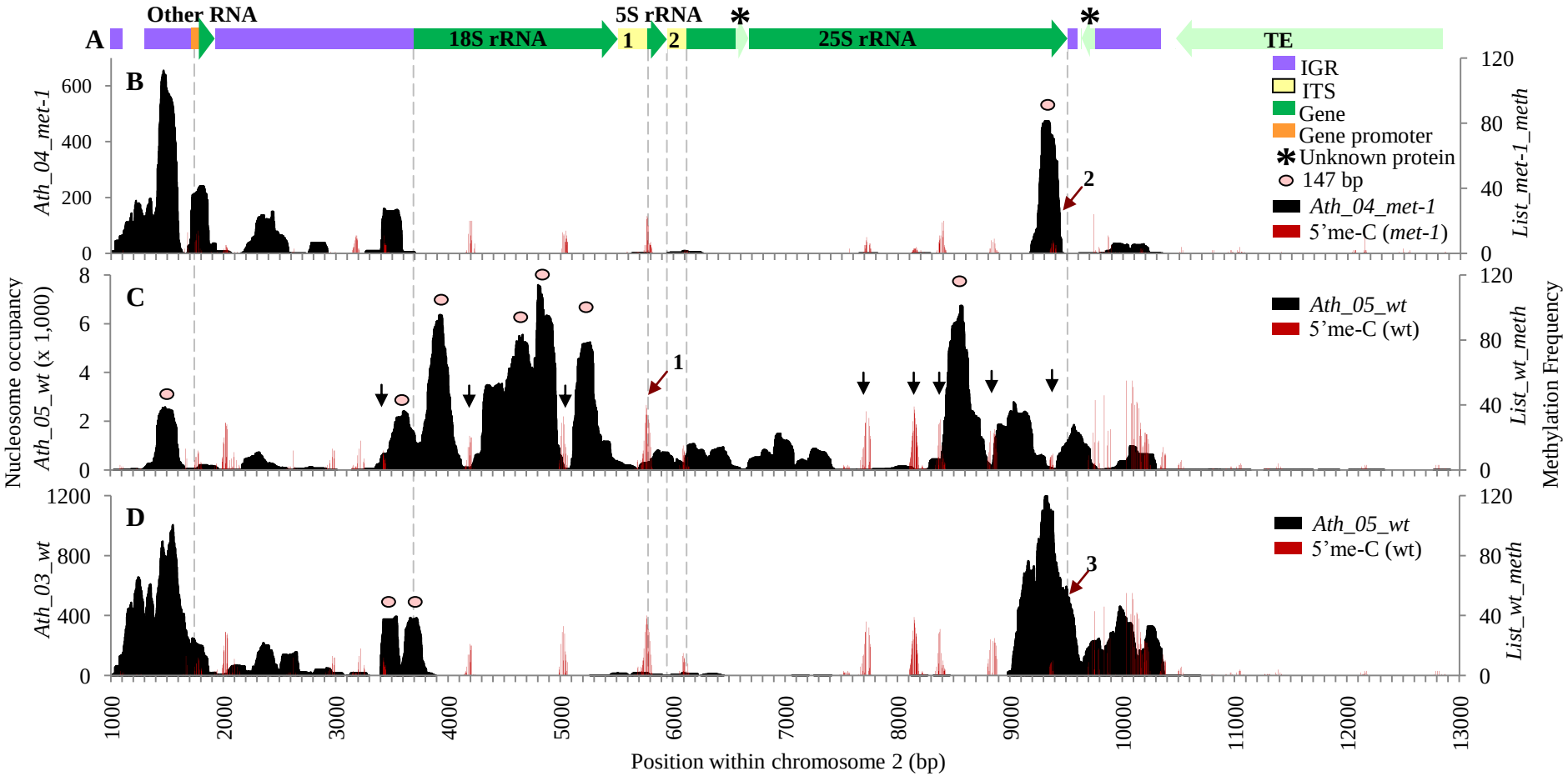

Figure 6.1 Position of 5me-C relative to nucleosome positions over the rRNA region on chromosome 2. A: relative positions of the rRNA genes, B: nucleosome dataset *Ath_04_met-1* and ^{5me}C levels in met-1 (Lister *et al.*, 2008), C: *Ath_05_wt* datasets and position of 5me-C in wild-type and D: the nucleosome positions of dataset *Ath_03_wt* and wild-type ^{5me}C. Black arrows indicate DNA methylation between nucleosome peaks. Red arrow (1) indicates high methylation at the 5' end of the 5S rRNA gene, red arrows 2 and 3 indicate methylation within the nucleosome peaks in datasets *Ath_04_met-1* and *Ath_03_wt*, respectively.

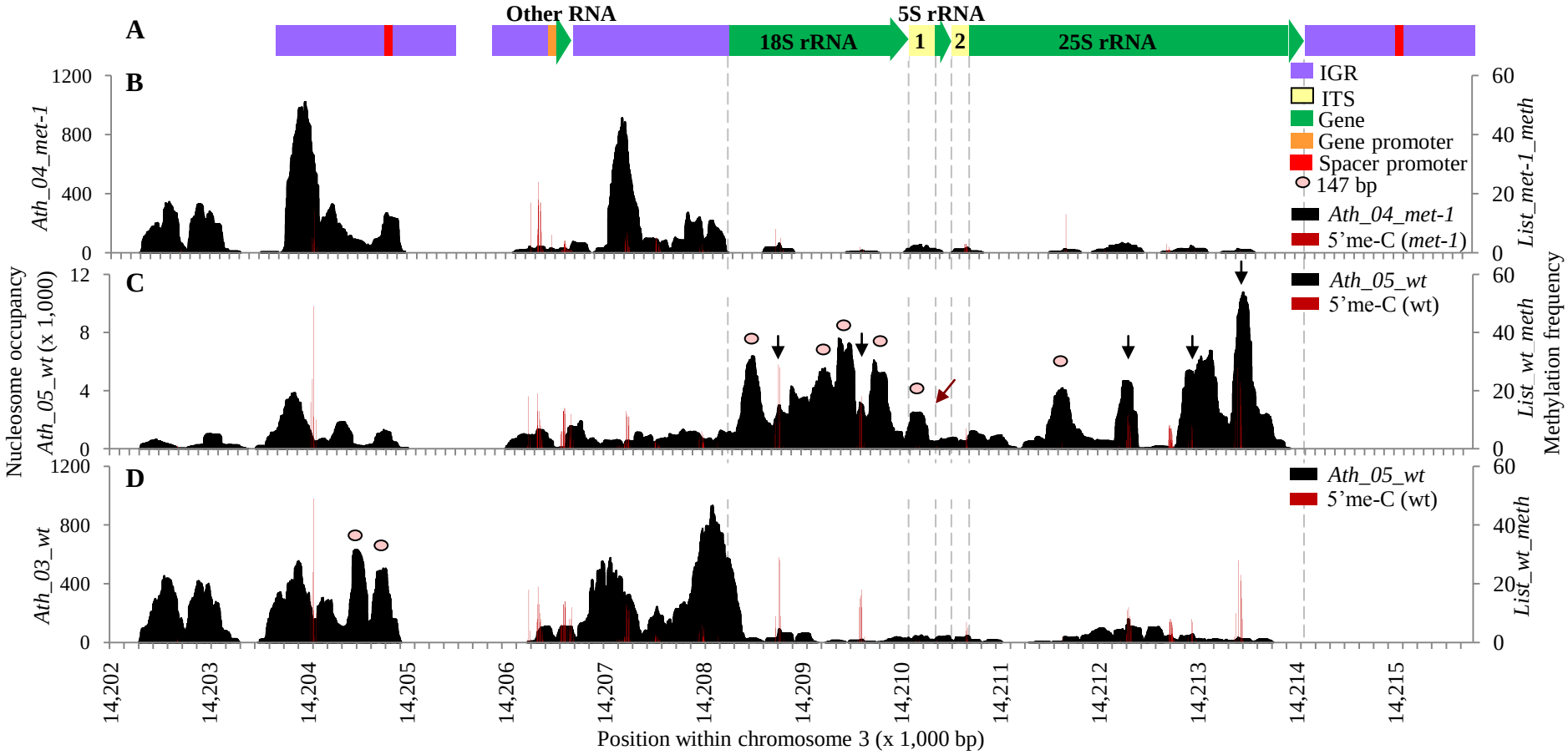

Figure 6.2 Position of 5me-C relative to nucleosome positions over the rRNA region on chromosome 3. A: relative positions of the rRNA genes, B: nucleosome dataset *Ath_04_met-1* and ^{5me}C levels in met-1 (Lister *et al.*, 2008), C: *Ath_05_wt* datasets and position of 5me-C in wild-type and D: the nucleosome positions of dataset *Ath_03_wt* and wild-type ^{5me}C. Black arrows indicate DNA methylation within nucleosome peaks. The red arrow indicates the position of the high methylation observed at the 5' end of the 5S rRNA gene on Chr 2 but is absent from Chr 3.

to the position of the DNA methylation appears to be different in the *Ath_04_MET1* dataset.

The DNA methylation in the *Ath_04_MET1* occurs predominantly at [CNG] sites (anti-sense *MET1* CG methylation <20% of wild-type levels) and is about half of the level of methylation in the wild-type, but appears to be in the same positions. The nucleosome peak adjacent to the 3' boundary of the 25S rRNA gene of datasets *Ath_03_wt* and *Ath_04_MET1* appears to coincide with a peak in the DNA methylation, (as opposed to the edges for Ath_05_wt) and is indicated by red arrows 2 and 3.

In contrast to Chr 2, for Chr 3, the DNA methylation occurs within the nucleosome peaks of the *Ath_05_wt* dataset (indicated by black arrows, Figure 6.2). Closer inspection shows that the methylation marks appear to be in similar positions on Chrs 2 and 3, with respect to the positions of the annotated rRNA genes. The levels of DNA methylation on Chr 3 appears to be about half of that on Chr 2. In addition, the DNA methylation observed at the 5' edge of the 5S rRNA gene on Chr 2 appears to be absent from the corresponding position on Chr 3.

6.4 Discussion

The rRNA genes annotated on chromosomes 2 and 3 show differences in the nucleosome positioning relative to the position of the ^{5m}C. The differences in position of the DNA methylation in relation to the nucleosome occupancy have not previously been demonstrated. The sequence of the coding regions of the rRNA genes is highly conserved in Arabidopsis (Riddle and Richards, 2002), however, variation in the epigenetic modifications have been shown (Woo and Richards, 2008).

The difference in methylation around the 5S rRNA gene between chromosome 2 and 3 was particularly interesting. The 5' and immediate upstream region of the 5S rRNA gene on Chr 2 appeared to be heavily methylated (compared to Chr 3). There also appears to be a denser region of DNA methylation near the gene promoter on Chr 2, whereas the DNA methylation is immediately upstream and downstream of the gene promoter on Chr 3. Previous studies have suggested that promoter-methylation occurs in genes which are developmentally regulated or transcriptionally silent (Zhang *et al.,* 2006). This suggests that there may be more transcriptionally active copies of the rRNA genes in the region annotated on Chr 3 than Chr 2.

The differences observed in nucleosome position between Chrs 2 and 3 may reflect differences in histone modifications between the two regions. Studies of nucleolar dominance in Arabidopsis allotetraploids have revealed that the rRNA genes from one parent are transcriptionally silenced, and that the silencing is reversed by treatment with DNA methylation inhibitors of histone deacytalase inhibitors (Earley *et al.*, 2006). This suggests that the silencing of rRNA genes in genetic hybrids has an epigenetic basis. It is hypothesised that similar mechanisms control rRNA gene dosage in Arabidopsis diploids (Lawrence *et al.*, 2004). It would be useful to compare these data with the positions of various histone modifications such as H3K9me^3 (Zhang *et al.*, 2007), and H2A.Z-containing nucleosomes (Zilberman *et al.*, 2008), associated with transcriptional repression and activation, respectively.

6.4.1 Summary

In summary, nucleosome occupancy over rRNA sequences representative of regions from two chromosomes was compared to the DNA methylation at the same sites obtained from previously published data. This revealed differences between the rRNA regions with respect to nucleosome positioning and DNA methylation not previously reported. DNA methylation was found to occur between peaks in nucleosome occupancy on Chr 2 and within peaks on Chr 3. It is likely that the differences represent differences in histone modifications over the two regions.

Chapter 7

General Discussion

General discussion

This thesis has presented an analysis of nucleosome positioning data for the Arabidopsis genome gathered by sequencing of nucleosome DNA fragments, using three different sequencing technologies: traditional Sanger sequencing, 454 FLX technology and Solexa technology. The datasets created by the 454 and Solexa sequencing technologies enable genome-wide study of nucleosome positioning preferences across the Arabidopsis genome and are comparable to the studies of nucleosome positioning in other taxa including yeast (Lee *et al.,* 2007), *C. elegans* (Johnson *et al.,* 2006), *Drosophila* (Mavrich *et al.,* 2008a) and human (Schones *et al.,* 2008).

7.1 Nucleosome positioning in Arabidopsis

It has been demonstrated in Chapter 3 that Arabidopsis nucleosome core sequences display similar characteristics to those isolated from other taxa, including chicken and yeast. In particular, the [AA/TT] distribution observed within the 147 bp of Arabidopsis nucleosome DNA fragments is a general characteristic of nucleosome sequences (Ioshikhes *et al.,* 1996; Satchwell *et al.,* 1986), although the base preference has been shown to vary between taxa (Kogan *et al.,* 2006). The presence of a nucleosome dinucleotide preference within Arabidopsis nucleosome DNA fragments suggests that, for at least a subset within the Arabidopsis genome, nucleosomes are rotationally positioned (or positioned with respect to the DNA sequence structure/curvature). Previous studies have suggested that nucleosomes tend to be rotationally positioned at the promoter of transcriptionally silenced genes and translationally positioned at transcriptionally active genes (Mahesh *et al.,* 2005; Chen and Yang, 2001). It is likely that there are contributions to the rotational positioning of nucleosomes other than the well-characterised [AA/TT] preference, and periodicities were detected in all of the complementary base combinations.

Evidence presented in Chapter 4 demonstrated a mean linker length for Arabidopsis which is similar to those described for other taxa (Yuan *et al,,* 2005; Kato *et al.,* 2003; and Satchwell and Travers, 1989). However, it is perhaps more useful to consider the range of linker lengths, as length distributions associated with smaller

datasets are non-normally distributed. While it has been reported that different cell types have different linker lengths (Lohr *et al*., 1977), it is likely that linker lengths also vary significantly between adjacent regions. It may be that linker lengths only appear uniform in the ordered arrays found in heterochromatin (Yang *et al.,* 2006), or in specific contexts of exons, dependent upon the level of transcriptional activation. A periodicity detected within the linker length distribution of ~10 bp has been reported previously (Widom, 1992) from sample sizes of 1,001 and 185, respectively. Although this was not detected in the Arabidopsis datasets, a shorter period of ~7-8bp was found. Following this discovery, the periodicities of the available datasets from other taxa were investigated. This revealed that the reported periodicity of 10 bp was only found in the chicken dataset (which also happens to be the smallest dataset), which may be skewed towards repetitive sequences. However, ~10 bp periodicities have been found in the spacing of nucleosome positioning signals within the human genome (Kato *et al.,* 2003). One reason for the observed variations in periodicity may be that the 'default setting' of nucleosome positioning signals are encoded in the DNA sequence and spaced at (multiples of) 10 bp periods, but the deviations from this detected in the Arabidopsis dinucleosome linker length datasets are a reflection of re-positioning at specific locations within the genome by active chromatin remodelling, such as around the TSS or at exon/intron boundaries. In addition, the difference between the mononucleosome and dinucleosome datasets suggests that the dinucleosomes extracted may represent a subset of nucleosomes with a different susceptibility to enzymatic digestion than the mononucleosomes. Therefore the estimation of linker length from dinucleosome subsets, may result in a bias towards inclusion of only a specific class of nucleosomes.

The difference in linker lengths of dinucleosomes aligned to the rRNA regions of chromosomes 2 and 3 was of particular interest. These differences most likely represent differences in the epigenetic landscape between the two regions.

DNA fragments which uniquely map to the rRNA genes on either Chr 2 or Chr 3 were analysed and some differences between the two regions were observed. There were two peaks within the distribution of linker length for fragments of Chr 3: one around 30 bp and one around 42 bp while the rRNA region on Chr 2 has a peak around 67 bp. These findings suggest that the rRNA regions annotated on Chrs 2

and 3 have different predominant chromatin structures, possibly relating to the transcriptional activity of these regions. Analysis of the histone modifications revealed that a fraction of the Arabidopsis rRNA genes are associated with the euchromatic mark H3K4me3 (Lawrence *et al.*, 2004). Furthermore, it was found that a fraction of the rRNA genes associated with heterochromatic H3K9me2, suggesting that some Arabidopsis rRNA genes copies are switched on and some are off. It is possible that the structure observed in this study over the rRNA regions may give some clues as to the proportions of each NOR which are active. Since the distributions overlap, it is likely that a dominant structure forms on each NOR. The lack of precision in current available sequence datasets for NORs in Arabidopsis currently hinders detailed analysis. However, it may be possible to test the relative abundance of transcripts with SNPs which correspond to either Chr 2 or 3 as there are likely to be variety of SNPs within the repeated regions. Experiments with Arabidopsis allotetraploids demonstrated that the rRNA genes inherited from *A. thaliana* in the tetraploid *A. suecica* were silenced and the contribution from *A. arenosa* were transcriptionally active (Chen *et al.*, 1998). However, in synthetic hybrids, under-dominance of the A. thaliana genes can take two generations to establish. It is likely that the transcriptional state, and therefore the chromatin structure, is altered with development.

Approximately half of the Arabidopsis genome was found to be occupied by nucleosomes. This is less than the previously reported 81 % in the 10 times smaller yeast genome (Lee *et al.*, 2007). Nucleosome occupancy was found to be higher in genic regions in wild-type Arabidopsis, which is supported by the finding that nucleosomal DNA has a higher GC content than the genome average (Chapter 3, section 3.3.3.ii). This also agrees with genomic nucleosome positioning data for yeast (Lee *at al.*, 2007). A specific pattern of occupancy in the vicinity of the TSS was demonstrated for Arabidopsis, with a peak (midpoint of the nucleosome) ~60 bp downstream of the TSS. This would position the TSS within the first 10 bp of the nucleosome, which associates with histone H2. A recent study of the interactions between the histone octomer and nucleosomal DNA suggest strong interactions around the dyad, with (weaker) interactions between the histone and DNA observed up to 60 bp from the dyad (Hall *et al.*, 2009). Thus the TSS in Arabidopsis is more likely to be within the region of weakest interaction between DNA and histone

octomer, and may result in Pol II having limited access to the TSS (rather than being excluded from it). Occupancy around the TSS has in the past couple of years been studied in a variety of non-plant eukaryotes, including yeast (Mavrich *et al.,* 2008b; Lee *et al.,* 2007; Albert *et al,* 2007), *Drosophila* (Mavrich *et al,* 2008a) and human (Schones *et al.,* 2008). Interestingly, these studies revealed taxa-specific differences in the organisation of nucleosomes in the vicinity of the TSS. Specific patterns of phasing of nucleosomes around the TSS were observed in all species. The TSSs in yeast tend to be buried ~ 13 bp inside the border of the +1 nucleosome the first nucleosome downstream of the TSS (Mavrich *et al.,* 2008b), whilst in Drosophila, the +1 nucleosome was found to be 180 bp upstream of the TSS, leaving the TSS unoccupied. Nucleosomes in the vicinity of the TSS in humans were found to be dependent on the transcriptional activity of the gene (Schones *et al.,* 2008). Phased nucleosomes were detected both upstream and downstream of the TSS of active genes, while only one nucleosome was detected in the TSS of non-transcribed genes, downstream of the TSS. Furthermore, the position of the 5' end of the +1 nucleosome was found to be 40 bp downstream of the TSS in active genes and 10 bp downstream in inactive genes. The position of the nucleosome boundary is thought to overlap the position of bound Pol II within inactive genes (Schones *et al.,* 2008). Therefore, the positioning around the TSS in Arabidopsis is similar to that of yeast, burying the TSS within the nucleosome, while the positioning of nucleosomes within the *Drosophila* and human genomes suggest a different mechanism of transcriptional control. It should be noted that the estimates of Arabidopsis positioning around the TSS presented here are based on a subset of the genome (12% of Chr 1), and consequently the distribution of occupancy was noisy. However, smoothing of the distribution was effective in revealing the trends of nucleosome positioning in the vicinity of the TSS. Extending the analysis of nucleosome occupancy across the entire Arabidopsis genome should negate the need for smoothing of the distribution.

7.2 Differences in nucleosome position resulting from loss of CG methylation

The Arabidopsis anti-sense *MET1* line was included in this study to investigate the effect of DNA methylation on nucleosome positioning. There were some systematic differences in the properties of nucleosome positions observed between anti-sense *MET1* and the wild-type datasets. One of the most striking differences was the

difference in linker length distributions. In general, the average linker length within the anti-sense *MET1* dataset was higher than those of the wild-type datasets. A periodicity of 9.2 – 9.8 bp was detected within the distribution of anti-sense *MET1*, which is close to the published value of ~10 bp for a wide range of organisms (Widom, 1992). One of the explanations for the discrepancies in the periodicity observed is that de-regulation of positioning allows the nucleosomes to revert to a DNA sequence-dependent 'default setting', with the DNA sequence becoming a more important positioning factor than in the wild-type that possesses ^{5m}CG. This might suggest that a greater proportion of nucleosomes in anti-sense *MET1* are rotationally positioned and the sequence periodicity within the nucleosome DNA reflects the sequence 10 bp periodicity found throughout the genome. This might be reflected in a stronger sequence preference within anti-sense *MET1* nucleosome DNA fragments; however the strength of the sequence preference was not tested here. Slight variation in dinucleotide frequencies within the linker regions was observed between anti-sense *MET1* and wild-type. This result suggests that, although the same distributions of dinucleotides are observed within the nucleosome DNA, the nucleosomes are (in general) in different positions in the wild type and anti-sense *MET1*. This could be due to the increased stiffness of the methylated wild–type DNA molecule, compared to the anti-sense *MET1*, which is brought about by DNA methylation (Virstedt *et al.*, 2004). Thus the primary consequence of DNA methylation marks may be to influence translational setting of nucleosomes. This is likely to occur through modulation of bendability within the minor groove of the DNA, affecting the ability of the DNA to interact efficiently with the histone octamer (Buttinelli *et al.,* 1998). Methylation of the nucleotides at the chicken β-globin gene has been found to inhibit nucleosome formation (Davey *et al.,* 1997).

The apparent de-regulation of nucleosome positioning in anti-sense *MET1* compared to the wild-type was demonstrated within small sections of Chr 1. Whilst the nucleosome occupancy appears to be higher within the coding regions compared to the intergenic regions in both anti-sense *MET1* and wild-type, the anti-sense *MET1* did have more nucleosome occupancy within intergenic regions compared to the wild-type. The analysis of the relative occupancies in the different annotation classes over Chr 1 seems to support this, suggesting that overall nucleosome occupancy within the Arabidopsis anti-sense *MET1* genome are more evenly distributed over the

different annotation classes. There are two possible explanations for the apparent de-regulation of nucleosome positioning in anti-sense *MET1*, (1) the effect of the interaction between DNA methylation within the nucleosome core and the core histones, and (2) the effect of the interaction between methylated DNA at the nucleosome boundary and linker histones. Linker histone H1 has shown a preference for methylated DNA (McArthur and Thomas, 1996). In addition, mutants with down-regulated linker histone levels exhibit phenotypes of the DNA hypomethylation methylation mutants (Wierzbicki and Jerzmanowski, 2004). Atomic Force Microscopy of chromatin fibres with normal and elevated levels of DNA methylation revealed that hypermethylated fibres are more compact than those with wild-type levels (only in the presence of linker histone (Karymov *et al.*, 2001), further suggesting that wild-type levels of DNA methylation are required for chromatin fibre condensation. If H1 has a higher affinity for methylated DNA, nucleosomes could be positioned at the nearest site of DNA methylation in the Arabidopsis anti-sense *MET1*. This may help to explain the de-localisation of nucleosome positions. It is thought that most nucleosomes are associated with a linker histone (~ 80 %) and that the association is transient (Lever *et al.*, 2000). The length of time that a linker histone is associated with the nucleosome seems to depend on the post-translational modifications. Acetylated histones had a shorter association with linker histones, which was thought to be due to a higher rate of exchange by chromatin remodelling (Mistelli *et al.*, 2000). The rate of chromatin remodelling could be tested by comparison of nucleosome occupancy at the same loci over different developmental tissues, and compare to the ^{5m}C within the different tissues to determine whether any interaction occurs between the two epigenetic marks.

The level of non- CG DNA methylation has been found to be higher and more extensive in the *met-1* mutant than in the wild-type (Cokus *et al.*, 2008). In the wild-type plants, methylation of the genes body occurs exclusively at the CG dinucleotides, whereas non- CG methylation (along with CG methylation) is found in the pericentromeric and centromeric regions, and within transposable elements (Cokus *et al.*, 2008). In addition, DNA methylation is known to affect the modifications of histones within nucleosomes. The complete loss of CG methylation from Arabidopsis has been shown to result in the loss of H3K9 methylation from

within constitutive heterochromatin (Tariq *et al.,* 2003). The H3K4 di and tri-methylation marks are associated with euchromatic transcriptionally active regions and tend to be absent from regions of DNA methylation in Arabidopsis. However, H3K4 mono-methylation is highly correlated with the presence of DNA methylation in the transcribed regions of genes (Zhang *et al.,* 2009). In addition, nucleosomes containing the histone variant H2AZ have been shown to be associated with transcriptionally active promoter regions (Albert *et al.,* 2007; Mavrich *et al.,* 2008; Schones *et al.,* 2008), and are excluded from regions containing DNA methylation (Zilberman *et al.,* 2008). A deficiency of CG methylation in the Arabidopsis anti-sense *MET1* may lead to the proliferation of euchromatic histone marks, resulting in a more relaxed, open chromatin conformation.

7.3 Higher-order chromatin structure

One of the questions raised at the beginning of this study was whether it is possible to predict chromatin higher-order structure from knowledge of nucleosome spacing, and if so, whether the higher-order structure can give any clues linking structure and transcriptional control. Early studies of nucleosome positioning and linker length variation postulated that increases in nucleosome repeat length were correlated with increases in transcriptional activity (Ull and Franco, 1986). In contrast, more recent studies of heterochromatin protein ACF suggest that such nucleosomes are spaced 50-60 bp apart in human heterochromatin (Yang *et al.,* 2006). It is possible that heterochromatic, or silenced, regions have a longer linker length and so a more efficient compaction can occur in the presence of linker histone, which would render more of the DNA inaccessible to regulatory proteins. With this in mind, the differences in linker length observed over the rRNA genes in Arabidopsis were intriguing (Chapter 4, Section 4.3.4). Models and measurements of 30 nm fibres suggest two classes (Robinson *et al.,* 2007; Wong *et al.,* 2007). This would place the linker lengths of Chr 2 NOR within the large class chromatin fibre and the smaller linker length represented by the rRNA genes on Chr 3 within two different classes in the smaller 30 nm fibre. Although an individual NOR is thought to be both active and inactive, these results may show a bias towards either the active or inactive state (or that more of the rRNA genes represented by chr3 are active and more on Chr 2 are silenced). One way to test this would be to study the positioning of nucleosomes

over the rRNA regions from condensed 30 nm fibres, a method which has been previously used for the study of solenoid *vs* zig-zag compaction (Staynov and Proykova, 2008). As active NORs remain decondensed during metaphase it may be possible to distinguish between the two states. It is worth noting that while there are dominant peaks within the distributions of linker length for each region, there is overlap of the distributions between the two regions. This indicates that there may be chromatin structures which are present in both rRNA regions in varying amounts.

In addition, it may be possible to provide other insights into higher-order chromatin structure from the different datasets generated in this study. Mononucleosome and dinucleosome datasets exhibited systematic differences in nucleosome positioning, which may reflect differences in chromatin condensation and so affect target sites for MNase digestion. Thus oligo-nucleosome bands present in MNase-generated nucleosome ladders may show increasing levels of compaction with increase in oligomer size. This phenomenon is worthy of investigation in the future as it may provide information on the hierarchies of nucleosome organisation present within *in vivo* chromatin.

7.4 Technological advances

During the duration of this project, sequencing technologies have progressed significantly. This enabled the various studies to be undertaken, at increasing precision but also with exponentially increasing size of datasets. In fact, analyses that seemed intractable at the beginning of the project, due to costs and starting material requirements became attainable towards the end. Much effort was expended in attempting to acquire nucleosome positioning data from the Arabidopsis tiling arrays, with four experimental (nucleosome) datasets and one control dataset being produced. Following extensive analysis with the tiling array datasets, a nucleosome signal which was uniformly distinct from the control could not be detected, possibly as a result of biases in probe binding efficiencies. However, these datasets were subsequently superseded by the high resolution high-throughput DNA sequence datasets.

7.5 Conclusions and future work

The large nucleosome positioning datasets generated in this study by high-throughput technologies allowed the elucidation of details which were not possible with the original smaller datasets (*Ath_01_wt*, n=475). However, at the time this dataset was generated, it was the first nucleosome positioning dataset generated from plant chromatin, and was larger than the chicken nucleosome positioning datasets (Drew and Travers, 1985). The *Ath_05_wt* dataset represents an average of 6.4 x coverage of the genome assuming an average linker length of 46 bp based on the average linker length of the two wild-type dinucleosome datasets. Attempts to determine any patterns of dinucleotide distribution within nucleosome sequences did not yield statistically meaningful results with the smaller *Ath_01_wt* dataset, and the coverage was insufficient to identify patterns of occupancy around the TSS.

In would be valuable to compare the nucleosome position datasets generated in this study with those associated with other epigenetic marks, for example H2A.Z (Zilberman *et al.,* 2008), H3K9me3 (Zhang *et al.,* 2007), and presence of H1 when this may become available. Such analyses are likely to provide a more comprehensive understanding of the interactions between 5mC marks, nucleosome rotational and translational positioning and transcriptional control of different gene families.

It is becoming clear that there are many factors involved in nucleosome positioning, and a complex relationship exists between, DNA sequence, DNA methylation histone modifications, and presence/absence of linker histone and that these factors are context dependant. The datasets generated in this project provide a comprehensive overview of the nucleosome landscape across the whole genome. This first study of the nucleosome positions in a plant genome provides a useful starting point for more detailed comparisons of nucleosome positioning within specific genes at different developmental stages, under different environmental conditions, and in response to biotic and abiotic stresses. Many examples of epigenetic-control of transcriptional regulation now exist in Arabidopsis. For example, FLC is one of the most studied epigenetically controlled developmental switches (Saleh *et al.,* 2008; Shindo *et al.,* 2006; Bastow *et al.,* 2004). In a wider context, understanding of the specific processes controlling genes involved in the

response to environmental stimuli, such as drought and salt stress, have practical applications for economically important crops.

Chapter 8

References

8 References

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

Alegre, C., Subirana, J.A. (1989) The diameter of chromatin fibres depends on linker length. *Chromosoma* **98**: 77-80

Allan, J., Rau, D.C., Harborne, N., Gould, H. (1984) Higher-order structure in a short repeat length chromatin. *The Journal of Cell Biology* **98**: 1320-1327

An, W., van Holde, K., Zlatanova, J. (1998) Linker histone protection of chromatosomes reconstituted on 5S rDNA from Xenopus borealis: a reinvestigation. *Nucleic Acids Research* **26**: 4042–4046

Anselmi, C., Bocchinfuso, G., De Santis, P., Savino, M., Scipioni, A. (1999) Dual role of DNA intrinsic curvature and flexibility in determining nucleosome stability. *Journal of Molecular Biology* **286**: 1293-1301

Baldi, P., Brunak, S., Chauvin, Y., Krogh, A. (1996) Naturally occurring nucleosome positioning signals in human exons and introns. *Journal of Molecular Biology* **263**: 503-510

Ball, D.J., Gross, D.S., Garrard, W.T. (1983) 5-Methylcytosine is localized in nucleosomes that contain histone H1. *Proceedings* of *the National Academy of Sciences of the USA.* **80:** 5490-5494

Bastow, R., Mylne, J.S., Lister, C., Lippman, Z., Martienssen, R.A., Dean, C. (2004) Vernalisation requires epigenetic silencing of FLC by histone modification. *Nature* **427**: 164-167

Beckman, J.S. and Trifonov, E.N. (1991) Splice junctions follow a 205-base ladder. *Proceedings* of *the National Academy of Sciences of the USA.* **88**: 2380–2383

Bernad, R., Sánchez, P., Losada, A. Epigenetic specification of centromeres by CENP-A, *Experimental Cell Research*. [Published online ahead of print Aug 3 2009] PubMed: PMID 19660450

Boeger, H., Griesenbeck, J., Stratten, J.S., Kornberg, R.D. (2003) Nucleosomes unfold completely at a transcriptionally active promoter. *Molecular Cell* **11**: 1587-1598

Boeger, H., Griesenbeck, J., Stratten, J.S., Kornberg, R.D. (2004). Removal of promoter nucleosomes by disassembly rather than sliding in vivo. *Molecular Cell* **14**: 667-673

Boeger, H., Bushnell, D.A., Davis, R., Griesenbeck, J., Lorch, Y., Stratten, J.S., Westove, K.D., Kornberg, R.D. (2005) Structural basis of eukaryotic gene transcription. *FEBS Letters* **579**: 899-903

Bolshoy, A., McNamara, P., Harrington, R.E., Trifonov, E.N. (1991) Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proceedings* of *the National Academy of Sciences of the USA.* **88**: 2312-2316

Bowler, C., Benevenuto, G., Laflamme, P., Molino, D., Probst, A., Tariq, M., Paszowski, J. (2004) Chromatin techniques for plant cells. *The Plant Journal* **39:** 776-789

Bradnam, K.R., and Korf, I. (2008) Longer first introns are a general property of eukaryotic gene structure. *PLoS One* **3**: e3093

Bustin, M., Catez, F., Lim, J.H. (2005). The dynamics of Histone H1 Function in chromatin. *Molecular Cell* **17**: 617-620

Buttinelli, M., Minnock, A., Panetta, G., Waring, M., Travers A. (1998) The exocyclic groups of DNA modulate the affinity and positioning of the histone octamer. *Proceedings* of *the National Academy of Sciences of the USA.* **95**: 8544-8549

Cao, H., Widlund, H.R., Simonsson, T., Kubista, M. (1998) TGGA repeats impair nucleosome formation. *Journal of Molecular Biology* **281**: 253-260

Caperta, A.D., Neves, N., Morais-Cecílio, L., Malhó, R., Viegas, W. (2002) Genome restructuring in rye affects the expression, organization and disposition of homologous rDNA loci. *Journal of Cell Science* **115**: 2839-2846

Chan, S.W., Zilberman, D., Xie, Z., Johansen, L.K., Carrington, J.C., Jacobsen, S.E. (2004) RNA silencing genes control de novo DNA methylation. *Science* **303**: 1336

Chen, Z.J., Comai, L., Pikaard, C.S. (1998) Gene dosage and stochastic effects determine the severity and direction of uniparental ribosomal RNA gene silencing (nucleolar dominance) in Arabidopsis allopolyploids. *Proceedings* of *the National Academy of Sciences of the USA.* **95**: 14891-14896

Chen, C. and Yang, T.P. (2001) Nucleosomes are translationally positioned on the active allele and rotationally positioned on the inactive allele of the HPRT promoter. *Molecular and Cellular Biology* **21**: 7682-7695

Cheng, X. (1995) Structure and function of DNA methyltransferases. *Annual Review of Biophysical and Biomolecular Structure* **24**: 293-318

Choong, M.K. and Yan, H. (2008) Multi-scale parametric spectral analysis for exon detection in DNA sequences based on forward-backward linear prediction and singular value decomposition of the double-base curves. *Bioinformation* **2**: 273-278

Chung, B.Y., Simons, C., Firth, A.E., Brown, C.M., Hellens, R.P. (2006) Effect of 5'UTR introns on gene expression in Arabidopsis thaliana. *BMC Genomics* **7**: 120

Chung, H-R. and Vingron, M. (2009) Sequence-dependent nucleosome positioning. *Journal of Molecular Biology* **386**: 1411-1422.

Cioffi, A., Fleurv, T.J., Stein, A. (2006) Aspects of large-scale chromatin structures in mouse liver nuclei can be predicted from the DNA sequence. *Nucleic Acids Research* **34**: 1974-1981

Cloutier, T.E., Widom, J. (2004) Spontaneous sharp bending of double-stranded DNA. *Molecular Cell* **14**: 335-362

Cohanim, A.B., Kashi, Y., Trifonov, E.N. (2005) Yeast nucleosome DNA pattern: deconvolution from genome sequences of S. cerevisiae. *Journal of Biomolecular Structure and Dynamics* **22**: 687-693

Cohanim, A.B., Kashi, Y., Trifonov, E.N. (2006) Three sequence rules for chromatin. *Journal of Biomolecular Structure and Dynamics* **23**: 559-566

Cohen P., Blackburn E.H. (1998) Two types of telomeric chromatin in Tetrahymena thermophila. *Journal of Molecular Biology* **280**: 327-344

Cokus, S.J., Feng, S., Zhang,X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., Jacobsen S.E. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452:** 215-219

Conconi, A., Widmer, R.M., Koller, T., Sogo, J.M. (1989) Two different chromatin structures coexist in ribosomal RNA genes throughout the cell cycle. *Cell* **57**: 753-761

Dalal, Y., Wang, H., Lindsay, S., Henikoff, S. (2007a) Tetrameric structure of centromeric nucleosomes in interphase Drosophila cells. *PLoS Biology* **5**:e218

Dalal, Y., Furuyama, T., Vermaak, D., Henikoff, S. (2007b) Structure, dynamics, and evolution of centromeric nucleosomes. *Proceedings* of *the National Academy of Sciences of the USA.* **104**: 15974-15981

Davey, C., Pennings, S., Allan, J. (1997) CpG methylation remodels chromatin structure in vitro. *Journal of Molecular Biology* **267**: 276-288

Davey, C.A. and Richmond, T.J. (2002) DNA-dependent divalent cation binding in the nucleosome core particle. *Proceedings* of *the National Academy of Sciences of the USA.* **99**: 11169-11174

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

Davey, C., Allan, J. (2003) Nucleosome positioning signals and potential H-DNA within the DNA sequence of the imprinting control region of the mouse Igf2r gene. *Biochimica et Biophysica Acta* **1630**: 103-116

Davey, C., Fraser, R., Smolle, M., Simmen, M.M., Allan, J. (2003). Nucleosome positioning signals in the DNA sequence of the human and mouse H19 imprinting control regions. *Journal of Molecular Biology* **325:** 873-887

Davey, C.S., Pennings, S., Reilly, C., Meehan, R.R., Allan J. (2004) A determining influence for CpG dinucleotides on nucleosome positioning in vitro. *Nucleic Acids Research* **32**:4322-4331

Denisov, D.A., Shpigelman, E.S., Trifonov, E.N. (1997) Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes. *Gene* **205**: 145-149

Dingwall, C., Lomonossoff, G.P., Laskey, R. A. (1981) High sequence specificity of micrococcal nuclease. *Nucleic Acids Research* **9**: 2659-2673

Doelling, J.H., Gaudino, R.J., Pikaard, C.S. (1993) Functional analysis of Arabidopsis thaliana rRNA gene and spacer promoters in vivo and by transient expression. *Proceedings* of *the National Academy of Sciences of the USA.* **90**: 7528-7532

Drew, H.R. and Travers, A.A. (1985) DNA bending and its relation to nucleosome positioning. *Journal of Molecular Biology* **186**: 773-790

Earley, K., Lawrence, R.J., Pontes, O., Reuther, R., Enciso, A.J., Silva, M., Neves, N., Gross, M., Viegas, W., Pikaard C.S. (2006) Erasure of histone acetylation by Arabidopsis HDA6 mediates large-scale gene silencing in nucleolar dominance. *Genes and Development* **20**: 1283-1293

Effron, B. and Tibshirani, R.J. (1994) An Introduction to the Bootstrap. Chapman & Hall, London, UK, 456 pages

Engelhardt, M. (2007) Choreography for nucleosomes: the conformational freedom of the nucleosomal filament and its limitations. *Nucleic Acids Research* **35**: e106

Fan, Y., Nikitina, T., Morin-Kensicki, E.M., Zhao, J., Magnuson, T.R., Woodcock, C.L., Skoultchi A.I. (2003) H1 linker histones are essential for mouse development and affect nucleosome spacing in vivo. *Molecular and Cellular Biology* **23**: 4559-4572

Flaus, A., Owen-Hughes, T. (2004) Mechanisms for ATP-dependent chromatin remodelling: farewell to the tuna-can octamer? *Current Opinion in Genetics and Development* **14**: 165-173

Finch, J.T., Noll, M., Kornberg, R.D. (1975) Electron microscopy of defined lengths of chromatin. *Proceedings* of *the National Academy of Sciences of the USA.* **72**: 3320-3322

Finch, J.T. and Klug, A. (1976) Solenoidal model for superstructure in chromatin. *Proceedings* of *the National Academy of Sciences of the USA.* **73**: 1897-1901

Finnegan, E.J., Peacock, W.J., Dennis, E.S. (1996) Reduced DNA methylation in Arabidopsis thaliana results in abnormal plant development. *Proceedings* of *the National Academy of Sciences of the USA.* **93**: 8449-8454

Fransz, P., ten Hoopen, R., Tessadori, F. (2006) Composition and formation of heterochromatin in *Arabidopsis thaliana*. *Chromosome Research* **275**: 3761-3771

Fraser, R.M., Keszenman-Pereyra, D., Simmen, M.W., Allan, J. (2009) High-resolution mapping of sequence-directed nucleosome positioning on genomic DNA. *Journal of Molecular Biology* **390**: 292-305

Gendrel, A.V., Lippman, Z., Martienssen, R., Colot, V. (2005) Profiling histone modification patterns in plants using genomic tiling microarrays. *Nature Methods* **2:** 213-218

Gilbert, N., Thomson, I., Boyle, S., Allan, J., Ramsahoye, B., Bickmore, W.A. (2007) DNA methylation affects nuclear organization, histone modifications, and linker histone binding but not chromatin compaction. *The Journal of Cell Biology* **177**: 401-411

Gottesfeld, J.M. and Melton, D.A. (1978) The length of nucleosome-associated DNA is the same in both transcribed and nontranscribed regions of chromatin. *Nature* **273**: 317-319.

Grigoryev, S.A., Spirin, K.S., Krasheninnikov, I.A. (1990) Loosened nucleosome linker folding in transcriptionally active chromatin of chicken embryo erythrocyte nuclei. *Nucleic Acids Research* **18**: 7397-7406

Grant, P.A. (2001) A tale of histone modifications. *Genome Biology* **2**: Reviews 1-6

Grover, A. and Sharma, P.C. (2007) Microsatellite motifs with moderate GC content are clustered around genes on Arabidopsis thaliana chromosome 2. *In Silico Biology* **7**: 0021

Hall, M.A., Shundrovsky, A., Bai, L., Fulbright, R.M., Lis, J.T., Wang, M.D. (2009) High-resolution dynamic mapping of histone-DNA interactions in a nucleosome. *Nature Structural and Molecular Biology* **16**: 124-129

Hsieh, T.F., Fischer, R.L. (2005) Biology of chromatin dynamics. *Annual Review of Plant Biology* **56**: 327-351

Henikoff, S., Furuyama, T., Ahmed, K. (2004) Histone variants, nucleosome assembly and epigenetic inheritance. *Trends in Genetics* **20**: 320-326

Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M., Trifonov, E. (1996) Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *Journal of Molecular Biology* **262**: 129-139

Jaeger, A.W. and Kuenzle, C.C. (1982) The chromatin repeat length of brain cortex and cerebella neurons changes concomitant with terminal differentiation. *The EMBO Journal* **7**: 811-816

Jeong, Y.M., Mun, J.H., Lee, I., Woo, J.C., Hong, C.B., Kim, S.G. (2006) Distinct roles of the first introns on the expression of Arabidopsis profilin gene family members. Plant Physiology **140**: 196-209

Johnson, S.M., Tan, F.J., McCullough, H.L., Riordan, D.P., Fire, A.Z. (2006) Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Research* **16**: 1505-1516

Kahmann, N.H. and Rake, A.V. (1993) Altered nucleosome spacing associated with Down syndrome. *Biochemical Genetics* **31**: 207-214

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., Segal, E. (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**: 362-366

Karymov, M.A., Tomschik, M., Leuba, S.H., Caiafa, P., Zlatanova, J. (2001) DNA methylation-dependent chromatin fibre compaction in vivo and in vitro: requirement for linker histone. *The FASEB Journal* **15**:2631-2641

Kato, M., Onishi, Y., Wada-Kiyama, Y., Abe, T., Ikemura, T., Kogan, S., Bolshoy, A., Trifonov, E., Kiyama, R. (2003) Dinucleosome DNA of human K562 cells: experimental and computational characterizations. *Journal of Molecular Biology* **332**: 111-125

Kato, M., Miura, A., Bender, J., Jacobsen, S. E., Kakutani, T. (2003b) Role of CG and Non-CG Methylation in Immobilization of Transposons in Arabidopsis. *Current Biology* 13: 421-426

Kato, M., Onishi, Y., Wada-Kiyama, Y., Kiyama, R. (2005) Biochemical screening of stable dinucleosomes using DNA fragments from a dinucleosome library. *Journal of Molecular Biology* **350**: 215-227

Kim, J.M., To, T.K., Ishida, J., Morosawa, T., Kawashima, M., Matsui, A., Toyoda, T., Kimura, H., Shinozaki, K., Seki, M. (2008) Alterations of lysine modifications on the histone H3 N-tail under drought stress conditions in Arabidopsis thaliana. *Plant and Cell Physiology* **49**: 1580-1558

Kiyama, R. and Trifonov, E. (2002) What positions nucleosomes? - A model. *FEBS Letters* **523**: 7-11

Kogan, S. and Trifinov, E.T. (2005) Gene splice sites correlate with nucleosome positions. *Gene* **352**: 57-62

Kogan, S.B., Kato, M., Kiyama, R., Trifonov, E.T. (2006) Sequence structure of human nucleosomal DNA. *Journal of Biomolecular Structure and Dynamics* **24**: 43-48

Köhler, C. and Makarevich, G. (2006) Epigenetic mechanisms governing seed development in plants. *EMBO Reports* **7**: 1223-1227

Längst, G. and Becker, P.B. (2001) ISWI induces nucleosome sliding on nicked DNA. *Molecular Cell* **8**: 1085-1092

Längst, G. and Becker, P.B. (2004) Nucleosome remodelling: one mechanism, many phenomena? *Biochimica et Biophysica Acta* **1677**: 58-63

Lawrence, R.J., Earley, K., Pontes, O., Silva, M., Chen, Z.J., Neves, N., Viegas, W., Pikaard, C.S. (2004) A concerted DNA methylation/histone methylation switch regulates rRNA gene dosage control and nucleolar dominance. *Molecular Cell* **13:** 599-609

Lee, C.K., Shibata, Y., Rao, B., Strahl, B.D., Lieb, J.D. (2004) Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genetics* **36**: 900-905

Lee, W., Tillo, D., Bray, N., Morse, R.H., Davis, R.W., Hughes, T.R., Nislow, C. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genetics* **10:** 1235-1244

Lever, M.A., Th'ng, J.P., Sun, X., Hendzel, M.J. (2000) Rapid exchange of histone H1.1 on chromatin in living human cells. *Nature* **408**: 873-876

Levitsky, V.G., Katokhin, A., Podkolodnaya, O.A., Furman, D.P., Kolchanov, N.A. (2005) NPRD: Nucleosome Positioning Region Database. *Nucleic Acids Research* **33**: Database issue D67-D70.

Li, G., Chandler, S.P., Wolffe, A.P., Hall, T.C. (1998) Architectural specificity in chromatin structure at the TATA box in vivo: nucleosome displacement upon beta-phaseolin gene activation. *Proceedings* of *the National Academy of Sciences of the USA.* **95**: 4772-4777

Lindroth, A.M., Shultis, D., Jasencakova, Z., Fuchs, J., Johnson, L., Schubert, D., Patnaik, D., Pradhan, S., Goodrich, J., Schubert, I., Jenuwein, T., Khorasanizadeh, S., Jacobsen, S.E. (2004) Dual histone H3 methylation marks at lysines 9 and 27 required for interaction with *CHROMOMETHYLASE 3*. *The EMBO Journal* **23**: 4286-4296

Lippman, Z., Genedrel, A.V., Black, M., Vaughn, M.W., Dedhia, N., McCombie, W.R., Lavine, K., Mittal, V., May, B., Kasschau, K.D., Carrington, J.C., Doerge,

R.W., Colot, V., Martienssen, R. (2004) Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**: 471-476

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., Ecker, J.R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**: 523-536

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B., Ecker, J.R. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322

Lohr, D., Corden, J., Tatchell, K., Kovacic, R.T., Van Holde, K.E. (1977) Comparative subunit structure of HeLa, yeast, and chicken erythrocyte chromatin. *Proceedings* of *the National Academy of Sciences of the USA.* **74**: 79-83

Lohr, D. and Van Holde, K.E. (1979) Organisation of spacer DNA in chromatin. *Proceedings* of *the National Academy of Sciences of the USA.* **76**: 6326-6330

Lorch, Y., LaPointe, J.W., Kornberg, R.D. (1987). Nucleosomes inhibit the initiation of transcription but allow chain elongation with displacement of histones. *Cell Research* **49**: 203-210

Lorch, Y. and Kornberg, R.D. (2004) Isolation and assay of the RSC chromatin-remodelling complex from Saccharomyces cerevisiae. *Methods in Enzymology* **377**: 316-322

Lubliner, S. and Segal, E. (2009) Modelling interactions between adjacent nucleosomes improves genome-wide predictions of nucleosome occupancy. *Bioinformatics* **25**: i348-355

Luschnig, C., Bachmair A., Schweizer D. (1993) Intraspecific length heterogeneity of the rDNA-IGR in Arabidopsis thaliana due to homologous recombination. *Plant Molecular Biology* **22**: 543-545

McArthur, M., Thomas, J.O. (1996) A preference of histone H1 for methylated DNA. *The EMBO Journal* **15**: 1705-1714

McStay, B. (2006) Nucleolar dominance: a model for rRNA gene silencing. *Genes and Development* **20**: 1207-1214

Marky, N.L. and Manning, G.S. (1995) A theory of DNA dissociation from the nucleosome. *Journal of Molecular Biology* **254**: 50-61

Marx, K.A., Zhou, Y., Kishawi, I.Q. (2002) Evidence for long poly (dA).poly(dT) tracts in *D. discoideum* DNA at high frequencies and their preferential avoidance of nucleosomal DNA core regions. *Journal of Biomolecular Structure and Dynamics* **23**: 429-446

Marx, K.A., Hess, S.T., Blake, R.D. (1994) Alignment of (dA).(dT) homopolymer tracts in gene flanking sequences suggests nucleosomal periodicity in *D. discoideum* DNA. *Journal of Biomolecular Structure and Dynamics* **12**: 235-246

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., Gilmour, D.S., Albert, I., Pugh, B.F. (2008a) Nucleosome organization in the Drosophila genome. *Nature* **453**: 358-362

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

Mengeritsky, G. and Trifonov, E.N. (1983) Nucleotide sequence-directed mapping of the nucleosomes. *Nucleic Acids Research* **11:** 3833-3851

Misteli, T., Gunjan, A., Hock, R., Bustin, M., Brown, D.T. (2000) Dynamic binding of histone H1 to chromatin in living cells. *Nature* **408**:877-881

Noll, M., Zimmer, S., Engel, A., Dubochet, J. (1980) Self-assembly of single and closely spaced nucleosome core particles. *Nucleic Acids Research* **8**: 21-42

Olins, A.L. and Olins, D.E. (1974) Spheroid chromatin units (v bodies). *Science* **183**: 330-332

Olins, A.L., Senior, M.B., Olins, D.E. (1976) Ultrastructural features of chromatin nu bodies. The *Journal of Cell Biology* **68**: 787-793

Ozsolak, F., Song, J.S., Liu, X.S., Fisher, D.E. (2007) High-throughput mapping of the chromatin structure of human promoters. *Nature Biotechnology* **25**: 244-248

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

Parthasarthy, A. and Gopinathan, K.P. (2006) Transcriptional activation of a moderately expressed tRNA gene by positioned nucleosome. *Biochemical Journal* **396**: 439-47

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

Pennings, S., Allan, J., Davey, C.S. (2005) DNA methylation, nucleosome formation and positioning. *Briefings in Functional Genomics and Proteomics* **3**: 351-361

Pruess, A. and Pikaard, C.S. (2007) rRNA gene silencing and nucleolar dominance: Insights into a chromosome-scale epigenetic on/off switch. *Biochimica et Biophysica Acta* **1769:** 383-92

Rabinowicz, P.D., Palmer, L.E., May, B.P., Hemann, M.T., Low, S.W., McCombie, R., Martienssen, R.A. (2003) Genes and transposons are differentially methylated in plants, but not in mammals. *Genome Research* **13**: 2658–2664

Raisner, R.M. and Madhani, H.D. (2006) Patterning chromatin: form and function for H2A.Z variant nucleosomes. *Current Opinion in Genetics and Development* **16**: 119-124

Rhodes, D. (1997) Chromatin structure. The nucleosome core all wrapped up. *Nature* **18**: 231-233

Richmond, T.J., Finch, J.T., Rushton, B., Rhodes, D., Klug, A. (1984) Structure of the nucleosome core particle at 7 Å resolution. *Nature* **17**: 532-537

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

Riddle, N.C. and Richards, E.J. (2002) The control of natural variation in cytosine methylation in Arabidopsis. *Genetics* **162**: 355-36

Robinson, P.J. and Rhodes, D. (2006) Structure of the '30 nm' chromatin fibre: a key role for the linker histone. *Current Opinion in Structural Biology* **16**: 336-343

Robinson, P.J., Fairall, L., Huynh, V.A., Rhodes, D. (2006) EM measurements define the dimensions of the "30-nm" chromatin fibre: evidence for a compact, interdigitated structure. *Proceedings* of *the National Academy of Sciences of the USA.* **103**: 6506-6511

Rose, A.B., Elfersi, T., Parra, G., Korf, I. (2008) Promoter-proximal introns in Arabidopsis thaliana are enriched in dispersed signals that elevate gene expression. *Plant Cell* **20**: 543-551

Routh, A., Sandin, S., Rhodes, D. (2008) Nucleosome repeat length and linker histone stoichiometry determine chromatin fibre structure. *Proceedings* of *the National Academy of Sciences of the USA.* **105**: 8872-8877

Salganik, R.I, Dudareva, N.A, Kiseleva, E.V. (1991) Structural organization and transcription of plant mitochondrial and chloroplast genomes. *Electron Microscopy Reviews* **4**: 221-247

Satchwell, S.C., Drew, H.R., Travers, A.A. (1986) Sequence periodicities chicken nucleosome core DNA. *Journal of Molecular Biology* **191**: 695-575

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

Satchwell, S.C. and Travers, A.A. (1989) Asymmetry and polarity of nucleosomes in chicken erythrocyte chromatin. *The EMBO Journal* **8**: 229-238

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.Y., Barski, A., Wang, Z., Wei, G., Zhao, K. (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**: 887-98

Schwabish, M.A. and Struhl, K. (2004) Evidence for eviction and rapid deposition of histones upon transcriptional elongation by RNA polymerase II. *Molecular and Cell Biology* **24**: 10111-10117

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

Saleh, A., Alvarez-Venegas, R., Yilmaz, M., Le, O., Hou, G., Sadder, M., Al-Abdallat, A., Xia, Y., Lu, G., Ladunga, I., Avramova, **Z.** (2008) The Highly Similar *Arabidopsis* Homologs of Trithorax ATX1 and ATX2 Encode Proteins with Divergent Biochemical Functions. *Plant Cell* **20**: 568-579.

Seoighe, C., Gehring, C., Hurst, L.D. (2005) Gametophytic selection in Arabidopsis thaliana supports the selective model of intron length reduction. *PLoS Genetics* **1**: e13

Sheather, S.J. and Jones, M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B.* **53**: 683-690

Shindo, C., Lister, C., Crevillen, P., Nordborg, M., Dean, C. (2006) Variation in the epigenetic silencing of FLC contributes to natural variation in Arabidopsis vernalization response. *Genes and Development* **20**: 3079-3083

Shivaswamy, S., and Bhargava, P. (2006) Positioned nucleosomes due to sequential remodelling of the yeast U6 small nuclear RNA chromatin are essential for its transcriptional activation. *Journal of Biological Chemistry* **281**: 10461-10472

Shrader, T.E., and Crothers, D.M. (1989) Artificial nucleosome positioning sequences. *Proceedings* of *the National Academy of Sciences of the USA.* **86**: 7418-7422

Smith, T.F., and Waterman, M.S. (1981) Identification of Common Molecular subsequences. *Journal of Molecular Biology* **147**: 195–197

Staynov, D.Z. and Proykova, Y.G. (2008) The sequentiallity of nucleosomes in the 30 nm chromatin fibre. *FEBS Journal* **275**: 3761-3771.

Stein, A. and Bina M. (1999) A signal encoded in vertebrate DNA that influences nucleosome position and alignment. *Nucleic Acids Research* **27**: 848-853

Strohner R., Wachsmuth, M., Dachauer, K., Mazurkiewicz, J., Hochstatter, J., Rippe, K., Längst, G. (2005) A 'loop recapture' mechanism for ACF-dependent nucleosome remodelling. *Nature Structural and Molecular Biology* **12**: 683-690

Tanaka, S., Zatchej, M., Thoma, F. (1992) Artificial nucleosome positioning sequences tested in yeast minichromosomes: a strong rotational setting is not sufficient to position nucleosomes in vivo. *The EMBO Journal* **11**: 1187-1193

Tanaka, T., Koyanagi, K.O., Itoh, T. (2009) Highly diversified molecular evolution of downstream transcription start sites in rice and Arabidopsis. *Plant Physiology* **149**: 1316-1324

Tariq, M., Saze, H., Probst, A.V., Lichota, J., Habu, Y., Paszkowski, J. (2003) Erasure of CpG methylation in Arabidopsis alters patterns of histone H3 methylation in heterochromatin. *Proceedings* of *the National Academy of Sciences of the USA.* **100**: 8823-8827

Tchurikov, N. A. (2005) Molecular mechanisms of epigenetics. *Biochemistry* (*Moscow*) **70**: 406-423

Tienari, P.J., Kuokkanen, S., Pastien, T., Wikstrom, J., Sajantila, A., Sandberg-Wollheim, M., Palo, J., Peltonen, L. (1998) Golli-MBP gene in multiple sclerosis susceptibility. *Journal of Neuroimmunology* **81**: 158-167

Ull, M.A. and Franco, L. (1986) The nucleosome repeat length of pea (*Pisum savitum*) chromatin changes during germination. *Plant Molecular Biology* **7**: 25-32

Unfried, I. and Gruendler, P. (1990) Nucleotide sequence of the 5.8S and 25S rRNA genes and of the internal transcribed spacers from Arabidopsis thaliana. *Nucleic Acids Research* **18**: 4011

Villar, C.B.R., Erilova, A., Makarevich, G., Trösch R., Köhler, C. (2009) Control of PHERES1 Imprinting in Arabidopsis by Direct Tandem Repeats. *Molecular Plant* **2**: 654-660

Virstedt, J., Berge, T., Henderson, R.M., Waring, M.J., Travers, A.A. (2004) The influence of DNA stiffness upon nucleosome formation. Journal of Structural Biology **148**: 66-85

Vignali, M., Hassan, A.H., Neely, K.E., Workman, J.L. (2000) ATP-dependent chromatin remodelling complexes. *Molecular and Cell Biology* **20**:1899-910

Vinogradov, A.E. (2005) Non-coding DNA, isochores and gene expression: nucleosome formation potential. *Nucleic Acids Research* **33**: 559-63

Waddington, C.H. (1953) Genetic Assimilation of an Acquired Character. *Evolution* **7:** 118-126

Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Düsterhöft, A., Stiekema, W., Entian, K.D., Terryn, N., Harris, B., Ansroge, W., Brandt, P., Grivell, L., Rieger, M., Weichselgartner, M., de Simone, V., Obermaier, B., Mache, R., Müller, M., Kreis, M., Delseny, M., Puigdomenech, P., Watson, M., Schmidtheini, T., Reichert, B., Portatelle, D., Perez-Alonso, M., Bountry, M., Bancroft, I., Vos, P., Hoheisel, J., Zimmermann, W., Wedler, H., Ridley, P.,

Langham, S.A., McCullagh, B., Bilham, L., Robben, J., Van der Schueren, J., Grymonprez, B., Chuang, Y.J., Vandenbussche, F., Braeken, M., Weltjens, I., Voet, M., Bastiens, I., Aert, R., Defoor, E., Weitzenegger, T., Bothe, G., Rose, M. (2000) Progress in Arabidopsis genome sequencing and functional genomics. *Journal of Biotechnology* **78**: 281-292

Wang, J.P., Fondufe-Mittendorf, Y., Xi, L., Tsai, G.F., Segal, E., Widom, J. (2008) Preferentially quantized linker DNA lengths in Saccharomyces cerevisiae. *PLoS Computational Biology* **4**: e1000175

Wendel, J.F., Cronn, R.C., Alvatez, I., Liu, B., Small, R.L., Senchina, D.S. (2002) Intron size and genome size in plants. *Molecular Biology and Evolution* **19**: 2346-2352

Weintraub, H. (1978) The nucleosome repeat length increases during erythropoiesis in the chick. *Nucleic Acids Research* **5**: 1179-1188

Widlak, P. and Garrard, W.T. (2006) Unique features of the apoptotic endonuclease DFF40/CAD relative to micrococcal nuclease as a structural probe for chromatin *Biochemistry and Cell Biology* **84**: 405-410

Widlund, H., Cao, H., Simonsson, S., Magnusson, E., Simonsson, T., Nielsen, P., Kahn, J., Crothers, D., Kubista, M. (1997) Identification and characterisation of genomic nucleosome-positioning sequences. *Journal of Molecular Biology* **267**: 807-817

Widom, J. (1992) A relationship between the helical twist of DNA and the ordered positioning of nucleosomes in all eukaryotic cells. *Proceedings* of *the National Academy of Sciences of the USA.* **89**: 1095-1099

Wierzbicki, A.T., and Jerzmanowski, A. (2005) Suppression of histone H1 genes in Arabidopsis results in heritable developmental defects and stochastic changes in DNA methylation. *Genetics* **169**: 997-1008

Winston, F. and Carlson, M. (1992) Yeast SNF/SWI transcriptional activators and the SPT/SIN chromatin connection. *Trends in Genetics* **8**: 387-391

Wong, H., Victor, J.M., Mozziconacci, J. (2007) An all-atom model of the chromatin fibre containing linker histones reveals a versatile structure tuned by the nucleosomal repeat length. *PLoS One* **2**: e877

Woo, H.R. and Richards, E.J. (2008) Natural variation in DNA methylation in ribosomal RNA genes of Arabidopsis thaliana. BMC *Plant Biology* **8**:92

Woodcock, C.L. and Horowitz, R.A. (1997) Electron microscopy of chromatin. *Methods* **12**: 84-95

Woodcock C.L., Skoultchi, A.I., Fan, Y. (2006) Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length *Chromosome Research* **14**: 17-25

Workman, J.L. (2006) Nucleosome displacement in transcription. *Genes and Development* **20**: 2009-2017

Yuan, G.C., Liu, Y.J., Dion, M.F., Slack, M.D., Wu, L.F., Altscluler, S.J., Rando, O.J. (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* **309**: 626-630

Yang, J.G., Madrid, T.S., Sevastopoulos, E., Narlikar, G.J. (2006) Chromatin remodelling enzyme ACF is an ATP-dept DNA length sensor that regulated nucleosome spacing. *Nature Structural and Molecular Biology* **13**: 1078-83

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E., Ecker, J.R. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell* **126**: 1189-1201

Zhang, L, Wu, C., Carta, R., Zhao, H. (2007) Free energy of DNA duplex formation on short oligonucleotide microarrays *Nucleic Acids Research* **35**: e18.

Zhang, Y, Shin, H., Song, J.S., Lei, Y., Liu, X.S. (2008). Identifying Positioned Nucleosomes with Epigenetic Marks in Human from ChIP-Seq. *BMC Genomics* **9**: 537

Zilberman, D. and Henikoff, S. (2005) Epigenetic inheritance in Arabidopsis: selective silence. *Current Opinion in Genetics and Development* **15**: 1-6

Zilberman, D., Coleman-Derr, D., Ballinger T., Henikoff S. (2008) Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. *Nature* **456**: 125-129